Contents

Chapter 1

Genetics Background

1.1 Basic Principles of Genetics

Reference: [Human Molecular Genetics, 4th Ed, Chapter 3], [Thomas, Statistical Methods in Genetic Epidmeiology, 2004, Chapter 3]

Chromosome structure:

- Cytogenetic map [Wiki]: Diagrams identifying the chromosomes based on the banding patterns are known as cytogenetic maps. Notations: p for short arm, and q for long arm, numbers indicat band, sub-band and sub-sub-band.
- Sex chromosomes:
  - Y-chromosome carries the gene sex determining region Y (SRY), the presence of which makes an embryo develop into a male.
  - X-chromosome recombines only in females. There are two regions on the human X chromosome that are homologous with and recombine with the Y chromosome. These are called pseudoautosomal regions (PAR) because they act like autosomes. The larger (PAR1) is at the tip of the short arm; it is 2.6 Mb long and contains 14 genes. The smaller (PAR2) region on the tip of the long arm is only 320?kb long and contains four genes.
  - Size of X-chr: 165 Mb, about 5% of human genome. Estimates of the number of genes on the X chromosome range from 600 to 1500. It has been claimed that the X is particularly rich in genes involved in intelligence and in sex and reproduction.

Chromosome abnormalities:

- If chromosome translocation or inversion happens that disrupts the disease gene, then it will lead to disease. Thus for patients with chrosome abnormalities in cytogenetic data, search for the breakpoints, which may reside in or close to the candidate genes.
- Large chromosome deletions: often delete multiple genes, thus the patient may experience multiple genetic disorders, or one disorder and mental retardation (many deletion could lead to mental retardation because of the large number of genes involved in fetal brain development).
- Array CGH: genome-wide meaurement of gene copy number amplifications or deletions.
- **Remark**: because chromosome abnormalities are rare, thus it is likely that when chromosome abnormality occurs in a diseaes patient, there is a causal relation between the two.

Principles of Mendelian Inheritance:

- Segregation of alleles: each person carries two copies of each gene, one from each parent. Alleles at any given gene are transmitted randomly and with equal prob.
- Independent assortment: the alleles of different genes (not linked) are transmitted independently.
- The expression of two copies of a gene is indpendent of which parent they come from. The exception is imprinting.

Hardy-Weinberg Equilibrium (HWE):

- HEW in sex-linked loci: if a gene is X-linked, let the allele frequencies be $f(A) = p$ and $f(a) = q$. Then in males, the frequencies of A and a would be p and q. In females, following HWE, the frequencies of AA, Aa and aa would be: $p^2, 2pq$ and q^2.
- HWE in general does not hold for loci associated with diseases. Thus deviation from HWE in cases of diseases is often taken as preliminary evidence of disease association.

Monogenic vs. multifactorial diseases:

- A spectrum of diseases, from monogenic to multifactorial/polygenic. In the intermediate, e.g. multiple loci with one making the major contribution. Note that the genetics of molecular traits (e.g. enzyme activity, gene expression) may be close to Mendelian traits.
- Dichotomous traits: suspectibility genes/loci. Continuous traits: eQTL. A trait that could be Mendelian or polygenic is called a complex trait.

Patterns of Mendelian inheritance in pedigrees: note that males are hemizygous for loci in X or Y chromosomes.

- Symbols for pedegrees: (1) Main symbols: males, females, mating, affection status. (2) Roman numbers for generations and Arabic numbers for individuals. (3) Other features: consanguineous mating, carrier status, probands (through which the family was ascertained.
- Autosomal dominant: an affected person usually has at least one affected parent.
- Autosomal recessive: an affected persons are usually born to unaffected parents.
- X-linked recessive: affects mainly males.
- X-linked dominant: an affected father will only transmit to daughters but not sons.
- Y-linked: affect only males.
- Genes in the pseudoautosomal region (2.6Mb homologous regions between X and Y): inheritance pattern is similar to autosomal genes.
- Mitochondrial inheritance: always transmitted from mothers (however, could be de novo mutations from unaffected mother). Also often highly variable clinical manifestation (a cell contains multiple mitochondrial and it is possible that some are normal).

X-inactivation:

- In females, one of the two X-chromosomes is inactivated in each cell, and remains so in the entire life. Occurs probably in 10-20 cell stage.
- Phenotypic consequence of X-inactivation (mosaicism): (1) if the trait is highly localized (e.g. sweat gland/hair), female carriers may show patches of normal and abnormal tissues. (2) For some traits, the cells expressing the mutant copy may die, as a result, all surviving cells in females may be normal (e.g. mature B cells). (3) For X-linked recessive diseases, females occasionally get severely affected because of X-inactivation in critical cells. (4) X-linked dominant diseases: females often show variable phenotypes because some cells express normal X.

- An explanation of IQs in males and females based on X-inactivation:

  http://www.economist.com/news/science-and-technology/21568704-geniuses-are-getting-brighter-and-ger
  It is easier for males to win the IQ lottery if they happen to get a single very strong version of the relevant genes on their X-chromosome. All female mammals are mosaics with respect to the X-chromosome. For a woman to have that same gene in all of her tissues, she would have to inherit the rare, very strong allele/s from both of her parents, which is statistically far less likely to occur. Similarly, males are also far more likely to lose this lottery, and wind up mentally retarded.

Complications to the basic Mendelian inheritance patterns:

- Locus heterogeneity: common in syndromes that result form a failure of a complex pathway. Ex. profound hearing loss, when two disease carriers marry, the children is often normal (complementation).
- Incomplete penetrance: define **penetrance** of a genotype, G, as $P(D|G)$, i.e. the probability of disease of an individual with genotype G. Not all diseases have 100% penetrance, especially for dominant diseases.
- Age-related penetrance in late onset diseases: if a patient dies before the onset of the disease, then he/she may show as normal in the pedigree, complicating the analysis.
- Variable expression: may depend on the genotypic background and environment.
- Imprinting: for some autosomal diseases, the disease gene is inherited only from the father or the mother. Epigenetic changes in father or mother.
- Male lethality may complicate X-linked pedigree: e.g. Rett Syndrome. Affected males are often not born (miscarriage).
- De novo mutations: for diseases that prevent affected people from reproducting but new cases keep occuring, it is likely that a significant fraction is de novo mutations. Severe X-linked recessive diseases may also show a large fraction of de novo mutations.
- Mosaicism: if a mutation occurs early in life, then a significant fraction of cells may express the abnormal gene. When a large number of germ-line cells are mutated, this could produce multiple affected offsprings - the pedigree would mimic the recesive inheritance.

Characterizing familial clustering of diseases: not all familial clustering is due to genetic factors, thus it is important to distinguish the two concepts

- Sporadic vs. familial cases: sporadic means isolated cases without a family history, and familial cases refer to cases with a positive family history.
- Hereditary cases: among the cases with a strong family history, those with a clear evidence of genetic factors.
- Example: breast cancer cases with genotyped BRCA1 and BRCA2. The majority of cases are sporadic. Among most familial cases, most are not hereditary.

Polygenic threshold theory: [Human Molecular Genetics, 4rd Ed, Chapter 3]

- History: debate betweeen Mendelians and biometricians (Francis Galton). A paper by R.A. Fisher in 1918 settled the debate: characters governed by a large number ofindependent Mendelian factors (polygenic characters) would display precisely the continuous nature, quantitative variation, and family correlations described by the biometricians. DS Falconer extended this model to cover dichotomous characters.

- Polyngenic theory: if a trait is determined by many loci (the sum of effects of all loci), then the trait distribution is approximately normal. The goal of this theory is to understand the influence of genetic and environmental factors, and the pattern of inheritance. However direct regression analysis is difficult, because neither genotype nor environmental variables are directly measured.

- Regression to the mean: one main observation in quantitative trait studies is: the offsprings of the individuals with extreme phenotypes, tend to have traits in the intermediate between the parent and the population mean. This is called "regression to the mean". This can be explained by that the parent may have many alleles that have extreme values (comparing with population mean), but because of random mating, the other parent has a smaller number of such alleles.

- Dicontinuous characters: assume there is an underlying continuous variable: susceptibility, and that there exists a threshold - when the suspectibility of an individual is higher than the threahold, the disease may develop. Susceptibility is a RV for individuals, similar to a quantitative character, while the liability threshold is fixed for a group. To model the risks: let t be the threshold relative to the population mean in the unit of standard deviation (i.e. the Z score), then the risk is:

$$P(X \geq t) = 1 - \Phi(t) \tag{1.1}$$

  where $\Phi(\cdot)$ is the CDF of normal distribution. Suppose we have a mutation in a gene, the individuals carrying this mutation have a different distribution of susceptibility - suppose its mean is μ and its standard deviation is the same (the mean is 0 for people without the mutation). Let Z_μ be the Z-score of the mean (i.e. measure of μ with the unit of standard deviation), then the risk of the mutant group is:

$$P(X \geq t) = P(X' \geq t - Z_\mu) = 1 - \Phi(t - Z_\mu) \tag{1.2}$$

  The relative risk is thus:

$$\text{RR} = \frac{1 - \Phi(t - Z_\mu)}{1 - \Phi(t)} \tag{1.3}$$

  For example, at $t = 2$, the risk is about 2.3%, and suppose $Z_\mu = 0.5$, the risk is 6.7%, and the RR about 2.9.

- Model different risks in different groups: e.g. some diseases have higher risk in one gender than the other. This can be modeled by allowing different thresholds in different genders.

- Genetic model: suppose we have genotypes, AA, Aa and aa. Typically, we model the effect of a in the liability scale as α, then the mean of the three genotypes are: $-\mu$, $\alpha - \mu$ and $2\alpha - \mu$, where μ is chosen so that the population mean is 0. It can be shown that $\mu = 2\alpha p$, where p is the allele frequency of a.

- Q: what is the relationship between the polygenic threshold model and probit model for binary response?

Penetrance function and genetic model:

- Penetrance function: defined as $P(Y|G)$, where Y is phenotype and G genotype. Note that the penetrance function takes different forms depending on the phenotype. If the phenotype is binary, the penetrance is multinomial; if the phenotype is quantitative, the penetrance represents multiple continuous (e.g. normal) distributions.

- To characterize the penetrance function, we define different genetic models. Let A be the wild type and a be the mutant allele, then dominant means $P(Y|Aa) = P(Y|aa)$, i.e. a single copy of a is sufficient to increase the risk; recessive means $P(Y|Aa) = P(Y|AA)$, i.e. need two copies of a to increase the risk; additive (or multiplicative, depending on the scale) means $P(Y|Aa)$ is intermediate between $P(Y|AA)$ and $P(Y|aa)$. Codminant: means $P(Y|AA) \neq P(Y|Aa) \neq P(Y|aa)$.

Genetic model of binary (dichotomous) traits: [UWash, b516, 2009] consider a locus with two alleles A_1 and A_2 (suppose A_2 increases the risk of diseaes), let q_1, q_2 be the frequency of the two alleles respectively.

- Penentrance: f_{ij} is the probability of disease of the genotype A_iA_j.
- Population prevalance of disease: also the weighted average of diseaes penetrance:

$$K_p = q_1^2 f_{11} + 2q_1q_2f_{12} + q_2^2 f_{22} \tag{1.4}$$

- Genotype relative risk: the risk of disease of a genotype relative to the population average: $R_{ij} = f_{ij}/K_p$.
- Dominant model: $f_{11} = k$, $f_{12} = f_{22} = k + c$. Recessive model: $f_{11} = f_{12} = k$, $f_{22} = k + c$.
- Additive model: $f_{11} = k$, $f_{12} = k + c$, $f_{22} = k + 2c$.
- Multiplicative model: $f_{11} = k$, $f_{12} = rk$, $f_{22} = r^2k$.

Genetic model of quantitative traits: consider a locus with 2 alleles Q_1 and Q_2. Typically, assume the trait of Q_1Q_1 (reference) is 0, and then the trait of Q_1Q_2 and Q_2Q_2 are $a+d$ and $2a$ respectively (or $a(1+k)$ and $2a$). We have: $k = d/a$, which measures the amount of dominance. Depending on the value of k, we have some special cases:

- $k = 0$: no dominance;
- $k = 1$: complete dominance;
- $k = -1$: recessive;
- $k > 1$: overdominance.

Population genetics of Mendelian diseases:

- Model: apply the theory of selection-mutation balance, the frequency of the disease allele is related to the mutation rate and the selection coefficient: $\hat{q} = \sqrt{\mu/s}$ for recessive traits; and $\hat{q} = \mu/(hs)$ for dominant traits (h is the degree of dominance). This allows one to estimate the mutation rate of diseases (e.g. for lethal diseases, $s = 1$).
- The explanations of polymorphism in disease-causing alleles for Mendelian diseases: high mutation rate, late onset of symptoms (e.g. Huntington disease), heterozygote advantage (e.g. for cyarix fibrosis, mutation in membrance chloride channel, heterozygotes may be more resistent to certain infections, similar to sickel-cell disease).

1.2 Genetic Epidemiology Overview

Reference: [Thomas, Statistical Methods in Genetic Epidmeiology, 2004, Chapter 1, 3, 4]

Problems of complex trait genetics:

- Technonology and analytic methods for detecting the loci of complex traits.
- Genetic architecture of complex traits: genetic vs non-genetic contribution, number of loci and effect size distribution, common vs rare variants, the comparison of genetic loci of related traits, etc.
- Mechanism of complex trait genetics: how the polymorphism at the loci affects the trait: coding vs non-coding (or cis- vs trans-), types of genes, the mechanism of hot-spots (pleotropy), etc.
- Application of the results of complex trait mapping: prediction of phenotypes, explanation of other complex traits, etc.

- Evolutionary forces on complex traits: e.g. whether alleles are under natural selection.

Genetic epidemiology background [Thomas04, Chapter 1]: an overview

- Descriptive epidemiology: forming the hypothesis of whether a trait/disease has a genetic component, by correlating trait with the properties that might represent genetics (e.g. race).
- Aggregation analysis: using family data to test if a trait is genetic, or "run in families".
- Segregation analysis: fit the genetic model using family model, how many genes are involved (major gene vs. polygene), dominant or recessive, etc.
- Linkaging analysis: within families, co-segregation of traits and markers (which are close to causal loci).
- Association mapping and functional studies.

Descriptive epidemiolgy:

- Principle 1. correlation of phenotypic traits and genetics. Typically, genes are not observed, we use "markers" that represent genetic similarity of groups, such markers often represent common ancestry, such as country of origin, race, etc.
- Principle 2. the key challenge is the confounders such as environmental factors, culture and history, etc. Find the type of data where one (gene or environment) is controlled, e.g. migration studies (genes controlled).
- Background: incidence rate is the number of people caught a disease per year, often measured using person-years. It is different from prevalence.
- Variation of incidence rates across countries/regions: one of the most common studies of descriptive epidemiology. Ex. breast cancer rates vary nearly 20 fold across countries.
  - Migration studies: the obvious problem with the country comparison is that many things are different. To control for that, study the incidence of migrants.
- Race and ethnicity: [Risch, 2002] argues the use of self-identified ethnicity in genetic epid. rather a purely genetic definition that would not encompass cultural factors that may be relevant. A special case of studies involving race is admixture studies, e.g. in African-Americans or Latinos. In particular, with such group, test the correlation between traits and the degree/gradience of certain genetic heritage. The advance of admixture studies is that the environmental factors are better controlled.
- Age effects: if a trait is genetic, then its age of onset tends to be earlier. Ex. comparison of age of onset in familial form vs. sporadic form of diseases.
- Time trends: because genetics cannot change in a short time, the changes of incidence in time are generally due to environmental factors.

Family aggregation: the question is how to test if a disease "runs in families"?

- Case-control studies: if a disease "runs in families", then if one member gets the disease, it means that the chance that other members may also have the disease is higher. Two specific statistical strategies:
  - Randomly sample cases and controls (matched), obtain their families and compare the incidence of case-families vs. control-families.
  - Treat the family history of case/control as "exposure risk", and test if the exposure increases the disease risk using the standard case-control comparison.

- Variance components method: for quantitative trait, statistical analysis of variance due to shared environmental or genetics.
- Twin studies: comparison of MZ and DZ twins. DZ twins are different genetically, but environment is controlled.
- Adoption studies: could be done in either ways:
  - Same gene, different environment: MZ twins but raised in separate families.
  - Same environment, different gene: biological child vs. adopted child in the same families.
- Inbreeding: correlation of incidence with the degree of inbreeding.
- **Lesson**: a major problem is to separate genetic and environmental effects, and this can be addressed to a large extent by the study designs. The easiest one is the case-control comparison where the two groups are matched. More generally, twin studies, adoption studies, admixture/inbreeding stuides all try to control one and study the effect of the other.

Other steps in genetic epidemiology studies:

- Segregation analysis: the goal is to fit the genetic model - single gene vs polygene model, dominant or recessive, etc. From the model, one can estimate the segregation parameters such as allele frequency and penetrance.
- Linkage analysis: the difference with the segregation analysis that both disease loci and markers need to be modeled.
- Fine mapping: the regions identified from linkage studies are usually broad, use LD mapping to locate the segments shared by multiple haplotypes in cases.
- Association stuides: in addition to the standard case-control, association studies on families can reduce the problem of population substructure. Two common designs: case-sibling design (unaffected siblings as controls) and trios (TDT).

Study designs in epidemiology:

- Cohort design: choose a cohort, record the exposures, and then do follow-up. Advantage: less bias (see below). Disadvantage: may require large sample size, since only a small fraction in general will become cases.
- Case-control design: from cohort, find cases and controls, then compare risk factors (in the past). To choose cases and controls:
  - Nested case-control design: for each case, find a best matched control: random sampling from the set of subjects who are at risk at the time the case occurred.
  - Case-cohort design: the controls are a random sample of the entire cohort, irrespective of whether or not they later become cases. The advantage: the same controls can be used for multiple case groups.

Common issues/biases in epidemiology studies:

- Selection bias: the selected samples are not representative of the entire population of interest. Ex. the controls may come from a particular geographic region that is different from cases in some unmeasured ways.
- Information bias: the information in cases and controls is somehow not comparable. Ex. recall-bias: cases and controls recall information in the past (risk factors) differently.

- Confounding: if a variable (1) is associated with the exposure of interest, and (2) conditional on the exposure, an independent risk factor of the disease, then this variable is a confounder.

Modeling disease risks:

- Motivation: the disease risk typically changes over time/age, thus we cannot use the standard Poisson process.
- Defintions: hazard and survival functions. Let $Y(t)$ be the status at time t, the absolute risk can be written as: $F(t) = P(Y(t) = 1|Y(0) = 0)$. The hazard function is similar to Poisson rate:

$$\lambda(t) = \lim_{dt \to 0} P(Y(t + dt) = 1|Y(t) = 0)/dt \tag{1.5}$$

  The survival function is $S(t) = 1 - F(t) = P(Y(t) = 0|Y(0) = 0)$.
- Relationship between hazard and survival functions: the total rate from 0 to t is the integration of $\lambda(u)$, and the survival function follows exponential distribution, so we have:

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right) \tag{1.6}$$

  On the other hand, the probability of having an event in $(t, t + dt)$ can be expressed in two ways. (1) No event up to t, followed by an event in $(t, t + dt)$, so this is $\lambda(t)S(t)dt$. (2) This is also: $P(Y(t + dt) = 1|Y(0) = 0) - P(Y(t) = 1|Y(0) = 0) = dF(t) = -dS(t)$. So we have:

$$\lambda(t)dt = -\frac{dS(t)}{S(t)} \tag{1.7}$$

  The relationship could also be derived using Fundamental Theorem of Calculus.
- Statistical problem of estimating hazard and survival function: our data would be time of cases for many subjects (or at multiple time points, the status of all subjects). The simplest strategy is to have multiple time windows, and the hazard at each window is constant, and can be computed simply by the frequency of cases. More complex statistical analysis is possible.

Modeling relative risks:

- Case-control design: the data is a 2 by 2 table: based on exposure status and affection status. Let the four entries b A, B, C, D respectively, then the simplest test is the χ^2 test or Fisher's exact test (for small samples). The odds ratio is, following definition:

$$OR = \frac{D/C}{B/A} = \frac{AD}{BC} \tag{1.8}$$

  Alternatively, it can be interrpeted as the ratio of frequency of exposure in cases vs. controls. To obtain the confidence interval, we have:

$$\text{Var}\left(\log OR\right) = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D} \tag{1.9}$$

  This follows from binomial distributions. The intuition is that the variance is dominated by rare events.
- Matched case-control design: equal number of cases and controls. We define a different 2 by 2 time: the exposure status in cases and in controls, and we test the null hypothesis if the diagonal elements (number of exposure in cases and controls) are equal. This is done by McNemar test.

- Cohort design: use the standardized incidence ratio (SIR). Let z be the exposure variable, s be other risk factors (e.g. age). We could define the hazard ratio as:

$$\gamma_s = \frac{\lambda_s(z=1)}{\lambda_s(z=0)} \tag{1.10}$$

The expected number of cases under null model is: $E_z = \sum_s \lambda_s T_{zs}$, where λ_s is the baseline rate and T_{zs} is person-year. And the expected number of cases under alternative model is: $\mathrm{E}(Y_s) = \sum_s \gamma_s \lambda_s T_{zs}$. This motivates the definition of SIR: when γ_s is a constant, we have: $\hat{\gamma} = Y_z/E_z$.

Modeling modifers/interactions:

- Single explanatory variable: suppose we have a binary phenotype y and an explanatory variable G (e.g. genotype). We use the logistic regression:

$$\log \frac{P(y=1|G)}{P(y=0|G)} = \beta_0 + \beta_1 G \tag{1.11}$$

Plug in $G=0$ and $G=1$, we have:

$$\log -\mathrm{odds}(G=0) = \beta_0 \qquad \log -\mathrm{odds}(G=1) = \beta_0 + \beta_1 \tag{1.12}$$

Subtracting the two: $\beta_1 = \log -\mathrm{OR}(G=1)$, thus β_1 has a simple interpretation: the log odds ratio of $G=1$.

- Modifer variable: suppose we have an additional variable E, we have the model:

$$\log \frac{P(y=1|G,E)}{P(y=0|G,E)} = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 G \cdot E \tag{1.13}$$

To see what the parameters mean, first, we plug in different combinations of G and E, and use $\log -\mathrm{odds}(G=0, E=0)$ as the baseline odds:

$$\log -\mathrm{OR}(G=1,E=0) = \beta_1, \log -\mathrm{OR}(G=0,E=1) = \beta_2, \log -\mathrm{OR}(G=1,E=1) = \beta_1 + \beta_2 + \beta_3 \tag{1.14}$$

From this, we see that β_3 represents the non-additive effect on the odds ratio:

$$\beta_3 = \log -\mathrm{OR}(G=1,E=1) - [\log -\mathrm{OR}(G=1,E=0) + \log -\mathrm{OR}(G=0,E=1)]. \tag{1.15}$$

Alternatively, β_3 has another interpretation representing the modification of the effect of G due to E. To see this, we use the conditional OR (the baseline is defined under the condition):

$$\log -\mathrm{OR}(G=1|E=0) = \beta_1 \quad \log -\mathrm{OR}(G=1|E=1) = \beta_1 + \beta_3 \tag{1.16}$$

Note that the later OR is $\mathrm{odds}(G=1,E=1)/\mathrm{odds}(G=0,E=1)$. Thus we have:

$$\beta_3 = \log -\mathrm{OR}(G=1|E=1) - \log -\mathrm{OR}(G=1|E=0). \tag{1.17}$$

Thus β_3 measures the strength of the modifier effect.

Matched design: conditional logistic regression [Thomas, chapter 4, page 85]:

- Problem: suppose we are testing the effect of Z on a response variable Y. However, there are potentially other confounding variables that might modify the risk Y. To deal with the problem, we use a 1:1 matched case-control, where each case and the matched control are identical in all aspects except that Z is different. We can prove that under this matched design, even though the baseline risk of each pair is different (reflecting the other variables), the likelihood doesn't depend on these baseline risks, so they can be ignored.

- Model: consider the i-th pair, let Y_{i1} and Y_{i0} be the response variables for case and control, and let Z_{i1} and Z_{i0} for the explanatory variables. Let α_i be the baseline risk of the i-th pair, and β the main effect to be estimated. We can show that:

$$L(\beta) = \prod_i P(Y_{i1} = 1 | Z_{i1}, Z_{i0}, Y_{i0} + Y_{i1} = 1) = \frac{\exp(Z_{i1}\beta)}{\exp(Z_{i0}\beta) + \exp(Z_{i1}\beta)} \tag{1.18}$$

Generalized Estimating Equation (GEE) approach to correlated data [Robert Weiss' lecture, UCLA, Biostatistics 411]:

- Motivation: we may have correlated data, examples: (1) longitude data: we have N subjects, and for each subject, we may take n measurements of a variable, clearly, the errors of these measurements tend to be correlated; (2) family data: we have N families, and within each family, the random error terms (e.g. of a quantitative trait) may be correlated - due to shared environment, etc. However, instead of modeling the full likelihood, we want an approach that is robust to the error distributions.

- Background: Estimating Equation (EE) for parameter estimation. For the GLS problem (see Statistics Notes, "Generalized Least Squares"), the EE can be written as:

$$X^T \Sigma^{-1} (y - X\beta) = 0 \tag{1.19}$$

  It has three components: (1) X^T: this comes from $d\mu/d\beta$ (the chain rule), where $\mu = X\beta$ for linear model is the mean of Y; (2) Σ^{-1}; (3) the residual $Y - X\beta$. This motivates the GEE approach: even when the likelihood is not satisfied, we solve this form of equation to estimate β.

- Marginal model: we have data Y_{ij}, where $1 \leq i \leq N$ represents families/subjects, and j represents individual observation within families/subjects. The mean, $\mu_{ij} = \mathrm{E}(Y_{ij})$ depends on explanatory variables X_{ij} through a link function: $g(\mu_{ij}) = X_{ij}\beta$. The marginal variance:

$$\mathrm{Var}(Y_{ij}) = v(\mu_{ij})\phi \tag{1.20}$$

  where $v(\cdot)$ is the variance function, and assumed known, and ϕ is a constant (1 for Poisson, Bernoulli; may need to be estimated). Correlation between Y_{ij} and Y_{ik} may depend on extra parameter α and means μ_{ij} and μ_{ik}. The idea is to use the three components from the marginal model in the EE equation above (for each family):

$$\sum_{i=1}^{N} \frac{d\mu_i}{d\beta} V_i^{-1} (Y_i - \mu_i(\beta)) = 0 \tag{1.21}$$

  where V_i is the "working covariance matrix". Solving this equation gives the GEE estimator, $\hat{\beta}_{\mathrm{GEE}}$. The working covar. matrix can be constructed as:

$$V_i = A_i + \mathrm{Corr}(Y_i) \tag{1.22}$$

  where A_i is the diagonal matrix of the marginal variance, and $\mathrm{Corr}(Y_i)$ is the correlation matrix, dependent on the parameter α. The variance of $\hat{\beta}_{\mathrm{GEE}}$ can also be analytically determined (sandwich estimator).

- Extensions in genetics: we may be interested in estimating the parameters in the covariance, thus we need to extend GEE to higher-order moments. For instance, let C_i be the observed covariance within families, and we are interested in estimating the parameters of the expected C_i: $\Sigma_i(\alpha) = \mathrm{E}(C_i|\alpha)$. Use the GEE:

$$\sum_{i=1}^{N} \frac{d\Sigma_i}{d\alpha} W_i^{-1} (C_i - \Sigma_i(\alpha)) = 0 \tag{1.23}$$

  In genetics, we may need to solve both GEE (in terms of mean and covariance) simultaneously.

1.3 Aggregation and Segregation Analysis

Reference: [Human Molecular Genetics, Chapter 15]

Aggregation analysis: deciding whether a non-Medelian character is genetic ("run-in-families"):

- Family clustering: genetic disease tends to occur in families. Measured by λ_R, the risk to relative R of an affected person (the entire population should be 1).
- Shared family environment: a confouding variable, family clustering alone does not necessarily suggest genetic contribution. To estabilish genetic cause (especially for behavior disorders), may need twin studies or adoption studies.

Segregation analysis: the mode of inheritance - Mendelian, oligogenic or polygenic, etc.

- Test: propose a genetic model (the number of loci, penetrance, etc.), compute the segregation ratio (the fraction of each phenotype) and compare the predicted vs and the observed ratio. Testing of statistical significance can be done with binomial/multinomial distributions.
- Example: three phenotypes with the genetic model (full penetrance): $C*G*$ - grey, $C*gg$ - black, $CC**$ - albino, and F1 is $CcGg$, then the segregation ratio is, grey : black : albino $= 9:3:4$.
- Ascertainment bias: this is a problem when only families with diseases are genotyped. Thus the true distribution is truncated binomial, and need correction (e.g. will not be $3:1$ for simple Mendelian traits).

1.3.1 Aggregation Analysis

Reference: [Thomas, Chapter 5]

Genetic relationship between relatives:

- IBD: for a pair of individuals, the probability of sharing 0,1 or 2 alleles IBD: $\pi_0, \pi_1, \pi_2(\nabla)$. The expected number of alleles sharing IBD is $\bar{\pi} = \pi_1 + 2\pi_2$ is called the coefficient of relationship.
- Kinship coefficient (ϕ): for a pair of individuals, the probability that a randomly selected pair of alleles, one from each individual, is IBD. Ex. for a sib pair, it is 1/4 (50% the pair is from the same parent, and 50% they are identical).
- The path method to determine the kinship coefficient: find all paths p from the pair to a common ancestor, and let M_p denote he number of meioses along that path. Then $\phi = \sum_p (1/2)^{M_p+1}$.

Testing family clustering of continous traits:

- ANOVO approach: the factor is family, and we are testing if the family means are equal (if so, then family has an effect on the trait). The phenotype of the j-th member of the i-th family is:

$$Y_{ij} = \mu + X_i + E_{ij} \tag{1.24}$$

  where X_i is the family effect and E_{ij} is the error (environment) term.

- Use family member correlation: for the i-th family, the covariance between members j and k can be expressed as (using components of covariance - see the Section on Quantitative Genetics):

$$\text{Cov}\,(Y_{ij}, Y_{ik}) = 2\phi_{jk}\sigma_A^2 + \nabla_{jk}\sigma_D^2 + \phi_{jk}\sigma_I^2 + \gamma_{jk}\sigma_C^2 + \delta_{jk}\sigma_E^2 \tag{1.25}$$

  where $\sigma_A^2, \sigma_D^2, \sigma_d^2$ are additive, dominance variance and interaction (between different loci) variance, σ_C^2 for shared environment and σ_E^2 for indepenent environment, and γ_{jk} for the propotion of shared env. influence and $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. Given this, we can construct the covariance matrix (e.g. for a nuclear family of four member) of a family, and use the data to fit the covariance matrix to estimate the variance terms.

1.3.2 Segregation Analysis

Reference: [Thomas, Chapter 6]

Ascertainment:

- Why could ascertainment be an issue? For example, suppose we are estimating the effect β_1:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1.26}$$

  where x is binary. Then $\beta_1 = \mathrm{E}(y|x=1) - \mathrm{E}(y|x=0)$. If we only ascertain the extreme values, then β_1 may be overestimated.

- Different ascertainment strategies for binary traits: define π as the probability that a random case in the population is chosen for study. Then we have complete ascertainment if $\pi = 1$, or single ascertainment if $\pi \to 0$, i.e. any case has a certain probability to be chosen. Under single ascertainment, multi-case families have higher probabilites of being ascertained.

Segregation analysis for sibship data:

- Example: suppose we are testing if the gene is dominant or recessive. If dominant, then the probability of being affected in a sibling is $p = 1/2$; and if recessive $p = 1/4$. So we are testing whether $p = 1/2$ or $1/4$. Given a sibship of size s, suppose r siblings are affected, then we have:

$$P(r|s,p) = \binom{s}{r} p^r (1-p)^{s-r} \tag{1.27}$$

  Suppose we have N_{rs} siblings for the cases of r affected in s siblings, then the likelihood is:

$$L(p) = \prod_r \prod_s P(r|s,p)^{N_{rs}} \tag{1.28}$$

  This allows us to test p.

- Ascertainment correction: suppose we have single ascertainment, then a family is ascertained if $r \geq 1$. We have:

$$P(r|s,A) = \frac{P(r|s)P(A|r,s)}{P(A|s)} \tag{1.29}$$

  The term $P(r|s)$ is the binomial probability, $P(A|r,s) = 1-(1-\pi)^r \approx \pi r$, which is roughly proportional to the number of affected siblings, and $P(A|s)$ can be computed by summing over r.

Segregation analysis in a general pedigree with major gene:

- Model: suppose we consider a pedigree, let Y_i be the phenotype of the i-th member, and G_i be its genotype (unobserved). Let G_{m_i} and G_{f_i} be the parent genotypes. The parameters of the model are $\Theta = (f, q)$ where f is the penetrance function and q allele frequency (in founders). The model can be written as:

$$P(Y|\Theta) = \sum_g \prod_i P(Y_i|g_i) P(g_i|g_{m_i}, g_{f_i}) \tag{1.30}$$

  Note that for founders, the conditional probabilities are replaced by allele frequencies.

- Ascertainment: we need to modify the likelihood according to the ascertainment model:

$$L_A(\Theta) = P(Y|\Theta, A, \pi) = \frac{L(\Theta)P(A|Y,\pi)}{P(A|\Theta,\pi)} \tag{1.31}$$

- Elston-Stewart peeling algorithm: assume conditional independence of members given the parents, the we could formuate a Dynamic Programming algorithm where the subproblems are: $P(Y|G)$, where G is the genotype of "linkers" (the individuals that link different families).

Segregation analysis in a general pedigree with polygene:

- Ideas: the traits are determined by the breeding values Z: the contribution of genes to the trait. And the dependency within families members can be captured by the correlation between Z's.
- The breeding values Z: correlated among families members. Ex. assuming additive model, we have $\mathrm{E}(Z) = \frac{1}{2}[\mathrm{E}(Z_m) + \mathrm{E}(Z_f)]$, the mean of Z is the average of parent means, and for the variance, $\mathrm{Var}\, Z = (\frac{1}{2}P_1 + P_2)V = \frac{1}{2}V$ where V is the variance of parents.
- Model: we have the likelihood

$$L(\Theta) = \int \cdots \int \prod_i P(Y_i|Z_i = z_i)P(z_i|z_{m_i}, z_{f_i})dz \tag{1.32}$$

  We can use MVN for the distribution of Z: its mean is vector μ, and covariance matrix C captures the dependency of relatives. For example, for a nuclear family with two offsprings, C is a 4×4 matrix, with the covariance between parent and child $1/2$, and the covariance between siblings also $1/2$. The trait can be written as a linear model of Z, or simply $\mathrm{E}(Y|Z) = \alpha + Z + \epsilon$ (then Z need to be scaled). It's can be shown then that Y follows MVN distribution.

1.4 Recombination and Linkage

[Hartl, Principles of Population Genetics, Chapter 2,3, Section 9.2]

Linkage disequilibrium:

- Model: consider two-loci, each with two alleles A, a and B, b. Let p_A, p_a be the frequency of A and a, p_B, p_b be the frequency of B and b, P_{AB} be the frequency of the haplotype AB and so so. We know that if two loci are independently assorted, then we always have $P_{AB} = p_A p_B$. If the two are linked, this may not be true. We define C as the fraction of recombination (defined as the fraction of recombinants, or the probability that any haplotype undergoes recombination), then we have the equation:

$$P'_{AB} = (1 - c)P_{AB} + cp_A p_B \tag{1.33}$$

  where P'_{AB} be the frequency in next generation. If we define $D = P_{AB} - p_A p_B$, called linkage disequilibrium (LD), then we have:

$$D_{t+1} = (1 - c)D_t \tag{1.34}$$

  Thus D_t converges to 0 at the rate $1 - c$, i.e. strong recombination leads to quicker linkage equilibrium ($D = 0$). For any two loci with $D > 0$, we call the two at LD.
- Frequency of the gamete types: given the allele frequency, only a single parameter D is sufficient to determine the gamete types:

$$\begin{aligned} P_{AB} &= p_A p_B + D \\ P_{Ab} &= p_A p_b - D \\ P_{aB} &= p_a p_B - D \\ P_{ab} &= p_a p_b + D \end{aligned} \tag{1.35}$$

- Remark: fraction of recombination: $0 \le c \le \frac{1}{2}$, the extreme case is the two loci are completely independent, e.g. when they are very distant. Note in this case, cannot conclude that always recombination, thus the fraction of recombinant is 1; in fact, there could be multiple recombinations between two loci, and the overall effect is one locus has no information of the other.

Measure of linkage disequilibrium (LD):

- D': D depends on allele frequencies, thus not comparable when allele frequency is different. Define D' as the fraction of D over the theoretical maximum of D (at that allele frequency), $D_{\max}$ if $D > 0$; and the fraction over the theoretical minimum if $D < 0$. The maximum and minimum of D is given by:

$$\begin{aligned} D_{\max} &= \min\{p_A p_b, p_a p_B\} \\ D_{\min} &= \max\{-p_A p_B, -p_a p_b\} \end{aligned} \tag{1.36}$$

  Note: D' has the property that $D' = 1$ if any of the four gametes has frequency 0, i.e. D' is sensitive to rare haplotypes.

- r^2: defined as $D^2/(p_A q_a p_B q_b)$. It has a simple interpretation: its square root is the correlation coefficient between the two alleles A and B (treating them as two RVs).

- Comparison of D' and r^2: D' measures the departure from linkage equlibrium, and r^2 measures the correlation/dependence of two loci. Thus at $D' = 0$ (equlibrium), we have $r^2 = 0$. However, as D' increases, r^2 may take a range of values - there are many ways of moving away from equlibrium. Treat this as the independence of two binary RVs: $D' = 0$ means two RVs are independent; r^2 is large only when two RVs are very correlated, or one variable encodes information of the other (which may not be the case, e.g. $A = 1$ suggests $B = 1$, but not the other way around).

- Example: when one haplotype is absent, $D' = 1$. r^2 on the other hand, depends on the genealogy/history: e.g. if a and b mutations happen to occur in the same chromosome in the ancestor, then they will show strong correlation; if mutations are new, then weak correlation.

Haplotypes:

- Definition: a combination of alleles at multiple loci that are transmitted together on the same chromosome.

- Haplotype blocks: defined according to r^2. In human the haplotype blocks are of size a few tens of kb; in Drosophila, a few kb; in Arabidopsis, the order of 100 kb (intense inbreeding).

- Haplotype resolution: in genotyping experiment, at each SNP, if it is heterozygous, it may not be clear which parent an allele is from, and as a consequence, it may be difficult to determine haplotype of a multi-SNP region.

Causes of LD:

- Population admixture: if allele frequencies are different, then one locus may show LD with another. Suppose p_A is large in one population, but not in the other, then an A allele will suggest that it is more likely to come from the first population, and then the allele in the B locus would tend to be that specific to the first population. An extreme case: suppose one population has fixed haplotype AB, and the other has fixed haplotype ab, then the mixture of two mutations will result in LD (no heterozygotes at the beginning).

- Reduced recombination: from inbreeding (e.g. in plants), chromosome inversion (making recombination more difficult), etc.

- Selection: a particular combination of alleles may have higher fitness.

- Recent mutations: if mutation is recent, and recombination is low, then the two loci may be at LD. This is especially important in human genetics, where human populations grow exponentially in a short amount time of time, and the recombination rate is low, about 1 crossover per 100 Mb per generation (thus many mutations in a neighborhood without enough time to shuffle them).

- **Remark**: the association of X and Y may come from the common association of X and Y with some confounding variable (e.g. the group where the data instance comes from). In population genetics: population structure is a common confounding variable.

Meiosis, recombination and origin of species [Peter Donnelly, May, 2016]

- Recombination: hot-spots about 1-2kb, 30K hotspots in human. Contains motif of Zinc finger gene PRDM9. PRDM9 binds the motif, place H3K4me3 mark, recruit recombination machinary, and create ds break. The break is then resolved in two ways: 10% cross-over, 90% non cross-over.
- PDRM9 in recombination: if there is a cis-change in one chromsome, asymmetry leads to inability of fixing ds break, and the offsprings have low fertility. Changing PRDM9 sequence would equilize both chromosomes and make the offsprings fertile again. Experimental proof from breeding two mouse strains.
- Role in speciation: each change at PRDM9 motif creates asymmetry and pressure of PDRM9 change. However, if PDRM9 does not change quick enough, potential for creating hybrid incompability.

1.5 Quantitative Genetics

Reference: [Felsenstein, Chapter 9], [Falconer & Mackay, Introduction to Quant. Genetics, 4ed]

Overview:

- Goal: the relative contributions of genetic and environmental influences; prediction of the traits of relatives; prediction of the response to selection (e.g. choosing parents of higher values of traits, what would be the distribution of offsprings); etc.
- Intuition: in general, we do not know the exact genetic basis of a quantitative trait. But we still be able to make predictions because: the traits of relatives are correlated (to different degrees depending on how close they are), and the extent of correlation reveals genetic vs environmental effects, and allows extrapolations (response to selection, other relatives, etc.).
- ANOVA perspective: Consider variation of some phenotype, P, it has two sources: genotype variation and environmental variation. We call the effect of genotype G, and the effect of environment E, then:

$$P = G + E \tag{1.37}$$

  From ANOVA perspective, G represents the group effect and E represents the intra-group variation. The total variance can be partitioned into the variance due to each source:

$$V_P = V_G + V_E \tag{1.38}$$

  We now have a way of defining the contribution of one source to the total variation, e.g.: the ratio of V_G over V_P is the importance of genetic effects on the phenotypic variation (heritability).
- Correlation between relatives: in general, the genotypes (groups) are not observed, so the standard ANOVA approach cannot be applied. The statistical idea is: the G component is shared between relatives (to different extents), and this creates correlation between relatives. This would allow one to infer V_G, V_E, h^2, etc.

1.5.1 Variance of Phenotypes

Quantitative genetic model:

- Additive model: the phenotype value P is the sum of effects contributed by each of the n loci, plus an environment effect:

$$P = \sum_{i=1}^{n} g_i + e \tag{1.39}$$

  It is assumed that $E(e) = 0$. The parameter g_i can be understood as the deviation from the mean introduced by the genotype at the i-th locus, i.e. the mean of the group defined by the genotype at the i-th locus, relative to population mean.

- Assumption 1: the effects of loci are additive, and there is no interaction among loci.
- Assumption 2: the genotypes at the n loci are independent of each other. Effectively assume linkage equilibrium of the n loci.
- Assumption 3: the environmental contribution to the phenotpye is indepdent of the genotype, and indepdent of environment contributions in other individuals. The later part of this assumption is the part that is most violated in practices, as environments of relatives are often correlated.
- Scale transformation: often need to transform the quantitative trait s.t. the additivity is held (and also for normality).

Mean phenotypes in relatives: the goal is to determine how mean phenotypes of relatives are related.

- Mean phenotype: use the assumption $E(e) = 0$, we have:

$$E(P) = \sum_i E(g_i) \tag{1.40}$$

  Thus we only need to consider each locus independently.

- Inbreeding effects: let the genotype frequencies of AA, Aa and aa be P, Q and R respectively (which can be determined from inbreeding coefficient); and the phenotypes are a_{11}, a_{12} and a_{22}. Then:

$$E(g) = Pa_{11} + Qa_{12} + Ra_{22} = p^2 a_{11} + 2p(1-p)a_{12} + (1-p)^2 a_{22} + fp(1-p)[a_{11} + a_{22} - 2a_{12}] \tag{1.41}$$

  Thus the mean phenotype is linearly related to the inbreeding coefficient f. The coefficient is proportional to the difference between the mean of the two homozygotes and the heterozygote.

- Inbreeding depression/hybrid vigor: for many traits, the heterozygote is often better than either parent (assuming homozygotes). This could be due to overdominance, or dominance at most of the loci (thus the heterozygote is always maximum between the two possible homozygote genotypes).
- Means of crosses and backcrosses: suppose we cross two inbred lines. We have $P_1 = a_{11}, P_2 = a_{22}$, and $F_1 = a_{12}$, but there may not be a simple relationship of P_1 and P_2 as we don't know dominance relation. The mean phenotypes of F_2 can be expressed in terms of the mean phenotypes of the parents and F_1:

$$F_2 = \frac{1}{4}a_{11} + \frac{1}{2}a_{12} + \frac{1}{4}a_{22} = \frac{1}{2}\left(\frac{1}{2}P_1 + \frac{1}{2}P_2\right) + \frac{1}{2}F_1 \tag{1.42}$$

Additive and dominance variances:

- Motivation: even through relatives share genetic contribution, not all genetic components can be inherited, so we need to partition the genetic effect into inheritable ones and the non-inheritable ones.
- ANOVA perspective: Suppose we have two factors, gene (represented by group G) and environment (represented by group E):

$$Y = G + E \tag{1.43}$$

  where G is the effect of gene, and E is effect of environment (since we are mainly interesed in gene, this is also the "error variance" within a group defined by genotype). To test if G has any effect, ANOVA decomposes the variance of Y into that due to G (inter-group) and that due to E (within-group). We can write this as:

$$\operatorname{Var} Y = \operatorname{Var} G + \operatorname{Var} E \tag{1.44}$$

  G is related to the underlying alleles: $G = \mu_i - \mu$ with probability p_i, where μ_i be the mean of the i-th group and μ is the mean of the entire population. Note that $\mathrm{E}(G) = 0$. The variance of G is given by:

$$\operatorname{Var} G = \mathrm{E}(G^2) = \sum_i p_i(\mu_i - \mu)^2 \tag{1.45}$$

From this we see that, large effect $\mu_i - \mu$ leads to large $\operatorname{Var} G$ - thus we focus on the variance paritition to estimate how large the genetic effect is.

- Additive and dominant genetic values (Fisher's decompoisition): we assume an additive genetic model, given by Equation 1.39. For simplicity, we consider a single gene with two alleles, i and j. We want to write the genotype effect, g, as the sum of three parts:

$$g = \mu + \alpha_i + \alpha_j + \delta_{ij} \tag{1.46}$$

where α_i, α_j are the departure from the mean due to the alleles i and j, respectively. And δ_{ij} is the "dominance deviation" from group mean that is unexplained by α_i and α_j. Define the additive genetic value $A = \alpha_i + \alpha_j$, and the dominance genetic value $D = \delta_{ij}$.
Note: our notation is α_i and α_j are RV's since i and j (allele) are random; however, α_1, α_2, δ_{11}, etc. represent constants, e.g. α_1 is the mean of the group (allele type 1), μ_1, minus the overall mean μ.

- Additive and dominance variances and heritability: in general we can write the phenotype as:

$$P = \mu + A + D + E \tag{1.47}$$

where A is called the "breeding value", and all terms are independent. Thus:

$$V_P = V_A + V_D + V_E \tag{1.48}$$

We define the broad-sense heritability as:

$$H^2 = \frac{V_G}{V_P} \tag{1.49}$$

and the narrow-sense heritability as:

$$h^2 = \frac{V_A}{V_P} \tag{1.50}$$

- Experimental determination of V_G and V_E:
  - In theory, one can determine V_E by measuring the phenotypic variance of genetically identical strains. Then suppose we have V_P from the normal variation in the population, then $V_G = V_P - V_E$.
  - Difficulities: V_E may depend on the genotype of the identical strains. Furthermore, inbreed strains somtimes have higher (or lower) variance than the normal cross-bred population.

Relating V_A and V_D to the genetic model: [Felsenstein, Chapter 9]

- Genetic model: suppose we have a locus with 2 alleles A_1 and A_2, with the frequency of A_1 being p and the other $q = 1 - p$. The genetic effect of the genotype (ij) is denoted as a_{ij}. Our goal is to relate V_A and V_D to p and a_{ij}'s. We first note that the genotype of a sample, i and j, are random, and so the genetic values (thus we can define the variance of genetic values). Also note that $\mathrm{E}(\alpha_i) = 0$ and $\mathrm{E}(\delta_{ij}) = 0$. We have the relation between V_A, V_D and genetic values:

$$V_A = E[\alpha_i^2] + E[\alpha_j^2] = 2E[\alpha^2] = 2\sum_{i=1}^{m} p_i \alpha_i^2 \tag{1.51}$$

$$V_D = E[\delta_{ij}^2] = \sum_{i,j=1}^{m} p_i p_j \delta_{ij}^2 \tag{1.52}$$

where m is the number of alleles (the equation can be applied to multi-allele locus as well). Our problem next is to determine α_i and δ_{ij}.

- Genetic values by group means: α_1 is simply the mean phenotype of the group where one alleles is A minus the population mean; and similary for α_2. The population mean is:

$$\mu = \sum_{ij} p_i p_j a_{ij} \tag{1.53}$$

assuming HWE. For two-allele case, we have:

$$\alpha_1 = pa_{11} + (1-p)a_{12} - \mu \tag{1.54}$$

$$\alpha_2 = pa_{12} + (1-p)a_{22} - \mu \tag{1.55}$$

And for the dominance genetic value, we have:

$$\delta_{ij} = a_{ij} - (\mu + \alpha_i + \alpha_j) \tag{1.56}$$

Plug-in α_1 and α_2, we have:

$$\delta_{11} = (1-p)^2(a_{11} - 2a_{12} + a_{22}) \tag{1.57}$$

The expression for δ_{12} and δ_{22} are the same, but with $(1-p)^2$ replaced by $-p(1-p)$ and p^2 respectively.

- Genetic values by regression: we could consider the equation of the genetic value as a regression where the predictor N is the number of a alleles in the parents:

$$g = \mu + 2\alpha_1 + (\alpha_2 - \alpha_1)N + \delta \tag{1.58}$$

where δ is the residual term. Thus the regression coefficients can be expressed through regression on three groups: the group AA with average trait a_{11}; the group Aa with average trait a_{12}; and the group aa with average trait at a_{22}. Solving this regression leads to the same equations of α_i and δ_{ij} above.

- Common parametrization of genetic model [Falconer, Chapter 7]: this is given by:

$$a_{11} = a \qquad a_{12} = d \qquad a_{22} = -a \tag{1.59}$$

If there is no dominance, $d = 0$; if A_1 is dominant over A_2, $d > 0$; if A_2 is dominant over A_1, $d < 0$. The degree of dominance is measured by d/a. Under this parameterization, we have the average effect of gene substitution:

$$\alpha = p(a-d) + q(d+a) = a + d(q-p) \tag{1.60}$$

And the additive genetic values of two alleles:

$$\alpha_1 = q\alpha \qquad \alpha_2 = -p\alpha \tag{1.61}$$

Extending the basic model of genetic and environment variance [Falconer, Chapter 8]:

- Correlation between genotype and environment: if the two variables are not independent, e.g. one genotype is more likely to be associated with one type of environment (e.g. in experimental populations, the treatment depends on the phenotype, thus the genotype of organisms), we have:

$$V_P = V_G + V_E + 2\text{Cov}_{GE} \tag{1.62}$$

Intuition: when G and E are correlated, V_P will be higher because the phenotype can now become more extreme (both G and E change in the same direction).

- Gene-environment interaction: when there is an interaction term, we have:

$$P = G + E + I_{GE} \tag{1.63}$$

And then:

$$V_P = V_G + V_E + 2\text{Cov}_{GE} + V_{GE} \tag{1.64}$$

Again, V_P is now higher because the interaction creates additional deviation from the additive model.

- Epistasis: suppose two loci control the genetic value, we could write genetic value as:

$$G = \mu_G + A + D + AA + AD + DD \tag{1.65}$$

where A is the average effect of single alleles, D is dominance deviation, AA, AD and DD the interactions. Variances:

$$V_G = V_A + V_D + V_{AA} + V_{AD} + V_{DD} \tag{1.66}$$

- Remark: in general, correlations and interactions among the components increase the variance.

1.5.2 Heriability and Resemblance of Relatives

Reference: [Felsenstein, Chapter 9], [Falconer & Mackay, Introduction to Quant. Genetics, 4ed]

Covariance between relatives:

- Indepdence among loci: suppose we consider the phenotypes of two relatives:

$$\begin{aligned} X &= g_1 + g_2 + \cdots + g_n + e \\ Y &= g'_1 + g'_2 + \cdots + g'_n + e' \end{aligned} \tag{1.67}$$

Assume the independence of loci, the independence of environment and genetic effects, we have:

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(g_1, g'_1) + \mathrm{Cov}(g_2, g'_2) + \cdots + \mathrm{Cov}(g_n, g'_n) \tag{1.68}$$

- Covariance in terms of V_A and V_D: we consider only a single locus, there are only three situations in g and g':
  - g and g' has no allele IBD: no relationship between the two, thus $\mathrm{Cov}(g, g') = 0$.
  - g and g' has one allele IBD: we have $g = \alpha_i + \alpha_j + \delta_{ij}$ and $g' = \alpha_i + \alpha_k + \delta_{ik}$, the only nonzero terms are $\mathrm{Cov}(\alpha_i, \alpha_i)$ and $\mathrm{Cov}(\delta_{ij}, \delta_{ik})$. It can be proven that the latter term is 0 (intuitively, this is clear, as the dominance term is entirely determined by A_j and A_k, but they are independent). Thus $\mathrm{Cov}(g, g') = V_A/2$
  - g and g' has two alleles IBD: this is the variance of g, $\mathrm{Cov}(g, g') = V_A + V_D$

Suppose the probability of the cases above are P_0, P_1, P_2, respectively, we have:

$$\mathrm{Cov}(g, g') = P_1(V_A/2) + P_2(V_A + V_D) \tag{1.69}$$

In terms of the covariance between the relatives:

$$\mathrm{Cov}(X, Y) = \left(\frac{1}{2}P_1 + P_2\right) V_A + P_2 V_D \tag{1.70}$$

Remark: this is the key equation, and it can be understood without invoking the underlying genetics, i.e. expression of α and δ terms in terms of the underlying genetic values (a_{11}, a_{12}, a_{22}) and allele frequencies. The underlying genetics provide a rigorous proof of the relevant relations, notably $\mathrm{Cov}(\delta_{ij}, \delta_{ik}) = 0$.

- Correlations and regression: the correlation and regression coefficient between two relatives X and Y:

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{1.71}$$

$$\beta_{XY} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \tag{1.72}$$

when $\sigma_X = \sigma_Y$ (normally true except the case of offspring-midparent regression), we have $\rho_{XY} = \beta_{XY}$.

- Some cases:
  - Parent and offspring: we have $P_1 = 1, P_2 = 0$, thus $\rho_{OP} = \beta_{OP} = \frac{1}{2}h^2$;
  - Between offspring and mid-parent: the covariance:

$$\text{Cov}\left(\frac{P_1 + P_2}{2}, O\right) = \text{Cov}\left(O, P\right) = \frac{1}{2}V_A \tag{1.73}$$

    The variance of mid-parent:

$$\text{Var}\left(\frac{P_1 + P_2}{2}\right) = \frac{1}{2}V_P \tag{1.74}$$

    Thus we have $\beta_{O\bar{P}} = h^2$.
  - Half siblings: the covariance:

$$\text{Cov}_{HS} = \frac{1}{4}V_A \tag{1.75}$$

    Thus $\rho_{HS} = \frac{1}{4}h^2$;
  - Full siblings: $P_1 = \frac{1}{2}, P_2 = \frac{1}{4}$:

$$\rho_{FS} = \frac{1}{2}h^2 + \frac{1}{4}\frac{V_D}{V_P} \tag{1.76}$$

    Note that the correlation between ful sibs is greater than between parents and offsprings (the possibility of exactly the same diploid genotype in the two).

General model: resemblance between relatives under gene-gene interactions and shared environment:

- Gene-gene interactions: in the absence of gene-gene interactions, suppose the covariance between two relatives is given by:

$$\text{Cov } = rV_A + uV_D \tag{1.77}$$

  With epistatic interactions, the covariance would be higher. Suppose we consider two relatives:

$$G = A + D + AA \qquad G' = A' + D' + AA' \tag{1.78}$$

  The covariance is thus:

$$\text{Cov}\left(G, G'\right) = rV_A + uV_D + r^2V_{AA} \tag{1.79}$$

  where r^2 comes from the fact that if the probability that the two share a single gene is r, then the probability that the two share two genes (s.t. epistasis is shared) is r^2. More generally, we have more interaction terms:

$$\text{Cov } = rV_A + uV_D + r^2V_{AA} + ruV_{AD} + u^2V_{DD} + r^3V_{AAA} + r^2uV_{AAD} + ru^2V_{ADD} + u^3V_{DDD} + \cdots \tag{1.80}$$

- Environmental covariance: for relatives, it may be important to model the shared environment, thus the environmental variance can be writted as:

$$V_E = V_{Ec} + V_{Ew} \tag{1.81}$$

  where V_{Ec} is the between-group variance (contributing to the similarity between relatives in the same group).

Estimating variance components and heritability: [Falconer, Chapter 10]

- Common procedures: parent-offspring regression; half-sib covariance: groups of half-sibs, each group with the same father; etc.

- ACE model for twin-studies: [Twin Study, Wiki] consider only the additive effect (no dominance, no epistasis):

$$P = A + C + E \tag{1.82}$$

where C is the contribution from the common environment and E from the unique environment. Then for mono-zygotic (MZ) twins, the covariance is:

$$\text{Cov}_{MZ} = \text{Cov}\,(A + C + E, A + C + E') = V_A + V_C \tag{1.83}$$

For di-zygotic (DZ) twins, the covariance:

$$\text{Cov}_{DZ} = \text{Cov}\,(A + C + E, A' + C + E') = \frac{1}{2}V_A + V_C \tag{1.84}$$

Solving V_A and V_C:

$$V_A = 2(\text{Cov}_{MZ} - \text{Cov}_{DZ}) \qquad V_C = 2\text{Cov}_{DZ} - \text{Cov}_{MZ} \tag{1.85}$$

In terms of heritability:

$$h^2 = 2(r_{MZ} - r_{DZ}) \tag{1.86}$$

- Limitations:
  - Shared environments among relatives: to control it, randomize the environment of samples in experimental populations.
  - Genotype-environment interactions.
  - Genetic interactions: among different loci.
  - Age and sex effects: e.g. parent-offspring regression, the traits may not be comparable, thus in general, half-sibs regression may be better.

1.6 Genetic Mapping of Complex Traits

Understanding genetics of complex trais: the fundamental problem is the genetic basis of individual variation.

- Genetic mapping: heritability of trait, finding the loci, environmental influences, the interactions among loci, etc.
- Mechansims: molecular explanation of the genetics: causal loci, gene-gene interactions, gene-environment interactions, etc.

Problems/challenges of quantitative trait mapping:

- Reference: [Mackay & Ayroles, NRG, 2009]
- Identifying causal genes or CREs and QTNs (quant. trait nucleotide): not straightforward, as the QTL confidence interval or haplotype block map may be large. One idea: the allele frequency of SNP may indicate the importance, e.g. low frequency suggest selection, thus prioritize the SNPs among all candidates.
- Increasing the power to detect QTLs and QTNs: especially a problem if a huge number of hypothesis is tested (in the case of eQTL and dense markers). One idea is to group correlated traits.
- How QTNs affect phenotypes through molecular networks: joining eQTLs and QTTs (quant. trait transcripts). Idea: (1) causal inference procedure; (2) use co-expression network of QTTs.

Strategies for human disease gene identification: mostly applicable to Mendelian traits [Human Molecular Genetics]

- Positional evidence: through genetic linkage/association, also information such as chromosome abnormality (below), and expanded repeats (often found in disease genes in patients) can provide clues on the positional information.
- Functional evidence: gene expression pattern, function in other species (animal model) or of paralogs, biochemistry and physiology, etc. Functional information is used for both priortization of genes (even with positional information), and for providing candidates.
- Positional cloning: the process of starting with positional information and reaching the disease genes. The steps: candidate genes in the region, prioritization (functional information), mutation screening, and confirming the genes.
- Confirming a candidate gene: require stronger evidence, most often mutation screening. Also restorization of function in vitro or the function in animal models.

Challenges of personal genomics [personal thoughts]:

- Finding causal variants: with whole-genome sequencing (WGS), very large number of variants may be found. How would causal variants of a complex trait be determined? This is particularly difficult given the heterogeneity of complex traits. Broad strategies:
  - Population genetics: it's possible that even though rare variants are rare in general population, they may be fairly common in patients.
  - Recognize relations among genes to better deal with heterogeneity through (1) functional interactions among genes; (2) multiple variants/genes may affect the same intermediate/molecular traits.
  - Stratification of phenotypes: once divide a phenotype into sub-phenotypes based on variables such as certain molecular signatures, the genetics of the sub-phenotype may be less heterogenous.
  - Functional evidence of candidate loci: this is important for validating the findings. If additional functional data is available, e.g. the loci is associated with another trait known to be important for the disease (e.g. LDL for metabolic diseases).
- Molecular mechanism of causal variants: understanding how a genetic change (or multiple ones) leads to diseases (increase of disease risk).
  - Mutations in non-coding sequences: what are the targe genes? Mechanism of how non-coding mutation leads to change of target genes: e.g. TFBS change, nucleosomes, etc.
  - Downstream effects: suppose X is a causal gene (a regulatory protein), then X may affect some functional genes (e.g. enzyme), which affect certain metabolite/stress level, and cause cell/tissue damage. Need to find such causal chain of events. Note that things such as stress level, cell damage may be reflected as levels of certain markers (e.g. DNA damage may be reflected by the level of DNA repair enzymes).
- Gene-environment interactions:
  - Inferrning causality from correlations between genotypes and different environmental variables.
  - Recognizing interactions: e.g. the effect of genetic changes is only manifested in the presence of some environmental influence, and vice versa.
  - Environmental effects may be manifested as changes of metabolites, gene expression, epigenetic states, etc.

- Prediction and personalization: risk prediction and prediction of response to treatments
  - A better disease model would help prediction and personalization: e.g. in pharmacogenetics, both drug metabolism and drug target level could affect the efficacy of a drug, such knowledge can be encoded into some (non-linear) model.
  - Better utilization of patient information: phenotype stratification, patient grouping, etc.

Strategies of improving genetic mapping of complex traits [personal thoughts]: according to the general framework of regression analysis

- Incorporating more features: using imputation to add more SNPs; rare variants; CNVs.
- Feature expansion: epistasis between SNPs.
- Structure in features: haplotype and regional test; pathway analysis.
- Prior knowledge of features: SNP annotation (conservation, MAF, etc.); gene annotation and relations.
- Group structure: exploiting family structure; multi-population mapping and meta-analysis; phenotype stratification (by family, population, age of onset, etc.).
- Sparsity: SNP selection and causal variant analysis.

Global analysis of genetic data:

- Motivation: when we do not have confidence for individual genes, how do we draw conclusion about the overall genetics of a complex trait? This may happen, e.g. when we have case/control data where there are many weak loci; or we have de novo mutation data where most causal genes have only zero or one mutation.
- Enrichment pattern: some signals may be collectively detectable. Ex. (1) suppose we compute the statistic of each gene, and compare the distribution of the statistic of all genes with null distribution; (2) de novo mutation data: the total number of mutations across all genes may be higher than expected by chance.
- Genetic architecture: formal analysis of the data may allow one to infer genetic architecture of the trait, including the number of causal genes, the average effect size (or effect size distribution), and the variance explained by the data.
- Gene network analysis: patterns in the gene network, e.g. pathway test, whether candidate genes are highly connected, whethere they are linked to known genes, etc.
- Reference: [De novo mutations revealed by whole-exome sequencing are strongly associated with autism, Sanders, Nature, 2012], [Patterns and rates of exonic de novo mutations in autism spectrum disorders, Neale, Nature, 2012]

Chapter 2

Population Genetics

2.1 Genetic Polymorphism: an Overview

Ref: [Hartl & Clark, Principles of Population Genetics, Chapter 1]

Challenges of population genetics:

- Basic problems: how the pattern of genetic variations (SFS, heterozygosity, etc.) are related to the population history, mutational process and natural selection? How do we reconstruct the history from extant data (including the inference of time of events)?
- Strategy: we start with simple scenario, uniform population with one gene (two alleles), then solve more complex problems, non-random mating, population subdivision/migration, selection, recombination between multiple sites, and so on.

Maintanence of polymorphism:

- Mutation-selection: most mutations are deleterious or nearly neutral, and are eliminated by natural selection (or random drift for neutral).
- Balance: overdominance (heterozygosity is favored), frequency-dependent selection (rare alleles are favored), etc.

Utility of polymorphism data:

- Polymorphism as markers: shared polymorphism may suggest shared ancestry. Thus polymorphism data can be used to infer the information about ancestry and history, e.g. migration/race, DNA fingerprinting and genealogy, phylogenetic tree reconstruction.
- Inference of evolutionary process: population history (e.g. growth), selection on DNA sequences, disease association.

Lessons/strategies of population genetics [peronsal notes]:

- Pattern of genetic relatedness or heterozygosity: basic process of random drift/inbreeding leads to increase of genetic relatedness or reduction of heterozygosity. So from pattern of relatedness, we can infer the underlying changes related to population size, population subdivision/migration, selection, and so on.
- Coalescence framework: the amount of genetic variations in a certain number of samples is due to mutations occuring during coalescence. So estimating coalesence time can lead to estimation of the amount of genetic variations.

Questions of population genetics [personal notes]:

- Proof of expected SFS under infinite-sites model from coalescence:

$$\mathrm{E}\left(S_i\right) = \frac{\theta}{i} \tag{2.1}$$

  We start with $\mathrm{E}\left(S_1\right) = \theta$. One possible strategy is: after each coalesence event, ask how many nodes are internal and external, and count the contribution to the external lineage from both type of nodes.

- Defining F_{ST} for genomewide data: from the relation of F_{ST} and heterozygosity, we should define F_{ST} using all loci, included the ones that are fixed in the population. However, comparing two populations, most of the sites should remain unchanged between the two and have $F_{ST} = 0$. Shall we include them?

Departure of HWE:

- HWE specificies the frequency of the heterozygotes, $P(Aa) = 2p(1-p)$, as a result of random mating. The variance of X (the number of a's in an individual) is given by: $\mathrm{Var}\, X = 2p(1-p)$.

- Loss of heterozygosity (LOH): some factors may reduce the frequency of heterozygotes, creating departure from HWE. This may include: (1) recessive diseases: thus in patients, most are homozygotes aa; (2) inbreeding; (3) population admixture: e.g. mixture of two populations, one fixed with A and the other fixed with a. The LOH due to population substructure is called the Wahlund effect.

  - Consequence of LOH: increase of the variance of X. This can be understood as: due to LOH, the population has more AA and aa's, thus more deviation from the mean (Aa).

Questions:

- Genetic variation under neutrality: does inifite-site model incorporate recombination? What would be the level of variation considering mutation and recombination?

- Testing selection using polymorphism data: selective sweep vs. test based on neutral theory (e.g. Fu-Li test)?

- Inferring age from polymorphism data: e.g. the HapMap common variants?

2.2 Evolution of Infinite Populations

Ref: [Hartl & Clark, Principles of Population Genetics; Felsenstein, Theoretical Population Genetics]

Goal: understand how evolutionary forces shape the long-term behavior of the population, as well as the dynamics of the changes.

Mutation:

- Model: consider a single locus with two alleles A and a. The mutation rates are μ (A to a) and ν (a to A) respectively. Let p_t be the frequency of A at generation t (and q_t be that of a), then p_t can be solved with discrete difference equation:

$$p_t = \frac{\nu}{\mu+\nu} + (p_0 - \frac{\nu}{\mu+\nu})(1-\mu-\nu)^t \tag{2.2}$$

- Result: because of recurrent mutations between the two, the population reaches equlibrium frequencies, whose ratios depend on the mutation rates. This should holds in the general case with multiple alleles. The rate of p_t convergence to equilibrium is geometric, with rate $1-\mu-\nu$.

Selection in haploids:

- Discrete generations: let p_A, p_a be the frequenices of A and a alleles, w_A, w_a be the fitness of A and a, and p'_A, p'_a be the frequencies in the next generation, we have:

$$\frac{p'_A}{p'_a} = \frac{w_A}{w_a}\frac{p_A}{p_a} \tag{2.3}$$

  Thus selection will drive the population towards the advantageous allele, with geometric rate equal to w_A/w_a. We could also write the dynamic equation in different ways, let $\bar{w} = w_A p_A + w_a p_a$, we have:

$$p'_A = \frac{p_A w_A}{\bar{w}} \qquad p'_a = \frac{p_a w_a}{\bar{w}} \tag{2.4}$$

  Or if we define $p = p_A$, we have the change of frequency in one generation:

$$\Delta p = p(1-p)\frac{w_A - w_a}{\bar{w}} \tag{2.5}$$

  The rate of change can be defined by the number of generations needed for the change of certain frequency:

$$t = \frac{\ln\left(p_A^{(t)}/p_a^{(t)}\right) - \ln\left(p_A^{(0)}/p_a^{(0)}\right)}{\ln(1+s)} \tag{2.6}$$

  which is roughly inversely proportional to s when s is small.

- Continuous generations: let r_A, r_a be the growth rates of A and a, we have:

$$\frac{dp}{dt} = (r_A - r_a)p(1-p) \tag{2.7}$$

Selection in diploids:

- Dynamic equations: similarly, the change of frequency can be written in three forms. Only need to replace w_A and w_a with the average fitness of the genotypes having at least one A and a respectively: $\bar{w}_A = p_A w_{AA} + p_a w_{Aa}$ and $\bar{w}_a = p_A w_{Aa} + p_a w_{aa}$. First, relative frequency ratio:

$$\frac{p'_A}{p'_a} = \frac{\bar{w}_A}{\bar{w}_a}\frac{p_A}{p_a} \tag{2.8}$$

  Second, the frequency in terms of average frequency:

$$p'_A = \frac{p_A \bar{w}_A}{\bar{w}} \qquad p'_a = \frac{p_a \bar{w}_a}{\bar{w}} \tag{2.9}$$

  Third, the change of frequency in one generation:

$$\Delta p = p(1-p)\frac{\bar{w}_A - \bar{w}_a}{\bar{w}} \tag{2.10}$$

- Multiplicative cases: the fitness of three genotypes are: $AA-(1+s)^2, Aa-(1+s), aa-1$. The result is exactly the same as the haploid case with selection coefficient s.

- Recessive and dominant case: (1) Recessive: $AA-1+s, Aa-1, aa-1$; (2) Dominant: $AA-1+s, Aa-1+s, aa-1$.

- Overdominance [Nielsen & Slotkin, Chapter 7]: the fitness of Aa is higher than both AA and aa, in this case, a stable polymorphism will be established. Overdominance is a special form of balancing selection. With overdominance, the frequency reaches equilibrium (stable polymorphism):

$$\hat{f}_A = \frac{s_{aa}}{s_{AA} + s_{aa}} \tag{2.11}$$

- Underdominance: the fitness of Aa is lower than AA and aa, the population will be fixed on A or a, depending on the initial frequency.

Dynamics of fitness: how the average fitness of the population changes over time. Intuitively this should always increase.

- Fisher's Fundamental Theorem of Natural selection: we consider haploid population with k possible alleles, we have:

$$p_i' = \frac{p_i w_i}{\bar{w}} \tag{2.12}$$

Use the equations of $\bar{w}$ and $\bar{w}'$:

$$\bar{w}' - \bar{w} = \sum_i p_i' w_i - \sum_i p_i w_i = \frac{1}{\bar{w}} \left[\sum_{i=1}^{k} p_i w_i^2 - \bar{w}^2 \right] \tag{2.13}$$

Thus we have the basic equation of Fisher's Fundamental Theorem of Natural selection:

$$\Delta \bar{w} = \frac{\mathrm{Var}(w)}{\bar{w}} \tag{2.14}$$

Thus, the increment of the mean population relative fitness is the ratio of the genetic variance in fitness to the mean fitness. This immediately implies that $\Delta \bar{w}$ will never decrease.

- Adaptive topography and fitness optimization: the function $\bar{w}(p)$ shows how the average fitness changes with the frequency of A in the population, called "adaptive topography". Thus the evolution of the population can be visualized as the movement in the adaptive topography. This leads to an alternative way of showing $\Delta \bar{w}$ never decreases. For the diploid case, we can derive the following equation:

$$\Delta p = \frac{p(1-p)}{2\bar{w}} \frac{d\bar{w}}{dp} \tag{2.15}$$

Then we have:

$$\Delta \bar{w} = \frac{d\bar{w}}{dp} \cdot \Delta p = \left(\frac{d\bar{w}}{dp} \right)^2 \frac{p(1-p)}{2\bar{w}} > 0 \tag{2.16}$$

Therefore, the behavior of the population can be described as fitness optimization: p will always move towards the direction where $\bar{w}(p)$ increases. However, this does not guarantee that the global optimum of $\bar{w}$ can be reached, and the final equilibrium depends on the initial frequency.

- Generality of the results: even though the Fundamental Theorem of Natural Selection and monotonicity of average fitness are proven only for simple cases, they are valid for much more general cases (the case of multiple alleles).

Selection in multiple alleles:

- Dynamics: let p_i be the frequency of the i-th allele, the dynamics:

$$p_i' = \frac{p_i \bar{w}_i}{\bar{w}} \tag{2.17}$$

which leads to:

$$\Delta p_i = p_i \frac{\bar{w}_i - \bar{w}}{\bar{w}} \tag{2.18}$$

- Equilibrium and fitness: at equilibrium, either $p_i = 0$ or $\bar{w}_i = \bar{w}$. However, not all equilibrium points are stable. Because $\bar{w}$ is nondecreasing, the local maximums must be stable equilibrium.

Mutation-selection balance: [Hartl, Section 5.4; Felsenstein, Chapter 3]

- Haploid model: suppose the fitness of A and a are 1 and $1-s$ respectively. Suppose the mutation of A to a is μ, and the back-mutation can be ignored. Suppose at t generation, the frequency of A is p, then after selection, its frequency becomes:

$$p^* = \frac{p}{1-(1-p)s} \tag{2.19}$$

This follows from a simple selection model of two genotypes. And after mutation, its frequency becomes:

$$p' = p^*(1-\mu) \tag{2.20}$$

Putting the two together, we have, in $t+1$ generation, the frequency of A is:

$$p' = \frac{p(1-\mu)}{1-(1-p)s} \tag{2.21}$$

At equilibrium, we have $p' = p$, solving this equation, and we have the frequency of a allele $(1-p)$:

$$q_e = \mu/s \tag{2.22}$$

- Diploid model: consider the case where A is advantageous. Let the fitness of AA, Aa and aa be 1, $1-hs$ and $1-s$, respectively. Suppose the rate of A to a mutation is μ, and the back-mutation can be ignored because of the low frequency of a in the population. Because of recurrent mutations, a cannot be completely eliminated. Let $\hat{q}$ be the equilibrium frequency of a, we have two cases: if a is completely recessive ($h = 0$):

$$\hat{q} = \sqrt{\frac{\mu}{s}} \tag{2.23}$$

If $h > 0$, then an approximation:

$$\hat{q} = \frac{\mu}{hs} \tag{2.24}$$

- Interpretation: when μ is very small (e.g. $2N\mu << 1$), the number of mutations per generation is much less than 1, and most will not survive next generations, so the population is essentially fixed at the advantageous allele A. In general, if $2N\mu$ is not substantially less than 1, then there will be more mutations created in each generation, and natural selection will reduce, but not completely eliminate all of them. So at equalibrium, there will be deleterious allels, and its frequency will depend on the ratio of (1) the rate of introducing new mutations, μ; and (2) the rate of eliminating mutations by selection ($\ln(1-s)^{-1} = s$).

2.3 Evolution of Finite Populations

Ref: [Hartl & Clark, Principles of Population Genetics; Felsenstein, Theoretical Population Genetics]. An introduction to population genetic theory [Crow & Kimura]

Random drift: Wright-Fisher model with no selection and recurrent mutation:

- Intuition of the process: at each generation, there is always a probability that some individuals leave no offspring. Thus over time, there is a growing tendency that all individuals come from a small number of ancestors; or the individuals are more likely to be genetically related over time.

- Inbreeding perspective: random drift will drive fixation of one allele in the population. This can be understood as the inbreeding effect: inbreeding is more likely within a finite population. Consider a population with two alleles, let f_t be inbreeding coefficient (the probability that two random gametes are IBD), then:

$$f_t = \frac{1}{2N} + (1 - \frac{1}{2N}) f_{t-1} \tag{2.25}$$

Solving this equation:

$$f_t = 1 - (1 - \frac{1}{2N})^t \tag{2.26}$$

Or equivalently, the heterozygosity (the probability that the two random gametes are different):

$$h_t = (1 - \frac{1}{2N})^t \tag{2.27}$$

Thus the heterozygosity decreases at the rate of $(1 - 1/2N)$ and approaches 0 as $t \to \infty$.

- Probability of fixation: equal to p when the initial frequency of the allele is p. Specifically, for new mutation, it is $1/2N$. The intuition is (coalescence): eventually, the population will be fixed with one ancestor, but since the chance is equal for all $2N$ ancestral alleles, the probability of fixing the allele of interest is $1/2N$. Another way to derive this: let X_t be the frequency of A alleles in time t, we have:

$$E(X_t) = X_{t-1} \tag{2.28}$$

from the Wright-Fisher process. Thus: $E(X_t) = \cdots = E(X_0) = p$, where p is the initial frequency.

- Time to fixation: for a new mutation that is fixed, the time to fixation is approximately $4N$ generations. This can be understood from coalescence. For a new mutation that is lost, the time of loss is approximately $2\ln(2N)$.

Mutation-drift balance: [Hartl, Section 5.7; Felsenstein, Chapter 7]

- Homozygosity under infinite-alleles model: with infinite allele model, the homozygosity is equal to the inbreeding coefficient, as every two alleles of the same type must be IBD. Let F_t be the homozygosity at time t, then at time $t + 1$, two random gametes have the same alleles if the same individual is sampled in t, or the two sampled individuals are IBD in t, and no mutations happened:

$$F_{t+1} = (1 - u)^2 \left[\frac{1}{2N} + \left(1 - \frac{1}{2N} \right) F_t \right] \tag{2.29}$$

At equlibrium:

$$F \approx \frac{1}{1 + 4Nu} \tag{2.30}$$

- Homozygosity under finite alleles: this could be analyzed similarly. Suppose there are K different alleles, the equilibrium homozygosity is given by:

$$F \approx \frac{1 + 4Nu/(K-1)}{1 + 4NuK/(K-1)} \tag{2.31}$$

- Two-allele model: suppose there are recurrent mutations between A and a. The idea is to analyze the fate of any new mutations.
  - $4N\mu \leq 1$: the PDF of the allele frequency is bi-modal at $x = 0$ or $x = 1$. At each generation, there are $2N\mu$ mutations, but only $1/(2N)$ of these if fixed, so on average, μ mutations will be fixed per generation, or in other words, one mutation per $1/\mu$ generations will be fixed. On the other hand, each fixed mutation takes $4N$ generations to fix, thus if $4N \leq 1/\mu$ (fixation occurs before new to-be-fixed mutations), the population will remain fixed most of the time. In general, if there are multiple alleles with possible mutations among each other, the population will follow a Markov chain of these alleles.
  - $4N\mu > 1$: with higher mutation rates, the probability that a popluation will be polymorphic is higher (the PDF of the allele frequency is single-peaked at intermediate frequency) - mutational equlibrium.

Diffusion approximation of Wright-Fisher model: [Hartl, Section 3.2, 3.3]

- Model: let $\phi(p, x; t)$ be the pdf of the allele frequency at x in time t, when starting at $x(0) = p$. Let $M(x)$ be the mean of the change of allele frequency (Δx) in one generation, and $V(x)$ be the varirance of Δx in one generation. $M(x)$ is driven by the systematic force such as selection and $V(x)$ by random drift.
- Kolmogorov forward equation: analyze $\phi(p, x; t)$ in terms of the frequency at $t - \Delta t$, which could be $x - \Delta x$, $x + \Delta x$ and x. Multiply the probability of the three cases (time 0 to time $t - \Delta$) and the probability of transition from the three case to x (in Δt), we could obtain the forward equation.
- Kolmogorov backward equation: starting from time $t = 0$, at time Δt, the frequency could be $p - \Delta p$, p and $p + \Delta p$. Similarly, multiply the probability of transition to three cases (in time Δt) and the probability of generating the final frequency from these three cases (time Δt to t), we could obtain the backward equation.
- Probability of fixation: at fixation, the time derivative is 0, and solve $u(p)$ in Kolmogorov backward equation (ODE of a single variable).

Mutation-selection-drift: [Section 5.7]

- Model: consider the life cycle of a population. Starting from zygotes, selection determines the frequency of genotypes before reproduction, then mutation changes the frequencies when producing gametes, and random drift (sampling) happens at the step of selecting zygotes for the next generation.
- Probability of fixation for finite populations: starting at frequency p:

$$u(p) = \frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}} \tag{2.32}$$

  The probability of fixation of a newly arised mutation is given by:

$$u(1/2N) = \frac{1 - e^{-2s}}{1 - e^{-4Ns}} \approx \frac{2s}{1 - e^{-4Ns}} \tag{2.33}$$

  Unless for extremely weak selection (say $4Ns < 2$), most new beneficial mutations are fixed with probability $2s$. The probability of fixation of deleterious mutations is low, however, heterozygosity of deleterious mutations can be quite high (as selection against heterozygotes is low or no selection).
- Fixation time: (not provided in the text) favorable selection will be fixed faster, however, the effect is limited. Most of fixtation time is spent at the beginning (coalescence process), however, selection is the weakest at the beginning. Thus fixation time would not be too different from $4N$.

Relative strength of the evolutionary forces: [Felsenstein, Theoretical Population Genetics, VII.10]

- Intuitions of evolutionary forces:
  - Time scale: defined according to how fast allele frequency is changed by these forces. Mutation - $1/\mu$, selection - $1/s$ (in fact, valid only when the allele reaches intermediate to large frequency), random drift - $4N$. Comparison of two time scales could suggest which one dominates, when two forces are compared.
  - Selection: most effective when the allele frequency is large. At the beginning (e.g. a new favorable mutation occurs), selection is not very effective.
- Scenarios: consider two alleles with mutation rates μ being equal between the two.

- $4N\mu > 1$: there will be more than 1 new mutation in each generation, and this is faster than what random drift or selection can eliminate (selection is weak at the beginning), thus there will be substantial polymorphism, even in the presence of selection. This leads to mutation-selection balance or simply mutation balance (when selection is weak).
- $4N\mu \leq 1$: either dominated by random drift (at weak selection, $|4Ns| < 10$) or selection (when $|4Ns| > 10$). In either case, population tend to be fixed (unless with overdominant selection that favors heterozygotes).
- An important case: weak selection of finite population (nearly neutral theory). This is applicable in small populations, e.g. human. The fate of any new mutations is: initially, dominated by random drift (a small number of a alleles, thus chance plays the important role); as the a alleles accumulate to higher frequency, selection will stop it from getting higher (at large number of a alleles, selection more important). Therefore, the probability of fixation of slightly deleterious mutations is very low, but the heterozygosity may be substantial.

2.3.1 Diffusion model of Wright-Fisher process

Reference: [Sethupathy & Hannenhalli, PRF tutorial, 2007], [Bustamante, Directional Selection and the Site-Frequency Spectrum, Genetics, 2001]

Goal: given a gene in a population under selection, mutation and random drift, how often a mutation is fixed or eliminated? What is the average time of absorption? In general, characterize the change of genetic composition.

Model: the fitness of the 2 alleles are 1, and $1 + s$ respectively. Let $f(x; , p, t)$ be the probability density function of the frequency of the derive allele (x) at time t when the initial frequency equals to p. Then it can be computed by summing over all possible daf in the previous generation conditioned on the same inital frequency p:

$$f(x + \delta x; p, t + \delta t) = \int_0^1 f(x; p, t) f(x + \delta x; x, \delta t) dx \tag{2.34}$$

Perform Taylor series expansion on both sides in δt and δx to derive the Kolmogorov forward equation:

$$\frac{\partial f(x; p, t)}{\partial t} = \frac{\partial^2 [b(x) f(x; p, t)]}{2\partial x^2} - \frac{\partial [a(x) f(x; p, t)]}{\partial x} \tag{2.35}$$

where: $a(x)dt = E(dx)$ and $b(x)dt = Var(dx)$, i.e. they are expectation and variance of the change of daf in one generation. In the neutral case, $a(x) = 0$, and $b(x) = x(1 - x)/(2N)$. Similarly, we could compute $f(x, p, t)$ by summing over all possible daf in the previous generation conditioned on the same future daf x:

$$f(x; p, t + \delta t) = \int_0^1 f(x; p + \delta p, t) f(p + \delta p; p, \delta t) d(\delta p) \tag{2.36}$$

Again do Taylor expansion and derive the Kolmogorov backward equation:

$$\frac{\partial f(x; p, t)}{\partial t} = b(p) \frac{\partial^2 [f(x; p, t)]}{2\partial p^2} + a(p) \frac{\partial [f(x; p, t)]}{\partial p} \tag{2.37}$$

Probability of extinction: the basic idea is to expression the extinction probability in terms of the function $f(x; p, t)$. In fact: the probability that daf is 0 at time t given initial daf is p can be expressed as:

$$P_0(p, t) = \int_0^{0^+} f(y; p, t) dy = F(0^+; p, t) \tag{2.38}$$

where 0^+ indicates $0 + \epsilon$ (ϵ is a very small number) and $F(x; p, t)$ is the cdf of the function $f(x; p, t)$. Integrating over dx in both sides of the Kolmogorov backward equation and plug in $x = 0^+$:

$$\frac{\partial P_0(p, t)}{\partial t} = b(p) \frac{\partial^2 [P_0(p, t)]}{2\partial p^2} + a(p) \frac{\partial [P_0(p, t)]}{\partial p} \tag{2.39}$$

Let $t \to \infty$, then LHS should be 0:

$$0 = b(p)\frac{\partial^2[P_0(p)]}{2\partial p^2} + a(p)\frac{\partial[P_0(p)]}{\partial p} \tag{2.40}$$

which could be analytically solved.

Probability of fixation: similar to the case of extinction, one could solve $P_1(p)$. Assuming selection, but no dominance, no recurrent mutation, $a(x) = \gamma x(1-x)$, $b(x) = x(1-x)$ where $\gamma = 2N_e s$ (N_e is the effective population size), this probability is:

$$P_1(p) = \frac{1 - e^{-4N_e sp}}{1 - e^{-4N_e s}} \tag{2.41}$$

Mean time of fixation or extinction: the basic idea is to relate the mean time of absorption to the fixation and extinction probabilities. Let $\phi(p,t)$ be the pdf of the absorption time, then:

$$P_0(p,t) + P_1(p,t) = \int_0^t \phi(p,t)dt \tag{2.42}$$

Or equivalently:

$$\phi(p,t) = \frac{\partial[P_0(p,t) + P_1(p,t)]}{\partial t} \tag{2.43}$$

Plug in the equations of $P_0(p,t)$ and $P_1(p,t)$ and take derivative of t:

$$\frac{\partial\phi(p,t)}{\partial t} = b(p)\frac{\partial^2[\phi 0(p,t)]}{2\partial p^2} + a(p)\frac{\partial[\phi(p,t)]}{\partial p} \tag{2.44}$$

The average time of absorption $\bar{t}(p) = \int_0^\infty t\phi(p,t)dt$, could be obtained from this equation via integration by parts (see Text):

$$-1 = b(p)\frac{d^2\bar{t}(p)}{2dp^2} + a(p)\frac{d\bar{t}(p)}{dp} \tag{2.45}$$

Define $t(p,x)dx$ be the mean time that the daf spends in the interval $(x, x+dx)$ before absorption occurs, then $\bar{t}(p) = \int_0^1 t(p,x)dx$, could solve $t(p,x)$ (see Text). Suppose $p = \frac{1}{2N_e}$ [Bustamante, Genetic, 2001], we have

$$f(x) = t(p,x) \approx \frac{1 - e^{-2\gamma(1-x)}}{(1 - e^{-2\gamma})}\frac{2}{x(1-x)} \tag{2.46}$$

where $f(x)dx$ is the expected time for which the daf is in the range $(x, x+dx)$ before absorption. If $s = 0$: $f(x) = \frac{2}{x}$. This can be shown by applying L'Hopital rule: limit of $f(x)$ when $\gamma \to 0$.

Remark:

- The derivation of $t(p,x)$ from the equation of $\bar{t}(p)$ is still unclear.
- The equation of $f(x)$ is from Bustamante. The one from the PRF tutorial has a constant N_e instead of 2 - probably wrong.

2.4 Coalescence Theory

Reference: [Nielsen & Slatkin, An Introduction to Population Genetics, Chapter 3. Hartl & Clark, Principles of Population Genetics, Chapter 3]

Motivation: understanding the patterns of genetic variations, how they are related to the underlying population genetic processes/parameters. Ex.

- Given two samples: the difference between two samples. Heterozygosity in a set of samples.

- Given many samples: how many variants are singletons, doubletons, etc. In general, the site-frequency spectrum.

Coalescence in a sample of two ($n = 2$):

- Motivation: (1) it is difficult to study the forward process, e.g. we can ask how long it takes to an individual to have more than one offspring. The number of offsprings of any indiviual follows $\text{Binom}(2N, 1/2N)$, which can be difficult to track. (2) Backward thinking is easier: every individual must have a parent, and it has probability $1/2N$ to be any specific one. So we trace the ancestors of each gene copy.

- Coalescence process: the coalescence events occur at a rate of $1/2N$ - at each generation, the probability that two offsprings share the same parent is $1/2N$. The waiting time for the coalesence is geometric distribution with mean $2N$. The continuous version is that the coal. process occurs at a rate $1/2N$, and the waiting time is expoential distribution with rate $1/2N$.

- Coalescence with mutation: infinite-site model. We make the assumption that coalecence and mutational processes are independent. This is not always true, e.g. when there is selection, the coalescence rate will not be a constant. We now have two processes in parallel then: coalescence and mutation. Under this assumption, the number of difference between two samples, π, is the number of mutations that have occurred since their most recent common ancestor (MRCA):

$$\mathrm{E}(\pi) = 2N \times 2\mu = 4N\mu = \theta \tag{2.47}$$

  where $2N$ is the expected time of coalesence, and 2μ is the mutation rate. π is thus called Tajima's estimator of θ. Furthermore we can obtain the distribution of the number of segregating sites S: the probability of $S = j$ is the probability of j mutational events occur before the coalescence. The two processes (coal. and mutation) occur at rates $1/2N$ and 2μ, respectively, thus the probability mutation occurs befores coal. is $\theta/(1+\theta)$. Effectively, we have j mutational. events before coal., and this probability is given by geometric distribution:

$$P(S = j) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^j \tag{2.48}$$

- Coalescence with mutation: infinite-allele model. Homozygosity between two copies is equivalent to coalescence before mutation. This probability depends on the relative rates of the two:

$$p = \frac{1/2N}{1/2N + 2\mu} = \frac{1}{1+\theta} \tag{2.49}$$

  The expected heterozygosity is:

$$H = \frac{\theta}{1+\theta} \tag{2.50}$$

  This can be viewed as a measure of the amount of genetic variation in a sequence: more variation, higher H. The value of H is monotonically dependent on θ: higher θ, larger H.

- Effective population size: defined based on coalesence rate. Ex. given a population of changing size, we find the effective population size s.t. the expected rate equals to the rate under the changing size. Suppose we have a population that fluctuates in sizes: it has size N_i in the p_i proportion of time. Then we have:

$$\frac{1}{2N_e} = \frac{p_1}{2N_1} + \cdots + \frac{p_k}{2N_k} \tag{2.51}$$

  So N_e is the harmonic mean of N_i's.

Coalescence in a sample of n:

- Bugs-in-a-box analogy of coalescence [Hartl & Clark]: random movement of n bugs in a box of size $2N$ (i.e. $2N$ grids), when two bugs meet at the same grid, one of them will eat another (thus reducing the number of bugs). The process of going back in time can be modelled as the dynamic process of the change of bug numbers over time.

- Coalescence tree: the basic process is that any pair coalescences with rate $1/2N$. Which pairs coalscenced at each step is random, so the shape of the tree is stochastic. We are interested in the time of coalescence (at each step and overall) to estimate the amount of genetic variations (mutations in the coalescence tree). Suppose we have k samples, the coalescence from k to $k-1$ samples happens at rate: $1/2N$ times $k(k-1)/2$ (the number of pairs). Thus the expected time of coalescence when there are k samples is:

$$\mathrm{E}\left(T_k\right) = \frac{1}{\frac{1}{2N} \cdot \frac{k(k-1)}{2}} = \frac{4N}{k(k-1)} \tag{2.52}$$

  The variance of T_k is: $\mathrm{Var}\left(T_k\right) = \frac{16N^2}{[k(k-1)]^2}$. The time to the most recent common ancestor (MRCA) is the sum of t_k:

$$\mathrm{E}\left(T_{\mathrm{MRCA}}\right) = \sum_{k=2}^{n} \mathrm{E}\left(T_k\right) = 4N \sum_{k=2}^{n} \frac{1}{k(k-1)} = 4N\left(1 - \frac{1}{n}\right) \tag{2.53}$$

  The total tree length is the sum of all lineages: from k to $k-1$, each of the k individuals has a waiting time of T_k

$$\mathrm{E}\left(T_{\mathrm{Tree}}\right) = \sum_{k=2}^{n} k\mathrm{E}\left(T_k\right) = 4N \sum_{k=1}^{n-1} \frac{1}{k} \tag{2.54}$$

  The expected number of segregating sites is:

$$\mathrm{E}\left(S\right) = \mu \cdot \mathrm{E}\left(T_{\mathrm{Tree}}\right) = \theta \sum_{k=1}^{n-1} \frac{1}{k} \tag{2.55}$$

  This leads to Watterson's estimator of θ: the number of segregating sites in n samples divided by $\sum_{k=1}^{n-1} \frac{1}{k}$.

- Site frequency spectrum (SFS): we are interested in S_i, the number of segregating sites with frequency i (i.e. occuring in i out of n samples). First, consider singletons. The number of singletons is equal to the number of mutations in the external lineage (lineage leading to a leaf node). The expected total length of external lineage is $4N$, so we have:

$$\mathrm{E}\left(S_1\right) = \mu \cdot 4N = \theta \tag{2.56}$$

  More generally, we have $\mathrm{E}\left(S_i\right) = \frac{\theta}{i}$.

- Tree shape as a function of population size: the tree shape reflects the pattern of genetic variations, and it depends on population size. When population size is small, the coalescence rate is high, thus the lineages are short. Thus one can infer the population size changes from the tree shapes (which need to be consistent with the pattern of genetic variations).

  - With exponential growth of population size, the branches near the root are shorter. In the extreme case, the tree resembles a star phylogeny: all lineges are persistent.

Coalescence with recombination:

- Ancestral recombination graph: the history of mutations and recombinations. With low mutation rates, it is possible to resolve recombination; not possible with high mutation rates (indistinguishable).

- Inference: determine the coalescence tree consistent with sample is extremely difficult - random trees will not match the sample properties, e.g. LD patterns. Sampling techniques such as MCMC have been used.

- Estimation of recombination rates: if coalesence approach is not applicable, one could compute some summary statistic that contains information of recombination rate. Ex. $4Nr$ (where r is the recombination rate) is reflected in the value of $\text{Var}(S)$, where S is the number of mismatches in every pair of nucleotide sequence (infinite-site model): recombination will reduce the variance of S.

2.5 Inbreeding, Population Subdivision and Inference of Population History

Reference: [Nielsen & Slatkin, Chapters 4-5; Felsenstein, Chapter 5; Hartl & Clark, Chapter 9,10; Laird & Lange, Chapter 3]

Motivating problems: we are generally interested in inferring the population history from extant pattern of genetic variations. Some examples:

- Given one population, infer the history of population size changes.

- Given two populations (e.g. European and African), infer the time of divergence; and/or estimating the level of genetic flow/migration between the two.

- Given many populations, inferring the history of divergence among these populations (the relationship), and the ancestral alleles of MRCA.

- For each of these problems, it is important to keep in mind that the inference (pattern of genetic variations) can be complicated by factors such as population size changes and selection.

General strategies for inferring population history and demography [personal notes; Nielsen & Slotkin, Chapter 5]

- Pattern of genetic variations reveal population history: intuitively, samples from the same (sub)population (more recent common ancestor) are more genetically similar than samples of different populations, thus one can study the genetic similarity between samples to infer history.

  - Formally, suppose we have two populations, then the average difference between samples of the same population is $\pi_s = \theta$ and between samples of the different populations is $\pi_D = 2\mu T + \theta$ where T is the divergence time.

- Explorative analysis: instead of modeling the population genetic process, study the relationship between samples to reveal their similarity and infer the history. Ex. cluster samples by their genetic similarity to find samples with common ancestry. One could have a probabilistic model, but the model may be phenomenoloical, rather than population genetic.

- Summary statistics: e.g. pairwise difference π, number of segregating sites S, F_{ST}, homozygosity. Their values often reflect the underlying structure/process. One often derives the expectation of these values under a population genetic model, and compare that with the observed values. To determine the distribution of parameters, one can use simlation (see below).

- Likelihood: directly model the genetic data to infer the parameters. The challenge is generally the unknown coalescence trees linking the samples, and may need expensive sampling approach.

Inbreeding:

- Inbred lines: the goal is to create experimental animals/plants that are genetically identical. Procedure: long-term inbreeding. If starting population is heterozygous (e.g. Aa), then at each generation, the fraction of heterozygotes will be reduced by half according to Mendel law of segregation. Thus over long time, most loci will become homozygous.

- The inbreeding coefficient of an individual is the probability that the two gene copies present at a locus in that individual are identical by descent, relative to an appropriate base population. In the base population, all gene copies are assumed not to be identical by descent.

- Inbreeding effect: given the allele frequency of the base population in HWE, p for the allele A, and the inbreeding coefficient (f), we could calculate the genotype frequencies. Thus a genotype is AA if two alleles are IBD and the ancestral allele is A, or if two alleles are not IBD and both ancestral alleles are A, and similarly for other genotypes:

$$\begin{array}{ll} AA: & p^2(1-f)+pf \\ Aa: & 2p(1-p)(1-f) \\ aa: & (1-p)^2+(1-p)f \end{array} \tag{2.57}$$

  With inbreeding, the value of f will increase over generations, and eventually reaches 1, i.e. most are homozygotes.

- Loss-of-heterozygosity (LOH) for inbreeding: heterozygotes occur when the two alleles are not IBD and are different:

$$P(Aa) = 2p(1-p)(1-F) \tag{2.58}$$

  It can be shown that the variance of X (number of a's individuals) is increased:

$$\operatorname{Var} X = 2p(1-p)(1+F) \tag{2.59}$$

Population subdivision: Loss-of-heterozygosity.

- Intuition of inbreeding and population subdivision: if for any reason, genetic exchange is limited to a smaller population, then there is always the effect of random drift, or increase of genetic relatedness (or LOH) over time. We can thus study LoH or genetic relatedness to explore/define inbreeding and population subdivision.

- Concepts: genetic differntation (difference of allele frequencies) defines subpopulations. Many populations have hierarchical population structure.

- Reduction of heterozygosity due to population subdivision (Wahlund effect): assume mating only occurs within subpopulations, then this is similar to inbreeding and reduces heterozygotes. Thus population subdivision will cause departure of HWE.

- Wright's F statistic: to measure the reduction of heterozygosity, define:

$$F_{ST} = \frac{H_T - H_S}{H_T} \tag{2.60}$$

  where H_S is the average heterozygosity of all subpopulations, and H_T is the heterozygosity of the total population. In general, if $F_{ST} < 0.05$: very little genetic differentation; 0.05 to 0.15: moderate genetic differentiation, etc.

- Causes of population subdivision: debate over natural selection (adaptation to local environment) or random drift (including fixation of different founder alleles).

- Analysis of LOH for population stratification: suppose there are K strata in the population, with allele frequencies p_k each. The heterozogosity is:

$$P(Aa) = 2p(1-p) - 2\text{Var}\, p_k \tag{2.61}$$

where $\text{Var}\, p_k$ is due to the difference of p_k across strata. The variance of X is: $\text{Var}\, X = 2p(1-p) + 2\text{Var}\, p_k$.

Population substructure: defined by the difference of allele frequencies across the population. Several types:

- Population stratification: within a population, the individuals can be subdivided into multiple geographical/ethnic groups.
- Population admixture: mixture of populations with different ancestry, often due to migration.
- Population inbreeding: due to isolation of subgroups. The inbreeding coefficient F, is the probability that the two alleles from an individual is IBD. An extreme case is self-fertilizing plant, $F = 1$.

Migration model:

- Wright-Fisher model of migration: two populations that exchange genes, and we study how AFs change at each population. Let $m_{1\to 2}$ be the probability that any gene copy in population 2 in one generation is replaced by some one in population 1, and similarly we have $m_{2\to 1}$. Let f_{A_1} and f_{A_2} be the frequencies of A in two populations. Then we have the recurrence of the allele frequency at generation $t+1$:

$$\text{E}\left[f_{A_1}(t+1)\right] = (1 - m_{2\to 1}) f_{A_1}(t) + m_{2\to 1} f_{A_2}(t) \tag{2.62}$$

The idea is that at each generation in population 1, each copy is either a migrant from population 2, or not. Similarly, we have:

$$\text{E}\left[f_{A_2}(t+1)\right] = (1 - m_{1\to 2}) f_{A_2}(t) + m_{1\to 2} f_{A_1}(t) \tag{2.63}$$

At equilibrium, we have $f_{A_1} = f_{A_2}$: eventually two population exchanging a lot of migrants will have the same allele frequencies.

  - Remark: there are variations of the basic scenarios of migration. For example, in the case of immigration, a small population is merged within a much larger population.

- Coalesence with migration: for simplicity, we assume that the two populations have equal size $N_1 = N_2 = N$, and the migration rates are also equal: $m_{1\to 2} = m_{2\to 1} = m$. We define $M = 2Nm$, as the number of migrants per generation. From the coalesence perspective, m is the probability that any gene copy, when moving back in time, is from a different popoulation. Consider the problem of $n = 2$, we want to learn the expected coalescence time. Two scenarios:
  - Two samples are from the different populations (D): the only way to coalescence is first they trace back into the same population, then coalescence. The rate of migration is m, thus the waiting time is $1/2m$ for migration, so we have:

$$E_D(t) = \frac{1}{2m} + E_S(t) \tag{2.64}$$

where $E_S(t)$ is the coalescence time when two samples are from the same population.

  - Two samples are from the same population (S): we have two processes, migration with rate m and coalescence with rate $1/2N$. The probability that coalescence occurs before migration is: $p = \frac{1}{2N}/(\frac{1}{2N} + 2m)$, and when this happens, the coalescence time is $2N$. The probability that coalescence occurs after migration is $1-p$, and when this happens, the coalescence time is $E_D(t)$. So we have:

$$E_S(t) = \frac{\frac{1}{2N}}{\frac{1}{2N} + 2m} \cdot 2N + \frac{2m}{\frac{1}{2N} + 2m} \cdot E_D(t) \tag{2.65}$$

Solving the two equations together we obtain:

$$E_S(t) = 4N \qquad E_D(t) = \frac{1}{2m} + 4N \tag{2.66}$$

Thus when the migration rate m is small, it can take a long time to coalescence for two samples in different populations. The result about $E_S(t)$ is somewhat surprising: it does not depend on m. The intuitions are: (1) it is bigger than $2N$ because there is a small probability of migrating out, which increases the coalescence time. (2) When m is small, it is a small probability to migrate out, but it takes longer to return, so the overall effect is independent of m.

- Remark: a special case is $m = 0$, and we should coalescence time equal to $2N$. However, when $m > 0$, even if it is small, it extends the coalescence time by two fold.

- F_{ST} and migration rates: to estimate F_{ST}, we need to determine heterozygosity, H_S and H_T. When two samples are from the same population, the average heterozygosity is simply the number of pairwise difference divided by the number of sites k (see below). Use $E_S(t) = 4N$, we have $\pi_S = 4N \cdot 2\mu = 2\theta$. So we have:

$$H_S = 2\theta/k \tag{2.67}$$

For H_T, again we use the coalesence time above to estimate the pairwise difference:

$$\pi_D = \left(\frac{1}{2m} + 4N\right) \cdot 2\mu = \left(\frac{1}{4Nm} + 2\right)\theta \tag{2.68}$$

This leads to calculation of H_T. The final result of F_{ST} is given by (general results):

$$F_{ST} = \frac{1}{1 + 4Nm_T} \tag{2.69}$$

where m_T is the total number of migrants into a population. For example, if we have d populations with m the pairwise migration rate, we have $m_T = (d-1)m$. The results say that F_{ST} decays as Nm_T increases: when there are significant migration, e.g. $Nm_T > 10$, F_{ST} is close to 0 - the population effectively evolves as a whole. When the migration is limited (e.g. with large geographic separation), F_{ST} could get large.

- Remark: the relation between heterozygosity and π. Previously, when studying coalescence model (without migration), we estimate heterozygosity using the infite-allele model, which compares two events, mutation and coalescence, to obtain the expected heterozygosity at $\theta/(1+\theta)$. When there are many sites in a sequence (k is large), θ per site is close to 0, thus heterozygosity is equal to θ, which is also the average number of pairwise difference, divided by the sequence length.

Divergence model:

- Process: an ancestral population splits into two populations at some point. Assumption: the population size does not change, i.e. $N_A = N_1 = N_2 = N$, for simplicity. The coalescence process is: each lineage is traced back to the divergence time, after that, the standard coalescence process. Given two samples, the coalescence time:

$$E_S(t) = 2N \qquad E_D(t) = T + 2N \tag{2.70}$$

where T is the divergence time. From this, we estimate the pairwise difference:

$$\pi_S(t) = \theta \qquad \pi_D(t) = (T + 2N) \cdot 2\mu = 2\mu T + \theta \tag{2.71}$$

The term $2\mu T$ is the amount of genetic difference due to divergence.

- F_{ST}: we first obtain $H_S = \pi_S = \theta$ (ignore the term $1/k$), then we have:

$$H_T = \frac{1}{2}(\pi_S + \pi_D) = \theta + \mu T \tag{2.72}$$

From these, we obtain:

$$F_{ST} = 1 - \frac{H_S}{H_T} = \frac{T}{T + 4N} \tag{2.73}$$

So when T is small, F_{ST} is close to 0 (a single population); when T is very large, F_{ST} is close to 1.

Isolation by distance:

- Observation: for many species, F_{ST} increases with distance in a way s.t. $F_{ST}/(1 - F_{ST})$ is linear wrt. distance. This is called "isolation-by-distance".
- Explanation by the migration model: if the migration rate m_T is a linear function of the distance (decreases with longer distance), this would explain the pattern. Alternatively, one can imagine many local populations, while migration occurs only between adjacent populations - this is called the Stepping-stone Model.
- Explanation by the divergence model: imagine a series of divergence events, each of which results in one spopulation splitting apart from another. And the new population occupies an area close to the parent population. This would lead to a linear relation between divergence time and distance, and the observed pattern of F_{ST}.
- **Remark**: in population genetics, there may be multiple scenarios that produces similar pattern of genetic variations. The reason is, for example, there may be multiple scenarios leading to change of effective population size (reduction of actual number, or limiting of genetic flow among individuals, etc.), which manifest as similar genetic patterns (increase of genetic relatedness).

Allele frequency spectrum under migration/divergence model [personal notes]:

- Problem: The model provides the expected relation between F_{ST} and migration/divergence parameters, e.g. divergence time. However, the pattern of genetic variation can be more complex, e.g. the same average F_{ST} in a region could be a result of: many loci with small F_{ST}, or a smaller number of loci with large F_{ST} (say fixed). How do we relate the frequency spectrum (e.g. the number of fixed sites at different alleles) with the migration/divergence parameters?
- Analysis of FS from coalesence: the time/position of mutational events in the coalesence tree determines its AF. Suppose we have n samples from one population, the time to MRCA is about $4N(1 - 1/n)$. If there are no mutations in this time, the locus remains fixed at ancestral allele. The time of divergence is T, if $4N(1 - 1/n) > T$, the mutation happens before divergence, thus the site will appear fixed as the derived allele.
- Analysis of FS from forward models: e.g. given two populations diverged at timet T, we want to estimate the number of sites that are fixed in two different alleles in two populations (one ancestral, one derived). The idea is that: we have μT new mutations in a region since divergence, and each mutation has probably $1/2N$ to be fixed, so roughly the number of fixed mutations is $\mu T/2N$.

Parametric bootstrapping to determine the distribution of estimators [Nielsen & Slotkin, Chapter 5]: suppose we estimate some parameter θ using summary statistics (e.g. Tajima's estimator for $\theta = 4N\mu$), and let $\hat{\theta}$ be the estimator

- Sample coalesence tree linking the samples: at each step, randomly sample time t from expoential distribution, then sample the two nodes to coalescence.
- Sample mutations along the tree: Poisson process with rate μ (or using $2N$ as time unit).

- Compute $\hat{\theta}$ from each sampled tree (with mutations).

Species tree and coalescence tree: (species could also mean populations within a species)

- Reciprocal monophyly: suppose we have multiple samples per species, but the divergence event is very ancient. Then at the point of divergence, all the samples within a species have coalescenced (thus can be treated as one individual). In this case, the species tree exactly matches the coal. tree.
- Incongruence (Incomplete Lineage Sorting): when the MRCA of all samples occurs before the divergence event (taking a long time back to reach the MRCA), then the species tree is incongruent with the coal. tree of all samples.
- Example of Incomplete Lineage Sorting: human-chimp-gorilla, the short divergence between MRCA of all three species to the ancestor of (human, chimp) can cause the problem.

Infering population history from gene trees: suppose we are inferring the history of multiple populations of a common ancestor (when the divergence occurs, the effective population size, migration, etc.)

- Analysis using divergence model: we know that $\mathrm{E}_S(t) = 2N$ and $\mathrm{E}_D(t) = T + 2N$, when T is large (or N small), this leads to reciprocal monophyly; so the species/population tree is easy to infer. We could also say that population bottleneck (small N) eliminates ancestral variation.
- Incomplete lineage sorting can be resulted from short divergence time (relative to $2N$), or the gene flows.
- Example: from human gene tree, we are testing the hypothesis of out-of-Africa or multi-regional hypothesis (small N of non-African group with some gene flow between African and non-African group).
- Effect of recombination: instead of one coalescence tree, we may have multiple trees for different regions.

Likelihood method for inferring population history:

- Felsenstein equation: let X be data, G be the genealogy, and θ be the parameters of the model. Then the data can be used to infer the underlying processes:

$$P(X|\theta) = \int_G P(X|G)P(G|\theta) \tag{2.74}$$

  where $P(G|\theta)$ is given by the coalesence tree, and $P(X|G)$ is a result of mutations.
- Inference: MCMC to sample G. Another strategy is Approximate Bayesian Computation (ABC). The intution is that to sample $\theta|X$, MCMC uses $P(\theta|X) \propto P(X|\theta)$, while ABC assess how good θ agrees with X based on summary statistics (simulate data from θ, compute smmary statistics, and see if they are close to the observed values to decide if θ should be accepted). The advantage of ABC is that it does not directly sample the trees G.

Inference of population structure through analysis of genetic similarity:

- Population assignment by AFs: different populations differ in AFs. So we can assign a sample to population through the AFs of its markers. See the STRUCTURE algorithm below.
- PCA: imagine each population has a "canonical" profile (PC direction), and then each individual haplotype can be expressed as a linear combination of these profiles (the coefficients are PCs of that individual: the population composition).

STRUCTURE algorithm: [Pritchard & Donnelly, Genetics, 2000]

- Assumptions: alleles are in HWE, and in LE with each other.
- Model without admixture: suppose there are K populations. The probability of the genotype of an individual, X_i is given by $P(X_i|Z_i, P)$, where Z_i is the population index of the individual and P is the allele frequencies of the populations. Let $x_l^{(i,1)}, x_l^{(i,2)}$ be the genotype of the i-th individual of the l-th locus, we have:
$$P(x_l^{(i,a)} = j|Z, P) = p_{z_i lj} \tag{2.75}$$
where p_{klj} is the frequency of allele j at locus l in the population k.
- Model with admixture: define $q_k^{(i)}$ be the proportion of the individual i's genome that originated from population k, and change Z to: $z_l^{(i,a)}$ be the population of origin of the allele $x_l^{(i,a)}$. Then:
$$P(x_l^{(i,a)} = j|Z, P, Q) = p_{z_l^{(i,a)} lj} \tag{2.76}$$
$$P(z_l^{(i,a)} = k|P, Q) = q_k^{(i)} \tag{2.77}$$
- Inference: Dirichlet prior of the allele frequencies, and sample Z and P (and Q for the model with admixture) through MCMC.

Spatial locations of ancestry [Anand Bhaskar, May, 2016]

- Problem of PCA: genetic distance does not match the spatial distance. Show by simulation. PCA minimizes the total pairwise genetic similarity.
- Related work: SCAT, SPA, SpaceMix.
- Model: let μ_l be the mean AF of locus l (uniform across space). Assume actual AF at two locations correlate with each other with:
$$\text{Cov}\left(Q_l(z), Q_l(z')\right) = \eta(z - z') \tag{2.78}$$
where η is some function. Ex. isotropic decay, then covariance decays at rate $\exp[-\alpha_1\|z - z'\|^{\alpha_2}]$. Another model is directional decay. Idea is to do Taylor expansion so that locally, genetic similarity is a simple function of spatial distance.
- Inference: use ideas from manifold learning. Basically, find pairs with smallest genetic distance, then create the spatial distance; then stitch the pairs to obtain the spatial distance of distantly related individuals.
- **Remark**: how the parameters of the Gaussian process model relates to the population genetic processes?

Introgressed tract of Nean. ancestry [Matthias Steinrcken, NHS meeting, 2018]

- About 3% of Nean. genome introgressed with human. To detect introgression: train CRF using African, European and Nean. as training; $S*$ statistics that detect divergence using panels.
- HMM: hidden states are ancestry, and transition represents introgressed Nean.
- Observation: less introgression in chr. X. Likely due to selection: DMI?

Spatial distribution of deleterious rare alleles [Dan Rice, Nov, 2019]

- Observed pattern of rare deleterious alleles: plot AF vs. location, for CVs, fluctuation across mean in large regions; for RVs, much more localized.

- Model of spatial AF of deleterious alleles: a mutant follows birth-death process, where the death rate (due to negative selection) is higher than birth rate. Also spatial random move. This can be modeled by a PDE, let $p(x, r, t)$ be the PDF of AF at x, location r at generation t. At a given location, we can use W-F process, and we just need to additional terms to capture spatial diffusion: in-flux from neighbors and out-flux to neighbors.
- Solving the PDE: marginalize locations (focus on AFs), and define the MGF of $p(x)$. Solve the MGF. Results: the spatial scale of AF distribution depends on $\sqrt{D/s}$, where D is diffusion rate and s selection coefficient.
- Applications: (1) Test selection acting on RVs: derive neutral distribution. (2) Rare variant association test: adjusting for population structure due to different spatial locations.
- Q: do we need selection to explain local spatial distribution of RVs?

2.6 Recombination and Linked Selection

LD in finite populations: [Hartl, Principles of Population Genetics, Section 9.2; Yang, Introduction to Statistical Methods in Modern Genetics, Appendix A]

- Intuition: even without mutation and selection, LD could emerge in finite populations. To see why, we view the haplotypes as single polymorphic sites, and notice that: (1) Why not LE: in finite populations, certain haplotype tends to be fixed, and this pushes the population away from LE; (2) why not fixation: the frequency of haplotypes change at each generation, in fact, high frequency haplotypes may be more likely to be replaced due to recombination. The situation is similar to mutation-drift balance.
- Model: consider two loci u and v, and define I as the probability that any two randomly selected gametes are IBD at u and no recombination has ever happened between u and v. We first claim without proof (see Yang) that: $I = r^2$. We could derive the recurrence equation of I, similar to the derivation of homozygosity under mutation-drift balance:

$$I_{t+1} = (1-c)^2 \left[\frac{1}{2N} + \left(1 - \frac{1}{2N} \right) I_t \right] \tag{2.79}$$

  where c is the recombination rate between u and v. At equilibrium, the expected LD measure r^2 is given by:

$$E(r^2) \approx \frac{1}{1+4Nc} \tag{2.80}$$

Consequences of linkage: because of linkage, the evolution of one site could influence the evolution of a neighboring site.

- Clonal interference (in asexual organsisms) and Hill-Robertson effect: even if a favorable mutation occurs in one site, it may not get fixed because of a perhaps stronger favorable mutation in a neighboring site.
- Genetic hitchhiking (selective sweep): Suppose a recent adaptive mutation occurs in the locus A, and is fixed very fast in the population, then an allele at a tightly-linked locus will also be fixed in the population, even if this allele may be neutral or deleterious. The consequence: a small region around the favored allele will be overrepresented in the population.
- Background selection: suppose we analyze a site under neutral mutation, but a neighboring site is under negative selection, then some recombinations will be elimiated by selection. Thus the overall effect on the site of interest is effectively a reduction of population size (which reduces the level of polymorphism).

- Evolutionary benefits of recombination: suppose there are multiple favorable alleles in the population, with low recombination or asexual organisms, they must be fixed in sequential order, thus effectively there is interference among favorable alleles. With recombination, different favorable allels can be brought together in rapid succession.

Testing selection by LD patterns and haplotypes: [Hartl, Section 9.2]

- Idea: selection of one site can change the polymorphism of the neighboring sites through hitchhiking. However, LD generated by hitchhiking persites for a relatively short amount of time, on the order of $0.4N$ generations, so the test is only applicable when selection is strong and recent, and linkage sufficiently tight.

- Example: (Figure 9.8) a sequence segment with one highly similar haplotype in multiple copies, while other haplotypes are much more variable, suggesting that this haplotype is under selective sweep.

Selective sweeps [Graham Coop, Population Genetics notes, Chapter 8]

- Fully linked variants during selective sweep: suppose we have a new beneficial allele with selection coefficient s, it takes $\tau = 4\log(2N)/s$ generations to fixation. This time is usually very short, comparing with $2N$. The effect is: loss of diversity after sweep, then slowly recover.

- Recombination during sweep: Figure 8.3, genealogy tree (Figure 8.4): the ones from the selected lineage coalescence (short tree), and other lineages coalescence further back in time, with possible variants/mutations in the branches. The effect is: loss of diversity close to the selected variant, and recover diversity moving away from the selected variant (Figure 8.5)

- Quantitative analysis of loss of diversity due to selective sweep: consider two haplotypes, and we need to compute $\mathrm{E}(\pi)$, this is given by T_2, the coal. time of the two. This depends on which one of them or both are selected lineage: two main possibilities (1) One selected and one neutral; (2) Both are selected. The coal. time is $2N$ and τ respectively. The probabilities of (1) and (2) are given by p_{NR}, probability of no recombination. Note that since $\tau << 2N$, most of the diversity is from (1), when recombination occurs. Results: the diversity at recombination distance r is given by:

$$\mathrm{E}(\pi_r) = 4N\mu(1 - e^{-r\tau}) = \pi_0(1 - e^{-r\tau}) \tag{2.81}$$

  where π_0 is the expected diversity if there is no linked selection.

- Loss of diversity vs. recombination distance: intuitive analysis. Diversity is 0 if both haplotypes are the selected one. Diversity is $4N\mu$ if one haplotype is new. The probability of this is just probability of recombination. Note that, the expected number of recombinations between selected variant and the variant at distance r is $r \cdot \tau$. So probability of recombination is given by $1 - e^{-r\tau}$.

- The extent of selective sweep: $s = 0.1\%$ will reduce diversity over 10's of kbs and $s = 1\%$ over 100kb.

- Other signals of selective sweep: (1) Fay-Wu test: at intermediate distance, excess of high frequency alleles (from the selected haplotype). (2) As neutral diversity recovers, excess of low frequency alleles. (3) Local peaks of F_{ST} between differentiated populations.

- Soft sweep: Figure 8.12, multiple mutations selected, none completely sweep; or single standing variant (already multiple haplotypes at the time of selection) also lead to incomplete (soft) sweep.

- Genomewide patterns of diversity due to linked selection: positive correlation of neutral diversity and recombination rates.

Background selection (BGS) [Graham Coop, Population Genetics notes, Chapter 8]

- BGS on fully linked variants: suppose we have a variant under negative selection. At this locus, we assume mutation-selection balance, so at equilibrium, we have: $q = \mu/(hs)$, where hs is selection on heterozygotes. Only $1-q$ haplotypes contribute to diversity since q deleterious haplotypes will be lost. The effect is similar to reduced population size by $1-q$. Putting this together, we have:

$$\mathrm{E}(\pi) = \pi_0(1-q) = \pi_0\left(1 - \frac{\mu}{hs}\right) \tag{2.82}$$

  where π_0 is the expected diversity without BGS.

- BGS with recombination: suppose we have a neutral locus that is at recombination distance r away from a BCS locus, the diversity is:

$$\mathrm{E}(\pi) = \pi_0\left(1 - \frac{\mu hs}{2(r+hs)^2}\right) \tag{2.83}$$

  More generally, a netural locus under the influence of multiple BGS loci at distance r_i, has diversity:

$$\mathrm{E}(\pi) = \pi_0 \prod_i \left(1 - \frac{\mu_i h_i s_i}{2(r_i + h_i s_i)^2}\right) \approx \pi_0 \exp\left(-\sum_i \frac{\mu_i h_i s_i}{2(r_i + h_i s_i)^2}\right) \tag{2.84}$$

- Approximation of BGS effects on linked variants: consider a large genomic region, with a neutral locus in the center with total recombination $R = \sum_i r_i$, and $U = \sum_i \mu_i$. We assume equal mutation (deleterious mutations only), recombination rates and strength of selection across all positions in the region. Under these assumptions:

$$\mathrm{E}(\pi) \approx \pi_0 \exp(-U/R) = \pi_0 \exp(-\mu_{BP}/r_{BP}) \tag{2.85}$$

  where μ_{BP} and r_{BP} are per base mutation and recombination rates. Average r_{BP} is about 2cM/Mb $= 2 \times 10^{-8}$.

- Genomewide pattern of BGS and B-factors: [McVicker, 2009] fits BGS with the genome, and estimates the effect of BGS (B-factors) across genome. The estimated mutation rate is high 7.4×10^{-8}.

- Distinguishing hitchhiking from BGS: hitchhiking has systematic effects on SFS, distorting it towards rare minor alleles (slow recovery after sweep). An example (Figure 8.19): dip of diversity around syn. sites driven largely by BGS, but some fraction of dip around non-syn. substitutions from positive selection/sweep.

- Analysis (personal notes) of how BGS affects SFS: suppose we have a neutral locus A, linked with BGS locus B. Our goal is to estimate SFS at A from PRF, and we need to study the survival time of an allele at A. Suppose we have AF at A equal to x, we need to derive the new diffusion equation of x, by modifying $\mathrm{E}(\Delta x)$ and $\mathrm{Var}(\Delta x)$ in one generation to incorporate deleterious variants at B.

Recombination rate variation within and between species [Molly Przeworski, 2017]

- Part I: Recombination hotspots: double-strand breaks (DSB), orders of magnitude more often than average.

- Hotspots are not conserved: e.g. human vs. chimp, about 10% are conserved. Also variations across individuals.

- Key role of PRDM9: N terminal Zn finger (DNA binding). Can predict PRDM9 binding sites from motif. Another domain has histone methyltransferase activity, H3K4me3, which often precedes recombination (help recruit recombination machinery). However, the mark itself is not sufficient to attract the machinery, so the recombination hotspots are not enriched near promoters.

- PRDM9 motifs are rapidly lost: if there are two alleles, one prefer PRDM9 binding, it will lead to differential transmission (high binding allele less likely to transmit) - meiotic drive. PRDM9 Zinc finger also evolves rapidly.
- Recombination (DSB) is reduced near TSS. This pattern is lost in PRDM9 KO mice.
- Part II: What drives recombination in species without PRDM9, e.g. chicken?
- Experiment: Sequence finches (N = 1-20) - no PRDM9, then infer recombination rates. (1) Found hotspots near promoters. (2) Hotspots are much more shared, 70

- Background: GC-biased gene conversion: in DSB repair, heteroduplex forms (AG), but is more likely to be fixed as G.
- Elevation of GC content in hostpots, enriched near CG islands. Also find similar pattern in yeast.
- Model: in species lacking PRDM9, recombination machinery is more likely to target promoters (e.g. open chromatin or motifs). And they tend to be conserved because of selection on promoter function.
- PRDM9 in 225 vertebrates: lost in amphibilians, and several cases, some domains (KRAB) are lost. Zinc finger evolves rapidly only if the gene is intact. In species with partial ortholog (fish), H3K4me3 correlates with recombination rate. PRDM9 binding motif (predicted from sequence) not associated with H3K4me3; and PRMD9 motif not associated with high recomb. rate. So the partial PRDM9 does not direct recomb. in fish.
- Part III: Evolutionary consequences of recombination rate hotspots?
- Background: Hybridization between species is very common.
- Human-Nean: there is some desert of Nean. ancestry. Model: D-M incompatibilities, Nean has low Ne thus more deleteious mutations.
- Recombination rate and minor parent ancestry: low recombination rate, the minor parent ancestry is lost because of DMI, and some retained in high reocmb rate regions. The pattern of correlation between recombination rate and ancestry is observed in Human-Nean.
- Lesson: PRDM9 function is related to histone modification. Meiotic drive can lead to rapid evolution of PRDM9 binding sites.
- Lesson: relationship between recombination rate and ancestry, in low recombination rate regions, DMI will remove DMI-incompatible regions (higher load of deleterious mutations due to small population).
- Q: Selection of recombination rates: optimal level? If too high, break the good combinations too often.
- Q: In species without PRDM3, why recombination machinary evolves to recognize H3K4me3 mark or promoters? Is this due to CpG islands?
- Q: In birds (no PRDM9), are hotspots not close to promoters also highly conserved? Answer: not clear, because of uncertainty of defining hotspots.

The Effect of Strong Purifying Selection on Genetic Diversity [Cvijovic and Desai, Genetics, 2018]

- Model assumption: a neutral site, perfectly linked to a site with deleterious mutation. The mutation is strong Ns at the level of 1000.

- BGS can distort SFS (Figure 1): (1) Excess of rare variants: even strong selection cannot purge deleterious alleles instantly. So at very low AF, the SFS is similar to neutral expectation under the original population size. (2) Excess of high frequency variants: similar to positive selection, once the alleles reach high frequency (few ancestral alleles), drift will dominate. Neutral alleles get fixed in a sweep-like fashion because they carry on average fewer deleterious mutations than the wild type.
- Implication of the distortion: not obvious in small samples, but the effect on high and low ends of SFS can be large with large sample sizes.
- Discussion: the model assumes single selection coefficient in BGS sequence s. The target sequence is neutral, but if negative selection, can be easily accommodated. If the BGS sequence has large variation of s, need more work.

Natural selection interacts with recombination to shape the evolution of hybrid genomes [Science, 2018]

- Hybrid genomes, assess the ancestry of two parents. Ancestry from the minor parental species is more common in regions of high recombination and where there is linkage to fewer putative targets of selection.
- Model: ancestry from the minor parental species is more likely to persist when rapidly uncoupled from alleles that are deleterious in hybrids.

2.7 Neutral Theory and Detecting Natural Selection

Reference: [Hartl & Clark, Principles of Population Genetics, Section 4.4, 4.5; Nielsen & Slatkin, Chapters 8-9]

Principles of Detecting Selection:

- Strategy: the goal is to infer the evolutionary forces, in particular selection, on protein or DNA sequences, from the pattern of polymorphism. The general strategey is to derive these patterns under the assumption of no selection, and the departure in real data from these expected patterns would indicate selection or other evolutionary forces. This departure can often be measured as the amount of genetic variation, for example, positive selection in a sequence can reduce its genetic variation.
- Measures of genetic variation: many measures such as H, heterozygosity, π, pairwise difference, S, number of segregating sites, and so on. More generally, we can use patterns of polymorphism to measure genetic variation: (1) allele-level pattern: the frequency of different alleles; (2) site-level pattern: how many sites are segregating and their frequencies; the number of different sites among pairs of sequences; etc. Two common patterns: allele frequency spectrum and site frequency spectrum (derived allele frequency, DAF, of each site).
- Influence of natural selection on the patterns of polymorphism and divergence: basis of statistical test of selection.
  - Within-species polymorphism: generally, selection affects SFS, which may change the metrics of genetic variation, such as π. In general, S is less sensitive to selection, while π is. Selective sweep: reduce intermediate frequency alleles, leading to a reduction of π. Balancing selection: increase π. Negative selection: excess of rare variants, thus reducing π and reduce S to a smaller extent.
  - Within-species haplotype: another signal of selection, nearby regions of selected sites show reduction of homozygosity.
  - Genetic differentiation between populations/F_{ST}: when a locus is adapted to a local environment, it may change AF significantly between population. This is a signature of positive selection. Ex. F_{ST} is elevated in the lactase locus, relative to adjacant region in European samples. Another example, AF difference at EPAS1 gene between Han and Tibetan.

  - Divergence: generally, positive selection makes it easier to reach fixation while negative selection makes it harder.
  - Combining patterns of polymorphism and divergence: they may result from the same underlying selective force, thus can be combined to increase the power of detecting selection.

- Application of a test depends on specific problems/contexts: e.g. when finding genes underlying population difference within a species, choose SFS type of tests. When testing genes being selected in a local population, choose AF-based tests such as F_{ST}. When searching for genes underlying species difference, use dN/dS test.

- Defining expected pattern of genetic variation: this generally depends on mutation rates and demography. While the absolute values can be difficult to obtain, sometimes we can still utilize information in them. For example, the relative mutation rates between sites (e.g. syn. vs. nonsyn.) and genes are possible to find.

- The importance of controls: often the expected pattern of genetic variation under neutrality is not known, e.g. $\theta = 4N\mu$ depends on the local mutation rate, which is often unknown. So we will need to define a control/contrast to test for selection. This is a very general idea that has many applications in population genetics and comparative genomics:
  - d_N/d_S test of selection using divergence data.
  - Variant density (e.g. number of singletons) in synonymous vs. nonsyn. sites: could be used to estimate selection. Intuitively, few nonsyn. sites would suggest negative selection. Under neutrality, the ratio of the two is determined by mutation rates.
  - Constraint in non-coding sequences: choose neutral sequences as those that are non-DHS in any known cell types.

- Confounding factors on detecting selection: demography (change of population size, in particular). For example, negative selection will create an excess of rare variants, but population expansion will also create excess of rare variants.

Neutral theory and negative selection:

- Neutral theory: postulate that most of the observed genetic difference is neutral, instead of adaptive. The theory does not exclude the role of negative selection, which does not contribute much to the observed difference.

- Negative selection and amount of genetic variation: in theory, negative selection could limit the amount of genetic variation. Ex. we consider θ (or H): with negative selection, the effective mutation rate (the effective portion of sequences that are allowed to mutate) is lower than the actual rate. However, when we compare the test sequence with the control (e.g. inter-species divergence), as long as the proportion of negatively selected sequence is the same, θ will be identical.

- d_N/d_S test and why it is not enough: the main limitation is that the selection within species may have changed since divergence. Another challenge for this test is that: it is possible that only a small proportion of sites are under positive selection while most of the sites are under negative selection.

- Main tests for negative selection: Intra- and inter-species variation (MK test). Intra-species: fraction of rare variants, variant density (more useful for strong selection). Tajima's D and its extensions are not designed for testing negative selection. Remark: need a better way to integrate information from density and FRV.

Pattern of genetic variations following positive selection: genetic hitchhiking

- Selective sweep: suppose we have a new advantageous mutation, then as it is getting fixed, the alleles linked to the mutation will also be fixed. Selective sweep will reduce genetic variation (H). Another way to understand it: because of selection, the coalesence time is shorter, thus fewer mutations. Recombination will reduce the effect of selective sweep: the further away from the selected locus, the less selective sweep. It can be shown that the recombination c and selection cofficient s determines the extent of sweep: at $c > s$, hardly any selective sweep.
  - Intuitive analysis: suppose we have complete LD. Initially, we have an advantageous allele a, linked to a neutral site b. As f_a increases from directional selection, f_b also increases, and eventually both reach fixation.
  - Recombination: with recombination, the efficiency of f_b increase will be lower because with any new a alleles, only a fraction of them have the b allele.
- Partial sweep: advantageous allele reaches intermediate frequency but has not been fixed. The pattern is: polymorphism at sites close to the advantageous alleles is reudced, and strong LD is created between alleles on the same chromsome as the advantageous allele. Ex. G6PD (Figure 6.1 in [Nielsen & Slatkin]).
- Associative overdominance: suppose we have overdominance (heterozygous advantage) at Aa, we ask what is its effect on a neighboring neutral site B/b. Intuitively analysis:
  - Complete LD: suppose we start with a single allele a, initally, f_a will increase becuase of heterozygous advantage, but it will reach equilibrium $\hat{f}_a$. Because of complete LD, the b allele also increases in frequency and reach stable polymorphism. The overall consequence is that the genetic variation at B/b locus increases.
  - Recombination: As f_a increases, f_b also increases, but as a reaches substantial fraction, b will swithc its partner from a to A due to recombination. Formally, we can think of two populations A and a, and there is migration of b between the two populations, the rate of which depends on recombination rate and f_a. This allows us to formally analyze heterozygosity at the b locus using the coalesence model of migration.

  The overall consequence of associative overdominance is the increase of genetic diversity (heterozygosity) at the linked neutral loci, opposite to the pattern of selective sweep.

Tests of selection from population genetic and inter-species data:

- HKA test: let S be the number of segregating sites within the population, intuitively, with positive selection, S should be reduced. We use F, the number of fixed difference between two species, to assess the change of S. For neutral sites, $\mathrm{E}(S)/\mathrm{E}(F)$ should be constant:

$$\frac{\mathrm{E}(S)}{\mathrm{E}(F)} = \frac{4N\mu\sum_i 1/i}{2T\mu + 4N_A\mu} \tag{2.86}$$

  where T is the divergence time and N_A the ancestral population size. So choosing two loci, we compare S_1, F_1, S_2, F_2 and do a χ^2 test. If S/F for a locus is significantly below the other one, this suggests selective sweep.
- MK test: similar to the idea of HKA test, instead of comparing two loci, we compare the syn. and nonsyn. sites of a region/gene. Specifically:
  - Under positive selection (selective sweep): F_N is increased (mutations get fixed across species) and S_N is reduced (less genetic variation), so $S_N/S_S < F_N/F_S$.
  - Under negative selection: (1) Moderate/weak selection: F_N is reduced (harder to fix mutations), and S_N not change much as moderate/weak selection does not affect much the number of variants, so $S_N/S_S > F_N/F_S$. (2) Strong negative selection: both F_N and S_N will be reduced (thus strongly deleterious mutations do not contribute), so MK test cannot detect.

- Remark: The measure of genetic variation is S, this does not take all the information in the data.

Testing selection [John Novembre, HGEN 469]

- I. Diversity vs. recombination rate: positive correlation. Divergence vs. r: no correlation, ruling out mutation. Neutral theory cannot explain this pattern. Possible explanation: genetic hitichiking; or background selection.
- Effect of positive selection on linked loci: use figure to explain selective sweep. Reducing nearby diversity. After fixation, then accumulation of mutations, increasing diversity. However, number of variants recover faster than π - signature of positive selection. The difference can be tested via Tajimas D.
- Explaining r-μ correlation using positive selection: low recombination rate region, larger impact of hitchhiking, thus lower diversity.
- Background selection: effective population size in a region reduced. Similar effects of reducing diversity. Thus low recombination regions, larger impact from negative selection and lower diversity in nearby regions.
- To distinguish the two: the site frequency spectrum could be different. In positive selection: excess of singletons. However, background selection can also result in similar, negative Tajimas D.
- Demography and selection: Tajimas D is negative with growing populations. How to differentiate the two?
- Fit a demography model using neutral data: then estimate the distribution of Tajimas D. Alternatively, fit the empirical Tajimas D from neutral sites.
- References for detecting SFS changes (extension of Tajimas D), SweepFinder, Nelson, 2005.
- MKPRF: combination of divergence and SFS. Estimation of distribution of fitness effects.
- II. Haplotype patterns for selection. Voight 2006. Idea: in a selected site, group by the alleles, at the site, homozygosity = 1, nearby regions show reduction of homozygosity. But the reduction is slow with sweep, comparing with neutral. Idea: assess the AUC of haplotype homozygosity, and normalize by neutral. IHS test.
- III. AF difference across populations. F_{ST} based scan.
- IV. comparison of methods. IHS better at detection of partial sweep. Detecting different processes: e.g. recent or older.
- CMS: Sebati. Combining multiple signatures.
- Lesson: population geneticists use genetic diversity as a main tool: how various forces affect diversity. In selective sweep, analysis of how diversity changes because of selection, and then how it changes due to mutations. Different patterns of different measures of diversity: number of variants and diversity, the latter is much more sensitive to selection.
- Remark: about teaching (1) ask questions, e.g. why positive correlation between recombination and mutations? Motivates the work. (2) Use figures, e.g. to show the dynamics/ change over populations. (3) Link to whats discovered. Ex. background selection, link to $4N\mu$.

Detecting negative selection using polymophism data [personal notes]:

- MK test: more powerful than Tajima's D and related test for detecting negative selection [Zhai & Slatkin, MBE, 2009]. However, using divergence data.

- INSIGHT: under negative selection, no intermediate and high-frequency variants. Weakness: not based on mutation rates, thus lower power; and strong assumptions.
- dN/dS test and variations: compare the number of NS and S sites within a gene, and compare with expectation. The expectation can be obtained from mutation rates (approximately 3.9:1).
- Mutation rate based: RVIS, ncRVIS and Samocha2014. Obtain the number of NS sites in a gene, and compare with expectation based on mutation rates (Samocha2014, ncRVIS) or gene length/total num. of variants (RVIS). For RVIS and ncRVIS: consider only the number of common NS sites - they are more sensitive to negative selection.
- Number of common variants reflects the strength of negative selection: we use PRF model to analyze the difference of SFS under neutrality or negative selection. We recall that the expected number of polymorphic sites with frequency x is $g(x) = 2N_e\mu f(x)$ and

$$f(x) \approx \frac{1 - e^{-2\gamma(1-x)}}{[N_e(1 - e^{-2\gamma})][x(1-x)]} \tag{2.87}$$

  where $\gamma = 2N_e s$. The ratio of $f_\gamma(x)$ and $f_0(x)$ (neutrality) is:

$$\frac{f_\gamma(x)}{f_0(x)} = \frac{1 - e^{-2\gamma(1-x)}}{1 - e^{-2\gamma}} \tag{2.88}$$

  Suppose we have $\gamma = -5$ and $x = 0.01$, the ratio is 0.9; but at $x = 0.2$, the ratio is 0.13! So negative selection strongly limits the number of common variants, but has little power of removing rare and very rare variants.

Molecular signatures of natural selection [Nielsen, ARG, 2005]

- Predictions of neutral theory:
  - On inter-species divergence: new mutations arise at the rate $2N_e\mu$, and each of them, if neutral, has a probability $1/2N_e$ to be fixed. Thus the number of fixed mutations between two species is proportion to μ, the mutation rate.
  - On intra-species polymorphism: both the number of segragating sites (S) and the nucleotide diversity (Π) is proportional to $\theta = 4N_e\mu$.
- Estimating selection and test of neutrality: general strategies:
  - For a gene to be tested, comparison of nonsyn. mutations vs. syn. mutations (neutral).
  - Using polymorphism data: under the neutral theory, the number of segregating sites (or average diversity/heterozygosity) is proportional to the mutation rate (roughly the length of sequence). Thus the number of segregating sites in syn. vs. nonsyn. sequences should be equal to the ratio of mutations of the two types of sequences.
  - Using polymorphism data: site-frequency spectrum (SFS) (Nielsen05, Figure 2). The tests such as Tajima's D identify the excess of rare variants (relative to neutral expectation), which is due to natural selection. However, the departure of SFS from neutral expectation can be also due to population growth, especially for human (inflated rare variants from recent expansion).
  - Extension of Tajima's D test: (1) Fu & Li extended this test to take information regarding the polarity of the information into account by the use of an evolutionary outgroup. (2) Fay & Wu suggested a test that weights information from high-frequency derived mutations higher.
  - Using divergence data: under the neutral theory, the subsitution rate between two species should be proportional to the mutation rate (the substitution rate is related to the number of fixed sites under a molecular evolution model, such Jukes-Cantor model). Thus if the gene is neutral, the d_N/d_S should be equal to 1.

- Test of neutrality using both divergence and polymorphism data (MK test). The general idea can be applied in different ways: e.g. the polymorphism may be measured using diversity (or average heterozygosity).

The genetics of human adaptation: hard sweeps, soft sweeps and polygenic adaptation [Pritchard, Curr Biol, 2010]

- Examples of human recent adaptation: height, e.g. pygmy, may be driven by food limitation or humidity. High altitude adaptation: to low O2 during pregnancy. Pigmentation: half a dozen genes found with strong genetic signals of selection.

- Conflicting patterns of selection learned from genome-wide data: (1) Genome-wide scan of selection: different methods often do not agree. (2) AF difference between populations: large AF diff. tend to be in genic than non-genic regions, suggesting selection (Figure 1A). (3) However, differentiation of AS closely matches historical population, suggesting drift (Figure 1B); and high F-ST SNPs do not show haplotype patterns of selection (Figure 1C), similar XP-EHH to random SNPs, and are often not fixed.

- Possible explanations: soft sweep model, where sweep occurs in a standing variant or multiple mutations. Consider mutation target sizes (number of possible beneficial mutations): show that when it is large, say > 100, usually there will be standard variations when a new selection force arrives. Soft sweep does not produce classical pattern of sweep. When mutation target size is small, there may be long waiting time.

- Model of polygenic adaptation (Figure 3): short-term adaptation, selection on standing variations at multiple loci, so the AFs of these variants will increase, but not get fixed. This may also explain the pattern of AF difference (genic $>$ non-genic) due to linked selection. Once the population reaches new optimum, weak frequency-dependent selection: downward drift of some alleles will have to be balanced by upward drift of other alleles.

- Analysis: why standing variants may be more important for adaptation? Suppose we have a reasonably large mutation target size, then when selection force arrives, some standing variants will respond. As these variants increase in frequency, selection will be reduced.

- How to find real selection signals? "place more weight on sweep signals thatinclude variation at likely functional sites. Improved external information will likely help greatly in the coming years; test whether, looking across many loci, there is a significant tendency for alleles that increase the phenotype value to increase (or decrease) in frequency together.

Recent Progress in Polymorphism-Based Population Genetic Inference [J Hered, 2012]

- Methods for testing the departure of neutrality: using the site frequency spectrum. Tajima's D, Fu-Li test.

- Methods for testing selective sweep:
  - e.g. Fay-Wu derives the expected frequency spectrum after a selective sweep.
  - Important to account for population demographic history.
  - Using haplotype (LD) information.

- Estimating the extent of purifying selection:
  - Estimating selection on nonsynonymous mutations [Loewe et al, 2006]: using data of two species. idea: variants subject to sufficiently strong purifying selection will not increase significantly as effective population size increases, whereas neutral diversity is expected to increase proportionally with population size (the comparison of syn. sites in two species).

– Simultaneous inference of selection and population growth from patterns of variation in the human genome [Williamson, PNAS, 2005]: Simultaneous inference of selection and population growth from patterns of variation in the human genome. Account for population growth.

– Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change [Eyre-Walker A, Keightley PD. MBE, 2009]: inference of selection, population history and account for beneficial mutations.

2.7.1 Infinite-Allele Model

Infinite-alleles model: often used to analyze the allozyme data.

- Model: each new mutation creates a new allele. The new alleles are constantly created from mutation, and also removed from the population by random drift (and selection, if it is modeled).

- Homozygosity: we are interested in the genetic variation of the population, which is commonly defined as the homozygosity, (probability that two random alleles are the same). Under infinite-alleles model, homozygosity is equal to probability of IBD. Let F_t be the probability of IBD at generation t, could derive the recurrent of F_t, and at equlibrium:

$$\hat{F} = \frac{1}{\theta + 1} \tag{2.89}$$

  where $\theta = 4N\mu$ (μ is the mutation rate of the protein). The intuition is that: when μ is low, most of the alleles are the ancestral form, thus the IBD is high.

- Allele-frequency spectrum: the polymorphism data of n sequences can be characterized by the allele-frequency spectrum: the frequency of each unique allele. Let k be the number of different alleles in a sample of n, then:

$$E(k) = 1 + \frac{\theta}{\theta + 1} + \frac{\theta}{\theta + 2} + \cdots + \frac{\theta}{\theta + n - 1} \tag{2.90}$$

  The probability of a particular allelic configuration is given by the Ewens sampling formula: the probability of a sample of size n containing k distinct alleles with n_i copies of type $i, 1 \leq i \leq k$ is:

$$P(n_1, n_2, \cdots, n_k, k) = \frac{n!\theta^k}{k! n_1 n_2 \cdots n_k S_n(\theta)} \tag{2.91}$$

  where $S_n(\theta) = \theta(\theta + 1) \cdots (\theta + n - 1)$.

Testing neutrality with infinite-alleles model: Ewens-Watterson test

- Allele frequence spectrum: e.g. excess of common or rare alleles. This could be compared with the expectation under Ewens-sampling formula.

- Homozygosity: the test statistic is F, the homozygosity. The statistical significance can be assess by simulation (sampling from a population according to the neutral infinite-alleles model). If F is higher than expected by chance (more homozygosity), this may suggest: purifying selection that removes deleterious mutations; or growing populations.

- Limitations: the test assumes that any two alleles that cannot be distinguished (e.g. by electrophoresis) must be IBD. The change of population size may also cause departure of the infinite-alleles model.

2.7.2 Infinite-Site Model

Infinite-sites model:

- Model: the sequence is infinitely long, and each new mutation occurs in a different site of the sequence. The pattern of interest is: (1) the number of segregating sites; (2) the nucleotide mismatches of any two sequences in a sample. Both (1) and (2) are defined at per site level.
- Number of segregating sites (S): this is the same as the number of mutations in the entire genealogy tree of n sequences. The expected number of mutations is:

$$E(S) = \mu E[\sum_{i=2}^{n} iT_i] = \theta \sum_{i=1}^{n-1} \frac{1}{i} \tag{2.92}$$

  Note that the mutation rate μ above is the mutation rate of the entire nucleotide sequence. The variance of S can be computed using the law of total variance (the distribution of the number of mutations conditional on the coalescence time). Let X_i be the number of mutations in time T_i, then (condition on T_i, the number of mutations in i branches are independent):

$$E(X_i|T_i) = i\mu T_i \qquad \text{Var}(X_i|T_i) = i\mu T_i \tag{2.93}$$

  Thus, plug in the mean and varirance of T_i, we have:

$$\text{Var}[X_i] = E[\text{Var}(X_i|T_i)] + \text{Var}[E(X_i|T_i)] = \frac{\theta}{i-1} + \frac{\theta^2}{(i-1)^2} \tag{2.94}$$

  The variance of S is given by:

$$\text{Var}(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \tag{2.95}$$

- Number of mismatches between any two sequences (Π): this is the number of mutations in the tree relating the two sequences:

$$E(\Pi) = 2\mu E(T_2) = \theta \tag{2.96}$$

  The variance of Π can be found in Equation (4.19) or [Tajima83].
- Nucleotide polymorphism and diversity: both S and Π depend on the sequence length. We could define nucleotide polymorphism as the average S over the sequence length L: $S^* = S/L$; similarly we define nucleotide diversity as: $\pi = \Pi/L$.

Testing selection using Site Frequency Spectrum (SFS): Figure 9.2 of [Nielsen & Slotkin]. Different scenarios and how the SFS are changed:

- Negative selection: the SFS will be skewed towards low-frqeuency alleles comparing with neutral sites. Note: the patterns are different for strong and moderate/weak selection. Strong selection: reduce number of variants; moderate/weak selection: AF change (higher fraction of rare variants) without a large effect on the number of variants.
- Positive selection: the opposite effect, SFS skewed towards high-frequency alleles.
- Selective sweep: alleles linked to the advantagous site will increase and reach higher ferquency, while the unlinked alleles will have lower frequency. The result: an excess of both high- and low-frequency alleles.
- Balancing selection: opposite to selective sweep, more alleles of intermediate frequency.

Tajima's D statistic: designed for testing selective sweep

- Intuition: both S and Π can be used to estimate θ. Thus under the neutral model: $\Pi - S/a = 0$ (where a is the harmonic series). With selection, the two would not be equal: S is not very sensitive to frequency of polymorphic sites (and thus less sensitive to selection), but Π is sensitive - very rare or very common variants contribute less to Π than intermediate frequency variants.
- Test: $D = \frac{\Pi - S/a}{\sqrt{V(\Pi - S/a)}}$. Needs simulation to get the null distribution of D.
- Selective sweep: creates more rare and very common alleles, and this leads to lower Π, thus $D < 0$. Other population process such as population growth, which leads to an excess of rare alleles.
- Balancing selection: leads to more nucleotide diversity (pairwse difference) relative to S than expected, thus $D > 0$. Another interpretation: population contraction.

Fu-Li test:

- Intuition: the numbers of singleton and non-singleton sites (η_i). With purifying selection, expect most polymorphic sites to be singleton, as it would be hard for the polymorphic sites to acqure more than one copy in the population due to selection.
- Test: the number of singleton sites is equal to the number of mutations in the external branch (η_e), and non-singleton sites equal to that in the internal branch (η_i), where the external branch is defined as the branch from an internal node to the tip. Could show that:

$$E(\eta_e) = \theta \qquad E(\eta_i) = (a-1)\theta \tag{2.97}$$

  Note that the results are independent of the sample size. The test statistic is $\eta_e - \eta_i/(a-1)$ normalized by its varirance. The P value is often determined by simulation.

2.7.3 Poisson Random Field Model

Reference: [Sethupathy & Hannenhalli, PRF tutorial, 2007], Directional Selection and the Site-Frequency Spectrum [Bustamante & Hartl, Genetics, 2001]

Goal: given a sequence of interest (a protein, some genomic region, etc.) and its polymorphism spectrum data (the number of fraction of polymorphic sites with different daf), how to detect natural selection?

- PRF model: assumption of unlinked loci. The expected frequency spectrum can be calculated directly using mathematical models. The model can be used for estimating selection coefficients for particular classes of mutations and test various hypotheses regarding selection.
- Williamson et al [PNAS, 2005] applies PRF for human data, taking population history into account.

Model: let X_i be the number of polymorphic sites with daf i/n, i.e. i copies of derived alleles within n samples, the problem is to determine the distribution of X_i.

- Survival time (transient distribution): from the diffusion model, we know that for a new mutation, its time of staying in frequency $(x, x+dx)$ is $f(x)dx$, where $f(x)$ is defined by Equation 2.46, which we repeat here:

$$f(x) \approx \frac{1 - e^{-2\gamma(1-x)}}{(1 - e^{-2\gamma})} \frac{2}{x(1-x)} \tag{2.98}$$

  where $\gamma = 2N_e s$. When $s = 0$, we have $f(x) \approx 2/x$.

- Relating number of polymorphic sites with survival time: we are interested in the number of polymorphic sites whose daf are in $(x, x+dx)$, denoted as $g(x)dx$. Intuitively, if any mutation has a long survival time in $(x, x+dx)$, we would expect more sites in this range; and if we have high mutation rates θ, we'd also expect more sites. We can show that $g(x) = 2N_e\mu f(x)$. Suppose mutations are non-overlapping (e.g. we look a small segment), then we can analyze one mutation a time. After each new mutation arises, it will stay at frequency x for $f(x)dx$ generations. But we do not have one mutation every generation, so we should multiply $f(x)dx$ by the probability we have a mutation in one generation $2N_e\mu$. The same results would also hold for longer segments. So we have:

$$g(x) = 2N_e\mu f(x) \tag{2.99}$$

- Finite sampling: We need to further consider the sampling process in analyzing the actual polymorphism data. Note that $g(x)dx$ is the expected number of polymorphic sites with AF in $(x, x+dx)$, among these sites, the percent that generates i copies in a sample of n sequences is $\text{Binom}(i; n, x)$. Thus the expected number of sites with i copies of derived allele is:

$$F(i) = \int_0^1 g(x)\text{Binom}(i; n, x)dx = 2N_e\mu \int_0^1 f(x)\binom{n}{i}x^i(1-x)^{n-i}dx \tag{2.100}$$

  The distribution of X_i is Poisson distribution with mean $F(i)$ because all sites are iid. Note that in the above equation, we have binomial sample of n, instead of $2n$ - one can show that this does not change the results.

- Likelihood: the probability of observing $X = (X_1, \cdots, X_{n-1})$ is given as:

$$L(\theta, \gamma) = P(X|\theta, \gamma) = \prod_{i=1}^{n-1} P(X_i = x_i|\theta, \gamma) \tag{2.101}$$

  This leads to a way of parameter estimation and LRT for neutrality.

- Extension to site-specific selection [Bustamante & Hartl, Genetics, 2001]
  - Mixture of selected and neutral sites: change $f(x)$ to $(1-\pi)f_0(x) + \pi f_\gamma(x)$, where π is the fraction of selected site, $f_0(x)$ and $f_\gamma(x)$ are the transient distribution of neutral and selected sites respectively.
  - Individual site: inferrence in a mixture model.

Understanding SFS from PRF model:

- Neutral SFS under PRF vs. coalescence theory: e.g. we consider the expected number of singletons $F(1)$. Under neutral theory, this is $\theta = 4N_e\mu$. We show that this is indeed the case with PRF:

$$F(1) = \int_0^1 2N_e\mu\frac{2}{x}\binom{n}{1}x^1(1-x)^{n-1}dx = 4N_e\mu n\int_0^1 (1-x)^{n-1}dx = 4N_e\mu \tag{2.102}$$

- Comparing SFS under neutral vs. selection.

Application to McDonald-Kreitman (MK) test:

- MK test
  Intuition: if there is negative selection, then there will be less polymorphism than divergence because more mutations are eliminated by natural selection than in the case of neutral; and similarly, there will be more polymorphism if there is positive selection. The polymorphism and divergence measures are both scaled by those of the neutral sites (synonymous sites).
  Procedure: construct 2-by-2 table with number of polymorphic sites and number of fixed substitutions for both synonymous and replacement sites, and test if the ratio of polymorphism between replacement and synonymous is significantly larger than that of divergence.

- application of PRF to MK test
  - number of polymorphic sites: apply the expected density function of daf and consider the sampling process (because some true polymorphic sites will not be sampled as polymorphic if only one allele is sampled). The expected number of polymophic sites with sample size m is:

$$H(m) = \int_0^1 g(x)P_m(x)dx = \int_0^1 g(x)[1 - x^m - (1-x)^m]dx \tag{2.103}$$

    Apply this equation to both the synonymous ($s = 0$) and replacement sites.
  - number of fixed substitutions: it has two parts, one from fixed substitutions (mutation multiplied by fixation probability) and the other from sampling:

$$2N_e\mu t_{div}P_{fixation} + \int_0^1 g(x)x^m dx \tag{2.104}$$

    where $P_{fixation}$ is given by Equation 2.41.

2.7.4 Inferring Negative Selection

Approximation to the Distribution of Fitness Effects across Functional Categories in Human Segregating Polymorphisms [Racimo and Schariber, PLG, 2014]

- Challenges of estimating distribution of fitness effects (DFE): confounded by demography; most methods depend on binary classification of neutral or selected genes.
- Estimating selection coefficients from SFS: use non-equilibrium demography to derive the expected SFS, let $f(x,t)$ be the frequency spectrum at frequency x and time t, and let $g(x,t) = x(1-x)f(x,t)$. Solve the PDE of $g(x,t)$ from diffusion equation. Then the probability that a given site in a sample of n has i copies of the derived allele is:

$$p_i(t) = \frac{f_i(t)}{\sum_{j=1}^{n-1} f_j(t)} \tag{2.105}$$

  where $f_i(t)$ is the frequency of variants with i copies of allelies at time t. Comparing with PRF, this does not explicitly model absorption (extinction or fixation). To fit selection coefficients to observed SFS: create bins by C-scores assuming one DFE value for the entire bin, then use MLE to estimate DFE for that bin.
- DFE for C-score bins (Figure 1A): substantial selection at $C > 30 - 35$ (C-scores are PHRED-scale), $s < 10^{-4}$ ($N(0) = 10000$ reasonable population size for human).
- Background selection: divide sequences into 10 B-score bins. The DFE-C score relationship is robust to BGS, except the two highest bins (strongest BGS). Solution: fit a neutral demography model only in the exome, and then do C-score to DFE mapping in these regions.
- DFE in different classes of elements:
  - Nonsynonymous variants: bimodal distribution. Note: PhastCons scores do not perform well as the scores are regional, thus cannot distinguish NS and S variants.
  - Synonymous: not neutral, however, could be due to background selection.
  - Non-coding elements: more uniform.

2.7.5 Inferring Positive Selection

Population differentiation as a test for selective sweeps (XP-EHH) [Chen and Reich, GR, 2010]

- Background: selective tests, within population, use AFS or haplotype (EHH test). Intuition of EHH: Figure 1A, selection leads to increase of DAF of sites linked to selected sites. Comparing this with neutral expectation leads to selection inference. Importantly, the change of DAF under neutrality depends on the region (LD) size: small region, old allele, so expect larger changes on average; and large regions, young alleles, so smaller changes.
- Background: Cross-population test of selection: (1) AF differentiation, or F-ST. However, F-ST has large variations under neutrality. (2) Use haplotype comparison: increased hom. at linked sites.
- Model idea: Figure 1B, similar to EHH test, expect that it considers difference of AF b/t populations.
- Model: of a single site, let p_1 be the AF of the objective population, and p_2 be the AF of the reference population. Our goal is to model p_1 distribution as a function of p_2, recombination rate r, selection coefficient s, and divergence time (time since selection event) ω. The change of AF follows Brownian motion model, the frequency is expected to be increased to $1 - c + cp_1^*$, where p_1^* is the frequency in the objective population before selection, and $c \approx 1 - q_0 r/s$, where $q_0$is the initial allele frequency of A in population 1. The likelihood $f(p_1|r, s, p_2, \omega)$ is given by Equation (4), see Figure 2 for different distributions of p_1 under neutrality vs. selection (closer to 0 and 1).
- Composite Likelihood ratio (CLR) test: test if $s = 0$ across all sites. They are not independent, so weigh the likelihood by their LD.
- Analysis: how does the model encode the idea of calibrating Δ AF with LD/allele age? The parameters are fit using all sites in a window. Intuitively, the method uses low LD sites to learn about neutral pattern, including ω, and use high LD sites to learn selection (expected to be deviated from neutrality).

Signature of multiple-merger coalescence in genomic diversity data [Daniel Rice, Sep, 2016]

- Background: nucleotide diversity π depends on $T_c\mu$ where T_c is the time since common ancestor and μ the mutation rate. Also π is spatially correlated because T_c of adjacent pairs are similar. The spatial correlation depends on recombination rate $T_c r$: higher r, lower spatial correlation.
- Background: genealogy tree, balanced tree topology. With population bottleneck or positive selection, one has "mergers", or star-like topology. This picture changes with recombination: one has merger in the tree, but then the subtree coalescence with other neighbors.
- Recurrent selective sweep: leads to multiple mergers in a tree. Signature: selective sweeps lead to increase of DAF, thus the variants with high DAF will be higher than expected.
- Motivation: detection of multiple mergers without using recombination map, ancestral alleles. Let ϕ_i be the number of sites with AF i. The idea is that, with merging, we create multiple high AF variants (consider multiple mutations before the merging: they all have high AFs). So we can use average correlation between ϕ_i's as our statistic: it will be generally negative under neutral model, but could be > 0 under multiple merge model.

2.8 Population Genetics of Complex Traits

A Population Genetic Signal of Polygenic Adaptation [PLG, 2014]

- Model idea: suppose we have multiple populations. Let Y_i be the average trait value in population i. We want to infer whether the trait is under selection. Define the neutral model of the trait, and test deviation. The idea is that we can obtain the neutral model of the AF, and the trait is related to the AF, so we can obtain the neutral model of the trait.

- Model: let p be the AF of an extant population and p_A be that of ancestral population. Then $p|p_A$ follows normal distribution, with variance dependent on the time. For the trait in population i, we have $Y_i = \sum_l \beta_l p_{il}$ where β_l is the effect size of SNP l, and p_{il} the frequency of SNP l in population i. One can show that Y_i follows MVN.

The osteoarthritis and height GDF5 locus yields its secrets [NG, 2017]

- Fine-mapping GDF5 locus: transgenic mice, define an enhancer of 2.5kb, driving expression in long bones.
- Signature of positive selection: high LD near the SNP, suggesting selective sweep. Use phylogenetic analysis to infer the history: compare African, European and Neanthandal. The allele arrives in Africa, then spread to both Neanth. and European during the old and later out-of-Africa migration.
- Remark: does this result mean that height is under positive selection? If so, wed expect many other loci showing evidence of selection. If not, whats the selective force driving GDF5 locus?

From Adaptation to Disease [Jeremy Berg, 2018]

- Part I. Detecting polygenic adaptation. Ref: Berg and Coop, 2014: polygenic score under selection, much stronger signature than individual SNPs.
- Detecting polygenic selection across populations: Let Z be the polygnic score, Z is weighted average of effect size (by AF). Define $V(Z)$ as variance of Z across populations, then $E(V(Z)) = VF_ST$. To see this, we can write $Z = G\beta$, where β is effect size, then:

$$\text{Var}(Z) = \text{Var}(G^T\beta) = \beta^T \text{Var}(G^T)\beta = \text{tr}(\text{Var}(G^T)\beta\beta^T) \tag{2.106}$$

  where $\text{Var}(G^T)$ is related to F_{ST}. Intuitively, large selection leads to large AF difference, thus large variance of Z. Its expected value under neutrality depends on gene flow, etc.
- New model: Racimo, Berg and Pickrell 2018. Selection happens at particular lineages. Analysis of phylogenetic tree of populations and how polygenic scores change. Ancient DNA data to help with understanding polygenic score changes.
- Q: Can we develop a model of H_1 and use it to identify specific populations where the trait under selection? The model is far more complex, e.g. need to consider population size, when selection happens, etc.
- Part II. Why disease alleles persist? Does genetic architecture depend on fitness cost?
- Model: $s = \alpha \cdot t(P) \cdot S$, alpha: allele effect, and t(P) density at threshold of liability (population distribution of liability), S trait fitness. s: selection coefficient. Intuition: when S increases, the distribution of liability shifts (so that prevalance is reduced), as a result, for any particular allele, its less likely to reach the liability, so $t(P)$ is weaker and overall, s is constant.
- Show with simulations: when S changes, s returns previous level after some generations.
- Remark: intuitively, with stronger fitness cost, deleterious variants will be eliminated more often. So any deleterious allele will be present in an individual with fewer other deleterious variants - thus this allele is further from liability threshold, weakening selection.
- Chalk Talk. Goal: given AF vs. effect size distribution, learn about selection on traits.
- Remark: PRF kind of model, the site frequency spectrum for variants under a given selection.
- For directional selection: Power is proportion to $\alpha^2 \cdot p(1-p)$. Strong selection p $1/(\alpha\beta)$, where β is selection.

- For stabilizing selection: p $1/\alpha^2$, so power is the same for different AFs. Remark: power is proportion to variance per SNP.
- Q: But in practice, power is higher for common variants, why?
- Directional selection: $\mathrm{E}(\Delta p) = \alpha\beta pq$, change of AF per generation. Disease: liability threshold, fitness depends on risk, which depends on liability.
- Stabilizing selection: $\mathrm{E}(\Delta p) = \alpha^2\omega^2 pq(.5 - p)$, where ω is strength of stablizing selection.
- Modeling Pleiotropy: treat pleiotropic effect as some additional term influencing selection. Long term, joint modeling of multiple traits.
- Discussion [personal notes]: Interaction can be very common under the liability threshold model: a variant near the threshold may have a large effect than a variant far away from the threshold. Interaction between variant and PRS: collider bias?
- Model: how natural selection shapes AF vs. effect size of a trait? If the trait is correlated with another continuous trait, then it is possible that selection for the second trait would drive the first trait to higher prevalence even if the first trait is deleterious. Ex. ASD and education.
- **Lesson**: two challenges of studying evolution of complex traits: (1) Epistasis: the selection on one locus, depends on genetic background (how many deleterious alleles are present). This could be thought of as Collider bias: both alleles act on a common trait. (2) Pleiotropy: selection on one allele depends on its effect on all phenotypes it affects.

Evolution of polygenic risk score using ancient DNAs [Maryn, NHS meeting, 2020]

- Problem: during evolution, the risk alleles may change from random drift, e.g. some risk alleles may get lost, and new alleles may emerge (from low frequency to higher frequency with measureable effects). This turnover will change the population level PRS distribution.
- Model: suppose we have L risk alleles, with effect β_l, and $X_l(\tau)$ be the genotype of a sample back in time τ. The genotype-phenotype model is written by:

$$Y(\tau) = C + \sum_l \beta_l X_l(\tau) \tag{2.107}$$

  The observed PRS can be written as:

$$\hat{Y}(\tau) = \hat{C} + \sum_s \beta_s \hat{X}_s(\tau) \tag{2.108}$$

  where we sum over all observed variants. We can thus consider the change of the mean of PRS and variance of PRS. The mean change is the bias of PRS due to the change of the intercept. To model how the PRS distribution changes, we denote $Z(\tau)$ the vector of AF of all variants at time τ and $Z(0)$ the AFs in the current time.
- Our primary analysis involves the computation of AF change conditioned on $Z(0)$. In particular, consider the cases where $Z(0)$ is rare (below GWAS AF cutoff) and $Z(\tau)$ is common and the opposite situation. Such cases lead to bias of PRS. Let $p(z, z')$ be the transition probabilities of population AF, we have:

$$\mathrm{E}(X_l(\tau)|Z_l(0)) = \int z(2z-1)p(z, z'; \tau)\frac{\rho(z)}{\rho(z')}dz \tag{2.109}$$

  where $\rho(\cdot)$ is the equilibrium frequency of z.
- Remark: the model considers one sample a time, and treat ancient AF as unknown and marginalized. When we have multiple samples in ancient, we have some information of ancient AF.

- Remark: PRS model also depends on LD. How LD structure evolves over time?

Mutation-selection balance of polygenic traits [Jeremy Berg, NHS, 2020]

- Motivation: in quantitative genetics, it is often assumed that the fitness effects of mutations are "symmetric": increase or decrease of a trait has the same fitness effects. But for polygenic disease trait, this is not true, as risk-increasing mutations are subject to stronger selections.
- Model idea: mutations mostly increase the liability and reduces the fitness. This forces is balanced by the smaller portion of protective mutations and negative selection. At equilibrium, the fitness distribution of the population does not change. This is similar to mutation-selection balance, but across many loci.
- Model assumptions: consider a single site, it can exist in two forms, protective or risk alleles. We consider the two alleles by a two-state Markov chain, with mutation rate μ. The effects of new mutations follow certain distribution, denoted ad $p(\alpha)$, for effect size α. At equilibrium, we expect many more sites exist in the protective allele form, and this proportion generally depends on α. We denote ρ_α the proportion of sites fixed for the risk (or protective?) allele.
- Model: how liability distribution changes in each generation? We would expect at equilibrium, the increase of liability by mutation is equal to the reduce of liability by selection:

$$\Delta_u \bar{z} = -\Delta_s \bar{z} \tag{2.110}$$

  where z denotes liability. Suppose there are L sites in total, We have $\Delta_u \bar{z} = 2L\mu b$, where b is the average effect on liability, over all sites (by their mutation effects). We write selection effect:

$$\Delta_s \bar{z} = \int p(\alpha)\alpha \langle \Delta x|\alpha\rangle d\alpha \tag{2.111}$$

  where x is the frequency of allele. The change of x for sites with effect α is given by:

$$\langle \Delta x|\alpha\rangle = \int \langle \Delta x|x,\alpha\rangle \Psi(x|\alpha)dx \tag{2.112}$$

  where $\Psi(x|\alpha)$ is the AF distribution. We have $\Delta x|x,\alpha\rangle = \pi_\alpha S x(1-x)$, where π_α is the risk-scale effect size, and $S_\alpha = \pi_\alpha S$ is the selection coefficient. And for $\Psi(x|\alpha)$, we have:

$$\Psi(x|\alpha) = (1-\rho_\alpha) f(x|-\gamma_\alpha,\theta) + \rho_\alpha f(1-x|\gamma_\alpha,\theta) \tag{2.113}$$

  where $\theta = 4N_e\mu$ and $f(\cdot)$ is given by the diffusion equation.
- What have we learned from the model? Assuming effects are small, the results are insensitive to fitness cost of disease S.
- Question: how is the model related to empirically observed effect size and AF relationship?

2.9 Pattern and Rates of Germline Mutations

Influences of mutation rates and patterns [personal notes]

- Local mutation rates are influenced by both mutational supply (e.g. exposure to carcinogens) and DNA repair. There are multiple mechanisms of mutations and repairs, each with possibly distinct signatures/mutational spectrum. Ex. some mutational mechanisms are dependent on replications, others not.

- Replication timing: strong influence on cancer/somatic mutation rates, moderate influence on germline rates. Mechanism: nucleotide store depleted during late replication, reducing the effectiveness of repair.
- Distance to telomere.
- Recombination rates.
- Low complexity regions/repeats.
- Transcription coupled repair. Signature: strand asymmetry.
- Chromatin structure: strong association of DNM with DHS in [Autism WGS, Cell, 2012]. Also found association with DHS in [Francioli, 2015], but not significant after correction of GC.
- Regulatory activities: association with K27ac in [Autism WGS, Cell, 2012]. However, this could be due to confouding factor: GC content or DHS.
- DNA methylation and CpG: cytosine deamination.
- Local sequence context: tri-nucleotide context. GC content: DNA repair needs to separate two DNA strands, it is easier for AT (2 H-bonds) than GC. So generally higher GC content associated with higher mutation rates.
- Hotspots: found in multiple studies. Also [Francioli, 2015] suggests that their mutational spectrum is different from non-clustered mutations.
- Individual-level variation of rates: paternal age. Also found the difference in mutational spectrum (and influence of other factors) in older fathers vs. younger ones [Francioli, 2015].
- Remark: to model/understand the pattern of mutations, consider all mechanisms, and the effect of each feature on each mechanism. Sometimes a feature may impact multiple mechanisms, creating complex relationship with observed mutation rates, e.g. chromatin structure can affect mutational supply and repair. Another example: transribed regions.

Properties and rates of germline mutations in humans [Campbell & Eichler, TiG, 2013], Determinants of Mutation Rate Variation in the Human Germline [Sgurel & Przeworski, ARG, 2014]

- Four strategies of estimating mutation rates:
  - Severe Mendelian disease: mutation/selection balance to estimate how frequent new mutations occur.
  - Comparative genomics: need estimation of generation time (large uncertainty). Also influenced by GC-biased gene conversation, which acts analogously to selection. Also highly sensitive to the assumptions of ancestral polymorphism.
  - Genetic variation within human population: may be affected by other processes such gene conversion.
  - Pedigree sequencing.
- SNV mutation rates
  - DNMs from families: estimate of 1.16E-8 per base pair per generation.
  - Human-chimp comparison: earlier estimate of 2.5E-8 in pseudogenes, and 1.82E-8 using inferred ancestry of nearby microsatellites.
  - Discrepancy may be due to (1) differences in filtering applied for SNVs or in sequencing methodology. (2) Change of life history in primates: number of spermatogenetic cell divisions per cycle, change of generation time, delayed onset of puberty and so on.

- Sources of mutations: replication or spontaneous (endogenous or exogenous).
  - Replication errors: incorrect base incorporation or slippage of polymerase (indels < 20 bp). The repair process is facilitated in regions where DNA strans can be easily separated (AT-rich).
  - Non-replicative errors: deamination of C is an important source.
- Nonrandom distribution of new mutations:
  - Shorter scale: Transitions outnumber transversions by twofold for de novo SNV. The rate of mutation at CpG dinucleotides has been observed to be ten- to 18-fold the rate of non-CpG dinucleotides, probably due to CpG methylation.
  - Intermediate scale: Low-complexity repetitive DNA may also be mutagenic, consistent with sequence-dependent replication slippage.
  - Broader scale: (1) The higher mutation rates in or near protein-coding regions: perhaps from the higher GC content of these regions in combination with the effects of transcription-associated mutations (TAM). Also strand asymmetry in transcribed DNA. (2) Recombination: may introduce point mutations.
  - Other factors such as nucleosome occupancy and DNase hypersensitivity are correlated with broad-scale mutational patterns, but the association could be due, at least in part, to confounders such as base composition.
- Nonindependence of mutations:
  - Over 10kb scale: mutations in a single transmission more clustered, from IBD studies.
  - In the range of 100bp or less: excess of mutations in a single replication. Mechanism: DNA lesion (e.g. oxidized base), requres special polymeraise such as Pol zeta, which is error prone.
- Inter-individual variation of mutation rates:
  - Replication driven mutations: likely dominate most of germline mutations. In males, after puberty, about 23 replications / year in sperms. The ratio of male to female mutations α is about 4. However, the ratio c based on number of cell divisions should be 9-14 depending on the time of conception.
  - Spontaneous mutations: proportional to time.
  - Possible explanations of $\alpha \neq c$: (1) Many mutations are spontaneous in both sexes. (2) Higher error rate per cell division before puberty.
  - Individuals may vary in their propensity to acqute mutations: (1) notably due to mutations in DNA mismatch repair genes. (2) Germline methylation.

The effects of chromatin organization on variation in mutation rates in the genome [Makova & Hardison, NRG, 2015]

- Why it is important to estimate local mutation rates? Inferring selection and interpreting somatic mutations.
- Regional variation in mutation rates (RViMR): demonstrated in multiple types of mutations using cross-species comparison, including base substitutions, small indels, TE insertions. Co-variation among different types was also found.
- Genomic features that contribute to RViMR: *GC content* (methylated CpG). Hight AT content - high sbustitution rates (more likely in *heterochromatin*). Increase in rates close to *telomeres* due to altered repairs in these regions. These genomic features (add male *recombination rates*, exon density, etc.) explain $> 50\%$ variability in substitution rates and 30% in indel rates.

- Chromatin structure/accessibility: depends on whether it has a greater impact on mutagenic or repair process. In general, increased access leads to increased exposure to DNA damaging agents, but also greater access of DNA repair enzymes. Overall, open chromatin has a lower mutational frequency, either due to fewer mutations or increased repair.
- Results of chromatin accessibilty on mutation rates:
  - Pairwise analysis: nucleosome occupancy may affect substitution rates. Some possible mechanisms: DNA in nucleosomes are less prone to cytosine deamination. But the substitution pattern may also be influenced by selection that maintain optimal GC in core and linker regions. In general, some studies support a link between open chromatin and repressed mutations (due to repair), while other studies found the opposite pattern.
  - Multivariate analysis: use base substitutions, insertions, deletions and microsatellite repeat number alteration. Classify the four types of mutations into different types of regions (6) with HMM model. Ex. the "hot" state where the rates of all mutations are elevated - characterized by open chromatin, and "warm" states where deletion and substitution rates were mildly elevated - characterized by closed chromatin.
- Transcription coupled repair (TCR): lead to mutation bias (lower in transcribed strand) and overall lower rates.
- Discussion: future work
  - Experimental manipulation of enzymes to study the mechanims of mutation.
  - The influence of biased gene conversion on the relationship between mutation and chromatin.
- Remark [personal notes]: the relationship between mutation rates and chromatin accessibility seems complex, nonlinear. This could be due to (1) accessibility has different effects on exposure to mutational agents and on repair (e.g. repair is saturated at some point, while exposure is monotonic); (2) additional factors not accounted for: heterogenietity of open chromatin regions. It might be interesting to apply **causal inference** techniques to distinguish direct and indirect effects of various genomic features on mutaiton rates.

Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation [Michaelson & Sebat, Cell, 2012]

- Background: for structural variation (SV), the mutation rates are predominantly driven by meiotic recombination - nonallelic homologous recombination (NAHR) between tandem segmental duplications lead to SV hotspots. Human and chimpanzee comparison have found evidence that regional SNV mutation rate is influenced by GC content, recombination rate and chromosome-banding patterns. Data: WGS of 10 MZ twin-family samples, 40x. MZ improves the power of detecting DNMs. A total of 581 germline DNMs were detected in ten MZ twin pairs. Our estimate is lower than theoretical estimates by a factor of two, but consisent with other estimates.
- Mutational clusters: defined as two or more DNMs within 100 kb. Found 10 clusters in all samples (one per family).
- Factors influencing the mutation rates: find features that discriminate DNMs and random genomic sequences. The features were obtained from ESCs (histone mark, DHS), Mammary Epithelial Cells (K27ac, nucleosome occupancy) and LCL (replication timing). The most significant features were DNase hypersensitivity, GC content, nucleosome occupancy, recombination rate, simple repeats, and the trinucleotide sequence surrounding the site. Train a sparse logistic regression model to predict mutation rate (MI).
  - Mutability was greatest for CpG dinucleotides.

  - Factors act on different scales: larger scales such as nucleosome occupancy (100 bp), recombination rate (10^4 bp), and replication timing (10^6 bp).

- Relation between mutability and evolutionary conservation: a distinctly U-shaped relationship between mutability and sequence conservation: hypermutability is correlated with highly conserved sequence and low genetic diversity.

Genetic Variation Meets Replication Origins [Cell, 2012] on [Koren, AJHG, 2012]

- Identify replication timing from WGS data of proliferating cells: early replication (S phase) tends to have higher sequencing depth than late replication ones.

- Replication time QTL (rtQTL): phenotypes can be: gain or loss of replication orgin, change of replication zone length, etc.

- Possible mechanisms of rtQTL: chromatin accessibility (affecting binding of replication proteins), enrichment of eQTL.

Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes [Chen and Xionglei He, Science, 2012]

- Nucleosomal DNA has fewer cytosine deamination, 50% decrease of the C > T mutation rate. The rates of G > T and A > T mutations were also about twofold suppressed by nucleosomes.

- Remark: what is the implication on mutation rates in open chromatin regions? No nucleosome, so expect higher mutation rates. But also more accessible by DNA repair enzymes.

A framework for the interpretation of de novo mutation in human disease [Samocha & Daly, NG, 2014]

- Model: trinucleotide context for each base. There are $64 \times 3 = 192$ rates for each mutation in every trinucleotide context.
  - Estimation of mutation rate per base: intergenic regions that are orthologous between humans and chimps. Tally the number of all 64 trinucleotides, and for each SNP, considered the chimp allele to be ancestral and determined the trinucleotide (XY1Z) to trinucleotide (XY2Z) change.
  - Calibration: use the model to determine the relative mutability, then set the absolute mutation rate by genome-wide mutation rate of 1.2E-8.
  - Per-gene probability of mutation: look up the table to obtain the rate for each base, then sum appropriate mutations.The probability of a frameshift mutation: multiplying the probability of a nonsense mutation by 1.25 (the relative rate from ASD WES data).

- Adjustment of the rates: use the number of synonymous singletons from ESP data.
  - Base model: gene length alone show $r = .835$, the base model, $r = 0.854$.
  - Depth adjustment for prob. of mutations in sequencing data (lower than theoreical rates because some bases are not covered well): multiply the estimated rates by a fraction 0.9 to 1, which is determined from the proportion of trios where every member has 10x coverage of a base. After this adjustment, $r = 0.891$.
  - Divergene adjustment: the idea is to adjust for regional variation. Define local divergence scores as the number of divergent sites between human and macaque over screened sites in a gene and up/down-stream 1M region. Then use linear models to determine the best equation to adjust the per-gene probabilities of mutation to incorporate the divergence score. Results: $r = .91$.
  - Replication time: no significant effect, so not used.

- DNM rate: in exome, 1.67E-8 per base per generation.

- Individual genes based on the estimated mutation rates: in 1,061 trios, (1) CHD8: 3 LoF and 1 missense, p.LoF = 1.76E-6, p.all = 3.2E-5. (2) TTN: 4 missense, p = 0.18.
- Validation using ESP data: correlate the number of rare syn. variants with predicted rates. Gene length alone showed high correlation (r = 0.880), but our full model showed significantly greater correlation, $r = 0.94$.
- De novo mutations and IQ: ASD cases with below average IQ have significant excess of de novo LoF mutations, while those with above average IQ have no such excess.
- **Using DNM rates to define constrained genes** (page 10 of Supplement): analogy of dN/dS test. For every gene, we let μ_S be the rate of syn. mutation and μ_{NS} be the rate of nonsyn. mutation, and similarly, N_S and N_{NS} be the observed number of variants. Our goal is to estimate the expected number of variants E_S and E_{NS}. Specifically:
  - Fit a linear model of N_S vs μ_S across all genes in the genome.
  - Apply the linear model to a specific gene to obtain E_{NS} using μ_{NS} from that gene.
  - Test deviation of N_{NS} from E_{NS} using chi-squared test. Then obtain the corresponding signed Z-score: positive score reflects negative constrainer (fewer nonsyn variants than expected).

  Note: the test uses missense variants. ExAC also has scores based on LoF variants, which is less reliable.
- Defining the excessively constrained genes: should have at least 5 syn. singletons (some genes have low coverage and not enough variants identified), the syn. mutation fits the mutation rates, and there is significant deficit of missense singletons ($p < 0.001$).
  - 5% of genes (1,003) are under strong selection (missense Z-score cutoff 3.09, $P < 0.001$), about half of which are in OMIM.
  - Constrained list: 2.3 fold enrichment of de novo ASD LoF, highly significant.
  - Missense Z scores: genomewide average around 0.94, genes with dn-LoF in ASD 1.68 and with dn-LoF in ID 2.46.
- Comparison with other metrics: simple dN/dS test using human data, only 377 highly onstrained genes ($P < 0.001$). The missense Z scores correlate with RVIS.
- Question: definition of χ^2 statistic, use only NS or both NS and S variants?
- Remark: the dN/dS test compares two counts (syn. and nonsyn.), while the current approach for testing selection estimates the mutation rates, instead of using counts.

Genome-wide patterns and properties of de novo mutations in humans [Francioli & Sunyaev, NG, 2015]

- Data: 250 Dutch parent-offspring families, 13-fold coverage. A total of 11,020 DNMs, with estimated sensitivity of 69% and specificity of 95%.
- Effect of paternal age: twich more DNMs in children of 40-year old father than 20-year old. Replication timing was significantly associated with paternal age, whereas chromatin states and recombination rates not. DNMs in younger fathers biased towards late replicating regions, also DNMs in older fathers are more likely to be in exonic regions.
- Regional variation: correlation with DHS and exonic regions (more enriched). However, no correlation after controlling for CpGs. Also no evidence of transcription-coupled repair (power is limited though).
- Clustering of mutations: beyond correlation with epigenomic variables: 78 clusters (up to 20kb) of 2-3 mutations. Both within and between individuals. Also a unique mutational spectrum in these clusters: enriched with C>G mutations.

- Within population variation: correlation with local recombination rates, after controlling for CpG sites and GC content.
- Correlation with human-chimp substitution rate (HCCG model): correlation with observed DNM rates $r = 0.18$. After adjusting for local recombination rates, $r = 0.37$. Also relative frequency of trinucleotide mutations are constant between DNMs and human-chimp divergence.
- Buidling mutation rate map: for any 1M region i and type t (8 types, including one CG dinucleotide mutation), let r_{ti} be the rate. The procedure:
  - Test correlation of human-chimp substitute rate with observed DNM rates for a type t, if significant, use the human-chimp rate as the baseline of r_{ti}; otherwise, use the global rate of that type of mutation.
  - Correct for local recombination rates ρ_i, linear regression of r_{ti} and ρ_i.
  - Add scaling factor f_t for each type of mutation, matching the observed frequency of type t.
- Remark:
  - The model is largely based on HCCG, however, even after justing for recombination rates, the correlation with DNM rate in human is only 0.37.

DNA methylation level of sperm is a major influence on the rate of de novo germline mutations [Li & Wu, 2015]

- Data: 4,470 exonic DNMs (4,143 DNSs and 327 DNIs) of 4,039 ASD trios and 2,061 DNMs (1,936 DNSs and 125 DNIs) of 2,299 control trios
- Distributions of DNMs:
  - C>T/G>A transition accounts for the majority (55%) of SNVs. Of these, 62% are located on CpG sites (CpGs), which is significantly higher than average level (12%).
  - Comparison of 6 subsitution types: only two types of bps (AT) or (CG), and each can mutate to 3 bases, so a total of 6. Next we define the context: CpG or non-CpG, so a total of 9 types of substitutions (only 3 for CpG context). DNMR of CpG C¿T/G¿A are 25-100 folds higher than other eight sub-types of DNSs (SNVs-8).
- Association of sperm DNA methylation levels with DNM rate:
  - For both C>T/G>A and SNVs8 (DNA methyl. levels are defined in the proximal CpGs), $R^2 > 0.9$ for both (previous studies found association at R^2 no more than 0.2). Method: define 5 levels of DNA methylation bins, then estimate DNMR for each bin.
  - DN indels: not significantly associated with DNA methylation
- Non-sperm results: the methylation level of non-sperm samples ($r < 0.75$) were less correlated with the reference sperm sample.
  - Table S6: CpG rate, correlation is close to 0.9 in non-sperm cells. But SNV-8: correlation is much lower (0.3 vs. 0.9).
- Gene-specific DNMR: Higher correlation with the number of syn. SNPs from ExAC, $r^2 = .92$. Comparison with other methods:
  - Samocha: 0.90. Full model: trinucleotide sites, sequencing depth, local divergence rates, and replication timing, 0.925.
  - Francioli: 0.90.

- DNA methylation in sperm explains most of mutability in DNMs using WGS data of 250 trios.
  - Studying the relationship between DNA methylation and DNMR using WGS data: We separately divided CpGs and non-CpGs into 20 groups, and counted the expected DNMR based on our DNA methylation model and the observed DNMR in each group.
  - Figure 4AB. Sperm DNA methylation level could explain 94.7% and 92.5% of mutability at CpGs and non-CpGs in whole genome.
  - Define clusters: If a 20-kb window contains > 2 mutations which showed significantly smaller distance than expected, we defined these mutations as clustered DNMs.
  - Figure 4C: clustered DNMs and non-clustered DNMs. It was showed that the expected site-level DNMR of non-clustered DNMs was significantly higher than that of genomic background, but lower than clustered DNMs.
- Candidate genes from TADA:
  - Use the mutation rates with TADA: 38 genes at FDR < 0.05 and 141 genes at FDR < 0.3, comparing with 2 and 17 in controls.
  - The ASD candidate genes were involved in synaptic function, chromatin remodeling, transcriptional regulation, and Wnt signaling.

Timing, rates and spectra of human germline mutation [Rahbari & Hurles, NG, 2016]

- Background: germline mosaicism. If mutations during development of germ cells, these mutations will appear in a fraction of germ cells (the earlier the mutations, the higher fraction). As a result, the DNMs in offsprings may share mutations. Also note that if the mutations occur before the separation of germ cells and soma, they will manifest as somatic mosaicism.
- Data: WGS of 3 families, about 20 individuals. 60-70 DNMs per genome, with rate 1.28×10^{-8}.
- Patenal age effect: for about half of DNMs, parental origin can be determined. Variation of paternal age effect: one family, 1.5 mutations per year, but other two, 3.2-3.6 per year.
- Mutation rate model during gametogenesis: deep sequencing 567x. About 1.3% DNMs of siblings are shared. This allows one to estimate that 3.8% of mutations in parental germ cells are shared. Infer the mutation rate change: Figure 6. Pre-PGC: about 0.2-0.6 mutation per replication; post-PGC: 0.5-0.7; after puberty: 0.1-0.2 (23 replications per year).
- Mutational spectra: No significant difference in the spectra of paternal and maternal mutations; and no difference in mutations from young and old fathers. Mutation signatures from cancer: signature 1 and 5 correlate well with the germline mutations.
- DNA methylation in testis cell line: 25% CpG are methylated (above 50% reads). 13 CpG sites overlap with DNMs, of which 12 were methylated.

An expanded sequence context model broadly explains variability in polymorphism levels across the human genome [Aggarwala & Voight, NG, 2016]

- Background: mutational spectrum. When we do not care about the absolute mutation rates, the pattern of mutations can be summarized in two ways:
  - Mutational spectrum: among all mutations, the frequency of X_1 to X_2 mutations, written as $f(X_1 \to X_2)$. So let the total number of mutations be n, the number of mutations of a type follows multinomial distribution.

– Conditional mutational spectrum: among positions with nucleotide X_1, how often it mutates to X_2, written as $f(X_1 \to X_2|X_1)$. The number of mutations of all sites with nucleotide X_1 is given by the multinomial distribution.

The two are related by:

$$f(X_1 \to X_2) = p(X_1) \cdot f(X_1 \to X_2|X_1) \tag{2.114}$$

where $p(X_1)$ is the frequency of base X_1.

- Model: use data from 1000 GP. For each position at the genome, say C, it has three polymorphism, $C \to X$. Let n_C be the number of sites with C in the reference genome, and n_{CX} be the number of polymorphic sites, then n_{CX} follows multinomial distribution. We model this distribution. Consider 6 mutational types, then 3-mer, 5-mer and 7-mer models. Show that the 7-mer model provides a significant better fit using LRT.

- Feature selection: for any of 6 mutational type $X_1 \to X_2$, let S be its sequence context (7-mer). The rate of $X_1 \to X_2$ mutations depend on S. We use a regression model: response variable is the mutational frequency of $X_1 \to X_2$ for base X_1 under a given S (i.e. conditional mutational spectrum defined above); and the explanatory variables are features of S. The intuition is that this mutational frequency depends on S. For each X_1, X_2, we then have 4^6 data points:

$$P(X_1 \to X_2|S) = \alpha + \beta_1 p_1^C + \beta_2 p_1^G + \cdots + \beta_n p_7^T + \epsilon \tag{2.115}$$

  For each S, we encode it as a set of features: first, single nt. feature, e.g. whether the first position is C or not. Next, incorporate interaction features. Use stepwise regression to selection features.

- Selected features from stepwise regression: first, show that the selected models explain 81% of variation of rates across contexts, comparing with 30% using only 3-mer model. Note: this ignores variation of rates within a certain 7-mer context. Some features include, polyA or T motif.

- Impact of DNA methylation: correlate the average mutation probability vs. average methylation level among all 7-mer contexts. Found a weak correlation $R^2 = 0.33$.

- Detecting selection of genes: using the estimated NS. mutation (probabilities), estimate the number of NS polymorphic sites per gene, and compare with the expected number (Z-score).

- Significant constraints in neuropsychiatric disease genes, OMIM genes, but less so in immune, OR genes. The scores work as well as Samocha in autism DNMs, and better than RVIS.

- Remark: the feature selection model does not use count data, rather, it uses linear regression of rates.

Discussion with Kelly Harris [2016]

- Polymerase zeta (low fidelity) may create clustered mutations (< 100bp): when high fidelity enzymes gets stuck, polymerase will fix the DNA damage, but often introduce wrong nucleotides (often A's). The most common signature: GC $\to$ AA. Overall, these clustered mutations explain about 2% of SNPs.

- Different mutation spectrum in different populations: eg. TCC 50% more common in Europeans than non-Euro. Possible explanation: match UV mutational signature.

Statistical methods for identifying sequence motifs affecting point mutations [Zhu & Huttley, review for Genetics, 2016]

- Motivation: suppose we want to study if neighborhood (context) affects mutation rates. Let's say $C \to T$ mutations. The possible confounder is the genomic regions. For example, suppose the mutations happen more often in exons (due to chromatin structure, etc.) than other regions, and exons are more rich in T's. Then we compare the context of C mutations vs. C non-mutations, we will find more T's near C mutations than non-mutations, even though the adjacant T's do not direct affect mutation rates.

- Idea: match each mutation with a non-mutation (of the same base) in 300 bp. Then compare the context.
- Model: let i be a base and j be status (mutaiton M or reference R), f_{ij} be the counts. Then model $\log f_{ij}$: the effect of base (some bases are more mutable), of status, adjacent bases and interactions. If there is an interaction between adjacent base and status, this suggests that context modifies the mutation rate.

Decoding germline de novo point mutations [NG, 2016; on Parent-of-origin-specific signatures of de novo mutations, NG, 2016]

- Data: 800 trios, 36K DNMs. Assign a substantial proportion of DNMs to their parental chromosomes of origin: 7k DNMs.
- 80% of mutations originating during spermatogenesis. Paternal effect: about 1 DNM per father age.
- Male and females: distinct mutational signatures. Evidence of non-replicative origin of DNMs in females.
- Some regions of the genome show an enrichment of maternally derived mutations. Perhaps some DNMs provide a survival advantage to mutant eggs in aging women?

Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans [Carlson & Zollner, biorxiv, 2017]

- Extremely rare variants (ERVs) from WGS of 3,000 subjects, about 36M ERVs. Advantages: earlier studies using variants of all AFs (ERVs only 25%), which are subject to other processes such as gBGC and selection. The ERVs are very young, estimated 1,200 years old.
- Defining relative mutation rate: consider a mutation of $X_1 \to X_2$, suppose it occurs in a context of S. The relative mutation rate is: $P(X_1 \to X_2|S)$, the probability that at S, how often it is mutated. Specifically, it is the number of times we observe ERV with $X_1 \to X_2$ divided by the number of times S (including X_1) occurs in the reference genome. Ex. 7,548 $C > T$ or $G > A$ autosomal singletons occurring in an ATACGCA or TGCGTAT 7-mer motif, and there are 53,314 such motifs in the autosomal reference genome, the relative mutation rate is 7,548/53,314 = 0.1416.
- Mutation rates in mutation subtypes (mutation type in a 1, 3, 5, 7-mer context): 3-mer context, 96 subyptes. 7-mer: 24,576 subyptes (256 times of 3-mer). 7-mer rates can vary 400 fold. Ex. NTT(A>T)AAA has 6-fold higher rate than average A > T.
- Genomic features: (1) Binary features: histone marks, DHS, CpG islands, exons, lamin-associated domains (2) Continuous features: recombination rate, replication timing and GC content, in nearby 10kb window centered at site.
- Model incorporating genomic features: train a genomewide logistic regression model. For a given subtype, each matched site in the genome is either a ERV $Z_j = 1$ or not $Z_j = 0$. Genomic features are explanatory variables and fit using `speedglm`. Estimating parameters for each subtype separately: for 84% of 24,576 subtypes, 10 times more ERVs than number of parameters, so OK.
- Influence of genomic features: (1) H3K9me3 (heterochromatin mark), recombination rate and late replication time, all associated with higher mutation rates across all subtypes. (2) H3K36me3, DHS, GC content and CpG islands: effects vary depending on mutation type and context. H3K36me3: regulate DNA MMR; CpG islands associated with lower DNA methylation. (3) For CpG > TpG subtypes: lamin-associated domain: higher rate (DNA hypermethylation); histone marks H3K4me1, me3 and K27ac: lower rate (DNA hypomethylation).

- Prediction of DNMs from trio-sequencing: classification of DNMs (46K) and non-DNMs (1M) with logistic regression, using predicted rate as covariate. Show that 7-mer model with genomic features (all included) performs the best (Figure 4).
- Prediction of genomewide relative mutation rates: use the logistic regression model (trained genomewide). All 14 genomic features are used - no feature selection.
- Possible mechanisms: (1) TFBS may increase mutation rate (limit the access of repair proteins): CEBP motifs within DHS show 2 fold lower mutation rates. (2) NTT(A>T)AAA: associated with L1 EN (retrotransposon)-induced damage.
- Remark: genomic features are derived from multiple somatic cells; better to use germline cells.

Reduced intrinsic DNA curvature leads to increased mutation rate [Duan and Qian, GB, 2018]

- Experiment: yeast mutant strains, only LoF mutations in a gene can survive. Sequence 1000 strains and map nonsense mutation freq.
- Calculation of DNA structure features: Figure S3. A table of 17 DNA structure features for each dinucleotide. Sliding window (10bp) to estimate the DNA structure features.
- Correlating DNA structure features with mutation rates: vary sliding window sizes. At 100 bp windows, DNA curvature shows strongest correlation.
- Yeast and TCGA data: similar patterns.
- Possible mechanism: mutation of mismatch repair genes does not alter correlation. Possibly due to different mutagen sensitivity.

Characterizing mutagenic effects of recombination through a sequence-level genetic map [Science, 2019]

- Background: most DSB ends up with lateral transfer of homologous segments (sometimes observable as gene conversions if there are heterzogyous markers). A small percent ends up as cross-overs.
- Data: 3000 WGS trios. 200K DNMs and 4M cross-overs.
- Human recombination rate map: 700bp resolution. Cross-over rates: vary with epigenomic factors.
- DNM rates: DNMs increases by 1.39 and 0.38 with every paternal and maternal age. Within 1kb of cross-overs: 50 times higher DNM rates. Females: higher DNM rates even within 40kb of cross-overs. Also change of DNM spectrum.
- GWAS of recombination rates: and other phenotypes, e.g. the average GC content within 500 bp of cross-over locations. Found 35 loci.
- Remark: mutation rates strongly correlate with cross-over events. However, cross-over events are random, though their distribution are highly non-random.

2.10 Evolution of Human Population

Reference: [Hartl, Principles of Population Genetics, Chapter 9, 10]

Overview: the patterns of genetic variations, including polymorphism, LD & recombination, divergence, in genome and across populations, reveal the underlying forces such as selection, mutational processes, human demography:

- Mutation rate: depends on sequence content, e.g. GC content (mutational bias).

- Recombination rate: depends on repeats (high repeat regions may have high recombination rates due to unequal cross-over and gene conversion), and chromosome location (e.g. telemere, centreme, etc.).
- Selection: coding sequences, promoters, 3' UTRs, introns, functional RNAs, etc.
- GC content: often garge variation across the genome: e.g. in human, varies from 35% to 60%. Call the regions of high local similarity of GC content isochores. GC content variation can be caused by utational bias, selection (different on different regions) and biased gene conversion.

Patterns of polymorphism and recombination in model organisms:

- Yeast: excess polymorphism (copy number variation) in subtelomeric regions. Also note that these regions are enriched for genes with functions in transport, fermentation, etc.
- Fly: low recombination rates and polymorphism in centromeres, Y chromosome, telemere. Note Y chromosome effectively has a smaller population size, thus low polymorphism.

Polymorphism and LD pattern in human genome (average):

- Data:
  - Earlier projects: sequencing of chosen genes/regions of multiple individuals.
  - HapMap project: about 6.1 million SNPs in samples from 90 Yorubans (Nigeria), 90 CEPH (European ancestry) and 45 Han and Japanese. Not phased. Indels underrepresented (about one per 10 SNPs).
  - dbSNP: all known SNPs and indel polymorphism.
- Variant density: (1) SNPs: 3M in each pair of individuals (1 in 1000 bp), 30M in our species. (2) CNVs: 3-7 large CNVs per individual, 5-10% individual have CNV $>$ 100 Kb, 1-2% have CNV $>$ 1Mb.
- Ascertainment bias of common variants: in HapMap, a SNP is genotyped in large populations only if it is found at least twice in the sample. Thus rare alleles are missed in the chosen SNPs. The bias can be seen easily from site frequency spectrum: uniform distribution (as opposed to large number of rare alleles as expected by mutation-drift). Need correction for analysis of polymorphism: the site frequency is weighted by the probability of discovery of a SNP.
- Site frequency spectrum: with full sequence data (not HapMap), the excess of rare alleles, or Tajima's $D < 0$, across the genome. Thus not caused selection, but by human population growth (recent mutations).
- Haplotype blocks: block structure (high LD, measured by pairwise r^2, within blocks, but low LD between blocks). However, note that, even with very close regions, r^2 could be small.
- Local recombination rates: recombination hotspots, 80% of recombination falls in 10-20% chromosome.

Polymorphism and LD patterns across different populations:

- Polymorphism: about six genetic clusters of human, corresponding to major geographical regions. F_{ST} is about 0.07. About 93-95% of total genetic variation occurs bewteen individuals within each group, with only 3-5% between groups.
- LD: African populations have larger effective population size, increased level of diversity, and reduced LD. Even though LD patterns are different, a common set of SNPs are highly predictive of linked SNPs in all human populations.

Culture in human evolution:

- Examples of culture influencing human evolution: in general, they relax selective pressure, and may be responsible for accumulation of deleterious mutations in human genome. Examples: lactose tolerance, Vitamin C synthesis in chimp (lost because of fruits in diet), short-sightedness and spectacles.
- Inferring the role of culture in human evolution: a critial question is when this happened, before or after major migrations out of Africa. One scenario: gene-culture co-evolution leading to genetic difference in response to culture difference among different populations (e.g. ashkenazi jews might have high IQ - very controversial hypothesis).

Understanding Site Frequency Spectrum in human population [personal notes]

- Goal: we estimate the SFS under neutrality, this would help interpret the empirical patterns from genetic variation data.
- Number of variants: this depends on the gene length, e.g. $L = 5000$ bp and $N = 10000$, $\mu = 10^{-8}$, we have:
$$\theta = 2N\mu L = 1 \tag{2.116}$$
So at $n = 1000$, we have 7.5 variants in the gene and 10 at $n = 10000$. Nevertheless, it was found that the actual number of variants is much higher (10 times), see [Nelson & Mooser, Science, 2012].
- Percent of rare variants among all variants: use $\mathrm{E}(S_i) = \theta/i$, we can show that:
  - At $n = 1000$ and AF < 0.05 as cutoff: percent of RVs is 60%.
  - At $n = 10000$ and AF < 0.05: percent of RVs is 69%.

  With natural selection, this ratio is higher.
- Percent of singletons: similarly
  - At $n = 1000$: 13% of variants are singletons.
  - At $n = 10000$: 10% are singletons.
- Rates of rare variants per individual: about 100 nonsyn. variants.

Most rare missense alleles are deleterious in humans: implications for complex disease and association studies [Kryukov & Sunyaeva, AJHG, 2007]

- Background: why deleterious mutations could accumulate in the human population?
  - Evolution is not effective at eliminating these mutations: late-onset diseases, balancing selection, change of environment and lifestyle (e.g. thrifty genes).
  - Mutation-selection balance: relatively weak selection combined with a high mutation rate. Implication: Common Disease-Rare Variants.
  - The distinction between the two hypothesis largely lies in the parameters: the rate of deleterious mutations (the mutation rate and the chance of being deleterious) and the strength of selection acting on them.
- Data: sequences of 37 genes in about 700 people (obese and healthy, but the genes are not obesity-related).
- Estimating the fraction of strong deleterious mutations among new missense mutations: "strong deleterious" is defined as LoF, similar to non-sense or splice site.
  - Idea: if all missense mutations are strong deleterious, then in the set of such mutations, it will be very frequent relative to the nonsense or splice site mutations (the ratio is determined by the mutation rates). Otherwise, the ratio would be much smaller.

- Result: among all known disease-causing mutations (HGMD, Mendelian disease), the ratio of missense over nonsense is $3.9:1$, but by mutation rate, the ratio should be about $19.7:1$. So the ratio of strong deleterious mutations (causing gene LoF) in missense is about $3.9/19.7 = 20\%$.

- Estimating the fraction of effectively neutral muations among new missense mutations:
  - Idea: these mutations will be fixed at the same rate as the syn. mutations. Thus by comparing the rate of fixation (using human-chimp comparison) of missense vs. syn. mutations, we can estimate this fraction.
  - Results: about 27% are effectively neutral.

- Estimating the fraction of mildly deleterious alleles among standing variants: we known that about 53% of new missense mutations are mildly deleterious (1 minus 20% and 27%). But what about the fraction in existing variants?
  - Using sequences of 37 genes in human: we use polymorphism data, specifically, the number of segregating sites (which should be proportional to mutation rate under neutrality). Focus on singletons in this analysis. The N_a/N_s ratio (the number of nonsyn. vs. syn. sites) is compared with the theoretical ratio (based on mutation rate, about 2.2). The observed ratio is 1.49, thus 32% (1-1.49/2.20) of missense mutations are deleterious.
  - Using existing SNP data in human: for common SNPs, N_a/N_s is close to neutral expectation. For rare variants (MAF less than 1%), the majority (52%-71%) of amino acid substitutions are mildly deleterious.

- **Summary: among missense mutations, about 20% are strongly deleterious, 27% neutral and the rest (53%) mildly deleterious**.

- Lessons/Remark:
  - **General strategy for estimating selection/fraction of deleterious mutations: extension of dN/dS test**. Under neutral model, the expected ratio of different types of variants (singletons) is determined by mutation rates (in some cases, these are known, e.g. ratio of syn. vs. nonsyn. mutations). The depature of this pattern is due to selection and can be used to estimate the strength of selection.
  - Limitation: assumption that under alternative model (selection), all singletons (variants) will be removed by selection.
  - Sources of data to estimate these fractions: human-chimp divergence (pattern of fixation vs. mutation), known disease-causing mutations, human polymophism data (pattern of singletons).

A map of human genome variation from population-scale sequencing, [1000 Genome Project, Nature, 2010]

- Data: 1000 Genome Pilot Project, low-coverage (2-4) WGS of about 100 individuals and a small number of trios and WES of about 600 individuals.

- Power analysis for variant discovery:
  - We estimated that 250 samples sequenced at low coverage would be needed to find 99% of the synonymous variants in an individual, and with 320 sequenced samples 98.5% of nonsynonymous and 96.3% of LOF variants would be found.

- Pattern of genetic variation:
  - We estimated that an individual typically differs from the reference at 10,000-11,000 nonsynonymous sites.

  - LoF variants per individual: in frame indels (190-210), premature stop codons (80-100), splice site disrupting variants (40-50), and deletions that shift reading frame (220-250).
  - Each genome is heterozygous for 50-100 variants classified by the Human Gene Mutation Database (HGMD)
  - Putative functional variants had an allele frequency spectrum depleted at higher allele frequencies
- Selection on the genes:
  - Comparison of diversity (measured by average heterozygosity) across different regions: lowest in exons, etc. reflecting selection.
  - Diversity/divergence: roughly constant for different elements (exons, introns, etc.), suggesting that variation in diversity can be explained that by divergence. Explanation: a certain fraction is under selection, the rest (polymorphic sites) are mostly neutral, thus both diversity and divergence are scaled by the same ratio (1 - the fraction of selected sites), so the ratio is constant. The results suggest that the common part of the allele frequency spectrum is dominated by effectively neutral variants.
- Applications of 1000 Genome data:
  - Impute untyped variants in GWAS: find better association. Ex. in one eQTL study, by imputation, the SNP significance is increased from P value 10^{-8} to $< 10^{-15}$.
  - Filter of variants in the study of Mendelian diseases.
- Remark: the results do not show that most rare variants are neutral. The diversity (average heterozygosity) is dominated by common variants, so the results only suggest that common variants tend to be neutral.

Genetic variation and population size: [Barton, PG, 2010] the neutral theory predicts that the nucleotide diversity (the number of mismatches between two sequences), $E(\Pi) = \theta = 4N\mu$, but the observed diversity varies only about an order of magnitude, while the population size varies far more. The possible explanations (for variations at very large populations):

- Effective population size: in the short-term, the factors such as sex ratio may reduce population size by one order of magnitude; in the long term, population bottlenecks limit the effective population size.
- Selective sweeps: genetic hitchhiking reduces genetic variation as linked alleles are fixed together. Evidence: regions of low recombination rate have lower level of genetic variation.
- Deleterious mutations: eliminated by selection, and reduce genetic variations.

Adaptation in large populations: [Barton, PG, 2010]

- Mutations at large populations: e.g. for Drosophila, the actual population size is much bigger than that suggested by the genetic diversity (about 10^6). Even in a single orchard, the population size could reach 10^6, and the size in a local area may reach 10^8. Thus assuming mutation rate $\mu \approx 10^{-8}$, effectively, the mutation rate per nucleotide is: $2N\mu \approx 2$, i.e. every possible mutation will occur in every generation in a local area.
- Adaptation of large populations: because of high mutation rates, benefical mutations (either from standing variation or from new mutations) are likely to appear independently mutliple times in a local area, following environmental changes (say, introduction of an insecticide). These mutliple mutations all contribute to adaptation and each affecting the surrounding DNA sequences (genetic hitchhiking). Thus after adaptation: the population is dominated by a few independent alleles, each associated with a distinct haplotype.

An integrated map of genetic variation from 1,092 human genomes, [1000 Genome Project, Nature, 2012]

- At the most highly conserved coding sites, 85% of nonsynonymous (NonSyn) variants and over 90% of STOP gain and splice-disrupting variants are below 0.5% in frequency , compared to 65% of synonymous (Syn) variants.
- The least conserved splice-disrupting variants show rare-variant load similar to synonymous and non-coding regions suggesting that these alternative transcripts are under very weak selective constraint.
- The NonSyn to Syn ratio among rare ($< 0.5\%$) variants is typically in the range 1-2 and among common variants in the range 0.5-1.5, suggesting that 25-50% of rare NonSyn variants are deleterious.
  - Certain groups (e.g., ECM-receptor interaction, DNA replication and pentose phosphate pathway) show a substantial excess of rare coding mutations, which is only weakly correlated with the average degree of evolutionary conservation.

Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes [Tennessen, Science, 2012]

- Data: ESP, 111x coverage in 2440 individuals of European (n = 1351) and African (n = 1088) ancestry.
- Purifying selection on rare variants:
  - Of the total SNVs, 57% (285,857) were singletons, and SNVs with three or fewer minor alleles accounted for 72% of all variants.
  - The overall site frequency spectra (SFS) and the SFS for both AAs (African) and EAs (European) are highly skewed, exhibiting a large excess of rare variants relative to the standard neutral model. Tajima's D was highly negative for both EAs (-2.12) and AAs (-2.10) and dropped precipitously as sample size increased.
- Functional impact of RVs: using Polyphen, SIFT, MutationTaster, GERP, PhyloP, etc.
  - About 47% of all SNVs (74% of nonsynonymous and 6% of synonymous variants) are predicted to be deleterious by one or more method.
  - Overlap among methods is modest. For example, only 1% of nonsynonymous variants are predicted to be functional by all seven methods, and variants predicted by any single approach are likely to have a high false-positive rate.
- Summary: most CV's are not under selection, but rare variants are. The evidence of RV selection comes from damaging effect analysis (complements the earlier ASHG'07 study that uses singlenton density and divergence data).

An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People [Nelson & Mooser, Science, 2012]

- Motivation: use sequencing data of 202 genes to study the pattern of genetic variation in human, including, excess of rare variants, inference of population demography, and natural selection.
- Data: 202 genes that are potential drug targets, in 14,002 people (from multiple disease cohorts), mostly European ancestry.
- Excess of rare variants (0.5%):
  - Among all varaiants, 95% are rare, and more than 74% observed in 1 or 2 subjects.
  - Watterson's estimator (θ_W) is nearly 10 times larger than pairwise estimator (Tajima, θ_π). Explanation: exponetial population growth, the coalescence have long terminal branches (a large number) and short roots - comparing with constant population size tree, there are more variants, leading to larger estimator.

- Inference of population demograhy and mutation rates: using only syn. sites.
  - Current: $N_e = 10,000$ based on pairwise estimator, however, $\theta_W >> \theta_\pi$, signature of rapid growth. Using likelihood model with genealogy (MCMC for inference), find 1.7% population growth rate, and effective European population size of 4.0 million.
  - Mutation rates: the same model can also estimate mutation rates. Median estimate: 1.38×10^{-8}.
- Natural selection on genes: using the approach of AJHG'07, comparison of NS vs. S in variants with different AFs. Estimate that roughly 70% of all NS singletons are sufficiently deleterious that they will never reach 5%. Only 13% of new NS mutations are so deleterious that they cannot be observed even as singletons in this sample size.
- Discussion: the estimation of demography ignores background selection on S variants.

2.11 Evolution of Multi-State Systems

Multi-state systems: the fitness of states and mutation rates among states could be very unequal, e.g. some states have high fitness, but mutationally less accessible. We are interested in which states/regions will be most likely: the high-fitness ones vs. the ones with high entropy. Two approximations:

- Mutation-selection balance: applicable if $4N\mu >> 1$. In the two-state case, the frequency of state a would be proportional to $\mu(b,a)/F(a)$, where $\mu(b,a)$ is the mutation rate from b to a, and $F(a)$ is its fitness. In the general case, mean-field approximation.
- Successive fixation: among all possible states, applicable if $4N\mu << 1$. In each time point, population is fixed at one state, and the system can be viewed as a Markov chain. The rate of switching state a to b is given by:

$$u(a,b) = 2N\mu(a,b)\frac{1-e^{-2[F(b)-F(a)]}}{1-e^{-4N[F(b)-F(a)]}} \tag{2.117}$$

  The rate of switching to a favorable mutation with selection coefficient s is approximately:

$$u(a,b) = \mu(a,b)\frac{4Ns}{1-e^{-4Ns}} \tag{2.118}$$

  The rate of switching to a deleterious mutation with selection coefficient s is approximately:

$$u(a,b) = \mu(a,b)\frac{4Ns}{e^{4Ns}-1} \tag{2.119}$$

  Putting the two together, we have $u(4N\Delta s)$ as a monotonic function of $4N\Delta s$ (Δs is signed): $u \approx \mu$ when $|4N\Delta s|$ is close to 0; u decreases to 0 about exponentially as $4N\Delta s$ reduces to negative values; and increases exponentially as $4N\Delta s$ increases to positive values.
- Low-fitness states: Under the mutation-selection balance model, a low-fitness state/region is likely if the mutational influx to this state/region is high. Under the population random walk model, fixation of low-fitness state/region is very unlikely, as probability of fixation of deleterious mutations is apprxoimately $4Ns/e^{4Ns}$ (very small if $4Ns > 10$). It should be understood however, under this scenario, even though low-fitness states are unlikely to be fixed, they could reach substantial frequency.
- Weakly deleterious selections: (the fate of new mutations) for large populations, they will be eliminated by selection; for small populations, we see that, the probability of fixation is (use $e^{4Ns} \approx 1 + 4Ns$): $u(a,b) \approx \mu(a,b)$, thus weakly deleterious selection almost as likely to be fixed as neutral mutations.

Successive fixation model: [Sella & Hirsh, PNAS, 2005]

- Assumption: $N\mu << 1$, the mutation is rare, thus the population will undergo successive fixation (the linked polymorphism can be ignored). The probability of fixation of any site does not depend on other segregating alleles. Also assume that the mutation rates between any two states are equal.

- Equlibrium distribution of states: let W_{ij} be the rate of substitution from state i to j, and f_i be the fitness of state i. Using the equation of substitution rate:

$$\frac{W_{ij}}{W_{ji}} = \frac{f_j^\nu}{f_i^\nu} \tag{2.120}$$

  where $\nu = 2N - 1$. This implies the detailed balance:

$$W_{ij} f_i^\nu = W_{ji} f_j^\nu \tag{2.121}$$

  Let $x_i = \ln f_i$ (called the additive fitness), the equilibrium distribution is given by:

$$P_i^* \propto f_i^\nu = e^{-\nu(-x_i)} \tag{2.122}$$

  This is analogous to the Boltzmann distribution: $-x_i$ can be viewed as the (negative) energy of a state, and ν is analogous to $\beta = 1/k_B T$. When population size is infinite (temperature at 0), the population will be fixed on the highest fitness state.

- Free fitness function: the dynamics of system (how probability distribution of states change) can be characterized by the free fitness function:

$$G = \langle \ln f \rangle + \frac{1}{\nu} S \tag{2.123}$$

  where the first term is the average fitness over all states, and S is the entropy of states, $S = -\langle \ln P \rangle$. So we could write G as:

$$G = \sum_i P_i \left[\ln f_i - \frac{1}{\nu} \ln P_i \right] \tag{2.124}$$

  The free fitness function should monotically increases, and converges to the equilibrium distribution defined above. The equlibrium state reflects the balance between the tendencies to increaese fitness and entropy (analogous to physical system at thermal equilibrium).

Evolution of sequences: let L be sequence length, the total mutation rate is thus $4NL\mu$:

- Short sequences (e.g. TFBSs): for large population ($4NL\mu > 1$), this can be approximated by mutation-selection balance; for small population, may be apprxoimated by a Markov chain.

- Long sequences (regulatory sequences or genes): generally $4NL\mu > 1$, thus substantial polymorphism. When $4NL\mu < 1$, and most mutations are neutral, then the population may be fixed at a single sequence most of the time (see below).

- Limitation of monomorphism assumption: even at $4NL\mu < 1$, the Markov chain model may be inadequate. If beneficial mutations are relatively common, then before fixation of the current mutation (at the level of $4N$ generations), a new "fixable" mutation may occur (more than once per $4N$ generations because of the increase of fixation probability of advantageous mutations).

Chapter 3

Linkage Analysis

3.1 Introduction to Linkage Analysis

Reference: [Thomas, Statistical Methods in Genetic Epidemiology, Chapter 7]
Recombination: [Human Molecular Genetics]

- Recombination fraction: defined as the fraction of recombinants in an individual with heterozygotes in both loci, denoted as θ. Always no more than 0.5 (when two loci are completely independent). Given a heterozygote (A_1B_1, A_2B_2), the frequency of four genotypes in the gametes are:

$$f(A_1B_1) = f(A_2B_2) = \frac{1-\theta}{2} \qquad f(A_1B_2) = f(A_2B_1) = \frac{\theta}{2} \tag{3.1}$$

- Genetic distance is defined in the units of cenimorgan (cM), which corresponds to 1% chance of recombination between two loci. In mouse/human, about $0.5 - 1$ cM/Mb.

- Mapping functions: map genetic distance to physical distance. This is not linear because of the possibility of multiple cross-overs (always no more than 0.5, but physcial distance can be arbitrarily large). Ex. Haldane's function: $w = -1/2 \ln(1 - 2\theta)$, where θ is the recombination fraction.

- Distribution of recombination: highly non-uniform. Ex. very high in the telomere regions of male chromosomes. In human, chromosomes consist of conserved blocks, typically, 20-50 kb, separated by 1-2 kb recombination hotspots (about 95% of all recombinations).

Principles of linkage analysis:

- Linkage mapping in families: consider a polymorphic marker and the disease locus, the ratio of recombinants and non-recombinants in the offsprings (which can be detected through the combination of marker genotype and the disease state) reveals the genetic distance between the two. In particular, the presence of non-recombinants indicate the linkage between the marker and disease (more non-recombinants, tighter linkage).

- Intuition of linkage mapping: suppose we have a fully penetrant locus and a linked marker, then from the data, we can infer the complete genotype of all individuals. Suppose phasing is also not a problem, then we have the haplotypes. This allows us to estimate the fraction of recombinant and non-recombinant haplotypes, thus comparing it with random expectation ($\theta = 1/2$); furthermore, the fraction reveals information of the distance.

- Limits of linkage mapping: when no recombination is present, either from the small sample size (e.g. in family studies) or from the low recombination rates in most part of the genotypes, there will be

only non-recombinants, therefore the linkage mapping cannot narrow down to smaller regions. Ex. with genetic distance $\theta = 0.01$, need at least 100 samples to have one recombination event, thus the resolution would be at the level of 1 Mb.

- Association with disease/LD mapping: the basic idea of enrichment of non-recombinants can be applied at the population level. When the marker and the disease allele is tightly linked, there is LD between the marker and the disease gene (the two co-segregate in a haplotype block). Since the disease genotype is not directly observed, this LD is manifested as the association of the marker(s) with the disease status.
- Linkage vs. assocition (LD) mapping: let M be a marker and D be a disease locus, linkage mapping relies on recombination between M and D within families, and association mapping relies on the LD between M and D at population level. The two methods are different:
  - It is possible to have linkage but not association: M and D may be close, but not in the same LD block, then association would not identify D, but linkage mapping can.
  - It is possible to have association but not linkage: population stratification may create false associations.

Direct counting method:

- Method: suppose we can infer the full haplotypes of all subjects, and find that the number of recombinants and non-recombinants are r and s respectively. Then we can use McNemar's test: $(r-s)^2/(r+s)$ follows χ^2 distribution under the null model.
- Phasing problem: In practice, even for fully penentrant diseases, the direct counting method generally cannot be applied because the phasing is unknown. Under special cases, however, the phasing may be inferred, e.g. when one parent has identical alleles, then which one is transmitted to the child does not matter.

Identity by state (IBS) and identity by descent (IBD) in sibling pairs: [Thomas, Chapter 7]

- IBD and type of relationships: Table 7.2 of [Thomas]. Let π_0, π_1 and π_2 be the probabilities of sharing 0, 1 and 2 IBD alleles. Then IBD for some common relationships are shown in Table 3.1
- Computing IBD from marker genotypes of sib-pairs: let Z_A be the number of shared IBD alleles at the marker A between the sib-pair, and M_o^A be the observed marker genotypes at A. We define:

$$Q_a^A = P(Z_A = a | M_o^A) \tag{3.2}$$

  To compute this, we first convert it to $P(M_o^A | Z_A = a)$ using Bayes Theorem. This term is then computed by summing over the parental genotype:

$$P(M_o^A | Z_A = a) = \sum_m P(M_p^A = m | p^A) P(M_o^A | M_p^A = m, Z_A = a) \tag{3.3}$$

  where p^A is the allele frequency in the population at the marker A and $P(Z_A) = (1/4, 1/2, 1/4)$. Note that not genotypes are possible: we should only consider those that are consistent with $Z_A = a$.
- Computing IBS between sib-pairs: suppose we want to compute the probability that IBD between the sib-pair is n. This is done by summing over parental genotypes and offspring genotypes:

$$P(\text{IBS} = n) = \sum_{M_p} P(M_p | p) \sum_{M_o} P(M_o | M_p) I(\text{IBS} = n | M_o) \tag{3.4}$$

  where p is the allele frequency and $I(\cdot)$ is the indicator function.

Relationship	π_0	π_1	π_2
MZ twins	0	0	1
Full sibs	1/4	1/2	1/4
Parent-offspring	0	1	0

Table 3.1: IBD for different relationships

- Computing conditional IBD at different loci between sib-pairs: if we know IBD at one marker, then we should be able to get some information of the IBD at an adjacent marker. In the extreme case, when the two are completely linked, the IBD should be identical. More generallly, the conditional distribution would depend on the recombination fraction θ. If we denote s and t the segregation indicators of the four alleles in the markers A and B, the conditional IBD distribution:

$$T_{ab} = P(Z_B = b|Z_A = a) = \frac{\sum_{s,t} I(Z_A = a|s)I(Z_B = b|t)\theta^{r_{st}}(1-\theta)^{4-r_{st}}}{P(Z_A = a)} \tag{3.5}$$

  where r_{st} is the number of recombinations, which is derived from s and t.

Questions of linkage analysis:

- How sequencing would benefit linkage analysis: because of the limit of resolution, it is hard to narrow down to individual genes even if we have the full sequences. The true benefits are probably the causal loci from the identification of very rare variants, and other information such as natural selection.
- In ASP analysis: how are the IBD status are determined?

3.2 Parametric linkage analysis

Reference: [Human Molecular Genetics, 3rd Ed, Chapter 13, 14; Yang, Introduction to Statistical Methods in Modern Genetics, Chapter 2; Thomas, Chapter 7]

Two-point mapping:

- Intuition: the fraction of non-recombinants (between the disease locus and the marker) indicates the extent of linkage between the two loci. To detect recombinants and non-recombinants, one of the parent must be a double heterozygote; if not the recombinantion will not change the probabilities of the haplotypes in the offsprings.
- Model: given the marker genotype and disease status of a pedigree, we test the hypothesis in terms of the genetic distance θ. H_0: the marker is not linked to the disease, i.e. $\theta = 0.5$ vs. H_A: the marker is linked to the disease, i.e. $\theta < 0.5$. This is done through the likelihood ratio test. We define g as genotype and x as phenotype. $P(x|g)$ is the penetrance function.
- Likelihood [Thomas]: let $\mathbf{Y}$ be the phenotypes and $\mathbf{M}$ be the observed marker genotypes. Furthermore, we have parameters θ for recombination, and $\omega = (f, q)$ be the segregation parameters (penetrance and allele frequency). The likelihood is:

$$L(\theta, \omega) = P(\mathbf{Y}|\mathbf{M}) = \sum_g \prod_i P_f(Y_i|G = g_i)P_\theta(G = g_i|M_i, g_{f_i}, g_{m_i}, M_{f_i}, M_{m_i}) \tag{3.6}$$

  where f_i and m_i are father and mother of i, respectively. The last term will be replaced by $P_q(G = g_i|M_i)$, if i is a founder. Usually most linkage analysis assume linkage equilibrium. The summation over g (four possible genotypes per subject) can be done by the Elsteon-Steward algorithm. The transition probabilities in one generation depend on recombination (Table 7.1 of [Thomas]). In general, phasing is unknown and some marker genotypes may be unobserved. When a marker has two alleles, there are 10 possible hayplotypes (Table 7.1).

- Example: Suppose we are analyzing the data where A and D represents the disease status (affected or unaffected). We use D, d for disease gene (recessive) and $1, 2, 3, 4$ for marker alleles. Consider the following observation:

$$U14 \times U14 \to A11 \tag{3.7}$$

  The likelihood computation consists of the following steps:

  - Parent genotype: since the disease is recessive and one child is affected, the parent genotypes must be Dd and Dd. Thus we have the mating type: $(Dd)/(14)$ and $(Dd)/(14)$ (where the parenthesis means the phase is unknown).
  - Phasing: each parent could have two phases $D1/d4$ or $D4/d1$, thus there are a total of 4 mating types.
  - Offsprings: conditioned on one mating type with phasing: $D1/d4 \times D1/d4$. The genotype of the offspring is $(DD)/(11)$ or $(Dd)/(11)$. The first can be only obtained through no-recombination and the second by recombination in one parent only, thus the probability is:

$$P(A11|D1/d4, D1/d4) = (1-\theta^2)/4 + \theta(1-\theta)/2 \tag{3.8}$$

- Parameter estimation: generally linkage analysis assumes that the segregation parameters are are known from separate segregation analysis, thus f and q are fixed in estimation. This simplifies parameter estimation. In addition, the segregation parameters may be biased if using only the genotyped families (ascertainment bias).

- Importance of extended families: more information regarding phase is provided. Ex. when the phase of an individual cannot be resolved: DA_1 or DA_2, where D is the disease allele; the genotype of A locus at the parent carrying D may help to resolve.

LOD score:

- Definition: the LRT score is defined as:

$$Q = -2 \ln \frac{L(\theta = 0.5)}{\max_\theta L(\theta)} \tag{3.9}$$

  It is common in genetics to use LOD score, defined as:

$$LOD = \log_{10} \frac{\max_\theta L(\theta)}{L(\theta = 0.5)} \tag{3.10}$$

  For a particular θ, often use:

$$Z(\theta) = \log_{10} \frac{L(\theta)}{L(\theta = 0.5)} \tag{3.11}$$

- Distribution: Q normally follows χ^2 distribution with df 1; in our case, since H_A is one-sided, the p value will be half of that under the two-sided alternative hypothesis. Thus Q follows a mixture distribution:

$$Q \sim \frac{1}{2}\chi^2 + \frac{1}{2}\{0\} \tag{3.12}$$

  Then $4.6 \times LOD$ follows the mixture distribution. The threshold for LOD score is usually $+3$ for acceptable (and -2 for exclusion). This is used even when the number of markers is high [Lander & Botstein, 1989].

Performance/power analysis: we study this with double backcross $(Aa)/(21) \times (aa)/(11)$, where A, a are trait alleles and $1, 2$ are marker alleles (phasing may or may not be known).

- Strategy: to assess the power of a test, we use the distribution of the test statistic under H_A. This can be done in two manners: expected LOD score and the sample size requirement to achieve a certain power. The key to derive the distribution is to express the LOD score (or $Z(\theta)$) under a given θ, in terms of probabilities that can computed from basic model parameters including θ.

- Complete penetrance with phasing known: the mating type is $A2/a1 \times a1/a1$. Suppose there are n offsprings from this type of mating (often in experimental animals/plants, or many families with the same mating types), we could test the recombination parameter θ using the fraction of non-recombinants. The sample size requirement, at the true $\theta = \theta^*$ vs. $H_0 : \theta = 1/2$ is given by (from power calculation of binomial distribution):

$$n = \left[\frac{z_\alpha \sqrt{\theta_0(1-\theta_0)} + z_\beta \sqrt{\theta^*(1-\theta^*)}}{\theta_0 - \theta^*} \right]^2 \tag{3.13}$$

  Note that the threshold calculation can also be performed on the LOD score (which follows χ^2 distribution, but the square root follows normal distribution).

- Incomplete penetrance: reduces the power (or increases sample size requirement). In this case where the penetrance is λ, we cannot use the simple binomial distribution. We have double backcross: $A2/a1 \times a1/a1$, the probability of the offsprings (let A be affected and U unaffected):

$$\begin{array}{ll} p_1 = P(A11) & = (1-\theta)\lambda/2 \\ p_2 = P(A12) & = \lambda\theta/2 \\ p_3 = P(U11) & = (1-\lambda+\lambda\theta)/2 \\ p_4 = P(U12) & = (1-\lambda\theta)/2 \end{array} \tag{3.14}$$

  The likelihood function is:

$$L(\theta) = \prod_{i=1}^{4} p_i(\theta)^{n_i} \tag{3.15}$$

  where n_i are the number of individuals of the type i. This would allow us to compute the expected LOD as (using $E(n_i) = np_i$):

$$ELOD = n \sum_{i=1}^{4} p_i \log_{10} \frac{p_i(\theta)}{p_i(0.5)} \tag{3.16}$$

- Unknown phasing: reduces the power. This is similar to the case above, but need to replace the probabilities taking unknown phasing into account. The distribution of Z is approximately normal because it is a sum of Z-scores for many families. The expectation and variance of Z can be computed from Equation 3.15.

Multi-point mapping: how can multiple markers help?

- More informative families: if there is only one marker, then in some families, the marker may happen to be monomorphic in the affected parents, thus these families are not informative. With more markers, it is likely that for any family, at least some marker may be polymorphic and thus informative.

- Rare events: with more markers, say two, it is likely to observe double recombination events. In general, rare events are more informative. Ex. in estimating the parameter of a Bernoulli distribution, p, the variance of the statistic:

$$\mathrm{Var}(\hat{p}) = \frac{p(1_p)}{n} \tag{3.17}$$

  It is clear that the variance is small at small value of p. Suppose we have two markers B, b and E, e, and disease alleles D, d. Then in the offsprings of BDE/bde, the haplotypes BdE or bDe require two recombination events.

Multipoint linkage analysis: [Thomas, Chapter 7]

- Order of markers and direct generalization of two point analysis: when the number of markers is small, we can directly generalize the two-point analysis. For simplicity, we consider only markers, say A, B, C. Consider one particular ordering ABC, the likelihood (consider only genotypes) would depend on θ_{AB} and θ_{BC} - bascially we consider all recombinations between A and B, and between B and C. We can also simplify the model by using a single parameter x, the location of B. Then both θ_{AB} and θ_{BC} can be written as functions of x using the Haldane map function.

- Challange of multipoint linkage: when there are multiple markers, the likelihood calculation would involve summartion of G, which goes exponential with the number of markers. Furthermore, there is clearly dependency between markers, and the likelihood cannot be factored into individual markers.

- Intuition: there is some conditional independence between markers: e.g. suppose two markers, A and B, between two subjects, are IBD, then one can infer that there is no recombination between the two markers and thus all markers between A and B are IBD. More generally, given the IBD status of two markers A and C, the IBD status of any marker B between A, C would be conditionally independent of all the other markers.

- Computing IBD probabilities by HMM (Lander-Green algorithm): following our notation in the section "IBS and IBD in sibling pairs", the IBD probabilities of B (a vector, π_0, π_1, π_2) in a sib pair is given by:

$$\pi_B = Q_A T_{AB} T_{BC} Q_C \tag{3.18}$$

  In sibpair analysis, we need to determine the IBD of the location of the disease locus x. Let it be Z_x, and we need to estimate $P(Z_x = z)$ where $z = 0, 1$ or 2. Suppose it's between marker $i-1$ and i, we can have the IBD status of x as:

$$\pi_x = P(Z_x|M, x) = \mathbf{1} Q_{1,l-1} T_{l-1,x} T_{x,l} Q_{l,L} \mathbf{1} \tag{3.19}$$

  where $Q_{l,m} = Q_l Tl, l+1 Q_{l+1} \cdots Q_{m-1} T_{m-1,m} Q_m$ represents the results of matrix multiplication of all markers between l and m inclusive. Now we can extend this to pedigrees. We use indciator variable Z for possible IBD configurations (for each subject in the pedigree, which ancestral allele its genotype comes from). The corresponding IBD probabilities are denoted as $\Pi(Z_x)$, and can be computed similarly.

- Parameteric analysis: the likelihood can be represented in terms of the location x and the segregation parameter $\omega = (f, q)$:

$$L(x, \omega) = P(Y|M) = \sum_g P(Y|G_x = g) P(G_x = g|M) = \sum_z P(Y|Z_x = z) \Pi(Z_x = z) \tag{3.20}$$

  Note that instead of summing over genotypes (which is exponential of the number of markers), we sum over the IBD configurations. The term:

$$P(Y|Z_x = z) = \sum_a \sum_{s|z} P(Y|G(s, a); f) P(a; q) \tag{3.21}$$

  where a is the founder alleles and s segregation indicators (which alleles are transmitted). The summation is over all segregation indicators that are consistent with the IBD configruation z.

- **Lesson**: the genotype information is essentially equivalent to the "history" of the alleles, i.e. which ancestral alleles is inherited. This is technically the IBD status (IBD: whehther two alleles, in pairwise comparison, come from the same ancestor). In contrast to genotypes, IBD follows simpler rules, determined by Mendelian law of segregation and recombination. E.g. in sib-pairs, IBD follows simple distributions, which IBS depends on parental genotypes.

3.3 Non-parametric Linkage Analysis

Reference: [Human Molecular Genetics, 3rd Ed, Chapter 15; Yang, Introduction to Statistical Methods in Modern Genetics, Chapter 2, 4]

Motivation: why do we need (or not need) non-parametric methods?

- Incomplete penetrance: when penetrance is low and complex (e.g. in complex traits, or age-dependent), the parametric likelihood method may be difficult to apply. The common solution is the Affected Sib Pair (ASP) design.
- Low resolution: with parametric linkage methods, when θ is small, the recombination events will be very rare in a few generations, thus it is not possible to estimate small value of θ. This limits the resolution of linkage methods. The idea is to effectively look at extended families or subpopulations (more recombinations), and look for association between markers and the disease status.
- Near-Mendelian families: even for complex traits, it may be possible that within families, the traits are Mendelian (e.g. breast-cancer). Thus parametric linkage analysis within families may identify genes of complex diseases.

Intuitions of non-parametric linkage analysis:

- If there is linkage, the markers should be correlated with the disease trait. To reformulate it, if we divide all individuals into two groups, affected and unaffected, then one allele of the marker should be overrepresented in the affected group. A special case is that among affected pairs of relatives, they tend to share the same alleles.

Affected sib pairs (ASPs):

- Strategy: suppose the marker is very polymorphic. A mating type is typically 12×34. Let S_m be the number of alleles shared by two affected siblings, then under H_0 of no linkage:

$$P(S_m = 0) = P(S_m = 2) = 1/4 \qquad P(S_m = 1) = 1/2 \tag{3.22}$$

  Let there be n sib pairs with n_0, n_1, n_2 pairs sharing 0, 1 and 2 alleles. The departure from the null distribution (the ratio of $1 : 2 : 1$) thus indicates linkage. The simple test would be the χ^2 test of departure. But this test is not very specific to H_A, so a better test is: the increased number of shared alleles if there is linkage:

$$Z = \frac{n_2 - n_0}{\sqrt{n/2}} \tag{3.23}$$

  It is simple to show that Z follows normal distribution under H_0.

- More than two affected sibs: the idea is similar, if the marker is linked to the disease locus, the number of shared alleles will be high. Consider the most frequent haplotypes, one from each parent, let N_i be the count of these two haplotypes in the i-th family, then the test statistic is $N = \sum_i N_i$.
- IBD: when the markers are not very polymorphic, it is better to use haplotypes. The method would then need to recognize IBD shared by affected sibs.

Extensions:

- Relative pair methods: for more distant pairs, IBD are harder to determine, so use IBS instead.
- ASP for quantitative traits: the idea is similar, pairs with similar traits should have more allele sharing. Let the difference between two siblings of a pair be: $D_i = Y_{i1} - Y_{i2}$, we could regress D_i with π_i the expected proportion of alleles shared IBD. More generally, we assume that the traits of a pair follow bivariate normal distribution with mean $\mu\mathbf{1}$, marginal variance σ^2 and correlation coefficient ρ, which depends on the IBD status.

High resolution mapping:

- High resolution mapping with extended pedigrees: within an extended family (e.g. with a lot of inbreeding), the disease alleles in all patients generally come from the same ancestral mutation, therefore, all these disease carriers will share the same haplotype blocks. To define these blocks that contain the disease locus:
  - Autozygosity mapping: for receissive diseases, a patient carries two identical blocks, thus the marker should be homozygous (in this case, also autozygous as they are identical by descent). Thus search for blocks where the markers are homozygous.
  - Shared haplotype blocks: search for such shared blocks in the haplotypes of all patients with the disease.
- Population association studies/LD mapping:
  - Association of markers and diseases: generally speaking, association of disease locus and disease, and LD between marker and disease locus, create non-random association between marker and disease. Specifically, if marker is in tight LD with disease allele, then after a certain number of generations, there will be still a relatively large fraction in the population where the marker and the allele are in the same block, thus creating association with the disease.
  - Remark: possilbe causes of association between markers and diseases - natural selection (if an allele helps survival of that disease), population stratification, LD with true disease loci, etc.

3.4 QTL mapping

Reference: [Lynch & Walsh, Genetics and Analysis of Quantitative Traits, 1998]

Principles of QTL mapping:

- Idea: (1) if we can measure the genotype, then we could test if the association bewteen the genetic variation at any locus and the phenotypic variation; (2) even if we do not have the complete genotype, the genetic markers close to the QTL can be a proxy of the true QTL, if the marker remains linked to the QTL. In QTL mapping, usually starts with crossing two inbred lines, thus creating LD (similar to population admixture).
- Experimental procedure: "create" genetic variations among different strains by crossing in different ways. In particular, crossing two inbred strains, and then apply backcross or intercross on F_1, and collect data on the many progenies. For any locus, the progenies will have one of several genotypes.
- Why does the procedure work, and how it depends on recombination? If a marker, M_1 is close to L (the true QTL), then M_1 will be associated with the trait as M_1 allele will be coupled with L; a distant M_2 will not be associated with the trait because of recombination (alleles of M_2 are randomized). Therefore, recombination allows one to narrow down the range of L: high recombination rates, large samples (thus more recombinations), and more markers will lead to more precise localization of L.

Experimental procedures of QTL mapping: crossing of two inbred lines, and F_1 is heterozygotes for both marker and QTL, with genotype MQ/mq, where M/m is for marker alleles and Q/q for QTL alleles. The gametes of F_1 however, could have four genotypes MQ, mq, Mq, and mQ with probabilities dependent on recombination rate c.

- Intercross (F_2 design): crossing F_1 with F_1.
- Backcross: crossing F_1 with one of the parental lines. Note that not all genotypes will be examined, e.g. with the parent $MMQQ$, the genotypes of the offsprings will have MM and Mm, but not mm, so this method is not good for dominance effects.

- F_t design: to increase the resolution of QTL mapping (more recombinations), by crossing F_1, F_2, etc.
- Recombinant Inbred Lines (RILs): take F_1 line through multiple rounds of selfing or multiple generations of brother-siste mating. The resulting lines have no within-ine genetic variance (each RIL represents a different multi-locus genotype).

Designing experimental procedures [personal notes]

- In model organism studies, one has the freedom to design experiments to best acheive the goal of mapping trait loci. Considerations of experimental design may include: power, resolution, non-additive genetic model (e.g. dominance), gene-gene and gene-environment interactions, etc.
- Power consideration: eg. in F_1 intercross, the AF is 50% and this leads to maximum power.
- Minimizing variance: e.g. small variance of trait leads to better power, so we control the environment to minimize environment-introduced variance. Or we choose two F_0 strains that are genetically similar, but different in our interest trait; this would minimize variance due to genetic background, when we analyze any particular locus.
- Resolution consideration: e.g. F_t design achieves higher resolution. Another example, QTL generally does not tell the causal gene; thus combine QTL with eQTL to find the causal gene (expression trait).
- Creating more genetic variations: because of selection, some important loci may not have natural genetic variations. So use artificial selection or other means to introduce more gentic variations, in more genes/loci.

Basic concepts for statistical methods of QTL mapping:

- Conditional probabilities of QTL genotypes: we are interested in knowing $P(Q_k|M_j)$, where Q_k is the QTL genotype and M_j is the genotype of the marker. This is written as:

$$P(Q_k|M_j) = \frac{P(Q_k, M_j)}{P(M_j)} \tag{3.24}$$

  We consider several cases:

  - Single marker in F_2 design: In F_1, we have:

$$P(MQ) = p(mq) = (1-c)/2 \qquad P(Mq) = P(mQ) = c/2 \tag{3.25}$$

    In F_2, we have the probability of MM, Mm, mm be $1/4, 1/2, 1/4$ respectively. And the joint probability can be easily computed from multiplying the probabilities of gametes. So we have for example:

$$P(QQ|MM) = (1-c)^2 \qquad P(Qq|MM) = 2c(1-c) \qquad P(qq|MM) = c^2 \tag{3.26}$$

  - Two markers in F_2 design: suppose the distance between M_1 and Q is c_1, and between M_2 and Q is c_2. Assume there is no recombination inteferance, then $c_2 = c_{12} - c_1$, where c_{12} is the distance between markers 1 and 2 (known). So there is only one unknown parameter. In F_1, we have:

$$P(M_1QM_2) = (1-c_1)(1-c_2)/2 \qquad P(M_1qM_2) = c_1c_2/2 \tag{3.27}$$

    Similarly, we could derive the conditional probabilities:

$$P(QQ|M_1M_1M_2M_2) = \frac{(1-c_1)^2(1-c_2)^2}{(1-c_{12})^2} \tag{3.28}$$

– Single marker in backcross design: with crossing with $MMQQ$. This can be done similarly:

$$\begin{array}{ll} P(QQ|MM) = 1-c & P(Qq|MM) = c \\ P(QQ|Mm) = c & P(Qq|Mm) = 1-c \end{array} \tag{3.29}$$

- Genetic model: we assume the following parameterization of the means of three genotypes of the QTL:

$$\mu_{QQ} = \mu + 2a \qquad \mu_{Qq} = \mu + a(1+k) \qquad \mu_{qq} = \mu \tag{3.30}$$

- Marker-class means: the mean of a marker genotype M_j can be expressed as:

$$\mu_{M_j} = \sum_k \mu_{Q_k} P(Q_k|M_j) \tag{3.31}$$

where Q_k is the k-th genotype of the QTL. For a simple case of single marker (two alleles), and F_2 design, we could compute μ_{MM}, μ_{Mm} and μ_{mm}, and have the following results:

$$(\mu_{MM} - \mu_{mm})/2 = a(1-2c) \tag{3.32}$$

$$\frac{\mu_{MM} - (\mu_{MM} + \mu_{mm})/2}{(\mu_{MM} - \mu_{mm})/2} = k(1-2c) \tag{3.33}$$

Thus one strategy is to test for significant differences between mean trait values associated with different marker genotypes.

Methods for single marker analysis:

- Linear models: the idea is to test association bewteen marker genotype and the trait with ANOVA or regression: the difference of traits between genotypes carries information of the effect size and the distance. Let z_{ik} be the trait value of the k-th individual of marker genotype i, we have:

$$z_{ik} = \mu + b_i + e_{ik} \tag{3.34}$$

- Maximum likelihood (ML) method:
  - Mixture model: the distribution of trait values under marker genotypes can be modeled as a mixture of Gaussian distribution:

$$P(z|M_j) = \sum_k P(z|\mu_{Q_k}, \sigma^2) P(Q_k|M_j) \tag{3.35}$$

where summation is over all possible genotypes of the QTL. The model has five parameters: marker means for three QTL genotypes, σ^2 and the recombination ratio c. Parameter estimation can be done through EM.

  - Confidence interval of c: this is needed for defining the boundary of the QTL. It can obtained from (1) asympotic distribution of MLE; (2) bootstrapping procedure: sample with replacement.

Interval mapping:

- ML interval mapping: similar to the single marker mapping, by replacing the term $P(Q_k|M_j)$ with the condition probabilities with two markers.

- Approximating ML interval mapping by regression: let M_i be the marker genotype of the i-th individual, the expected trait value for M_i is given by the marker-class means equation:

$$\mu_{M_i} = (\mu + a)P(QQ|M_i) + (\mu + d)P(Qq|M_i) + (\mu - a)P(qq|M_i) \tag{3.36}$$

where we have the genetic model: $\mu_{QQ} = \mu + a, \mu_{Qq} = \mu + d, \mu_{qq} = \mu - a$. So regression coefficients of the trait values, z_i (which is μ_{M_i} plus an error term), on the marker genotype M_i, will give the relevant parameters, a, d and c.

Lessons from QTL mapping [personal notes]:

- Generally 100-200 individuals with 20-100 markers are able to detect QTLs.
- Genetic architecture: 222 traits, almost half (45%) of all traits had a QTL accouting for at least 20% of the total phenotypic variance ($R^2 > 0.2$). When all detected QTLs are considered, the ratio was raised from 45% to 84%. However, it should be noted that when the QTL power is low, the effect may be (severly) estimated.
- Dominance is common, but epistatis is rare. Possible explanations: the methods screen with significant single-locus effects, thus may miss many epistatic effects; the sample size for multi-locus genotype is small, thus reducing the power.

Comments on QTL mapping [personal notes]:

- One major limitation of QTL (and eQTL for mapping GRNs): the crucial genes of the corresponding process may be highly conserved in the population, thus not enough genetic variations.
- Comparison of linkage mapping and association studies: in linkage mapping, only a small number of genetic variations are explored, thus not powerful in detecting common variants. Also linkage mapping relies on recombination events within the pedigree, thus more coarse-grained than association mapping, which relies on LD between nearby variants.

Multiple loci in QTL mapping: [Sen & Churchill, Genetics, 2001]

- Introduction:
  - Motivation: a signal in QTL mapping may represent multiple linked QTLs, thus need a method to test and separate linked QTLs. Also need to model the effect of epistasis.
  - Strategy: similar to single marker analysis, the trait values are determined by the QTL alleles, and the QTL alleles are related to the marker genotypes in a manner dependent on the QTL position.
- Genetic model: often assume additivity among different QTL, i.e. the mean of the normal distribution, μ, of a genotype g is:

$$\mu_{g_1 \cdots g_p} = \mu + \sum_{i=1}^{p} \Delta_j g_j \tag{3.37}$$

  where g_j is the allele at the j-th locus.
- Probabilistic model: let y be the trait value, m be the marker genotype, these are the observed data. The unobserved data are: the QTL genotype, denoted as g, and there could be multiple QTLs in the region being tested. The unknowns are the parameters of the genetic model μ and the QTL locations, denoted as γ (specified through recombination rates). The joint distribution:

$$P(y, m, g, \mu, \gamma) = P(\mu)P(y|g, \mu)P(\gamma)P(m, g|\gamma) \tag{3.38}$$

  where $P(y|g, \mu)$ is specified by the genetic model, and $P(m, g|\gamma)$ is specified according to the recombination distances.
- Inference:
  - Sampling QTL genotypes:

$$P(g|y, m) \propto P(g|m)P(y|g) \tag{3.39}$$

    Sampling could be performed with importance sampling: the proposal distribution is $P(g|m)$, and the weighting according to $P(y|g)$.

- Sampling QTL locations and effects: $P(\gamma|y,m)$ and $P(\mu|y,m)$.
- Number of QTLs: determined via Bayesian model selection.

Mapping functional allele series in multiparental populations [Wes Crouse, postdoc interview, 2020]

- Motivation: haplotype based test can be more powerful than single marker analysis because: (1) Multiple causal variants: then the haplotype with multiple causal alleles can have a large effect. This is similar to block-level analysis in GWAS. (2) Epistasis: this would further increase the difference between haplotypes.
- Idea: if we know which haplotypes have similar effects, we should group them, and do a contrast analysis. However, they are not known. We define haplotypes with the same effects as a functional allele. The goal is to find functional allele series (grouping of haplotypes).
- Model: define M as the group structure of haplotypes, then our model is $y = XM\beta + \epsilon$ where X is haplotype and β the effects of functional alleles. They key is to define a prior on M: assume causal mutations occur during the phylogeny of haplotype, and haplotypes with the same mutations belong to the same group. Let T be the local phylogenetic tree (may vary across loci because of recombination), L be the branch length and π be when causal mutation occurs. Use a coalescence tree prior for T, and causal mutations follow Poisson process, then the prior of M marginalize T and L and π. Show that M follows Chinese Restaurant Process.
- Note: the prior depends quite strongly on the rate of causal mutations α. This parameter is chosen to have mostly 0 or 1 mutations.
- Part II. mediation analysis in QTL. Find mediators of trans-expression QTL. Motivation: current mediation analysis requires several tests, which is not statistically satisfactory. Also Sorbet test cannot distinguish partial and complete mediation.
- Model selection approach: define several models, including complete mediation, partial mediation, independent effects. Use a uniform prior for all models.
- Results: some example, eg. trans-pQTL of a gene, mediated by cis-gene.
- Remark: not account for reverse causality. Confounders may be OK: unless they are heritable.

Chapter 4

Association Mapping

4.1 Introduction to Association Mapping

Reference: [A tutorial on statistical methods for population association studies, Balding, NRG, 2006; Genome-wide association studies for complex traits: consensus, uncertainty and challenges, McCarthy & Hirschhorn, NRG, 2008; Handbook of Statistical Genetics, Chapter 36,37]

Concepts of GWAS analysis:

- Indirect association between markers and traits: association between causal variants and genetic markers through LD and association between causal variants and traits.
- Population genetic perspective: in the association studies, the individuals are not really unrelated. In fact, they share a small number of common ancestors, and disease-causing mutations may descend from a small number of ancestral mutations.
- CDCV hypotheiss: Only common variants are genotyped in association studies (e.g. above 5%). The fundamental assumption of GWAS is that the disease traits are influenced by common variants (Common Disease-Common Variants, CDCV). There is no statistical power to detect rare variants in GWAS (e.g. if only 1%, then in a study involving 1,000 people, only 10 people will have the rare allele, not enough to detect association).
- Linkage vs association studies: two main problems with linkage studies (limited to families) for complex diseases: (1) not concentrated within families; (2) the causal variants are not shared among affected family members. Thus for complex disease, association studies are more powerful.

Justification of CDCV: if some ancestral mutation has small effect on fitness, then it may accumulate in the population and reach relatively high frequency. The possible causes of common varirants:

- Many mutations have small effects on fitness due to the robustness of the system, the uncertainty of environment (time-varying selection), etc.
- Human population size is small: thus selection is weak.
- Selection against many late-onset diseases may be small.

Design of association studies:

- Design: family-based association studies: transmission-disequilibrium test (TDT), and population-based case control: GWAS. Multi-stage replication design (at each stage, narrow down candidate SNPs and increase sample size).
- Control selection: should minimize the possible bias in case vs control group.

- Tag SNP selection: because of LD, only a subset of SNPs need to be genotyped (50-70% reduction). How tag SNPs are chosen depends on the LD structure. E.g. use $r^2 = 0.8$ as the cutoff for redundent SNPs that are adjacent.
- Sample size: if n is the sample needed for causal variants, then n/r^2 for indirect association where r^2 is between the marker and causal variant.
- Prospective and retrospective design: prospective design - inidividuals are followed forward in time and disease events are recorded. Most studies are retrospective design.

Possible biases of case-control studies:

- Selection bias: bias of sampling of cases and controls. E.g. cases are obtained nationally while controls are obtained locally.
- Information bias: the measurement errors could be different in cases and controls.
- Stratification or admixture: potentially confounding variables including ethnic groups/population structure and environmental variables.

Preliminary analysis:

- HWE: often used for data control (discase SNPs not in HWE), but may also indiate disease association. Test of deviation: Pearson's χ^2 test.
- Missing genotype data and haplotype reconstruction: possible if LD is strong.
- LD and recombination rate: LD between two loci is commonly measured by D' or r^2. A disadvantage of D' is: sensitive to rare alleles. r^2 gives the sample size needed to detect the disease association (relative to the sample size required if directly type causal SNP). Summarizing LD in a region requires estimation of recombination rates.
- Tag SNPs: could reduced the SNP set in the analysis, thus reduce d.f. of a test.

Diagnostic plots:

- Quantile-quantile (Q-Q) plot: comparison of the null distribution of the test statistic (e.g. χ^2) with the observed distribution. Population stratification will result in deviation from the null across the entire distribution, whereas large-effect disease loci generate deviation at the highly significant end of the range.
- Genome-wide Manhattan plot: test statistic or P value across the genome positions.

Test of association: single SNP

- Two d.f. test: Pearson χ^2 test or Fisher's exact test (preferred with small sample size/rare alleles).
- One d.f. test: assume the risk is additive. Armitage test: test if the ratios of case/control are the same across all three genotypes (test slope of the line). Armitage test sacrifices power if the genotypic risks are far from additive in order to obtain better power for near-additive risks.
- Linear/logistic regression: the linear predictor is equal to β_0, β_1 or β_2. LRT against the null hypothesis $\beta_0 = \beta_1 = \beta_2$ has 2 d.f. (equivalent to Pearson 2-df test with large sample size). If assume the coefficients are linear, i.e. $\beta_1 = (\beta_0 + \beta_2)/2$, a 1-df test that is equivalent to Armitage test.

Test of association: multiple SNPs. This can be used to test, e.g. whether a gene is associated with the disease.

- SNP-based logistic regression: suppose there are L SNPs, treat each of them as a predictor and do regression. LRT: the null hypotheis requires for every SNP, $\beta_0 = \beta_1 = \beta_2$ (test with additivity assumption is similar). Step-wise selection or shrinkage method or tagging SNPs (feature selection) can be applied to deal with redundant SNPs.
- Haplotype-based method: avoid the problem of having too many predictors.
  - Basic methods: test of 2 by k contingency table (k is the number of haplotypes); comparison of the proportion of case and control in k haplotypes; regression anallysis with haplotypes treated as categorical variables.
  - Haplotype clustering: impose a structure on haplotype space to exploit possible evolutionary relationship among haplotypes. Ex. cluster haplotypes that are assumed to share a common ancestry and therefore convey a common disease risk.

Multiple testing:

- Permutation procedure: use simulation to estimate the FP rate.
- FDR: the expected number of false positives among significant associations. Treat actual distribution of P values as a mixture of distributions under null (uniform) and alternative (skewed towards zero) hypothesis. Let N be the number of SNPs, α be the significance level (uncorrected), and k be the number of SNPs with p value less than α, then: $FDR = N\alpha/k$.
- Bayesian: traditional approach discourages additional tests, because all tests will suffer from the multiple-testing penalty. Bayesian approach may have advantages, as the prior probability of association should not be affected by what tests are chosen to carry out.

Epistatis:

- Test SNP pairs: in the logistic regression framewok, allow interaction terms of two SNPs (four terms corresponding to additive/dominance contributions to epistasis).
- Two-stage approach: single SNP test, then choose SNPs above a certain threshold for pair testing.
- Bayesian model averaging.

Strategies for dealing with confounding:

- Stratification: of the confounding variaible and test for association within each strata. To test for association: (1) assume the association parameter is constant across strata and allow the disease distribution varies across strata, and do contingency table test. (2) Logistic regression: allow disease status to depend on the stratum (additional categorical variable).
- Group or individually matched study: may help stratification problem.

Population stratification:

- Genomic control: the idea is that the scores of the SNPs are inflated, so make a correction by dividing a factor λ. The value of λ is estimated by: compute the test statistic (e.g. Armitage test) of null SNPs, and compare the distribution with the expected null distribution. Limitation: only applicabile to the simplest single-SNP test.
- Structured association: (1) infer the population structure - K specific ancestral strata inferred using the STRUCTURE algorithm (MCMC); (2) test for associations conditional on subpopulations. However, the correct number of subpopulations is generally unknown.

- Regression: incorporate null SNPs (or known race/ethnicit) as covariates in regression (these null SNPs will capture the effect of stratification: e.g. one null SNP may be a marker of some subpopulation). Could also use population structure variable (e.g. PCs from PCA) as covariates. This has the benefit of incorporating the effect of different environmental or demographic factors: such as diet. Thus in this sense, the marker SNPs are simply surrogates of these (unmeasured) factors.
- Hierarchical model: estimate kinship, with or without subpopulation effect. The population structure can be diagnosed by PCA.

Replication:

- Evaluate the selected signals with additional independent samples, under the same allele, the same phenotype and genetic model. Replication should be separate from fine-mapping: avoid using additional SNPs (e.g. those adjacent to the selected ones). However, in practices, investigators often combine the two studies.
- In general, replication experiments only genotype the SNPs that are significant (or close) in the first (scan) phase of the study. Ex. see [Barrett et al., Crohn's disease GWAS, NG, 2008]
- Heterogeneity: not all signals can be replicated because of heterogeneity. The possible causes include: the difference in LD, in the allele distribution, non-additive interactions with other genetic variants or environmental exposures, etc. Ex. FTO gene is found to be associated with type 2 diabetes in one study (through association with weight), but not in a replicate study (which controls the weight of samples).

Follow up analysis:

- Candidate genes or regulatory elements: the intervals containing putative causal SNPs are defined in terms of the flanking recombination hot spots. Usually the intervals contain multiple genes, or distant from any known genes.
- Biological knowledge: e.g. expression of a gene is associated with the disease.
- Resequencing and fine mapping: detect the putatively causal variants.
- Application of GWAS results: (1) novel biological insights: which may suggest new biomarkers or therapeutic targets; (2) personalized medicine: which may improve diagnostics/prognostics and treatment (however, the predictive power, even all SNPs combined, is still very low).

Remark:

- Question: the environmental influences in GWAS (similar to population stratification)? E.g. if a disease is strongly associated with a certain diet, and the case and control groups are imbalanced with this diet, then there will be false associations with the SNPs that tend to occur in the people with one type of diet.
- The possible answers: (1) if the environmental influence is known, then model with multi-variate regression, and explain the trait using the environmental variable. (2) Otherwise, depending on whether environmental variable is associated with genetic backgrounds, if there is no association, then there wouln't be many false associations; if there is association, most interestingly, people with one genetic background tend to have similar life style/be subject to the same environmental influence, and tend to have certain diseases or not, then there will be false associations. A population stratification problem.

Genetic mapping in human disease [Altshuler & Lander, Science, 2008]:

- Principles of GWAS:

  - Common disease/common variants assumption (CD/CV): common diseases are caused by genetic variations at common variants (normally defined as allele frequencies $> 1\%$). The alternative hypothesis is: common diseases/rare variants (CD/RV), where common diseases are often caused by variantions at rare variants. The CD/CV assumption is the foundation of GWAS, as generally GWAS only measure genotypes at common variant SNPs.
  - Human haplotype structure (Figure 1): the heterozygosity rate is about 1 in 1000 bases, and about 90% are common variants. The haplotype of SNPs are often highly correlated within regions with low recombination, forming the haplotype blocks. This reflects the fact the evolutionary history of human populations, the recent genetic variations in closely linked loci have not been shuffled by recombination. Therefore, a high level of LD within human genome (this allows ous to genotype only one SNP within a haplotype block). Particulary, the LD between markers (SNPs or structural variants, such as copy number variations, or CNVs) and disease loci means that the information of markers reveal information of the disease loci.
  - Analysis of GWAS: since the physical location information is not quantitatively known (when the founder mutation occurs, etc.), methods such as interval mapping cannot be applied. Instead, testing association between each SNP and the trait.

- Statistical power of GWAS:
  - Problem of population structure: if the case group and the control group have different population structure, e.g. the case group is enriched with African people, then any SNPs typical of the African population will show significant associations. The idea is that: population structure is revealed by the patterns across a large number of SNPs.
  - Because of haplotype structure (the strong LDs among nearby SNPs), only 500,000 SNPs (out of milliions of SNPs) provide excellent power to test $> 90\%$ of common SNP variations.
  - Sample size requirement: achieving 90% power to detect an allele with 20% frequency and a factor of 1.2 (increase the disease risk by 20%) at a statistical signifance of 10^{-8} requires 8,600 samples. The sample size of the current GWAS is usually smaller (thousands).
  - Pooling data could yield higher power.

- Biological lessons from GWAS:
  - In the vast majority of cases, the estimated effects are small, a factor of 1.1 to 1.5 per associated allele. And the total variants explain only a small fraction of inherited risk of disease, e.g. 10% for Crohn's disease.
  - Many associations implicate non-protein coding regions, e.g. the regions at 8q24 associated with cancer, 300kb from the nearest gene.
  - Often multiple disease are associated with the same regions.

- Future directions:
  - Increase the statistic power by: increasing sample size; the structural variants. The goal is to identify rare variants (this may be particularly important for disease).
  - Identify causal mutations and disease mechanism: GWAS typically yield regions of 10 to 100 kb. E.g. through resequencing or fine-mapping. And further studies of the disease mechanism (creating disease model in human cells or non-human animals).
  - Gene-gene interactions ang gene-environment interactions.
  - Disease-related intermediate traits: e.g. gene expression.
  - Genome sequencing of a large sample.

Statistical Challenges in GWAS [Cantor & Sinsheimer, AJHG, 2010; Moore & Williams, Bioinfo, 2010]:

- Meta-analysis:
  - Types of meta-analysis: (1) combine studies in two diverse populations; (2) combine studies that differ in covariates or in measure of traits (e.g. binary and continuous).
  - Considerations of meta-analysis/heterogenity: ascertainment of the sample (including ethnic stratification), the definition/measure of the trait, the statistics that summarize the association results.
  - Methods:
    * SNP imputation
    * Traditional approach: combine p values, or other test statistics. Ned weighting to reduce the effect of heterogenity: fixed-effect model (effect size is constant) and random-effect model (effect size also varies).
    * Bayesian hierarchical models: parameters as random variables sampled from prior distributions; also incorporate the LD data and other evidence (biological functional information) as prior information.
- Epistatis:
  - Difficulty: the large number of possible interactions. May need to restrict the search: e.g. only those with significant marginal effects. Involve a trade-off between miniminizing computation and the number of tests, and maximizing power.
  - Parametric methods: test departure from the additive model (between two loci). Penalized regression is one powerful way of selecting a small number of significant interactions.
  - Non-parametric methods: these methods directly or indirectly model interactions, including decision tree, multifactor dimensionality reduction (MDR), combinatorial partitioning, entropy/conditional entropy measures, logic regression, Bayesian partitioning, etc.
    * Random forest: the standard implementation is conditional on marginal effects. The interpretation is not straightforward as the interactions are hidden in the tree.
    * MDR: designed to detect interaction even in the absence of strong marginal effect. The idea is to construct features by pooling genotypes from mulitple SNPs, and these features will make it easier for methods such as logistic regression to detect attribute dependencies.
    * Relief, RelieF, TuRF: attribute selection method. Assess the weight/quality of an attribute based on whether the nearest neighbor of a randomly selected instance from the same class and the nearest neighbor from the other class have the same or different values.
- GWAS pathway analysis (GWASPA):
  - Motivation: genetic heterogeneity of complex traits - mutations of the same gene, or different genes of the same pathway, can lead to the same disease (especially in different ethnic groups). Thus GWASPA can increase power by combining these evidences.
  - Statistical considerations for GWASPA: sources of biases may include: SNP density in genes; gene size; pathway size. If not correcting for these factors, there will be bias toward finding the pathways that are large, well-known, with more genes that are densely covered by SNPs.
  - Methods/issues of GWASPA:
    * SNP assignment: the most frequent approach is to select a SNP with the strongest association signal for a gene. Not optimal because still bias with large genes, and multiple association signals within a gene may be lost.
    * Score pathways: most method aggregate P values of genes in the pathway in some way such as GSEA, k most significant genes, etc.
    * Permutation test: the score is often assessed by permutation test, which may also adjust for many sources of bias.

– Challenge: SNP assignment - consider multiple independent association signals of a gene; correction of testing multiple pathways; epistasis in genes of a pathway; P-value based test (because genotype data are not always available); capture both rare and common variants.

4.2 Statistical Analysis of Association: Background

Reference: [Laird & Lange, The fundamentals of Modern Statistical Genetics, Chapter 7; Handbook of Statistical Genetics, Chapter 36,37]

Linkage and association: [Thomas, Chapter 9]

- Linkage: between disease locus and marker (co-segregation in families). The consequence of linkage (and co-segregation) is that the affected members in families tend to share markers.
- Association: LD between disease locus and marker.
- Comparison: for linkage analysis, we are testing $H_0 : \theta = \frac{1}{2}$, where θ is the fraction of recombination, and for association analysis, we are testing $H_0 : D' = 0$, where D' is the measure of LD.

Disease risk and genotype frequenices:

- Disease penetrance: for a genotype, the probability that the individual carrying this genotype is disease, denoted as $f_i = P(D|i)$, where $i = 0, 1, 2$ is the number of A alleles (suppose A is the disease allele and a is the normal allele), and D and U denote disease or not.
- Genotype frequency: the probability of a genotype in the population (including both cases and controls), denoted as g_i. If HWE holds, then g_i can be calculated from allele frequencies.
- Relative risk: measure the strength of association. The increase of disease probability relative to some reference genotype, denoted as: $\gamma_i = f_i/f_0$.
- Odds and odds ratio: The odds of a genotype is defined as:

$$\text{odds}(i) = \frac{f_i}{1 - f_i} \tag{4.1}$$

  The odds ratio is defined as the odds relative to some reference genotype (similar to relative risk), e.g.:

$$OR(i) = \frac{f_i}{1 - f_i} / \frac{f_0}{1 - f_0} \tag{4.2}$$

  For most diseases that are rare in population, we have $1 - f_i \approx 1$, thus $OR(i) \approx \gamma_i$, i.e. odds-ratio is approximately equal to relative risk.

- Genotype frequencies in case and control: let $p_i = P(G = i|D)$ be the genotype distribution of case, and $q_i = P(G = i|U)$ be that of the control. We have the following relations using Bayes's Theorem:

$$p_i = P(i|D) = \frac{P(D|i)P(i)}{P(D)} = \frac{f_i g_i}{\sum_i f_i g_i} \tag{4.3}$$

$$q_i = P(i|U) = \frac{P(U|i)P(i)}{P(U)} = \frac{(1 - f_i)g_i}{\sum_i (1 - f_i)g_i} \tag{4.4}$$

  One could define $K = \sum_i f_i g_i$, which is the disease prevalence in the whole population. A consequence of these relations is that the relative genotype frequency of case/control is proportional to the disease odds.

- Hardy-Weinberg equilibrium (HWE): in general not hold for loci associated with diseases. Thus deviation from HWE in cases of diseases is often taken as preliminary evidence of association.

Genetic models:

- Multiplicative penetrance model: the relative risk is multiplicative of the number of copies of the allele A, i.e. $f_i = \gamma_i f_0$ and $\gamma_2 = \gamma_1^2$. Under this model, association with disease does not lead to deviation from HWE.

Testing for association at a single locus: two basic approaches based on different intuitions:

- Disease risk perpsective: genetic variant increases disease risk. This corresponds to discriminant models in machine learning ($X \to Y$).
  - Hypothesis testing based on genotype partitioning: the case/control ratios are different across genotypes. Armitage trend test: test if the ratios are the same across three genotypes (or whether slope is zero vs slope not equal to 1 in regression). It has df equal to 1 (effectively assuming additive risk in H_A).
  - Likelihood method: model the probability distribution of disease conditional on genotypes. Logistic regression: under the null hypothesis, the coefficient of the genotype variant is zero.
  - Bayesian: priors on genetic model (additive, dominant, etc.) and effect size.
- Genotype distribution perspective: disease group has a different genotype distribution than the case group. This corresponds to generative modesl in machine learning ($Y \to X$).
  - Hypothesis testing based on genotype distribution: the genotype distributions are different in case and control groups - χ^2 or Fisher's exact in contingency table.
  - Likelihood method: model the probability distribution of genotypes conditional on case/control group. Multinomial distribution: the coefficients of the genotypes under case are different from those under control.
  - Bayesian: priors on the genotype distribution.
- Analogy: this is similar to text classification problem. The main statistical challenge of much more parameters than observations is essentially the same. And the different methods here also parallle those in text classification: discriminative models such as SVM where response is a function of predictors; and generative models such as Naive Bayes.

Allele test for additive model: [LL, Section 7.2]

- Problem: Consider a 2 by 3 table and let r_i be the number of cases, and s_i be the number of controls, $i = 0, 1, 2$. And the total number of cases is R, the total number of controls is S, and the number of genotype i is n_i. And the total number is $N = R + S$. The null hypothesis is that the number of A alleles is equal in case and in control.
- Idea: in the case group, the number of A alleles follows a binomial distribution with sample size R and in the control group, it follows an independent binomial distribution with sample size S. The problem is thus the test whether the probability parameters of two binomial distributions are identical.
- Test: let p_{case} and p_{control} be the A frequency in the case and control group respectively, we are testing $H_0 : p_{\text{case}} = p_{\text{control}}$. Let $\bar{p}_{\text{case}}$ and $\bar{p}_{\text{control}}$ be the average A frequency in case and control, and our test would be based on $T = \bar{p}_{\text{case}} - \bar{p}_{\text{control}}$. Under H_0, $E(T|H_0) = 0$, and under the HWE assumption (each copy of allele is independent), the variance of $\bar{p}_{\text{case}}$ and $\bar{p}_{\text{control}}$ is the variance at each copy of allele divided by sample size. So:

$$\text{Var}(T) = \bar{p}(1 - \bar{p}) \left(\frac{1}{2R} + \frac{1}{2S} \right) \tag{4.5}$$

where $\bar{p} = (2n_2 + n_1)/N$ is the average frequency of A. Thus we have the test: $Z = T/\sqrt{\text{Var}(T)}$ following standard normal distribution, or $X^2 = T^2/\text{Var}(T)$ following χ^2 with df equal to 1.

Armitage trend test: [Armitage test derivation, statgen.org; LL, section 7.2]

- The null hypothesis, the trends of cell counts in the case and in the control are the same. Or more precisely, if we assume $\bar{X}_{\text{case}}$ and $\bar{X}_{\text{control}}$ be the average number of A alleles per individual, then under H_0, the two should be equal.
- Idea: the numbers of three genotypes in the case group follow a multinomial distribution $\text{Mul}(R; p_0, p_1, p_2)$; and similarly, they follow an independent distribution $\text{Mul}(S; q_0, q_1, q_2)$. We are thus testing if $p_i = q_i, i = 0, 1, 2$.
- Test: we first standarize the table s.t. the number of cases and controls are equal (RS). Our test statistic measures the difference of the number of A alleles in cases and in controls (weighted by the sample size of case and control):

$$U = \sum_{i=0}^{2} x_i(Sr_i - Rs_i) \tag{4.6}$$

  Under $x_i = i, i = 0, 1, 2$, the test statistic becomes:

$$U = N(r_1 + 2r_2) - R(n_1 + 2n_2) \tag{4.7}$$

  Clearly, under H_0: $E(U) = 0$. The variance of U can be computed by: (1) only r_i and s_i are variables and the expression contains only their variance and covariance; (2) the variance and covirance of r_i and s_i under H_0 can be computed using properties of multinomial distribution: $(r_0, r_1, r_2)|H_0 \sim \text{Mul}(R; n_0/N, n_1/N, n_2/N)$ and $(s_0, s_1, s_2)|H_0 \sim \text{Mul}(S; n_0/N, n_1/N, n_2/N)$.

$$\text{Var}(U) = \frac{(N-R)R}{N}\left[N(n_1 + 4n_2) - (n_1 + 2n_2)^2\right] \tag{4.8}$$

  At large N, we form the χ^2 test statistic (or normal):

$$X^2 = \frac{N\left[N(r_1 + 2r_2) - R(n_1 + 2n_2)\right]^2}{(N-R)R\left[N(n_1 + 4n_2) - (n_1 + 2n_2)^2\right]} \sim \chi_1^2 \tag{4.9}$$

- An alternative way of deriving the test [LL, section 7.2]: we have $\bar{X}_{\text{case}} = (2r_2 + r_1)/R$ and $\bar{X}_{\text{control}} = (2s_2 + s_1)/S$. Under H_0, the difference of the two is 0, and the variance can be computed as:

$$\text{Var}(\bar{X}_{\text{case}} - \bar{X}_{\text{control}}) = \text{Var}(X)\left(\frac{1}{R} + \frac{1}{S}\right) \tag{4.10}$$

  The variance of X (the number of A allele) under H_0 can be easily calculated using that $X \sim \text{Mul}(N; n_0/N, n_1/N, n_2/N)$. This leads to the same test.

Estimation of effect size: [LL, section 7.5]

- Notation: Suppose we are given a genotype and a reference genotype, denoted as E (exposed) and U (unexposed) respectively. The numbers of cases in E and U groups are a and b respectively, the numbers of controls in E and U are c and d respectively.
- The estimated odds ratio is given by:

$$\text{OR} = \frac{a/c}{b/d} = \frac{ad}{bc} \tag{4.11}$$

  The variance of the $\log \text{OR}$ is (see the section "Estimating odds-ratio in a 2 by 2 table" in Statistics Notes):

$$\text{Var}(\log \text{OR}) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{4.12}$$

- Remark: it is possible to test association using effect size (check if 0 is in the confidence interval) - this test is not equivalent to the trend test (which is based on the difference of risk, not ratio); in practice, usually do association test first, and if significant, calculate the effect size.

Logistic regression:

- Model: let G be genotype, y be phenotype, and π_i be the probability of disease given G_i:

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \tag{4.13}$$

  where η_i is the linear predictor dependent on G_i. The log-likelihood is then given by:

$$\ln f(\mathbf{y}|\mathbf{G}, \beta) = \sum_i [y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i)] \tag{4.14}$$

  Typically, the linear predictor encodes additive and dominance effect of two alleles:

$$\eta_i = \beta_0 + \beta_A z_{(A)i} + \beta_D z_{(D)i} \tag{4.15}$$

  where β_A denotes the additive effect of the allele m (minor allele), β_D the dominance effect, $z_{(A)i}$ is an indicator variable representing the additive component of i-th genotype (the number of copies of the minor allele), and $z_{(D)i}$ the dominance component of G_i (only present for heterozygous genotype). Note: the model can be easily extended to incorporate other (e.g. env.) variables simply by adding these in the linear predictor.

- Significance: β in the logistic regression model is the log-OR, and the significance of β can be tested using Z test or LRT (see section of Logistic Regression).

- Genetic models:
  - Multiplicative (additive) model: the disease odds are multiplicative for the a allele, thus we have: $\text{odds}(Aa) = \alpha(1 + \theta)$ and $\text{odds}(aa) = \alpha(1 + \theta)^2$.
  - Dominant model: $\text{odds}(AA) = \alpha$, $\text{odds}(Aa) = \text{odds}(aa) = \alpha(1 + \theta)$.
  - Recessive model: $\text{odds}(AA) = \text{odds}(Aa) = \alpha$, $\text{odds}(aa) = \alpha(1 + \theta)$.

Background for Summary Statistics [personal notes]:

- Relation of effect size, Z-score and PVE under simple regression: $y = x\beta + \epsilon$, let N be sample size, σ_x, σ_y be the variance of x and y. The variance of $\hat{\beta}$ is given by:

$$s^2 = \frac{\sigma_y^2}{N\sigma_x^2} \Rightarrow s = \frac{\sigma_y}{\sqrt{N}\sigma_x} \tag{4.16}$$

  The PVE of x is given by:

$$\text{PVE} = \frac{\hat{\beta}^2\sigma_x^2}{\sigma_y^2} \Rightarrow \sqrt{\text{PVE}} = \frac{\hat{\beta}\sigma_x}{\sigma_y} \tag{4.17}$$

  When both x and Y are normalized, then PVE is simply the square of $\hat{\beta}$. The Z-score of $\hat{\beta}$ is given by:

$$Z = \frac{\hat{\beta}}{s} = \frac{\sqrt{N}\hat{\beta}\sigma_x}{\sigma_y} \Rightarrow Z = \sqrt{N} \cdot \sqrt{\text{PVE}} \tag{4.18}$$

  Thus Z score is simply $\sqrt{N}$ times the square root of PVE - this is true regardless of when x and y are normalized.

- Z-score is normalized: it is easy to check that $\text{Var}(Z) = 1$.

4.2.1 Power of Association Studies

Power analysis: [Neale, Statistical Genetics, Chapter 21]

- Armitage trend test [Slager & Schaid, Hum Hered, 2001]: suppose p_i and q_i are genotype frequencies of cases and controls respectively. We know that under H_A, $r_0, r_1, r_2 \sim \text{Mul}(R; p_0, p_1, p_2)$ and $s_0, s_1, s_2 \sim \text{Mul}(S; q_0, q_1, q_2)$. Our statistic is:

$$U = \sum_i x_i \left(\frac{S}{N} r_i - \frac{R}{N} s_i \right) \tag{4.19}$$

  U follows normal distribution asymptotically. Taking expectation of U (using the expectation of r_i and s_i from multinomial distribution):

$$E(U) = \mu_1 = N\phi(1-\phi) \sum_i x_i (p_i - q_i) \tag{4.20}$$

  where $\phi = R/N$. The variance of U (again using the properties of multinomial distribution):

$$\text{Var}(U) = \sigma_1^2 = N\phi(1-\phi)^2 \left[\sum_i x_i^2 p_i - (\sum_i x_i p_i)^2 \right] + N\phi^2(1-\phi) \left[\sum_i x_i^2 q_i - (\sum_i x_i q_i)^2 \right] \tag{4.21}$$

  The terms p_i and q_i can be related to the genetic parameters, according to Equation 4.3.

Calculation of power in case-control studies:

- Distribution of test statistic: to calculate the power, need to compute the distribution of test statistic, χ^2 in case-control studies, under H_A. This is equivalent to finding the distribution of the counts of two alleles in cases and controls respectively. The distribution of the χ^2 statistic can be analytically computed under the simple additive model with only a single causal SNP. When the numbers of cases and controls are equal, this is given by [Spencer & Marchini, PG, 2009]:

$$E(\chi^2) \propto N\gamma^2 p(1-p) r^2 \tag{4.22}$$

  where N is the number of cases and controls, γ the effect size, p the allele frequency of the risk variant and r^2 the correlation between the marker and the causal SNP.

Factors affecting the power of case-control studies: power is assessed by the sample size required for a given significance and power level. See [Neale, Table 21.1-21.3]

- Genetic model, risk allele frequencies and genotype relative risk: these factors affect the genotype frequencies $P(G|D)$. Rare alleles require considerably more samples. The recessive models typically require many more samples.

- TDT vs. case-control: about the same number of TDT trios and case-control pairs. So the sample size requirement of case-control is 2/3 of TDT. The main advantage of TDT is the control of population stratification.

- D' and marker allele frequency: in the calculation of power, replace the genotype frequency of disease alleles with those of marker allele frequencies (by summing over the disease alleles, whose probabilities are related to marker allele frequencies through LD). Large D' increases the sample size requirement; marker allele frequencies are optimal when they match the causal allele frequencies.

- Quantitative traits: discretization leads to a substantial loss of power.

Simulating genotypes for power analysis: HAPGEN [Spencer & Marchini, PG, 2009]:

- Generative model of LD: suppose we want to model n haplotypes $h_1, \cdots, h_n$, where the haplotype is size S. The idea is that given k haplotypes, the next one can be viewed as a mosaic of existing haplotypes. We have:

$$P(h_1, \cdots, h_n|\rho) = P(h_1|\rho)P(h_2|h_1, \rho) \cdots P(h_n|h_1, \cdots, h_{n-1}, \rho) \tag{4.23}$$

  where ρ represents the recombination parameters of the region. When we have k haplotypes, in generating h_{k+1}: we have a hidden variable X_j for the j-th position, indicating where it comes from, $X_j \in \{1, 2, \cdots, k\}$. X_j is modeled as a HMM, where the transition probability depends on ρ. Also, the mutation process is modeled in the term $P(h_{k+1}|h_1, \cdots, h_k, X)$.

- Simulation procedure:
  - Sample a locate in the region as the disease locus.
  - For a given disease model, and effect size, sample the genotype at the locus according to the disease status. If case, the genotype frequencies can be calculated from the risk allele frequency in the population (control) and the effect size.
  - Starting at the disease locus, sample the rest of the two haplotypes: first sample X_d for the disease locus, the probability should be high for the haplotypes with the same genotype at d; next sample X_j for other positions using the HMM described above; Then the haplotype is sampled according to X_j for all j's, with mutations allowed.

- Power analysis: suppose we want to evaluate a method, or a SNP set, we simulate the data: at each simulation, chooses a disease locus, and set the effect size (and the disease model), then simulate the genotypes according the procedure above, and apply the test method on the simulated data. The procedure is repeated many times, with different disease loci. Finally, to obtain power for a given threshold of the test statistic, compute the fraction of times when the test statistic (at some locus in the region) exceeds the threshold.

4.2.2 Meta-analysis

Methods for meta-analysis:

- Reference: [LL, The Fundamentals of Modern Statistical Genetics, Chapter 11], [Bakker & Voigh, Human Mol Genetics, 2008; Zeggini & Ioannidis, Pharmacogenomics, 2009], [Evangelos & Ioannidis, NRG, 2013]

- Statistics background: see "Meta-analysis" in Statistics Notes.

- Fisher's method: drawbacks include, direction of effects, weighting of studies, inability to produce a summary effect.

- Z-score based: The most commonly used is the Liptak method, let Z_i be the Z-score of the i-th study (the sign of Z_i should reflect direction of effect),

$$Z_{\text{meta}} = \sum_i w_i Z_i \tag{4.24}$$

  where the weights should satisfy $\sum_i w_i^2 = 1$. It can be proved (using sum of normal random variables) that if $Z_i \sim N(0, 1)$, then $Z_{\text{meta}} \sim N(0, 1)$. w_i is often chosen to be $\sqrt{\frac{N_i}{N_{\text{total}}}}$. This is also equivalent to: inverse of the standard error of the regression coefficient. From linear model, we know that $\text{Var}\,(\hat{\beta})$ is proportional to inverse of total variance in predictor, which is $2N_i p(1-p)$ for SNPs. So the standard error is proprotional to $1/\sqrt{N_i}$.

  - Correction for sample imbalance: [Bakker & Voigh] the effective sample size (assume case and control are balanced) is determined by power analysis: should be equivalent to the actual sample (imbalanced) in terms of power.
  - Effective sample size formula: for any individual study, it is given by [Willer et al, METAL, Bioinformatics, 2010]

$$N_{\text{eff}} = \frac{4}{\frac{1}{N_{\text{case}}} + \frac{1}{N_{\text{control}}}} \tag{4.25}$$

- Fixed Effect model: inverse variance method, where the effect size is: log-OR for binary trait, and β (regression coefficient) for quantitative trait. Advantage over p-value or Z-value based methods: maximize power. Most existing meta-studies (70%) use the fixed effects model.
- Random-effect models: model the between-study heterogeneity: the effect in each study is a RV from a common distribution. Formal Bayesian methods have been developed for random-effect models.
  - Often not used in discovery effects because it loses power.
  - More appropriate in studies with expected heterogeneity, e.g. across different populations.
- Representation of meta-analysis results: forest plot, drawing the effect of each study (with confidence interval), and the summary effect.
- Diagnosis of meta-analysis: the goal is to understand the heterogeneity of results: e.g. whether a large effect in one study is an outlier or a true effect. Techinques: sensitivity analysis - how sensitive the results are to a single study; meta-regression: explain the observed between-study heterogenity using additional covariates; etc.
- Test of between-study heterogenity: Crochran's Q statistic, and I^2. However, both are underpowered at < 20 studies.
- Ranking the results in GWAS meta-analysis by effect size: while conceptually it is the right thing to do, in practice, the effect size is often so modest that this is difficult to.
- Examples:
  - Crohn's disease meta-analysis: [Barrett, NG, 2008; Franke, NG, 2010] compute Z-scores from individual scans, and combine "the scores across all six datasets (inversely weighted by variance)".
  - T2D meta-analysis: [Ziggini, NG, 2008] Liptak method, where effective sample size is calculated by power analysis: choose parameters s.t. the power if 50%, and choose effective sample size that reaches the same power.
  - Remark: the details of meta-analysis are not clear in either cases: (1) CD meta-analysis: inverse variance weighting is usually applied to effect size (OR), not Z-scores; (2) T2D meta-analysis: power of a study depends on parameters such as allele frequencies, effect size, significance level (type I error), not clear how parameters are chosen for "equivalent power".

Practical issues of meta-analysis:

- Data quality issue: e.g when different studies use different platforms.
- Imputation: it is common to perform imputation on a common set of SNPs (e.g. HapMap SNPs) with multiple studies. However, the uncertainty of imputation results should be considered.
- Dealing with inflation/population stratification in meta-analysis. In each individual study: should apply the correction (genomic control or PCA, etc.), before performing meta-analysis. After meta-analysis: may also need correction because of issues such as sample overlap, etc.

- Generally, use the same criteria for SNP filtering: AF $< 1\%$, $P < 10^{-5}$ for HWE, imputatino quality index < 0.3 will be removed [Evangelou].

Criteria of replication test: [LL, Section 11.5]

- Sample size of replication: generally need to be sufficiently large, i.e. at least 80-90% power. Otherwise, high false negative, and dismissal of true findings.
- Genetic model: need to be the same effect, and the same genetic model in the replication test.
- P-value criterion: usually should be nominally significant, i.e. $P < 0.05$.

Between-study heterogeneity:

- Possible causes:
  - Phenotype definition [Evangelos]: may affect the estimated effect sizes. Harmonization is often desirable.
  - Population ancestry: assessment shows that GWAS-discovered variants show modest correlation in MAF between ancestries and different effects in different ancestry. A trans-ethnic meta-analysis that considers the distance between different ethnic groups.
  - Population structure.
  - Gene-gene or gene-environment interactions.
  - Difference of designs of studies: differnet platforms, thus the same causal SNPs may be linked to tagged SNPs to different degrees. The phenotype defintion may be different: e.g. FTO association with T2D (significant association with body mass index, but not if body mass index removed).
  - Sex difference.
  - Winner's curse: earlier studies may suggest stronger effects, and the most significant SNPs are likely to exhibit some regression-to-the-mean upon replication.
  - Other biases or errors: e.g. sample selection (e.g. how control group is chosen, whether disease subjects are rigorously excluded).
- Assessing heterogeneity: e.g Corchran's Q statistic:

$$Q = \sum_i w_i (d_i - d^F)^2 \tag{4.26}$$

  where d_i is the study-specific effect size, d^F is the summary effect size, and w_i is the weight of each study. Note the limitations of each statistic: sensitivity to outliers, to the number of studies, etc.

Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies [Han & Eskin, AJHG, 2011]

- Motivation: in GWAS meta-analysis, it is often true that random-effect model (RE) is not as powerful as fixed-effect model (FE). Confirmed in simulation that even when there is heterogeneity, RE still loses power. Why?
- Problem of traditional random-effect method: traditional RE first estimates the summary effect and its confidence interval, then obtain its p-value. The problem is that when doing this, the model implicitly assumes $H_0 : \mu = 0$ with the same τ^2 as H_1, as confidence interval is based on RE model. The true H_0 should be $\mu = 0, \tau^2 = 0$.

- New RE model: first show that LRT is equivalent to FE model, where we use CI to derive p-value: $S_{FE} = Z_{FE}^2$, where Z_{FE} is the Z-score for the summary effect. Next, for traditional RE, the LRT statistic is equivalent to testing $\mu = 0$ assuming H_0 and H_1 share the same τ^2: $S_{RE} = Z_{RE}^2$. In the new RE model, do LRT, where $H_0 : \mu = 0, \tau = 0$:

$$S_{New} = S_{FE} + S_{Het} \tag{4.27}$$

  where S_{Het} test heterogeneity, and is similar to Cochran's Q. The distribution of S_{FE} is chi-square, and S_{Het} is mixture of 0 and 1-df chi-square.

- Results: when the heterogeneity is low, new RE has similar power to FE; when heterogeneity is high, new RE has higher power. Tradition RE always has lowest power.

4.2.3 Imputation and Haplotype Methods

Reference: [Neale, Statistical Genetics, 2007, Chapter 17, 18]
Genotype imputation [Li & Abecasis, ARGHG, 2010]

- Imputation in related individuals:
  - Intuition: family memebers would share long stretches of haplotypes (several Mb) - IBD regions. So we will need only a small set of genetic markers to cover IBD regions, and if we identify the IBD region, we can impute untyped markers in the regions.
  - Application in linkage analysis: for disease-related loci, regions of IBD would be more similar in phenotypes than other regions.
- Imputation in unrelated individuals: assume reference panel of haplotypes are available (100-200 kb in length, typically 10-20 genotyped markers), then we infer which haplotype generates given SNP regions. If we know the haplotypes, we can infer the untyped SNPs.
- Accuracy of imputation:
  - Procedure of estimating accuracy: masking the typed markers and impute. Comparison of imputed and masked SNPs.
  - Overall error rate (discrepancy): about 1-2%.
  - Measuring accuracy: r^2 between the observed and imputed allele counts. It is directly translated to power calculation: e.g. 1000 imputed samples equivalent to 930 typed samples if $r^2 = 0.93$.
- Application: increasing power and fine-mapping. Ex. in a LDL GWAS.
- Application: meta-analysis of multiple stuides genotyped using different platforms.
- Practical considerations:
  - Non-European samples: for some populations, use specific panel from HapMap, e.g. YRI for west Africa and JPT + CHB for east asian. For other samples, evalaute which reference panel works best (by masking/cross-validatin). Or use multiple reference panels.
  - Determining accurate imputed genotypes: use r^2 measure, the posbility that an imputed genotype is correct - this is not comparable across different MAFs (e.g. if MAF is low, then most of time it will be correct if one just assigns the major allele).
  - Association testing: allele count based association testing.
- Application: low-coverage sequencing data. Use imputation to combine information across individuals who share a haplotype stretch. Ex. 400 sequenced at 2x, using imputation, sites with MAF $> 2\%$ can be genotyped with $> 99\%$ accuracy.

Genotype imputation for genome-wide association studies [Marchini & Howie, NRG, 2010]

- Intuitions of imputation: each sample is a mosaic of multiple reference haplotypes. The challenge is to infer the source haplotype of every study sample. These source haplotypes should have these properties:
  - They are more likely to be from frequent haplotypes in the reference panel.
  - There should be higher level of mosaicism in high recombination regions.
  - The source haplotypes should be similar to the sample genotypes.

  These intuitions can be encoded by HMM model.

- IMPUTE v1: let G be the genotype data and H be the reference haplotypes. Each haplotype region in the study sample can be viewed as mosaic of the haplotypes in the reference sample. Let Z_i be the source haplotype of study sample i, θ be the mutation rate parameter and ρ the recombination rate parameters, we have:
$$P(G_i|H,\theta,\rho) = \sum_{Z_i} P(G_i|Z_i,\theta)P(Z_i|H,\rho) \tag{4.28}$$
  The term $P(G_i|Z_i,\theta)$ encodes the intuition that the genotypes should be similar to reference haplotypes (few mutations), and the term $P(Z_i|H,\rho)$ encodes the intuition about haplotype frequencies and recombination. The latter is based on a HMM. In inference, θ is fixed and the program estimates ρ (allow variation across regions). The parameter N_e must be specified by the user.

- IMPUTE v2: in v1, the haplotype phasing in the study sample is intetraged out(Z term). In v2, the program tries to leverage study samples in addition to reference samples, so it estimates haplotype phases explicitly: it alternates between phasing (of both study and reference samples) and imputation given phases.

- FastPhase: the idea is that at large sample size, there are many possible haplotypes, and that makes the imputation programs slow (quadratic of number of haplotypes). So assume haplotypes can form a small number of clusters. And do similar HMM for the haplotype clusters. Essentially, the model has two parts: first how genotype relates to the haplotype cluster; second how observed haplotypes relate to the clusters. Let α denotes the weights of clusters at each site l (similar to haplotype frequencies):
$$P(G_i|\alpha,\theta,r) = \sum_{Z} P(G_i|Z,\theta)P(Z|\alpha,r) \tag{4.29}$$
  where r is recombination rate. And
$$P(G,H|\alpha,\theta,r) = \prod P(G|\alpha,\theta,r)P(H|\alpha,\theta,r) \tag{4.30}$$

- Factors affecting imputation accuracy: error rates about 5%. Error rates depend on MAF (lower frequency higher error). Reference panel important: how close they are to the study sample.

- Association testing using imputed data:
  - Frequentist approach: use expected allele count in regression model. Score test: likelihood method that marginalize the missing data.
  - Bayesian approach: let D be the data, then we marginalize the imputed genotypes. Let $p_{ijk} = P(G_{ij} = k)$ be the imputation probability of the genotype of sample i at site j. Then we have model evidence:
$$P(D|M) = \int \left[\prod_i \sum_k P(D|G_{ij=k},\theta)p_{ijk} \right] P(\theta|M)d\theta \tag{4.31}$$

- Joint imputation and testing: the problem of separate imputation and testing is that the effect size will be underestimated as the data are imputed under the null model. However, the improved performance from using study samples in imputation outweights the advantage of joint association and testing.

Bayesian methods for imputation: the idea is to use the probability distribution $P(H)$, where H are the hidden haplotypes, to encode the criterion of fewer mutational events among haplotypes [Stephens & Donnelly, AJHG, 2001].

- Inference: Gibbs sampling, sample haplotype of one individual given haplotypes of all other individuals. The algorithm consists of repeated steps of sampling H_i from $P(H_i|G, H_{-i})$, where H_{-i} stands for the haplotypes of all other individuals in the sample:

$$P(H_i|G, H_{-i}) \propto P(H_i|H_{-i}) \propto \pi(h_{i1}|H_{-i})\pi(h_{i2}|H_{-i}, h_{i1}) \tag{4.32}$$

- Conditional distribution $\pi(h|H)$: the sipmler form would be multinomial distribution. A better form is to use coalescent modeling: the haplotypes should be clustered, i.e. only a few haplotypes can generate all observed ones; or the new haplotype h should be the same or similar to some of the existing haplotype in H. Formally, the probability is computed from sampling of an ancestral haplotype, α, and applying the mutational events (perhaps both substituation and recombination events):

$$\pi(h|H) = \sum_{\alpha} P(\alpha)P(\alpha \to h) = \sum_{\alpha}\sum_{s} P(\alpha)P(s)P(\alpha \to h|s \text{ events}) \tag{4.33}$$

  where s is the number of mutations from α to h (follow geometric distribution). The first term is r_α/r, where r is the sample size and r_α the number of haplotypes of type α; the second term encodes the geometric distribution of number of mutations:

$$P(s) = \left(\frac{\theta}{r+\theta}\right)^s \frac{r}{r+\theta} \tag{4.34}$$

  where θ is the mutation rate. And the last term is the αh entry of the matrix P^s, where P is the mutation matrix.

- Other ideas:
  - Partition ligation (PC): for large region, first do haplotype estimation in short segments (e.g. 10 bp), then ligate the segments.
  - Parental information: can be used to improve the inference, in particular, for heterozygote children, if the parent is homozygotes at multiple markers, the phasing may be determined.

- Remark:
  - Generally highly accurate: the best method is PHASE v2.1: less than 6% of errors of unrelated individuals, and 0.2% for trio (parents) data. Other methods are slightly worse.
  - Limitations: not directly model the coalescence. E.g. for the PHASE method, the coalesence is treated at the level of prior.
  - Idea: use HMM to learn the block structure in the data. Ex. one state per marker (one state for each allele), the transition probabilities among the states would suggest which markers form haplotypes.

Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data [Li & Stephens, Genetics, 2003]

- Motivation: block structure is not often obvious from pairwise LD. Need a model that uses all the data to study recombination.

- Problem: suppose we have a set of haplotype data, $h_1, \cdots, h_n$, how can we estimate the recombination rates?
- Intuitions of the model (Figure 2): we need to specify the conditional distribution $P(h_k|h_1, \cdots, h_{k-1})$. Intuitively, h_k should be mosaic of haplotypes h_1 to h_{k-1}. This distribution should have these properties:
  - The next haplotype is more likely to match a haplotypes that has already been observed many times.
  - The probability of seeing a new haplotype decreases as k increases.
  - The probablilit of seeing a new haplotype increases as θ (recombination) increases.
  - The haplotype h_k tend to differ from existing haplotypes by only a few mutations.
  - The next haplotype should be somewhat similar to existing haplotypes.
- Model: the total probability is $P(h_1, \cdots, h_n) = P(h_1)P(h_2|h_1) \cdots P(h_n|h_1, \cdots h_{n-1})$. To obtain the conditional distribution $P(h_k|h_1, \cdots, h_{k-1})$, assume that h_k is a mosaic of existing haplotypes, and let X be the variables of which haplotypes for h_k. We use a HMM for X. Specifically, the transition probability at position $j+1$:

$$P(X_{j+1} = x'|X_j = x) = (1 - \exp(-\rho_j d_j/k))/k \tag{4.35}$$

  when $x \neq x'$. The rate of total recombination at position j is: $\rho_j d_j$, where ρ_j is the recomibnation rate ($4Nc$ where c is recombination rate) and d_j the distance. But the more existing haplotypes we have, the less likely we will need recombination (i.e. the more likely h_k is an exact copy of some existing haplotypes), so the actual rate is $\rho_j d_j/k$. Next if we do have a recombination (choose a different haplotype), each of the k haplotypes has equal probability. Similarly, we can derive the transition probability when $x' = x$ (skipped).
- Remark:
  - The model depends on the order of haplotypes in inference. In practice, try multiple orders and averaging.
  - See Discussion in Appendix A about the choice/justification of ρ/k in the transition rate.

A new multipoint method for genome-wide association studies by imputation of genotypes [Marchini & Donnelly, NG, 2007]:

- Idea:
  - The LD patterns would allow one to infer the recombination history in the interested SNP, meanwhile, the genotype can be inferred if the recombination events are known: if no recombination, the SNP is determined from the haplotype; if one recombination, from the other haplotype (if there are only two); etc.
  - Approximating population genetic process: to infer the complete history in the genealogy is difficult. For each individual, the SNP depends on its immediate neighbor and the probability that a recombination event occurs between the two SNPs. This can be modeled as a simple HMM.
- Model: N haplotypes (from HapMap), and K individuals with G_i the genotype of i-th individual: $G_{il} \in \{0, 1, 2, \text{missing}\}$ for the l-th SNP. Need to infer missing genotypes G_M from the observed genotypes G_O: $P(G_O|G_M, H)$. Assume that: (1) G_i is sampled independently. (2) The probability $P(G_i|H)$ is given by a HMM. Specifically, given G_{il} at l-th locus, sample $G_{i,l+1}$ according to whether recombination occurs: if no recombination, from the current haplotype; if recombination, sample with equal probability from any of the N haplotypes. The hidden state at a locus in HMM is thus the haplotype where the locus is sampled from. Suppose $Z_{il}^{(1)}$ and $Z_{il}^{(2)}$ are the two haplotypes of the i-th individual at l-th locus, we have:

- Transition probability: $P((Z_{il}^{(1)}, Z_{il}^{(2)}) \to (Z_{i,l+1}^{(1)}, Z_{i,l+1}^{(2)}))$ is given by the recombination probability. The probability of the same haplotype is equal to the probability of no recombination plus the probability of recombination with the same haplotype. The former is given by: $\exp(-\rho_l/N)$, where $\rho_l = 4N_e r_l$ (r_l is the genetic distance per generation). This could be understood intuitively as per generation rate times number of generations (about $2N_e/N$: generating $2N_e$ copies from N haplotypes). The later is: $(1 + \exp(-\rho_l/N))/N$.
- Emission probability: $P(G_i|Z_i^{(1)}, Z_i^{(2)}, H)$ mimics the effect of mutation. Estimate the probability of mutation using an approximation to population genetics model.
- Remark: since the HMM needs to track two states (diploid) per step, thus the complexity is $O(n^2)$, where n is the number of haplotypes.

- Testing association within a region: suppose there are W SNPs in a region, we test the hypothesis: M_0 - no assication; M_1 - some SNP is associated with the disease. Let $P(S_i)$ be the probability that S_i is associated given M_1 is true, then we have:

$$BF_{region} = \frac{P(D|M_1)}{P(D|M_0)} = \frac{\sum_i P(S_i)P(D|M_1, S_i)}{P(D|M_0)} = \sum_i P(S_i)BF(S_i) \tag{4.36}$$

  Thus BF_{region} is a weighted average of BF of all SNPs. An alternative method is simply the maximum of BF in the region.

- Simulation: for each of the SNPs in the ENCODE region (HapMap), create simulated case-control data assuming it is a causal SNP. Only show a subset of SNPs to any program (Affymetrix SNPs). The programs are assessed by the power at different type I error rate.

- Results:
  - Data: WTCCC, using 10-Mb window for imputation.
  - In simulation: imputation significant increases power for rare non-tagged SNPs, as well as common non-tagged SNPs (with smaller effect). Imputation makes small difference in the tagged SNPs.

- Remark:
  - Haploptypes: the model assumes all N haplotypes are equally frequent for computation of transition probabilities. This may be invalid.
  - Testing association in a region: under M_1, if one SNP is associated, then its nearby SNP will show a pattern different from M_0 because LD, so not independent. This may create bias: e.g. a weak causal SNP in a region with high SNP density and large LD may create signals in multiple nearby SNPs, thus the average BF is high in this region.

A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies (IMPUTE v2) [Howie & Marchini, PLG, 2009]

- Motivation: the performance of imputation algorithms largely depend on the haplotypes. Existing methods often fix a set of haplotypes from reference sample. The idea here is to use the study sample as well to reconstruct better haplotypes.

- Method: let G_i be the genotype of i, and H_i be the haplotype indicators (state path in HMM). The idea is to sample the HMM paths, instead of integrating them out (i.e. phasing the study samples). This way, we'll be able to learn more haplotypes from combined reference and study samples. We alternate two steps:
  - Step 1: sample H_i from: $P(H_i|G_i, H_{(-i)}, H_R)$ where $H_{(-i)}$ are all haplotypes in study samples except i and H_R are reference haplotypes. This step needs sampling diploid states, so time complexity $O(n^2)$.

  - Step 2: given the haplotypes, we need to impute the missing genotypes. At this step, we only deal with haplotypes, and for the missing genotypes, we need to evaluate the probability of data given each different possible genotype, so we use forward-backward algorithm with running time $O(n)$.

- Multiple reference samples: the typed SNPs may be different, and this will be takend into account. Also the reference samples may be unphased (diploid).

MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes [Li & Abecasis, GE, 2010]

- Model: instead of fixing a set of haplotypes, the method infers the haplotypes from study samples. Specifically, let G be gentoype data and S be the mosaic states underlying the unphased genotype (two values, one for each chromosome), then we have:

$$P(G, S) = P(S_1) \prod_j P(S_j|S_{j-1}) \prod_j P(G_j|S_j) \tag{4.37}$$

  The model then update S of each individual at each step, using the current set of haplotype estimates for all individuals as template using Li-Stephens model.

Motivations for haplotype-based methods for association mapping:

- At single locus level, the association of SNP with the disease may be very weak. The association can be improved by using haplotypes [Zaykin et al, HTR].

- If group subjects by haplotypes (i.b.d.), then these subjects tend to be genetically homogeneous, and thus it would be easier to locate genetic factors (comparison within a group: better control).

- Regional test: it may be desirable to test on the level of a region encompasing multiple SNPs, for taking advantage of the information in multiple SNPs, or for testing at the gene level. In these tests, haplotypes may not be explicitly modeled, instead, the LD structure is taken into account while modeling/weighting multiple markers [Mingyao Li & Chun Li, ATOM, Bioinfo, 2008].

Haplotype-based analysis:

- Method: the linear predictor in logistic regression is a function of haplotype:

$$\eta_i = \beta_0 + \beta_{H_{i1}} + \beta_{H_{i2}} \tag{4.38}$$

  where H_{i1} and H_{i2} are the two haptoypes of the i-th individual. Typically these methods assume that the risks from two haplotypes are multiplicative. Since the haplotypes not observed, they are either resolved, or integrated over by treating as hidden variables (distribution determined from the genotype G_i). The likelihood can be written as:

$$f(y|G, \beta, H) = \prod_i \sum_{H_i} f(y_i|H_i, x_i, \beta) f(H_i|G_i, H) \tag{4.39}$$

  where G is phenotype data, H are haplotypes (unobserved), H_i is the haplotype of the i-th individual. Also, the additional covariates can be easily incorporated into regression. The standard LRT can then be applied (both the parameters and the hidden haplotypes are maximized).

- Tests: commonly test the null hypothesis that the regression coefficients of all haplotypes equal to zero (omnibus test), or the coefficient of a particular haplotype is 0 (haplotype-specific test).

- Two-stage strategies: first estimate the haplotype frequencites using both case and control data, and then estimate other parameters where the haplotype frequencies are fixed.

- Ascertainment bias: the affected individuals are overrepresented in the case-control samples, thus the estimation of high-risk haplotypes will be inflated. One could use only data from the case group, but information on rare high-risk haplotypes is missed. [Stram & Thomas, Hum Hered, 2003] proposed a correction of the likelihood function.

Haplotype clustering:

- Idea: group similar haplotypes according to some similarity measure, and assign the same risk to all haplotpes within the same cluster. Essentially an approximation to the population genetic process.
- Allowing dominance effects: the underlying causal variants may have non-multiplicative effects. [Morris AP, AJHG, 2006] models the presence of causal variant as a hidden variable in the haplotype. All haplotypes in a cluster have the same probability of carrying a causal variant.

Issues of haplotype-based methods: the performance of these methods vary with the strength of LD, the haplotype diversity (e.g. the fraction of common haplotypes), etc.

- Haplotype diversity: sample size is effectively reduced in the partition based on haplotypes. In particular, the rare haplotypes increase the d.f. of the test, thus reducing the power.
- Comparison of single-locus and haplotype methods: single-locus can be as powerful or more than multi-locus test when permutation test is used to control for LD [Roeder & Devlin, GE, 2005]
- Remark: haplotype based methods are similar to regression on multiple markers, allowing for marker interactions.

Haplotype pattern mining (HPM) [Toivonen & Kere, AJHG, 2000]

- Haplotype patterns: to avoid the problem of having too many haplotypes, define haplotype patterns allowing wildcard symbols, e.g. $(*, 2, 5, *, 3, *, *, *, *, *)$. Then a haplotype, $(4, 2, 5, 1, 3, 2, 6, 4, 5, 3)$ would match this pattern. For any pattern, its association with the trait can be defined using the similar metric, e.g. χ^2.
- The search of best patterns can be done using e.g. *apriori* algorithm.

Haplotype trend regression (HTR) [Zaykin & Ehm, Human Hered, 2002]

- Motivation: haplotypes may be in higher LD with the causal varirant than the individual markers. Ex. haplotypes of three loci: $A_1B_1C_1$, $A_2B_1C_2$, $A_1B_2C_2$ and $A_2B_2C_1$, suppose A is the causal locus, clearly, A is in LE with B and C, but is in LD with the BC haplotype.
- Methods:
  - Simple test: if the haplotypes are certain, then we have $2 \times L$ table, where L is the number of haplotype, and can use the standard χ^2 test on the table.
  - HTR: better to incorporate the uncertainty of the haplotype inference. Thus in logistic regression (genotype to trait), instead of integer count (of the number of minor alleles), use fractional count, which is the expected number under the posterior distribution.

Haplotype clustering allowing dominance effects: [Morris, AJHG, 2006]

- Intuition: similar haplotypes are clustered; furthermore, a cluster may be associated with the causal variant (thus disease risk) and the genetic model of the (hidden) causal variant can be implemented for haplotypes. Let C_k be the k-th cluster, $Z_i^{(a)}$ be the a allele (first or second) of the causal polymorphism of the i-th individual, then the association of a cluster to the causal variant is defined by: $\phi_k = P(Z_i^{(a)} = A|C_k)$, where A is the causal variant.

- Notations: our data consists of the phenotype y, the genotype G, and other covariates x, the hidden variables are: C - clusters 1 to K; H - haplotypes, $H_i = (H_{i1}, H_{i2})$ is the haplotype of the i-th individual; and Z - the causal variant, $Z_i = (Z_{i1}, Z_{i2})$ is the causal variant genotype of the i-th individual. The parameters are: h - haplotype frequencies; θ - model parameters.

- Model: several components:

  - $C_k \to H_i$: haplotype clustering.
  - $H_i \to G_i$: the consitency of G_i and H_i.
  - $C_k \to Z_i$: association of haplotype cluster and causal variants, determined by $\phi_k = P(Z_i^{(a)} = A|C_k)$.
  - $Z_i, x_i \to y_i$: disease risk is a function of the causal variant and covariates. Genetics models are implemented in this distribution (e.g. dominance effects).

  The likelihood function:

$$\begin{array}{rcl} P(y|G,x,h,\theta) & \propto & \prod_i \sum_{H_i} P(y_i|H_i,x_i,\theta)P(H_i|G_i,h) \\ & \propto & \prod_i \sum_{H_i} \sum_{Z_i} P(Z_i|H_i,\theta)P(y_i|Z_i,x_i,\theta)P(H_i|G_i,h) \end{array} \tag{4.40}$$

  Note that the term $P(Z_i|H_i,\theta)$ is computed by summing over all possible clusters for H_i.

- Inference: MCMC algorithm to sample $P(\theta|D,M) \propto P(y|G,x,h,\theta)P(\theta|M)$. The key variables of interest are Ψ_j: the probability that the j-th haplotype carries a causal variant (by summing over all possible clusters of haplotype j).

4.3 Family-Based Methods

Reference: [Thomas, Chapter 9], [Laird & Lange]

Using family members as controls:

- Idea: the case-control studies suffer from the problem of population stratification (cryptic relatedness, etc.). If we can use match every case with a family member as control, then the other confounding variables (diet, culture, etc.) will be matched. By Conditional Logistic Regression, the different baseline risks for different families can be ignored with a matched design.

- Design: often use siblings, pseudo-sibs (e.g. TDT) and cousins for controls. Not using other members to control age and sex. Note: pseudo-sibs mean the expected genotypes that are allowed by the parent mating types.

Transmission disequilibrium test (TDT):

- Motivation: TDT is designed to provide the internal control - the background of the same parents. Basically TDT compares the allele frequencies of markers transmitted to affected children: if no association, a particular allele is equally likely to be transmited or not, to the affected children.

- Test: for a particular alleles of a marker, let T be the number of times it is transmitted to an affected child, and R to be the number of times it is not. Under the null hypothesis, we expected $T = R$. We thus define the test statistic:

$$Q_1 = \frac{(T-R)^2}{T+R} \tag{4.41}$$

  With multiple alleles, we use T_i and R_i for the i-th allele respectively, and define the test as:

$$Q_m = \sum_{i=1}^{m} \frac{(T_i - R_i)^2}{T_i + R_i} \tag{4.42}$$

- Likelihood model of TDT: let β be the effect size, and $R_G(\beta)$ for the RR of the genotype G. We have N families (trios), and G_i, Y_i for the genotype and phenotype of the child i and P_i be the parent genotypes. The likelihood:

$$L(\beta) = \prod_i P(G_i|Y_i = 1, P_i) = \prod_i \frac{P(Y_i = 1|G_i)P(G_i|P_i)}{\sum_{G_i^*} P(Y_i = 1|G_i^*)P(G_i^*|P_i)} = \prod_i \frac{R_{G_i}(\beta)}{\sum_{G_i^*} R_{G_i^*}(\beta)} \tag{4.43}$$

  where G_i^* is the permitted genotype based on the parent genotype. Use the simple log-additive model of RR (the RR of a genotype is the product of the RR of each allele), the likelihood can be written as:

$$L(\beta) = \prod_{ij} \frac{r_{g_{ijt}}}{r_{g_{ijt}} + r_{g_{ijn}}} \tag{4.44}$$

  where r_g is the RR of the allele g, g_{ijt} is the allele transmitted from parent j of subject i and g_{ijn} is the corresponding nontransmitted allele. From the likelihood, we see that it is a function of the number of transmitted and nontransmitted alleles.

- TDT as a score test [Thomas, Chapter 9, p274]: from the likelihood function, assume $R_{G_i}(\beta) = \exp(g_i\beta)$, the log-likelihood:

$$l(\beta) = \sum_i \left[g_i\beta - \log \sum_{g \in G} \exp(g\beta) \right] \tag{4.45}$$

  It can be shown that the score:

$$U = \sum_i \left[g_i - \mathrm{E}\left(g_i|p_i\right) \right] \tag{4.46}$$

  where $\mathrm{E}\left(g_i|p_i\right)$ is the expectation of g_i given the parental genotypes p_i. So the TDT is a test of departure of the actual transmitted genotypes in the affected children vs. what is expected from Mendel's Laws.

- TDT is a test of both linkage and association: it can be shown that under either situation, $\theta = 1/2$ or $D' = 0$, the expectation of TDT statistic is 0.

Analysis of TDT:

- Power analysis: we need to compute the distribution of T under H_A: the marker is linked with the disease locus with $\theta < 1/2$. Suppose there are two alleles at the disease locus s, and m alleles at the marker. We denote TM the transmitted marker and NM the non-transmitted marker and A the disease status ($A = 1$: affected). We define:

$$t_{ij} = P(TM = i, NM = j|A = 1) \tag{4.47}$$

  Then the the probabilities of transmitting and non-transmitting i to an affected child (to compute the distribution of T and R) are:

$$P(TM = i|A = 1) = \sum_{j=1}^{m} t_{ij} \qquad P(NM = i|A = 1) = \sum_{j=1}^{m} t_{ji} \tag{4.48}$$

  The probability t_{ij} can be computed by summing over the transmitted disease allele (TD), denoted as s:

$$P(TM = i, NM = j, A = 1) = \sum_{s=1}^{2} P(TH = si, NM = j)P(A = 1|TD = s) \tag{4.49}$$

  where TH stands for the transmitted haplotype. The first term can be computed from the meiotic process:

$$P(TH = si, NM = j) = p_{si}p_j(1 - \theta) + p_{sj}p_i\theta \tag{4.50}$$

and the second term from summing over the possible transmitted disease allele from the other parent (OTD):

$$P(A=1|TD=s)=\sum_{u=1}^{2}P(A=1|TD=s,OTD=u)P(OTD=u) \tag{4.51}$$

Family-based Association Test (FBAT) [LL, Appendix B]

- Model: suppose we have multiple families, and in each family, we have data of multiple offsprings. Let x_{ij} be the genotype of the j-th sibling of the i-th family, and y_{ij} be the phenotype. Let P_i be the parent genotypes of the i-th family. It can be shown from the likelihood (similar to TDT) that the score test

$$U=\sum_{i,j}\left[y_{ij}-\mathrm{E}\left(y_{ij}\right)\right]\left[x_{ij}-\mathrm{E}\left(x_{ij}|p_i\right)\right] \tag{4.52}$$

  So the test is effectively a covariance between phenotypes and genotypes (departure from the mean - in the case of genotype, mean is conditional on parental genotypes). The variance:

$$\mathrm{Var}\left(U\right)=\sum_{ij}\left[y_{ij}-\mathrm{E}\left(y_{ij}\right)\right]^2\mathrm{Var}\left(x_{ij}|P_i\right) \tag{4.53}$$

  Both $\mathrm{E}\left(x_{ij}|P_i\right)$ and $\mathrm{Var}\left(x_{ij}|P_i\right)$ are computed using Mendel's Law.

- Special case: when y_{ij} is binary and only cases are considered, this reduces to TDT.

A Bayesian approach to genetic association studies with family-based designs [Naylor & Lange, GE, 2010]

- Model ideas: in FBAT, only transmission information is used (inference is conditioned on affected states and parental genotypes). However, there is additional information in the parental genotypes and child phenotypes: if parents carrying minor alleles are also likely to have affected children, then this SNP is likely associated. More formally, let x be the child genotype, y be their phenotypes and P be the parental genotypes:

$$P(x,y,P)=P(x|y,P)P(y,P) \tag{4.54}$$

  The first part uses information in transmission, and the second information of association between parental genotypes and child phenotypes.

- Model: we are testing hypothesis using posterior odds:

$$\frac{P(H_1|x,y,P)}{P(H_0|x,y,P)}=\frac{P(x|y,P,H_1)}{P(x|y,P,H_0)}\times\frac{P(H_1|y,P)}{P(H_0|y,P)} \tag{4.55}$$

  We compute the BF of the first part. The model is $y_i\sim N(\mu+ax_i,\sigma^2)$, where a is prior effect size. For the second part, we have $P(H_i|y,P)\propto P(y|H_i,P)$, for $i=0,1$. The marginlal likelihood is computed by:

$$\mathrm{E}\left(y_i|P_i\right)=\mu+a\mathrm{E}\left(x_i|P_i\right) \tag{4.56}$$

  and summing over $x_i|P_i$.

Rare Variant Analysis for Family-Based Design [De & Laird, PLoS ONE, 2013]

- Multi-marker FBAT: for the j-th marker, let x_{ij} be its genotype at the i-th subject. The FBAT statistic can be expressed as:

$$U_j=\sum_{i}\left(y_i-\mu\right)\left[x_{ij}-\mathrm{E}\left(x_{ij}|P_{ij}\right)\right] \tag{4.57}$$

And its variance:

$$\mathrm{Var}\left(U_j\right) = \sum_i (y_i - \mu)^2 \mathrm{Var}\left(x_{ij}|P_{ij}\right) \tag{4.58}$$

To extend to the multi-marker case, we let $(U_1, \cdots, U_M)$ be the vector of FBAT statistics, where M is the number of markers. Each of U_i is normally distributed under H_0, but U_i's are correlated because of LD in a region. So U would follow MVN and we need to obtain the covariance matrix of U. We first compute V_E, with its element:

$$e_{jk} = \sum_i (y_i - \mu)^2 \left[x_{ij} - \mathrm{E}\left(x_{ij}|P_{ij}\right)\right]\left[x_{ik} - \mathrm{E}\left(x_{ik}|P_{ik}\right)\right] \tag{4.59}$$

The covariance matrix is then obtained from V_E with appropriate normalization (see the paper). The test statistic $T = U^T V_A^{-1} U$ follows χ^2 distribution with dof equal to the rank of V_A.

- Rare variant FBAT: we cannot use the multi-marker test because it will lose power (too many degrees of freedom). So we effective collapse the rare variants in the region. The test statistic is:

$$W = \sum_i (y_i - \mu) \left[\sum_{j=1}^{M} (x_{ij} - \mathrm{E}\left(x_{ij}|P_{ij}\right))\right] \tag{4.60}$$

The variance of W now needs to take the correlation among RVs into account:

$$\mathrm{Var}\left(W\right) = \sum_i (y_i - \mu)^2 \left[\sum_{j=1}^{M} \mathrm{Var}\left(x_{ij}|P_{ij}\right) + \sum_{j \neq k} \mathrm{Cov}\left(x_{ij}, x_{ik}|P_{ij}, P_{ik}\right)\right] \tag{4.61}$$

Also note that weighting can be allowed, i.e. we can have $W = \sum_j w_j U_j$, where w_j is the weight of the j-th variant and U_j the FBAT statistic of j.

Utilising Family-Based Designs for Detecting Rare Variant Disease Associations [Preston and Dudbridge, Ann Hum Genet, 2014]

- The advantage of family studies: no population structure, enrichment of risk variants.
- Study design: treating trios as case-controls (1) Pseudo-case-control (PCC); (2) Unrelated controls (UCC). Use trios, and enriched trios (one affected child from each multiplex family).
- Main results: (1) Simplex family (trios): family tests (score tests) and case-control tests have equal power. (2) Multiplex family (enriched trios): UCC gives the best power. Explanation: in multiplex families, pseudo-controls are not typical of general population, and devoid of genotypic variations - this reduces power.
- Comparison of various RVAT methods: SSU family (including C-alpha and SKAT) performs best at low causal fraction $< 40\%$, while KBAC best at higher proportions.

The nature of nurture: Effects of parental genotypes [Kong, Science, 2018]

- Model of genetic nurturing: considering a single locus/SNP. Let T and NT be the genotype of transmitted and non-transmitted alleles of the parents, respectively. Let X_O be the phenotype of interest of the offspring. We assume that genotypes can affect X_O via parental nurturing, which is mediated by parental phenotypes Y_P. Note that Y_P are mostly unobserved. Let δ be the effect of the transmitted alleles on the offspring phenotype (direct effect), our model can be written as:

$$X_O = T \cdot \delta + Y_P \cdot \gamma + \epsilon_{X_O} \qquad Y_P = (T + NT) \cdot \alpha + \epsilon_{Y_P} \tag{4.62}$$

We can then plug in Y_P into the equation of X_O:

$$X_O = T \cdot (\delta + \gamma\alpha) + NT \cdot (\gamma\alpha) + \epsilon \tag{4.63}$$

Denote $\eta = \gamma\alpha$ be the genetic nurturing effect. If we do regression of X_O on T and NT, the expected effects of T and NT are given by:

$$\mathrm{E}\,(\hat{\theta}_T) = \delta + \eta \qquad \mathrm{E}\,(\hat{\theta}_{NT}) = \eta \tag{4.64}$$

The difference of the two estimates gives genetic nurturing effects. If we use a prior on δ, we can estimate the contribution of genetic nurturing on phenotypes.

- Effect sizes from GWAS: in GWAS, we consider only transmitted alleles, so the GWAS effect sizes include both direct and genetic nurturing effects. This is true for any single SNP, and for PRS. So we can partition the PRS into those from direct effects and those from genetic nurturing.
- Estimating genetic nurturing in Education Attainment (EA): first construct GWAS PRSs. Then given family data, we can transmitted PRS and non-transmitted PRS, $poly_T$ and $poly_{NT}$ (using transmitted and non-transmitted genotypes, but the fixed weights). Then we can compute the association of $poly_T$ and $poly_{NT}$ with EA status. This gives the estimated $\hat{\theta}_T$ and $\hat{\theta}_N T$, respectively. For EA, the effects are 0.22 and 0.067 (or PVE 0.05 and 0.025).
- Remark: the PRS model takes estimated weights/effects. However, the weights already include genetic nurturing effects. Shall we use the data to estimate these effects for each SNP?
- Incorporating associative mating with direct effects only: two loci case. Because of associative mating, genotype of one parent is correlated with genotype of the other parent in the same locus (cis), or any other loci (trans) - Figure 2. Consider two loci A and B. Let A_{TM}, A_{TP} be the genotype of maternal and paternal transmitted alleles. Similar for B_{TM}, B_{TP}. Let δ and δ_B be the direct effects of the two loci. We now have the model:

$$X = \delta(A_{TM} + A_{TP}) + \delta_B(B_{TM} + B_{TP}) + \epsilon \tag{4.65}$$

The correlation of X with $A_T = A_{TM} + A_{TP}$ and $A_{NT} = A_{NTM} + A_{NTP}$ gives the estimated effects $\hat{\theta}_T$ and $\hat{\theta}_{NT}$. To compute these covariance/correlation, we note the dependency of the genotypes in the above equation include:

  - Cis-correlation: A_{TM} vs. A_{TP}, A_{TM} vs. A_{NTP}, A_{TP} vs. A_{NTM} and A_{NTM} vs. A_{NTP}. Note that A_{TM} vs A_{NTM} and A_{TP} vs A_{NTP} are independent.
  - Trans-correlation: A_{TM} vs B_{TP}, and A_{TP} vs. B_{TM}. Note that A_{TM} vs. B_{TM} are mostly independent (ignored in the derivation).

We also introduce some notation, for the variance of genotypes:

$$\mathrm{Var}\, A_{TP} = \mathrm{Var}\, A_{TM} = \mathrm{Var}\, B_{TP} = \mathrm{Var}\, B_{TM} = v \tag{4.66}$$

and the same value for non-transmitted alleles. With these assumptions, we can now show that:

$$\mathrm{Var}\, A_T = \mathrm{Var}\,(A_{TP}) + \mathrm{Var}\,(A_{TM}) + 2\mathrm{Cov}\,(A_{TP}, A_{TM}) = 2v(1 + \mathrm{cor}(A_{TP}, A_{TM})) \tag{4.67}$$

where $\mathrm{Cov}\,(A_{TP}, A_{TM}) \neq 0$ due to associative mating. And

$$\mathrm{Cov}\,(A_T, A_{NT}) = \mathrm{Cov}\,(A_{TP} + A_{TM}, A_{NTP} + A_{NTM}) = 2v \cdot \mathrm{cor}(A_{TP}, A_{TM}) \tag{4.68}$$

where we use $\mathrm{Cov}\,(A_{TP}, A_{NTP}) = \mathrm{Cov}\,(A_{TM}, A_{NTM}) = 0$. And the covariance of A_T and B_T:

$$\mathrm{Cov}\,(A_T, B_T) = \mathrm{Cov}\,(A_{TM} + A_{TP}, B_{TM} + B_{TP}) = 2v \cdot \mathrm{cor}(A_{TM}, B_{TP}) \tag{4.69}$$

To get estimated effect of T on trait, we compute:

$$\mathrm{Cov}\left(X, A_T\right) = \mathrm{Cov}\left(\delta A_T + \delta_B B_T, A_T\right) = \delta \mathrm{Var}\left(A_T\right) + \delta_B \mathrm{Cov}\left(A_T, B_T\right) = 2v\left[\delta(1+\mathrm{cor}(A_{TM}, A_{TP})) + \delta_B \mathrm{cor}(A_{TM}, B_{TP})\right] \tag{4.70}$$

Using these results, we can obtain that the extra effect due to asociative mating, i.e. $\hat{\theta}_T = \delta + \phi_\delta$, with

$$\phi_\delta = \frac{\delta_B \mathrm{cor}(A_{TM}, B_{TP})}{1 + 2\mathrm{cor}(A_{TM}, A_{TP})} \tag{4.71}$$

Note that the denominator is close to 1. So the main contribution comes from δ_B, and the extent of associative mating, characterized by $\mathrm{cor}(A_{TM}, B_{TP})$.

- Associative mating using PRS: the analysis above is based on two loci. Now we suppose A captures PRSs and B the remaining loci. Then the effect sizes δ and δ_B are not equal, we denote: $\pi = \delta_B^2/\delta^2$ as the ratio of PVE. We can also view $\sqrt{\pi}$ as the ratio of number of independent alleles in B vs. A. Then we have:

$$\mathrm{cor}(A_{TM}, B_{TP}) = \sqrt{\pi}\mathrm{cor}(A_{TM}, A_{TP}) \tag{4.72}$$

With this, we can obtain:

$$\frac{\phi_\delta}{\delta} = \frac{\pi \mathrm{cor}(A_{TM}, A_{TP})}{1 + 2\mathrm{cor}(A_{TM}, A_{TP})} \tag{4.73}$$

Now we can estimate ϕ_δ/δ for EA by: (1) Estimation of π: we know that direct effects explain 17.0% PVE (using heritability analysis). And PRS explains 2.45%. So $\pi = (17.0 - 2.45)/2.45 = 5.94$. (2) Using transmitted genotype data, we can estimate $\mathrm{cor}(A_{TM}, A_{TP}) = 0.012$.

- Incorporating associative mating in the full model: we can expand the analysis above to the full model with both direct and genetic nurturing effects. The results:

$$\frac{\phi_\eta}{\eta} = 2 \times \frac{\phi_\delta}{\delta} \tag{4.74}$$

The factor of 2 is from non-transmitted alleles have the same nurturing effect as the transmitted allele. In summary, we have the measured effects as:

$$\mathrm{E}\left(\hat{\theta}_T\right) = \delta + \phi_\delta + \eta + \phi_\eta \qquad \mathrm{E}\left(\hat{\theta}_{NT}\right) = \phi_\delta + \eta + \phi_\eta \tag{4.75}$$

- Estimating genetic nurturing and associative mating in EA: $\hat{\eta}$ accounts for 75% of $\mathrm{E}\left(\hat{\theta}_{NT}\right)$, and is 32% of $\hat{\delta}$.
- Parent-of-origin: we can do the analysis on separate parents to estimate genetic nurturing effects via P or M. For 7 traits, only height show difference in η between parents, and it accounts for 45% of $\mathrm{E}\left(\hat{\theta}_T\right)$.
- Impact of genetic nurturing on h_g^2 analysis: by definition h_g^2 should include only direct effects. But using GREML will include the nurturing effects, so the results will be biased.
- Nature of genetic nurturing effects: in the EA cases, parental EA may mediate some of the nurturing effects, but is only a small fraction.
- **Remark**: is it possible that a fraction of GWAS associations are entirely driven by genetic nurturing effects?

Estimating genetic nurture with summary statistics of multi-generational genome-wide association studies [Wu and Qiongshi Lu, review for PNAS, 2020]

- Problem: can we estimate direct and indirect (genetic nurturing) effects, including paternal and maternal ones, from GWAS summary statistics of offspring phenotype vs. offspring genotype, maternal genotypes and paternal genotypes, respectively?

- Model: Fig. 1. Our goal is to express the estimated effects from GWAS summary stats as functions of the true effect sizes, including indirect effects. Let Y_O be the offspring trait, G_O be the offspring genotype. Let G_M and G_P be the maternal and paternal genotypes. We have:

$$Y_O = \beta_{\text{dir}} G_O + \beta_{\text{ind\_}M} G_M + \beta_{\text{ind\_}P} G_P + \epsilon \tag{4.76}$$

Now we can plug in $G_M = T_M + NT_M$ and $G_P = T_P + NT_P$, we have:

$$Y_O = (\beta_{\text{dir}} + \beta_{\text{ind}}) G_O + (\beta_{\text{ind\_}M} NT_M + \beta_{\text{ind\_}P} NP_M) + \epsilon \tag{4.77}$$

We now use $\hat{\beta}_O = (G_O^T G_O)^{-1} G_O^T Y_O$. Plug in the equation of Y_O and use the fact:

$$\text{Cov}\left(G_O, G_{M,P}\right) = p(1-p)(1+\alpha) \tag{4.78}$$

where p is AF and α is the correlation of maternal and paternal alleles from associative mating. From this, we have:

$$\text{E}\left(\hat{\beta}_O\right) = \beta_{\text{dir}} + \left(1 + \frac{\alpha}{2+\alpha}\right) \beta_{\text{ind}} \tag{4.79}$$

Similarly, we can derive $\hat{\beta}_M$ - summary stats of maternal GWAS, and $\hat{\beta}_P$. From these equations, we can derive the MOM estimators of $\beta_{\text{dir}}, \beta_{\text{ind\_}M}$ and $\beta_{\text{ind\_}P}$. The paper describes several cases, depending on whether we have all three GWAS, or just two (e.g. by merging maternal and paternal GWAS). See Table 1.

- Standard errors of the estimators and accounting for sample overlap: see "Variances and covariances among effect size estimators" in Supplements. First show $\text{Var}\left(\hat{\beta}_{\text{dir}}\right)$ as function of variance of $\hat{\beta}_{O,M,P}$ (given) and pairwise covariance among the three. Then to obtain covariance, model sample overlap, using LDSC intercept term.

- Correlation among the estimators: there are correlations b/c they are all functions of the same summary statistics, $\hat{\beta}_{O,M,P}$. (1) Correlation between direct and indirect effect estimators. (2) Correlation between maternal and paternal indirect effects.

- Analysis: is the model equivalent to the use of individual level data? The benefit of individual data is to do joint regression. However, note that T_M and NT_M (also paternal) are independent, so it makes no difference. The dependency of T_M and T_P due to associative mating is accounted for in the summary stats model.

- Simulations: use direct effect of $\beta = 0.02$ or PVE 0.04% (roughly in line with 10K causal SNPs, toatl h2g = 0.4). Indirect effect is somewhat smaller. Show the estimator is unbiased, and type 1 error calibrated.

- Application to birth weight: Fig. 3, show that using GWAS with complete sample overlaps (between O and M), the results are similar to the results using orthogonal phenotypes. The standard errors would be larger if not account for sample overlap.

- Application to EA and study of genetic correlations of EA and 45 other traits: using GWAS-O and GWAS-MP to estimate EA direct and indirect effects. No genome-wide significant associations. Do correlation with 45 traits: found some pairs with large correlation with indirect effects, and large correlation of direct effect with ASD.

- Correlation of indirect effects of EA vs. other traits: (1) EA-smoking relationship: indirect effect correlates with less smoking. (2) EA indirect effect correlates with lower BMI, larger height, and lower risk of RA.

- EA-ASD relationship: only direct effect has positive correlation with ASD risk. Confirmed by transmission bias of EA PRS in ASD probands.

4.4 Polygenic Modeling

Heritability [personal notes]

- Why heritability matters? If some factor is important for a trait, then it should explain the variation of that trait. Thus heritabilty is a fundamental way of quantifying the importance of one factor over some trait. The idea can be applied in many contexts: e.g. measuring the importance of one type of variances over another (say coding vs non-coding).
- Perspective of genetic and phenotypic similarity: if a trait is very heritable, then genetically similar individuals should have similar phenotypes, so the ratio of phenotypic covariance and genetic covariance reflects heritability.
- Perspective of effect sizes: if there are many variants of large effect sizes, then a large fraction of phenotypic variance can be explained by genetic variants.
- General considerations of heritability estimation: genetically similar individuals tend to be raised in similar environment, then the genetic and environmental effects are coupled/confounded. If we do not remove the confounding, we will overestimate heritability.

Strategies of mapping heritability:

- Partition of variance: we write the phenotype as:

$$y = \mu + g + \epsilon \tag{4.80}$$

  where g is the random genetic effect of an individual (see below: could view genotype of any individual as random; or a combination of genotype and effect size). From this, we obtain variance of y:

$$V_P = \text{Var}(y) = \text{Var}(g) + \text{Var}(\epsilon) = V_A + V_E \tag{4.81}$$

  where V_A is interpreted as genetic variance, or variance explained by genetic variation. One can imagine two individuals with identical genotype (monozygotic twin), then their phenotypic covariance is V_A.

- Estimating V_A through genetic relation/similarity: the idea is that covariance between phenotypes of individuals (related ones) carry information of V_A, so we derive the covariance matrix of y (vector) in terms of V_A.
- Estimating V_A through effect sizes: if we write g as a product of genotype (Z) and effect sizes (u):

$$y = \mu + Zu + \epsilon = \mu + \sum_j Z_j u_j + \epsilon \tag{4.82}$$

  Then obviously covariance of y depends on u_j's. So we can estimate u_j's through covariance of y. To relate effect sizes with V_A, we use basic relation from regression model:

$$\text{SSR} = \sum_j \hat{u_j}^2 \text{Var}(Z_j) \tag{4.83}$$

- Remark: the problem is basically making inference of a random effect model, marginalizing the random effects g_i. Becuase the joint distribution of y (of many individuals) follow MVN, the key is to estimate the covariance matrix of y.

Problems of estimating heritability [personal notes]:

- GCTA/LMM: assumption of random effect (non-sparse). How would this affect the estimation if the assumption does not hold (most likely)? Intuitively, if there are a small number of large effect loci, then LMM would estimate small σ (most loci have no effects), and this would underestimate h^2?

- Bayesian method: could use a mixture prior. Matthew's comment: the sparse Bayesian prior is sensitive to the non-sparse scenario (perform badly), while GCTA is relatively robust to the assumptions. Question: what if we use a mixture prior, with flexible π?
- Impact of SNP heritability (effect size) prior: GCTA and LDSC vs. LDAK prior. Justification of LDAK prior was based on tagging in the paper, however, this is not satisfactory. Possible explanation of the observation that low LD regions (high recombination rates r) tend to explain more heritability than expected based on population genetics: causal variants are likely deleterious
  - Linked selection view: given a deleterious variants, in low r regions, it reduces the effective population size, hence strength of selection and this will lead to more deleterious variants (higher effect size) in the region.
  - Epistasis view: Given a deleterious variant, in low r regions, a nearby SNP is less likely to be deleterious because the haplotype containing two deleterious SNPs will be under strong purifying selection.

  What explains the discrepancy?

Quantitative genetic background:

- Random-effect model of phenotypes: we typically write the trait as $P = g + e$ where g is genetic effect (breeding value) and e the environmental effect. The genetic effect can be decomposed as:

$$g = \mu + \alpha_i + \alpha_j + \delta_{ij} \tag{4.84}$$

  where α_i and α_j are the effects of the two alleles, and δ_{ij} the dominance effect. We consider the model of P as a random effect model: it has a group effect (genetic) and individual error (environmental). The groups can be thought of possible genotypes, and the group effect is random because which group an individual belongs to is random. An analogy is: treating individuals with different dosages (groups) - this is usually a fixed effect model; but if the dosage is randomly assigned, then a random effect model.
- How breeding values are related to IBD? Let g and g' be the breeding values of two individuals, then the covariance between the two is from the decomposition of g above:

$$\mathrm{Cov}\left(g, g'\right) = \left(\frac{1}{2}P_1 + P_2\right) V_A + P_2 V_D \tag{4.85}$$

  where P_1 and P_2 are the probabilities of sharing one or two alleles IBD between the two individuals.
- Kinship coefficient: probability that two randomly chosen alleles of two individuals are IBD. It's easy to see that:

$$\phi = \frac{1}{4}P_1 + \frac{1}{2}P_2 \tag{4.86}$$

  When the two individuals share one IBD allele, there is only probability of $1/4$ that the two randomly chosen alleles are these two ones; when two alleles share two IBD alleles, the chance is $1/2$. Some special case: MZ twins, $P_2 = 1$, thus $\phi = 1/2$.

Using expected genetic relationship (pedigree) to estimate heritability [personal notes]

- Model: for the i-th individual, let y_i be its phenotype, we have:

$$y_i = \mu + u_i + \epsilon_i \tag{4.87}$$

  where u_i is the genetic effect (breeding value) and ϵ_i the environmental effect. Notes that the equation ignores the dominance term. To compute the covariance of two subjects: $\mathrm{Cov}\left(y_i, y_j\right) = \mathrm{Cov}\left(u_i, u_j\right)$,

we first note that this covariance is twice of the covariance from alleles. Next, for a pair of alleles, its covariance is 0 if independent, and σ_a^2 if IBD. Let ϕ_{ij} be the probability of IBD, so we have:

$$\text{Cov}(y_i, y_j) = \text{Cov}(u_i, u_j) = 2\phi_{ij}\sigma_a^2 \tag{4.88}$$

The kinship is given by: $\phi_{ij} = P_1/4 + P_2/2$. When $i = j$, we have:

$$\text{Var}(y_i) = \sigma_a^2 + \sigma_e^2 \tag{4.89}$$

In the matrix form, we have the covariance matrix:

$$\text{Var}(y) = 2\sigma_a^2\Phi + \sigma_e^2 I \tag{4.90}$$

- Estimating variances and heritability: when we have multiple individuals, then we can fit the MVN to the data, where the covariance matrix is given above. When the relationship among individuals are known, the kinship matrix is given, and one only need to estimate σ_a^2 and σ_e^2 terms. When the kinship matrix is unknown, one needs to estimate it first.
- Remark: the derivation assumes one source of breeding values. But we can show that when we have multiple sources, the same results hold. In this case, the expected genetic relationship ϕ_{ij} between any two individuals is the same across all loci, and σ_a^2 is the sum of contribution of all loci.
- Remark: this result is based on expected genetic sharing, and on known pedigree; in the general case of population design, we can still use similar idea, but the kinship matrix would have different interpreatations.

Heritability and polygenic scores/effect sizes [personal notes]

- Polygenic scores: defined as the sum of estimated sizes of all variants present in a sample. Polygenic scores can be used for a number of applications, e.g. estimating heritability, disease risk prediction, and so on. It is closely related to the mixed model approach to heritability.
- Intuition: we are interested in variance of trait explained by genetics (Proportion of variance explained, or PVE). This is similar to R^2 in regression, and it is related to the effect size estimates. Intuitively, with higher heritability, we should see many variables or large effect sizes; conversely, if effect size estimates are close to 0, then heritability is low.
- Heritability in terms of effect size estimates: assuming we have n independent markers, then using the results from [Relationship between R^2 and regression coefficients, Statistics notes], we have:

$$\hat{\sigma_a}^2 = \text{SSR} = \sum_j \hat{\beta_j}^2 \text{Var}(x_j) \qquad \hat{\sigma_e}^2 = \text{SSE} = \hat{\sigma}^2 = \text{Var}(\hat{\beta})(X^T X) \tag{4.91}$$

This leads to an estimate of h_a^2.

- Comparison of random effect model based estimate vs. summary statistics based estimate: both are based on the idea that if heritability is high, similar genotypes should lead to similar phenotypes. The difference is: (1) LMM uses implicity genotype similarity, measured by kinship coefficient; (2) effect size approach uses the explicity genotypes.
- Remark: this estimate represents only chip-heritability (explained by tagged SNPs), not narrow-sense heritability.

Heritability in the genomics era - concepts and misconceptions [Visscher & Wray, NRG, 2008]

- Why most heritability is based on additive? Most relatives share only 1 or 0 IBD, thus dominance that is based on sharing two copies is not relevant.

- Heritability is not constant. Genetic variance depends on segregation in a population of the alleles that influence the trait, the allele frequencies, the effect sizes of the variants and the mode of gene actions. All these variables can differ across populations.

- Comparison of heritability of traits? Ex. morphological vs. fitness traits (higher heritability). Heritabilty also higher in more favorable environments.

- Why so much genetic variance is additive and why h2 is so large? Theory predicts that additive genetic variance should be depleted because of natural selection, and biology tells us that genes work in interactive pathways, which implies non-additive interaction variance. Possible answers: strong interactions not a problem if gene frequency is near 0 or 1.

A Tool for Genome-wide Complex Trait Analysis (GCTA) [personal notes; Yang and Visscher, AJHG, 2011]

- Motivation: In the general case, we may not know the pedigree relationship; furthermore, even with pedigree, the actual genetic similarity (realized genetic sharing) may not be equal to the expected similarity inferred from IBD; so we will need to derive the heritability in terms of the realized genotypes.

- Theoretical framework: linear mixed model (LMM). For any indvidual, its trait:

$$y_i = X_i\beta + Z_i u + \epsilon_i \tag{4.92}$$

  where X_i is the vector of fixed effects, including covariates such as age and ex and known genotypes of causal variants, and Z_i the genotypes of causal variants(standardized with variance equal to 1). The coefficients β and u represent the effect sizes. We makes the assumption that u_j of the j-th variant is random: $u_j \sim N(0, \sigma_u^2)$. To do LMM analysis, we first subtract the fixed effects, ans assuem for now that we only deal with the remaining terms. The covariance between two individuals is:

$$\mathrm{Cov}\left(y_i, y_k\right) = \mathrm{Cov}\left(Z_i u, Z_k u\right) \tag{4.93}$$

  We expand the genetic loci:

$$\mathrm{Cov}\left(Z_i u, Z_k u\right) = \sum_j \mathrm{Cov}\left(Z_{ij}u_j, Z_{kj}u_j\right) = \sum_j Z_{ij}Z_{kj}\mathrm{Cov}\left(u_j, u_j\right) = Z_i Z_k^T \sigma_u^2 \tag{4.94}$$

  where we uses the random effect of u_j: $\mathrm{Cov}\left(u_j, u_j\right) = \sigma_u^2$. When $i = k$, we have:

$$\mathrm{Var}\left(Z_i u\right) = Z_i Z_i^T \sigma_u^2 \tag{4.95}$$

  This allows us to write the covariance matrix of y:

$$\mathrm{Cov}\left(y\right) = ZZ^T\sigma_u^2 + I\sigma_e^2 \tag{4.96}$$

  This is the basic relation between covariance of traits and the effect size distribution. The model is also called "variance component" model, because the covariance of phenotypes is partitioned into two parts, one from genetic variation, and the other from environmental variation (or more generally, across-group variation and within-group variation, where group means genotype here).

- Alternative derivation: write in the vector form, $Y = X\beta + Zu + \epsilon$, our prior is $u \sim N(0, I\sigma_u^2)$. Ignoring $X\beta$, and treat u as random and Z fixed, we use the result from MVN:

$$\mathrm{Var}\left(Y\right) = \mathrm{Var}\left(Zu\right) + I\sigma_e^2 = Z(I\sigma_u^2)Z^T + I\sigma_e^2 = ZZ^T\sigma_u^2 + I\sigma_e^2 \tag{4.97}$$

- Remark: the interpretation is, suppose we have given genotypes, but our effect sizes u are random, then on average, what do we expect about the relationship of phenotypes of two individuals?

- Estimating heritability: to relate σ_u^2 to heritability, we note that $Z_i u$ is a linear combination of a random vector, so we can obtain its variance (i.e. the variance explained by SNPs):

$$\sigma_a^2 = \mathrm{Var}\,(Z_i u) = Z_i Z_i^T \sigma_u^2 = \sigma_u^2 \sum_j Z_{ij}^2 \approx m\sigma_u^2 \tag{4.98}$$

where we use the fact that Z_{ij} is normalized s.t. its variance is equal to 1. To understand this, we could derive this by using the basic relationship between RSS (explained variance) and effect size from regression model:

$$V_A = SSR = \sum_j u_j^2 \mathrm{Var}\,(Z_j) = \sum_j u_j^2 \approx m\sigma_u^2 \tag{4.99}$$

where we uses the fact that $\mathrm{E}\,(u_j^2) = \mathrm{Var}\,(u_j) = \sigma_u^2$ and m is the number of causal SNPs. Note that this derivation makes the assumption that SNPs are independent, which is not actually needed. We define the genetic relationship matrix (GRM) A as:

$$A_{ik} = \frac{1}{m} \sum_{j=1}^{m} \frac{(x_{ij} - 2p_j)(x_{kj} - 2p_j)}{2p_j(1-p_j)} \tag{4.100}$$

where x_{ij} and x_{kj} are raw genotypes. Then $A = ZZ^T/m$, so we can write covariance matrix:

$$\mathrm{Cov}\,(y) = A\sigma_a^2 + I\sigma_e^2 \tag{4.101}$$

where $\sigma_a^2 = V_A$ is the additive genetic variance. So in practice, we first estiamte A from data, then fit the data (covariance) to estimate σ_a^2 and σ_e^2.

- Chip heritability and related individuals: in the model, we use only genotyped SNPs, and the heritability is those explained by these SNPs, hence called chip-heritability. To see this, we note that the model is correct if the covariance matrix from causal SNPs (ZZ^T) is equal to that estimated from genotyped SNPs (A). This is OK if the samples are independent. However, when there are related subjects, even if we do not tag all the causal variants, we may still estimate the matrix ZZ^T (only need a fraction of variants to estimate relatedness). Then we cannot say that the heritability is explained by our genotyped SNPs. So we should remove close relatives in the analysis (e.g. 0.025 as cutoff).

- Partition of genetic variance: suppose we partition our SNPs by groups, e.g. chromosomes, then each group explains a small amount of genetic variance ($\sigma_g^2 = m\sigma_u^2$, so roughly it is proportional to the number of SNPs). To estimate each of them, we have:

$$\mathrm{Cov}\,(y) = \sum_i A_i \sigma_i^2 + I\sigma_e^2 \tag{4.102}$$

where σ_i^2 is the contribution of the i-th group and A_i is the GRM estimated from the SNPs in this group. The model is generally identifiable since A_i's would be different in different pairs: e.g. when estimating σ_1^2, we use the pairs with the highest A_1.

- The impact of LD (personal notes): GCTA does not require independence of SNPs, as it accounts for all SNPs simultaneously. In the two key steps in the derivation: $\mathrm{Var}\,(Y)$ in terms of σ_u^2, and $\sigma_a^2 = m\sigma_u^2$, neither requires independence. LD only becomes a problem when causal variants and non-causal variants have different LD and MAF.

Estimation and partition of heritability in human populations using whole-genome analysis methods [Vinkhuyzen & Visscher, ARG, 2013]

- History: Galton, resemblance of relatives. Fishers theory: reconcile two schools. BLUE and BLUP for estimation and prediction in animal and plant breeding.

- Three designs: contrasting individuals of different genetic relatedness. All these methods share a common conceptual framework: for genetically similar individuals, how similar they are phenotypically. The difference lies on how genetic similarity is defined: expected or realized; family or population.
  - Twin design: use comparison of MZ twins and DZ twins. MZ twins should be more similar than DZ twins, if heritability is high.
  - With-family design: use full-siblings.
  - Population design: genetically similar individuals from population (due to population history, they may share part of the genome, even if they are unrelated).
- Genetic relationship matrix (G): the accuracy of estimation methods depend on the estimateion of G. In general, the sampling variance of G is small for close relatives, large for distant ones (Figure 1). However, bias is more of a problem for close relatives (confounding with environment).
- Population design: we first need to estimate Genetic relationship matrix (GRM) - the matrix G in the LMM framework. The term G_{ik} is the average genetic relationship (covariance) of causal variants. If we assume the causal variants are indistinguishable from genetic markers (in terms of genetic relateness), then a simple estimation of G as the average correlation of genotypes:

$$G_{ik} = \frac{1}{m}\sum_{j=1}^{m}\frac{(x_{ij}-2p_j)(x_{kj}-2p_j)}{2p_j(1-p_j)} \tag{4.103}$$

  where p_j is the AF of the marker j. The estimation can be improved by using IBD inferred from haplotypes. Once we have G, we use ML of phenotypes or Equation **??**. The heritability obtained in this way is determined by the tagged SNPs in a chip, thus called "chip heritability".
  - Closely related relatives: should be removed. Shared environment.
- Comparison of designs:
  - Pedigree or family design: estimated heritability has small sampling variance, but bias (due to environmental confounding) is a seriour concern.
  - Population design: large samples so could obtain small sampling variance; and less bias (indpendent samples, so no shared environment). In fact, population design uses all pairwise comparison: so even if each pair carries relatively little information, the number of pairs is large (quadratic).
- The impact of LD:
  - If causal variants not in LD with tag SNPs, matrix G is not the same as the true G. Methods were proposed to weigh SNPs based on LD.
  - Population design: estimation of h^2 is driven by LD between casual and tag SNPs; or by distant IBD between unrelated individuals. The two descriptions are equivalent - causal variants not tagged will not contribute. For family design: estimation is driven by shared IBD in relatives estimate total h^2 because markers track all causal variants.
- Example: height.
  - Pedigree design: about 80%
  - Significant GWAS loci: 10%
  - Using estimated genetic similarity from all SNPs: 50%. The implication is that SNPs with small effects remain undetected. Incomplete LD with GWAS SNPs explain the gap (using simulation).
  - We should call "hidden heritability" instead of "missing heritability".

- Future directions:
  - Analysis of multiple phenotypes: Genetic covariance can be estimated from subjects where only one phenotype is measured.
  - The properties of causal variants as a class: effect sizes, allele frequency spectrum. This can help understand the evolutionary process and guide experiment design.
- Discussion: (Dan Nicolae) environmental effects in population design. When we use PCA to remove the effects of population substructure, do we also remove the true genetic effects, thus underestimate heritability?
  - Thought: PCs are surrogates of culture, ethnicity, and so, so we are removing the environmental effects, not the genetic ones. However, different ethnicities are genetically distince, so we do remove some common genetics. The idea is that we remove the main difference between ethnicities, but use the remaining the genetic variations to estimate heritability (i.e. controlling/stratfying ethnicity).
- **Question**: using CVs can estimate G, but does it mean we "explain" the variation of traits? Even if we miss many causal variants, we may still be able to estimate G well. Ex. imagine we are applying the method to family data, then we don't need all SNPs to estimate IBD.

Advantages and pitfalls in the application of mixed model association methods [Yang & Price, NG, 2014]

- Applications of LMM: include
  - Estimation of h^2 due to common variants.
  - Control for genetic background (even if the samples are unrelated) to increase power.
  - Correct for population substructure and cryptic relatedness.
  - Prediction of phenotypes.

  The paper focuses on the first three applications, and in particular, how different choices of using LMM affect the three issues (estimation of h^2, power and control for population substructure).
- Correct for population substructure: the idea is that the control of genetic background removes the effect of relatedness. In practice, this also controls for geographical population structure (controlling for the ancestry markers is similar to control for PCs).
- Computational costs of LMM: three steps, building GRM, estimating variance components and compute association statistics for each SNP. Different computational strategies
  - Exact strategy: compute the variance components for each candidate marker being tested. Exists more efficient methods of doing this.
  - Computing variance components only once and use them for all markers: OK if all markers have small effects. The difference between exact and approximate strategies is large with pervasive relatedness and large effect sizes.
- Issue 1: including the candidate marker (MLMe) or not (MLMi). Mathematically, one should use MLMe, as MLMi control for the candidate while testing its effect. Use of MLMi would lead to a lower power: this can be shown by studying the mean association statistics λ under different models. Using only linear regression:

$$\lambda_{mean}(LR) = 1 + Nh^2/M \tag{4.104}$$

  where N is sample size and M the number of markers. This is bigger than 1 due to polygenecity, however, when N is not too big, it is close to 1. For MLMi,

$$\lambda_{mean}(MLMi) = 1 \tag{4.105}$$

For MLMe:

$$\lambda_{mean}(MLMe) = 1 + \frac{Nh^2}{(1 - r^2h^2)M} \tag{4.106}$$

where $r^2 \approx Nh^2/M$. The difference between MLMe and MLMi can be understood as testing different H_0: a SNP has no effect (MLMe); or a SNP's effect is explained by the random effect $N(0, \sigma_u^2)$.

- Issue 2: use a subset of markers. Conceptually, one can use a subset of markers to estimate σ_u^2, thus speed up the computation. Two ways of selecting markers: top M_T markers, or random M_R set of markers. Whether use one of the two (or all markers) depends on the two (somewhat competing) goals: (1) increase the power by controlling for genetic background; (2) correct for population substructure. Specifically:
  - If the goal is to increase the power, choosing top M_T markers is OK, and the value of M_T can be chosen by maximizing the (out-of-sampe) prediction accuracy.
  - Choose a small M_T or M_R however may be insufficient to control for population substructure.
- Issue 3: ascertained case-control data. The control of genetic background may lead to loss of power in ascentained case-control samples, when the disease prevalence (f) is low.
  - Intuition [The Covariate Dilemma, PLoS Genetics, 2012]: suppose we have a non-confounding covariate (independent of test marker, but correlate to the trait), then include it may reduce power. Explanation: when f is small, the test marker and the variable may become correlated in cases, so controlling for the variable also removes some of the effect of the test marker, reducing the power.
- Recommendations:
  - Excluding candidate markers (MLMe) in preference to including them (MLMi).
  - For randomly ascertained quantitative traits: general include all markers. When population stratification is not a significant concern, could use top markers.
  - Genome-wide significant markers should be conditioned out as fixed effects.
- Directions:
  - Distinguishing between polygenic effects and incomplete correction for stratification (both lead to $\lambda_{mean} > 1$).
  - MLM methods for ascertained case-control data.
  - Use better prior model of effect sizes: mixture distribution.
- Questions: if large effect from population structure, then the effects of population ancestry markers are large. Including them as fixed effect should be more powerful?

Concepts, estimation and interpretation of SNP-based heritability [Yang and Visscher, NG, 2017]

- GREML (GCTA): not confounded by environment and epistasis.
- LD in GREML: accounted for because all SNPs are fit together. LD pruning is thus unnecessary; with LD pruning, possible that the MAF spectrum of SNPs changes.
- Bias due to LD or MAF: difference in LD and MAF of causal vs. non-causal variants can affect. Recommended GREML-LDMS: stratify by LD and MAFs and estimate h2 separately, at the cost of more parameters. LDAK: (1) Earlier version: implicitly RVs explain 10 times more variance than CVs. (2) Later version: more similar to GREML-LDMS.

- Under neutral model: variance explained by MAF bins is proportional to the MAF bin size. Can use this to test for negative selection.
- LDSC: sensitive to genetic architecture, cannot estimate contribution from rare variants, biased estimate due to incorrect LD matrices.
- Simulation study of LDSC: one major gene explains half of PVE, the rest SNPs other half (true PVE = 0.5). Sample size 13K and 500K SNPs. Both GREML and HE regression give estimated PVE close to 0.5, but LDSC gives 0.37. Possible explanation: for all other SNPs, their relationship of effect sizes and LD scores can be fit by a line with slope 0.25. Having a single large-effect locus may not be able to overcome the bias in the slope.

Improved Heritability Estimation from Genome-wide SNPs [Speed and Balding, AJHG, 2012]

- GCTA model makes several assumptions: polygenicity, normal prior, the relationship of effect size and MAF, and independence of LD (each SNP makes equal contribution regardless of LD pattern).
- Simulation to investigate how these assumptions may affect the estimated h^2 for GCTA: real genotypes from 2,500 cases and 2,500 controls, choose causal variants (different scenarios), and simulate phenotypes. The true h^2 is 0.5 or 0.8.
- Modeling effect of LD on $\hat{h}^2$: let w_j be the weight of variant j, choose w_j s.t. $w_j + \sum_j w_j r_{jj}^2 e^{-\lambda d_{jj}}$ is constant over j where d_{jj} is the distance between variants. The motivation is that the sum represents the total amount of tagging in a SNP. To implement this weighting, standardize genotype by changing X_j (SNP j) to $\sqrt{w_j} X_j$. Then update the kinship matrix in GCTA.
- Polygenicity: different numbers of causal variants from 1 to ALL. In almost all cases, the estimated h^2 is unbiased. However, when number of causal SNPs is small, GCTA estimates of standard error of h^2 is too low. The problem is fixed with weighted kinship matrix.
- Relationship of effect size and MAF: Var $\beta_j \propto [p_j(1-p_j)]^\alpha$, set α from -2 to 1. The results are relatively robust to the value of α used in the model (scaling genotype).
- Normal prior: simulation under normal exponential gamma (NEG) prior. The results are robust.
- Impact of LD on $\hat{h}^2$: (Figure 3) when causal variants are in weak LD regions, we underestimate contribution to h^2; in high LD regions, we overestimate contribution to h^2. This can be fixed with weighted kinship matrix.
- Why pattern of LD matter? It is strongly linked to MAF: the signals from low-MAF variants are less replicated (tagged), comparing with high-MAF variants. So in diseases where rare variants dominant (e.g. bipolar disorder), we tend to underestimate h^2 and in diseases where common variants and/or high-LD regions (e.g. MHC for AIDs), we tend to overestimate h^2.

Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis (BOLT-REML) [Loh and Price, NG, 2015]

- Background: methods to fit variance component model, include:
  - First-derivative methods: gradient descent, e.g. EM.
  - Second-derivative methods: calculation or approximation of Hessian matrix. While faster than first derivative methods, they are less robust when far from the optimum. Newton-Raphson (NR) method: Zhou and Stephens use faster method based on EVD. Average information (AI): average of Hessian and Fisher information matrix.

  Also Monte Carlo methods: applied to both EM and second derivative.

- Model: write the model as:
$$y = \sum_k \sigma_k Z_k u_k + \sigma_0 u_0 \tag{4.107}$$
where Z_k are normalized genotype, u_k are normalized random effects, and σ_k variance component parameters. The variance of y is given by:
$$V = \sigma_0^2 I_n + \sum_k \sigma_k^2 Z_k Z_k^T \tag{4.108}$$
This leads to the log-likelihood function:
$$l(\sigma_0^2, \sigma_1^2, \cdots, \sigma_K^2) = -\frac{1}{2}(\log\det V + y^T V^{-1} y) \tag{4.109}$$

- Computing gradient and Hessian: (1) Monte Carlo approximation of graident: MC REML. The gradient $\partial l/\partial\sigma_k^2$ is a function of $Z_k^T V^{-1} y$. This is BLUP estimates of u_k, effect sizes. In computation, replace expectation with Monte Carlo samples. (2) Approximation of Hessian: Average Information (AI).

- Optimization: (1) Trust region methods: find regions where local quadratic model of log-likelihood breaks down. They can be detected by comparing true vs. approximate log-likelihood. (2) Convergence: MC AI REML truly converges rather than jump around parameters.

- Standard errors: note that heritability and genetic correlations need to be scaled.

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies (LDSC) [Bulik-Sullivan and Neale, NG, 2015]

- Model: let χ_j be the chi-square of SNP j, $l_j = \sum_k r_{jk}^2$ be the LD score of SNP j. We have this equation: $E(\chi_j^2) = N l_j h^2/M + Na + 1$, where h^2/M is the per SNP heritability and a the effect due to population stratification or other confounding variables.

- Derivation from random effect sizes (personal notes): one possibility is to use RSS and with normal prior of effect sizes. Another way (equivalent) is: consider the normalized effect size, u and estimated effect $\hat{u}_j$ for SNP j. We are interested in the chi-square statistic, which is just variance of $\hat{u}_j$. We use Law of Total Variance, treating u as random:
$$\operatorname{Var}\hat{u}_j = \operatorname{E}_u[\operatorname{Var}(\hat{u}_j|u)] + \operatorname{Var}_u[\operatorname{E}(\hat{u}_j|u)] \tag{4.110}$$
The first term is simple: with standardized effect, its se = 1, so $\operatorname{Var}(\hat{u}_j|u) = 1$, and its expectation is 1. The second term, we have $\operatorname{E}(\hat{u}_j|u) = (Ru)_j$, where R is the LD matrix. We consider the variance:
$$\operatorname{Var}(Ru) = R \cdot \operatorname{Var}(u) \cdot R^T = \sigma_u^2 RIR^T = \sigma_u^2 R^2 \tag{4.111}$$
We now have: $\operatorname{Var}[(Ru)_j] = \operatorname{Var}(Ru)_{jj} = \sigma_u^2 R_{jj}^2$.Let $l_j = (R^2)_{jj}$ be the LD score of SNP j. So we have:
$$\operatorname{Var}\hat{u}_j = 1 + \sigma_u^2 l_j \tag{4.112}$$
Now plug in $h^2 = \sigma_u^2 M/N$, where M is number of SNPs and N sample size. This gives the LD score regression results.

- Derivation in the paper: also use Law of Total Variance. However, instead of treating effect size as given, it treats X (genotype) as given:
$$\operatorname{Var}(\hat{\beta}_j) = \operatorname{Var}_X[\operatorname{E}(\hat{\beta}_j|X)] + \operatorname{E}_X[\operatorname{Var}(\hat{\beta}_j|X)] \tag{4.113}$$
It then uses the random effect size in the equation.

- Remark: the LD score of a SNP can be thought of the effective number of SNPs tagged by that SNP. Under polygenicity assumption, the more SNPs tagged, the higher observed effect the SNP should have.
- Regression estimation: we regress the chi-square of each SNP with its LD score. The slope would suggest per SNP heritability and intercept the confounding effect. In practice, define LD scores in 1cM blocks.
- Two statistical issues: (1) SNPs are not independent; (2) Variance of the chi-square different: SNPs with high LD scores have higher variance. To address this, use weighting of SNPs by $1/l_j$.
- Obtaining standard error: knife procedure, each time removing a block of 2,000 SNPs and re-estimate.
- Sensitivity of LDSC on various factors:
  - Low frequency variants: effects cannot be captured by LD scores. Thus they will drive up the slope.
  - Long-range LD: leads to underestimation of LD scores, and as a result, drive up the slope and/or intercept.
  - Very large effect variants: drive up the slope. "SNPs with very large effect sizes can result in large LD Score regression standard errors with an unconstrained intercept" because "linear regression deals poorly with outliers in the response variable" (from the cross-trait LDSC paper).

  All these SNPs should be filtered.
- Sensitivity of LDSC on LD matrices: (1) Difference of ref. and target LD is just noise (mean 0): increase intercept, and reduce slope. (2) Systematic difference: if LD in ref is smaller than LD in target on average, the intercept will be upward biased.
- Interpreting LD score results: Generally, if intercept is close to 0, no confounding. (Table 1) for the null SNPs, their mean χ^2 is the intercept term. It is very close to λ_{GC} when all SNPs are null (in null simulations). If the mean χ^2 of all SNPs is much larger than the intercept, it means that the results are mostly driven by polygenecity rather than confounding.
- Simulation on polygenic architecture: unrelated cohort of 1,000 Swedes. Intercept close to 1, and the estimates of h2 is unbiased at different levels of causal proportions. However, when the proportion is very low, the s.e. becomes very large.
- Simulation on confounding only: a useful result, under null model with population structure, the mean χ^2 is given by:
$$\bar{\chi}^2 = 1 + bNF_{ST} \tag{4.114}$$
  where b is the correlation of phenotype and ancestry, N is the sample size. To simulate confounding:
  - Continent level: choose one cohort as cases and another controls.
  - Country level: obtain PCs, and use first three PCs as phenotypes.

  In simulations, λ_{GC} is large, and is close to intercept, but the slope is close to 0.
- Simulation on polygenic and confounding: partition the chromosomes, with the first half containing causal SNPs, and the second half only null. Then simulation phenotypes: causal SNPs and environment correlated with PCs.
- Intuition of LDSC is PCs already capture LD, so SNPs in high LD vs. low LD regions are equal in PC-space. So their inflated effects are independent of LD.
- Remark: LDSC model holds even if prior effect sizes are not normally distributed.

Relationship between LD Score and Haseman-Elston Regression [Brendan Bulik-Sullivan, BiorXiv, 2015]

- HE regression: the covariance between two individuals depends on heritability and their genetic relationship:
$$\mathrm{E}\left(y_h y_i | X\right) = h_g^2 A_{hi} \tag{4.115}$$
where $A = XX^T/M$ is the GRM. This leads to a closed form estimator of h_g^2, which is the sample $\mathrm{Cov}\left(y_h y_i, A_{hi}\right)$ divided by sample $\mathrm{Var}\left(A_{hi}\right)$.

- Equivalence of HE regression and LDSC: the idea is that the numerator of the HE estimator can be related to the marginal SNP effect, and the denominator expressed as LD. So the HE regression is equivalent to LDSC with intercept constrained to 1 and regression weights $1/l$.

- Possible to incorporate fixed effects.

- Simulation results: $h^2 = 0.5$ on chr 2, 2000 samples. REML: no bias, SD = 0.05. LDSC with no intercept: bias = 0.02, SD = 0.06 (or MSE = 0.063). LDSC with intercept: bias = 0.02, SD = 0.09.

Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data (HESS) [Shi and Pasaniuc, AJHG, 2016]

- Treating effect sizes as fixed: e.g. suppose we are interested in $\sum_j \beta_j^2 = \beta^T\beta$, how can do estimate that when our power of estimating individual β_j is low? In general, we avoid estimation of individual β_j, but construct an estimator of $\beta^T\beta$.

- Strategy for estimating local heritability: treating genotype as random and effect sizes fixed. Given $y = X\beta$, we have:
$$\mathrm{Var}\left(y\right) = \beta^T \mathrm{Cov}\left(X\right)\beta + \sigma_e^2 = \beta^T V \beta + \sigma_e^2 \tag{4.116}$$
where V is the LD matrix. Assuming $\mathrm{Var}\left(y\right) = 1$, then $\beta^T V \beta$ is the heritability. Our problem is then to estimate this function defined on β with observed effect sizes and LD matrix.

- Estimator of local heritability: first obtain that $\hat{\beta} \sim N(V\beta, V(1-h^2)/n)$, where n is sample size. This leads to MOM estimator:
$$\mathrm{E}\left(\hat{\beta}\right) = V\beta \Rightarrow \hat{\beta}_{\mathrm{MOM}} = V^{-1}\hat{\beta} \Rightarrow (\beta^T V\beta)_{\mathrm{MOM}} = (V^{-1}\hat{\beta})^T V (V^{-1}\hat{\beta}) = \hat{\beta} V^{-1} \hat{\beta} \tag{4.117}$$
However, its expectation is not exactly $\beta^T V \beta$. With some correction, we obtain the unbiased estimator and its variance, Equations (5) and (6) in the paper.

- Dealing with LD matrix rank deficiency: Psuedo-inverse, V^{pi}, which effectively consider the q eigenvectors of the EVD of the LD matrix, if the rank is $q < p$, where p is the number of SNPs.

- Regularization of LD matrix when it's obtained from external reference samples: we perform EVD of the LD matrix, and the estimator can be rewritten in terms of the eigendecomposition as:
$$\hat{\beta}^T V^{pi} \hat{\beta} = \sum_{i=1}^{q} (1/w_i)(\hat{\beta} u_i)^2 \tag{4.118}$$
where w_i and u_i are eigenvalues and eigenvetors of V respectively. The interpretation is: it is the weighted sum of the square of projected effect sizes along eigenvectors. The term is dominated by the first k eigenvectors, so to regularize, consider only k eigenvectors for LD derived from reference samples.

- Simulation: 50K samples, chr. 22. When using external reference, biased estimator (as in LDSC), but $k = 30 - 50$ is optimal. When k is small, generally underestimate h^2: intuitively, we fail to capture all the true effects in a LD block.

- Discussion: reference-based LD, often estimated from smaller number of individuals, often miss subtle information in in-sample LD, so have lower ranks.

- **Lesson**: possible to deal with over-parameterization in frequentist statistics, by considering the problem of estimating some function defined over the parameters.

- **Lesson**: LD projected effect sizes. Ex. suppose we have two SNPs in strong LD, $\hat{\beta}_1$ and $\hat{\beta}_2$ would be both large and highly correlated. Now we project on eigenvectors (rotating the coordinates), then the second direction has small projected effect. Its contribution to heritability would be low.

A united framework for variance component estimation with summary statistics in genome-wide association studies (MQS) [Xiang Zhou, AAS, 2017]

- Motivation: REML is slow, require summary statistics and lead to down-ward bias in case-control data.

- Model and notation: let X_i be the genotype of all SNPs in category $i, 1 \leq i \leq k$ and y be phenotype. We assume they are all centered to mean 0. Our model of phenotype is:

$$y = \sum_{i=1}^{k} X_i \beta_i + \epsilon \qquad \epsilon \sim N(0, \sigma_{k+1}^2 M) \tag{4.119}$$

  where $M = I - 1_n 1_n^T / n$. The effect size $\beta_i \sim N(0, \sigma_i^2 I / p_i)$, where p_i is the number of SNPs in category i. With this model, we can marginalize β_i's. Let $g_i = X_i \beta_i$ be the total contribution of SNPs in category i:

$$g_i \sim N(0, \sigma_i^2 X_i X_i^T / p_i) = N(0, \sigma_i^2 K_i) \tag{4.120}$$

  where $K_i = X_i X_i^T / p_i$ is the $n \times n$ GRM from SNPs in the category i.

- MINQUE estimator: we have $k+1$ parameters σ^2 to estimate. Using MOM, we have, for any matrix A_j:

$$\mathrm{E}\left(y^T A_j y\right) = \sum_{i=1}^{k} \mathrm{tr}\left(A_j K\right) \sigma_i^2 + \mathrm{tr}\left(A_j\right) \sigma_{k+1}^2 \tag{4.121}$$

  It is easy to prove this equation using the result about quadratic form of random vectors. Note that RHS is amount of variation or covariance of y's (scalar). Intuitively, we can choose A_j to "extract" correlation of any two pairs of samples, then we have MOM estimation that matches sample covariance with expected covariance.

- Choosing A_j matrices: we choose A_j to minimize squared error (MINQUE criterion). The optimal A_j is given by Equation (5), however, it depends on the values of σ^2, which are unknown. Possible strategy: iterative update I-MINQUE, which is equivalent to REML. In order to use summary statistics, use A_j of the form:

$$\tilde{A}_j = X_j W_j X_j^T / p_j \qquad \tilde{A}_{k+1} = M \tag{4.122}$$

  where W_j is pre-specified $p_j \times p_j$ diagonal matrix. To simplify algebra, scale W_j s.t. the average weight is 1. Depending on the choice of W_j, this leads to either HE regression or LDSC.

- Equivalence to Hazeman-Elston regression: we choose W_j be identity matrix (equal weights to all SNPs). This leads to HE regression. Now we have $\tilde{A}_j = X_j X_j^T / p_j = K_j / p_j$. We have:

$$y^T K_j y = \sum_j = \sum_j \sigma_j^2 \mathrm{tr}\left(K_j^2\right) + \sigma_{k+1}^2 \mathrm{tr}\left(K_j\right) \tag{4.123}$$

Suppose $k = 1$, then LHS is $y^T KY = \sum_{i,j} K_{ij} y_i y_j$, and RHS is $\sigma_1^2 \text{tr}\,(K^2) + \sigma_2^2 \text{tr}\,(K)$. When phenotypic covariance matches genetic covariance, we have large value of $y^T Ky$, and this leads to large value of σ_1^2. Another way to understand this is: (ignoring constant p_j), we have

$$y^T \tilde{A}_j y = y^T X_j W_j X_j^T y = (X_j^T y)^T W_j (X_j^T y) = Z_j^T W_j Z_j \tag{4.124}$$

where Z_j measures covariance of X_j and y (Z-scores). When $W_j = I$, this term is now $Z_j^T Z_j$, which is the sum of Z-scores of all SNPs in category j. So if this value is large, it suggests that σ_j^2 is large.

- Estimator and confidence interval: we can solve the linear system analytically and the solution is given by $\hat{\sigma}^2 = S^{-1} q$, where S is roughly LD matrix, and q is k-dim. vector. See Equation (13) and (14). It is easy to obtain the variance of $\hat{\sigma}^2$ as a function of $V(q)$ and S. Use approximation to compute $V(q)$, Equation (18).

- Subsampling of S: exact computation of S takes $O(pn^2)$ which dominates computation. Approximate S using a subset of m samples (or external panel). Possible to account for sampling error of S, however, it is generally much smaller than $V(q)$.

- Using summary statistics: both q and S can be expressed as summary statistics: Z scores of SNPs, and pairwise LD. Two strategies: (1) use only summary statistics: require some additional summary statistics; (2) block jackknife strategy in LDSC. However this can work poorly if blockwise independence is not satisfied.

- Simulation results: comparison with BOLT-REML, LDSC, MQS-HEW (HE weighting) and MQS-LDW (LD weighting). LDSC not efficient. MQS-HEW and LDW are more efficient for small h^2 than large h^2, and works better when samples are independent. Type 1 error of testing if $\sigma^2 = 0$ is slightly inflated with normal asymptotic instead of mixture of chi-square.

Reevaluation of SNP heritability in complex human traits (LDAK) [Speed and Balding, NG, 2017]

- Background: under current models, GCTA and LDSC, heritability per SNP is constant.

- Background: LDSC tend to have standard error 25-100% higher because it has an extra parameters and it is moment-based.

- Motivation of dependency of SNP heritability on LD: Figure 1. High LD regions, tag fewer causal variants, comparing with low LD regions.

- LDAK prior: let h_j^2 be the heritability of SNP j, we assume:

$$\text{E}\,(h_j^2) \propto [f_j(1 - f_j)]^{1+\alpha} \times w_j \times r_j \tag{4.125}$$

where f_j is MAF, w_j weight of SNP which penalizes SNPs in high LD regions, and r_j genotype uncertainty. α is a parameter controlling the relationship of SNP heritability with MAF: constant effect means $\alpha = -1$.

- Strategy for comparing priors: (1) Log-likelihood of data: compare with different α and different LD models. (2) SNP partitioning: into bins (100 Kb segments) of MAF and LD, then estimate the heritability of each bin with GCTA.

- Effect of MAP on prior: Figure 2, LL of GWAS traits under different α, for most traits, it is maximized around $\alpha = -0.25$. However, the impact of using $\alpha = -1$ on PVE estimation is small.

- Effect of LD on prior: Figure 4, partition SNPs into low LD and high LD halves. Show that low LD half explains more than 50% heritability and is more consistent with LDAK model than GCTA for most of the traits. Figure S12: generally, quantitative traits show less effect of LD on SNP heritability.

- Results: in 19 traits, h^2 on average 43% higher than GCTA. Also, DHS only explains 24% heritability (instead of 79%).

Linkage disequilibriumdependent architecture of human complex traits shows action of negative selection [Gazel and Price, NG, 2017]

- Background: h2g enriched in DHS, etc, which have low LD. On the other hand, regions of low recombination rate, and high LD, are enriched with exonic deleterious variants (opposite).
- Levels of LD (LLD): low LLD regions have higher heritability (Figure 1). Use S-LDSC with continuous annotations.
- Explanation: driven by more recent common variants having lower LLD (positive correlation of age and LLD): the youngest 20% of common SNPs explain 3.9 times more heritability than the oldest 20%. For fixed set of variants, older alleles have low LLD (negative correlation), however, for new variants, we have positive correlation. Imagine a background of many variants, now new variants appear: the youngers one are generally in lower LLD with existing ones (positive correlation).
- Recombination and h2g: low recombination rates, selection less effective (BGS reduces population sizes), so variants more likely to be deleterious.
- Gazel LD annotations: MAF adjusted allele age, MAF adjusted LLD in African, and other LD related annotations, B-scores, nucleotide diversity, etc.
- Summary: both low LLD and low recombination rates can be associated with higher per SNP h2g, for different mechanisms.

Better estimation of SNP heritability from summary statistics provides a new understanding of the genetic architecture of complex traits (SamHer) [Speed and Balding, Biorxiv, 2018]

- MOM estimation of heritability with LDAK prior: $\mathrm{E}\left(h_j^2\right) \propto q_j$, where q_j is given by LDAK. Let Z_j^2 be the chi-square statistic of SNP j, we modify the LDSC equation by:

$$\mathrm{E}\left(Z_j^2\right) = 1 + n_j(h_j^2 + \sum_{l \in N_j} r_{jl}^2 h_l) \tag{4.126}$$

  We plug in the LDAK prior, and normalize per SNP heritability:

$$\mathrm{E}\left(Z_j^2\right) = 1 + u_j h_{\mathrm{SNP}}^2 \text{ where } u_j = \frac{q_j + \sum_{l \in N_j} q_l r_{jl}^2}{Q} \tag{4.127}$$

  where $Q = \sum_j q_j$ is a normalizing constant. Estimation is done by MOM, with weighting of SNPs and estimation of standard error similar to LDSC.
- Estimating confounding bias: use $\mathrm{E}\left(Z_j^2\right) = C(1 + 1 + u_j h_{\mathrm{SNP}}^2)$, and estimate C.
- Comparing heritability models: use log likelihood of the SNPs assuming a diagonal matrix.
- Enrichment: generalization of stratified LDSC.
- Results of enrichment analysis: conserved regions are highly enriched (13 fold) for GWAS effects in LDSC, but only 1.7 fold by SamHer.
- Prediction of risks: very modest (1 or 2%) improvement of prediction accuracy (correlation) using LDAK vs. GCTA.

Reconciling S-LDSC and LDAK functional enrichment estimates [Gazel and Price, NG, 2019]

- LDAK model: use different ways of modeling LD. Baseline-LD: baseline annotations and LD related annotations. Gold standard: baselineLD and LDAK.
- Comparison of likelihood: LDAK lower than Gazal-LD and baseline-LD. Remark: not very comparable if the number of parameters are different.
- Simulation to assess enrichment estimates: S-LDSC + LDAK gives robust estimates (unbiased).
- UK biobank traits: LDAK consistently underestimate enrichment of functional annotations, because it assigns 0 h2g to 85% of SNPs.
- Baseline annotations: from [Finucane, NG, 2015]. DHS, H3K27ac, etc.: union of ENCODE/Roadmap annotations from all cell types.

Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture [Hou and Pasaniuc, NG, 2019]

- Background: (1) multi-component REML by MAF and LD stratification is very resource intensive, and "it is unclear whether multi-component methods based on summary statistics produce accurate estimates of total SNP-heritability". (2) Models that explicitly model MAF-LD dependency: sensitive to model mis-specification.
- GRE estimator: $\beta_i \sim N(0, \sigma_i^2)$. The goal is to estimate h_g^2, defined as PVE by all variants, or $\text{Var}(x^T\beta)/\text{Var}(y)$. Assuming y is normalized with variance 1, then h_g^2 is just $\text{Var}(x^T\beta)$. Use Law of Total Variance to partition this:

$$\text{Var}(x^T\beta) = \text{E}(\text{Var}(x^T\beta|\beta)) + \text{Var}(\text{E}(x^T\beta|\beta)) = \text{E}(\beta^T\text{Var}(x^T)\beta) + \text{Var}(\text{E}(x^T)\beta) \tag{4.128}$$

  It is easy to see that the second term is 0, and let $V = \text{Var}(x^T)$ be the LD matrix. So we have:

$$h_g^2|\beta = \beta^T V \beta \tag{4.129}$$

  where we treat β as given. Of course β is not given, so marginalizing β gives $h_g^2 = \sum_{i=1}^M \sigma_i^2$.
- Deriving GRE estimator: first we consider β as fixed. We know from RSS the distribution of $\hat{\beta}|\beta$, so we use $\hat{\beta}$ to obtain the unbiased estimator of $\beta^T V\beta$. Roughly, we should use $V^{-1}\hat{\beta}$ to approximate β, so we have the estimator $\hat{\beta}^T V^{-1}\hat{\beta}$. This estimator is not unbiased, but with modest change, we have the unbiased MOM estimator (GRE estimator) as:

$$\hat{h}_{GRE}^2 = \frac{N\hat{\beta}^T V^{-1}\hat{\beta} - M}{N - M} \tag{4.130}$$

  where N is sample size and M number of variants.
  Proof: see Equation (1)-(3) of the paper. Briefly, treating $\hat{\beta}$ as RV, we have $\text{E}(\hat{\beta}) = V\beta$ and $\text{Cov}(\hat{\beta}) = \sigma_e^2 V/N$. Then we use the result of the expectation of Quadratic form of a random vector to obtain $\text{E}(\hat{\beta}^T V^{-1}\hat{\beta})$. It is given by:

$$\text{E}(\hat{\beta}^T V^{-1}\hat{\beta}) = \frac{M}{N}\sigma_e^2 + \beta^T V\beta \tag{4.131}$$

  We note that $\sigma_e^2 + \beta^T V\beta = 1$, plug in and we can solve the GRE estimator.
  While we assume β as fixed effect, it is easy to show that the estimator, when treating β as random, is still unbiased.
- Analysis: do we require $N > M$? The paper said, when $M > N$, we get negative estimates. However, when this happens, both denominator and numerator of the GRE estimation Equation 4.130 are negative. And it's easy to plug in the expectation of $\hat{\beta}^T V^{-1}\hat{\beta}$ to see that $N - M$ term cancels out. The real reason that GRE works well in large samples is probably related to estimation error of GRE. In fact, if we ignore the LD, then the estimator is roughly $\sum_j \hat{\beta}_j^2$. So it does not do any shrinkage, and with large sample size, would be unbiased. At smaller N, however, the variance of $\hat{\beta}_j$ is large.

- Genome-wide approximation: $N > M$ may not always satisfy, so we divide the genome into blocks, and estimate h_g^2 for each block then add. Let p_k be the number of SNPs in block k, then we should use p_k instead of M in the equation above.
- Dealing with rank deficient LD matrix: We should use pseudoinverse instead, and also use q_k, the rank, instead of number of SNPs in block k. So our final formula is:

$$\hat{h}^2_{GRE} = \sum_k \frac{N\hat{\beta}_k{}^T \hat{V}_k^{-1} \hat{\beta}_k - q_k}{N - q_k} \tag{4.132}$$

- Simulation setting: test a number of different genetic architectures: different proportion of causal variants and MAF and how per SNP h2g depends on MAF and LD. The genetic architecture is defined as (assuming genotypes are standardized):

$$\sigma_i^2 \propto c_i w_i^\gamma [2f_i(1-f_i)]^{1+\alpha} \tag{4.133}$$

  where c_i is causal status of SNP i, w_i the LD weight and f_i the AF. With $\gamma = 0$: no dependency on LD, and $\gamma = 1$: higher LD scores imply smaller effects. For α: if $\alpha = -1$, it is the standard normal prior. Recommended value: $\alpha = -0.25$ by LDAK.
- Simulations at large sample size: N =400K, 0.5M variants. GRE is unbiased under various scenarios (Fig. 2). LDSC, S-LDSC and SumHer all have bias under various architecture.
- Simulations at small sample size: N =8K, 14K variants. Still unbiased, but large variations comparing with GREML (Fig. 3). LDAK may have large bias under some settings. The best method, GREML with different components, stratified by MAF and LD (GREML-LDMS-I).

4.4.1 Heritability from Family Studies

Relatedness disequilibrium regression estimates heritability without environmental bias [Young and Kong, NG, 2018]

- Background: (1) Kinship method: correlate genetic correlation (GRM by expected kinship) (2) Sib-regression: use random deviation from 50% sharing, free from environment.
- Intuition: relatedness of two individuals i and j (IBD sharing), $[R]_{ij}$, is largely determined by the relatedness of their parents, $[R_{\text{par}}]_{ij}$. However, due to random segregation, $[R]_{ij}$ may deviate from its expectation, $[R_{\text{par}}]_{ij}/2$. This deviation can causal additional phenotypic covariance between i and j, and allow one to estimate heritability (higher heritability, higher additional covariance).
- Variance decomposition: let Y_i be the phenotype of sample i, and g_i be direct genetic effect, p_i be the environmental effect correlated with parental genotypes, including indirect genetic effect, and e_i be the residual environmental effect (not correlated with parental genotype). We have:

$$Y_i = g_i + p_i + e_i \tag{4.134}$$

  Note that g_i and p_i are correlated. So the variance is:

$$\text{Var}(Y) = \text{Var}(g_i) + \text{Var}(p_i) + \text{Cov}(g_i, p_i) + \text{Var}(e_i) \tag{4.135}$$

  The terms corresponding to Eq (1) in the paper.
- Covariance decomposition: We consider two subjects with $Y_i = g_i + p_i + e_i$ and $Y_j = g_j + p_j + e_j$. Covariance has these components: $\text{Cov}(g_i, g_j)$ which relates to R_{ij}; $\text{Cov}(p_i, p_j)$ which relates to $[R_{\text{par}}]_{ij}$; $\text{Cov}(g_i, p_j)$ and $\text{Cov}(g_j, p_i)$, which relates to $[R_{\text{o,par}}]_{ij}$ (individual i vs. parent of individual j).

- RDR regression: to fit the model, we regress cov. between Y_i and Y_j vs. R_{ij}, $[R_{\text{par}}]_{ij}$ and $[R_{\text{o,par}}]_{ij}$. The reason that it is robust to environment effects is: the coefficient of R_{ij} must be due to random Mendelian segregation since we control $[R_{\text{par}}]_{ij}$.

- Simulation results: compare with kinship and sib-regression methods. Kinship is sensitive to genetic nurturing, maternal environment. Sib-regression has large s.e., and sensitive to epistasis and dominance.

- Analysis: in GWAS, SNP effect should be the sum of direct and indirect genetic effects. Can we disentangle the two so that we can learn which variants may act indirectly? Let P_i be parental genotype, G_i be genotype of sample i and Y_i be phenotype, then we have:

$$Y_i = \beta G_i + \gamma P_i + \epsilon_i \tag{4.136}$$

  where β, γ are direct and indirect effects, respectively. We can thus learn the separate effects when P_i is available (measured or imputed from relatives).

4.4.2 Methods of Linear Mixed Model

EMMAX: Variance component model to account for sample structure in genome-wide association studies [Kang & Eskin, NG, 2010]

- Goal: testing individual SNPs while correcting for population substructure and cryptic relatedness.

- Mixed effect model: suppose we are testing one SNP at a time, let it be k. For th i-th individual, let $x_{i,k}$ be the SNP's genotype, and β_k be its effect size. The phenotypic trait of the individual is:

$$y_i = \beta_0 + \beta_k x_{i,k} + \eta_{i,k} \tag{4.137}$$

  where $\eta_{i,k} = \sum_{s \neq k} \beta_s x_{is} + \epsilon_i$ is the total effect of genetic background and environment. In general, the effect of a single SNP is small, so $\eta_{i,k} = u_i + \epsilon_i$, where u_i is a random effect, as in Equation 4.87.

- Procedure: three steps:

  - Estimation of kinship matrix $\hat{S}_N$.
  - Estimation of σ_a^2 and σ_e^2 using the random effect model, Equation 4.87. Test if $\sigma_a^2 = 0$.
  - If $\sigma_a^2 \neq 0$: the model is reduced to linear model with dependent error term:

$$y_i = \beta_0 + \beta_k x_{i,k} + \eta_i \tag{4.138}$$

    where $\text{Var}(\eta) \propto \hat{\sigma}_a^2 \hat{S}_N + \hat{\sigma}_e^2 I$.

FaST linear mixed models for genome-wide association studies [Lippert & Heckerman, NM, 2011]

- Model: following the standard LMM, we have

$$y \sim N(X\beta, \sigma_g^2 K + \sigma_e^2 I) \tag{4.139}$$

  where σ_g^2 is the genetic variance and σ_e^2 environmental variance, K is the kinship matrix. The challenge is that the likelihood function involves determinant and inverse of the covariance matrix, which is computationally expensive.

- Computational speedup: let $\delta = \sigma_g^2/\sigma_e^2$, then we write $\sigma_g^2 K + \sigma_e^2 I = \sigma_g^2(K + \delta I)$. Our idea is to factorize the matrix s.t. we do not have to recompute determinant and inverse of this matrix each time

we update δ (the main parameter). We use the Spectral Decomposition of K, $K = USU^T$, where S is diagonal and U is orthogonal. Then:

$$K + \delta I = USU^T + \delta I = U(S + \delta I)U^T \tag{4.140}$$

The determinant of this matrix:

$$\det(U(S + \delta I)U^T) = \det U \cdot \det(S + \delta I) \cdot \det U^T = \det(S + \delta I) \tag{4.141}$$

where we use the fact that $\det U \det U^T = 1$ since U is orthogonal. The inverse of the covariance matrix:

$$[U(S + \delta I)U^T]^{-1} = U(S + \delta I)^{-1}U^T \tag{4.142}$$

Note that S is diagonal, so both determinant and inverse of $S + \delta I$ is a simple function of δ. Let $S = \text{diag}(S_{ii})$,

$$\det(S + \delta I) = \prod_i (S_{ii} + \delta) \qquad (S + \delta I)^{-1} = \text{diag}\left(\frac{1}{S_{ii} + \delta}\right) \tag{4.143}$$

- Equivalent linear regression model: with the spectral decomposition, we can show that the original regression model (where errors are correlated) has the log-likelihood model that is exactly the same as the following model where errors are independent:

$$U^T y = U^T X\beta + E, \qquad E \sim N(0, \sigma_g^2(S + \delta I)) \tag{4.144}$$

  Note that the covariance matrix of E is diagonal. So if we rotate the data y, X with U, then the new regression model is just a weighted linear regression. The log-likelihood is then simply weighted sum of squared error. See Equation 2 of the paper.

- Optimization: first solve β and σ_g^2 in terms of δ (closed form); then do 1D optimization of δ.

- Running time analysis: suppose we test s SNPs with sample size n. Then computation involves Spectral Decomposition $O(n^3)$, data rotation for each of s tested SNPs, $O(n^2 s)$, and optimization for each SNP $O(Cns)$. If we fix δ (do not restimate for each SNP), then the optimization step takes $O(Cn)$.

- When K is low-ranked (k): we can exploit that to further speed up the computation. Roughly we have $O(n^2 k)$, $O(nks)$ and $O(C(n + k))$ for each of the three steps.

SNP Set Association Analysis for Familial Data [Schifano & Lin, GE, 2012]

- Motivation: SKAT (SNP-set test) does not account for family structure. Use a random effect model to dependency of relative.

- Model: let s_{ij} be the SNP set (r SNPs) of the j-th indiviudal in the i-th family, x_{ij} be the covariates and y_{ij} be the phenotype of the individual i, j. The phenotype can be written as a mixed effect model:

$$y_{ij} = x_{ij}\alpha + h(s_{ij}) + b_{ij} + \epsilon_{ij} \tag{4.145}$$

  where $h(s_{ij})$ is the random effects of SNPs (each SNP has a random effect), and b_{ij} the random effect due to family background. The genetic random effects are correlated among family members, and we assume $b_i = (b_{i1}, \cdots, b_{in_i})^T \sim N(0, 2\Phi_i\sigma_b^2)$, where Φ_i is the kinship matrix. One can then form the score test.

Efficient Bayesian mixed-model analysis increases association power in large cohorts (BOLT-LMM) [Loh and Price, NG, 2015]

- Why LMM? Even if our goal is to perform association test (per SNP), there are two advantages: (1) Correcting for population structure, including cryptic relatedness. (2) By adjusting for all other SNPs, increase the power.

- LMM for association test: suppose we test one SNP a time, for a SNP with genotype x,

$$y = x\beta + g + e \tag{4.146}$$

where g is the genetic effect and e the error. We use a random effect model for g. Note that g should not include SNPs in LD with the test SNP. Using the variance component approach, we have:

$$V = \text{Cov}(y) = \sigma_g^2 K + \sigma_e^2 I \tag{4.147}$$

where K is the GRM. Now this becomes a problem of Generalized Least Square (GLS), where the error term are correlated, or write in MVN: $y \sim N(x\beta, V)$. Using GLS result, we have:

$$\hat{\beta} = (x^T V^{-1} x)^{-1} (x^T V^{-1} y) \tag{4.148}$$

We can think of $V^{-1}y$ as the phenotype, corrected for dependency between samples. We can test if $\beta = 0$ using the score test, and the test statistic is:

$$T = \hat{\beta}^2 / \text{Var}(\hat{\beta}) = (x^T V^{-1} y)^2 / (x^T V^{-1} x) \tag{4.149}$$

It follows chi-square distribution with df 1. Reference: Rapid variance componentsbased method for whole-genome association analysis, NG, 2012

- BOLT-LMM: replace normal prior of effect size with mixture of normal prior, inference of the posterior mean of effect size for each SNP, and regress out the predicted phenotype when testing each SNP. When test each SNP, remove all SNPs in the same chromosome in modeling the random effects.

- Step 1: estimating the hyperprior parameters and the genetic effects. Major components:
  - First fit a standard variance component model to estimate PVE.
  - With a two mixture of normal prior (three parameters), constraint it using the total PVE, so there are two free hyperprior parameters. For each parameter combination, inference is based on cyclic coordinate descent: for each SNP, update its effect conditioned on the estimated effects of all other SNPs.
  - Use cross-validation (based on prediction performance) to select among 18 parameter combination.

  The result is residual phenotype (for each test SNP), regressing out the remaining chromosome.

- Step 2: association testing. This involves slight modification of Equation 4.149 above. We replace y in the equation with the residual phenotype obtained from Step 1.

- Q: if we explicitly regress out residual phenotype, why do we still need the GLS, which treats the residual as random?

4.4.3 Heriability Studies

Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder [Purcell, Nature, 2009]

- Problem: assess the contribution of common variants of small effects to a complex disease.

- Defining scores (page 19-20 of Suppl): divide the data into a discovery set, and a target set (the paper used males as discovery and females as target). From the discovery set, obtain the loci using a liberal threshold, e.g. $p < 0.1$ - these are called "score alleles" (instead of risk alleles). Then assess the collective contribution of these score alleles in the target set. The score is defined as: "total score for

each individual as the number of score alleles weighted by the log of the odds ratio from the discovery sample". Formally, for each individual i in the target samples, its score is defined as:

$$S_i = \sum_j x_{ij} \cdot \log \lambda_j \tag{4.150}$$

where the summation is over all score alleles, λ_j is the OR of the allele j in the discovery set, and x_{ij} the genotype of j in subject i.

- Testing for association between score and disease in the target sample (page 21): regression in the target sample, estimate R^2 by comparing a model including the score and covariates versus a model including only the covariates (needed for controlled population stratification).
- The relation between score and disease risk [Wray, Prediction of individual genetic risk to disease from genome-wide association studies, GR, 2007]: the risk of getting affected in individual i with genotype G_i is given by

$$P(D_i|G_i) = f_0 \prod_j \lambda_j^{x_{ij}} \tag{4.151}$$

where f_0 is the baseline penetrance.

Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits [Zaitlen & Price, PLoS Genet. 2013]

- Problem: in a dataset with some related and some unrelated invididuals, how do we estimate heritability?
- Model idea: define a threshold based on GRM, when $IBS > t$, consider them related, we use IBD estimated from haplotype to obtain h^2; when $IBS < t$, we use GRM to obtain h^2. The statistical problem is: we cannot simply partition all pairs of individuals into related and unrelated, so our model should account for that.

Most genetic risk for autism resides with common variations [Gaugler & Buxbaum, NG, 2014]

- Background: the relation between explained variance, effect size/relative risk and allele frequency. Suppose we have a mutation/variant with known RR and AF, how much variance of the liability is explained by this variant in the population? We know that the proportion of variance explained (PVE) is given by (from linear model):

$$R^2 = \frac{\sum_j \beta^2 \mathrm{Var}\,(x_j)}{\mathrm{Var}\,(y)} = \frac{\beta^2 p(1-p)}{\mathrm{Var}\,(y)} \tag{4.152}$$

In liability model, we often assume y is normalized, i.e. variance equal to 1. Next, the effect size β is related to RR by this equation:

$$\mathrm{RR} = \frac{1 - \Phi(t - \beta)}{1 - \Phi(t)} \tag{4.153}$$

where t is the threshold (determined from the baseline penetrance). For autism, assume baseline penetrance of 1/68, we have $t = 2.18$.

- Overall heritability: using a large sample from Sweden, 1.6M families, including 14K ASD cases. Estimate heritability from recurrence in MZ twiwns to first cousins: 54%.
- Heritability due to common variants: 3K subjects include 466 cases and 2.5K controls, use GCTA to estimate, remove all related samples, result 49.4%.
- Heritability due to both common and rare variants: include more closely related individuals, result 52.4%.

- Heritability due to non-additive effects: 3%.
- Variance of liability due to de novo mutations: 2.6%. For example, for LoF mutations, we use RR = 2.42 and exposure (q) 0.053, we obtain the variance explained by LoF as 1.1% (my own calculation 0.7%). For missense mutations, we use RR = 1.1 and exposure 0.397, we have the explained variance 0.04%.
- Question: how to estimate the heritability due to common and rare variants?
- Remark: different studies could obtain very different estimates of heritability. Ex. for common variants, the estimate from PGC is only 17%. The general issues of consideration for estimating heritability: ascertainment, prevalence (could be different in different samples/populations), control of environment.

Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases [Gusev & Price, AJHG, 2014]

- Partitioning heritabilty in the presence of LD: suppose we have multiple categories of variants (similar to chromosomes), and we want to partition h^2 into categories, but variants in different categories could be in LD. We use similar partition when the variants are not in LD:

$$y = \sum_i \sum_{s \in S_i} W_s \beta_s^i + e \tag{4.154}$$

  where i is the i-th category, S_i the set of all variants in i, W_s is the genotype of SNP s, and β_s^i the effect size. The random effect model: $\beta_s^i \sim N(0, \sigma_i^2)$ and σ_i^2 is category-specific effect size parameter. We could do the joint estimation similar to partition into chromosomes. Under this model, LD is taken into account: whenever we estimate the contribution of i-th category, it is conditioned on all other categories, automatically accounting for LD.
- Estimating enrichment of disease loci from h^2 estimation: the ratio of h^2 from a category i vs. the ratio of category i in terms of number of SNPs defines the enrichment of i. It is equivalent to the the relative risk that a SNP in category i is causal in comparison to an average SNP.
- Estimating enrichment of disease loci from summary statistics: count the enrichment of number of SNPs above a threshold vs. random expectation.
  - Remark: this estimation does not account for the power of studies, i.e. even if a SNP is causal, it may not reach the p-value threshold. So the enrichment estimated in this way will always be lower than the true enrichment.
- Annotation data: 6 primary categories, coding, UTR, promoter, DHS in 217 cell types (only distal ones), intronic and intergenic.
- Simulation studies: assume 10% of SNPs are causal, randomly sample them from one category only, and see if the enrichment methods can recover the category. Results:
  - h^2 based enrichment always recovers the true category, however, p-value based enrichment fails for DHS (large categories). Explanation: p-value enrichment is less sensitive, and the low LD in DHS region also makes it hard to detect enrichment.
  - Using imputed SNPs always better than only genotyped SNPs.
  - When all causal variants are in coding regions, UTR and promoters are still enriched in p-value approach but not h^2 approach.
- Enrichment analysis in 11 common diseases (WTCCC) using heritability and p-values: variants with MAF $> 1\%$, using h^2 enrichment, DHS is 5.1x enriched (79% of h^2) and coding sequence is 13.8x, or 8% of h^2 (Figure 3.). Using p-value enrichment, depending on the p-value cutoff, the enrichment of coding is highest (around 2 at stringent p-value), and the enrichment of DHS is very modest, around 1.1-1.2 at very stringent p-values.

- Contribution of enhancers: show that for DHS, a large fraction of heritability is due to enhancers.
- Lessons:
  - Enrichment of causal variants can be significantly underestimated when the study power is not taken into account: with the p-value approach, at a given threshold, many true loci will not pass the threshold, leading to an underestimation.
  - How to control for LD: could be a problem, e.g. promoters and coding sequences may be in close LD, and thus can be difficult to separate out the effects using p-value approach (or simple regression in general). The strategy is to include them in a single regression model, thus adjusting for other variants in LD.
  - Enrichment of causal variants: can be done using heritability or PEV (proportion of explained variance) under a linear model (instead of logistic regression). The results can be used as a prior for Bayesian anaysis of functional variants.

Heritability estimation: comparison of GCTA vs. Sparse prior [Xiang Zhou, UMich, 2019]

- H2g estimation: (1) Low polygenicity: normal prior has large s.e., and sparse prior works very well (small s.e.). (2) Mid to high polygenicity: normal prior is unbiased, but sparse prior has downward bias, because sparse prior will miss many effects.
- Prediction: biobank scale. Large effect SNPs, treated as fixed effects, then the rest with random effects (normal prior).

4.4.4 Polygenic Prediction

Disease classification from genotype [Jakobsdottir & Weeks, PG, 2009]

- Two basic ways of measuring a set of SNPs for disease association/diagnosis:
  - Risk analysis: typically in a setting where multiple SNPs are identified, and the SNPs are assessed by the increased risk/odds of the (combined) SNPs.
  - Classification performance: suppose we use a genotype G to classify: i.e. declare case if the genotype is G; declase not if the genotype is not G. Let the disease status be A and U (affected, or unaffected). The performance is measured by two quantities: true positive and false positive fractions:
    $$\begin{aligned} TPF &= P(G|A) \\ FPF &= P(G|U) \end{aligned} \tag{4.155}$$
    Thus the performance of the classifer is determined by the genotype frequencies in the cases and controls.
- Relation between odds ratio (OR) and classification performance:
  - Use the relation between odds and genotype frequencies (G_0 is the reference genotype):
    $$OR = \frac{\text{odds}(G)}{\text{odds}(G_0)} = \frac{P(G|A)/P(G|U)}{P(G_0|A)/P(G_0|U)} \tag{4.156}$$
    Note that $P(G_0|A) = 1 - TPF$ and $P(G_0|U) = 1 - FPF$. Solving this equation for a given OR leads to the ROC of performance (only one of TFP or FPF can be fixed).
  - Analysis: as an approximation, we assume the reference genotype does not increase disease risk and equally distributed in cases and controls, then $\text{odds}(G_0) = 1$, and $P(G_0|A)/P(G_0|U) = 1$, then we have:
    $$\frac{P(G|A)}{P(G|U)} \approx OR \tag{4.157}$$

Even for large OR, its predictive power depends on the genotype frequencies (in control): (1) common alleles: large $P(G|U)$, thus high FPR, and high TPR; (2) rare alleles: small $P(G|U)$, thus low FPR, but also low TPR. Ex. AUC = 0.76 for a huge OR of 50 (e.g. at $FPF = 2.4\%$, the TPF is only 55%).

- Why low risk alleles are not good predictors of diseases? In most GWAS studies, the alleles have OR below 2. We consider an allele with relatively low frequency in the control (if too common, FPR would be too high), say, $P(G|U) = 0.1$, and $OR = 2$, then according to our approximation, $P(G|A) = 0.1 \times 2 = 0.2$, i.e. the sensitivity is only 20%.
- Why multiple low risk alleles may not be sufficient for disease classification? Suppose we have two risk alleles with $P(G|U) = 0.2$ and $OR = 2$ for both, and we follow AND rule for classification (OR rule would give too many false positives). With single risk allele: $FPR = 0.2, TPR = 0.4$; with both alleles: $FPR = P(G_1 \wedge G_2|U) = 0.2 \times 0.2 = 0.04$, and $TPR = P(G_1 \wedge G_2|A) = 0.04 \times 2^2 = 0.16$. In general, AND rule would lead to low TPRs.

- Classification performance for several common diseases: e.g. in AMD data set, using an additive model of three variants (highly significant association) leads to AUC at 0.79. In T2D, a model of 12 SNPs has AUC of only 0.64.

Multi-ethnic polygenic risk scores improve risk prediction in diverse populations [Mrquez-Luna & Price, review for PLG, 2017]

- Motivation: how do we predict phenotype in a population with small sample size? There is larger data from a different population, but the LD patterns are different, which could reduce the accuracy.

- Background: Polygenic risk score (PRS), usually do LD pruning and thresolding.

- Analysis: assume no change of true causal variants and their effects. Consider a region, and let i be its causal variant. Suppose in the training data, j is the top SNP (could be equal to i) and its observed effect $\hat{\beta}_j$. In the target data, the true phenotype of an individual is $y = x_i\beta_i$, and the predicted: $\hat{y} = x_j\hat{\beta}_j$. We take expectation, and use the RSS model (assume the same standard errors in i and j): $\mathrm{E}\,\beta_j = R_{ij}\beta_i$. We have:

$$\mathrm{E}\,(\hat{y}) = x_j R_{ij}\beta_i \qquad y = x_i\beta_i \tag{4.158}$$

  In the target data, LD between x_i and x_j is R'_{ij}. We note that the different LD pattern could cause problem: suppose R_{ij} is large, but R'_{ij} is small, then even if j is a good proxy of i in the training data, it is not in the testing data.

- Analysis: the impact of ancestry. Ancestry could be correlated to phenotype, and adding it could improve accuracy. Intuitively, in the target population, because of random drift, j might no longer be a good proxy of i. But the overall ancestry may still be correlated with many causal variants.

- Model: train effect size model using both popluations. Then use the weighted average of PRS scores from the two effect size models. The weights are learned by (1) use the validation data, and compute adjusted R^2 to account for additional dof (when comparing methods in the validation data). (2)cross-validation.

- Model: incorporating ancestry. Similar idea, use linear combination of PRS and the top PC. The weight parameters are learned from validation data.

- Simulation study: use EUR and LAT data from WTCCC2. Show that EUR and LAT have somewhat similar performance even though EUR sample size is much larger. EUR + LAT improves upon both, and EUR + LAT + PC further improves by 10% or so.

- T2D risk prediction: similar to simulations. The PC part has little improvement over EUR + LAT, probably due to the fact that EUR effect sizes already captured most of the genetic ancestry.

Improved polygenic prediction by Bayesian multiple regression on summary statistics [Lloyd-Jones and Visscher, NC, 2019]

- Model: similar to RSS, let b be the estimated effect sizes, and β be the true effects, the model is given by:

$$b = D^{-\frac{1}{2}} B D^{\frac{1}{2}} \beta + D^{-1} X^T \epsilon \tag{4.159}$$

  where D is diagonal matrix of $x_j^T x_j$, and $B = D^{-\frac{1}{2}} X^T X D^{-\frac{1}{2}}$. Using ASH prior for β_j. The model makes inference with Gibbs sampling. The conditional distribution of β_j given others is in page 38 of Supplement. Let δ_j be the class indicator of β_j, and define $l_{jc} = x_j^T x_j + \sigma_e^2/(\gamma \sigma_\beta^2)$, then

$$f(\beta_j | \delta_j = c, \theta_{-\beta_j}, y) \propto \exp\left[-\frac{1}{2}\frac{(\beta_j - \hat{\beta}_j^2)}{\sigma_e^2/l_{jc}}\right] \tag{4.160}$$

  where $\hat{\beta}_j$ is the posterior mean of β_j. It is given by $\hat{\beta}_j = x_j^T w / l_{jc}$, where w is the residual regressing out all other parameters except β_j. In the actual model, sample δ_j first, marginalizing all parameters (PIPs), then sample β_j.

- Right-hand site (RHS) update: note that to in each update, need to only compute $\hat{\beta}_j$. The computation of residual w is expensive, however, we only need $r_j = x_j^T w$, so we only need to store $r^* = X^T y - X^T X \beta$, which is $p \times 1$ vector. Once we have r^*, we can compute r_j as:

$$r_j = r_j^* + x_j^T x_j \beta_j \tag{4.161}$$

  Note that computation of r^* is faster and requires much less memory: since $X^T y = Db$ is proportional to b (summary statistics), and $X^T X$ is just the LD matrix, which is local/sparse. At each iteration, once we update β_j using the current r_j, we can update r^* as:

$$\Delta r^* = X^T x_j \Delta \beta_j \tag{4.162}$$

4.5 Gene-Gene and Gene-Environment Interactions

Reference: epistasis [Cordell, NRG, 2009; Carlborg & Harley, NRG, 2004]; multi-locus methods [Hoh & Ott, NRG, 2003; Onkamo & Toivonen, Human Genomics, 2006]

Discussion with Margit Burmeister [Umich, 2019]

- Examples of gene-environment interactions: a variant about additive behavior: in smokers, decrease BMI, because it makes one more additive to smoking, which reduces appetite. But it increase BMI in non-smokers.
- Confounding of genetics and environment: e.g in UK, some areas, low SE status, and different genetic groups. So genetics/regions/environment become confounded.
- Taste preference: highly genetic, e.g. broccoli bitterness. This may influence the risk of other traits, e.g. BMI (because one avoids bitter vegetables).
- Testing gene env. interactions via PRS and MR: e.g. PRSs of taste or smoking, and use as a surrogate for env.
- Remark: big question is how gene-environment interaction works? Is mediation by env. a common mechanism?

Multi-locus methods: motivations and benefits

- Regional test: use all information in a region (a gene) to test association, this is a generalization of haplotype-based tests.
- The SNPs may have non-linear interactions s.t. single SNP-based test may not be sensitive to detect them.
- Fine mapping: in a candidate region, multiple SNPs may be associated with the trait because of LD. Using multi-locus methods may help to identify the causal variants (removing the effect of correlated, but non-causal SNPs).

Background on epistatis:

- Definition of epistasis: this is broadly defined as the effect of one allele on the phenotype depends on the allele at another locus. Example: the coat color of pig: the effect of the allele at the MC1R locus is dominated by the KIT locus, i.e. MC1R allele expresses phenotype only with a particular allele in the KIT locus.
- Biological basis of epistatis: many possible mechanisms of epistasis, including (but not limited to):
  - Redundancy: e.g. two proteins play similar function, thus only when mutating both proteins, an effect can be seen.
  - Dominance/masking: e.g. in a biochemical pathway, one enzyme is rate limiting, thus mutation of another enzyme is invisible unless the rate limiting enzyme is also mutated.
  - Synergism: e.g. two proteins may be synergistic (in signal transduction or gene regulation), thus mutating two proteins has a effect similar to mutating only one of them.
  - Network behavior: the positive and negative feedbacks in the network create non-linearity, e.g. negative feedback may make the system robust to single mutations, but not to double mutations.
- Ubiquitous nature of epistasis:
  - From the study of QTL: the proportion of the genetic variance in F2 that results from epistasis ranged from 0 to 81% with a mean of 38% for the 18 traits studied.
  - Difficulty of detecting epistasis: from the perspective of genotype partitioning, epistatis is detected from the frequencies of genotype combinations, which are exponentially lower than the frequencies of single genotype. Therefore, considerably larger sample size is generally needed.

Testing for epistasis/interactions: two basic ideas of statistical interaction

- Disease risk perspective: the disease risk of the genotype combination is not the same as the sum of the risks of individual genotypes.
  - Genotype paritioning: e.g. multifactor dimensionality reduction (MDR) method, compare the case/control ratio under different genotype combinations.
  - Likelihood method: e.g. epastasis option in PLINK: comparison of different logistic regression models, where the interaction term(s) are set to zero.
- Genotype distribution perspective: the frequency of genotype combination is not the product of that of individual genotypes (the genotypes at two loci are not independent).
  - Dependency/correlation of different predictors (loci) in case-only studies: this test has higher power than the regression models. However, two loci may be correlated for other reasons: LD, or genotype combinations may be related to viability.
  - Likelihood method: the data are sampled from multinomial distributions. Under null model (no interaction): the probabilities of a genotype combination can be factorized, or equivalently, the probabilities of multiple loci can be multiplied. Ex. BEAM.

Issues of testing epistasis:

- Alternative epistatis models: there are a number of cases where epistatis may occur. A particular important situation to consider is: whether the two loci have marginal effects. It is possible that two loci has strong epistais but each loci itself has no main effect, but it is not clear how common this situation is.
- Test for interaction vs. test for association allowing for interaction: in some cases, the later may be more interesting. A natural approach is to average over all possible interacting loci of the locus of interest.
- Single locus test as filter: this is related to the issue whether a single locus has marginal effect. If not, filtering based on marginal effect may have low power. Ex. stepwise logistic regression model performs poorly under the experiment of S [Zhang & Liu, NG, 2007].
- Multiple testing: permutation test can be theoretically used, but not feasible if the number of tests is large.

Comparison of methods: (some methods are evaluted by using WTCCC data of Crohn's diease, pre-filter with p value < 0.1 to limit the number of SNPs to about 10,000)

- Random forest: No clear sharp signals (thus even worse than single-locus analysis). And no clear greater insights than single-locus analysis.
- MDR: use TuRE for attribute selection, then apply MDR. Very sensitive to random seed, thus the perofmrance does not seem very stable. CPM is similar, but for quantitative traits.
- BEAM: similar to single locus analysis, find a single significant SNP interaction, however, the two SNPs are tightly linked, thus probably from LD.
- SVM: [Waddell, SIGKDD, 2005] 3,000 SNPs in 40 cases/controls. The resulting SVM is not very interpretable (150 SNPs).

Overview of gene-environment interactions:

- Graphical model perspective of gene-environment interaction analysis: suppose we have genotype G, environment E and disease phenotype D, then we have the simple model:

$$G \to D \leftarrow E, \qquad G \leftrightarrow E \tag{4.163}$$

  where the second relation describes the possible dependence between G and E. This perspective could help some questions in G-E analysis. For example, even if G and E are independent in the population, and there is no interaction between G and E, conditioned on D, they are dependent (easy to prove using three normal RVs and check the conditional correlation between G and E).

Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies [Piegorsch & Taylor, Stat Med 1994]:

- Motivation: the standard test of gene-environment interaction is the logistic regression model testing the coefficient of the interaction terms. However, this test has three d.f. (the main effects of gene and environment, interaction effect), can we have a test with a single d.f.?
- Intuition: suppose G and E are independent in the population, and G and E work independently to raisk the disease risk, then in the disease group (cases), when we see G as a risk allele, we don't have more information of E (beyond what we konw from the disease state D, thus E should be higher), i.e. G and E are independent in cases.

- Relation between interaction term and ORs: consider the discrete model where genotype $G = i$ and environment exposure $E = j$ (with $i = 1, j = 1$ the reference state), and the disease state D or $\bar{D}$. Let Ψ_{ij} be the OR of the genotype i and environment j combination, i.e.:

$$\Psi_{ij} = \frac{P(D|G=i,E=j)}{P(\bar{D}|G=i,E=j)} \cdot \frac{P(\bar{D}|G=1,E=1)}{P(D|G=1,E=1)} \tag{4.164}$$

  Similarly, we could define Ψ_{i1} and Ψ_{1j}. Under the logistic model, the interaction term is related to the OR:

$$\exp(\gamma_{ij}) = \frac{\Psi_{ij}}{\Psi_{i1}\Psi_{1j}} \tag{4.165}$$

  When G and E act independent, we have $\Psi_{ij} = \Psi_{i1}\Psi_{1j}$. Our goal is to show that this is equivalent to: G and E are indepndent in cases.

- Case-only test: we apply the Bayes rule to the odds of (i, j) combination:

$$\frac{P(D|G=i,E=j)}{P(\bar{D}|G=i,E=j)} = \frac{P(G=i,E=j|D)P(D)/P(G=i,E=j)}{P(G=i,E=j|\bar{D})P(\bar{D})/P(G=i,E=j)} = c \cdot \frac{P(G=i,E=j|D)}{P(G=i,E=j|\bar{D})} \tag{4.166}$$

  where $c = P(D)/P(\bar{D})$ is a constant. Plug in this into the equation of the interaction term:

$$\frac{\Psi_{ij}}{\Psi_{i1}\Psi_{1j}} = \frac{P(G=i,E=j|D)P(G=1,E=1|D)P(G=i,E=1|\bar{D})P(G=1,E=j|\bar{D})}{P(G=i,E=1|D)P(G=1,E=j|D)P(G=i,E=j|\bar{D})P(G=1,E=1|\bar{D})} \tag{4.167}$$

  If the disease is rare, and G and E are independent in the population, then the four terms conditioned on $\bar{D}$ will be canceled out. We have (rearranging the terms and use the conditional probability terms $P(E|G, D)$):

$$\frac{\Psi_{ij}}{\Psi_{i1}\Psi_{1j}} = \frac{P(E=j|D,G=i)P(E=1|D,G=i)}{P(E=1|D,G=i)P(E=j|D,G=1)} \tag{4.168}$$

  which is exactly the OR of genotype i treating E as the phenotype.

- Higher power using the case-only test: it can be shown that the case-only test is more powerful or has a lower standard error when estimating the interaction than the standard logistic regression test. The reason is probably due to (1) lower df. in the case-only test (no the main effect terms); (2) using the independence of G and E in the population (similar to the fact that modeling the joint distribution may be more powerful than the regression approach modeling conditional distribution only.

- Summary of the case-only test of interaction: if G and E are independent and the disease is rare, then to test if G and E have interactions to influence the disease risk is equivalent to test if G and E are independent in the case-only samples, and the case-only test is more powerful.

Multifactor dimensionality reduction (MDR) [Ritchie & Moore, AJHG, 2001]

- MDR procedure: (1) Prediction rules: for any multi-factor genotype class (e.g. genotype combination of two loci), determine for each combination, whether it increases or decreases the risk. And the rule for predicting disease is simply the union of all combinations that are associated with high risk. Thus this is a simple Boolean function. (2) Assessing a combination of multiple loci: predictive accurancy under 10-fold cross validation.

- Data: 200 breast cancer patients, about 10 polymorphic loci in the coding regions of the candidate gene(s).

- Results: a four-locus interaction associated with breast cancer.

Tree-based association of alcoholism: [Ye & Zhang, BMC Genet, 2005]

- Data: 1,614 family members, 4,720 SNPs, response: alcohol dependence.
- Classification method: decision tree with the covariates: sex, parental phenotypes, and the SNP markers.
- Results: The pruned tree at the significance level of 0.00001: the top nodes are sex and mother phenotype, the rest are SNPs.

Comparison of tests of association between a region and a trait: [Roeder & Devlin, GE, 2005]

- Problem: given a set of SNPs in substantial LD, test if the set is associated with a trait. The idea is that by using all information in the LD region, it may perform better than the single locus test.
- Tests:
  - T_P: suppose T_l is the test statistic of the l-th locus, then define $T_P = \max\{T_1, \cdots, T_L\}$. The threshold is determined by permutation of cases and controls.
  - T_S: a smoothed version of T_P, fit a smooth curve $g(\cdot)$ to $T_1, \cdots, T_L$, and let $S_l = \hat{g}(b_l)$ be the value of the l-th SNP at the curve. Define $T_S = \max\{S_1, \cdots, S_L\}$.
  - T_R: let Y, the phenotype, be a linear function of genotypes in all loci: $g(\mu) = \beta_0 + X\beta$, where g is the logistic function, $\mu = E(Y)$ and β are the regression coefficients of all loci. T_R is defined as the test statistic of the hypothesis: $\beta = 0$ simultaneously for all loci. This is also the Hotelling T^2 statistic: the genotype frequency of all loci are equal under case and control.
- Results/Discussion:
  - T_R may be more powerful than haplotype-based association test [Chapman & Clayton, Hum Hered, 2003]. Need to check what version of haplotype-based test is used: e.g. parsimony of the haplotypes. T_R suffers from a very flexible alternative hypothesis, thus a penalty is power.
  - T_P: perform at least as well and usually better than other test statistics.
  - T_S: the effect of sommothing is that it is more robust than other statistics, most powerful when a dense set of SNPs is genotyped and a causal variant is located with this region. But it has little power when the signal is very local.
  - T_P and T_S: increase power with more SNPs; but T_R will reduce power with more SNPs (more flexible alternative hypothesis).
- Question: in T_R, the genotype frequencies do not follow Guassian distribution, how would Hotelling T^2 test apply?

Two-stage approach for testing epistatis [Marchini & Cardon, Genome-wide strategies for detecting multiple loci that influence complex diseases, NG, 2005]

- Simulation models: consider three models (Figure 1). Ex. two-locus interaction with multiplicative effect, the odds of the genotype with both loci minor allele is: $\alpha(1+\theta)^4)$, where α is the odds of baseline genotype, and θ the increase of odds by any single minor allele (in either locus).
- Simulation: assume that only one pair of loci has interaction. $L = 300,000$ markers, $n = 1,000$ to 4,000 cases and controls; MAF varies from 0.05 to 0.5; r^2 from 0.5 to 1.0; and λ (the marginal effect) from 0.2 to 1.0. In general, the results will depend on these parameters, so need to examine their influences. Note that from genotype frequency (MAF and HWE) in the control, we can estimate the frequencies in the case group (by using the disease odds). For each parameter setting, repeat the simulation 1,000 times. The results are evaluated by the power of tests, i.e. the fraction of successfully detecting association in 1,000 simulations.

- Tests: the two-stage model first selects SNPs using single-locus test, with some weak threshold. Then test any pair of loci using LRT where the alternative hypothesis would use full interaction model, i.e. logistic regression with 9 parameters (intercept, additive and dominant terms per loci, and four interaction terms). The threshold is determined by Bonferroni correction. Thus for interaction test, the penalty is very high.
- Results:
  - With no-interaction model, the single-locus test has significantly higher power than two-stage or full interaction tests, especially at low MAF.
  - With the interaction models: the two-stage and full interaction methods perform similarly, and both better than single-locus test.

Bayesian inference of epistatic interactions in case-control studies (BEAM) [Zhang & Liu, NG, 2007]

- Model:
  - Assume the genotype data are generated from a probabilistic model. Three groups of loci: group 1, no effect; group 2, main effect only; group 3, interactions. The index variables (which loci belong to which group) are unknown and to be inferred. Assume independence of loci, however, modeling LD is also possible by using a Markov chain, where the emission of the genotypes would depend on the adjacant marker.
  - Likelihood/evidence of model: the data is sampled from a multinomial distribution with three parameters for control. For case: (1) In group 1: case distribution is the same; (2) In group 2: the control distribution is a different multinomial distribution; (3) In group 3: multiplying probabilities over genotype combinations. In all computations, the parameters in the multinomial distribution are assumed to have Direchlet prior, and can be integrated out.
  - B statistic: log. of Bayes factor as the test statistic for any marker set.
- Results: BEAM outperforms the two-stage logistic regression methods on the models with no main effects, using simulation data. No epistatis is found in AMD data (small size with less than 100 cases/controls.
- Remark: the method is sensitive to LD between two markers: the probability of the genotype at two loci will not be independent, thus the test will lead to FP results.

Epistatic interactions in GWAS using PPI network [Emily & Schierup, EJHG, 2009]

- Analysis: test epistaitic interactions on SNP pairs, where associated genes have PPIs. The test is based on LRT of two models: H0 - independent effect of two SNPs; H1 - H0 and interaction between two SNPs. The logistic regression is used to map genotype to phenotype.
- Data: WTCCC data. PPI network from STRING database containing 71k interactions. After filtering (non-HWE, etc.), the number of SNP pairs (with PPI in associated proteins) for each disease was about 3 million. In contrast, the number of all possible SNP pairs is 125 billions.
- Results: 4 epistatic interactions were identified. in Crohn's disease: 8 significant SNP pairs, all from a single pair of genes: they are in LD with genes APC and IQGAP1. Both have function in cell adhesion and regulating cell migration, and both regions have been related to CD before.

Gene-Environment Interaction in Genome-Wide Association Studies [Murcray & Gauderman, Am J Epiderm, 2009]:

- Motivation: the case-only test of gene-environment interaction is more powerful, however, it assumes the independence of G and E in the population. The standard regression test does not make the assumption, but is less powerful. The idea is to use the case-only test test to filter, then apply the regression test only on the filtered list (no independence assumption, and less penalty of multiple testing).

- Two-step procedure:

  - Step 1. combine the cases and controls, and test the indepdence of G and E, i.e. $\text{logit}P(E = 1|G) = \gamma_0 + \gamma_g G$, test if $\gamma_g = 0$. Suppose there are m SNPs with $P < \alpha_1$.
  - Step 2. the standard test:

$$\text{logit}P(D = 1|G, E) = \beta_0 + \beta_g G + \beta_e E + \beta_{ge} GE \tag{4.169}$$

    Test if $\beta_{ge} = 0$. The significance level is α/m, where α is the desired type I error.

- Independence of the Step 1 and Step 2 tests: if the two tests are dependent, then we will have inflated type I error. It can be shown that, the two test statistics are asymptotically independent (using the normal approaximation of the test statistic, and compute the correlation coefficient). Note that: this is not true if we use cases only in Step 1.

- Remark: intuitively, the two tests are independent realizations of the interaction term. The first test is about the independence of G and E, and the second test is about how G and E together influence the risk of D.

Screen and clean: a tool for identifying interactions in genome-wide association studies [Wu & Roeder, GE, 2010]:

- Computational procedure(Figure 2):

  - Step 1: a set of SNPs to analyze
  - Step 2: Lasso screening retains a subset of SNPs
  - Step 3: the new dictionary involves all the subset of SNPs plus all pairwise interactions
  - Step 4: Lasso screening again
  - Step 5: cleaning, traditional hypothesis testing on the remaining terms from Step 4.

- Simulation: generate SNPs following specified LD. For simpler simultations: 1000 SNPs split into 200 blocks (5 SNPs per block); for large simulations: 100K SNPs with LD by Markov chain. The genetic models tested include: main effect only, interaction effect with varying number and strength of interactions.

- T1D analysis with WTCCC data: three pairs of SNPs, two in MHC region. The SNPs are close to each other in all three cases, though LD is very low.

A Flexible Bayesian Model for Studying Gene-Environment Interaction [Yu & Liang, PLoS Genetics, 2012]:

- Model: gene-environment interaction by assuming there are multiple clusters (sets of genotypes), and the environmental effect may be different in different clusters.

  - Inference: the number of clusters (K) unknown.

- Remark: The better performance in simulation (comparison with single SNP or PC test): relies on multiple SNPs interacting with environment. Not clear how this is true in general.

SBERIA: Set Based gene EnviRonment InterAction test for rare and common variants in complex diseases [Jiao & Peters, submitted to GE, 2013]:

- Background: difficulty of aggregating signals: tell signals from noise and how to determine the direction of the signals. To deal with this issue in the SNP set test:
  - Han and Pan 2010: used the signs of the marginal effect to determine the direction of the main effect.
  - Lin and Tang (2011) used the corresponding regression coefficient plus a constant as the weight for each marker.
  - Cai, Lin and Carroll 2012: proposed to weight each marker based on the z-score of its effect.

  One common characteristic of these methods is that the statistics used to weight the markers are not independent of the main effect test. Hence, permutation is needed to estimate the null distribution and maintain the correct type I error, which is computationally intensive.

- Background: correlation screening to identify GxE interactions, and the test is independent of the usual regression-based test [Murcray & Gauderman, Am J Epiderm, 2009].

- Basic model and benchmark methods: suppose we are testing the interaction of a group of SNPs with environment (E). Let G be the genotype vector, we have the model:

$$\text{logit}(D) = \alpha_0 + \alpha_1 E + \alpha_2 G + \sum_j \beta_j (EG_j) \tag{4.170}$$

  where β_j is the coefficient of the j-th interaction term. The null hypothesis is $\forall j : \beta_j = 0$. The simplest tests are: the LRT with df. m (the number of SNPs) - χ^2 distribution, and the individual SNP test (min-P value) - use permutation to assess significance.

- Model: similar to the aggregate test of the main effect (burdent test, EREC, etc.), we assume $\beta_j = w_j \rho$, with w_j the weight of the j-th SNP. The model becomes:

$$\text{logit}(D) = \alpha_0 + \alpha_1 E + \alpha_2 G + \rho \sum_j w_j (EG_j) \tag{4.171}$$

  Our test is $H_0 : \rho = 0$. Similar to EREC, it is better to learn w_j from data, instead of using fixed weights. However, this creates inflation of type I error. To address this problem, use the correlation screening to test each SNP, and set w_j to 1 or -1 (depending on the direction) or 0 (if not passing a certain threshold). The threshold is chosen s.t. under H_0, about 10% of SNPs pass the threshold.

- Simulation: consider four different scenarios, with the first threee of GWAS (common variants) setting and the last rare variant setting.
  - Scenario 1: 21 SNPs in LD, no main effect, two SNPs have interaction effect. Consider three settings $\beta_1 = \beta_2, \beta_1 = -\beta_2, \beta_2 = 0$.
  - Scenario 2: 20 independent SNPs, each has main effect (e.g. pathway). Two SNPs have interaction effect, and similar setting for β_1 and β_2.
  - Scenario 3: 21 SNPs in LD, no main effect, two SNPs have interaction effect with similar β. The difference, E is correlated with one of the two interaction SNPs.
  - Scenario 4: 10 independent SNPs (rare), each has main effect. 2-8 SNPs have interaction effects (could be both positive and negative, the ratio is varied).

  In S1-3, compare with min-P and LRT. In S2, compare with the genetic risk score (GRS) method: test the interaction of GRS and E. In S4, compare with the simple version of burden test: the burden is simply defined as the sum of rare alleles.

- Application: in a cancer data of 10K cases and controls, 25 known SNPs associated with the phenotype. Treat the 25 SNPs as a group and test the interaction with smoking. Found a SNP with low p-value (less than 0.01), while the GRS method gives a higher p-value (about 0.05).
- Analysis: why the method performs better than the simple LRT? Similar to EREC, it uses w_j from the data, thus reducing the df. of the test: in simulations, the naive LRT has df. of 21. The question is that: in these situations, unlike the burden test, there is no obvious aggregation of signals across SNPs (e.g. in S1, only two causal SNPs with opposite signs). Where does the power come from?
  - Similarlity to James-Stein estimator: better estimator when one tries to simulateously estimate multiple parameters, even if the parameters are completely independent. However, there is no obvious shrinkage?
  - Bayesian interpretation: w_j is (almost) entirely from the data, thus the model is similar to a noninformative Bayes prior.

Test for interactions between a genetic marker set and environment in generalized linear models (GESAT) [Lin & Lin, Biostatistics, 2013]

- Model: let μ_i be the expectation of phenotype Y_i, X_i be covariates, E_i be scalar environment variable, G_i be the genotype vector and $S_i = (G_i E_i)^T$ be the interaction term. The model:

$$g(\mu_i) = X_i\alpha_1 + E_i\alpha_2 + G_i\alpha_3 + S_i\beta \tag{4.172}$$

  Use the variance component model, assuming $\beta_j \sim N(0, \tau^2)$, and we are testing $H_0 : \tau = 0$. This can be tested using the score test.

Polygenic model of epistasis [Andy Dahl, NHG, 2020]

- Defining coordinated epistasis: our phenotype model is

$$y \sim \sum_s G_s\beta_s + \sum_{s,s'}(G_s \cdot G_{s'})\omega_{s,s'} + \epsilon \tag{4.173}$$

  where β are single SNP effect, and Ω are interaction effects. A simple model is to use a normal prior for Ω on all SNPs or a subset of SNPs. Our goal is to model the dependency of Ω on β: intuitively, for instance, SNPs with large/non-zero β should also tend to have large/non-zero interaction terms. Define:

$$\gamma(\Omega, \beta) = \mathrm{Cor}_{s\neq s'}(\Omega_{ss'}, \beta_s\beta_{s'}) \tag{4.174}$$

- Pathway interpretation of coordinated epistasis: imagine phenotype depends on several latent variables z_k and there are interactions among z_k's. Assuming SNPs act on z_k, we can see that the interact term of SNPs depends on the interaction between pathways. Ex. suppose there is only interaction term between pathways k and k', then we have:

$$\Omega_{ss'} = \beta_s^{(k)}\beta_s^{(k')}\alpha_{kk'} \tag{4.175}$$

  where the first two terms are SNP to pathway effect and the last term is pathway interaction.

- Implications of coordinated epistasis, non-zero γ on evolution: deviation of trait distribution from normal in the population. If each SNP segregate independently, we will see an excess of large values of trait.
- Inference of γ: MOM estimator, even/odd chromosome tests. Let P_E and P_O be the PRS of even and odd chromosomes, respectively. Fit the regression model:

$$y \sim \beta_E P_E + \beta_O P_O + \gamma P_E P_O \tag{4.176}$$

  One can prove that γ estimates coordinated epistasis.

- Results: UKBB 26 traits, several of them show strong signals of coordinated epistasis.
- Remark: if the relationship between z_k to the trait is non-linear, e.g. threshold function, it may also cause interactions among SNPs in the same pathway.

4.6 Extensions of Association Analysis

Bayesian genetic mapping [personal notes]

- Statistical background: we have a regression model with the prior:

$$y|\beta \sim N(X\beta, \sigma^2 I) \qquad \beta \sim N(\beta_0, V_0) \tag{4.177}$$

  Marginalize β:

$$y \sim N(X\beta_0, \sigma^2 I + XV_0X^T) \tag{4.178}$$

- Individual-level model: we follow the model of BIMBAM. We assume genotype is centered, but not normalized. Suppose the configuration is γ. We have:

$$y|\beta_\gamma \sim N\left(X\beta, \frac{1}{\tau}I\right) \qquad \beta \sim N(0, \sigma_a^2/\tau I_\gamma) \tag{4.179}$$

  where I_γ is the diagonal matrix with diagonal terms given by the indicators γ. This leads to the marginal likelihood:

$$y|\gamma \sim N\left(0, \frac{1}{\tau}I + \frac{\sigma_a^2}{\tau}XI_\gamma X^T\right) \tag{4.180}$$

  Simlarly, we have $y|\gamma = 0 \sim N(0, I/\tau)$. One can then compute BF for a configuration.

- Summary level model: we have $\hat{\beta}_j$ and s_j from unnormalized genotypes, with $\beta_j \sim N(\hat{\beta}_j, s_j^2)$.

Relationship between PVE and prior effect sizes [personal notes]

- Individual level data: usually, one uses a regression model:

$$Y = X\beta + \epsilon \qquad \epsilon \sim N(0, 1/\tau) \tag{4.181}$$

  following BIMBAM. The prior effect size is usually defined on the scale of $1/\tau$, assuming BVSR, we have, for causal variants, $\beta \sim N(0, 1/\tau \cdot \sigma_a^2)$, and $\beta = 0$ for non-causal variants. Suppose genotypes are normalized, the variance explained by a single causal variant is: $\pi\sigma_a^2 \cdot \frac{1}{\tau}$. This leads to the PVE across all m SNPs as:

$$\text{PVE} = \frac{m\pi\sigma_a^2 \cdot \frac{1}{\tau}}{m\pi\sigma_a^2 \cdot \frac{1}{\tau} + \frac{1}{\tau}} = \frac{m\pi\sigma_a^2}{1 + m\pi\sigma_a^2} \tag{4.182}$$

  If we define prior effect size on the scale of phenotypic variance, i.e. $\beta \sim N(0, \sigma_a^2\sigma_y^2)$, where σ_y^2 is the variance of phenotype, then we have: $\text{PVE} = m\pi\sigma_a^2$.

- Summary level effect sizes and standard errors: we should assume that effect sizes are defined on the scale of residual variance or σ_y. Even though they are unknown, they will disappear in PVE and in calculation of statistical evidence (BF of a SNP or a configuration in fine-mapping). We have the same relationship of PVE and σ_a^2 above.

- Summary level Z-scores: we have $\hat{\beta}_i \sim N(\beta_i, s_i^2)$, where $\hat{\beta}_i$ and s_i are estimated effect size and standard error, respectively. Let N be sample size and $v_i = 2f_i(1 - f_i)$ be the variance of genotype, we know $s_i^2 = \sigma_y^2/(Nv_i)$. From this, we have the PVE:

$$\text{PVE} = \sum_i \beta_i^2 v_i/\sigma_y^2 = \frac{1}{N}\sum_i \frac{\beta_i^2}{s_i^2} \tag{4.183}$$

Now we defined standardized effect size as: $u_i = \beta_i / s_i$, and assume $u_i \sim N(0, \sigma_u^2)$ for causal variants and 0 for non-causal variants. So $\mathrm{E}(u_i^2) = \pi\sigma_u^2$, and we have:

$$\mathrm{PVE} = \frac{1}{N} m\pi\sigma_u^2 \tag{4.184}$$

So when summary statistics are Z-scores, we should consider effect size at the Z-score scale, which depends on sample size.

4.6.1 Basic Fine Mapping

Background [personal notes]:

- Prior variance and PVE explained: let $\beta_j \sim N(0, \sigma_\beta^2 \sigma_y^2)$ be the effect of SNP j. If genotype data is normalized (FINEMAP paper), the PVE explained by j is given by:

$$\mathrm{PVE}_j = \frac{\mathrm{Var}(\beta_j)\mathrm{Var}(X_j)}{\mathrm{Var}(Y)} = \frac{\sigma_\beta^2 \sigma_y^2}{\sigma_y^2} = \sigma_\beta^2 \tag{4.185}$$

  Typically, even a causal SNP explains a small amount of PVE. Ex. default parameter of FINEMAP is $\sigma_\beta = 0.05$, or explained PVE 0.25%.

From genome-wide associations to candidate causal variants by statistical fine-mapping [Schaid, NRG, 2018]

- Initial GWAS results: LocusZoom plot, lead or index SNPs.
- Why need fine-mapping? Simulations with 1000 cases and 1000 controls, at effect size 1.5 and AF 50%, the causal variant has 79% chance of being the lead SNP; but at effect size 1.1 and AF 5%, only 2.4% chance.
- LD and SNP statistics: ideally, the causal variant has the lowest p-values and the p-values of other SNPs decay with LD. However, in practice, often not the case. Ex. APOE locus for LOAD, the association statistics change in complex patterns.
- Remark: this is possible. Consider a relatively large haplotype, there may be recombination events within the block, leading to non-monotonic relation of LD.
- Forward regression to determine if there are multiple signals in a region: a challenge is to determine the threshold in later steps, some use the same 5E-8, some uses more liberal thresholds of 1E-5 or 1E-4. Limitations of this approach: (1) Multiple testing: if there are m SNPs, then after k steps, mk tests are performed. (2) Low power of detecting secondary signal: dramatic loss of power if the correlation is 0.2 or higher (Figure 3).
- Heuristic LD approach for fine-mapping: top SNP and all within LD threshold; or all SNPs in the haplotype block of the lead SNP. Limitation: arbitrary threshold; block boundaries not easily defined.
- Penalized regression: better than forward regression. Limitation:high correlation combined with sparse models reduces the chance of selecting the causal variant.
- Bayesian approach (Figure 2C): Model marginal likelihood. (1) PIP: caution when there are multiple highly correlated SNPs, then PIP of each SNP may be small. Use the number of causal SNPs, or sum of PIPs (2) Credible set:the minimum set of SNPs that contains all causal SNPs with probability α. Significant advantages of Bayesian fine-mapping: PIPs, higher power of mapping SNPs with smaller effects.

- Trans-ethnic fine-mapping: SNP effects are often consistent across populations. Multiple European ancestries: little gain of power. Incorporating African ancestry helps because of much narrower LD. Trans-ethnic analysis often done with random effect model on summary statistics withMETASOFT [Han and Eskins, AJHG, 2011]; followed by Bayesian fine-mapping (Figure 2D).

- Remark: this approach assumes a single causal configuration. A better approach may be to model possibly different causal variants and effect sizes across populations.

- Incorporating annotations in Bayesian fine-mapping: most use a small set of annotations. CAVIAR-BF. To assess the performance: compare the size of Bayesian credible set. In PAINTOR paper, reduce from 17 to 13 SNPs per region.

Identifying Causal Variants at Loci with Multiple Signals of Association: Cavier [Hormozdiari & Eskin, Genetics, 2014]

- Review of current fine-mapping methods: (1) Single causal SNP assumption. (2) Conditional approach: iterative selection of SNPs, at each step, conditioned on the ones previously chosen. The problem: selection of SNPs in close LD is arbitrary.

- Cavier: the idea is to choose causal SNP set that covers most of the posterio prob. Let S be the statistics of SNPs, and c be indicator vector. Assume $P(c)$ follows binomial distribution, and $P(\hat{S}|c)$ is given. The evidence of a c is given by its posterior $P(c|\hat{S})$. Given a set of SNPs K, the total posterior prob. of K is given by the sum of posterior of all c consistent with K, called confidence level of K. Our goal is to choose the minimum set K with confidence level above a threshold.

- Algorithm: to reduce computational burden, two ideas (1) Limit to at most 6 causal SNPs per blocks. (2) at each iteration, choose a SNP that increase the posterior prob. the most.

FINEMAP: efficient variable selection using summary data from genome-wide association studies [Benner and Pirinen, Bioinfo, 2016]

- Summary statistics on binary traits: use Z scores, the phenotypic variance for quantitative traits is 1; for binary traits, $\sigma^2 = 1/(\phi(1-\phi))$ where ϕ is the proportion of cases among all samples.

- Prior model: let γ be the indicator configuration and λ be the effect sizes. For a causal variant, $\lambda \sim N(0, s_\lambda^2 \sigma^2)$ where σ^2 is phenotypic variance and s_λ^2 is user-defined parameter, default 0.05^2 for quantitative trait. The prior for γ is: let p_k be the probability of having k causal variants, and each configuration is equally likely given k. The distribution of p_k is basically binomial with probability of success $1/m$, where m is the number of variants in the region.

- Computational problem of evaluating marginal likelihood: in a region with m SNPs, the likelihood involves $m \times m$ matrix. By factorize the probability into causal and non-causal SNPs, we can reduce the computation to $k \times k$ matrix, where k is the number of causal SNPs.

- Shotgun Stochastic Search (SSS): once reach a configuration, compute the unnormalized prob. of all neighboring configurations (defined by simple operations); then move to one configuration with probabilities determined by these unnormalized probabilities. All these prob. computation results will be stored for future use (hash table). Comparison with MCMC: once in a dense region, SSS explores the entire neighborhood.

Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies [Benner and Pirinen, AJHG, 2017]

- APOE causal variants of LDL-C: true results from in-sample LD, two causal SNPs. With ref. LD from 99 individuals (Finnish from 1000GP): several FP SNPs, and one likely causal variant has PIP close to 0.

- Intuition why mismatched LD can cause problems: if the external panel underestimates LD, it will try to explain some association as independent causal SNPs (FPs). If it overestimates LD, it will not be able to distinguish two SNPs, so the true causal SNP will have to split evidence with another SNP (both reduced PIP for causal, and FP for another SNP).
- Simulation study (Figure 4): (1) GWAS of sample size 5000: ref. panel of 100 individuals significantly lower AUC, but 1000 is sufficient; (2) GWAS of sample size 50K (UK Biobank): ref. panel of 100 is not enough; 1000 will have significantly lower AUC; 5,000 will be close to optimal.
- LDstore software: to share LD.

4.6.2 Bayesian Statistical Methods

Reference: Bayesian statistical methods for genetic association studies [Stephens & Balding, NRG, 2009]

Motivation for Bayesian methods:

- p value does not take into account the power of test (i.e. how likely the alternative hypothesis is). Ex. testing two SNPs, one with high MAF, the other low MAF. For the former, the test is simpler because of larger sample size (if it is associated) thus the same p value should be attached with higher confidence than the latter.
- Also, to determine a threshold of significance, one should take into account the number of true associations, not the number of tests (our probability should not be affected by how many tests we perform).

Bayesian testing of a single SNP in GWAS:

- Bayes factor (BF) and posterior probability of association (PPA): under Bayesian framework, the quantity of interest is $PPA = P(\text{there is association}|D)$. We first obtain posterior odds (PO): $PO = BF \cdot \pi/(1-\pi)$, where π is the prior probability of association, and the ratio is the prior odds. In general, π is taken to be in the range of 10^{-4} to 10^{-6}, therefore BF generally needs to be large, say, greater than 10^5 to be significant.
- Computation of BF: need to compute $P(D|M_1)$ and $P(D|M_0)$ assuming the logistic regression model.
  - BF for the additive model [WTCCC, Nature, 2007]: suppose M_0 denotes the model of no association, M_1 denotes the model with an additive effect on the log-odds scale. Then we have:

$$BF_1 = \frac{P(D|M_1)}{P(D|M_0)} = \frac{\int P(D|\theta_1, M_1)P(\theta_1|M_1)d\theta_1}{\int P(D|\theta_0, M_0)P(\theta_0|M_0)d\theta_0} \tag{4.186}$$

    where the parameters θ are log. of odds ratio. We use a logistic regression model for the likelihood. Let N be the sample size, Z_i, Y_i be the genotype and phenotype of the i-th individual respectively, then:

$$P(D|\theta) = \prod_{i=1}^{N} p_i^{Y_i}(1-p_i)^{1-Y_i} \tag{4.187}$$

    where for M_1, we have: $\theta_1 = (\mu, \gamma)$, and $\log p_i/(1-p_i) = \mu + \gamma Z_i$; for M_0, we have: $\theta_0 = (\mu)$, and $\log p_i/(1-p_i) = \mu$. The prior distributions are specified by $\mu \sim N(\alpha_1, \beta_1)$, in practice, choose $\mu \sim N(0,1)$; and $\gamma \sim N(\alpha_2, \beta_2)$. To evaluate the marginal likelihood $P(D|M)$, use the Laplace approximation, and the result is a function of $\hat{\theta}$ (MAP estimator of θ).
  - Averaging genetic models: could consider multiple models: additive, dominant, recessive, and general model. Weighting the models by their prior probabilities: even small weights on non-additive models can allow the identification of large non-additive effects without substantial reduction in the ability to detect near-additive effects.

  - Averaging effect size: the prior distribution of the effect size (log-OR) is usually assumed to be $N(0, \sigma^2)$ ($\sigma^2 = 0.2$ is a common choice). However, this distribution decays rather fast; to allow for large effect size, replace with a mixture of normal distribution.

- Dependence of BF: one factor is the MAF, the same p value corresponds to large BF if MAF is large (or the power is high). Another factor is the prior standard deviation (σ) of the effect size: could have a big impact if MAF is low.

- Why Bayesian methods do not require multiple testing correction? Suppose we are testing n SNPs in a region, assume that the global prior odds is fixed (e.g. there is one disease SNP among n SNPs), then the prior odds of each SNP is $1/n$. This amounts to correction of testing multiple hypothesis (as n increases, the BF needs to increase to reach the same level of posterior odds).

Imputation-based analysis of association studies: candidate regions and quantitative traits. (BIMBAM) [Servin & Stephens, PG, 2007]

- Model: phenotype is associated with genotype (q SNPs) by a standard regression.

$$y_i = \mu + \sum_j \beta_j x_{ij} + \epsilon_i \tag{4.188}$$

  where $\epsilon \sim N(0, 1/\tau)$. Assume the genetic effects of the j-th SNP are: $(0, a_j(1+k), 2a_j)$ for three genotypes, respectively.

- Prior distributions:

  - Prior for phenotype mean and variance (μ, τ): two priors are considered. The D_1 prior is defined as $p(\mu, \tau) \propto 1/\tau$.
  - Prior for QTNs: p_0 is the probability that no QTN is present; $p(l)$ is the distribution that there are l QTNs, in this paper, choose $p(l > 2) = 0$.
  - Prior for effect size: assume $a_j \sim N(0, \sigma_a^2/\tau)$. σ_a should be relatively small. The paper choose $\sigma_a = 0.5$, i.e. one allele may increase the trait by 0.5σ. It was suggested in Discussion that a smaller σ_a should be used. The default BIMBAM averages BF for four values: $\sigma_a = 0.05, 0.1, 0.2, 0.4$.

- Inference: Let M_γ be a specific model of which SNPs are associated with the disease, where γ encodes q-dim vector (each SNP associated with the disease is called QTN). Consider the special case where only 1 QTN is associated, then there are q possible models: H_j - SNP j is the QTN. Then:

$$BF = \frac{1}{q} \frac{\sum_j P(\mathbf{y}|\mathbf{G}, \mathbf{H_j})}{P(\mathbf{y}|\mathbf{G}, \mathbf{H_0})} \tag{4.189}$$

  Thus the overall BF is the mean of the single-SNP BFs.

- "Bayes/non-Bayes compromise": use the p value of BF through permutation.

- Bayesian analysis with imputation: imputation of untyped SNPs using standard methods. The uncertainty of imputation can be incorporated into the inference, e.g. summing over the untyped SNPs in computation of $P(\mathbf{y}|\mathbf{G}, \mathbf{H_j})$ (where $\mathbf{G}$ contains only typed SNPs).

Application of Bayesian approach in other tasks:

- Imputation: the uncertainty of imputed SNPs can be taken into account using Bayesian approach.

- Fine mapping: the goal is to determine the truely associated (or causal) SNP(s) in a region of LD. This is a problem of Bayesian variable selection where the predictors may be correlated. The advantage of Bayesian method is: the probabilities of each variable can be quantified. This is often necessary because given the correlations, the true feature can be difficult to determine (the methods such as Lasso will choose a few, but this provides a supferfically simple solution).
- Meta-analysis: the random-effects model, where the effect across different studies/population is treated as random variable.

Polygenic modeling with Bayesian sparse linear mixed models [Zhou & Stephens, 2013]

- BSLMM model: the model can be written as:

$$y|X,\beta \sim N(X\beta,\tau^{-1}) \qquad \beta_j \sim \left[\pi N\left(0,\frac{\sigma_a^2}{\tau}\right)+(1-\pi)\delta_0\right]+N\left(0,\frac{\sigma_b^2}{\tau}\right) \tag{4.190}$$

  We can write $\beta_j=\tilde{\beta}_j+\gamma_j$ where $\tilde{\beta}_j$ is the sparse genetic component, and γ_j the polygenic component. We can rewrite the model as:

$$y=X\tilde{\beta}+u+\epsilon \qquad u\sim N(0,\sigma_b^2\tau^{-1}K) \qquad \epsilon\sim N(0,\tau^{-1}I_n) \tag{4.191}$$

  where $K=XX^T/p$ is the GRM (columns of X are centered but not standarized).

- Prior specification: the parameters are $\pi,\sigma_a^2,\sigma_b^2$. The prior of π is similar to BVSR. For the effect size parameters, we specify the prior in terms of PVE and PGE (the fraction of sparse effects among all genetic variance). Let h be the PVE and ρ the PGE, both have $U(0,1)$ prior. The expected genetic and environmental components are:

$$V_B=\frac{\sigma_a^2}{\tau}p\pi s_a \qquad V_P=\frac{\sigma_b^2}{\tau}s_b \qquad V_E=\frac{1}{\tau} \tag{4.192}$$

  where s_a is the average variance of SNP genotype, and s_b the average of the diagonal term of K. This leads to Equations 16 and 17 in the paper.

- Estimation of PVE: comparison of LMM, BSLMM and BVSR. When the number of causal SNPs is relatively large, BVSR performs poorly, significantly underestimating the PVE than LMM and BSLMM.
- Prediction of phenotypes: in real data, for some diseases, HT, CAD, T2D and BD, BVSR does not perform well and BSLMM and LMM have similar performance. In other diseases, CD, RA and T1D, BVSR and BSLMM both perform well, better than LMM.

4.6.3 Gene, Pathway and Network Association Test

Goal: from GWAS data, find pathways/gene modules that are associated with diseases.

Problems of gene and pathway association test:

- Gene/region test: test the significance of a gene/region, combining the evidence of all SNPs in the gene/region. Need to address the bias problem: the size, SNP density, etc. of genes are different.
- Testing given gene groups: in general, assign P-values of each gene in the group, and test the significance of group (e.g. GSEA).
- Finding significant groups in a list of significant SNP/loci: assign genes to the SNPs first and find significant gene groups. The first step may not assign the unique gene to a SNP.
- Network test: in the gene network (e.g. PPI) with connectivity information, find significant subnetworks.

Challenges of pathway association test:

- Gene bias: the size, SNP density and LD structure are different in different genes. Larger genes are more likely to be close to some significant SNPs, so the method of assigning most significant SNPs in the neighborhood to a gene may be biased. Need normalization: convert the statistic of any gene (e.g. its most significant SNP) into some value that normalizes the gene size and LD, e.g. permutation of case and controls and compute the p value of the statistic of that gene [Holmans09].
- Assign multiple genes adjacent to a single SNP: if this is not allowed, this is probably not realistic; if this is allowed, however, this may skew the results. Ex. one SNP is adjacent to 10 genes that are functionally related, then these 10 genes may show enrichment in GWASPA.
- Multiple independent association signals/causal variants: not a seriour problem according to [Wang07], as most often if a gene has multiple significant SNPs, they are generally located in the same LD block. However, it is not clear if this is also true for weakly associated SNPs (as multiple rare causal variants should be common).
- Heterogeneity of effects: need to consider different cases: (1) one gene has the dominant effect, or (2) a number of genes each has modest effect. The two scenarios are indistinguishable in the test of [Wang09]: e.g. whenever a gene set contains a MHC variant, the gene set is significant, even if the rest of genes are completely random.
- Verification of pathways: the same idea of replication may be applied.
- Other issues:
  - Epistatis of genes in a pathway: not explored.
  - Rare variants: how would the analysis extend to rare variants?

Approaches of gene and pathway tests and analysis: we focus on the analysis using summary statistics, since these are most useful in practice [personal notes]

- Difference between gene and pathway tests: under the gene test, there could be strong LD between SNPs, and usually the true number of causal SNPs is small (though not necessarily 1). While at the pathway leve, the statistics of genes are usually independent, and the proportion of causal genes can be relatively large.
- Existing approaches to gene test: two general approaches: (1) The top SNP or more generally, top H SNPs. (2) Some combination of the test statistics of all SNPs, including Fisher's method, summation of χ^2 statistic, and so on. Also in general, need to correct for type-1 error.
- Analysis: depending on the true model, different method may perform better. Ex. if there is a single causal SNP, then the best SNP methods may perform the best. In this case, the sum of χ^2 statistic would likely be dominated by noises, especially when the number of SNPs is large. On the other hand, if there are multiple causal SNPs (each with modest signal), the best SNP method would suffer.
- Existing approaches to pathway test: more similar to anaysis of gene sets in gene expression data. Could use top genes, or use overall departure of p-value distribution from null (e.g. KS test).
- Ideas: explicit model the alternative model, and use Bayesian approach for the gene-level test.

Analysing biological pathways in genome-wide association studies [Wang & Hakonarson, Nat Rev Genet, 2010]

- General procedure:
  - Mapping SNPs to genes: different thresholds

  - Pruning SNPs: identify independent SNPs from LD regions.
  - Gene-based test statistic: often min-P, some methods use multiple markers
  - Pathway enrichment: distribution of test statistics or hypergeometric test
  - Adjust for gene size and pathway size: often permutation.

- Self-contained test: use pathway and null background. Competitive: comparison between pathways. The concern of self-contained test: inflation.

- Challenges of pathway analysis

  - Effect heterogeneity: distinguish the case where one major gene drives the effect. Idea: remove known associated genes and test pathway, e.g. remove TCF7L2 from Wnt signaling, for T2D.
  - Bias introduced by permutation: esp. for permutation of p-values.
  - Multiple pathways that are non-independent. FWER may be overly conservative.

Functional and genomic context in pathway analysis of GWAS data [Mooney & Wilmot, TIG, 2014]

- List of existing software for GWAS pathway analysis (Table 1).

Pathway-based approaches for analysis of genomewide association studies. [Wang & Bucan, AJHG, 2007], Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. [Wang & Hakonarson, AJHG, 2009]:

- Assign P-values of genes: Each SNP is assigned to its closest gene if it is within 500 kb upstream/downstream, otherwise, the SNP will not be assigned to any gene. For each gene, choose the most significant SNP, according to the test statistic (χ^2 value), and assign this test statistic to this gene.

- Enrichment score (ES): similar to GSEA, test for enrichment of the query gene group in the top-ranked list. Rank all N genes by their test statistic: $r_{(1)}, \cdots, r_{(N)}$, and for any gene set S, compute the enrichment score, $ES(S)$, as the weighted Kolmogorov-Sminov like running sum statistic.

- To address for possible biases (e.g. gene size bias): do permutation test. In different gene sets, the gene sizes are different, thus some sets may have higher ES scores simply because they have larger genes. Random shuffling of case and control labels and repeat the analysis, calculate the ES of the gene group for any permutation, $ES(S, \pi)$. The score of each group is then normalized by the mean and standard deviation under random shuffling:

$$Z(S) = \frac{ES(S) - \text{mean}[ES(S, \pi)]}{\text{SD}[ES(S, \pi)]} \tag{4.193}$$

- Multiple hypothesis testing correction: the FDR for a threshold Z^* is computed as:

$$FDR = \frac{\%\text{ of } (S, \pi) \text{ with } Z(S, \pi) \geq Z^*}{\%\text{ of observed } S \text{ with } Z(S) \geq Z^*} \tag{4.194}$$

- Results: Parkinson disease: new glycan-related pathways. The relationship between glycobiology and nuero-degenerative disease has been reported recently.

Pathway analysis of seven common diseases assessed by genome-wide association [Torkamani & Schork, Genomics, 2008]

- Assign P-values of genes: Each SNP was assigned to a gene within 5 kb, if mapped to multiple genes, assign to a single one according to coding > intron > upstream, etc. And for each gene, choose the most significant SNP.

- Enrichment test: take top 2.5% genes and use hypergeometric test for each gene group.
- Results: WTCCC data. (1) Bipolar disorder: sulfotransferase, HDL metabolism, glutamate receptors; (2) Hypertension: dopamine signaling, PKA and cAMP signaling pathways, calineurin signaling (regulation of lipid metabolism), cell-cell interactions and cytoskeletal remodeling.
- Biological lessons from pathway analysis:
  - Most of seven diseases have weak risk factors in general signaling pathways: G-protein, cAMP, calcium signaling, etc. Also TFs, transcriptional regulatory factors, membrane recetpros are common risk factors.
  - Each disease often has multiple aspects: metabolic, neurological or inflmamatory. Ex. Bipolar disorder: sulfotransferase is related to clearing of dopamine in extracellular space; Alzheimer's disease: ApoE.

Pathway and network-based analysis of genome-wide association studies in multiple sclerosis [Baranzini & Barnes, HMG, 2009]

- Assign P-values of genes: lowest P value of all SNPs mapping to a given gene. Different ways of combining multiple P values of SNPs to a gene-wise P value are tested, including Fisher's method of combining P values, and a method that corrects for the number of SNPs and for LD, but none performed better.
- Enrichment test: load the P values of genes to PPI network, and search for subnetworks where significant genes are overrepresented [Ideker02]. The modules are then tested for GO terms and KEGG pathways.

[Elbers & Onland-Moret, GenetEpid, 2009]

- Assign P-values of genes: SNPs are mapped to haplotype blocks (defined by $r^2 > 0.25$). Within a block, most have multiple genes. To find the best genes within each block: use a reference gene network and find genes related to genes in other selected regions with an interaction P value below 0.05.
- Enrichment test: once genes are assigned to pathways, use the standard tools to test if genes in a certain pathway are overrepresented in the list of candidate genes.
- Results:
  - DGI and WTCCC data of T2D, with a weak threshold on individual SNPs ($P < 0.003$). The remaining about 2,000 SNPs were then used for analysis: gene assignment and pathway enrichment test. The results from the two datasets overlapped in pathways, but frequently different genes in the same pathway were picked. Thus may be difficult to replicate, but shared pathways.
  - Randomly selected genes: usually result in significantly overrepresented pathways.
  - Possible biases: large genes, genes in large LD blocks, pathways with more genes. That large pathways are preferred may also come from the fact that with larger set, statistical evidence is stronger. Permutation and bootstrapping should improve the results.

[Holmans & Craddock, AJHG, 2009]

- Assign P-values of genes: a SNP is assigned to a gene if it lies within 20kb of 5' or 3' of that gene. Allow a SNP to be assigned to multiple genes.
- Enrichment test: [Holmans09-Fig1]
  - Test statistic: first define the list of significant genes (use all significant SNPs, and extract the associated genes), of size N. For each gene list, the test statistic T is the number of significant genes in that list.

– Statistical significance: randomly sample N genes (by random sampling of SNPs) M times, and count T, the number of significant genes in the test list, in each of the M replicates.

[Yu & Chatterjee, GE, 2009]

- Combination of P values: (test the null hypothesis that none is significant) the rank-truncated product (RTP) method:

$$W(K) = \prod_{i=1}^{K} p_i \tag{4.195}$$

where K is a predetermined integer (truncation point). The adaptive RTP (ARTP) method chooses the K that gives the minimum P-value. The significance of the ARTP statistic needs to be assessed by permutation test to account for the multiple testing of different values of K.

- Strategy of pathway association: (1) obtain the gene-level P values using ARTP; (2) combine the gene-level P values of the pathway using ARTP. If step (1) uses the P value based on known null distribution, then no permutation on (1) is needed; otherwise, would need multi-level permutation to assess the significance of the final statistic.

Logistic kernel machine model [Wu & Lin, AJHG, 2010]:

- Logistic regression of the trait on all SNPs in a gene, with a kernal function. The kernel can measure the similarity of genotypes of individuals. The df. is determined s.t. the high LD region has a lower df.

A Versatile Gene-Based Test for Genome-wide Association Studies (VEGAS) [Liu & Macgregor, AJHG, 2010]

- Test statistics for a gene: let χ_i^2 be the 1-df chi-square statistic of the i-th SNP of the gene, the gene-level test statistic is simply $\sum_i \chi_i^2$.

- Determining null distribution: use simulation to obtain the null sample, then calculate the test statistic. To get the null sample, generate MVN sample $Z \sim N(0, \Sigma)$: this is accompolished by first sample independent standard normal random vectors, then multiply that by C, where C is the Cholesky decomposition of Σ: $CC^T = \Sigma$.

MISA: Bayesian model search and multilevel inference for SNP association studies [Wilson & Schildkraut, Ann Applied Stat, 2010]

- Model: phenotype is associated with genotype (q SNPs) by a standard regression. Let M_γ be a specific model of which SNPs are associated with the disease, where γ encodes q-dim. vector. The BF is given by:

$$BF(H_A : H_0) = \sum_{M_\gamma} BF(M_\gamma : H_0) P(M_\gamma | H_A) \tag{4.196}$$

The posterior probability of a single SNP, or a gene, can be marginalized from $P(M_\gamma)$.

- Prior: BetaBinomial distribution of the size of M_γ. Choose: BetaBinomial$(1, \lambda S)$ in this work where S is the total number of SNPs and λ is a parameter, the global prior odds are $1/\lambda$. At $\lambda = 1$: the global prior odds of there being at least one association of 1, and the marginal prior odds of any single SNP of $1/S$.

- Inference: for any model, define its fitness as the log of the (unnormalized) posterior probability. Sample models based on their fitness using the Evolutionary Monte Carlo algorithm.

- Data: genotype only in candidate genes, 2129 women at 1536 SNPs in 170 genes on 8 pathways.

Gene-Based Tests of Association [Huang & Arking, PLG, 2011]:

- Motivation: Gene-based and related multi-marker association tests have generally under-performed single-locus tests when assessed with real data. A general drawback of methods that attempt to exploit the structure of LD to reduce the number of tests, for example through PCA, is the loss of power to detect low-frequency alleles. Methods that consider multiple independent effects often require that the number of effects be pre-specified, which loses power when the tested and true model are different.
- Methods:
  - A model M is defined as the subset of K SNPs in a gene with P total SNPs that are permitted to have non-zero regression coefficients. The prior $P(M)$ assumes that each of the SNP is equally likely to be causal (f). The effective number of SNPs (T) is smaller than the number of SNPs because of LD. Equation (5).
  - The model likelihood: $P(Y|M, X)$. Integration over β: dependent on the SSE at MLE of β (similar to Laplacian approximation?) Integration over τ ($1/\sigma^2$): Gamma function or steepest descent approximation (?). The final log-likelihood is given by Equation (10): it is a function of N, K, Σ (genotype matrix), MLE of σ, A and B (the limit of τ and β respectively). (Note: SSE is absorbed to MLE of σ^2.)
  - GWiS strategy: find M that maximize $P(M|X, Y)$, Equation (12). The terms involving K provide a Bayesian penalty for model performance, but also make this an NP-hard optimization problem
  - Optimization: First is a greedy forward search, essentially Bayesian regularized forward regression, in which the SNP giving the maximal increase to the model likelihood is added to the model sequentially until all remaining SNPs decrease the likelihood. The second is a similar heuristic, except that the initial model searches through all subsets of 2 SNPs or 3 SNPs.
  - GWiS is designed to select a single model for each gene. An alternative related approach would be to test for the posterior probability of the null model against all other models (BIMBAM). Unfortunately, the number of terms increases exponentially fast with model size, and the brute-force approach does not scale to genome-wide applications.
  - Models tested: minSNP - p-value for the best single SNP; BIMBAM - very similar to minSNP; VEGAS; LASSO - genes with at least one SNP selected.
- Results:
  - The performance of GWiS depends on the genetic architecture of a disease or trait: higher power if genes house multiple independent causal variants, and lower power if each gene has only a single causal variant. In practice, the loss of power was so slight as to be virtually undetectable.
  - Of the other methods, minSNP-P and BIMBAM had similar performance that degraded as the true model included more SNPs. The VEGAS test did not perform well, and the LASSO method performed worst.
- Discussion:
  - By gathering multiple independent effects into a single test, GWiS has greater power than conventional tests to identify genes with multiple causal variants. GWiS also retains power for low-frequency minor alleles.
  - Bayesian methods can be computationally expensive. GWiS minimizes computation by evaluating only the locally optimal models of increasing size in a greedy forward search. This appears to be an approximation compared to previous Bayesian methods that sum over all models
  - Why VEGAS does not perform well? Presumably because the sum over all SNPs creates a bias to find causal variants in LD blocks represented by many SNPs and to miss variants in LD blocks

with few SNPs. Also VEGAS is sensitive to low frequency alleles: its power drops two fold when MAFs drop from 50% to 5%. Reason: low-frequency SNPs lack correlation with other SNPs, reducing the contribution to the VEGAS sum statistic.

  - Why LASSO does not perform well? Also sensitive to low-MAF SNPs. Lasso shrinks the regression coefficients, thus a SNP with large regression coefficients but low MAF may be missed. For the same variance explained, this SNP is penalized more than SNPs with smaller regression coefficients but higher MAF (not a problem for GWiS whose penalty is based on the number of selected SNPs).

- Remark:
  - Lesson: approximation of model likelihood in Bayesian: Laplacian approximation?
  - Lesson: searching for models is an important bottleneck of Bayesian approach, heuristic search.
  - Need to check how well the method performs with rare variants.

MAGENTA: Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits [Segre & Altshuler, PLG, 2010]

- Gene-level statistics: (1) MinP, and obtain gene score (Z-score); (2) Control for confounders: regression of gene scores vs. possible confounders e.g. gene-size. Step-wise multiple linear regression.
- Gene set enrichment: variation of GSEA. (1) Number of genes with p < a threshold, use 95 percentile. (2) Gene set p-value: fraction of genes with p < threshold for a gene set, and compare with random gene sets of the same size.

GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure [Li & Sham, AJHG, 2011]

- Motivation: problems of existing approaches for gene-based test, (1) best single-SNP statistic: not normalize by gene size. (2) Fisher's method (or similar) to combine test statistics of SNPs: independence assumption not hold, thus need simulations to get correct type I error.
- Simest test: let $p_{(1)}, \cdots, p_{(m)}$ be the p-values of the m SNPs of a gene, we define the gene-leve p-value as:
$$p_G = \min_j \frac{m_e p_j}{m_{e(j)}} \tag{4.197}$$
where m_e is the effective number of SNPs, and $m_{e(j)}$ is the effective number of SNPs among the top j SNPs.
- Incorporating prior weights of p-values: similar to the Simes test except that:
$$p_G = \min_j \frac{m_e p_j}{\sum_k^j w_{(k)}} \tag{4.198}$$
where $w_{(k)}$ are non-negative weights summed to m_e.

Incorporating Biological Pathways via a Markov Random Field Model in Genome-Wide Association Studies [Chen & Zhao, PLG, 2011]:

- Idea: if a gene is associated with a disease, other genes in the same pathway tend to be associated as well, since both would disrupt the same pathway. Furthermore, genes that are directly linked in a network should have the same tendancy of association.

- MRF model: given n genes in a network $G = (V, E)$ (edges given), suppose S_i is the association status (binary) of the i-th gene, we first define the prior distribution of S under G:

$$P(S|\theta_0) = \frac{1}{z(\theta_0)} \exp\left[h \sum_i I_1(S_i) + \tau_0 \sum_{(i,j)\in E} (w_i + w_j) I_{-1}(S_i) I_{-1}(S_j) + \tau_1 \sum_{(i,j)\in E} (w_i + w_j) I_1(S_i) I_1(S_j) \right] \tag{4.199}$$

  where θ_0 are hyperparameters, and $I_1(S_i)$, $I_{-1}(S_i)$ are indicator variables, $w_i = \sqrt{d_i}$ is the square root of the degree of the i-th gene. A typical parameter setting $(h, \tau_0, \tau_1) = (-1, 0.25, 0.01)$ (used in simulation) would penalize a lot of associations (prefer simpler model), and reward the edges connecting two associated genes. And the degree parameters w_i would put more influence on highly connected genes.

- Posterior distribution: suppose we have the gene-level statistics y_i (in the experiment, derived from PCA regression at gene level). Assume given $S_i = 0$, $f_0(y_i) \sim N(0, 1)$. Further assume that $f_1(y_i)$ follows a normal distribution $N(\mu_i, \theta_i^2)$, with conjugate priors of μ_i and σ_i^2, we could obtain the distribution of y_i under hyperparameters θ_1. The posterior distribution of S given y is:

$$P(S|y, \theta_0, \theta_1) \propto f(y|S, \theta_1) P(S|\theta_0) \tag{4.200}$$

- Parameters: values chosen based on simulations. However, they could be estimated with an empirical Bayes approach. The marginal likelihood is:

$$L(\theta_0, \theta_1|y) = \sum_S f(y|S, \theta_1) P(S|\theta_0) \tag{4.201}$$

  Thus choose parameters to maximize L. In practice, summing over S is slow, so do maximization over S.

- Inference: Maximize the posterior distribution S. This could be done via maximizing the conditional probability (MCP), the label of S_i given observed data and labels of all the other nodes. FDR control: direct posterior probability approach.

- Experiment:

  - Simulation: given a network, and select genes for labels. Sample genotypes and phenotypes (require parameters of MAF and effect size). The performance is evaluated by the AUC of classifying genes.
  - Crohn's disease: compare the method (posterior means) with the p-value based approach for ranking genes. Improved AUC (Figure 6).

Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets [Xiong & Furey, GR, 2012]:

- Idea: Pathway analysis. Give more weights to genes that are differentially expressed in cases vs. controls.

- Methods: for any gene, define (1) SNP set association score: maximum statistics and (2) differential expression score. The gene association score is some combination of these two scores (e.g. Z-score sum). The pathway association score is then defined as the weighted K-S statistic.

Integrated Enrichment Analysis of Variants and Pathways in Genome-Wide Association Studies Indicates Central Role for IL-2 Signaling Genes in Type 1 Diabetes, and Cytokine Signaling Genes in Crohn's Disease [Carbonetto & Stephens, PLG, 2013]

- Testing enrichment of a single pathway: we do the analysis one pathway a time. For each SNP, let a_j be its annotation (of the test pathway, binary), the inclusion prob. (prior) of SNP j in association studies is: $\text{logit}(\pi_j) = \theta_0 + a_j\theta$. To test if $\theta > 0$, we use BF by comparing two models: $\theta > 0$ vs. $\theta = 0$.
- Testing enrichment of a combination of pathways: after the first, choose multiple pathways, then test combination of pathways. Only OR logic is considered, i.e. set a_j to 1 when SNP j is assigned to at least one of the enriched pathways.
- Use pathway enrichment to redo association analysis: after we have enrichment pathways (or combinations), we compute Posterior inclusion probability (PIP) of each SNP, taking into account the data and prior (annotations). For each SNP, we have multiple PIPs, one from each enrichment pathway. The PIPs are combined by averaging, weighted by BFs.
- Remark: limitations
  - Only consider OR logic when combining pathways. In practice, it is likely that AND of pathways is more important.
  - Combine PIPs from multiple pathways: the results can be easily dominated by single enriched pathway (weighted by BF). The average PIP is never more than each of the PIPs, thus a gene with multiple enriched pathways does not benefit from multiple annotations.

Detecting Association With Network (DAWN) [Li Liu thesis defense]:

- Network construction from co-expression: partial neighborhood selection (PNS) algorithm. First choose a set of likely candidates: nodes with small p-values from TADA, further filter out isolated notes. Next, for each node, choose its neighbors using Lasso (predict the expression of this node). Choose regularization parameter by fitting the power law degree distribution (the best value that leads to scale-free network).
- Hidden MRF (HMRF) model: the Ising prior, however, only reward when both nodes are risk genes and no penalty for non-risk neighbors.
- Remark:
  - PNS algorithm: miss the correlated genes.
  - Scale-free to select network parameters: unproven?
  - HMRF: no penalty term, thus not normalize the degree of nodes.

A Method for Gene-Based Pathway Analysis Using Genomewide Association Study Summary Statistics Reveals Nine New Type 1 Diabetes Associations [Evangelou & Wallace, GE, 2014]

- Goal: gene test and pathway test using only summary statistics.
- Gene test: the statistic has two alternatives. (1) Minimum p-value; (2) Fisher's method of combining p-values:
$$FM = -2\sum_j \log p_j \tag{4.202}$$
In both cases, using MVN to sample the p-values of SNPs under H_0.
- Pathway test: let r_i be the ranks of the genes (divided by the total number of genes), Fisher's method: $-2\sum_i \log r_i$. An alternative is Adaptive rank truncated product method (ARTP): similar to Fisher's method except that only the top H gene will be used.

Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics [Lamparter & Bergmann, review for AJHG, 2015]

- Motivation: both gene-based test and pathway-based test. For gene-based test, VEGAS is a popular method but it relies on simulation to obtain p-values. For pathway-based test, existing methods rely on threshold and hypergeomic test, which depends on the threshold parameter.

- Gene-based test: the basic idea is to obtain the analytic distribution of VEGAS test. Specifically, let Z_i be the Z-score of the i-th SNP, then under null,

$$Z \sim N(0, \Sigma) \tag{4.203}$$

  where Σ is given by the LD. We form the test statistic as the sum of χ^2 over all SNPs:

$$T_{sum} = \sum_i z_i^2 \tag{4.204}$$

  To obtain its distribution, the idea is to conver the vector z to a vector of independent random variables. Let the eigenvalue decomposition of Σ be $\Sigma = \Gamma\Lambda\Gamma^T$, then we define

$$y = \Lambda^{-1/2}\Gamma^T z \tag{4.205}$$

  It is easy to show that $y \sim N(0, I_n)$ ($\Lambda^{-1/2}$ is for scaling). We have:

$$\sum_i z_i^2 = z^T z = z^T\Gamma\Gamma^T z = y^T\Lambda y \sim \Sigma_i \lambda_i \chi_1^2 \tag{4.206}$$

  where λ_i is the i-th eigenvalue. The result is a weighted sum of independent χ^2 distribution. Another test statistic is:

$$T_{max} = \max_i z_i^2 \tag{4.207}$$

  Its distribution can be determined by: $T_{max} \geq t$ iff $z_i \geq \sqrt{t}$ for each i, so this amounts to a rectangular integration over multivariate normal.

- Pathway test: first make the gene-level statistic independent. To do that, do "gene-fusion", i.e. treat close genes (or LD) as a single "fusion gene". Next, to determine the pathway score, convert the p-value of each gene to χ_1^2 statistic, then use the sum as the test statistic of the pathway. Two variations: chi-squared-method or empirical sampling method. The difference is likely: (1) chi-squared-method assumes independence of genes (so that the null distribution is valid); (2) gene-score p-values are determined empirically, instead of from theoretical null (more conservative).

- Results: fusion method avoids the inflation problem (using simulated random phenotype data to show that) of existing pathway tests. The pathway test also is more powerful than existing pathway tests, hypergeometric or rank-sum: using replication type of analysis (the truth from a larger dataset), and using the number of significant pathways as the metric.

- Lessons: (1) The analytic distribution of sum of chi-square: convert to independent random variables; (2) Binary test loses information and power.

Biological interpretation of genome-wide association studies using predicted gene functions (DEPICT) [Pers and Franke, NC, 2015]

- Reconstitute gene sets: about 14,000 gene sets, each gene has a membership probability.

- Gene prioritization: get a set of genes in the trait-associated loci S. For each gene, a vector representing its membership probability of all gene sets. Then scoring one gene: correlation of the membership prob. vector with all genes in S.

- Gene set enrichment test: for any gene set, sum over the membership over all genes in S, then test the significance.

- Remark: a poor mans way of doing Bayesian hierarchical model: learn likely gene sets from all trait-associated loci; then score a gene by its similarity with likely genes.

MAGMA: Generalized Gene-Set Analysis of GWAS Data [de Leeuw, PLG, 2015]

- Procedure: (1) Gene level analysis: obtain gene p-values. (2) Pathway analysis: use gene p-values and gene correlation matrices (accounting for LD between genes).
- Gene-level analysis using individual level data: for SNPs in a gene, do PC first, and use the top PCs to represent the gene. Regression of PCs with phenotype, and use F-test for the gene, H0: no PC is associated with phenotype.
- Gene-level analysis using summary statistics: use mean or max. chi-square.For mean χ^2: an approximate distribution is available. For max χ^2: use permutation - permute phenotype labels and do association.
- Gene set analysis: let Z be Z-score of a gene, regression of Z vs. features of genes. Self-contained: use only one gene set a time, test if deviation from $N(0,1)$. Competitive test: use gene set as a feature, do regression analysis.
- Generalized gene set analysis: default in MAGMA, testing gene set, conditioned on gene size, gene density (number of SNPs, or number of genotype PCs).
- To account for LD between genes, the error model of Z are correlated, using Generalized Least Square.$\epsilon \sim MVN(0, \sigma^2 R)$. The correlation matrix R: approximated by using the correlations between the model sum of squares (SSM) of each pair of genes from the gene analysis multiple regression model, under their joint null hypothesis of no association.
- Lesson/Remark: dimensionality reduction in testing variable set association. What are alternative approaches? E.g. how to do testing with Lasso or Bayesian? Perhaps G-prior for Bayesian variable selection?

Pathway analysis using RSS [Xiang Zhu, 2015]

- RSS model: the likelihood is given by RSS model:

$$\hat{\beta}|\beta, S, R \sim N(SRS^{-1}\beta, SRS) \tag{4.208}$$

  The model is applied to LD blocks. Use large LD blocks with shrinkage estimation s.t. the LD matrix is block-diagonal. The LD blocks tend to be large: some many include > 100 genes, a total of about 1,000 blocks.

- Testing association of pathway: let β_j be the true effect of variant j. Use sparse parior

$$\beta_j \sim (1-\pi_j)\delta_0 + \pi_j N(0, \sigma^2), \quad \text{logit}(\pi_j) = \theta_0 + a_j\theta \tag{4.209}$$

  where a_j is the pathway annotation of j: 1 if j belongs to the pathway being test, and 0 otherwise. The results are $P(\theta|\hat{\beta}, S, R, a)$, the BF of pathway (if $\theta = 0$) and evidence of each SNP/gene $P(\beta_j|\hat{\beta}, S, R, a)$.

- Computational problem: need to integrate out β_j, but this cannot be done analytically, so rely on Variational Bayes.
- Analysis of height GWAS: $\theta_0 = 2.0$ (prior 0.01) and $\theta = 0.75$ (OR = 5.6) for the top pathways. Note: compare this with Carbonette & Stephens: the posterior was inflated, $\theta_0 = 0.001$ and $\theta = 10^3$ for T1D.
- Multiple pathways: if a gene/SNP belongs to multiple pathways, only use the strongest pathway, instead of additivity assumption.

4.6.4 Other Tests

A Novel Test for Recessive Contributions to Complex Diseases Implicates Bardet-Biedl Syndrome Gene BBS10 in Idiopathic Type 2 Diabetes and Obesity [Lim & Daly, AJHG, 2014]

- Model: our data are the counts of aa in cases and in controls (also the counts of non-aa). The expected frequency of aa in cases depends on the population frequency of aa and the relative risk. We derive this expected frequency for cases (controls are simpler), then perform the LRT using binomial distribution. In the LRT, the parameter γ (RR) needs to be determined.
- Estimating $P(aa)$: simplest case, $P(aa) = P(a)^2$. Two exceptions:
  - Population substructure that causes systematic departure from HWE: $P(aa) = FP(a) + (1 - F)P(a)^2$.
  - Local departure from HWE due to hemizygous deletions or systematic genotyping errors.

  If the observed rate of aa in control subjects exceeds the expected corrected $P(aa)$, simply use the observed rate (to be conservative).
- Remark: the power comes from estimated $P(aa)$. Normally this would be $P(a)^2$, which is very small for rare variants.
- Questions: the parameters under the model: (1) population frequency (would be a problem for rare variants), e.g. aa might have 0 count in controls. (2) Relative risk: maximization? (3) Why need EM?

Benchmarker: An Unbiased, Association-Data-DrivenStrategy to Evaluate Gene Prioritization Algorithms [Fine and Hirschhorn, AJHG, 2019]

- Background: limitations of current approaches for evaluation of gene prioritization algorithms. (1) Gold standard gene sets: often hard to obtain, and may be biased towards well-characterized genes. (2) Independent GWAS datasets: not always available.
- Benchmarker strategy (Figure 1): compare several methods, use LDSC to estimate the enrichment of heritability in the prioritized genes. (1) Obtain list of prioritized genes: take one chromosome (e.g. chr1), train on all other chromosomes, then prioritization on chr1 (correlation of membership vector vs. genes in trait-associated loci in other chr's). (2) For all prioritized genes: assess enrichment of h2 using S-LDSC. Use SNPs near genes.
- Results: three variations of DEPICT, performance of gene sets > GEO co-expression and GTEx co-expression. Relatively small overlap. Intersection of genes by multiple methods performs better.
- Comparison of DEPICT vs. MAGMA: similar performance, but gene overlap is not high. Intersection much better than each alone.
- Comparison with NetWAS: NetWAS uses PPI network to prioritize genes. Both DEPICT and MAGMA significantly better than NetWAS.

4.7 Incorporating Variant Annotations

Strategies of testing and incorporating annotations [personal notes]

- Basic strategy: if a feature is predictive of causal variants, then the variants with this feature should generally have higher signifiance in GWAS than those without this feature.
- Complication: (1) LD across variants; (2) bette to compare the distribution of summary statistics: enrichment analysis often requires discretization.

- Analysis of how LD may affect the enrichment analysis (in the general sense): two scenarios where LD can lead to inflation of enrichment:
  - A causal SNP in an enhancer, and the enhancer contains multiple SNPs in LD. Then a causal SNP may contribute multiple times in computing the fold enrichment.
  - A SNP in an enhancer, and in LD with a causal SNP (not in an enhancer): proper analysis should remove the enhancer-SNP as its effect is due to causal SNP.
- Statistical strategy: general idea is that the annotations increase the prior of a variant (indicator or effect size). And the likelihood of data (summary statistics) depends on the indicators or effect sizes.
- Polygenic model: the first scenario, the model would assume that the enhancer contains multiple causal SNPs, so each one would receive lower effects. In the second scenario, the model would infer that the true causal SNP (not in enhancer) has larger effect, and that may be due to LD with enhancer-SNP. The model does allow one to explain "away" SNPs in LD with causal SNPs that are in enhancers.

Review of methods [personal notes]

- fgwas: one causal variant per LD.
- PAINTOR: use only associated loci. Model distribution of test statistics (Z-scores) as MVN.
- LD-score regression: annotations change effect size variance. Cannot model penalty.
- CAVIAR: only fine-mapping, finding the interval that covers all causal variants with high probability.
- CAVIAR-BF: similar to CAVIAR, combine the idea of BIMBAR.
- Schaid paper: similar to PAINTOR, with different prior for causal indicators (L1, L2 and elastic net).
- Kellis group [Yue Li]: logistic prior, the innovation is cross-trait. However, only capture cross-trait correlation via annotation weights. Should model how effect sizes are correlated.
- PICS.

Enriching the analysis of genomewide association studies with hierarchical modeling: [Chen & Witte, AJHG, 2007]

- Idea: for each SNP, also consider its extra information: conservation, functional category, etc, which can weight SNPs. To assess the effect of the extra information, use a hierarchical model where the effect of a SNP depends on these general evidences.
- Model: for M SNPs, let β_m be the effect of the m-th SNP (regression coefficient). Then β_m depends on the K factors, including conservation (PhastCons score), the functional category (mRNA, UTR, or intron, etc.), and all these types of information in adjacent SNPs. Model β_m through a regression:

$$\beta_m = \sum_{j=1}^{K} \pi_j Z_{mj} + u_m \tag{4.210}$$

  where π_j is the effect of the j-th factor, Z_{mj} is the j-th factor of the m-th SNP, and u_m follows normal distribution with mean 0. The variance of u_m thus indicates heterogeneity of β_m. The model can be extended to allow dependence among SNPs: u_m are no longer independent, instead U follows a multivariate normal distribution.
- Result: test association of SNPs and gene expression, choose a few genes where the causal SNPs are known (from functional annotation, e.g. in promoters). Compare the ranking of SNPs from single-marker analysis and the current method: those close to the true causal SNP should rank higher in the new method.

- Remark: the hierarchical model here is unusual, β_m is modeled separately for each SNP. A probably better way is the model in [Veyrieras & Pritchard, PG, 2008]

Disease gene identification by PPI [Lage & Brunak, NBT, 2007]:

- Goal: rank candidate genes in the linkage interval of diseases.
- Methods:
  - Idea: if a protein X is involved in a disease, then a protein interacting with X may be involved in a related disease.
  - Data: training data - 963 genes involved in 1,404 distinct phenotypes. For each phenotype, construct a linkage interval (randomly chosen the size s.t. the average number of genes in one interval matches the data in the application data). Application data: 870 intervals with no candidate genes assigned.
  - Disease similarity: from literture mining - similarity of the associated profiles of semantic terms (e.g. UMLS).
  - Gene scoring in a linkage interval: for each gene, extract all interacting partners, and score a gene according to the number of partners that are known to be associated with similar diseases (thus considering both PPI strength and disease similarity). The score is a Bayesian probability that the gene is involved in the disease given the data (PPI and partner-disease association).

  Results:

  - Evaluation: precision and recall in 1,404 phenotypes.
  - Application: on 870 unassigned intervals. Assign high confidence genes to 91 intervals.

All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs [Shork, Dale, PLG, 2013]

- Enrichment of loci in different categories: consider exon, 5' UTR, 3' UTR, upstream sequence (1k or 10k) and downstream sequence. Stratified Q-Q plot to show the pattern of enrichment of signal in different categories.
- Quantification of enrichment: true discovery rate (TDR) is estimated as $1-p/q$, where p is the fraction of expected SNPs under a given thrshold and q the observed number. Another measure is: sample mean z^2-1.
- Stratified FDR (s-FDR) approach: effectively use different π_0 (or threshold) to control FDR for each category. Show that at $\alpha = .05$, the increased proportion of SNPs due to sFDR ranges from 20% (height) to 300% (SCZ).

Partitioning heritability by functional category using GWAS summary statistics: Stratified LD score regression (S-LDSC) [Finucane & Price, NG, 2015]

- Assumption: the heritability of a SNP is the sum of the heritabilities of the categories that this SNP belongs to.
- Model: suppose both genotype and phenotype are standarized, define heritabilty of a category C as $h^2(G) = \sum_{j\in C}\beta_j$. The summary statistic is $\hat{\beta}_j = X_j^T y$, plug in $Y = X\beta + \epsilon$, we have:

$$\hat{\beta}_j = \sum_k \hat{r}_{jk}\beta_k + \epsilon'_j \tag{4.211}$$

where $\hat{r}_{jk}$ is the LD between SNP j and k, and ϵ'_j has mean 0 and variance σ_e^2/N. The assumption of the effect size: it has mean 0 and variance

$$\operatorname{Var} \beta_j = \sum_{c:j\in c} \tau_c \tag{4.212}$$

where c is an index of category. This allows us to derive the expectation of $\chi_j^2 = N\hat{\beta}_j^2$. The results:

$$\mathrm{E}\left(\chi_j^2\right) = N \sum_{c:j\in c} \tau_c l(j,c) + 1 \tag{4.213}$$

where $l(j,c)$ is the weight of category c, $l(j,c) = \sum_{k\in C} r_{jk}^2$. This is a system of equation of τ_c and we solve it via multiple regression. Additional details of the regression: e.g. weighting SNPs, block jackknife for obtaining standard errors.

- How would the model work? Suppose we have only two categories (one feature): then for SNPs not in LD with this category, its $\tau_C = 0$, thus we expect generally lower χ_j^2. For SNPs in LD with this category, we would expect higher χ_j^2.
- Results: enrichment of main annotations across 9 phenotypes (Figure 4). The enrichment is defined as proportion of heritability of a category divided by the proportion of SNPs. The strongest categories are conserved (12), enhancer (4), fetal DHS (3), coding (7), etc.
- Figure 7. comparison of methods for testing enrichment, causal proportion is 0.005, fgwas performs poorly (very low power).
- Remark/questions:
  - Additivity of effect size variance: cannot model the negative effect (some annotations reduce the effect size). Additivity assumption cannot capture AND logic (interaction terms) - similar to logistic regression, but it is relatively easier to incorporate interaction terms.
  - Distribution of χ^2: the model fitting part uses linear regression, which is based on normal error model (not true here). Because of this, the confidence interval is obtained from jackknife.

Weighting sequence variants based on their annotation increases power of whole-genome association studies [Sveinbjornsson & Stefansson, review for NG, 2015]

- Motivation: incorporate weights of variants in multiple testing, where weights depend on annotations (damaing effects).
- Intuition: if we can collect all causal variants, and their categories of annotations, we can simply calculate the enrichment of causal variant in each category. Since we do not know, we can collect all associated loci, and then model the uncertainty of causal variants.
- Estimating category enrichment: assume all variants are partitioned into disjoint categories. For category c, we define q_c be the probability of a non-causal variant being in category c, and p_c the probability of causal variant being in c. The enrichment is defined as $e_c = p_c/q_c$. To estimate this enrichment, we model the likelihood of data (summary statistics). Suppose we are given association loci, and assume each locus contains exactly one causal variant (let k_i be the index/position of the causal variant of locus i). Let c_m be the category of a variant at m. The likelihood at locus i:

$$P(y|g) = \sum_m P(y|g_m, k_i = m)p(k_i = m) \tag{4.214}$$

where $p(k_i = m)$ is given by:

$$p(k_i = m) \propto p_{c_m} \prod_{m'\neq m} q_{c_{m'}} \tag{4.215}$$

as $k_i = m$ iff m is causal (with probability p_{c_m}) and all other variants are non-causal (with probability $q_{c_{m'}}$). Effectively, we treat categories a variant belongs to as data, and model the probabilities of these "category variables". We multiply this likelihood over all loci and do MLE.

- Weighted Bonferroni correction: for a variant j, use weight $w_j = e_{c_j}$ where c_j is the category of j. It is easy to prove that sum of w_j is 1 (actually sum of p_c, over all categories). Then follow KMR's weighted multiple testing procedure.

- Data: WGS of 2,636 Icelanders, 96 quantiative and 123 case-control phenotypes. Use MAF $> 0.1\%$, find 14.2 M variants. Bonferroni correction threshold would be 3.5×10^{-9}.

- Results of category enrichment: first find all variants with $p < 10^{-8}$, then for each of them, extract LD variants with $r^2 > 0.2$. A total of 700 or so association signals (loci). Fitting the model lead to estimation of enrichment. LoF: 186, missense: 51, synonymous: 6.2, upstream/downstream 5k: 5.0, 5' or 3' UTR: 2.8, intronic: 0.6, intergenic: 0.4. Regulatory annotations: DHS - 2.4 and 3.6; enhancers (DHS and ChromHMM) - 7.4 and 7.0.

  - The confidence interval for synonymous and UTR are reall large and so the enrichment of these two categories may not be significant. It's likely that this is due to LD: they are very close to nonsyn or LoF variants, and so their effects are harder to disentangle from nonsyn. variants.
  - Conservation: CADD scores > 5 show no enrichment, GERP scores show modest enrichment: 1.4 fold with scores ≥ 2 in non-coding sequences.
  - Comparison with LD score regression: the main difference is LD score regression considers only coding variants, while this method divides them into LOF, missense and synonymous, which have very different enrichments.

- Results of association analysis: first find all settled associations ($p < 10^{-10}$), then among the rest, use standard Bonferroni finds 146, and the weighted finds 172. Thresholds reduced, e.g. LoF, 6.3×10^{-7}. Almost all new coding variants are real.

- Remark: the enrichment defined in this way is similar/equivalent to the enrichment in [Gusev, AJHG, 2014], which define enrichment as: the probability that a variant is causal given that it is in category c vs. the probability that a variant is causal. Let Z_j be the indicator variable of variant j, then $p_c = P(j \in c | Z_j = 1)$ and $q_c = P(j \in C | Z_j = 0)$, we have:

$$P(Z_j = 1 | j \in c) = \frac{P(Z_j = 1) P(j \in c | Z_j = 1)}{P(j \in c)} \Rightarrow \frac{P(Z_j = 1 | j \in c)}{P(Z_j = 1)} \approx \frac{p_c}{q_c} \tag{4.216}$$

- Remark: the method of estimating enrichment requires a large number of association loci, so it cannot be used to estimate the enrichment for individual GWAS data.

- Lesson: frequentist framework of incorporating priors, weighted multiple testing. To determine weights, use something similar to empirical Bayes. In the GWAS context, the enrichment of causal variants in a particular category can be directly used as weights.

Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci (GoShifter) [AJHG, 2015]

- Motivation: enrichment analysis that control for gene density, LD (trait-associated SNPs often mapped to high LD regions). Also colocalization of annotations.

- Local annotation shifting: (1) Assessing overlap of GWAS SNPs with annotation X: for each GWAS index, extend to SNPs in LD $r^2 > 0.8$. Estimate the proportion of these SNPs overlapping X. (2) Obtain null distribution: the idea is to randomize the positions of X. This is achieved by shifting X within each locus (defined by GWAS SNPs and LD proxies).

- Stratified annotation shifting: testing enrichment of X conditioned on another annotation Y.
- Application: use GWAS significant SNPs associated with height, RA and breast cancer.
- Remark: the method does not account for LD uncertainty, and it only uses genome-wide significant SNPs.

GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach [Schmidt and Willer, Bioinfo, 2015]

- Overview: LD pruning to obtain a set of independent GWAS index SNPs. Next we evaluate the overlap of these SNPs with a feature and compare that with a set of matched SNPs. To account for LD: expand to LD proxies (r^2¿0.7), only for testing if a SNP overlap with a feature.
- Method (Table 1): let s be the number of index SNPs overlapping a feature (where overlapping allows LD proxy). The problem is to obtain the significance of s. For each index SNP, we obtain a set of m matched SNPs with the same LD, MAF and gene distance. Then for SNP set i, by chance index SNP i will overlap the feature with probability p_i, which is the proportion of feature overlap among (m+1) SNPs in SNP set i. Then the null distribution of s is the sum of Bernoulli RVs $\sum S_i$, where $S_i \sim \text{Bern}(p_i)$. Show that significance from permutation results are similar to analytic results.

GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction (GARFIELD) [review for NG, 2016]

- Goal: testing enriched features/annotations among GWAS loci.
- Input: GWAS summary statistics, annotations, LD statistics and MAF, TSS distance.
- Data preparation: annotations of independent set of SNPs
  - LD pruning: the goal is to find an approximately independent set of SNPs. Two SNPs are independent if $r^2 < 0.1$ We start with most significant SNPs, remove variants with $r^2 > 0.1$ and within 1Mb (all SNPs in LD). And repeat this process with the next significant SNPs. About 6% of variants were left after this step.
  - Annotation: a variant would have a feature if any of its tagged SNPs $r^2 > 0.8$ and within 500kb (including itself) has that feature.
- Testing feature enrichment: choose SNPs at given p-value thresholds from 0.1 to 10^{-8}. Assessing statistical significance of fold enrichment: based on permutations, match MAF, distance to TSS and number of tagged LDs ($r^2 > 0.8$). Specifically, for N variants, we have N p-values, we just permute these p-values. But to control for confounders, match SNPs in MAF, etc, (125 bins) and only permute SNPs within a bin.
  - Comparison with Maurono: (1) LD pruning; (2) null set: match features. In Maurono, only LD tagging association.
- Multiple testing correction: take into account the correlation among features. Estimate the number of independent annotations, then Bonfonerrino correction.
- Joint modeling of multiple annotations (in Revision): choose a threshold T for SNP p-values. Do a logistic regression of SNP association (passing threshold or not) with annotations, adjusting for MAF, TSS, and number of proxies (LD). Model selection: forward variable selection.
- Data: GWAS of 27 phenotypes including 3 diseases and 24 quantiative traits.
- Remark: Analysis of procedure: If LD pruning is too aggressive, many causal SNPs may be lost. Scenarios:

  - Most sign. SNP is causal, but there is another causal SNP (in enhancer) in LD $r2 = 0.2$ with the causal SNP.
  - Start with most significant SNPs: the assumption is that they are causal SNPs. Possible that SNP in an enhancer (causal), but a nearby SNP ($r2 = 0.7$) has higher significance. Then after pruning, causal SNP is lost. In tagging step, the highest SNP would not receive the annotation.

  Other confounding variables in fold enrichment analysis:

  - LD: in high LD regions, better tagging, thus likely to have higher effects and higher density of significant SNPs. High LD might also be correlated with enhancers.
  - MAF: higher MAF means higher power, and thus likely to have higher density of sig. SNPs. It may also be correlated with enhancers: e.g. mutation rates may be higher, and thus affect the mutation age of alleles (hence MAF).
  - Distance to TSS: near TSS, likely more sig. SNPs, and also enriched with enhancers.
  - Additional ones? Ex. GC content of regions, mutation rates.

  GARFIELD analyzes effectively one annotation a time (another reviewer), and this may cause false discoveries when annotations are highly correlated.

- Remark: the main difference with other approaches is: obtain a set of approximately independent variants with the annotation (LD tagging), then in null simulation, also match the independent variants.

Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types [Finucane and Price, NG, 2018]

- Stratified LD score: $\beta_j = N(0, \tau_0 + \sum_k \tau_k I(i \in C_k)$, τ_0 is the average effect. Continuous extension of LD score: $l(i,k) = \sum_j a_k(j) r^2(i,j)$.
- Gene expression matrix: for each gene in a tissue, t-test, regression of expression over tissue (category, e.g. brain). For every tissue, choose top 10% genes as tissue-specific genes. Then expand to 100kb nearby.
- Test tissue-specific annotation, controlling promoters, coding and other general annotation.
- Validation with chromatin marks: use all sequences active in a tissue.
- Results: BMI, only brain is enriched. T2D, only pancreas, but not adipose. The test likely has low power.

4.7.1 Fine Mapping with Variant Annotations

Joint analysis of functional genomic data and genome-wide association studies of 18 human traits (fgwas) [Pickrell, The American Journal of Human Genetics. 2014]

- Goal: test enrichment of a certain annotation (e.g. enhancers in a particular tissue type) in GWAS, and inference of indidivual SNPs ("re-weighting" based on functional annotation).
- Idea: an indicator variable of each SNP (or block, considering LD), and the prior of the indicator depends on functional annotation. Similar to Sherlock, where the prior depends on the status of eQTL.
- Model: let y be the data, the summary statistics of all SNPs. We divide the genome into blocks s.t. the blocks can be considered independent (2.5Mb on average). For the k-th block, let Π_k be the prior that it contains one causal SNP, the likelihood:

$$P(y) = \prod_k \left[(1 - \Pi_k) P_k^0 + \Pi_k P_k^1\right] \tag{4.217}$$

where P_k^0 and P_k^1 are the probabilities of the block under null and alternative hypothesis. Within a block, we assume there is only one causal SNP, then the probability of a block is a sum of probability over all SNPs, weighted by the prior of SNPs:

$$P_k^1 = \sum_k \pi_{ik} P_{ik}^1 \tag{4.218}$$

where π_{ik} is the prior of SNP i in block k and P_{ik}^1 is the prob. of this SNP under H_1. The prior depends on the functional annotations of SNPs or blocks. For the k-th block:

$$\text{logit}(\Pi_k) = \kappa + \sum_l \gamma_l I_{kl} \tag{4.219}$$

where I_{kl} indicates if block k has the l-th block-level annotation, and γ_l is the effect of l annotation. Similarly we can define the prior of π_{ik}.

- Computing BFs: the model would require computation of BFs per SNP. This is done by Wakefield's approximation, which depends on the estimated β, its standard error V_i, and a prior of the effect size: $\beta \sim N(0, W)$. In the paper, the prior parameter W is fixed at 0.1 (small effect sizes). For case-control studies, approximate V_i from Wakfield paper.
- Inference: penalized log-likelihood (penalty because many annotations are used). The penalization parameter is learned using cross-validation: split all chromosome region 10-fold.
- Reweighting: for any SNP, estimate its posterior prob. of association (PPA), that uses annotations as prior. The PPA is effectively "reweighting" of GWAS summary statistics. The PPA threshold (0.9) is chosen so that the true discovery rate is comparable to Bonferroni correction.
- Data: use ImpG to impute the summary statistics of 18 traits. Annotations: 450 including 402 DHS maps.
- Enrichment or depletion of GWAS loci: use HDL GWAS as an example, the enriched categories are enhancers, exons, TSS. Other examples: for platelet volume/count, open chromatin annotation is enriched with $\log_2$ enrichment betwen 2 and 3. A general pattern: repressed chromatins are depleted of GWAS loci: $\log_2$ from -1 to -2.
- Results of reweighting: increaes of 5% of more SNPs from using annotations.
- Question: how to do cross-validation for general estimation problems when there are no labels involved?
- Remark:
  - The model assumes one causal variant per region, which is much bigger than a single LD region (5,000 SNPs). Furthermore, the only regional level annotation seems to be gene-density, which is not very informative.
  - LD issue: the method is not aimed for fine-mapping, thus it does not model the LD structure of a locus.
  - Robustness to the effect size prior W: the paper shows the estimation of annotation parameters is robust, but what about PPA for individual SNPs?

Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies (PAINTOR) [Kichaev & Pasaniu, PLG, 2014]

- Motivation: integrating functional annotation and GWAS summary statistics to do fine-mapping. The difference with existing approaches: using summary statistics; aiming at fine mapping (fgwas does not model LD).

- Model: consider a locus, let Z be the summary statistics of all SNPs, and A be annotations. Define C as the latent indicator variable for which SNPs are causal (at most three), the distribution of C depends on A through a logistic model with parameters γ. The distribution of Z given C follows MVN: for the alternative model, it depends on the effect size parameter λ (or non-centrality parameter). To simplify, assume λ is equalt to Z-scores. The likelihood per locus:

$$P(Z|\gamma, \lambda, A) = \sum_C P(Z|C, \lambda)P(C|\gamma, A) \tag{4.220}$$

- Comparison of PAINTOR and fgwas for estimating annotation parameters: similar results, but fgwas is less efficient (higher standard error) and does not always converge.

- Results of fine-mapping: by using annotations, reduce the number of variants per locus from an average of 17.5 to 13.5 (90% confidence set).

GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation [Chung & Zhao, PLG, 2014]

- Motivation: use both multiple GWAS datasets and annotations to improve the finding of causal variants.

- Model: let P be the p-values of SNPs, and A be annotations. The idea is that causal variants have different distributions of P and A comparing with null variants. Let Z be the true underlying association variables. When we consider multiple phenotypes, Z of each variant is a configuration (true for one trait, false for another, etc.). Let j be an index of SNPs, the likelihood:

$$P(P, A) = \prod_j \sum_l P(P_j|Z_j = l)P(A_j|Z_j = l) = \prod_j \sum_l P(P_j|Z_j = l) \prod_d P(A_{jd}|Z_j = l) \tag{4.221}$$

  where l is one configuration of association states. The distribution $P(P_j|Z_j)$ is: uniform under null model and Beta distribution under alternative model. The distribution $P(A_{jd}|Z_j)$ follows Bernoulli distributions.

- Application: five GWAS data of psychiatric diseases, and annotations include CNS gene (expression in CNS), eQTL and TFBS.

- Remark: the model assumes conditional independence of multiple annotation dataset, and this does not work well when there are mutliple correlated annotation data.

Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation [Wen, PLoS Genet, 2015]

- Model: same as Torus model [Wen, AJHG, 2016].

- EM-MCMC algorithm: let G, Y be genotype and phenotype data, and D be annotation data. Let Γ be the configurations, and α be the enrichment parameters of the annotations. The algorithm computes the MLE of α, treating Γ as missing data. The complete data log-likelihood is given by:

$$\log P(Y, \Gamma|G, D, \alpha) = \log P(\Gamma|D, \alpha) + \log P(Y|G, \Gamma) \tag{4.222}$$

  Now we take expectation over $\Gamma|Y, G, D, \alpha^t$ to compute $Q(\alpha|\alpha^t)$. Note that the last term does not have α in it, so it's a constant term when maximizing α, so we can ignore it. The resulting algorithm is: use MCMC to sample $\Gamma|Y, G, D, \alpha^t$, and then take expectation (PIP) and do logistic regression of PIP vs. annotations. See Section S.2 in Text S1.

Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors (Torus and DAP) [Wen, AJHG, 2016]

- Model: let γ be the configuration, the prior of $\gamma = 1$ is related to the annotation by logistic regression, with α the enrichment parameters.

- Inference: (1) Estimation of α: using EM algorithm, treating γ as missing data; (2) Locus level discovery; (3) Fine-mapping on these loci: compute $P(\gamma|D, \alpha)$, where D is full data and α enrichment/prior parameters.

- EM algorithm for parameter estimation: in E-step, the method computes PIP for each SNP; in M-step, regression of PIPs (response variables) vs. variant annotations. Note that in the E-step, to compute PIP, we essentially need to fine-mapping for each block. This is difficult, so this computation is done either by MCMC (earlier work, MCMC-embedded EM), or by DAP-1. In DAP-1 approximation, the PIP of SNP i is given by Equation C2. Let π_k be the prior of SNP k, and B_k be its BF. Then the PIP is:

$$P(\gamma_i = 1|y, G, \alpha) = \frac{\sum_k \pi_k B_k}{1 + \sum_k \pi_k B_k} \cdot \frac{\pi_k B_k}{\sum_k \pi_k B_k} \tag{4.223}$$

  The first term is the probability that there is at least one causal variant, and the second term is the probability of SNP i is causal given that there is at least one. Note: in the denominator of the first term, we have 1 instead of $1 - \sum_k \pi_k$, this is because null model and single-effect model differs by a constant π_k.

- Approximation of posterior: our goal is to infer $P(\gamma|D, \alpha)$. This is given by:

$$P(\gamma|D, \alpha) = \frac{P(\gamma|\alpha)BF(\gamma)}{\sum_{\gamma'} P(\gamma'|\alpha)BF(\gamma')} \tag{4.224}$$

  where $BF(\gamma)$ is the BF of a configuration. The difficulty is to evaluate the normalizing constant C (denominator), which involves summing over all γ's. This is done by considering only models (set of causal SNPs) that cover most of the probability mass.

- Adaptive DAP algorithm: let s be the size of a model. We need to evaluate the normalization constant:

$$C_s = \sum_{\|\gamma\|=s} P(\gamma|\alpha)BF(\gamma) \tag{4.225}$$

  Suppose we have Ω_s, the set of models whose size $\leq s$. For next step, we add a SNP if its posterior (conditioned on all chosen SNPs) $\geq \lambda$, with default $\lambda = 0.01$. Stopping condition: if for a value of s, adding more SNPs does not change C much.

- Behavior of DAP (notes): suppose we have three causal variants, A, B and C. In addition, A' is in close LD with A. At $s = 2$, we should choose: $(A, B), (A, C), (B, C), (A', B), (A', C)$. If we choose only the best model at each step, which is A at $s = 1$, we will miss $(A', B), (A', C)$. In general, a model reported by DAP should be conditionally independently.

- Simulation results: 1,500 genes, with 50 cis-SNPs each, $n = 343$. Use $\alpha_0 = -4$. Vary α_1 (enrichment parameter). Compare: best-case: use true labels to regress with annotations in the EM step; DAP-1 and adaptive DAP. Results: for α_1 estimate, DAP-1 has somewhat larger SE than adaptive DAP, but not much larger (Fig. 1), both are roughly unbiased. α_0 is slightly under-estimated, -4.6 vs. -4.0 - expected because of power limitation.

- **Remark**: in real data, causal SNPs are not uniformly distributed, they are likely clustered around causal genes. DAP-1 approximation may significantly underestimate α_0, but it is not clear its estimates of enrichment parameters are biased.

Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics (Caviar-BF) [Chen and Schaid, Genetics, 2016]

- Model: similar to DAP, $y = X\beta + \epsilon$, where $\epsilon \sim N(0, 1/\tau)$. Let configuration be c, use logistic prior for c_i:

$$\log \frac{P(c_i = 1|A, \gamma)}{1 - P(c_i = 1|A, \gamma)} = \gamma_0 + A_i\gamma_1 \tag{4.226}$$

  where A_i are annotations of SNP i (vector). Use normal prior for γ_1: $N(0, \lambda^{-1})$, where λ is penalization parameter. The prior effect size variance is defined as $N(0, \sigma_a^2 1/\tau)$.

- Summary statistics version: BF calculation, the BF of a configuration c depends on S (variance of genotypes), Σ_X (LD), z-scores, N, sample size and ν, the prior effect size, or $\beta_i \sim N(0, \nu 1/\tau)$ for causal variants.

- MAP estimation of γ (annotation parameters): EM algorithm, the M step is effectively penalized logistic regression. The term in the E-step is the PIP of SNPs.

- Selection of annotations: use L2, L1 and elastic net. Parameters chosen by BIC, AIC or Cross-validation. Shown that BIC has too much penalty, AIC better, and CV best.

- Multiple loci: for individual level model, could be multiple loci of the same trait or same locus of multiple traits - the former has different regression model (adjusting all other loci when fine-map one). For summary level data, two scenarios are the same.

- Running time: exhaustive search of SNPs, 13 hours to fine-map a locus with 3 causal variants.

- Results: comparison of causal variants identified, show CAVIAR-BF with elastic network, 5-fold CV, has the best performance. Q: how is this defined? Likely by PIP.

- Calibration of PIPs: compare the true proportion of causal variants vs. average PIPs, show that PAINTOR has inflated PIP.

A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies (bfgwas) [Yang and Abecasis, AJHG, 2017]

- Background: PAINTOR is very slow, can be used to fine-map small regions only.

- Model: $Y = X\beta + \epsilon$, where $\beta_i \sim \pi_i N(0, \tau^{-1}\sigma_i^2) + (1 - \pi_i)\delta_0$. Consider K non-overlapping annotations, and all variants in category q share the same (π_q, σ_q^2). Let A be annotations, our goal is to infer $\pi, \sigma, \gamma, \beta|Y, X, A$, where γ is the indicators of causal variants (configuration). The posterior is given by:

$$P(\pi, \sigma^2, \beta, \gamma|Y, X, A) \propto P(\pi)P(\sigma^2)P(\gamma|\pi, A)P(\beta|A, \gamma, \sigma^2)P(Y|X, \gamma, \beta) \tag{4.227}$$

- EM-MCMC idea: given π, σ^2, we can sample γ, β for each genomic block independently (OK if each blocks explain small phenotypic variation). And once γ, β are obtained (some form of summary, e.g. posterior mean), we can estimate π, σ^2 efficiently (similar to MLE).

- E-step: MCMC of γ, β given π, σ^2. The conditional distribution of γ, β:

$$P(\beta, \gamma|Y, X, \pi, \sigma^2) \propto P(\gamma|\pi)P(\beta|\gamma, \sigma^2)P(Y|X, \gamma, \beta) \tag{4.228}$$

  This the Bayesian linear regression and we can integrate out β analytically. The problem then becomes sampling from the conditional distribution of γ. At each step, the proposal distribution does one of three things (with prob 1/3 each):

  - Randomly add a new variant, with higher probability of sampling top SNPs based on marginal association statistics.
  - Randomly delete a SNP in the current γ.

– Randomly switch a SNP in the current γ: replace it with a neighboring SNP. The selection of SNP is based on conditional association statistics of the SNPs (conditioned on all current SNPs except the switch SNP).

From the MCMC, we can be obtain the posterior summary (mean) of γ and β.

- M-step: inference of π, σ^2 given the results from the E-step. We obtain the MAP first, and then use Fisher information to approximate the posterior. We have the conditional posterior of π as:

$$P(\pi|\gamma) \propto P(\pi)P(\gamma|\pi) \tag{4.229}$$

The expected log-posterior-likelihood of π can be obtained by integrating out γ above. So the objective function of π also depends on the posterior mean of γ for each SNP. Similarly, for σ^2, we express its expected log-posterior-likelihood in terms of posterior mean of γ and β and maximize.

- Implementation: typically 5 iterations of EM, and 50K MCMC per block. Running time: for 30K individuals, 9M SNPs, require 5K CPU hours.

- Lessons: computational efficiency is gained by MCMC, using EM (marginalizing indicators for SNPs); taking advantage of block structure.

Genetic fine mapping incorporating functional annotation: a Random Effects approach [Fisher and Liu, review for AJHG, 2018]

- Motivation: use LDSC to estimate enrichment of heritability across annotations, then use that prior to obtain the posterior of effect sizes. Then test if effect size is 0 using Wald statistic.

- Model: use LDSC prior, $\text{Var}(\beta_j) = k\sum_c a_{jc}\tau_c$ where τ_c is the effect of annotation c and a_{jc} the c-annotation of SNP j. k is scale parameter: the enrichment of heritability in a region, comparing with the genome-wide average. The method will first estimate τ_c using LDSC, then compute the posterior mean of β_j. Treating it as a statistic, and obtain the variance of the estimator, and do the Wald test.

- Remark: ignoring the k term, and make some simplifying assumption (each subject has the same residual variance), the results reduce to RSS-p.

- Simulation: only one locus, Hapgen2 to simulate large number of genotypes. The annotation parameters we chosen from GIANT study of BMI. Choose one causal SNP in the region: at different LD (high or low), or different prior effect size $\sum_c a_{jc}\tau_c$. Then simulate the phenotype.

- Simulation results: better than GWAS, Lasso, PAINTOR, GenoWAP (Hongyu's method). GenoWAP very poorly (often much worse than just plain GWAS). PAINTOR seems to overemphasize annotations (if the causal SNP has high annotation scores, PAINTOR works well).

- Remark: a problem with polygenic model is that the prior is non-zero for every SNP, and the posterior is always non-zero. The solution here is to treat posterior estimate as test statistic, and do hypothesis testing of whether it equals to 0.

Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps [Mahajan and McCarthy, NG, 2018]

- Data: large T2D with 74K cases. 243 loci, and 403 signals.

- Use conditional regression from GCTA to infer multiple independent signals in a locus: use only summary statistics and LD. The LD is estimated from 6,000 unrelated individuals in UK Biobank of white British origin. Ex. TCF7L2 has 8 signals.

- Unweighted fine-mapping method: on 380 distinct signals (excluding 23 signals), considering 500kb nearby region on either side of each signal. For each SNP, obtain its effect size and standard error. In regions with multiple signals, the effect size and standard errors were obtained from conditional analysis. Then use Wakefield's formula to obtain BF for each SNP, Λ_j, with prior effect variance 0.04. The PPA of variant j is then given by:

$$\pi_j = \frac{\Lambda_j}{\sum_k \Lambda_k} \tag{4.230}$$

  The 99% credible set was then constructed by: ordering all variants by decreasing PPA, and include variants until the cumulative PPA reaches 99%.

- Fine-mapping: number of signals per locus (Figure 3a). 99% credible set of all loci (Figure 3b). Median of 42 variants per credible set. PPA distribution of variants in credible set (Figure 3c). At 51 signal, one variant has $> 80\%$ PIP.

- Impact of reference panel on fine-mapping: compare HRC vs. 1000GP, the results are similar.

- Enrichment of enhancers: fgwas, log-OR 1-2 for active enhancers in islet and adipose; Or 2-8 fold in islet promoters, enhancers and coding sequences.

- Fine-mapping with fgwas: 15 chr. annotations, and use backward variable selection. Fine-mapping (modified fgwas): (1) regions with single signal: use 1Mb region around lead SNPs. (2) Regions with multiple signals: analysis of each of the distinct signals.

- Results of fine-mapping with functional annotations (Figure 6): median credible set size reduced from 42 to 32. Number of variants with PIP $> 80\%$ increases from 51 to 73.

- Examples: for fine-mapped SNP, use cis-eQTL to find the target gene and validate its effect.

- Lesson: to assess fine-mapping results, use all variants included in credible sets, and see how their PIPs change by priors, reference panels, etc.

Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci [Thurner and McCarthy, eLife, 2018]

- Data: ATAC-seq, 17 samples and DNA methylation 10 samples in islet cells.

- Enrichment of open chromatin regions in GWAS of T2D and fasting glucose (FG): Figure 3B-D. Single-feature enrichment and joint enrichment test. Include the islet epigenomic annotations and CDS, TSS, conserved.

- Fine-mapping of T2D loci: found significant regions, and 99% credible sets. Using epigenomic data helps with fine-mapping: reduction of credible sets and increase of PIP of the top SNP (Figure 4AB). Assess the contribution of annotations (e.g. islet enhancers) to PPA of each locus: e.g. for some locus, signal is primarily driven by islet enhancers, suggesting importance of insulin secretion, rather than insulin resistance (Figure 4C).

- Several examples: Allele imbalance in top SNPs (PIP > 0.1): 3/20 show signals. In three examples (Figure 5), likely causal variants are found: high PIP, eQTL evidence, motif changes, Hi-C interaction (in addition to allele imbalance).

- Lesson: (1) To demonstrate the value of epigenomic annotations: change of PIPs and credible sets. Assess the contribution to PIPs. (2) Case analysis: independent evidence of causal SNPs, such as eQTL, motif disruption, Hi-C, allele imbalance.

Functionally-informed fine-mapping and polygenic localization of complex trait heritability (PolyFUN) [Weissbrod and Price, BioRxiv, 2019]

- Model: assume the variance of prior effect size is the same across SNPs, then the probability of being causal is proportional to the variance of effect size. Let a_i be annotations and β_i be effect, we have: $\text{Var}(\beta_i|a_i) \propto P(\beta_i \neq 0|a_i)$. The problem is then to estimate the polygenic effect size of $\beta_i|a_i$. To do this: partition SNPs into non-overlapping bines, and estimate the SNP heriability by each bin b, then then specify $\text{Var}(\beta_i|a_i)$ for all SNPs in that bin.
- Details: (1) using regularized (L2) S-LDSC to estimate heritability for each bin. Training with even chromosomes. (2) Estimation of heritability of SNPs in odd chromosomes. Avoid Winner's curse. (3) Reestimation of h2 for target chromosomes: ensure robustness to model mis-specificiation.
- Simulations: real genotype from UKBB. 3Mb blocks, 10 cuasal loci explaining 0.05% h2. (1) Calibration of PIPs: false discoveries of SNPs above PIP threshold (e.g. 0.5). CAVIARBF significantly inflated. (2) Power: number of true causal SNPs above a PIP > 0.5. PolyFun + FINEMAP slightly more than PolyFun + SuSiE, and best (37% more than others). PolyFun + SuSiE much faster.
- Discussion: examples of coding and non-coding SNPs near the same genes (Table S22).
- Lesson: to evaluate fine-mapping methods, simulate blocks with causal variants, then assess calibration/FDR and power at given PIP cutoff. Note: not need to simulate null blocks.

4.8 Population Structure and Association Studies

Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data (PESCA) [Shi and Pasaniuc, 2020]

- Background: summary statistics model in terms of Z scores. Let Z be Z-score, R be LD matrix and γ be configuration, then we have:
$$Z|\gamma \sim N(0, n\sigma^2 R \cdot \text{diag}(\gamma) \cdot R + R) \tag{4.231}$$
where n is sample size, $\sigma^2 = h^2/|\gamma|$, $|\gamma|$ is the number of causal SNPs.
- Background: EM algorithm in Bayesian variable selection. If our goal is to estimate hyperparameters, then we can use EM, however, the Q function would need to sum over all configurations. This is done by MCMC (bf-gwas) or DAP-1 (DAP).
- Background: multivariate Bernoulli distribution (MVB). Related to logistic regression, but in multivariate case.
- Model: let f be parameters of MVB distribution, and $C_i|f$ be the configuration of SNP i (2-dim vector). Given j population, we have $Z_j|C_j$ are conditionally independent. Let V_j be the LD matrix of population j. We have:
$$Z_j|C_j \sim N(0, V_j + \sigma_j^2 V_j \text{diag}(c_j) V_j) \tag{4.232}$$
This gives the marginal likelihood $P(Z_1, Z_2|f)$, summing over all C_j's.
- Inference: by EM algorithm: E-step, in calculating Q, averaging over all configurations (posterior of configurations) - simply do L times. M-step has a closed form solution. How?
- Simulation: 9000 SNPs in chr. 22. Genotype normalized. (1) Sample configurations c. (2) Sample effect size of causal variants: $\beta_i|c_i \sim N(0, c_i h^2/|C|)$, where $|C|$ is the number of causal SNPs. (3) Sample error terms: total variance of trait is 1, so error is sampled from $N(0, 1-h^2)$. (4) Compute Z-scores of all SNPs.
- Results: European and Asian, 9 traits. (1) for BMI, 10% SNPs and 2-3% for lipids - seems too high. (2) Most of variants $> 80\%$ have shared effects in both populations, and effect directions are highly correlated.

- Remark: not model correlation of effect sizes. In estimation of f: not clear if h^2 is estimated.
- Remark: standard error of f (Table 1): Seems to be too small.
- Remark: how to reconcile with lack of portability of PRS? Possible explanation: large effect SNPs are shared, as shown in the paper, but small effect ones are less shared, which contribute to the PRS difference.

4.8.1 Population Stratification

Reference: [New approaches to population stratification in genome-wide association studies, Price & Patterson, NRG, 2010]; [Laird & Lange, Chapter 8]

Overview of population stratification:

- Problem: the case and control group may have different proportions of different subpopulations, thus a SNP that has different allele frequencies in two subpopulations may exist at different frequencies in cases and controls, causing false association signal.
  - Ex. suppose African population are more likely to eat junk food and have less exercise because of lower income, thus more likely to develop obesity, then African population may be overrepresented in the case group.
  - In general, the individuals are more likely to be related in the case group (since they share the genetic disorder) than in the control group.
- Statistical characterization of the problem: suppose genotype x is independent variable and Y (disease) response variable. We have a confounding variable Z, which represents the race/ethnicity. If Z is correlated with x (SNPs are correlated with race), and correlates with an independent risk factor (diet, culture), then Z is a confounding variable that needs to be controlled. If not, there will be false association. Note that for a confouding variable to create false associations, we need both conditions.
- Detection: in general, if there is a population bias in the cases vs. controls, many SNPs will have different allele frequencies thus associated with the phenotype (thus low p values). So the in QQ plot, the observed distribution of the association statistic will depart from the theoretical null distribution. This could be measured using genomic control λ_{GC}, defined as the median χ^2 across SNPs divided by the median under the theoretical null distribution.
- Basic strategies of dealing with population bias:
  - Filtering putative related individuals.
  - Genomic control: the idea is to discount the statistic of a SNP. Intuitively, if a SNP has different AFs in the different subpopulations, and the subpoputions are different in cases/controls, then the statistic of the SNP should be discounts.
  - Ancestry matching: if we know the ancestral subpopulation of subjects, then we could test the association between SNP and phenotype, conditioned on the same subpopulation. This could be implemented via regression with subpopulation as a covariate, or other forms of ancestry matching. PCA is one form of ancestry matching.
- The challenge of crypic relatedness: population stratification can be caused by two sources: (1) Additional covariates (that correlate with SNPs) such as diet, culture. (2) Family genetic background: e.g. the cases contain a family, then all unique SNPs in this family may be enriched. PCA is a way of controling (1), however, crypic relatedness in (2) cannot be solved by PCA.

Correcting for population structure by LMM [personal notes]:

- The standard approach for correcting population structure and relatedness. Use a random effect to capture genetic background: the effect is correlated between samples, matching their genetic relatedness (similar to group structure in a typical LMM setting).
- Does LMM capture env. confounders? Not directly. But it is reasonable to believe that the genetic random effects have similar group/relatedness structure as env. confounders.

Sources of population stratification:

- Genetic drift: since the population divergence. This produces systematic shift of the observed distribution, and can be addressed by genomic control.
- Natural selection: produces markers with unusal allele frequency differences that lie outside the expected distribution. Genomic control is inadequate.
- Family structure and cryptic relatedness: this may be a more important explanation of spurious association [Devlin & Roeder, Biometrics, 1999].

Genomic control: because of the relatedness of individuals, particularly in the cases, the test statistic may be inflated [Devlin & Roeder, Biometrics, 1999]

- Idea: correct the null distribution of the test statistic by modeling its distribution under population structure and cryptic relatedness. Let T be the numerator of the Armitage tend test statistic (i.e. the difference of the numbers of A in cases and controls), we could derive the distribution (variance) of T under two scenarios: population structure and cryptic relatedness.
- Population strcture: suppose there are m populations and $a_1, \cdots, a_m$ in the cases and $b_1, \cdots, b_m$ in the controls. And let p_1, p_2 be the population frequency of A_1 and A_2 alleles and R the total number of cases. F is the inbreeding coefficient. We could derive the distribution of T by considering the distriution of the number of A_1 alleles in each case or control:

$$\text{Var}(T) = 4Rp_1p_2(1+F) + 4Fp_1p_2 \sum_k [a_k(a_k-1) + b_k(b_k-1) - 2a_kb_k] \tag{4.233}$$

- Cryptic relatedness: let F_1 and F_2 be the inbreeding coefficients in the cases and controls respectively, then we could derive:

$$\text{Var}(T) = 2Rp_1p_2\,[2 + (F_1+F_2)(2R-1)] \tag{4.234}$$

- Genomic control: defined as the ratio of $\text{Var}(T)$ over the theoretirc distribution under H_0. It can be estimated with a large number of SNPs. Let Y^2 be the Armitage trend test statistic, and Y_i^2 be the statistic of the i-th locus. Then $Y_i^2/\lambda \sim \chi_1^2$. So the estimate $\hat{\lambda}$ can be obtained by taking the median of Y_i.
- Correction: divide the χ^2 statistic by λ_{GC} (and then compare with the theoretical null distribution). $\lambda_{GC} \approx 1$, no stratification; $\lambda_{GC} > 1$, stratification. Generally, $\lambda_{GC} < 1.05$ is considered benign. Note that inflation in λ_{GC} is proportional to the sample size.
- Remark:
  - A common inflation factor is applied to all SNPs, however since the SNPs differ in their allele frequencies across ancestral populations, doing this will lose power. Ideally, we want the SNPs whose AF differ a lot are heavily discounts, while other SNPs are not penalized much.
  - Will address genetic drift, but not unusual markers. Also not maximize power when family structure or cryptic relatedness is present.

Genomic control: [LL, Section 8.2]

- Motivating example: suppose the sample consists of a mixture of two populations, however, the mixing fractions are different in cases and in controls (λ and λ' respectively). Givne a locus unrelated to disease, its allele frequencies in the two populations are p_A and q_A, then the total frequency in the cases is: $\lambda p_A + (1-\lambda)q_A$, and in the controls: $\lambda' p_A + (1-\lambda')q_A$. If $\lambda \neq \lambda'$, the two frequencies are different, then the expectation of U (in trend test) would not be equal to 0.

- GC correction: suppose X^2 is the trend test statistic of markers, and λ is the inflation factor (across all control markers) indicating the extent of population stratification. λ can be estimated from the test statistic of L control markers (not associated with the phenotype):

$$\hat{\lambda} = \text{median}(X_1^2, \cdots, X_L^2)/0.456 \tag{4.235}$$

  We could adjust test statistic by: $X^2/\hat{\lambda} \sim \chi_1^2$.

Association Mapping in Structured Populations: [Pritchard & Donnelly, AJHG, 2000]

- Test: infer the subpopulations in the sample and the association statistics are computed by stratifying the subpopulations. The idea: the effect of a SNP is assessed only within a stratum/subpopulation, and the effects of all strata are combined. Specifically, let P_0 be the allele frequencies in the control, and P_1 be those in the case:

$$P_0 = \langle p_{kj}^{(0)} \rangle, 1 \leq k \leq K \qquad : \text{the frequency of } j\text{-th allele in the } k\text{-th subpopulation at control} \tag{4.236}$$

  And similary for P_1. The hypothesis of $P_0 = P_1$ can be performed by LRT:

$$\Lambda = \frac{P(C|\hat{P}_1, \hat{P}_0, \hat{Q})}{P(C|\hat{P}_0, \hat{Q})} \tag{4.237}$$

  where C is the genotype, Q is the population origin of individuals: $q_k^{(i)}$ is the proportion of genome of the i-th individual from the k-th subpopulation. The distribution of Λ is obtained through simulation (under the MLE of the parameters).

- Remark: the STRAT approach (allowing fractional membership) is computationally intensive, and may not be applicable to genome-wide studies.

Principal components analysis corrects for stratification in genome-wide association studies (EIGENSTRAT): [Price & Reich, NG, 2006]

- Population structure analysis: clearly many SNPs are highly correlated (LD or shared ancestry), so a small number of latent variables for ancestral population (subpop.) may be needed. Suppose we have u and v as latent variables, the j-th SNP is:

$$x_j = \beta_j u + \gamma_j v + \epsilon_j \tag{4.238}$$

  So β_j is roughly the "average" genotype of the subpop. corresponding to u (allele frequency), and γ_j the average genotype of v. In this case, the eigenvectors represent the genotype/allele frequency (with appropriate standardization) of subpopulations.

  - Note: if there are k subpopulations, only $k-1$ latent variables will be needed (as the total fraction sums to 1).

- Testing association: the PCs of subjects can be used as covariates in the testing of genotype-phenotype correlation. Ex. suppose we are regression x_i (the SNP) on y_i (phenotype), we add one covariate, u_i, the PC:

$$x_i = \beta y_i + \gamma u_i + \epsilon_i \tag{4.239}$$

  Then we test if $\beta = 0$ or not. Equivalently, we could regression x_i on u_i and obtain the residual, and y_i on u_i and obtain the residual, and test the association between two residuals.

- Remarks:
  - Comparison with GC: the SNPs are discounted differently. If a SNP does not have different AF in a subpopulation (relative to the population average), then x_i is independent of u_i in the equation above, then no need of population stratification/discount.
  - Comparison with STRAT: continuous axis of genetic variation, i.e. u and v vary continuously.
  - These approaches are good for population-level confounding, but inadequate for family structure and cryptic relatedness. Even though these approaches may detect the bias in case vs. control, the power may be lost as any putative disease association will resemble a strong instance of the bias.

Testing for genetic associations in arbitrarily structured populations [Song & Storey, NG, 2015]

- Idea of Genotype Condtional Association Test (GCAT): the existing method (LMM) to correct for population structure does not correct for environmental variables that confound with structure. Correct for population structure by estimating the expected AF under the ancestry of any individual: then test if the AF changes with disease states.
- Logistic factor analysis (LFA): estimating the AFs under given ancestry. Consider m variants and n individuals, let x_{ij} be the genotype of SNP i and individual j, and π_{ij} be the corresponding AF (indvidual-specific AF). The idea to estimate π_{ij} is that it depends on the ancestry of the individual j, and how SNP i depends on the ancestry groups. Let $L_{ij} = \text{logit}(\pi_{ij})$, then we use a factor model:

$$L = AH \tag{4.240}$$

  where A is $m \times d$ matrix representing how SNPs are determined by the population structure, and H is $d \times n$ matrix is the projection of individuals on population structure. To estimate A and H, use the model: $x_{ij} \sim \text{Bin}(2, \pi_{ij})$.
- Association testing: test if x_{ij} correlates with y_j with logistic regression (inverse regression), adjusting population structure with $\text{logit}(\pi_i j)$:

$$\text{logit}\left[\frac{\text{E}\left(x_{ij}|y_j, z_j\right)}{2}\right] = a_i + b_i y_j + \text{logit}(\pi_i j) \tag{4.241}$$

  Test $H_0 : b_i = 0$ using LRT.
- Questions:
  - Validation of LFA model for estimating π_{ij}: most markers may not be ancestry informative, does it make sense to assume that its logit is a linear function of the latent factors representing the population structure?
  - Advantage of LFA over PCA?
  - How LFA controls for envrionmental variables that confound with genetic population structure?
- Remark:
  - Similar to UNICORN in modeling the dependency of AF on population structure. How to incorporate population control in the GCAT model?
  - Specical form of the inverse regression model, and relation to TADA2 model: the idea of individual-specific AF?

Correcting subtle stratification in summary association statistics, [Bhatia and Price, bioRxiv, 2016]

- Motivation: significant inflation in summary statistics (from LDSC) due to residual population stratification.
- PC loading regression: the phenotype y depends on both PC and SNP of interest:

$$y = \beta_{PC} PC + \beta x + \epsilon \tag{4.242}$$

  Meanwhile, PC can be written as: $PC = \gamma x + \Psi$ where γ is the loading of SNP. From here, we obtain the inflated (uncorrected) effect size is related to the true effect and extent of stratification:

$$\beta_{\text{STRAT}} = \beta_{PC}\gamma + \beta \tag{4.243}$$

  So if we regress the uncorrected effect size with PC loading γ, the slope would be β_{PC}, the extent of stratification and the corrected/true effect size β.
- Issue: using estimated PC loading from external reference γ leads to biased estimate and reduce power. In practice, this limits the number of PCs one can use in the regression.

4.8.2 Admixture Mapping

Principles of admixture mapping [personal notes, Guimin's talk]

- Principle: if a causal locus is associated with another locus in some way (not only genotype, could be any other properties of the locus), then the linked locus will be associated with disease.
- Admixed genome and local ancestry: genomes of admixed population are "mosaic genomes" of two populations, e.g. for African American, it would admixture of West African and European.
  - Global ancestry: for any individual we can define its global ancestry as, on average, the proprtion of genome that comes from population 1.
  - Local ancestry: for any individual, at any point in the genome, we can define the similar proportion. For any individual, clearly, local ancestry could vary greatly across locations. However, averaging over many people, local ancestry is generally close to the genomewide average (global ancestry).
- ALD and BLD: local ancestry of adjacent loci should be correlated (from the same ancestry block that has not been broken by recombination), and we can this ALD block. The genotypes of loci in ALD blocks are not necessarily correlated. BLD block: represent background correlation in genotypes.
- Concept of admixture mapping: it relies on two associations:
  - At causal locus, if AFs in the two ancestry populations are different, say $q_1 > q_2$, then in the cases, since the locus will be enriched, it means it is more likely from population 1, leading to elevated local ancestry at the causal locus.
  - ALD block: other loci in the ALD block will have correlated local ancestry, thus their local ancestry will also be associated with diseases.
- Example: suppose we have a locus that is present only in population 1 but not 0, $q_1 = 1, q_2 = 0$. In admixture population, the fraction of population 1 is 0.2, then the average AF would be 0.2. In the cases, it will be higher, let's say 0.3. Now because the allele is only present in population 1, it means that the local ancestry at this site is 0.3, higher than average (0.2).
- Formal analysis: let q_1, q_2 be the AF in two ancestral populations, let π be the fraction of population 1. Then in admixed population, the AF is:

$$q = \pi q_1 + (1 - \pi) q_2 \tag{4.244}$$

The same equation means that if given q, q_1, q_2, we can estimate local ancestry as:

$$\pi = \frac{q - q_2}{q_1 - q_2} \tag{4.245}$$

Now in the cases, the AF will be elevated, so its AF becomes γq. The local ancestry becomes:

$$f = \frac{\gamma q - q_2}{q_1 - q_2} \tag{4.246}$$

It is easy to check that if $q_1 > q_2$, $f > \pi$, and if $q_1 \approx q_2$, $f \approx \pi$. So admixture mapping depends on the difference of AF in ancestral populations.

- Local ancestry graph: typically, we estimate average local ancestry in the cases at each point of the genome. For most places, this average is equal to global average (when sample size is large enough). However, the disease ALD will show elevated or reduced average local ancestry. Effectively: local ancestry is similar to genotype, and average local ancestry is similar to AF.

Admixture Mapping Comes of Age [Winkler, ARGHG, 2010]

- Comparison of ALD with LD: usually much shorter, because the admixture is much more recent.
- Inferring local ancestry (Figure 6): suppose the AF (of allele 1) is higher in Ancestral population A and lower in B. Intuition: if we have a blocks of 1s, more likely to be from A. Block of 0s more likely from B. Implementation with HMM: emission, AF. Transition: recombination and population proportion.
- Admixture mapping: correlation of local ancestry with the phenotype. Similar to normal GWAS, except that we are correlation local ancestry instead of genotype. Rationale: cases, disease allele is enriched, which means that local ancestry is more likely population 1. So in the graph of local ancestry, the fraction of population 1 is elevated above global ancestry (Figure 3).

Combining admixture mapping and association test [Guimin Gao, Dec, 2015]

- Association test controlling ancestry: in association testing, one should control for global ancestry (population stratefication). Also, at local regions, control for local ancestry: otherwise any loci in ALD may be falsely associated with disease. Let D be the disease locus and A be a locus in ALD: then genotype of $A \leftrightarrow$ local ancestry of $A \leftrightarrow$ local ancestry of $D \leftrightarrow$ genotype of $D \leftrightarrow$ disease, where $\leftrightarrow$ means correlation. So we need to control for local ancestry in association testing (covariate).
- Admixture test: correlation of local ancestry with phenotype. Beause ALD is often much larger than LD, the admixture test can only identify large regions.
- Combining two tests: the idea is that using admixture test to find local regions, and then use association test to better define the SNP. Use Generalized Sequential Bonfonerri (GSB) procedure: let p_i be the p-value of i-th SNP in admixture test, then set $w_i = 1/p_i$, and do weighted multiple testing correction on the association results.
  - Variation: smoothed weights, $w' = (1 - \lambda)w + \lambda\alpha$ where α is a global parameter.
- Remark: w_i acts as a soft filter (very strong effect on non-significant SNP). Thus if admixture test has low power, it will lose information. Generally need to derive the optimal weights to maximize the power of study.

4.9 Sequencing Studies and Methods for Rare Variants

Sequencing studies in human genetics: design and interpretation [Goldstein, NRG, 2013]

- Applications of NGS: (1) Mendelian diseases (refractory to linkage); (2) undiagnosed childhood diseases; (3) common diseases.
- Functional prioritization of variants: "narrative potential". Ex. a control genome, all variants $< 1\%$ AF, 237 rare missense variants, 86% are considered damaging by one of four algorithms (PolyPhen, SIFT, GERP, Blosum62), and 32% can be connected to some phenotypes in OMIM or HGMD.
- Aggregate test of association: "the impact of incorporating differnet types of prior information into different types of tests has not been systematically evaluated".
- Functional evaluation: animal models and iPSC. Ex. iPSC-derived cardimocyte platform for assessing the effect of mutations.

Reference: [Bansal & Schork, Statistical analysis strategies for association studies involving rare variants, NRG, 2010], [Cooper & Shendure, Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data, NRG, 2011], [Kiezun & Sunyaev, Exome sequencing and the genetic basis of complex traits, NG, 2012]

Importance of rare variants: the evidence can be summarized:

- Population expansion: likely to have resulted in a large number of segregating, functionally relevant, rare variants that mediate phenotypic variation.
- Purifying selection [Kiezun12]: the number of observed variants is much higher than is predicted by the neutral model with constant population size. This is partly explained by population growth, but also by purifying selection. Thus rare variation is enriched for evolutionarily deleterious, and thus functional, variants. Among rare variants, missense variants predicted to be damaging are more prevalent than variants predicted to be benign.
- Cancer genetics: the discovery of rare independent somatic mutations within and across genes that contribute to tumorigenesis may parallel the functional effects of inherited variants that contribute to congenital disease.
- Mendelian diseases: the identification of multiple rare variants within the same gene that contribute to largely monogenic disorders such as cystic fibrosis and BRCA1- and BRCA2-associated breast cancer.
- Sequencing studies: that focus on specific genes have shown that collections of rare variants can indeed associate with particular phenotypes.

Scenarios where common and rare variants influence phenotype: [Figure 2, Bansal, NRG, 2010]:

- Variants at a single locus with common alleles are more frequent in cases then controls.
- Multiple rare variations contribute to the phenotype such that the collective frequency of these variations is greater in cases. This would create a greater diversity of haplotypes or DNA sequences among the cases. This is the extreme allelic heterogeneity (EAH) setting.
- Multiple rare variations contribute to the phenotype but act in a synergistic fashion, such that cases are likely to have more similar DNA sequences compared to controls.
- Multiple rare variations contribute to a phenotype but the variations contributing to the phenotype reside in specific genomic regions. This situation would create greater sequence diversity among the cases, but only in the relevant genomic regions.

Basic strategy of testing:

- Collapsing strategy: in its simplest form, counting the frequency of rare variants at any position in the genomic region of interest, in cases and controls, and compare the differences. Better strategies would weight the variants in some way (e.g. by allele frequency).
- Functional predictions: use protein structure information, sequence conservation and motif conservation to build models that generate a probability that a particular variant is functionally important. For example, nonsense variants should be prioritized above non-conserved missense variants. Similarly, missense variants should be prioritized above synonymous variants. A number of tests allow the inclusion of prediction scores in test statistics, including the VT test, KBAC, SKAT, the rare variant weighted aggregate statistic (RWAS) and the likelihood ratio test (LRT) [Kiezun12].
- Different methods differ in the way rare variants are weighted, or the assumptions about the effect sizes.
  - WSS test [Madsen09] assumes effect size proportional to $1/x(1-x)$, where x is the allele frequency.
  - The sequence kernel association test (SKAT) [Wu11] simulation framework uses effect size proportional to $-\log(x)$.
  - Variable threshold (VT) test [Price10] simulations use a demographic history model with a range of possible values of strength of selection leading to different relationships between effect size and x.
- Regression-based collapsed variant and conditional tests: If a set of rare variants each individually explain only a small fraction of the variation of the trait, they could be combined into a single predictor variable, e,g. a dummy variable. Could also include other factors in regression model, such as covariate effects, the effects of previously identified common variants or other collapsed sets of rare variants.

Methods based on summary statistics:

- CAST method [Morgenthaler07]: a version of the collapsing approach in which the frequency of individuals carrying any one of several rare variants is contrasted between case and control groups. Then use the standard contingency table-based chi-square or Fisher's exact tests for obtaining p-values. An extension of the CAST method is combined multivariate and collapsing (CMC) method [Li and Leal].
- Weighting by frequencies: Madsen and Browning proposed a statistic for testing a pre-specified collapsed set of variants that leverages weighting of each variant by its frequency, thus allowing one to include variants of any frequency into the collapsed set.
- Optimal or variable weighting [Price, AJHG, 2010]: in a procedure resembling that of Madsen and Browning. Price et al. showed that their method is more powerful than approaches that consider fixed weights. In addition, they argued that the use of the predicted functional impact of each individual non-synonymous coding variant could be leveraged in their model.
- Incorporating the direction of effects [Han and Pan]: for example, protective or deleterious, this can be implemented in a regression model framework.
- Haplotype analyses: comparing haplotype frequencies between, for example, case and control groups. Haplotype analyses require phase information, which is not trivial to obtain for genotyped rare variants or variants derived from sequence data.

Approaches based on similarities among individual sequences:

- Motivation: the general nucleotide background or context within which a rare variant can influence a phenotype may be important.

- Assessing the strategy: such strategies can be as powerful, if not more so, than some traditional tests of association in many settings involving common variations. However, the performance of these methods when many rare variants and no common variants are considered is unknown. In addition, a limitation of these methods is that a specific DNA similarity or distance measure or metric must be chosen and this can be problematic.

- Searches for optimal sets of variations: one could potentially search for a subset of variants that maximally discriminates between cases and controls. Such methods are problematic in that the determination of an optimal subset of variants based on group differences can be computationally intensive.

- Choosing a DNA sequence similarity measure: difficult because ultimately, functional nucleotide content determines gene activity, rather than the phylogenetic origins of those nucleotides. Thus, in theory, similarity measures that build off the functional features and functional capacities of the affected genes associated with DNA sequence are likely to be more appropriate for association studies.

- Family-based linkage analyses: consider the consistency of within-family sharing of specific transmitted chromosomal segments among affected family members rather than the consistency or similarity of the nucleotide content of these segments across different families. However, not all approaches to linkage analysis are very powerful, and this is especially true for non-parametric approaches involving small families.

Multiple regression and data-mining methods:

- Basic method [Morris and Zeggini]: the use of a simple tally of the number of rare variants possessed by an individual across a large region as a predictor of a phenotype against the use of a simple indicator of the possession of any rare variant.

- Problems with the simple regression approach: LD, multicolinearity, many potential predictor variables to choose from if many individual common and rare variants, as well as collapsed sets of variants, are considered.

- Newer regression methods: (1) regularization and shrinkage methods to control for colinearity and overfitting; (2) One possible solution to this problem is to devise methods that combine elements of many different regression procedures, such as the 'bridge' (GPS) regression procedure of Friedman; (3) "ensemble" methods that combine the results of different regression and prediction methods.

- Logic regression: may be a particularly attractive regression-based approach, at least in theory, for the analysis of rare variants. Use additional variables, constructed from logical operators such as 'AND' and 'OR' that connect and combine sets of variants into potential predictors of the phenotype. The issues include computational burden; difficulty in obtaining p-values for each potential independent variable (or individual rare variant compared to a collapsed group of rare variants); and the identification of the optimal, and hence the biologically most plausible, set of genetic predictors.

Related issues:

- Multiple hypothesis testing [Kiezun12]: could use permutation to obtain the threshold. For larger sample sizes, the permutation threshold would be closer to the Bonferroni threshold, asymptotically approaching it as the sample sizes increase.

- Power of methods [Kiezun12]: most existing studies (up to early 2012) are underpowered. May require up to 10,000 samples to obtain satisfactory power.

- Evaluation and simulation [Bansal11]: need to be assessed in a wide variety of contexts, not just the EAH setting. The best approach will be to take real sequence data obtained from many individuals (e.g. 1000 Genomes Project) and simulate phenotypes based on variants in those sequences, making assumptions only about phenotypic effect sizes and interactions between variants.

Directions:

- Methods that can accommodate covariates, previously identified genetic factors, allelic heterogeneity and different sets of collapsed variants simultaneously are clearly advantageous.
- Methods that can account for subtle synergistic effects of many loci within a defined region and/or different forms of variation that might contribute to gene function, such as those rooted in sequence or functional similarity, are also likely to be appropriate.
- Identifying causality of rare variants in a set: may be more pronounced than it is in assigning causality to a single common variant.

Evaluating the functional impact of rare variants [Cooper11]:

- Goal: two related but separable questions: whether a given variant has a functional effect at the molecular level and, if so, whether that functional alteration is deleterious to the organism.
- Evolution as the best measure of deleteriousness:
  - Two considerations are essential. First, sequence conservation is not a predictor of deleteriousness per se, but rather it is conservation in excess of neutral expectations. Second, the 'phylogenetic scope' of the compared sequences has substantial effects.
  - The assumption of purifying selection: functional divergence will lessen the correlation between past constraint and present-day deleteriousness.
- Predicting the effects of protein-coding sequence changes:
  - Nonsense and frameshift mutations are the most obvious candidates
  - Considering non-synonymous variants, the simplest and earliest approaches to estimate deleteriousness use discrete biochemical categorizations such as 'radical' versus 'conservative' amino acid changes
- The case for non-coding variation analysis.
  - However, non-coding variants constitute the overwhelming majority of human genetic variation, and most weak-effect causal variants are non-coding.
  - Additionally, evolutionary analyses demonstrate that approximately fivefold more non-coding positions exist than coding positions in human genomes that have been subject to purifying selection
  - As for protein-altering variants, comparative genomics is a central component in deleteriousness prediction for non-coding variants.
- Experimental approaches:
  - Projects such as the Encyclopedia of DNA Elements (ENCODE) are applying diverse assays in many cell types and conditions to generate functional annotations at a genome-wide scale, including protein-coding genes, non-coding RNAs and cis-regulatory elements
  - New strategies whereby variants in regulatory sequences, RNAs and proteins can be studied in a highly multiplexed fashion.
  - Detailed but generically assayed molecular phenotypes may be useful to capture and measure protein function. For example, cells may be perturbed by overexpression or knockdown of specific genes and subsequently subjected to high-throughput assessments, such as RNA-seq.
- Important directions:

- A unified, quantitative and predictive framework to estimate the prior probabilities for any given mutation to be both functionally relevant and disease relevant.
- Variant interactions: protein-protein interaction, gene co-expression networks, coupled with both literature and automated annotation of pathways and gene functions, are crucial to tackle this challenge.

Summary of rare variant association tests [personal notes]

- Multivariate test: the test statistic has M degree of free, where M is the number of variants. When M is large, the test loses power.
- Combining summary statistics of multiple variants: best when signal is sparse. When there exists joint effect of multiple variants, the test loses power.
- Random effect model (variance component): this increases power (over the multivariate test) by borrowing information across variants, effectively reducing DF (especially when variants are highly correlated?). However the test loses power when the signal is sparse, as it has to pay for the many non-informative variants (because the prior distribution assumes that many variants have effects).

Exome sequencing and the genetic basis of complex traits [Kiezun & Sunyaev, NRG, 2012]

- The relative power to detect association depends on factors such as the number and proportion of causal variants, their population frequencies and their effect sizes, as well as the directionality of effects, the number of genes contributing to the trait, etc.
- Effect size assumptions:
  - WSS test assumes effect size proportional to $1/q(1-q)$
  - VT: a demographic history model with a range of possible values of strength of selection leading to different relationships between effect size and q.
  - Sequence kernel association test (SKAT): effect size proportional to Beta function of q.
- A number of tests allow the inclusion of prediction scores in test statistics, including the VT test, KBAC, SKAT, the rare variant weighted aggregate statistic (RWAS), and the likelihood ratio test (LRT).
- To date, no published candidate gene study reported P values that would be significant in the context of the complete exome. This is particularly notable, because some candidate gene studies used much larger sample sizes (thousands of individuals) than ongoing exome sequencing studies (hundreds of individuals).
- Extrapolation of effect sizes and frequencies from published studies shows that thousands of individuals are required to reach acceptable statistical power.

In search of low-frequency and rare variants affecting complex traits [Panoutsopoulou & Zeggini, HMG, 2013]

- Population isolates: rare variants may have drifted up in frequency (random drift) and linkage disequilibrium (LD) tends to be extended. Example: Iceland-based deCODE study. Association of a rare functional variant (R19X) in the APOC3 gene with HDL-C and triglycerides levels was first detected in the Amish founder population.
- Rare variant reference panels: 1000GP, ESP, UK10K (high-depth WES of 6000 and low-depth WGS of 4000 well-phenotyped individuals).
- Variant weighting: imputation quality. MAF. Function prediction.

- Meta-analysis: based on study-specific summary statistics rather than individual-level data.
  - SKAT meta-analysis: They combine single-variant score statistics first across studies and then within a region. They also require between-variant covariance-type relationship statistics (such as LD structure) for each region, as well as MAF of variants.
  - Meta-analysis at gene level.
- Rare variants show increased population specificity. Existing methods to correct for population stratification at common variants such as principal component analysis and genomic control have not been shown to effectively control stratification at rare variants.

Rare-Variant Association Analysis: Study Designs and Statistical Tests [Lee & Lin, AJHG, 2014]

- Notation: GLM with link function $h(\mu)$

$$\mathrm{E}\left(h(\mu_i)\right) = \alpha_0 + \alpha^T X_i + \beta^T G_i \tag{4.247}$$

  where X_i is covariate and G_i genotype. The score statistic of the marginal model for variant j is:

$$S_j = \sum_{i=1}^{n} G_{ij}(y_i - \hat{\mu}_i) \tag{4.248}$$

  where $\hat{\mu}_i$ is the estimated mean of y_i under H_0.

- Burden test: it can be shown that the score statistic under $H_0 : \beta = 0$ is:

$$Q_{\text{burden}} = \left(\sum_j w_j S_j \right)^2 \tag{4.249}$$

  Adaptive burden test: use MLE of β as the weight of a variant.

- Variance component test:

$$Q_{\text{SKAT}} = \sum_j w_j^2 S_j^2 \tag{4.250}$$

- Omnibus test: for $0 \le \rho \le 1$,

$$Q_\rho = (1-\rho) Q_{\text{SKAT}} + \rho Q_{\text{burden}} \tag{4.251}$$

  where ρ can be interpreted as pair-wise correlation among the genetic effect coeffients β_j. Adaptive test that uses an optimal value of ρ that gives that minimum p-value.

- Comparison of test: gene based test can lose power when a very few of the variants in a gene are associated, when many variants have no effect, and when causal variants have low frequency.

Rare variant association studies: considerations, challenges and opportunities [Auer & Lettre, Genome Medicine, 2015]

- Advantages of population isolates:
  - Population bottleneck: rare variants may reach higher frequencies because of founder effect (larger genetic drift, loss of genetic diversity).
  - Environmental and culture homogeneity.
- Family studies:

  - General idea: co-segregation analysis of variants and phenotypes. The challenge is that one will have a large number of pathogenic variants. Ad hoc filtering; also the MAF in the population.
  - Comparison of TDT vs. case-control: similar power at low frequency.

- Problems with stratification: RVs may have different patterns, and often stronger stratification.
- Imputation: with large HRC, 30,000 projects, one can impute variants with MAF as low as 0.01%.

The increasing importance of gene-based analyses [Circulli, PLG, 2016]

- Why do we need the gene based test? We cannot focus only on case-only variants, and then say that they are extremely rare in controls.
- Family studies: co-segregation analysis, requires LOD 3.3 or higher. Multiple families with different co-segregation variants in the same linked gene can provide strong support (Ref 3).
- Coverage imbalance between cases and controls can create false signals.

Genetic architecture: the shape of the genetic contribution to human traits and disease [Timpson and Richards, NRG, 2018]

- Genetic architecture differs between phenotypes (Figure 1): T1D vs. T2D, while T2D has mostly low effect SNPs, T1D has some SNPs with large effects. Also comparison of Vitamin D vs. LDL: Vitamin D is mostly oligogenic while LDL is highly polygenic.
- Limitations of region-based rare variant testing: difficulty with replication; methods should be tailored by genetic architecture, which is unknown and varies; most WGS variants have no effects, and multiple testing burden; the direction of effects unknown.
- Review of rare variant studies: (1) UK10K: WGS, across 60 traits, burden and variance-component tests on regions. Not a single new region not already found by single SNV analysis. Note: low coverage. (2) T2D: WES in 12K samples. No regions found. (3) Height: 83 rare variants from single SNV analysis vs. 3 regions. (4) MI: no new regions beyond single SNV test.
- Summary: "We anticipate that, of the methods currently available, this method (single SNV test) will enhancer our knowledge of genetic architecture the most".
- Utility of small effect SNVs: could still lead to drug targets, e.g. PCSK9. Just need large perturbation of proteins.
- Gene-environment interactions: BMI study, environmental variables explain 14% variation. But only smoking shows interaction with genotype, and other covariate genetic interaction effects account for less than 1% of total phenotypic variance.
- Remark: all the limitations of region-based testing can be addressed by a Bayesian framework, which learns trait-level parameters and incorporates priors on the variant effects.

4.9.1 Rare Variant Association Tests

Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data [Li & Leal, AJHG, 2008]

- Motivation: In the presence of allelic heterogeneity, although the power of linkage analysis is not influenced, association studies based on LD mapping will inevitably be low-powered. Low frequencies of functional variants result in low r^2 values.
- Genetic model: used for simulation:

- Independence of rare variants: Usually, rare mutations occur on different haplotypes within a locus, therefore, correlation between variants is low.
- Penetrance of the locus of the wild type: denoted by f_0, is the probability of an individual being affected if the genotypes across all variant sites are wild-type aa. If the assumption is made that wild-type genotypes at different sites have the same penetrance, the relationship can be simplified to $f_0 = Mf_{0i}$, where M is the number of rare variants.
- Risk model: Rare variants also high-risk, and they independently affect phenotype. At each variant, multiplicative, dominant or recessive model.

- Methods:
  - Single marker test: standard chi-square test. Remark: correction for many hypothesis.
  - Multiple-Marker Test: e.g., the Fisher product method, Hotelling's T^2 test, or multiple logistic regression. Hotelling's T^2 test is used here, with the null hypothesis that none of the variants is associated with disease susceptibility. Remark: a large degree of freedom.
  - Collapsing method: Due to the rarity of variants, the probability of carrying more than one variant for an individual is low, and the method collapses genotypes across all variants, such that an individual is coded as 1 if a rare allele is present at any of the variant sites and as 0 otherwise. Then the data consist of a set of 0's and 1's in the cases and in the controls, and the test is whether 1's are enriched in cases than in control. A standard χ^2 test can be applied.
  - CMC method: combines collapsing and multivariate tests. Markers are divided into subgroups on the basis of predefined criteria (e.g., allele frequencies), and within each group, marker data are collapsed (individual coded as 1 or 0 depending on whether he has a rare variant). A multivariate test (e.g., Hotelling's T^2 test) is then applied.

- Results: comparison of methods
  - Single marker test: lowest power when there is no misclassification of variants. Not only does this test pay a penalty for multiple testing, but also affected by the low allele frequency at each variant, where the power for each individual test is low.
  - The power for Hotelling's T^2 test is superior to that for the single-marker test but is less powerful than that for the collapsing method. The improvement of power for the collapsing method is due to an enrichment of signals across variants and the single univariate test performed.
  - However, collapsing methods are not always robust to misclassification of nonfunctional variants, and power loss can be substantial. multivariate tests are more robust in the presence of misclassification of nonfunctional variants
  - In the presence of common variants, it can be advantageous to analyze both common and rare variants simultaneously with the CMC method; including rare variants in the analysis can greatly increase power if the rare variants have high genotypic RRs and are either numerous or not extremely rare.

- Other considerations: choosing variants, collapsing, etc.
  - Before statistical analysis of sequence data can be carried out, the first step is quantifying which variants are potentially functional or neutral, e.g. from bioinformatic analysis.
  - Choosing functional variants: a number of studies have demonstrated that alleles with a wide range of frequencies are involved in disease etiology. If high-frequency functional variants are removed from or high-frequency nonfunctional variants are included from the analysis, the effect on power can be very detrimental.

- Criteria for collapsing: the CMC method can be used to analyze data on the basis of allele frequencies or certainty of functionality. Even when classification is made on later, it is still inadvisable to collapse rare and high-frequency variants because, as previously discussed, if functionality classification is incorrect, then a large penalty in power can be incurred.
- Direction of effect: When all of the functional variants confer high risk or are protective, collapsing will enrich the signal. However, the signal will be weakened if some variants are protective whereas others increase disease risk. Although this situation is probably uncommon, when prior information is available on high-risk and protective variants it should be taken into account when deciding how to collapse variants.
- For rare variants, it is reasonable to assume that within a locus they reside on different haplotypes. The application and the validity of the single-marker test, Hotelling's T^2 test, the collapsing method, and the CMC method are not altered by the presence of LD.

Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies [Hoggart & Balding, PLG, 2008]:

- Motivation:
  - Multi-marker analysis: can improve over single-SNP tests, since a weak effect may be more apparent when other causal effects are already accounted for, but also because a false signal may be weakened by inclusion in the model of a stronger signal from a true causal association.
  - Computation: Bayesian stochastic search methods have been used to tackle variable selection problems, typically using the "slab and spike" prior formulation. However, normal MCMC appraoch is too slow for GWAS-scale problems.
- Method:
  - Shrinkage through Bayesian prior: continuous prior distributions with a sharp mode at zero, often referred to as "shrinkage" priors, to the regression coefficients. Consider two prior distributions, the Laplace, or double exponential distribution (DE) and a generalisation of it, the normal exponential gamma distribution (NEG), which has a sharper peak at zero and heavier tails.
  - Inference: seek only the posterior mode(s) rather than the full posterior distribution of the regression coefficients.
- Results:
  - Application to T2D GWAS: captured the same significant loci as the single-SNP analysis but at the cost of many fewer false positives.
- Discussion:
  - The main benefits: our analysis returns only the best SNP characterising the effect of a single detectable causal variant, thus suitable for fine-mapping. The NEG analysis improves on the single-SNP ATT (Armitage trend test) analysis, most notably in terms of false positives, and also in terms of power.

Lasso logistic regression in GWAS [Wu & Lange, Bioinfo, 2009]

- Application of Lasso in GWAS with large number of SNPs: (1) Set the penalty parameter λ s.t. effectively only a given number of predictors will appear in the regression; (2) after learning (the small number of) predictors using Lasso, reestimate the parameters; (3) assessing significance of each predictor: LRT with this predictor has zero effect or not, and compute P value based on χ^2 distribution (statistically incorrect).

- Interaction effect: only the predictors predicted from the previous steps (ie. those with large marginal effects) will be used in the test.
- FDR: B-H procedure to correct for multiple testing.
- Results:
  - Largely similar with the univariate test: similar P values for most SNPs. The main difference: multiple SNPs associated with the trait in LD region. Under Lasso procedure, none of these SNPs have high significance (as removing any of these can be compensating by the remaining SNPs).
  - Interaction effect: most of them are insignificant.
- Remark: Lasso is plausible in GWAS data, but few benefits are demonstrated.

A groupwise association test for rare mutations using a weighted sum statistic [Madsen & Browning, PLoS genetics, 2009]:

- Idea: extend the naive collapsing method. Accentuate mutations that are rare in the unaffected individuals, so that the test is not completely dominated by common mutations.
- Model: let the weight be the j-variant be w_j, it is given by:

$$w_j = \sqrt{q_j(1-q_j)} \tag{4.252}$$

  where q_j is the MAF of the j-th variant (Bayesian posterior mean) in the controls. Then the genetic score (load) of the i-th individual is:

$$\gamma_i = \sum_j \frac{x_{ij}}{w_j} \tag{4.253}$$

  where x_{ij} is the genotype of the j-th variant of the i-th individual. One could then compare γ_i in cases vs. controls, specificially, using the total rank of cases among all subjects as the test statistic (same as Wilcoxin test). Obtain p-value by permutation.
- Model explanation:
  - Weighting: the weight of the j-variant is in fact the standard deviation of x_{ij}, which follows $\text{Ber}(q_j)$. Thus the genetic score is the normalized count of all rare variants.
  - Permutation test: q_j is calculated from the controls only (if using both cases and controls, the estimate of q_j will be deflated/higher than the real values for risk variants). However, this may result in inflation of p-values, so need to use permutation to obtain the null distribution.

An evaluation of statistical approaches to rare variant analysis in genetic association studies [Morris & Zeggini, GE, 2010]:

- Methods:
  - Rare variant test 1 (RVT1): the predictor is the proportion of rare variants at which an individual carries a minor allele.
  - Rare variant test 2 (RVT2): the predictor is the presence/absence of a minor allele at any rare variant within an individual.
- Results: Our simulations clearly indicate that tests based on the accumulation of minor alleles at rare variants are always more powerful than conventional tests applied to SNPs present on GWA chips, particularly in the presence of substantial allelic heterogeneity.

Pooled association tests for rare variants in exon-resequencing studies [Price & Sunyaev, AJHG, 2010]:

- Motivation:
  - Variable allele-frequency threshold: [Li & Leal, AJHG08] pick a fixed allele-frequency threshold and perform an association test on the set of variants below that threshold, giving them each equal weight. The potential value of a statistical approach that uses a variable allele-frequency threshold.
  - Computational predictions of the functional effect of amino acid changes. The test gives higher weight to allelic variants predicted to be functionally significant.
- Log likelihood ratio and the weighted genetic score: we first note that the test statistic of the weighted collapsing test can be defined as the total burden in cases:

$$T = \sum_j C_j w_j \tag{4.254}$$

where C_j is the count of the j-variant in cases, and w_j its weight. To see its relation to LRT, let R_j be the OR of the i-th SNP. p_j is the allele frequence in controls, and q_j in cases; The two are related by R_j:

$$R_j = \frac{\frac{q_j}{1-q_j}}{\frac{p_j}{1-p_j}} \tag{4.255}$$

Assume independence of SNPs, and a generative model of genotypes, we could compute the LLR of causal (with specified ORs) vs null model.

$$L = \prod_{j=1}^{D} \frac{q_j^{C_j}(1-q_j)^{(N_+-C_j)}}{p_j^{C_j}(1-p_j)^{(N_+-C_j)}} = \left(\frac{1-q_j}{1-p_j}\right)^{N_+} \prod_{j=1}^{D} R_j^{C_j} \tag{4.256}$$

where N_+ is the sample size of cases. Take the log., we have:

$$l = \sum_j C_j \log R_j + N_+ \sum_j \log \frac{1-q_j}{1-p_j} \tag{4.257}$$

Thus the weight of the j-th variant is $\log R_j$ (log-OR).

- Methods:
  - Fixed threshold approach: $\sum_i C_i \xi_i$, where ξ_i is an indicator variable that is equal to 1 if the frequency of SNP i is below a specified threshold (1% or 5%) and is equal to 0 otherwise.
  - Weighted Approach: ξ_i is the inverse square root of expected variance based on allele frequencies p_i computed from controls only.
  - Variable-Threshold Approach: The intuition is that there exists some (unknown) threshold T for which variants with (MAF) below T are substantially more likely to be functional. Thus, we compute a z-score $z(T)$ for each threshold T, define $z_{\max}$ as the maximum z-score across values of T, and assess statistical significance of $z_{\max}$ by permutations on phenotypes,
  - Incorporating functional effects: use the PolyPhen-2 probabilistic score S, and convert it to posterior probabilities $p(S)$ of being functional for each SNP. The $p(S)$ is estimated from two distributions: the neutral set (substitutions that were fixed in the human lineage after divergence from chimpanzee) and the damaging set (known disease-causing missense mutations). These recalibrated posterior probabilities $p(S)$ were applied as weights in the regression.

Association screening of common and rare genetic variants by penalized regression [Zhou & Lange, Bioinformatics, 2010]:

- Motivation:
  - The problem with Lasso: the lasso is too stringent for rare variants. Shifting some of the lasso action to a group Euclidean penalty makes it easier for weak or low-frequency predictors to enter a model.
  - When we pass to penalized estimation, model selection is emphasized over hypothesis testing. The multiple hypothesis testing problem reappears in replication, but in a more benign form because the number of genes and SNPs of interest drop dramatically.
  - Here, we discuss how to incorporate group penalties that make it easier for related predictors to enter a model once one of the predictors does.
- Lasso regression background:
  - To put the regression coefficients on an equal penalization footing, all predictors should be centered around 0 and scaled to have approximate variance 1.
  - Logistic regression is handled in a similar manner. Instead of equating the loss function to a sum of squares, we equate it to the negative loglikelihood.
  - Group Lasso: if a parameter enters a model, then it does not strongly inhibit or encourage other associated parameters entering the model. In other words, the local penalty around 0 for each member of a group relaxes as soon as the regression coefficient for one member moves off 0.
  - Problem with group Lasso: Euclidean group penalties run the risk of selecting response-neutral predictors. As soon as one predictor from a group enters a model, it opens the door for other predictors from the group to enter the model.
- Methods:
  - If SNP j belongs to group G, it should experience penalty $\lambda_E||\beta_G||_2 + \lambda_L|\beta_L|$. If it belongs to no group, it should experience penalty $\lambda|\beta_j|$, where $\lambda = \lambda_E + \lambda_L$. The objective function is given by:

$$f(\theta) = L(\theta) - \lambda_L||\beta||_1 + \lambda_E||\beta_G||_2 \tag{4.258}$$

  - As we demonstrate, both kinds of penalties (Lasso and Eucledian) are compatible with coordinate descent, which is by far the fastest optimization method in sparse regression.
- Results:
  - Breast cancer data: candidate gene study, with genes in DNA repair (DSBR) pathway. 399 Caucasian: There were 196 affected and 203 unaffected individuals. 148 SNPs from the DSBR pathway were grouped by gene (17 genes). Although most of the SNPs in this dataset are common, 4 have MAFs $< 1\%$, 5 have MAF between 1% and 5% and 13 have MAF between 5% and 10%.
  - In the case of the pure lasso, SNPs enter the model singly, and in the case of the pure group penalty, genes enter the model with their full sets of SNPs. In the mixed cases, we see that either single SNPs or sets of SNPs grouped by gene enter the model. When a group enters in the mixed cases, it need not contain all of the SNPs in that gene.
- Remark:
  - The main benefit is: choose causal genes (accompolished by group level penalty) and the functional variants within group (by individual Lasso penalty).
  - Why group Lasso selects genes? Intuition: a group is not selected if none of the features is associated, this is similar to Hotelling's T^2 test: $\beta_1 = ... = \beta_k = 0$ within the group.

A data-adaptive sum test for disease association with multiple common or rare variants. [Han & Pan, Hum. Hered, 2010]:

- Background:
  - Global test: jointly testing on the multiple β_j parameters with the null hypothesis $H_0 : \beta_1 = ... = \beta_k = 0$ by one of the three asymptotically equivalent tests: the likelihood ratio test, the Wald test and the score test. Under H_0, any of the three test statistics has an asymptotic chi-square distribution with DF $= k$. The generalized Hotelling's T^2 test is closely related to the score test. A potential problem with the above tests is the power loss due to large DF.
  - In contrast to a global test, another extreme is to conduct a single-locus test for each SNP. The method is equivalent to choosing the univariate test for $\beta_{M,j}$ with the minimum p value, and is hence also called UminP method.
  - The Sum Test: while using all the SNPs, it adopts a key and generally incorrect working assumption that all SNPs are associated with the disease with a common OR. This is equivalent to regressing Y on a new covariate that is the sum of the genotypes of the multiple SNPs.
- Methods:
  - If the signs of $\beta_{M,j}$ (MLE of β_j) are quite different, it may result in power loss. Hence, before applying the Sum test, based on some ad hoc heuristic, one needs to choose the codings of the SNPs to maximize the number of their positive pairwise correlations.
  - A Data-Adaptive Sum Test: Hence, a natural approach is to choose the coding of each SNP j based on the sign of β_j (MLE), which is data-adaptive and may lead to inflated type I error rates if no adjustment is made with the null distribution.

Testing for an Unusual Distribution of Rare Variants [Neale & Daly, PLG, 2010]:

- Idea: under H_0, each variant is equally likely in cases or controls; under H_1, a fraction of variants may be more likely in cases than in controls (riks variants), or the opposite (protective variants). The difference lies in the overdisperson: more extreme (unbalanced) variants under H_1 than under H_0
- Test: contrasts the variance of each observed count with the expected variance, assuming the binomial distribution. The test statistic:

$$T = \sum_i T_i = \sum_i [(y_i - n_i p_0)^2 - n_i p_0 (1 - p_0)] \tag{4.259}$$

  where y_i is the count of the i-th variant in cases, and n_i is the total count of the i-th variant in both cases and controls, and p_0 is the null model (e.g. 1/2 if there are equal numbers of cases and controls).
- Remark: fail to take the effect size into account, e.g. a LoF variant with total count = 3, the imbalance is very small, thus contribute little to the test statistic (even if all are in cases).

Rare variant association testing for sequencing data using the sequence kernel association test (SKAT) [Wu & Lin, AJHG, 2011]:

- Model: GLM of the phenotype y_i as a function of genotype G_i and covariates X_i. Consider the quantitative trait:

$$y_i = X_i \alpha + G_i \beta + \epsilon_i \tag{4.260}$$

  The naive test of $H_0 : \beta_1 = \cdots = \beta_D$ suffers from high dof., thus low power. Assume the effect size follows the distribution, $\beta_j \sim N(0, w_j \tau)$, where w_j is the weight of the j-th variant (the model can be generalized to any distribution with mean 0, variance $w_j \tau$). Test the hypothesis $H_0 : \tau = 0$. This is the test of random effect under the variance component model.

- Test: we define the $n \times D$ matrix of the genotype, $G = (G_{ij})$, where G_{ij} is the genotype of the j-th variant of the i-th subject. Let $W = \text{diag}(w_1, \cdots, w_D)$ be the weights of the variants. The score test statistic is given by:

$$Q = (y - \mu)^T K (y - \mu) \tag{4.261}$$

  where μ is the predicted mean under H_0, and $K = GWG^T$. We can write K as:

$$K(G_i, G_{i'}) = \sum_j w_j G_{ij} G_{i'j} \tag{4.262}$$

  thus $K(\cdot, \cdot)$ is the kernel function, measuring similarity between the subjects i and i'. Under the null hypothesis, Q follows a mixture of chi-square distributions.

- Weighting: choose the form $\sqrt{w_j} = \text{Beta}(q_j; a_1, a_2)$ where q_j is the MAF of the j-th variant (cases and controls combined). When $a_1 = a_2 = 1$, this leads to $w_j = 1$ for any j. When $a_1 = a_2 = 0.5$, this corresponds to: $w_j = 1/(q_j(1 - q_j))$. The default/recommended choice is $a_1 = 1, a_2 = 25$, which would put some weights on the relatively common (MAF between 1% to 5%) variants.

- Model analysis: relation to the tests of individual variants. Let g_j be the vector of the genotype of the j-th variant. Q is a weighted sum of the individual score statistics for testing for individual variant effects:

$$Q = \sum_j Q_j = \sum_j w_j S_j^2 \tag{4.263}$$

  where $S_j = g_j^T (y - \mu)$ is the individual score statistic for testing the marginal effect of the j-th marker under the individual linear or logistic regression model.

- Kernel function: could incorporate additional prior information or epistatic effects: the weighted linear kernel (the basic model), the weighted quadratic kernel (epistatic), and the weighted identity by state (IBS) kernel.

A probabilistic disease-gene finder for personal genomes: VAAST [Yandell & Reese, GR, 2011]:

- Motivation: The utility of Amino Acid Substitution (AAS) approaches for variant prioritization is well established (Ng and Henikoff 2006); combining AAS approaches with aggregative scoring methods thus seems a logical next step.

- Methods:

  - Basic model: suppose there are k sites, each of which has one RV. The k sites are unlinked. Generative model of genotypes: the genotype at i-th site is simply a Bernouli trial. Under H_0: the probability is a constant (the MAF in the total samples); under H_A: two probabilities in cases and in controls respectively. The test is LLR.
  - Collapsing: to increase power, m collapsing categories: the intuition is when sampling at the i-th category, may sample an RV at any of the site within the category. Equation (1).
  - Incorporating AA substitution: under the generative model, add the probability of sampling a certain type of changes. Under H_0: at the i-th site, the probability that this change will not affect phenotype; under H_A: the probability that this change will affect the phenotype. Equation (2).
  - P-value: if unlinked, LLR follows chi-square distribution. If linked, do permutation test to get P-value.
  - Synonymous and noncoding variants: Normalized Mutational Proportion (NMP) in vertebrate (primate) genome alignments.

- Results:

- AAS frequencies among known disease-causing alleles in OMIM and AAS frequencies in healthy personal genomes differ from the BLOSUM model of amino acid substitution frequencies.
- VAAST scores a wider spectrum of variants than existing AAS methods. SIFT (Kumar et al. 2009), for example, examines nonsynonymous changes in the context of multiple alignments of homologous proteins. Because not every human gene is conserved and because conserved genes often contain unconserved coding regions, an appreciable fraction of nonsynonymous variants cannot be scored.
- Test on Miller's syndrom (a few genomes): the true gene is ranked in the top.
- Test on Crohn's disease: NOD2, with both rare ($< 5\%$) and common variants. Vary the number of individuals and assess the power by using 500 bootstrapped samples. Its estimated power is 89% for NOD2 using as few as 150 individuals ($\alpha = 0.05/21,000 = 2.4 \cdot 10^{-6}$, where 21,000 is the number of genes). By comparison, the power of GWAS is $< 4\%$ at the same sample size.
- Test on LPL, a gene implicated in hypertriglyceridemia (HTG): a data set of 438 re-sequenced subjects, rare variants only. Similar trend, the VAAST methods are better than GWAS and MB test [Madsen & Browning, 2009].

- Remark:
  - Very similar to CMC in overall apporach: collaping variants, and the tests are similar. In CMC: $H_0 : \beta_1 = ... = \beta_p$ and H_A use the MLE of β; in VAAST, $H_0 : p_j^A = p_j^U$, and H_A uses MLE of p_j. Since β_j and frequcy ratio of p_j are closely related, the two methods are very similar.
  - Contributions: use of AA subtitution data; synonymous and non-coding variants and other features such as use of pedigree.

A New Testing Strategy to Identify Rare Variants with Either Risk or Protective Effect on Disease [Ionita-Laza & Lange, PLG, 2011]

- Idea: test the allele frequency difference in cases vs. controls for each variant, then combine the results using Fisher's method.
- Model: let (k, k') be the number of copies of minor alleles in controls and in cases for a variant. The data can be then summarized as a table of $n_{kk'}$, the number of variants falling into the (k, k') cell. To test each variant, we use a (variation) of p-value: the probability that the variant occurs less than or equal to k times in controls, and more than or equal to k' times in cases, under H_0:

$$p(k, k') = \text{ppois}(k, \hat{f}) \times (1 - \text{ppois}(k', \hat{f})) \tag{4.264}$$

  where $\hat{f}$ is the expected frequency under H_0. The information of all variants are combined:

$$S = \sum_{k,k'} -n_{kk'} \log[p(k, k')] \tag{4.265}$$

- Extensions: to incorporate protective variants, using the two-sided test to derive p-values. To incorporate externtal information, let $\phi(v)$ be the probability that v is a risk variant, multiply $\phi(v)$ in the equation of the test statistic above.

A general framework for detecting disease associations with rare variants in sequencing studies [Lin & Tang, AJHG, 2011]:

- Contribution: relative to [Li08], [Madsen & Browning, PLG, 2009], [Price10], a score test that enhancers power, normal approximation (thus no need of permutation) and accomodate covariates.
- Basic model:

- Notation: n subjects with m SNPs. Y_i: phenotype of i-th subject; X_{ji}: j-th SNP of i-th subject; Z_{ji}: j-th covariates of i-th subject. Y_i are related to X_i and Z_i through logistic regression. We can write $\beta_j = \tau\xi_j$, where τ is a scalar constant, and ξ_j is called the weight function (assume given in the testing).
- Test: the score statistic for testing the null hypothesis H_0: $\tau = 0$. Under H_0, the test statistic $T = U/V^{1/2}$ is asymptotically standard normal.

- Forms of weight function: The true value of the weight function is unknown and must be determined biologically or empirically. If we set $\xi_j = 1$, then T is a burden test, which counts the total number of rare mutations each subject carries.
  - C test: the constant weight function.
  - F_u test: $\xi_j = [p_j(1-p_j)]^{-1/2}$, where p_j is estimated from unaffected individuals. This weight function was proposed by [Madsen and Browning, 2009].
  - F_p test: similar to F_u test except that p_j is estimated from pooled samples of affected and unaffected individuals.
  - Fixed threshold tests: assume common variants are not associated with the phenotype, set $\xi_j = 0$ if $p_j > c$, where p_j is the (MAF) of the jth SNP, and c is a given threshold. At $c = 0.01$, T_1 test; at $c = 0.05$, T_5 test.

- Variable threshold test: We consider K choices of ξ, which could correspond to different thresholds or different types of weight function, or both. Under H_0, the random vector $T(U_1, \cdots, U_K)$ is approximately K-variate normal with mean 0 and covariance matrix. For the two-sided test, we consider the maximum of the absolute test statistics.

- EREC (estimated regression coefficients) method for weight function: estimate weight function from data (important if the mutations have opposite effects on phenotypes).
  - If the choice of the weight function is not proportional to β or ξ is estimated from the data, then U is no longer the score statistic. However, the test statistic T is asymptotically standard normal under H_0 regardless of how ξ is determined.
  - Naive method: The optimal choice of ξ_j is β_j, which is unknown. We can estimate β_j from the data, e.g. its MLE. The main problem is that β_j's are highly variable (because the individual variants are very rare) and can be quite different from the true values.
  - Compromise: set $\xi_j = \hat{\beta}_j + \delta$, where δ is a given constant. The corresponding test statistic T will be asymptotically standard normal as long as δ is nonzero.

- Permutation test: important in small samples. In the absence of covariates, we simply permute the phenotype values Y_i's and calculate the test statistic T for each permutation.

- Discussion:
  - Comparison: Wald tests tend to be overly conservative (resulting in substantial loss of power) whereas likelihood ratio tests tend to be too liberal (resulting in excessive false-positive findings), especially for small n and low MAFs.
  - VT approach: improves upon that of [Price10] in three aspects: (1) it uses more powerful test statistics, (2) it can accommodate covariates, (3) it can be implemented by normal approximation instead of permutation.
  - EREC test: recommended for general use. Similar power to the tests assuming the same direction of effects when that assumption holds and is much more powerful than the latter when that assumption fails. Outperforms the HP, C-alpha and SKAT tests.

  - Our theory allows incorporation of any prior knowledge into the weight function. Efficient use of functional or bioinformatics information requires further investigation. It would be worthwhile to explore Bayesian methods.

- Remark:
  - Normal approximation: works well for large samples (in experiment, about 2,000 subjects). This is not the case in many NGS studies.
  - Not allow one to update the weight of SNPs: weighting is based on a sum of MLE and a constant. However, weights (the optimal values should be the actual effects of SNPs) should be part of the inference process.

Hierarchical Generalized Linear Models for Multiple Groups of Rare and Common Variants: Jointly Estimating Group and Individual-Variant Effects [Yi & Liu, PLG, 2011]

- Background: Bayesian Analysis of Rare Variants in Genetic Association Studies [Yi & Zhi, GE, 2011], using Bayesian GLM, rather than predetermining the weights of variants as previous methods, our approach jointly models the overall effect and the weights of multiple rare variants and estimates them from the data.
- Idea: define multiple groups of variants (a group could be all rare missense variants), and assume each group acts as a whole and could have a different effect. Conceptually, group can be thought of as a latent variable. Next, the effect of a variant to a group is not fixed: variants may have different effect in a group.
- Model: suppose we have K groups, the effect of the k-th group to the trait is g_k. For a variant j in the group k, its effect on this group (the burden of this group) is α_j. We also have covariants x_{ij} for the j covariate of the subject i, and the effect β_j. The trait of the i-th subject is given by:

$$y_i = \sum_j x_{ij}\beta_j + \sum_k g_k \sum_{j \in G_k} \alpha_j Z_{ij} + \epsilon_i \tag{4.266}$$

  where Z_{ij} is the genotype at marker j. For binary trait, we use GLM and replace y_i as η_i (the linear predictor). The prior of g_k is normal with mean 0. The prior of α_j is normal with mean μ_j, whose value is pre-specified (otherwise, g_k and α_j are coupled, and unidentifiable).

Testing Rare Variants for Association with Diseases: a Bayesian Marker Selection Approach [Zhang & Deng, Ann Hum Genet, 2012]

- Naive Bayes model: variant-specific prior of q, and relative risk:
  - Risk variant: $1/\gamma \sim \text{Beta}(\gamma_{l1}, \gamma_{l2})$,
  - Protective variant also allowed.
- Choose Beta(0.5,0.5) as the prior.
- Remark:
  - The default prior of RR makes little sense. At this prior, the relative risk is often as large as 100 for risk variants. For protective variants, most of the prob. mass of this prior is either close to 0 (highly protective) or close to 1.
  - In general, the prior is critical, and the method does not demonstrate the benefit of a good prior, and does not show how to obtain a good prior.

Incorporating prior biologic information for high-dimensional rare variant association studies. [Quintana & Conti, Hum Hered, 2012]

- Contributions:
  - Bayesian risk indeix (BRI) method: uncertainty of variants, and multi-level inference of both gene/region and individual variants.
  - iBRI method: integrate prior knowledge of variants as covariates, and allow multiple regions.
- BRI model: n subjects with p variants and q covariates. Note all variants are causal, so we let γ be the subset of causal variants, and M_γ denote the corresponding model. To incorporate direction of effects, let $\gamma_v = 1$ (risk) or -1 (protective), if the variant v is in the subset γ. Given a model M_γ, the burden of the i-th subject is called risk index, defined as $X_{i,\gamma} = \sum_{v=1}^{p} \gamma_v G_{iv}$, where G_{iv} is the genotype of the i-th subject at variant v. The model is then:

$$\text{logit}(Y_i = 1) = \beta_0 + \beta Z_i + \beta_\gamma X_{i,\gamma} \tag{4.267}$$

- The model can be easily extended to allow multiple regions.
- iBRI model: suppose we have covariates of variants, then we could have a second-level model relating the prior probability of whether v is included to the covariants. The probit model is used.
- Inference: first, the prior of M_γ is set s.t. the probability of at least one variant is included stays constant with the number of variants increased. Second, model selection tasks are: at least one variant in a gene is included (gene-level test using BF) and for any particular variant, it is associated (variant-level test). Finally, to sum over all models M_γ: to MH algorithm, and use Gibbs sampling to sample the second-level model parameters.
- Application to breast cancer data: 640 cases and 1272 controls, test BRCA1 gene (more than 100 rare variants). Gene-level association is extremely high with iBRI model, and 30 with non-informative BRI model. Table 1 shows the top variants: variant BF and supporting counts in cases/controls. In particular, the BF of BRI model of a variant with 2/0 counts is 13, but with iBRI, the BF increases to 600. This is because many LoF mutations exhibit associations, thus the prior probability is increased.
- Running time: 3h for 100K iteration on a single gene region (134 variants).
- Remark: comparison of our Bayesian approach:
  - Extremely slow, cannot be applied to genomewide.
  - Very simple prior information used: LoF and missense.

Identifying genetic marker sets associated with phenotypes via an efficient adaptive score test [Cai & Carroll, Biostatistics, 2012]

- Model idea: the variance component test loses power when the signal is sparse, so we estimate the effect of each SNP using data, and then down-weigh the SNPs likely non-informative.
- Remark: using a better prior to reflect the sparsity of informative SNPs, e.g. Slice-and-slab prior, can address the problem of variance component model?

The Value of Statistical or Bioinformatics Annotation for Rare Variant Association With Quantitative Trait [Byrnes & Li, GE, 2013]

- Goal: evaluation of weighting schemes for RVAT.
- Phenotype independent weighting scheme:
  - Collapsing approach (0 or 1) and burden approach
  - Madsen-Browning frequency weighting

- Phenotype dependent weighting scheme:
  - EREC-like weighting: first do regressoin to find $\hat{\beta}_j$ (either single-variate or multi-variate), then do burden test, with the wegiths equal to $\hat{\beta}_j$'s.
  - Penalized regression: Lasso, Elastic net, SCAN (penalize smaller coefficients more heavily than larger ones). After regression, the estimated coeff. will be used as weights.
- Statistical test: score test, and the significance determined through permutation.
- Simulation: for each region/gene, m causal variants, each variant is either risk or protective (with probability r). The effect size depends on q (the link functions). Also simultate the setting where bioinformatic tools can predict functionality of variants - the true variants has 90% prob. of being functional. In addition, there are a certain number (at least 1/3) of functional but non-causal variants.
- Results: in the absence of bioinformatic tools, vaiable selection methods (Lasso, EN) significantly outperform the rest. With such tools, the power of burden-type of tests is significantly boosted.
- Remark:
  - Phenotype dependent weighting using variable selection technique: better than independent weighting scheme, as expected (similar to random effects). However, all methods are slow because one needs permutation.
  - A major challenge is to combine statistical variable selection with bioinformatic predictions.
  - The effect-MAF functions in simulation are not realistic: small q should imply a larger effect size. Also under the simulation setting, power often reaches close to 100%, clearly unrealistic.

A unified mixed-effects model for rare-variant association in sequencing studies. MiST [Sun & Hsu, Genetic Epidemiology, 2013]

- Motivation: incorporation of variant characterisics and modeling variant-specific effect (heterogeneity).
- Model: let X_i be the covariates, G_i be genotypes, we have the usual GLM: $g(\mathrm{E}(Y_i)) = X_i\alpha + G_i\beta$. A prior model of β_j as:
$$\beta_j = Z_j\pi + \delta_j \tag{4.268}$$
where Z_j is a vector of variant annotations and $\delta_j \sim N(0, \tau^2)$. Thus the effect of a variant depends on its characteristics, but also allow individual variant effect. To allow for weighting, let $Z_j = w_j$, which depend on MAF of j. The null hypothesis is $H_0 : \pi = 0, \tau = 0$.
- Inference: instead of LRT, which is difficult for alternative model (estimating τ^2), use score test. Two test statistic, one for $\pi = 0$, the other for $\tau = 0$. To combine the two (independent) test, use Fisher's method (more powerful).
- Simulation: fix number of variants (10), with different MAFs. Simulate different models: all deleterious, both positive and negative effects, and a small fraction of causal deleterious variants (3 out of 10). Show MiST has better power than SKAT-O.
- Dallas Heart Study results: use annotations: missense, nonsense, frameshift. Found a couple of significant genes,with lower p-values than not using annotations. Also, p-value for τ is large, suggesting that variant effect heterogeneity is not important.
- Remark: limitations of the model
  - No borrowing information across genes: the ability to estimate π for each individual gene is probably limited in practice.
  - Additive effect of annotations.

– Prior of effect size: mean 0, but for rare variants, most causal variants should increase the risk.

A Variational Bayes Discrete Mixture Test for Rare Variant Association [Logsdon, GE, 2013]

- Model: the idea is to model the heterogenity of effects, or a mixture of causal and neutral variants. Suppose there are a proportion of causal variants in a gene, let Z_j be the indicator of causal variants. Z_j follows Bernoulli distribution with prob. p. The model regresses y with the sum of Z_j. Test the model $p = 0$ or $\theta = 0$ (the effect of causal variants).
- Remark: this is a special case of mixture prior (Slice-and-slab prior): β_j is a mixture of 0 and non-zero distribution. Here it uses the simplest non-zero distribution (1).

Meta-analysis of Gene-Level Associations for Rare Variants Based on Single-Variant Statistics [Hu & Lin, AJHG, 2013]

- Idea: obtain single variant statistics, and test the null hypothesis that none of the variant is associated with the trait. The single variant statistics can be easily combined across studies with meta-analysis.
- Model: suppose we have a gene with m variants, let Y be the phenotype of n individuals and G_j be the genotype of n individuals, then roughly speaking, the score statistic of variant j is:

$$U_j \propto \mathrm{Cov}\left(Y, G_j\right) = \sum_i Y_i G_{ij} \tag{4.269}$$

  Under H_0, U follow MVN distribution, with mean 0 and covariance dependent on LD. Specifically, the covariance of U_j and U_k under the null model is a linear combination of terms $\mathrm{Cov}\left(G_j, G_k\right)$, which is basically the LD matrix.
- Remark: the paper also derives the results under various burden tests. The burden test statistic is a function of U, and the MVN distribution of U should allow us to derive the distribution of other test statistics.

Design of DNA pooling to allow incorporation of covariates in rare variants analysis [Weihua Guan, PLoS ONE, 2014. Review]:

- Idea of DNA pooling: pool all DNA of cases and controls, respectively, then sequence all cases, and all controls. From the reads, infer the AFs in cases and in controls, respectively. Test the difference assuming binomial distributions.
- Limitations: individual identification is lost, and cannot control for covariates.
- Model idea/procedure: divide the samples based on covariates (i.e. group samples with similar covariates). Then sample genotypes from each group (pool) based on estimated AF in the group from the reads data. The last step is association test with regression: imputed genotypes and the covarates in each group.
- Estimating AF in each group: let p be the AF in a group of K subjects, and m be the number of alleles of a. Suppose we have n reads in the pool, x of which have allele a. Suppose the error rate is e, then the probability of observing a read with allele a is:

$$d = \frac{m}{2K}(1-e) + (1-\frac{m}{2K}) \cdot e \tag{4.270}$$

  The conditional distribution $x|m$ follows $\mathrm{Bin}(n, d)$. The distribution of m is simply $\mathrm{Bin}(2K, p)$. From these, we can estimate $P(m|x)$, the actual AF (counts) given the read data.

The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease, [Moutsianas & McCarthy, PLG, 2015]

- Some known cases of rare variants: NOD2 for Crohn's Disease. PCSK9 for coronary heart disease. LPL for hypertriglyceridemia (154 missense variants with MAF $< 0.1\%$, T1 association test). MTNR1B for T2D (13 variants, OR = 5.5, KBAC test).
- Simulation of genotype: HAPGEN2, mimic the observed SFS in 202 genes. Average length of coding sequences.
- Simulation of genetic architectures: sample up to 35 causal variants per loci, VE = 1% (variance explained, 1% is among the strongest loci found, TCF7L2 for T2D). Three main scenarios (AR1-AR3): strong, moderate or weak selection on causal variants, leading to inverse correlation between AF and effect size. AR4-AR5 are variations where only rare variants are causal. AR6: 50% risk and protective variants.
  - Frequency-effect size distribution: from simulations under selection on disease (T2D parameters: prevalence 8% and heritability 45%)
  - Sampling causal variants: Variant effects were sampled until the cumulative variance explained (VE) on the liability scale by each locus reached the desired threshold
- Power of methods: 1.5K cases and 1.5K controls
  - Power at 2.5E-6 is low ($< 20\%$) for all genetic architectures. Also low at $\alpha = 0.05$. MiST and SKAT-O are best, followed by KBAC.
  - At $\alpha = 0.05$: KBAC is best. This high sensitivity can be used, e.g. for candidate gene studies.
- Comparison of single-variant and gene-based tests: the comparative advantage of gene-based tests was most evident when there is strong purifying selection against causal alleles. Ex. AR4, power of single variant test is $< 5\%$ vs. gene-based test 20%. Under AR2 and AR3, single-variant test performs better, though each method detects unique loci.
  - Conclusions: single variant and gene-based association methods should be jointly employed for maximal power across divergent locus architectures (union in Figure 3).
- Sensitivity to fraction of causal variants: burden tests are highly sensitive to the fraction, while MiST and SKAT-O are relatively immnue.
- Effect of sample size: Even in 10K case-control samples, power remains modest (60% at $\alpha = 2.5\times10^{-6}$).
- Concordance between methods: high concordance between SKAT-O and MiST (0.92), between SKAT-O and KBAC (0.86 under AR2). The scenario where KBAC performs best: aggregate skew of case vs control counts.

4.9.2 Rare Variant Studies

Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease [Cruchaga & Goate, Nature, 2014]; TERM2 variants in Alaheimer's disease [NEJM, 2013]

- Strategy: use WES in families to identify candidate genes first, then use WES in case-control to find the risk genes.
- Family study: families with history of LOAD in four or more members. Choose 14 (out of 800) families to do WES, then choose candidates by: segregation of variants with status; filtering by MAF and function; shared by multiple families. Identify a single variant in PLD3.
- Case-control study to map PLD3: sequencing in 2,000 cases and 2,000 controls. Justification of PLD3: expression lower in LOAD, PLD1 and PLD2 previously implicated in APP trafficking and LOAD. Also functional study in N2A cells that express the gene: show that it affects APP and amyloid-β.

- Pattern of rare variant in PLD3 and TERM2:
  - Exists 1-2 relatively common (0.5-1.5%) variants that drive the burden. But by themselves, not strong enough.
  - Relatively common variants could be neutral, e.g. TERM2, a variant 25/31. This reduces the power of burden: TERM2 $p = 0.02$.
  - Possible RVs with large effects: e.g. 4/0 and 3/0 in PLD3 and 6/0, 3/0 in TERM2.

Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol [Lange, AJHG, 2014]

- Data: 2000 subjects, including 300 with extremely high and 247 extremely low LDL levels.
- RVAT: (1) 5 burden (CMC) test: including burden test on non-silent variants at 4 AF threshold, and burden test on LoF; (2) SKAT-O: LoF, LoF and missense, LoF and probably damaging.
- Single variant results: only one variant in APOE.
- Burden test: 3 genes reach threshold in stage 1 and several sub-threshold ones replicated. Find the best test for each gene. The patterns of association:
  - Relatively high frequency variants (about 1%): could drive association signal. But often neutral.
  - Very low frequency variants (0.1%): there are cases with clear signal, e.g. 5/1, 6/1, 8/0 in PCSK9.
  - For some genes, the signal is dominated by a few relatively common variants, e.g. PCSK9, PLD3; for other genes, overall burden of very rare variants, e.g. LDLR (non-silent), and APOB (LoF).

Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction [Nature, 2015]

- Study design: single variant association using arrays (less than 5%); rare variant burden test.
- Data: about 4000-6000 cases (early-onset MI) and controls, respectively.
- Negative results with single variant test. Gene-based test: limit to MAF $< 1\%$, several categories, NS, deleterious (PPH2, broad, strict), and LoF.
- Pattern in APOA5 and LDLR: (1) APOA5: burden in both relative common variants and singletons. (2) ORs: about 2 for NS, and 4-13 for LoF. (3) PPH2 and other annotations: sometimes increase the OR, but sometimes not (APOA5).

A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants [Fritsche, NG, 2016]

- Study design: exome chip. Do sequencing to find out RVs of some candidate genes, then include these RVs in the exome chip. Sample size: 17k cases and 17k controls.
- Single variant analysis: 34 independent loci. Most associated variants are common, and 7 are rare: frequency from 0.01% to 1%, and OR from 1.5 to 47.6.
- RVAT: conditioned on single variants, four genes show significant burden.
  - CFH: 88/38, with seven RVs with signal 6/0, 10/0, 8/0, 5/0, etc.
  - CFI: 213/82, with many RVs with signal, 17/2, 9/0, 6/0, 36/10, etc.
  - TIMP3: 29/1, with clear RVs 14/0, 5/0, 4/0, etc.
  - SLC16A8: driven by a signle splice variant 370/278.

- Gene priortization score (GPS) table: gene annotations include expression in retina, eye phenotype in mouse, RVAT burden, drug target, etc.
- **Remark**: a simple strategy of combining CV and RV is, limit the analysis to genes with some CV signal, thus reducing multiple testing burden.

Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7 [Luo & Anderson, NG, 2017]

- WGS low coverage on 4K IBD patients. Imputation from these sequences into existing GWAS cohorts of 16K cases and 18K controls.
- Single SNP analysis: a 0.6% missense SNP in ADCY7 that doubles the risk of UC.
- Burden analysis: low coverage, so correct for sequencing depth. A significant burden of damaging variants in a known Crohn's disease gene NOD2.
- Lesson: low-coverage sequencing, then imputation on very large cohorts, seems a good practical strategy.

4.10 Copy Number Variations and Structural Variations

Statistical tests of CNVs [personal notes]

- CNV burden analysis: test if CNVs (deletions or duplications, or separately) are enriched in cases than in controls. Standard test: Fisher's enrichment test. Multiple metrics of burden: CNV number/rate, CNV size and gene content.
- Testing individual CNVs or CNV regions (CNVRs): Fisher's enrichment test: individuals with a CNV vs. without a CNV. To control for systematic difference bewteen cases and controls, use regression analysis [Pinto, Nature2010; Raychaudhuri, PLG2010; CCRET].
- Gene set analysis: similar to testing individual CNVRs, define a variable for gene set disruption due to CNV and test its correlation with phenotype status using Fisher's enrichment test or regression [Pinto, Nature2010; Raychaudhuri, PLG2010; CCRET].
- Remark: CNV-level analysis and gene set analysis are typically done separately in literature. The ideal approach is to combine them: e.g. first infer CNVs, then translate the knowledge of CNVs into genes.
- Common problems:
  - Multi-gene CNVs: a CNV spanning multiple genes provides evidence of multiple genes, and the true evidence of a single gene is lower than CNV-level evidence.
  - Clustering of genes in a pathway: if we count the number of times a gene is disrupted, this could create spurious signal at the pathway level (as one CNV is counted multiple times).
  - Bias of CNV number and size in cases vs. controls: need to account for when analyzing a specific pathway.

Copy-number variation and association studies of human disease [McCarroll & Altshuler, NG, 2007]:

- Challenges of genotyping CNVs:
  - Most reported CNV locations actually correspond to the locations of CNV-containing regions (CNVRs), generally the genomic coordinates of a BAC probe, set of oligonucleotide probes, or fosmid from which a variant was discovered. A reported CNVR is consistent with a large number of potential variants

  - Until approaches for genome-wide CNP genotyping mature, a placeholder strategy may be to rely on raw hybridization measurements as an approximation to an unknown, underlying genotype.

- Using SNPs as markers for CNPs:
  - One would genotype the CNP in the HapMap (or other reference) samples and assess whether nearby SNPs were able to capture the CNP through linkage disequilibrium. Using available SNP data and PCR-based genotyping of deletion polymorphisms, initial studies found that deletion polymorphisms are generally ancestral and are tagged by SNPs.
  - CNV calling algorithms: iPattern implements a non-parametric density-based clustering model that integrates intensity data across samples to assign individual samples to distinct copy number states. QuantiSNP uses an Objective Bayes (OB) Hidden-Markov Model (HMM) approach for CNV callings

- Testing disease association of common CNPs:
  - For common CNPs, statistical tests will involve a straightforward comparison of allele frequencies between affected individuals and controls in a population cohort; between transmitted and untransmitted chromosomes in families with affected offspring; or between affected and unaffected siblings.
  - Some CNPs seem to involve more than three copy-number classes, and therefore more than two copy-number alleles. Need to extend methods: reduce to binary case (e.g. copy number < 4 vs. ≥ 4); logistic regression; etc.

- Testing disease association of rare CNPs:
  - Difficulty: given the existence of hundreds of rare CNVs with apparent frequencies of less than one percent, even in a well designed study it will frequently occur that a CNV is present (for example) in 3/200 cases and 0/200 controls. Such results are expected to occur by chance in a genome-wide search.
  - Collapsing: In the case of rare coding SNPs, a framework is typically used in which nonsynonymous SNPs are examined based on their a priori likelihood of functionality. In the case of CNVs, similar paradigms may be useful: for example, pooling just those CNVs confirmed as affecting a candidate gene's coding sequence and nearby highly conserved elements.
  - CNV burden analysis [Pinto et al, Nature, 2010]: use CNV size, CNV rate, etc. as measurs of burden, then compare the burden measures in cases and in controls. Ex. in cases, for a gene to be tested, there may be 10% chance of having a rare CNV, while the chance is 2% in the controls.

Human copy number variation and complex genetic disease [Girirajan & Eichler, ARG, 2011]

- Definition and detection of CNVs: CNVs typically between 1Mb (chromatomal aberrations) and 50 bp (indels). Detection relies on SNP microarrays, or CGH. The limit of CNVs using SNP microarrays is about 20-30 kb.

- CNVs in population:
  - Between any two individuals, bp difference due to CNV is > 100 larger of SNPs.
  - Common CNPs (copy number polymorphism), $> 1\%$, often in multicopy number states (0 to 30).
  - Rare CNVs (many may be de novo), often larger (> 100 kb), under strong selective pressure.
  - Large CNVs are individually rare (more than 71% CNVs larger than 100kb are rare), but collectively common. 65 - 80% individuals carry at least a large CNV, 1% carry a large CNV > 1Mb in size.
  - De novo CNV rates: 8K - 25K bases of genomic DNA were added or lost during each transmission.

- Mechanisms of CNVs: commonly caused by nonallelic homologous recombination (NHAR) between high-identity segmental duplications, causing "hotspots". The frequency of these events partly determined by the size and sequence identity of the flanking segmental duplications.
  - Segmental duplication: large (> 1kb) blocks that have nearly identical sequences (> 95%), as a result of duplication. SDs can be tandem or interspersed, and can be interchromosomal or intrachromosomal.
  - A large fraction of CNV disease burden is contributed by other nonhomologous mechanisms (nonrecurrent or lower frequency), such as microhomology/microsatellite- mediated break-induced repair. Establishing the disease relevance is harder for these CNVs.
- Evidence supporting association of rare CNV and diseases:
  - Frequency of pathogenic CNVs (known to be associated with diseases, called "genomic disorder") in children with developmental delays (Figure 2). Ex. 16p11.2 appears in 4%. Many of them are recurrent, and highly penetrant.
  - In individuals with developmental delays or IDs, 10% have de novo CNVs of size 500kb to 12Mb.
  - Many recurrent CNVs are associated with *variable penetrance and expressivity.* Ex. 15q13.3 with developmental deley, autism, SCZ, epilepsy.
  - Example of pathogentic nonrecurrent CNVs: in a SCZ study, 362-kbp microduplication overlapping VIPR2 was detected in 2.5% cases, but only 0.03% of controls.
- Linking rare CNVs to rare diseases:
  - Clinical diagnosis: large CNVs are considered pathogenic if they arise de novo in proband, are rare (< 1%). Diagnositc yield between 5-20%.
- Linking rare CNVs to common complex diseases:
  - De novo CNVs in Autism: Sebat et al, 165 case families with 99 control families, 7.2% de novo CNVs in cases vs. 1.02% in controls.
  - De novo CNVs in SCZ: some studies fround 2-3 times burden in SCZ, while another study fails to find using 1,000 cases and 1,000 controls in SCZ. One study of 359 SCZ families, 8-fold enrichment in de novo CNVs, and 1.5 fold in rare inherited CNVs. In Simplex families, de novo CNVs more common, in multiplex families, burden of rare inherited CNVs.
  - Possible sources of difference in CNV burdens across studies: sample quality, cell line artifacts, probe resolution, GC content, lack of genotype in- formation, subphenotype characterization and clinical heterogeneity, age of onset of disease, and platform-specific biases.
  - A model of de novo and inherited CNVs in autism: (1) One study found evidence of strong dominant transmission. Model: de novo CNVs in low-risk families, while inherited CNVs (from mother) in high-risk families. (2) ANother study found 4-fold enrichment of de novo CNVs in cases in multiplex families. Model: sensitized genetic background, requires de novo CNVs.
- Methods for studying rare CNVs in complex diseases (Figure 4):
  - De novo CNVs (FIgure 4a).
  - Comparion of CNV frequency in cases and in controls. Figure 4b.
  - Gene-based or sliding window approach: for each gene, compare the number of CNVs covering this gene in cases vs. in controls. Figure 4c.
  - Comparison of CNV rates in cases vs. controls at a given CNV size (Figure 4d): could be used for estimating OR.

  - Pathway-analysis: Figure 4e.

- Model of rare CNVs (genetic architecture): multiplex cases occur in high-risk familes. In low-risk families, autism may occur with de novo high-risk CNVs.

- CNPs in human population: CNPs are 4-10 fold enriched in regions of segmental duplications. Bi-allelic or multicopy CNPs.

  - Mechanisms leading to CNPs: microhomology of sequences at the breakpoints, NAHR,andL1 retrotransposition. Almost all cases of CNPs are associated with SDs.
  - CNPs are enriched with immune genes and environment response pathways, suggesting possible positive selection. Some example that CNP frequency difference across population may be driven by positive selection: CCL3L1 CNV in HIV, more common in Afrian ancestry; OCLN, which is required for hepatitis C viral entry; a taste receptor cluster on chromosome 12.

- Role of CNPs in diseases:

  - Assessing CPNs. For SNP arrays: often lack of probes in duplication rich regions, almost half of deletions are not captured by high-density SNP arrays. Also all array approaches are based on reference genome, missing many variations (e.g. novel insertions).
  - Methods for studying CNP association with disease: (1) For low copy numbers: first obtain integer copy numbers, then test association (Figure 5a). (2) For high copy number CNVs: may not be able to assign integer copy numbers, compare the distribution in cases vs. controls.
  - Using SNPs to test association of CNPs: mostly focus on biallelic CNPs, the majority of which are in high LD with SNPs. However, CNPs in segmental duplication-rich regions show less LD to SNPs.
  - CNPs are often associated with immune diseases (Table 1), but also examples of T2D, obesity, CHD.

Autism genome-wide copy number variation reveals ubiquitin and neuronal genes [Glessner & Hakonarson, Nature 2009]

- Data: Autism Case-Control (ACC) cohort of 859 ASD cases and 1,409 controls. For replication, use AGRE (1,336 cases) and 1,110 controls. On average, 15.5 CNV calls per individual (similar in cases and in controls).

- 16p11.2: relatively common in controls (3), and not statistically significant.

- Segment-based scoring approach: scan the genome for SNPs, and measure the SNP copy number changes in cases vs. controls using Fisher's exact test. Then for a CNV region (CNVR), find the local minimum and use permutation to assess p-values.

- Report 8 CNVs (Table 2): most CNVs have 0-1 copies in controls, and 3-4 in cases. OR about 5 for those that occur in controls. Most have P-values from 0.001 to 0.04. (Unadjusted). Each CNV contains about 1-2 genes. Four genes are ubiqutin gene family.

- Gene-based scoring approach: for each gene, consider the CNV calls affecting this gene region in cases and in controls. Identify 7 further genes at incresaed frequency of CNVs in ASD cases vs. controls.

Functional impact of global rare copy number variation in autism spectrum disorders. [Pinto & Scherer, Nature, 2010]

- Data: rare CNVs (less than 1%, greater than 30k in size) in 996 ASD cases (876 trios) and 1,287 controls.

- CNV burden analysis in cases vs. controls: use three measures of burden per individual, the number of CNVs, the estimated CNV size, and the number of genes affected by CNVs. Use permutation to assess the burden. Only the last one (genic CNVs) show a burden of 1.19 (1.26 if deletions only).
- Methods for gene set burden analysis:
  - Compare CNV frequency in cases vs. controls: define the proportion of indivdiuals containing CNVs overlapping the test gene set, and compare using Fisher's exact test.
  - Regression analysis for CNV effect on phenotype: regress of phenotype on the number of genes in the test set overlapping CNVs.
- CNV burden analysis with ASD gene lists: known ASD genes (36) - burden is 1.8, or 4.3% in cases versus 2.3% in controls. ID genes (> 100): burden 2.1.
- Gene set burden analysis: Novel sets include: GTPase/Ras signaling, microtubule cytoskeleton. Functional map of gene sets: gene set as nodes and overlap between sets as edges.
- ASD candidate genes: based on de novo CNVs in cases but not controls, SHANK2, SYNGAP1 and DLGAP2.

Accurately Assessing the Risk of Schizophrenia Conferred by Rare Copy-Number Variation Affecting Genes with Brain Function [Raychaudhuri & Daly, PLG, 2010]

- Motivation: possible biases for CNV-based gene set analysis if the analysis is performed at gene level:
  - Brain genes tend to be bigger: case only analysis to compare the relative frequency of gene sets could be biased.
  - Different CNV rates and sizes in cases vs. controls: if CNVs are more common in cases, then any gene set may have higher chance of being more frequent in cases than in controls.
  - Clustering of genes in a single pathway in the chromosome: if we count the number of genes affected by a CNV in cases vs. controls, we may double count the same genes.
  - Loss of information of CNV rates affecting a gene: if we compare the number of genes in a pathway affected by CNVs in cases. vs. the number in controls, a gene overlapped CNV in both cases and controls would not contribute, regardless of the rates of CNVs in cases or controls.
- Model: let y_i be the phenotype of individual i, and g_i be the count of gene within a pre-specified gene set affected by a cnv in i. We regress y_i on g_i, while controlling for confounders: c_i, the number of CNVs of individual i and s_i, the average size of these events.
- Simulations: simulate CNVs of cases and controls. Randomly place genes but brain genes are bigger. Introduce bias: e.g. larger CNVs in cases, or higher CNV rates in cases. Test the enrichment in brain genes: high type I error in the models that do not account for the bias.
- Gene size bias: brain genes tend to be bigger. In CNV controls, genes affected by rare CNVs are involved disproportionately in brain function. Fisher's exact test: using no. of genes (fraction of disrupted brain genes vs. fraction of brain genes in the genome).

Convergence of genes and cellular pathways dysregulated in autism spectrum disorders [Pinto & Scherer, AJHG, 2014]

- Data: 2,845 ASD familes and 2,640 control families. Rare CNVs: 1% and 30k or larger in size. 6.8K CNVs in total.
- Global burden in rare CNVs: 1.41 fold. CNV sizes: 188kb in cases and 159kb in controls.

- De novo CNVs: 102 rare de novo CNVs, 4.7% of cases have at least one de novo CNV vs. 1-2% in controls. The average length of de novo CNVs: 1.17 Mb is larger in probands than siblings (0.55Mb), and affect more genes (3.8 fold) in cases than in controls. Gene content: about 18 in de novo CNVs.

- CNV burden in genes implicated in ASD and ID: OR = 4.1 for CNVs overlapping dominant or X-linked genes. OR = 12.6 for Pathogenic CNVs and OR = 23.1 for pathogenic deletions. More than 50% of pathogenic CNVs are de novo.

- CNV burden for gene sets: FMRP targets (n = 842) and PSD genes (n = 1,453) carried a significant excess of both deletions (OR = 2-3) and duplications (OR = 1.5-2) in affected subjects. Gene sets show excess in deletions but not duplications: high brain expression (OR = 1.89), dominant neurological diseases and orthologous genes associated with abnormal phenotypes in heterozygous knockout mice (OR = 2.9), haploinsufficient genes (pHI > 0.35, OR = 1.4)

- Predictive model of disease risk: identify factors (burden) that correlate with disease status. Use logit model, where the explanatory variables are: number of genes in CNVs, number of deletions, number of brain-expressed genes in deletions, etc, and the outcome is the disease status. Results: number of deleted brain-expressed genes correlate with disease status (Figure 3C).

- De novo SNVs (LoF) and de novo CNVs converge on functional gene networks: use NETBAG to identify genes in de novo CNVs, then assess the overlap of this gene list with de novo LoF genes. Found 11 common genes, such as NRXN1, SHANK2 and RIMS1 (synpatic rab3a-binding protein, new finding).

- Lessons:

  - **Key statistics of CNVs**: (1) De novo CNVs: about 4% in probands and 1-2% in controls; size about 1Mb and 18 genes. (2) Inherited CNVs: rare CNVs about 1-2 per subject; size 150-200 kb and 1-3 genes.
  - **Logit modeling**: as alternative way of burden analysis. Good for controlling other variables.

Refining analyses of copy number variation identifies specific genes associated with developmental delay [Coe & Eichler, NG, 2014]

- Data: aCGH data of 29,085 primarily pediatric cases with intellectual disability, developmental delay and/or ASD in comparison to 19,584 adult population controls.

- Burden of rare CNVs (1%, $\geq$ 500kb): burden for deletions, OR = 5.09, for duplications, OR = 1.76. Analysis of 2,086 transmissions showed that likely deleterious CNVs were transmitted preferentially from mothers.

- CNV analysis: some CNVs are associated with SDs, and they are recurrent, so we use standard case-control comparison. Other CNVs may overlap, but have different breakpoints in different subjects (Multiple Breakpoints, or MB), and we need a different analysis.

  - CNV level analysis: counts in cases and controls, Fisher's exact test (Table S2). 2,000 CNVs in 55 regions. 19 regions were likely pathogenic and reach nomial significance. Most (40/55) are genomic hotspots flanked by segmental duplication.
  - Gene-level analysis: 1,945 genes enriched for deletions and 2,633 genes enriched for duplications at $P < 0.01$, one tailed Fisher's exact test.
  - Region analysis: many genes were clustered, so perform region level analysis, using Fisher's exact test. Regions are defined using rare case CNVs of > 250kb: find unique breakpoints to define the boundaries. Use simulation to obtain p-values. Found 14 significant regions, most not flanked by SDs.

- Methods for combined CNVs and SNVs: for each gene, count the number of times the SNV (or CNV) occurs in cases vs. in controls. Let a and b be the number of cases with and without CNVs, and c and d the numbers of controls. Define similar numbers a_2, b_2, c_2, d_2 for SNVs. Then test statistic is:

$$Z = \frac{a}{a+b} + \frac{a_2}{a_2+b_2} \tag{4.271}$$

  Assess Z using controls: hypergeometric distributions.

- CNV and SNV analysis:
  - In genes with single LoF (247 genes): 43 show excess of deletion in CNV case-control data, the burden is OR = 1.15, not significant. In 21 genes with 2 or more LoF mutations, OR = 2.72.

- Remark:
  - The combined test: the treatment of CNVs is exactly the same as that of SNVs. Ex. if SNVs and CNVs data have the same sample size, then the test statistic is basically the total number of events affecting a gene (SNV and CNV). No effort has been done to address the fact that a multi-gene CNV only "partially support" a gene.
  - The de novo SNV information is not used, which is significantly more informative.

Novel Findings from CNVs Implicate Inhibitory and Excitatory Signaling Complexes in Schizophrenia [Pocklingtong & Owen, Neuron, 2015]

- Data: 11K cases and 16K controls from three datasets. CNVs called from genotyping arrays. Analyses are based on large, rare CNVs ($>$ 100 kb, frequency $< 1\%$).

- Background: common variants from GWAS explain only 1/3 of heritability, and the 108 GWAS loci explain only a small fraction.

- Intuition: to test if a gene is associated with risk, we check among all CNVs affecting this gene, how many are in cases vs. in controls. If this ratio is large, it suggests that this gene is likely a risk gene. But we need to compare this ratio with some background. Also we should control for the possible difference between the background CNVs and the CNVs covering this gene: e.g. CNV size. This is best done by a regression model.

- Testing association of a feature (gene or gene set) with phenotype: suppose we have n CNVs (CNV occurrences, more precisely). For each CNV, it has some features, gene size, whether it contains a gene, etc. We want to test if a feature increases the chance that the CNV occurs in cases. Use logistic regression model.

- Gene set and single-gene analysis: use the same logistic regression model. (1) 134 predefined CNS gene sets from MPI Mammalian Phenotype (MP) database. (2) For single gene test, the problem: the p-values will be highly correlated for genes of the same recurrent CNV.

- Enrichment of de novo nonsyn. mutations in gene sets: define a minimum CNS gene set (capturing all gene set enrichment singal). In this set, a burden of about 2 in deletion only, and duplication only.

- Remark/Question:
  - In Table 5, why enrichment of DNM is lower in combined set, about 1.2, vs. deletions or duplications along (about 2)?
  - Problem of applying logistic regression on single gene anlaysis: correlation of adjacent genes.

A New Method for Detecting Associations with Rare Copy-Number Variants (CCRET) [Tzeng, PLG, 2015]

- Motivation: while cnv-enricnment-test (PLINK) addresses some of the biases in CNV analysis, it does not address the effect heterogeneity issue: between-locus (across CNVs) and within-locus (deletion vs. duplication of the same CNV).
- Problems: (1) test CNVR dosage effect (DS): could be applied to one CNVR (CNV region) or a group or all CNVRs, to test if dosage disruption has any effect. (2) Test gene set (GI: gene interaction): test if disruption of genes in a set has any effect.
- Model idea: similar to PLINK CNV test, control for difference of CNV size and numbers between cases and controls using fixed effects. But instead of collapsing all genes of the same pathway, consider the heterogeneity of possible effects. Similarly, for burden analysis of CNV dosage effect, consider heterogeniety of dosage effects across CNVs.
- GI effect model: we are testing a particular pathway. Let Y_i be phenotype of individual i, and $\mu_i = \mathrm{E}(Y_i)$. Let $\tilde{Z}_i^{len}$ the average length of i. Let Z_{im} be the disruption indicator of gene m in i. We have the model:
$$g(\mu_i) = \beta_{Len}\tilde{Z}_i^{len} + \sum_m \beta_m Z_{im} \tag{4.272}$$
where $g(.)$ is the link function, $\beta_m \sim N(0, \tau^2)$ is the random effect of gene m. The model is equivalent to a model where we use a random effect variable that is correlated across individuals (covariance matrix is given by the kernel).
- DS effect model: similar to GI model, but use CNVRs instead of genes.
- Simulation: CNV data of 2,000 cases and 2,000 controls, sampled from > 6K individuals in TwinGene study. Identify 1,757 CNVR from 2,000 samples.
  - DS model study: sample 600 causal loci from 1757 CNVRs. For each causal CNV, sample its effect β_m from a constant (no heterogeneity), or from a mixture of risk and protective effects (change the proportion). The effect size/OR is from 1-7.
  - GI model study: sample from 69 PSD genes that intersect with 1700 CNVRs (among a total of 668 PSD genes). Then do similar sampling for their effects: OR is chosen 1.5.
- Remark: some issues with the model
  - The GI model does not address the issue that a CNV carries information of multiple genes. So a CNV covering three genes can contribute to the evidence of all three genes.
  - Multiple counting problem in gene set analysis: Z_{im} in the DS model encodes whether gene m is disrupted in subject i. However, when two genes in a pathway are physically close and disrupted by a CNV, the two genes will be counted twice.
  - CNVR is defined as (sort of) the union of multiple overlapping CNVs. However, what is more interesting is the intersection of multiple CNVs with supporting evidence. By focusing on CNVR, we lose information and actually make it harder to identify causal genes.

CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits [Mace, NC, 2017]

- Data: UKBB (110K) and GIANT (56K), BMI, height, etc. Association test: infer dosage for each probe, and test each of about 1M probes. Multiple testing correction: 29K independent tests.
- Genomewide significant hits: 7 CNVs, with AF 0.01-0.07%.
- Example: 16p11.2 region, about 9 genes, similar p-values for many probes in this region (Figure 2).
- MC4R region: also coding SNP show association. Independent signals.

- For all CNVs found, except 16p11.2, CNV signal and common SNPs are independent.

Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects [Marshall and Sebat, NG, 2017]

- Data: 20K cases and 20K controls. CNVs: at least 10 probes, > 20kb and MAF < 1%.
- Global burden and gene set burden: use number of genes as burden metric, global burden = 1.2 (strongest). Use number of CNVs as burden metric, OR = 1.03. Significant gene burden (esp. deletions) in 15 out of 40 gene sets.
- Individual loci/genes: (Table 1) 16 loci at BH FDR < 0.05. About half are known. AF in controls: 0-7, and OR 3-20. Also a small set of protective CNVs.
- CNV breakpoint analysis: use breakpoint as analysis unit.

The impact of structural variation on human gene expression [Chiang and Hall, NG, 2017]

- Data: 147 GTEx samples ini 13 tissues. Test 9000 common SVs. Found 5000 associations at 3K genes, and 1.6K SVs.
- 10% of SV-eQTLs: change exons, most show effect consistent with SV classes.
- Estimating contribution of common SVs to gene expression: (1) Joint analysis with SNVs: estimate about 3.5% of eQTLs are due to SVs. (2) Use LMM to estimate h2g from SVs: estimate about 7%. (3) For individual SVs: 30-50 times more likely to be eQTLs vs. SNVs.
- GWAS: 52 loci where causal SV-eQTLs are in LD (>0.5) with GWAS variants. Some examples (Fig. 4): 1.4Kb deletion in an intron, LD (0.7) with GWAS SNP of RA.
- Rare SVs on gene expression: rare SVs are 16 times more likely to be close to gene expression outliers than SNVs.
- Lesson: about 7% contribution of eQTLs are from common SVs. Some SNP associtions are likely driven by common SVs. Rare SVs probably also important, though under-powered to detect.

Properties of structural variants and short tandem repeats associated with gene expression and complex traits [Jakubosky and Frazer, NC, 2019]

- Data: 350 iPSC samples, WGS, and association of SNVs, indels and SVs with expression. Classes of SVs: DEL, DUP, MEI (mobile element insertion), STR, and mCNVs (multi-allelic). Total of 42K common SVs.
- LD between common SNVs and SVs: for DEL and STR, 80% common SVs are tagged by common SNVs, however, for DUP, only 30%.
- Mapping eQTLs: use all variants or use only SVs. About 6000 eGenes shared: about 14% of them lead eVariants are SVs. This suggests that at least a certain fraction of eQTLs are explained by common SVs.
- Association with SV sizes: for SVs >50kb, the chance of being lead eVariant has OR = 3.
- Spatial distribution of eQTLs: 2% in exons, 9% in promoters and 17% introns.
- Enrichment of eQTLs (SVs) near chromatin loops: use Pc-HiC data from iPSC. Variants overlap with distal anchors (the other anchor is promoter) are 3-5 times more likely to be eQTLs, about 5-6% vs. 1-2% by chance (Fig. 5e).

- Multi-gene SVs: SVs that are in chromatin loops of multiple genes: more likely to be multi-gene eQTLs.
- Comparison with GWAS: 40% of eVariants (SVs) associated with two or more eGenes were in strong LD with a GWAS variant vs. 20% for eVariants associated with only one eGene.

Mapping and characterization of structural variation in 17,795 human genomes [Abel and Hall, Nature, 2020]

- Data: 17K WGS, 40% European, 9% Finnish, 24% African, 16% Latino. NIH Common Disease (CCDG), PAGE (population) and Simons Diversity Panel.
- Computational workflow: Ext. Figure 1. per sample variant discovery, merging and break point refinement, per sample re-genotyping and copy number, cohort level VCF.
- SV distribution per sample: Figure 1b, 4000 SVs, 35% deletions, 25% mobile element insertions (MEIs), 11% tandem duplications. AFS similar to SNVs, with most common. About 100 rare SVs per sample, and 10 ultra-rare SVs. Total of 200K rare SVs.
- Burden of rare deleterious SVs: coding sequences. Total of 42K rare SVs affect genes: 9K deletions alter gene dosage, 26K function changes (e.g. single exon deletion), 7K increase gene dosage. Most rare SVs are deletions (55%) and insertions (42%). 23% of all SVs affect ≥ 2 genes. Mean 4.2 rare SV-altered genes per sample, vs. 33.6 by small indels and SNVs.
- Method for analyzing burden of rare deleterious SVs: for a given variant class (SNV, indels, del, dup), use singleton rate (percent of singletons among all variants) as a measure of selection. Also use CADD and LINSIGHT to define impact scores of variants. Then assess singleton rates among different variant classes and different impact scores. Use syn. SNVs as control: about 40% singleton rates
- Burden of rare deleterious SVs: noncoding sequences. Figure 3c: high-confidence LOF in SNVs, singleton rate > 0.6 and at the top impact score, singleton rate > 0.7. For Del: singleton rates in coding are even higher, for noncoding lower, but similar to high-confidence LoF. For Dup: weaker, in noncoding, similar to missense SNVs. In all variant classes, impact scores modest increase of singleton rates.
- Estimation of rare deleterious variants in an individual: Figure 3d. Use the impact-score threshold based on LOF SNVs to choose a cutoff for defining deleterious rare variants. 120 deleterious rare variants per individual, $> 60\%$ are SNVs, and 17% SVs (majority are deletions). A given rare SV is 841-fold more likely to be strongly deleterious than a rare SNV. Median length of rare deleterious SVs is 4.5kb vs. 2.8 kb for all rare SVs. Top 50% of noncoding rare SV-deletions are as deleterious as LOF SNVs and indels: measured by singleton rates.
- Ultra-rare SNVs: about 0.01 per sample, mega-sized.
- Deletion and duplication sensitivity scores: (1) for each gene: based on the observed frequency of CNVs. (2) For 1kb windows: scores correlate with LINSIGHT scores and other annotations. Scores: roughly, for each window, define presence and absence of SVs. Also each window we have annotations, so we can define log-OR for each annotation.

Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia [Halvorsen and Sullivan, NC, 2020]

- Data: WGS from 1162 Swedish schizophrenia cases and 936 ancestry-matched population controls.
- Burden analysis: URVs defined as once in the study, no in Gnomad. Do aggregated burden analysis.

- SNVs and indels: (1) coding: burden in LOF only, and larger in pLI genes, OR = 1.2. (2) noncoding: promoters, conserved sequences (CDTS and GERP) within putative CREs - FIRE, ATAC, etc. No burden.
- Ultra-rare SVs: total of 7K DEL, 2K DUP and 700 INVs. Overall burden in DEL, OR = 1.08. Specific classes: TAD boundary (brain), H3K27ac, H3K4me3, ATAC, FIRE, Hi-C loops (intersected with different gene sets, including GWAS, FMRP, etc.). Only TAD boundary shows significant enrichment (94 SVs) OR = 1.6, CTCF OR = 1.4, K27ac OR = 1.2.
- Estimation of h2g by WGS: pedigree based about 0.6. SNP heritability: common, about 0.48. Use WGS, about 0.52, however, large SE.

Chapter 5

Cross-Phenotype Analysis and Mendelian Randomization

5.1 Multi-Trait Analysis

Multi-trait analysis: an overview [personal notes, Debashree Ray's paper]

- Problem: suppose we have K traits, in n subjects. We are interested in the association of a SNP with the K trait, let β_k be the effect of the SNP on the k-th trait. The K traits can have correlation structure. The model can be written as:

$$Y_{ik} = \beta_k x_i + \epsilon_{ik} \tag{5.1}$$

  where i is the index of subject and k that of the trait. The error term of the i-th subject is iid: $\epsilon_i \sim N(0, \Sigma)$, where Σ captures the covariance among traits.

- Intuition why testing multiple traits can be more powerful than univariate test: (1) suppose we have a pleiotropic SNP, and we consider the case $K = 2$, and for simplicity, assume $\beta_1 = \beta_2 = \beta$. So comparison of Y_1 and X, and Y_2 and X, both provide information of the shared parameter β. In particular, when Y_1 and Y_2 are uncorrelated (after removing X), both Y_1 and Y_2 provide independent information of β. (2) Surprisingly [Stephens13], even if only $\beta_1 > 0$, if Y_1 and Y_2 are correlated, incorporating Y_2 can be still useful: when testing Y_1 on X, we conditioned on Y_2, this removes some of the variance of Y_1, thus increase the power. A more formal analysis: in this case, our model is:

$$Y_1 = \beta_1 X + \epsilon_1 \quad Y_2 = \epsilon_2 \tag{5.2}$$

  And ϵ_1 and ϵ_2 are correlated. Then the model can be rewrite as something like: $Y_1 = \beta_1 X + \gamma Y_1 + \epsilon_1'$. So the joint model is equivalent to conditional test, controlling for Y_2.

- Example: suppose we are testing association of a SNP with height, and we have weight data, which is correlated with height. Height varies in the population, not only by genetics, but also by environment such as diet. While we do not have diet, weight is a good marker of diet. So incorporating weight as covarite would remove the variation of height due to environment, thus improving the power. When weight itself may be associated with the same SNP, controlling weight when testing association with height would remove some signal, so multivariate analysis would be preferred.

- Univariate tests: univariate test of each trait, then combine the results together. Ex. let p_k be the p-value from the k-th trait, then we can combine them using Fisher's method, or minimum p value. The null distribution however needs to take into account the dependence between traits.

- MANOVA: we test the model in Equation5.1. Consider the case of $K = 2$. We have:

$$Y_1 = \beta_1 X + \epsilon_1 \quad Y_2 = \beta_2 X + \epsilon_2 \tag{5.3}$$

  where ϵ_1, ϵ_2 follows MVN with covariance matrix specified by σ^2 and $\rho\sigma^2$. The hypothesis being tested: $H_0 : \beta_1 = \beta_2 = 0$, and $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$. The test can be done using LRT, which is equivalent to the MANOVA test statistic (Wilks Lambda), see [Ray & Basu, 2015]. Intuitvely, test statistic is the sum of squared error, divided by the total variance of y (sum of squared error plus the explained variance of y) - this is similar to the F-test for regression: explained variance divided by the mean squared error (unexplained). When test statistic is large, it means that X can explain a lot of variance, thus H_1 should be accepted.

- Comparison of uni- and multi-variate approach: we consider four scenarios (true model) and understand which approach is preferred:

  - Scenario 1: $\beta_1 > 0, \beta_2 = 0, \rho \to 0$, Y_2 does not provide any information of Y_1, and we still need to pay for extra d.o.f. under the multivariate test, so the univarite test is more powerful.
  - Scenario 2: $\beta_1 > 0, \beta_2 = 0, \rho \to 1$: Y_2 provides some information of Y_1, thus conditioned on Y_2, we can reduce the variance of Y_1, increasing the power of detecting β_1. In 2D plot of Y_1, Y_2, the data points of different x would be more separable for X in Y_1, when we condition on Y_2. In [Stephens, PLoS ONE, 2013], this is captured by Y_U: unassociated traits, but important to adjust.
  - Scenario 3: $\beta_1 = \beta_2 > 0, \rho \to 0$: pleiotropic SNP with independent traits, both trait provide independent information of β. So the multivariate test is more powerful.
  - Scenario 4: $\beta_1 = \beta_2 > 0, \rho \to 1$: the two traits are redundant, so adding Y_2 does not help much with learning β. The test will lose power [Ray & Basu, 2015]. Alternative scenario is the "indirect association" model of [Stephens, PLoS ONE, 2013]: when $Y_1 = \beta_1 X + \epsilon_1, Y_2 = Y_1 + \epsilon_2$, then $\beta_1 = \beta_2$, and if $\epsilon_1 >> \epsilon_2$, we have $\rho \to 1$. With indirect association, multivariate tests also loses power, since Y_2 is entirely redundant/noninformative for β_1.

- Test that incorporates trait correlation only through distribution: the test statistic is some kind of combination of the multiple tests, but consider only the correlation through the distribution of the test statistic. For example, Sum of Squared Score (SSU) test. Let U_k be the score statistic of the k-th test, and we have:

$$U_k = \frac{1}{\hat{\sigma}_0^2} Y_k^T X \tag{5.4}$$

  where $\hat{\sigma}_0^2$ is the MLE of σ^2, Y_k is the n-dim vector of the k-th trait, and X the n-dim vector of genotype. The SSU statistic is $T = \sum_k U_k^2$. Under H_0, T has an approximate asymptotic scaled and shifted chi-squared distribution.

- Comparison of MANOVA and SSU-type of approach (not univarite, but incorporate correlation only through the test distribution): the SSU statistic will have a large variance under H_0 because of the correlation ρ between two traits (intuitively, when U_1 is large by chance, U_2 will also be large because of correlation, thus T will be even larger).

- Importance of phenotypic covariance: n many practical problems: e.g. eQTL, the covariance among multiple traits are NOT due to biological pleiotropy, and removing this covariance (or source of trait variance) would be important to achieve higher power.

- Remark:

  - In the model here, Σ reflects the covariance of traits after removing the effects of the SNP; it may be different from the "marginal" covariance among traits.
  - To visualize: plot Y_1 and Y_2, and use different colors for values of X (discrete). The value of β is reflected through the difference of mean of Y_1 or Y_2, between different groups defined by X.

Questions/ideas about multi-trait analysis:

- Can we determine which test to use, univariate or multi-variate, before we do the test? One complication is that in the model ρ is defined as correlation after correcting for x, but we would not know this until we test for the SNP.
- Idea: use shared heritability of two traits (instead of phenotypic correlation) to set the prior of effect sizes.
- How do we find eQTL hotspot? The vast majority of expression traits are not correlated (even though considerable covariance exist) and for each SNP, it is associated with only a small number of traits. This seems to suggest Scenario 1 is most applicable, thus we prefer univariate test. Can we do clustering analysis first, then for the cluster, apply the multivariate test?
- Group-level test: suppose we are interested in testing if a SNP affects a set of traits, can we use a random effect model of β_k, and test the mean of β_k? Also, can we use the random effect model to improve the power of testing indivudial trait (borrow information from the group)?

Moving toward system genetics through multiple trait analysis in genome-wide association studies [Shriner, Frontiers in Genetics, 2012]

- Dimension reduction: PCA, and association on eigen-traits.
- Multivariate methods: (1) MANOVA, Seemingly Unrelated Regression (similar to MANOVA, except that the predictors can be different/unrelated). The common assumption is that the errors are independent across individuals, but correlated for different traits of the same individual. (2) GEE.
- Categorical and mixed outcome:
  - Bivariate logistic regression: four parameters π_{ij} where $i, j = 0, 1$. Only three are independent: 2 marginal prob and the OR that relates the two variables.
  - Bivariate probit regression.
  - Mixed continuous and binary traits: one possible model: we use a bivariate normal distribution, one for continuous trait, the other as underlying hidden variable for binary trait. Allow correlation between the two traits.
- Biological pleiotropy: a small number of loci affect many traits, but most affect a small number of traits.

Pleiotropy in complex traits: Challenges and strategies [Solovieff & Smoller, NRG, 2013]

- Examples of cross-phenotype (CP) associations: cancer, autoimmune diseases and psychiatric diseases.
  - AID: 44% of SNPs associated with one AID associated with another AID.
  - CACNA1C: CP effect on bipolar and SCZ.
- Biological pleiotropy: true effects on multiple phenotypes. Multiple scenarios: single causal variant, or different causal variants of the same gene. Interpretation of CP: could be distinct effects of the same allele in different cell populations; or truly mulitple consequences.
- Mediated pleiotropy: causal variant affects P_1 which affects P_2. Ex. SNPs to LDL to MI risk.
- Spurious pleiotropy: design artefact; or causal variants in different genes (but in LD). Design issues may be:
  - Ascertainment bias: e.g. patients with two diseases may be more likely to be recruited in the study.

  - Phenotype misclassification: e.g. SCZ and bipolar.
  - Shared controls: e.g. population stratification or batch effect, that affect shared controls but not cases.

- Establishing genetic overlap between traits: often the first step in the CP analysis. Approaches: polygenic scores, or LMM.

- Multivarate approaches: require all phenotypes to be measured on the same individuals and often require individual level data.

  - Multivaraite regression (MANOVA) for continuous traits. Extension: GEE for to allow non-normally distributed phenotypes.
  - Multiple categorical traits: log-linear model, Bayesian network.
  - Ordinal regressino: genotype as outcome and phenotypes as predictors.
  - Dimension reduction: PCA or canonical correlation analysis, find the linear combination of traits that explain the covariation with genetic variants.

- Univaraite approaches:

  - Step-wise analysis: use significant SNPs from one trait, then test its association with another trait (use smaller number of SNPs). Underpowered.
  - Fixed effect meta-analysis. Problem: same directions of SNPs.
  - Random-effect meta-analysis: allow heterogeneity, but still has problem with effects of opposite directions.
  - Cross-phenotype meta-analysis (CPMA): explicit testing of CP: null hypothesis of no additional association, vs. alternative hypothesis of two or more associations. Also deal with opposite effects.
  - Subset based meta-analysis: exhausitvely evalaute all possible combinations of non-null models for association.
  - Testing multiple associations using minimum (univariate) p-values across multiple traits. TATES (Extended Simes).
  - Regional test: use the number of trait associations with p value less than a threshold in a region.

- Example of CP analysis in practice: PGC study (Box 2).

  - Polygenic analysis: find significant overlap between SCZ, bipolar and MDD. Also ASD and SCZ, to a lesser extent.
  - Univariate GWAS of five traits.
  - Fixed-effect meta-analysis: identify some CP effects (bar-plot).
  - Identify which phenotypes drive the association: use multivariate approach to search for subset of traits (model selection) - log-linear model approach.

- Characterizing CP:

  - Fine-mapping to distinguish biological pleiotropy and spurious pleiotropy.
  - Identifying mediated pleiotropy and MR.

- Applications of CP analysis:

  - Disease classification.
  - Drug repurposing: target common biological pathways in related disorders.

– Drug side effects: e.g. anti-TNF therapy for Crohn and UC, but may increase the risk of multiple sclerosis.

Pervasive Sharing of Genetic Effects in Autoimmune Disease [Cotsapas and Daly, PLG, 2011]

- Cross-Phenotype Meta-analysis (CPMA): the idea is to do something like fixed-effect analysis, but use p-values. Under H_0, p is uniformly distributed, thus $-\ln(p)$ is exponentially distributed with rate 1. Under H_1, we assume $-\ln(p)$ follows exponential distribution with $\lambda > 1$. We assume λ is the same for all diseases, thus allow us to do LRT:

$$CPMA = -2 \ln \frac{P(D|\lambda = 1)}{P(D|\lambda = \hat{\lambda})} \tag{5.5}$$

  CPMA follows χ^1 with df 1 under H_0. Comparing with Fisher's method of combining p-values, it does not pay the penalty of high df.

- Analysis of 7 AID data: start with 107 SNPs associated with at least one disease, a total of 44 SNPs have CPMA $p < 0.01$, while the expected number is only 1.

- Remark: the key assumption is λ is the same across multiple studies, or p-value distributions under H_1 are the same. This is not true, for example, when the power of the studies are different.

A Unified Framework for Association Analysis with Multiple Related Phenotypes [Stephens, PLoS ONE, 2013; Michael Turchin talk]

- Model: suppose we test association of a SNP with d traits, possibly related. The key idea is to distinguish direct and indirect associations. Let γ be a parition of traits into three groups: direct association Y_D, indirect association Y_I, and unassociated Y_U. The relationship:

$$g \to Y_D \to Y_I, \quad Y_U \to Y_D, \quad Y_U \to Y_I \tag{5.6}$$

  Given γ, we can obtain the marginal likelihood (and BF) of γ. This allows us to test any specific hypothesis by model averaging: e.g. whether g is associated with any trait.

- Testing association: to assess if a SNP is associated with some trait, we compute the BF for a particular γ, where H_0 is the SNP is not associated with any trait:

$$BF_\gamma = \frac{P_1(Y_D|Y_U, g)}{P_0(Y_D|Y_U)} \tag{5.7}$$

  It can be shown that the BFs for different γ can be compared to assess which γ is more likely. Also note that: for different SNPs, γ may be different.

- Model with summary statistics: the model needs covariance between phenotypes, how do we obtain that from summary statistics? Idea: consider estimated effect sizes of a SNP wrt. the two phenotypes: $\beta_1 \propto x^T y_1$ and $\beta_2 \propto x^T y_2$. If y_1 and y_2 are correlated, then β_1 and β_2 of null SNPs are also correlated. In fact, one can show the relationship of two correlations. So the procedure is: estimation of how effect sizes under null are correlated (removing large effect SNPs).

- Importance of distinguishing Y_I and Y_U: when we have multiple traits, but SNP is only associated with some of them, and the traits are not highly correlated, doing multivariate analysis actually loses power. So association analysis should be done only on Y_D, while explaining Y_I or incorporating Y_U in association.

- Relation to MANOVA: suppose we ignore Y_I for now. Consider two traits, Weight W and Height H. There are two alternative models γ: $g \to W, g \to H$ (both direct association), and $g \to W \leftarrow H$ (one indirect association). The first is MANOVA, and the second is testing association with W while controlling H. The two tests are similar, when g explains only a small fraction of variance in W.

- Regression model: the term $P(Y_D|g)$ is given by Bayesian multivariate regression. Consider a case with a signle SNPLet V be the covariance matrix of traits, the prior of β (one for each trait) follows $N(0, cV)$, where c is a constant.
- Question: when Y_D and Y_I each has multiple traits, the relationship could get complicated, e.g. a trait in Y_I depends on some trait in Y_D but not other ones.
- **Remark**:
  - The model analyzes each SNP independently, without enforcing certain relationship between phenotypes.
  - The assumption of prior in Bayesian regression: the prior of effect size follows the same covariance as the traits. This may not be the best prior model.

Genome-wide Association Analysis for Multiple Continuous Secondary Phenotypes (SMAT) [Schifano & Lin, AJHG, 2013]

- Motivation: if we have multiple correlated secondary phenotypes in case-control, we can jointly analyze the data to increase the power (shared effects of one SNP on multiple traits) and also detect pleiotropic effects. To model the correlation among the traits, use GEE (similar to longtidual data analysis).
- Basic model: let y_{ij} be the j-th phenotype of the i-th subject, x_i be the covariates, s_i be the genotype, we have:
$$\mathrm{E}\left(y_{ij}|x_i, s_i\right) = x_i\beta_j + s_i\alpha_j \tag{5.8}$$
where β_j is the effects of covariates on the j-th phenotype, and α_j the effect of SNP on the j-th phenotype (assuming heterogeneous effects on the phenotypes). The correlation among y_{ij}'s can be modeled using GEE.
- Scaled marginal model: if the traits are positively correlated and measure the same underlying trait, we can assume a one-DF model. Let $\sigma_j^2 = \mathrm{Var}\left(y_{ij}|x_i, s_i\right)$ be the phenotype-specific variance, then we assume that the SNP has a shared common effect on the means of the scaled phenotypes:
$$\mathrm{E}\left(y_{ij}|x_i, s_i\right)/\sigma_j = x_i\beta_j + s_i\alpha \tag{5.9}$$
And we can test $H_0 : \alpha = 0$.
- Need to account for ascertainment issue (case-control sampling).
- Remark: can we use a random effect model of SNP effects on the M phenotypes, assuming that the effect of the same SNPs on different phenotypes tend to be similar?

Identification of risk loci with shared eff ects on five major psychiatric disorders: a genome-wide analysis [PGC, Lancet, 2013]

- Data: 33K cases and 27K controls of European ancestry, five disorders, ASD and ADHD for children, and BPD, MDD and SCZ for adult-onset diseases.
- Fixed-effect meta-analysis: combine all five disorders to increase the power of detection. Use weighted Z-score approach, where the weight is given by the inverse of standard error (or square root of the sample size). Results: 4 independent regions reaching genomewide significance.
- Omnibus test: meta-analysis allowing different effects in different disorders. Testing $H_0 : \beta_1 = \cdots = \beta_5 = 0$ vs. $H_1 : \exists i, \beta_i \neq 0$, using 5 df. LRT. The test statistic is effectively the sum of χ^2 for all five disorders. Results: no significant loci.

- Model selection: for SNPs reaching genomewide significance, and for SNPs previously identified for BPD and SCZ, we want to test if they have pleiotropic effects on other disorders, and what is the best-fit model (i.e. which subset of disorders are associated). We compare 13 different models (Table S1), and choose the one using BIC. A model describes a configuration of the risk of a SNP in all disorders, e.g.

$$\text{ASD} = \text{SCZ} \neq 0,\ \text{ADHD} = \text{BP} = \text{MDD} = 0 \tag{5.10}$$

  In other words, suppose for a SNP, its effect in five diseases are $\beta_1, \cdots, \beta_5$, each model specifies some constraint on β_j's. And the likelihood is given by $P(y_j|x_j, \beta_j)$, where y_j and x_j are phenotype and genotype data of j-th disorder.
  Results: forest plot, four significant SNPs all have pleiotropic effects. For other less significant SNPs or earlier identified BP, SCZ SNPs, a moderate fraction of SNPs have pleiotropic effects in at least one more disorder.

- Polygenic risk score analysis across disorders: define one disorder as discovery set (e.g. SCZ), find all score alleles at a p-value threshold, then assess their contributions in the target set (e.g. ASD). Results:
  - Three adult-onset disorders have larger overlap, especially SCZ and BP. The overlap of ASD and SCZ and ASD vs. BP, are still significant, but reduced.
  - In all cases, R^2 explained variance from SNPs in another disorders are small, generally $< 2\%$.

- Limitations of the study: (1) Model selection: in some cases, the best fit and second-fit models are similar; (2) Diagostic misclassification may happen, though it cannot explain all the observed disease overlap.

An atlas of genetic correlations across human diseases and traits [Bulik-Sullivan and Neale, NG, 2015]

- Definitions: genetic covariance defined as $\sum_j \beta_j \gamma_j$ where β_j, γ_j are effect sizes of SNP j on the two traits. Genetic correlation: normalized, -1 to 1, it is asymptotically proportional to MR estimate.

- MOM estimation: $E(z_{j1} z_{j2}|l_j)$ is a linear function of l_j (LD score), and the intercept represents the overlap among samples from two traits.

- Derivation [personal notes]: we have $y_1 = X\beta + \delta$ and $y_2 = Z\gamma + \epsilon$, where β, γ are effect sizes. Note that we use different genotype variables, since the samples are mostly independent. The summary statistics:

$$\begin{aligned} \hat{\beta}_j &= (X_j^T X_j)^{-1}[X_j^T X\beta + X_j^T \delta] \\ \hat{\gamma}_j &= (Z_j^T Z_j)^{-1}[Z_j^T Z\gamma + Z_j^T \epsilon] \end{aligned} \tag{5.11}$$

  Now we consider the covariance of the two, using Law of Total Covariance. We consider the conditional distributions with given β, γ, then marginalize them.

$$\text{Cov}\,(\hat{\beta}_j, \hat{\gamma}_j) = \text{E}_{\,\beta,\gamma}[\text{Cov}\,(\hat{\beta}_j, \hat{\gamma}_j)|\beta, \gamma] + \text{Cov}_{\,\beta,\gamma}(\text{E}\,(\hat{\beta}_j|\beta_j), \text{E}\,(\hat{\gamma}_j|\gamma_j)) \tag{5.12}$$

  The first term captures covariance of $\hat{\beta}_j$ and $\hat{\gamma}_j$ due to sample overlap, and the second term captures the covariance due to genetic correlation of β and γ. For the first term, we have:

$$\text{E}_{\,\beta,\gamma}[\text{Cov}\,(\hat{\beta}_j, \hat{\gamma}_j)|\beta, \gamma] = \text{Cov}\,(X_j^T \delta, Z_j^T \epsilon) = \text{Cov}\,\left(\sum_i X_{ij}\delta_i, \sum_{i'} Z_{i'j}\epsilon_{i'}\right) \tag{5.13}$$

  For all non-overlapping samples, the term is 0. And for the second term, we have:

$$\text{Cov}_{\,\beta,\gamma}(\text{E}\,(\hat{\beta}_j|\beta_j), \text{E}\,(\hat{\gamma}_j|\gamma_j)) = \text{Cov}_{\,\beta,\gamma}(X_j^T X\beta, Z_j^T Z\gamma) \tag{5.14}$$

  Note: we could also just take Equation 5.11, and compute the covariance of the two, marginalizing β, γ, i.e. treating both β, γ and δ, ϵ as random.

- Results in PGC: same individual-level data. Comparison with REML: LDSC and LDSC with constrained intercept. LDSC has significantly larger std error. LDSC with constrained intercept has somewhat larger standard error and the mean estimate sometimes differ quite a bit. The difference is larger for traits with small sample sizes.

Detection and interpretation of shared genetic influences on 40 human traits [Pickrell, NG, 2016]

- Data: summary statistics of 43 GWAS, including some unpublished ones from 23andMe. First map all variants associated with each trait separately: predefined blocks, and one variant per block (fgwas).
- Variants associated with pairs of traits: method to detect such variants is similar to COLOC. Estimate π_j's (for each model) from all blocks. Note: π_i here are sensitive to sample size, and represent detectable shared genetic influence. Observation that many variants are related to multiple traits, with no obvious relationship of effect sizes.
- Overlap between pairs of traits: Use the proportion of variants that affect one trait, would also affect the other trait (asymmetric).
- Causal relationship between traits: all variants ascertained in one trait X, and compare the effect sizes in two traits: $\hat{\beta}_{XX}$ and $\hat{\beta}_{XY}$. Obtain the rank correlation ρ_X. Similarly, obtain the rank correlation using variants ascertained on trait Y, ρ_Y. Intuition is that if $X \to Y$, then ρ_X should be large, but ρ_Y is close to 0 (Y can be affected by many factors other than X). Testing is simple: likelihood model of 2 rank correlations, under causal model, one of them should be 0; under non-causal model both are 0 or the two are equal.
- Remark:
  - Estimating genetic overlap: the gain of power comes from a large prior for the model where a SNP is shared between traits. Comparing this with an alternative prior based on correlated effect sizes.
  - Causality inference: same difficulties are Sherlock. Ex. other possible non-causal model not included is that the two traits share some genetic influence.

Playing Musical Chairs in Big Data to Reveal Variables Associations [Ashard, NG review, 2016]

- Why current multivaraite methods such as MANOVA are not successful? (1) Testing composite null: $\beta = 0$ for all traits, not informative. (2) When a SNP is associated with only a small fraction of traits, multivariate test could lose power (Scenario 1 in the univariate vs. multivariate analysis: other traits are not informative, but we have to pay for high d.o.f).
- Analysis: when we test one trait X on Y, we decide if we adjust for covariates (other traits). When a covariate C is uncorrelated with X, but explains some variance of Y, then adjusting for C is advantageous (Figure 1A, B). However, when C itself is affected by a SNP X, then adjusting for C can lead to errors:
  - Suppose X has an effect on Y: When $X \to C$, and $C \to Y$ (or both C and Y affected by some hidden U), then adjusting for C would remove some effect of X on Y. This loses power. Figure 1C.
  - Suppose X has no effect on Y: since C and Y are correlated, let's say it is created by a confounder U: $Y \leftarrow U \to C$. But we have $X \to C$, this leads to the collider case: $U \to C \leftarrow X$. So adjusting C will create dependency of U and X, but $U \to Y$, then X and Y becomes correlated conditioned on C.

  The general idea is thus: for any covariate C associated with Y, if it is independent of SNP X, we adjust for C, otherwise not.

- Notation: let β be the effect of X on Y. Let γ be the correlation between Y and C. We let δ be the association of X and C. Our model can be written as:

$$\mathrm{E}(Y) = X\beta + C\gamma \qquad \mathrm{E}(C) = X\delta \tag{5.15}$$

- Problem with p-value based filtering: we could test if X has a non-zero effect on C. However, p-value from this test is used to reject the null (zero-effect), while our goal is to reject the alternative (zero-effect). Say, we reject C at 95% level, but there could be many covariates that we fail to reject because of limited power. Call these type 1 covariates (associated with X). There is an additional problem, however, caused by type 2 covariates (not associated with X). If there covariates happen to have large $\hat{\delta}$, we will not include these covariates (fail to adjust).

- Claim: failure of adjusting for type 2 covariates using p-value filter would lead to inflated type 1 error when testing association between X and Y. We consider the null model: $\beta = 0$ and $\delta = 0$ (type 2 covariates). First consider the case with only one covariate. Suppose $\hat{\delta}$ is large. Under H_0, we have $Y = C\gamma + \epsilon_Y$, but since $\hat{\delta}$ is large, C and X would appear associated. Then Y and X would appear associated.

- MC algorithm: we should then include C if $\hat{\delta}_l$ is not rejected. We consider two intervals, in the unconditional test, we obtain inclusion area for $\hat{\delta}_l$ (95% confidence interval). But we also need to consider the conditional test of $\hat{\delta}_l$: when we know $\hat{\beta}$ (marginal effect of X on Y) and $\hat{\gamma}$ (correlation of C and Y), we can know δ better. Intuitively, $\hat{\beta}$ and $\hat{\delta}$ are correlated because Y can C are correlated: the larger γ is, the higher $\hat{\beta}$ and $\hat{\delta}$ are correlated. So we consider the conditional distribution $\hat{\delta}_l|\hat{\beta}, \hat{\gamma}$, under null model $\beta = 0, \delta = 0$:

$$\mathrm{E}(\hat{\delta}|\hat{\beta}, \hat{\gamma}, \beta = 0, \delta = 0) = \hat{\gamma}\hat{\beta} \tag{5.16}$$

  The two tests have two inclusion areas, and we take the union of the two as the inclusion area (OK to include/adjust C_l if its $\hat{\delta}_l$ is in the area).

- Results: work on a large number of phenotypes.

- Comparison of MC algorithm and [Stephens, PLoS ONE, 2013]: the covariates C that are included are similar to Y_U in [Stephens13]. The covariates that are not included (due to effect of SNP on them) are treated as multiple Y_D's, and multivariate analysis is used in Stephens13.

- Q: Figure 3 unclear: is the unconditional inclusion area in (b,c) the same as the one in (a)? If so, the inclusion area is larger than the simple case. This does not address the problem of limited power? Also,

- Remark: a general question is how do we estimate an effect when we have both direct measurements and indirect ones. Ex. suppose we want to estimate $X \to Y$, and we can estimate their association directly, say β. But we also have Z, which correlates with Y, and we have $X \to Z$. There could be multiple such Z's. How do combine all the information to estimate β?

- **Lesson**: in general, we should adjust for covariates associated with Y to increase the power and avoid false association, e.g. adjusting for ancestry PCs or PEER factors in eQTL. However, when the covariates are genetic, we should not adjust for them.

Genome-wide associations for birth weight and correlations with adult disease [Horikoshi and Freathy, Nature, 2016]

- GWAS of birth weight: 150K samples, 60 significant loci. Explain 2% of variation, chip-heritability 15%.

- Genetic correlation of BW and other traits: positive with anthorpometric and obesity related traits; but negative with CAD.

- Pattern of genetic correlations (effect sizes) across traits (Figure 2): choose BW loci, and plot their effect sizes on a number of related traits. The effect sizes of SNPs show cluster patterns (generally sparse).
- Understand genetic correlation between BW and blood pressure: mostly negative correlation across loci, however, some level of heterogeneity. Specific locus: e.g. a SNP with large effect on both BW and blood pressure, it is cis-eQTL of CYP17A1.
- Lesson: overall genetic correlation hides the heterogeneous pattern of effect sizes.

Multi-trait analysis of genome-wide association summary statistics using MTAG [Turley and Benjamin, NG, 2018]

- Model: for SNP j, its estimated effect (vector) $\hat{\beta}_j|\beta_j \sim N(\beta_j, \Sigma_j)$, where Σ_j captures correlation of estimation error. And the prior $\beta_j \sim N(0, \Omega)$.
- Inference: from Bayesian perspective, we can infer $p(\beta_j|\hat{\beta}_j, \Omega)$. MTAG uses Generalized Method of Moment (GMM) estimator: let estimator of β_j be linear combination of $\hat{\beta}_j$, then solve $\mathrm{E}(\hat{\beta}_j) = \hat{\beta}_j$.
- Estimation of Σ_j: use LDSC to estimate the intercept term for the diagonal element of Σ_j. Then use bivariate LDSC to construct the non-diagonal terms.
- Estimation of Ω: the marginal distribution $\hat{\beta}_j \sim N(0, \Sigma_j + \Omega)$. So $\mathrm{Cov}(\hat{\beta}_j) = \hat{\beta}_j\hat{\beta}_j^T = \Sigma_j + \Omega$. Then for all SNPs, estimate the average Ω.
- Analysis: potential problem, when SNPs are truly null for one trait, but non-null for others, the effect for the null trait would be biased away from 0. This may lead to high FDR when the GWAS for the two traits have very different power.
- Simulation study: the bias from ignoring sampling variation in estimators of Ω and Σ_j are small.
- Results on three behavior traits: (1) Increase of power on individual traits: number of significant loci (after LD clumping) for each of the three traits (Figure 4). (2) Improving polygenic prediction using independent data: MTAG has better R^2 than GWAS (Figure 6A).
- Remark: a major limitation is the homogeneous assumption of Ω. The method is only applied to traits with high genetic correlation 0.7.

Genome-wide association study results for educational attainment aid in identifying genetic heterogeneity of schizophrenia [Bansal and Koellinger, NC, 2018]

- Background: Education attainment (EA) is negative correlated with SZ, but has positive genetic correlation. Why?
- Proxy phenotype analysis: ascertain on EA GWAS, 500 SNPs, about 130 associated with SCZ at $p < 0.05$, and 21 passing Bonf. threshold (highly enriched).
- Sign inconsistency of 21 loci: half, half. Also about half show colocalization (using fine-mapping).
- Prediction of clinical phenotypes: if using SZ PRS, not correlated. If group them by sign consistency with EA, significant correlation with SCZ severity.
- Lesson: use PRS defined on subset of related SNPs may better capture biologically relevant factors

Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis [Udler, PLoS Med, 2018]

- Background: previously, unsupervised/hier. clustering, using half of SNPs here. Subtyping T2D: in contrast to our clusters of genetic loci, these clusters are defined using clinical data and biomarkers at the time of diabetes diagnosis.
- Data: 94 T2D variants (with LD pruning) and 47 T2D related traits, including glycemic traits from MAGIC, anthroprometric traits from GIANT (e.g. height, BMI), adipose tissues, birth weight and lipid levels, leptin, and fatty acid traits (e.g. omega 3). Also 10 clinical outcomes such as stroke/CAD and kidney disease.
- Bayesian NMF clustering: to deal with the non-negativity constraint, duplicate the columns (47 traits) to have both positive and negative. Then NMF on 94 x 94 matrix. Learn 5 clusters.
- Testing association of clusters with clinical outcomes: for each cluster, select SNPs based on weights in NMF using cutoff of 0.75. To test association: "genetic risk score (GRS) for each cluster with each GWAS trait or outcome (GWAS GRS) was performed using inverse-variance weighted fixed effects meta-analysis using summary statistics from GWAS.
- Clustering results: 5 dominant factors/clusters. No. selected SNPs per cluster: 5-30.
- Association of cluster GRSs vs. clinical outcomes: beta-cell cluster associated with stroke, lipid related cluster associated with blood pressure.
- Clusters are distinctly enriched for tissue enhancers or promoters (Figure 2): e.g. cluster 1 (beta-cell) are enriched in enhancers of islet and liver.
- Application of clusters to patients with T2D: T2D patients with many phenotypes. First, association of cluster GRS with traits, e.g. beta-cell cluster with BMI and CRP. Ex. those with extreme GRS in the Beta Cell cluster (N = 1,068) had decreased BMI, HC, and WC comparing with all patients.
- Analysis: why NMF is a poor tool for GWAS factor analysis? (1) Non-negativity constraint: not applicable to GWAS. (2) Sparsity: important: variant effects on factors should be sparse. (3) Using Z-scores: not scaled properly. The linear relationship should be defined in terms of effect sizes, just as in MR. Z-scores of variants depend on allele frequencies, which vary across variants.

Pleiotropic mapping and annotation selection in genome-wide association studies with penalized Gaussian mixture models (iMAP) [Zeng and Xiang Zhou, Bioinfo, 2018]

- Model: let β_j be the effect of SNP j on two traits (vector). It follows mixture of bivariate normal distribution with prior weights $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$ with covariance Σ. Fit the parameters, including β, γ and Σ by EM.
- Incorporating annotations: prior of mixing weights π_{jk} for SNP j, pattern k (4 patterns, 00, 01, 10 or 11), depends on X_j (annotations) by a multinomial logistic regression with linear term $X_j b_k$ for annotation k. Use penalized regression L_1 norm on b_k's. The objective function is the sum of $L(b)$ the log-likelihood, defined in terms of $\mathrm{E}(\gamma_{jk})$ (response variable in mlogic regression), and penalty of b_k's. Use general optimization technique.

Discovery of shared genomic loci using the conditional false discovery rate approach (condFDR) [Hum Genet, 2019]

- Motivation: given two phenotypes, if they show significant polygenic overlap, how do we leverage that to find more associations?
- Step 1: conditional QQ plot. QQ plot of primary phenotypes, conditioned on p-values of SNPs in the second phenotype.

- Step 2 (Box 1): Computing FDR for each strata of SNPs, based on p-values in the second phenotype. Compute the local FDR for each SNP at each bin, based on the distribution of p-values in that bin.

Genomic SEM Provides Insights into the Multivariate Genetic Architecture of Complex Traits [Grotzinger, Nature Hum Behavior, 2019]

- Model: $y = \Lambda\eta + \epsilon$, where y is $k \times 1$ effect vector on k phenotypes, and Λ is the loading matrix, and η is $m \times 1$ vector (SNP to factor effect).
- Fitting SEM: obtain the covariance matrix $\Sigma(\theta)$ of SNP to phenotype effects, where θ are parameters (loading matrix) and equate to the observed covariance matrix S. Estimation of S is based on LDSC, allowing for sample overlap (residual correlation).
- Simulation: (1) Use only summary statistics - details not clear. (2) Use GCTA to simulate genotype and phenotype data: 10K causal variants.
- Learning SNP effects: once factor model is learned, plug in individual SNP (one at a time), and learn their effects on factors.
- Confirmatory factor analysis (CFA) on 5 psychiatric traits: a single general latent factor, called p-factor.
- Exploratory factor analysis (EFA): use 2 or 3 factors, and do model comparison.
- Multivariate GWAS: 128 loci of the p-factor in the case of 5 psychiatric traits, 27 new loci.
- Polygenic scores of factors: use the general latent factor in 5-phenotype case, show it predicts better the symptoms.
- Analysis: how this compares with FLASH? The model still assumes homogeneity of SNP effects - same distribution on factors across all SNPs. Imagine we have two factors, A has large effects on group 1 traits, and B large effect on group 2 traits. But a SNP can only affect factor A or B, but not both. For a SNP affecting A: it will have large effect on group 1, but no effect on group 2. However, with polygenic model, the SNP is expected to affect B, so expected to have some effects on group 2 as well, so in order to explain the observed lack of effect of the SNP on group 2, the model will have to fit with a smaller effect of factor B on group 2 traits. In simpler words: a SNP acts on only A should not be used when we estimate the effects of factor B on other traits; but by including them and assume non-zero effects, we may push down the estimates of factor B on other traits.
- Analysis: alternatively, in truth, we expect 2 distinct clusters of SNPs, some have large effects on group 1 only, and the other cluster large effect on group 2 only. But the model expects a homogenous group of SNPs.

Characterisation of the genetic architecture of immune mediated disease through informed dimension reduction [Burren and Wallace, biorxiv, 2020]

- Motivation: (1)the SNP effect estimates must be on the same scale. (2) Variable correlation between input dimensions (SNPs) due to LD; (3) All SNPs are expected to show small deviations between studies due to random noise, different genotyping platforms and data processing decisions.
- Background: DeGA method, LD-thinning and p-value cutoff $p < 0.001$. However, this will make the results dominated by larger GWAS.
- Model: use PCA. To deal with challenge with (1), use $\hat{\beta}$ divided by σ_{MAF}, the variance of estimation from MAF, but not sample size. To deal with the challenges (2) and (3), favor SNPs that are likely causal variants. Do fine-mapping separately on all traits, and compute the weighted average of PIPs, w. For a SNP, its input to PCA is: $\hat{\gamma} = w\hat{\beta}/\sigma_{MAF}$.
- Selection of driver SNPs for PCs: most elements are close to 0, so use hard thresholding.

- Projection of new GWAS dataset to PC space: learn the contribution of PCs to the new trait.
- Importance of weighting SNPs (Figure 2): if do not use weighting, in PCA, traits are ordered by their data source (UKBB would be clustered), rather than trait types.
- Application to immune traits: 14 traits, PC1 is about autoimmune axis.
- Application to a large set of traits in UKBB: group the traits by their projections (Figure 3). Some clear patterns: IBDs form a cluster, some cancers form a cluster. Also use correlation of PCs with the phenotypes to interpret PCs.
- Remark: the SNP weighting by PIP is not a good strategy. In high LD regions, PIPs would be small. So low PIP would reflect LD, rather than true effects.

5.2 Mendelian Randomization

Questions about MR [personal notes]

- Suppose we have a valid IV of X to Y, and we know a confounder Z, should we adjust for Z in the MR analysis? Conceptually, MR addresses exactly the problem of confounding.
- Power of MR: how the power depends on the strength of IV (measured by PVE)?

Problems of MR: summary data [personal notes]

- Adjusting for confounders: if we know the confounder U, and we adjust U when estimating $\hat{\beta}_{X|G}$ and $\hat{\beta}_{Y|G}$, then the ratio estimate is unbiased.
- Confounding assumption: we cannot directly test if $G \to U$.
- 2-sample MR: is weak IV still a problem?
- 2-sample MR: if we use data of exposure to select the strong IV(s), then do MR, we can suffer from selection bias.

Analysis of MR and its issues:

- The effect of not including confounding variables in MR. Our true model is:

$$X = \alpha G + \gamma_X U + \epsilon_X \qquad Y = \beta X + \gamma_Y U + \epsilon_Y \tag{5.17}$$

  where ϵ_X and ϵ_Y are independent. We plug in the first equation into the second:

$$Y = \beta\alpha G + (\beta\gamma_X + \gamma_Y)U + (\beta\epsilon_X + \epsilon_Y) \tag{5.18}$$

  Note that G and U are assumed to be independent, however, in finite sample, they can be correlated. This leads to biased estimate when regressing Y vs. $\hat{x} = \alpha G$ (see weak IV bias in the MR book). Note that we assume $\hat{\beta}_X = \alpha$, which does not account for finite-sample.

- Reverse causation: MR does not address this and it is possible that the results are due to reverse causation. This violates MR assumption (G affects Y without going through X), but it may not be detectable. One can show that the 2SLS or ratio estimate leads to estimated effect of $1/\beta$, where β is the effect of Y on X. Let the true model be $G \to Y \to X$,

$$Y = \alpha G + \epsilon_Y \qquad X = \beta Y + \epsilon_X \tag{5.19}$$

  We have $\hat{\beta}_{X|G} = \beta\alpha, \hat{\beta}_{Y|G} = \alpha$, so the ratio is $1/\beta$.

- Pleiotropy without causal effects (independent of confounding): suppose our model is $X \leftarrow G \rightarrow Y$, without relationship between X and Y. Then MR will give non-zero estimate.

Weak IV bias in MR [personal notes]

- Causal model: let α_j be the effect of SNP j on X, and β the causal effect of X on Y. Let α_U, β_U be the effect of U on X and Y.
- Weak IV bias under 2SLS model: assuming there is a single IV. Let ΔG be difference of genotypes, ΔX and ΔY be the difference of exposures and outcomes. We first show that in the first regression model, the estimate of α_1 (G to X effect) is biased:

$$\hat{\alpha}_1 = \frac{\Delta X}{\Delta G} = \frac{\alpha_1 \Delta G + \alpha_U \delta U}{\Delta G} = \alpha_1 + \frac{\alpha_U \Delta U}{\Delta G} \tag{5.20}$$

  So when fitting the first regression model, we introduce a bias to α_1. Let $\tilde{X} = \hat{\alpha}_1 G$ be the predicted X. Next, we consider the second regression, the estimated causal effect is:

$$\hat{\beta} = \frac{\Delta Y}{\Delta \tilde{X}} = \frac{\Delta Y}{\hat{\alpha}_1 \Delta G} = \frac{\beta \Delta X + \beta_U \Delta U}{\Delta X} = \beta + \frac{\beta_U \Delta U}{\alpha_1 \Delta G + \alpha_U \Delta U} \tag{5.21}$$

  So the situation is similar to the analysis using ratio estimator. When α_1 is small, this leads to a bias close to β_U / α_U (non-zero).

- Weak IV bias under 2SLS model with multiple IVs: likely the same problem: when estimating $\hat{\beta}$, both ΔY and $\Delta \tilde{X}$ have the term ΔU.
- Can we correct weak IV bias using cross-validated prediction? Probably not. As shown above, the problem is introduced by having an estimated α_1 that incorporates ΔU. Leave-one-out prediction model would change little of $\hat{\alpha}_1$.

Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. [Evans & Smith, Annual Review Of Genomics And Human Genetics, 2015]

- Concept of MR: the genotype (IV) randomizes subjects, i.e. all possible covariates are balanced out. Think of this as: subjects randomly assigned to genotype-defined groups, then assess outcome.
- **Assumptions of MR**: (Figure 1) let Z be genetic instrument, X exposure and Y outcome. The causal diagram:

$$Z \rightarrow X \rightarrow Y \qquad X \leftarrow U \rightarrow Y \tag{5.22}$$

  where U is a confounder. The three assumptions are: (1) Z must be associated with X, (2) Z not associated with U (lack of edge between Z and U), (3) Z not associated with Y except through X (lack of edge from Z to Y directly).

- Difficulty of MR: no pleiotropic effect of IV. Ex. robust association of genetic risk scores of CRP and cancer, however, pleiotropy cannot be ruled out.
- Two-sample MR: the estimated effect is $\hat{\beta}_{GY}$ divided by $\hat{\beta}_{GY}$, where the two estimated effects may come from two studies.
- Two-step MR (Figure 3) and mediation: motivation is to study if the effect of an exposure on outcome is mediated by methylation. Two step MR involves the use of IV for both exposure and intermediate variable.
- Studying trait pairs with MR: start with finding traits with genetic correlations. The simple approach would use genetic risk scores as IVs, but it almost always violates the MR assumption.

- Network MR: multiple SNPs and risk factors, learn the relation of variables, and use IVs to estimate causal effects.

Fulfilling the promise of Mendelian randomization [Pickrell, biorxiv, 2015]

- Skepticism of MR: Arguably no new causal relationship has been identified with this approach and subsequently verified in a RCT.
- Skepticism of MR: assumption that the genetic variants have a direct effect on one trait (the causal trait), but only an indirect effect on the other (the caused trait). This assumption is hard to validate a priori, and we have many examples of pleiotropy.
- Simulations to show that using multiple loci may lead to false positives. Model: $G \to U$, where U is a confounder, and $U \to X, U \to Y$. Even if only 5-20% of all loci of X influences U, MR will often find $X \to Y$.

Problems of MR: individual level data [personal notes]

- Testing MR assumption: (1) if U is observed, we can test if an IV is valid, by testing if it's associated with U. However, we need to take power into account: even if the result is negative (no association), it does not prove that the IV is valid. (2) We may not know if $U \to X$ or $X \to U$, if it's the latter, then this does not violate the assumption of MR (U lies the path from X to Y).
- Adjusting for confounders: when U is observed, we can adjust for U in regression models. When U is unobserved, we can model X_i and Y_i, and allow the error terms to be correlated.
- Residual (pleiotropic) effect of $G \to Y$. Suppose we have a single IV. It is easy to show that a model with $G \to Y$ effect is not identifiable.
- Weak IV bias. Q: can we remove weak IV bias, by modeling the correlated errors (hence confounders) between exposure and outcome?
- Multiple correlated IVs (LD).

Core concepts and limitations of Mendelian randomization [de Leeuw and Posthuma, review for NRG, 2020]

- Basics of MR: for variant j, let γ_{Xj} and γ_{Yj} be the estimated effect of j on X and on Y. (1) NOME assumption: IV effect on X is given and no measurement error. (2) Estimation: IVW, where weight of the estimated causal effect β_j is proportional to $1/s_j^2$, where s_j is the s.e. of γ_{Yj}. 2SLS: very similar to using PRS as IVs.
- Violations of MR:
  - Horizontal pleiotropy (InSIDE) Figure 1E.
  - A shared factor: Figure 1F, and Figure 2C. Note: Figure 2C is a special case of Figure 1F b/c X would almost always have other variants. Also Figure 1G and Figure 2D.
  - Population structure (Figure 1H).
  - LD (Figure 1I)
  - Reverse causality: Figure 2A and 2B. Note that with reverse causality model, we expect that X should have some other variants that do not act through Y. Overall, we would still have heterogeneity of causal estimates.
- Instances of exposures and outcomes are imperfect proxies for those directly involved in the causal effect (Box 3, Figure 2e-i). This could be due to measurement error, canalization (outcome pushed back to equilibrium), different tissue context of gene expression.

- Collider bias: two important case. Figure 2I: G and D (confounder) acts on R, conditioning on R induces correlation of G and Y even in the absence of X to Y effect. Figure 2J: X affects R, then conditioning on R removes causal effects.
- Testing implied constraints to validate MR assumption:
  - Reverse causality: if effect sizes are normalized, and either direction is true, then we would expect that the causal effect is within -1 to 1. To see this:

    $$Y = X\beta + \epsilon \tag{5.23}$$

    Then $\beta^2 \cdot \operatorname{Var} X \leq \operatorname{Var} Y$, so $\beta^2 \leq 1$. This allows one to determine direction. Methods for doing this: Steiger test, BayesMR.
  - Testing potential confounders: difficult in practice. If D (confounder) is observed, then we can adjust D when estimating γ_{Xj} and γ_{Yj}. A variant that acts on X and Y through D will have $\gamma_{Xj} = \gamma_{Yj} = 0$, so they will not be used as IVs.
  - Negative control population: exposure is constrained to a single value, so exposure cannot mediate genetic effect of a variant on the outcome. So if G_j is associated to Y, it must act independently of X (so invalid IV). Ex. to test alcohol consumption vs. mortality, use a population who do not drink alcohol. There are methods designed to use negative control population: Pleiotropy-robust MR (PRMR) method.
- Methods: testing heterogeneity, let $\hat{\beta}_j$ be estimated causal effect of variant j, test if they are all equal - Q statistic. MR-PRESSO. GLIDE: very similar. Then pruning: individual deviation for each variant can be used to determine if it should be pruned.
- Valid subset methods: assuming the majority of variants are instrument. Weighted median method. ZEMPA (zero modal pleiotropy assumption): the largest homogenenous subset of variants are valid instruments. Modal MR method: estimate the mode of a smoothed distribution of $\hat{\beta}_j$.
- Methods: modeling deviations. MR-Eggar, BayesMR. CAUSE.
- MR in practice: 76% studies evaluate heterogeneity to some extents, but few did this in testing, and only one study used a robust estimation method.

5.2.1 Mendelian Randomization: Methods for using Genetic Variants in Causal Estimation [Burgess & Thompson]

Instrumental Variable (IV) approach: [Chapter 1,2]

- IV approach: suppose we want to study the causal effect of exposure X on the outcome Y from observational data. Consider a variable that satisfies these conditions:
  - It is associated with X;
  - it is not associated with any confounder;
  - it does not affect the outcome, except possibly via its effect on exposure.

  Then it is an IV. Suppose we divide the data into groups by IVs, then we can compare the group with high X vs. the group with low X: if the outcomes are different, we show that X causally influence Y. The intuition is that *if the conditions are satisfied, then the groups defined by IV are random (wrt to Y), except the difference of X. So any difference in the outcome must be due to X.*
- Examples of IV: geographical locations are often used as IV. Ex. two adjacent states that are very similar in all aspects, and one of them implements a policy and we want to know if the policy affects outcome. Another example of IV: policy changes.

- Analogy to Randomized Clinical Trial (RCT): IV essentially generalizes RCT. In RCT, each subject is assigned randomly into one of two groups, treatment or control. The key of RCT is that the assignment of groups (exposure) is independent of the outcome. We can define G, the group where a subject belongs to, as an IV. Then RCT can be viewed as an IV approach: which group (genotype) an individual belongs to (receive) is random (wrt phenotype), determined by Mendelian segregation. Difference: in RCT, the effect of X on Y can be directly obtained, while in IV approach, it needs to be inferred, as the IV might have a small effect on the exposure (as in the case of genetic variants), which does not reflect the effect of exposure on outcome.

- Confounding and endogeneity: when there is a variable Z that is associated with X, and also may have an effect on Y, then Z is a confounder. If confouner is not accounted for, this leads to endogeneity: the regressor X will be correlated with the error term in regression (due to Z), and this leads to a biased estimate the influence of X on Y.

Mendelian randomization: basic concepts [Chapter 1,2]

- Genetic variants as IVs: most genetic variants are found to distribute randomly in the population. In other words, if we divide the population by alleles, then the groups do not different substantially in a way that could impact the outcome of interest (Note: of course, for causal alleles, the groups would differ - but the assignment of alleles to individuals are random).

- Violation of the assumptions: the conditions for genetic variants as IV would include random mating and lack of selection. This could be violated, e.g. if a variant is under selection, then the groups G and g will differ in some substantial way, e.g. g has lower fitness, then this violates IV assumptions (the variant could affect a trait through fitness). Another example, variant g is lethal in males (X-chromosome), but not in females, then the variant G/g will differ in gender composition.

  - Remark: if IV of X, say G, have broad impact (not through affecting X), then it is not a valid IV of X. On the other hand, it is OK if G is pleiotropic per se, as long as all these effects are mediated via X.

- Examples:

  - Study if CRP level has a causal influence on the risk of coronary heart disease (CHD): using cis-SNPs of CRP as IVs.
  - Study if alcohol influences the risk of cancer. Alcohol consumption is often correlated with smoking, so difficult to study. Use Mendelian randomization, the variant in the gene ALDH2 (alcohol metabolism).

Different views of IV approach [Chapter 3]

- Causal inference from manipulation: "no causation without manipulation". To distinguish conditional proability $Y|X = x$ with the statement that X has a causal effect on Y, use $Y|do(X = x)$, where $do()$ operators means manipulate the value of X.

- Counterfactual view: X has an effect on Y, means that suppose X takes a different value (from the observed one), will Y be different? The challenge is that this conterfacutal (alternative universe where X is different) is not observable - the "fundamental problem of causal inference". The strategy is to use *"exchangable" data that mimics the alternative universe*. For RCT, the treatment and control group are exchangable, so we can say, even though the control group does not receive treatment, we can use the information from the treatment group to see the alternative universe where the control group does receive treatment. Similar logic applies for IV.

- Probabilistic graphic model (DAG): causal model can be represented by DAG, where each edge has a causal interpretation, and no edge means conditional independence. The basic IV approach can be represnted as:

$$G \to X \to Y, \qquad X \leftarrow C \to Y \tag{5.24}$$

  where C is confounder. IV must satisfy the conditions: not associated with C, and not affect Y (except through X). G is not a valid IV if G affects Y through some intermediate variable, or there exists D s.t. $D \to G$ and $D \to Y$. G is still valid if there exists a collider E. In general, we summarize the condition of IV as: d-separation between G and Y, or intuitively: **if G is associated with some variable D, and $D \to Y$ (not through X), then G is not a valid IV** (because the groups defined by G may differ in D, which could influence Y).

Possible violations of IV in Mendelian randomization [Chapter 3]: basically the existence of D s.t. D associates with G, and $D \to Y$ independently of X.

- Biological mechanisms: pleiotropy. Ex. to test BMI $\to$ T2D, we can use SNP associated with BMI (e.g. FTO) as IV. However, if the FTO variant has broad effects, e.g. blood pressure, then this is invalid.
- Canalization: genetic variation may trigger compensatory changes.
- Non-genetic inheritance: LD (e.g. variants associated with the causal variants may affect other variables), effect modification (the effect of X on Y depends on some covariate, which is not accounted).
- Study design problems: the study design might introduce variables D that violate IV. Examples: population stratitification (ancestry is D) and ascertainment.

Validation of IVs in in Mendelian randomization [Chapter 3]

- Statistical validations: check the three conditions. Ex. test if an IV is associated with known confounders. The problem: IV may associate with a covariate that lies in the causal path from X to Y (mediator), and in this case, association with the covariate is OK.
- Biological considerations: Bradford Hill criteria (Table 3.1), some important ones:
  - Consistency: using multiple genetic variants (IVs) lead to the same estimate of causal effects. Or use non-genetic variables, e.g. a drug that manipulates the exposure.
  - Specificity: the IV should not change many different things.
  - Gradient: the genetic effect on the exposure and the genetic effect on the outcome are proportional.
- Example: suppose we are studying LDL $\to$ heart disease risk and we use a SNP associated with LDL as IV. To make sure it is a valid IV, we check if SNP is associated with known risk factors of heart disease, e.g. obesity. If it is the case, and the effect is not mediated via LDL, then the SNP is invalid IV. On the other hand, if the SNP is associated with a risk factor downstream of LDL, e.g. HDL, then it is OK.

Estimating causal effect: overview [Chapters 3 and 4]

- Testing causal effect: when G is a valid IV of X, then to test X on Y, we simply test if Y is associated with G.
- Assumptions of causal effect estimation: monotonicity, the effect of X on Y is monotonic.
- An example of analyzing causal effect estimation problem: exposure is smoking, outcome cancer rate. The unobserved confounder is diet: poor diet may correlate with smoking, and increase the rate of cancer.

Ratio of coefficient method for estimating causal effect [Chapter 4]

- Idea: the strength of causal effect can be written as:

$$\frac{\Delta Y}{\Delta X} = \frac{\Delta Y/\Delta G}{\Delta X/\Delta G} \tag{5.25}$$

- Binary G: the coefficients are then simply $\bar{Y}_1 - \bar{Y}_0$ and $\bar{X}_1 - \bar{X}_0$, and the ratio of the two (slope in the plot of $Y \sim X$) is the causal effect.
- Intuition why this works: in the smoking example, we cannot simply correlate cancer rate and smoking, because of the confounder. Now suppose we create groups by genotype G, then compare $\bar{Y}_1$ and $\bar{Y}_2$, then with each group, the mean of Y averages out the confounder (which on average is no different in the two groups).
- Continuous G: it is the ratio $\hat{\beta}_{Y|G}/\hat{\beta}_{X|G}$. Note that this ratio is derived from the estimated effect size: thus it alreadys controls for covariates, can use summary statistics (without individaul level data), and can be estimated from two different datasets.
- Confidence interval: could be derived from the ratio of normal RVs. Or use bootstrapping.
- Problem of the ratio of coefficient method: the mean of the ratio is not finite. It reflects the fact that the denominator has a certain probability of being 0.
- Issue with case-control studies: when the disease is rare, case enrichment will induce correlation between IV and confounders, thus estimation of the causal effect is biased.
- Power of the method: suppose IV explains 2% of the exposure, and we need 100 samples to study the correlation between X and Y, then we need $100/2\% = 20{,}000$ samples for the IV approach.

Two-stage and likelihood methods [Chapter 4]

- Intuition: generalization of the ratio method when G is binary. When G has multiple groups, intuitively, we should obtain $\bar{X}_k$ and $\bar{Y}_k$ for the k-th genotype, and regress the two. In the general case, we have one group per data point, so we will need to regression y_i against $\hat{x}_i$, where $\hat{x}_i$ is the mean of exposure under the given G_i (genotype). So this leads to the following method.
- Two-stage least square (2SLS) method: it can be used for multiple IVs. First, we regress x_i on G_i: $x_i = G_i\alpha + \epsilon_{X_i}$, then we have the predicted value $\hat{x}_i$, then we regress y_i on $\hat{x}_i$:

$$y_i = \beta_0 + \beta_1\hat{x}_i + \epsilon_{Y_i} \tag{5.26}$$

  where ϵ_{X_i} and ϵ_{Y_i} are indepenent error terms. Justification: suppose $y_i = \beta_0 + \beta_1 x_i + \epsilon_{Y_i}$, then we plug in X, we have: $y_i = \beta_0 + \beta_1(G_i\alpha) + \beta_1\epsilon_{X_i} + \epsilon_{Y_i}$, so the regression coefficient of Y against $G\alpha$ is still β_1.
- Problems of two-stage methods: (1) this model does not account for confounders. (2) The uncertainty of the first-stage regression has to be taken into acccount. While this does not affect the mean estimate, the standard error term needs to correct for this uncertainty.
- Full information maximum likelihood model (FIML): we simply have two regression models,

$$x_i = G_i\alpha + \epsilon_{X_i} \quad y_i = \beta_0 + \beta_1 x_i + \epsilon_{Y_i} \tag{5.27}$$

  The difference with the previous model is that ϵ_X and ϵ_Y are not independent, and the model will need to estimate the covariance matrix. The intuition that the full likelihood model works is that: the confounding variable is adjusted through the correlated error term.

- Limited information maximum likelihood (LIML): find β_1 that minimize RSS of $(y_i - \beta_1 x_i)$ vs. G_i. The intuition is that under the MR causal model, $Y \perp\!\!\!\perp G|X$, thus if we adjust X in Y, the residual and G_i should be uncorrelated.
- Bayesian method: the model is similar to FIML, except that y_i depends on the expected value of x_i not x_i:

$$x_i = G_i\alpha + \epsilon_{X_i} \quad y_i = \beta_0 + \beta_1 G_i \alpha + \epsilon_{Y_i} \tag{5.28}$$

  The joint distribution of X_i, Y_i is bivariate normal, with the error term correlated. Inference can be done via MCMC (possible in WinBUGS).

Statistical issues for IV analysis [Chapter 4]

- Covariates: if a covariate is not correlated with the IV and not in the causal path between exposure and outcome, then we should include covariates. This is called *exogenous regressor* in ecometrics. This will improve the precision of the estimate - more efficient. It is easy to check that not including covariates will not create a biased estimate of causal effect.
- Weak IVs: literature has used F-test as a measure of strength, $F < 10$ is considered weak. However, this is not recommended for several reasons. Ex. F-statistic depends on sample size.
- Overidentification test: when there are multiple IVs, we can test if some IV has effect not through the exposure. However, these tests generally have low power in detecting violation of IV assumptions.
- Endogeneity test: test if the observational and IV estimates are the same. Not recommended because of power (even if the result is not significant, it does not mean that the two estimates are the same).

Examples of MR analysis [Chapter 5]

- Fibrinogen and CHD: IV is a SNP in the promoter of Fibrinogen. However, the IV is also associated with ApoB/A1 with $p = 0.01$. It does not pass multiple testing threshold (ambiguous). Possible that ApoB/A1 acts between fibrinogen and CHD.
- BMI and blood pressure: use two SNPs of BMI (FTO and MC4R) as IVs.

Weak instruments and finite-sample bias [Chapter 7]

- Example of weak IV bias: partition the data into smaller subsets, and do MR, and use meta-analysis to combine the results from smaller studies [Table 7.1]. Very large bias in the estimated effects.
- Intuition of weak IV bias: in MR, G is supposed to capture difference of exposure, but may capture difference of confounders by chance (a finite-sample effect). Our estimator is $\beta = \Delta Y/\Delta X$, where Δ is the difference of X or Y in different genetic subgroups. If true value of ΔX is known, there should be no bias.
- Analysis of weak IV bias: our true model is:

$$X = \alpha_1 G + \alpha_2 U + \epsilon_X \qquad Y = \beta_1 X + \beta_2 U + \epsilon_Y \tag{5.29}$$

  where ϵ_X and ϵ_Y are independent. The ratio estimator is given by:

$$\hat{\beta}_1 = \frac{\Delta Y}{\Delta X} = \beta_1 + \frac{\beta_2 \Delta U}{\alpha_1 + \alpha_2 \Delta U} \tag{5.30}$$

  where Δ means the difference in genetic subgroups. When sample size is large, because G and U are independent, then $\Delta U \to 0$, so unbiased. With finite sample, ΔU may not be 0. When α_1 is large (strong IV), the bias is small. But when α_1 is small relative to $\alpha_2 \Delta U$, the bias is close to β_2/α_2. In summary, in the one-sample case, finite sample (variation of U) leads to difference of Y in genetic groups, and the direction is the same as the confounder effect. When true causal effect is 0, this leads to the increased FP rate.

- Analysis: the situation is opposite in the two-sample setting. In this case, the chance variation of ΔY and ΔU would be uncorrelated, so the bias is in the direction of null.

Possible strategy to overcome weak IV bias [Chapter 7]

- Choose large IVs using F-statistics (for IV). If use single IV, in general, if $F > 10$, generally OK. However, if one has multiple IVs (2SLS method), and choose the stronger IVs by F-test can lead to selection bias (winner's case).
- Multiple IVs: not always better than single IV, because some IVs may be weak and lead to bias.
- Adjusting for measured confounders improves estimate.

Using multiple IVs [Chapter 8]

- In general, using multiple IVs helps: reduce weak IV bias, and improves the power, which depends on how much variation of exposure is explained by IV.
- Allele scores: if use the weights from regression in the same data, this is the same as 2SLS. However, this suffers from Winner's curse. Better to use external data to obtain weights. If external data not available, could use cross-validation: calculate allele scores for an individual using weights estimated from independent samples.

Summary statistics in multiple studies [Chapter 9]

- Problem setting: summary statistics of $\hat{\beta}_X$ and $\hat{\beta}_Y$ for possibly multiple variants in multiple studies. In some studies, only $\hat{\beta}_X$ or $\hat{\beta}_Y$ are available.
- Multiple variant single study: for variant k, we have:

$$\begin{pmatrix} \hat{\beta}_{Xk} \\ \hat{\beta}_{Yk} \end{pmatrix} = N\left(\begin{pmatrix} \xi_k \\ \eta_k \end{pmatrix}, \begin{pmatrix} \sigma_{Xk}^2 & \rho\sigma_{Xk}\sigma_{Yk} \\ \rho\sigma_{Xk}\sigma_{Yk} & \sigma_{Yk}^2 \end{pmatrix}\right) \qquad \eta_k = \beta_1\xi_k \tag{5.31}$$

  where σ_{Xm} and σ_{Ym} are standard errors. The coefficient ρ captures the correlated effects of $\hat{\beta}_{Xk}$ and $\hat{\beta}_{Yk}$ in the same samples. To see this, we consider the estimated effects:

$$\hat{\beta}_X = (G^TG)^{-1}G^TX \qquad \hat{\beta}_Y = (G^TG)^{-1}G^TY \tag{5.32}$$

  When there is confounder U, then X and Y are correlated (we consider G as fixed/given), so $\hat{\beta}_{Xk}$ and $\hat{\beta}_{Yk}$ are correlated. The inference can be done via likelihood ratio and to test β_1 with LRT. The parameters ξ_k and η_k and ρ are nuisance parameters.

- Fixed effect model: To combine across studies, we assume the model above for each study, but ξ_k, η_k may differ for each variant with the same β_1. A study with only exposure and outcome can still add information by having the model:

$$\hat{\beta}_{Xm} = N(\xi_m, \sigma_{Xm}^2) \tag{5.33}$$

  for study m.

- Random effect models: instead of treating ξ_{km} and η_{km} as parameters, we assume a common normal distribution. Similarly we can use a normal prior for $\beta_{1m} \sim N(\beta_1, \tau^2)$.
- Special case: 2-sample MR, with one study of exposure and the other outcome. The same model can be used. When there is no sample overlap, we have $\rho = 0$.

5.2.2 MR, Mediation and Related Methods

Mediation analysis in Genetics [personal notes]

- Ref: GMAC paper [Yang and Lin, GR, 2016].
- Model: suppose we are testing if a trans-eQTL is mediated by a cis-gene, we have SNP G, X and Y with the model:
$$G \to X \to Y \qquad G \to Y \tag{5.34}$$
where $G \to Y$ describes the unmediated effect (horizontal pleiotropy from MR perspective).
- Adjusting for possible confounders: suppose we have U (observed) associated with X and Y, there are two scenarios:
$$X \leftarrow U \to Y \qquad X \to U \leftarrow Y \tag{5.35}$$
In the first case, U is a confounder, and should be adjusted. But in the second case, U is a collider, and should not be adjusted. To distinguish the two, note that when U is a confounder, L_i and U should be uncorrelated. When U is a collider, L_i should affect U.
- The impact of LD: suppose L_i is not a causal variant, is adjusting L_i itself sufficient? No. Let L_j be the true causal variant, which may have unmediated effects. Our causal digram:
$$L_i \to X \to Y \qquad L_i \leftarrow V \to L_j \qquad L_j \to \{X, Y\} \tag{5.36}$$
where V is a hidden variable for LD. If there is no $L_j \to Y$ effect, then adjusting L_i would block all backdoor path from X to Y. However, when L_j has unmediated effect, adjusting L_i is not enough, since there is a backdoor path: $X \leftarrow L_j \to Y$
- Does significant mediation imply causality? No. See the notes in `Statistics.pdf`, "Does Mediation imply causality?" Significant mediation could happen if:
$$G \to X, \qquad X \leftarrow U \to Y, \qquad G \to Y \tag{5.37}$$
Biologically, this may happen when G is eQTL of a gene X, and is also in LD with a causal variant of Y.
- Comparison with MR: mediation analysis allows G to influence Y through other paths (with effect size δ). Under the MR assumption, without this pleiotropic effect, association of G with Y implies $X \to Y$.

An integrative genomics approach to infer causal associations between gene expression and disease (LCMS) [Schadt & Lusis, NG, 2005]:

- Motivation: QTL studies often map to regions with multiple genes. Even if a single gene can be mapped, additional evidence is needed to show that the gene is causal to the trait. The idea is to use expression data to bridge the gap:
  - Intuitively, if the expression of a gene in QTL is also correlated with the trait, then this gene is likely to be involved in the trait.
  - If the expression of a gene (somewhat related to this process) is linked to the QTL, then this QTL is functional (as opposed to other QTLs that may be be false positives).

  The issue is to decide the causality.

- Model: let L be QTL (genotype), R be transcript, and C be clinical trait, both eQTL and cQTL are mapped to the same loci, thus we know that change of L leads to change of R and C, the problem is to distinguish three models:

$$M_1 : L \to R \to C \qquad M_2 : L \to C \to R \qquad M_3 : L \to C, L \to R, C \leftrightarrow R \tag{5.38}$$

  The three models are denoted as: causal, reactive and independent model, respectively. The likelihood based causality model selection (LCMS) chooses the model by computing the likelihood of the three models:

$$\begin{array}{ll} P(D|M_1) = P(L)P(R|L)P(C|R) & P(D|M_2) = P(L)P(R|L)P(C|R) \\ P(D|M_3) = P(L)P(C|L)P(R|L,C) & \end{array} \tag{5.39}$$

  Since the three models have different complexities (one more parameter in M_3), use AIC to choose a model.

- An alternative test: conditional independence test. Under M_1, $L \perp\!\!\!\perp C|R$, so the corresponding partial correlation coefficient (PCC) should be 0; similarly under M_2, $L \perp\!\!\!\perp R|C$. Test the two PCCs to determine which test should be accepted: if one PCC is 0, accept M_1 or M_2; if neither, accept M_3. The problem is: if only one PCC is used, the test is incomplete, e.g. suppose we test PCC between L and C given R, but when M_2 is true, the test loses power; if both tests are used, it is not clear how they are combined.

- Intuition of LCMS: suppose M_1 is true, and suppose L is strongly correlated with R, but R correlation with C is weak. Then M_2 is a poor fit, because the correlation of both L and C and the correlation of C and R are both weak. For M_3, it is more complex, but it does not add much explanation beyond M_1. In general, suppose we have three variables with $X \to Y \to Z$, with the linear model between the two edges (coefficient β and γ), then

$$\mathrm{Var}\,(Z|x) = \gamma^2\sigma_Y^2 + \sigma_Z^2 \tag{5.40}$$

  which is larger than both σ_Y^2 and σ_Z^2. So a model with the edge $X \to Z$ would not be a good fit (large variance). In general, the likelihood method will try to put the variables with strong correlation into direct edges.

- Data: 111 F_2 mouse from crossing two strains. Each mouse is: (1) genotyped for 139 microsatellite markers; (2) the liver expression data is profiled; (3) is phenotyped for the disease trait (obesity).

- Multi-step procedure: (Figure S3)
  - Step 1: define the QTL of obesity. Found four QTL.
  - Step 2: find a list of diff. expressed genes (correlated with phenotype). 440 genes were identified.
  - Step 3: test if the four QTL are also eQTL of these 440 genes. There were 113 genes which overlap with at least two QTL (a total of 267 eQTL-gene pairs). The FDR of having two overlap is 0.4% (compared with 15% when requiring only one shared QTL), by permutation of genotypes.
  - Step 4: LCMS test of 113 genes, the ratio of three models are 50% (causal), 9% (reactive) and 41% respectively.

- Remark:
  - The main contribution to success comes from QTL overlap: with the requirement of having two or more QTL, the gene list is limited to 113 genes, and half of them were tested causal.
  - Correlation between genes is a major problem: LCMS test cannot distinguish them. That's why 4 QTL of obesity supports the finding of more than 50 genes. A related issue is that the model does not distinguish cis- and trans-eQTL.

– Multi-step procedure: information in many weaker eQTL and cQTL are not used.

Bias in causal estimates from Mendelian randomization studies with weak instruments [Statistics in Medicine, 2011]

- Weak IV problem from a causal model perspective: we know that correlation of G and U leads to a backdoor path: $G \leftrightarrow U \rightarrow Y$. Even when G and U has no confounder, stochastic/chance correlation can still lead to the backdoor path, and association between G and Y.
- Simulation to investigate weak IV bias: setting, no causal effect of X to Y. Error of U, X, Y: variance 1. Effects of U on X and $Y = 1$. Vary α_1, the IV strength, from 0.05 to 0.55 (Figure 1, Table 1). At $\alpha_1 = 0.05$, or F-statistics 1.07 (slightly above the null expectation), often non-zero β_1 (causal effect). But at F-statistics at 8, relatively free of bias.
- Multiple IVs: reduce the variance of estimator, but increase the bias.
- Discussion: IVs should be ascertained before doing MR (estimate effect sizes to both X and Y).

Mendelian randomization analysis with multiple genetic variants using summarized data [Burgess & Thompson, GE, 2013]

- Problem: given summarize statistics of variants X_k and Y_k for variant k, how do we estimate the effect of X on Y, assuming the variants are valid IVs?
- Likelihood model (Eq. 3): similar bivariate model as Sherlock2. Let ξ_k be the true effect of k on trait X, then the expected effect on trait Y is $\beta\xi_k$. The difference is that: not allow residual effect of k on trait Y; no prior on ξ_k (treated as fixed).
- Comparison of summary statistics vs. individual level data: similar performance when assumptions hold. The summary method overstates precision with LD. Weak instrument bias.
- Remark: we could investigate weak instrument bias. Our model is $X = \alpha_1 G + \alpha_2 U$ and $Y = \beta_1 X + \beta_2 U$. We can derive $\hat{\beta}_{IV}$ and show that its expectation is not the same as β_1.

Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. [Burgess and Thompson, AJE, 2015]

- Motivation: often hard to find valid IVs because they have pleiotropic effects on multiple risk factors. Ex. LDL, HDL and TG.
- Different scenarios: Figure 2, vertical and functional pleiotropy. Also there could be causal effects among risk factors (Figure 3).
- Model: Two stage LS with individual level data. First do multiple regression to build predictive model of multiple risk factors using all putative IVs. Then regression of trait vs. predicted risk factors. Note: this can be done when the risk factor genetic data and trait genetic data are collected in different samples.
- Likelihood method when we only have summary statistics: Equation (1), we model the summary statistics of SNP to each risk factor, and SNP to trait. The joint distribution is MVN, where correlation among different risk factors and traits capture sample overlap. To incorporating LD, summary statistics of related SNPs are correlated.
- Results on HDL, LDL and TG on CHD risk: in joint analysis with likelihood model, only LDL and TG show non-zero effects.
- The impact of causal effects among risk factors: MVMR estimates direct effects, misleading results on total effects.

Bias due to participant overlap in two-sample Mendelian randomization [Burgess and Thompson, GEpi, 2016]

- Weak IV bias under 2SLS: consider two stage model:

$$X = G\alpha_1 + \epsilon_X \qquad Y = \tilde{X}\beta_1 + \epsilon_Y \tag{5.41}$$

  where $\tilde{X} = G\hat{\alpha}_1$ is the least square estimator of X. When G and U are correlated in finite samples, ϵ_X is now correlated with U and ϵ_Y is also correlated with U, leading to correlation of ϵ_X and ϵ_Y. Let their covariance be σ_{XY}. Also define μ be the concentration parameter, a measure of IV strength and K the number of IVs used. The expected bias is then:

$$\text{Bias of 2SLS} = \frac{\sigma_{XY}(K-2)}{\sigma_X^2 \mu} = \frac{\sigma_{XY}}{\sigma_X^2 \mathrm{E}\,(F)} \tag{5.42}$$

  where σ_X^2 is the variance of X and $\mathrm{E}\,(F)$ is the expected F-statistic of the 2SLS estimator of X.

- Analysis: weak IV bias is a form of overfitting. It is clear that the problem is the prediction of X is correlated with the error of X (which captures U). In this case, over-fitting is due to weak IV. In BLUP prediction [GBAT paper, Xuanyao Liu], over-fitting is due to the BLUP model.

- Weak IV bias under 2SLS in two-sample setting: in the 2-sample setting, σ_{XY} should generally be close to 0, so the bias is close to 0.

- Question: does 2SLS reduce the bias? The claim that multiple IVs increase the bias [Burgess and Thompson, Stats in Med, 2011].

Orienting the causal relationship between imprecisely measured traits using GWAS summary data (MR-Steiger) [Hemani, PLG, 2017]

- Comparison of Mediation and MR: (1) Mediation: fit $Y \sim G + X$, where X is observed. (2) MR: the causal diagram can be understood as: $G \to \tilde{X} \to X \to Y$, where $\tilde{X}$ is the genetic component of X due to G. We fit $Y \sim \tilde{X}$. So the differences are:
  - Mediation is affected by confounding and measurement error, since it models X.
  - Mediation based Causal Inference Test (CIT): test both directions.

- CIT: a set of four conditions to conclude the direction is $X \to Y$. The conditions reject $G \to Y \to X$ and $X \leftarrow G \to Y$. It establishes $G \to X \to Y$ by CI of G and Y given X.
  Problem of CIT: if there is a confounder U affecting both X and Y, then G and Y are dependent when conditioned on X, because X is a collider.

- MR causal test: two tests, first MR test to assess causal relationship. Next, Steiger test: compares correlation of G and X, ρ_{gx} vs. correlation of G and Y ρ_{gy}. The idea is that if $G \to X \to Y$, then G and X would be more correlated.

- Mediation-based CIT is sensitive to measurement error: let X_O be the observed X, our model is:

$$G \to X \to Y \qquad X \to X_O \tag{5.43}$$

  It's clear now that X_O no longer blocks the path from G to Y, thus CI does not hold.

- Remark: Steiger filtering refers to: filtering variants by correlations.

- **Lesson**: CIT that relies on CI is problematic in practice because of confounder and measurement error.

A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. [Bowden and Thompson, Stat Med. 2017]

- Notation: standard MR, let β be the causal effect, γ_j be the IV effect size on exposure, and Γ_j be the IV effect on outcome. Let α_j be the direct effect on outcome, and Ψ_j be the effect on confounder U. Also let κ_X and κ_Y be the effect of U on exposure and outcome, respectively.
- Model: assume γ_j is known (measured). The estimated effect on outcome:

$$\hat{\Gamma}_j = \alpha_j + \beta\gamma_j + \epsilon_j \qquad \epsilon_j \sim N(0, \sigma^2_{Yj}) \tag{5.44}$$

  For simplicity, assume $\alpha_j = 0$. Let $\hat{\beta}_j = \hat{\Gamma}_j/\gamma_j$ be the estimated causal effect from SNP j, the IVW estimator is weighted average of $\hat{\beta}_j$, where weight is given by:

$$w_j = \frac{1}{\text{Var}\,(\hat{\beta}_j)} = \frac{\gamma_j^2}{\sigma^2_{Yj}} \tag{5.45}$$

  So the weight depends on both IV strength and the s.e. of the effect estimate on outcome. Putting this together, we have the IVW estimator:

$$\hat{\beta}_{\text{IVW}} = \frac{\sum_j \gamma_j \Gamma_j \sigma^{-2}_{Y,j}}{\sum_j \gamma_j^2 \sigma^{-2}_{Y,j}} \tag{5.46}$$

  Note: see Equation (1) of [Burgess and Thompson, Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data, GE, 2013].

- Random effect IVW: if we assume $\alpha_j \sim N(0, \sigma^2_\alpha)$, and use standard random effect meta-analysis, this is the additive random effect IVW. However, this is not good in MR setting: as this model generally put more weights on the outliers. Recommend to use multiplicative random effect model: where the the model can be written as:

$$\hat{\Gamma}_j = \alpha_j + \beta\gamma_j + \phi^{1/2}\epsilon_j \qquad \epsilon_j \sim N(0, \sigma^2_{Yj}) \tag{5.47}$$

  where ϕ is estimated from the heterogeneity.

Distinguishing genetic correlation from causation across 52 diseases and complex traits (LCV) [OConnor & Price, NG, 2018]

- Model: see Figure 1. Let q_1 and q_2 be the effects of L (latent factor) on Y_1 and Y_2 respectively (q_k are scalars). Let π be the SNP to L effect and γ_1, γ_2 be non-mediate effect of SNP to Y_1, Y_2 (these are distributions). Effect sizes are normalized: "α and π (but not γ) are normalized to have unit variance, and all random variables have zero mean".

$$\text{Var}\,(\pi) = 1 \qquad \text{Var}\,(\pi q_k + \gamma_k) = 1 \Rightarrow q_k^2 + \text{Var}\,(\gamma_k) = 1, \text{ where } k = 1, 2 \tag{5.48}$$

  The interpretation of q_k^2 is then the proportion of heritability of trait 1 (or 2) that is mediated by the latent factor.

- Assumptions of LCV: needs several independence assumptions, but weaker than independence of $(\pi, \gamma_1, \gamma_2)$. Most notably, it requires: SNPs with a large mediated effect do not tend to also have an additional non-mediated effect.
- Genetic causality proportion (GCP): when $q_1 = 1$, $\gamma_1 = 0$, this means trait 1 is causal to trait 2. Similarly, when $q_2 = 1$, we have trait 2 is causal to trait 1. In general, $q_1 > q_2$ means that trait 1 is "partially causal" to trait 2, through the latent factor. To define this, use gcp, defined as:

$$\text{GCP} = \frac{\log|q_2| - \log|q_1|}{\log|q_2| + \log|q_1|} \tag{5.49}$$

  Equivalent definition is: $q_2^2/q_1^2 = (\rho_g^2)^x$.

- Auxiliary test: estimate proportion of heritability explained by the correlated component. It performs poorly when two traits have unequal power and unequal polygeneicity, as expected (Table S2): highly inflated type I error.

- Inference with MOM estimator: Use this equation (derivation in Methods):

$$\mathrm{E}\left(\alpha_k^2\alpha_1\alpha_2\right) = \kappa_\pi q_k^2 q_1 q_2 + 3\rho_g, k = 1, 2 \tag{5.50}$$

  where $\kappa_\pi = \mathrm{E}\left(\pi^4\right) - 3$ and ρ_g is genetic correlation. Intuitively, if $q_1^2 > q_2^2$, we should have $\mathrm{E}\left(\alpha_1^3\alpha_2\right) > \mathrm{E}\left(\alpha_1\alpha_2^3\right)$, this allows us to infer the difference of q_1 and q_2, hence GCP. Specifically, first estimate ρ_g, and estimated mixed fourth moments accounting for errors in $\hat{\alpha_1}$ and $\hat{\alpha_2}$. We define, for $k = 1, 2$:

$$\hat{\kappa_k} = \mathrm{E}\left(\alpha_k^2\alpha_1\alpha_2\right) - 3\rho_g \tag{5.51}$$

  If there is no population structure or sample overlap, this term can be easily estimated, just observed moments subtracting 3 ρ_g. In real data, we use this result (Equation 6 of paper):

$$\mathrm{E}\left(\hat{\alpha}_1\hat{\alpha}_2^3|\alpha_1, \alpha_2\right) = \alpha_1\alpha_2^3 + 3\alpha_1\alpha_2\sigma_2^2 + \alpha_2^2\sigma_{12} + 3\sigma_{12}\sigma_2^2 \tag{5.52}$$

  where σ_2 is the intercept of LDSC and σ_{12} the intercept of cross-trait LDSC. This allows us to estimate $\mathrm{E}\left(\alpha_1\alpha_2^3\right)$ by computing expectation of $\hat{\alpha}_1\hat{\alpha}_2^3$ subtracting the last three tersm of the RHS of this equation. Intuitively, the difference of $\hat{\kappa_1}$ and $\hat{\kappa_2}$ should reflect the value of GCP. We could obtain a MOM estimator of GCP x (not what the paper does). From Equation (2) of the paper, we have:

$$\mathrm{E}\left(\hat{\kappa_1}\right) = \kappa_\pi q_1^3 q_2 \qquad \mathrm{E}\left(\hat{\kappa_2}\right) = \kappa_\pi q_1 q_2^3 \tag{5.53}$$

  We notice that $q_2^2/q_1^2 = (\rho_g^2)^x$, so we have:

$$\frac{\hat{\kappa_1} - \hat{\kappa_2}}{\sqrt{\hat{\kappa_1}^2 + \hat{\kappa_2}^2}} = \frac{q_1^2 - q_2^2}{\sqrt{q_1^4 + q_2^4}} = \frac{1 - \rho_g^{2x}}{\sqrt{1 + \rho_g^{4x}}} \tag{5.54}$$

  This allows one to estimate x from test statistics, $\hat{\kappa_1}$ and $\hat{\kappa_2}$, and ρ_g. Note that the LHS of this equation is very similar to the test statistic $S(x)$ in the paper (Equation 7).

- Dealing with LD: in the estimation of moments above, e.g. $\hat{\alpha}_1\hat{\alpha}_2^3$ over all SNPs, we weigh the SNPs by $\max\{1, 1/l_i\}$ where l_i is the LD score of SNP i. This weighting scheme will reduce the weights of SNPs with high LD (otherwise, they would be over-counted).

- Inference with likelihood: the paper uses a different estimator. It normalizes $\hat{\kappa_1} - \hat{\kappa_2}$ using x s.t. the resulting distribution is approximately standard normal when x is the true value of GCP. However, I am not able to show this. Suppose it is true, let $S(x)/\sigma_{S(x)}$ be the "normalized" test statistic, then $S(x)/\sigma_{S(x)} \sim N(0, 1)$ when x is the true value. Since x is unknown, we compute $S(x)/\sigma_{S(x)}$ for each value of x, and we should have: the true value of x gives standard normal (high degree t-distribution in the paper), but not other values of x. This leads to the likelihood-based estimator of the paper.

- Remark: to understand the estimation procedure, consider the simple problem of estimating μ in a normal distribution $N(\mu, \sigma^2)$. We define $T(\bar{x}, \mu) = (\bar{x} - \mu)/\sigma$, then it should follow $N(0, 1)$ under the true value of μ, assuming σ is known. To estimate μ, we compute $T(\bar{x}, \mu)$ for different values of μ, and the true value of μ is obtained by maximizing the likelihood: $P(T(\bar{x}, \mu)|N(0, 1))$ under different values of μ.

- Accounting for LD: in the estimator, a SNP is weighted by the inverse of its LD score.

- Methods compared: (1) Two-sample MR: ascertain SNPs significant for exposure, then estimate correlation coefficient with intercept 0. (2) MR-Egger: similar, but fitted intercept. (3) Bidirectional MR: ascertain SNPs for both exposure and outcome, then for each estimate r_1 or r_2, and test the difference.

- Analysis of MR models: Two-sample MR captures any genetic correlation between two traits, even if gcp = 0. Bidirectional MR: simply consider the difference of r_1 and r_2, it does not account for different genetic architecture and power of the two traits.

- Simulation procedure: Section 1.4 in Supplement. Assume q_1 and q_2 are given. Step 1. Simulate π and γ_k ($k = 1, 2$). Sample π from spike-and-slab distribution with mean 0 and variance 1. Specifically:

$$\pi \sim p_\pi N(0, 1/p_\pi) + (1 - p_\pi)\delta_0 \tag{5.55}$$

  where p_π is a specified parameter. It's easy to check that variance of π is 1 using the results about variance of mixture distribution [Wiki, mixture distribution]. Sampling of γ_k is similar: use spike-and-slab distribution with mean 0 and variance $1 - q_k^2$, specifically:

$$\gamma_k \sim p_{\gamma_k} N(0, (1 - q_k^2)/p_{\gamma_k}) + (1 - p_{\gamma_k})\delta_0 \tag{5.56}$$

  Step 2. Simulate effect sizes in the usual scale. Let β_k be the effect size of trait 1 and 2. Per SNP heritability, or the variance of β_k (assuming standardized genotypes), should be h_k^2/M, where h_k^2 is hertiability and M number of SNPs. So should scale α_k from the previous step as:

$$\beta_k = \frac{h_k}{\sqrt{M}}(q_k\pi + \gamma_k) \tag{5.57}$$

  Note: the equation of effect size scaling in this step is wrong, but the code is correct. Step 3. Simulate summary statistics without LD: $\hat{\beta}_k \sim N(\beta_k, s_k^2)$, where s_k is the standard error of the estimator, which depends only on sample size for a given AF of SNP. Note: under LCV, variance of effect sizes of shared SNPs, q_1^2/p_π, and SNPs acting only on trait 1, $(1 - q_1^2)/p_{\gamma_1}$, can be different.

- Connection with CAUSE: when h_k^2 and M are given, LCV has five free parameters, $p_\pi, p_{\gamma_1}, p_{\gamma_2}, q_1, q_2$, while CAUSE has four: proportion of causal variants p_M, p_Y, and sharing parameters q, η. Note that the effect size variance of causal variants σ_M^2 and σ_Y^2 (for simplicity, we use spike-and-slab instead of ASH), are determined once p_M and p_Y are given:

$$p_M\sigma_M^2 = h_M^2/M \qquad p_Y\sigma_Y^2 = h_Y^2/M \tag{5.58}$$

  The difference of the number of parameters is because in LCV, the shared variants and variants acting only on trait 1 could have different effect size variance. Now suppose we have LCV parameters, p_π, p_{γ_k}, q_1, q_2 and we will determine CAUSE parameters. First, the proportion of causal variants for trait M and Y:

$$p_M = p_\pi + p_{\gamma_1} \qquad p_Y = p_\pi + p_{\gamma_2} \tag{5.59}$$

  The sharing parameters of CAUSE: q is the proportion of shared variants among all variants acting on M, and for η, it is the ratio of effect size of shared SNPs on trait M and Y, respectively

$$q = \frac{p_\pi}{p_\pi + p_{\gamma_1}} \qquad \eta = \frac{q_2\sqrt{h_2^2/M}}{q_1\sqrt{h_2^2/M}} = \frac{q_2 h_2}{q_1 h_1} \tag{5.60}$$

  Now suppose we know CAUSE parameters and we will determine LCV parameters. The proportion of causal variants:

$$p_\pi = p_M \cdot q \qquad p_{\gamma_1} = p_M(1 - q) \qquad p_{\gamma_2} = p_Y - p_M q \tag{5.61}$$

  Next we consider effect size variance of shared variants on trait $1/M$: it should be σ_M^2 under CAUSE, and $1/p_\pi$ (variance on L) times q_1^2 (variance on trait 1) times h_1^2/M (scaling parameter). So we have:

$$q_1^2 \cdot \frac{1}{p_\pi} \cdot \frac{h_1^2}{M} = \sigma_M^2 \tag{5.62}$$

We use the equation $p_M\sigma_M^2 = h_M^2/M$, and $p_M q = p_\pi$ to obtain:

$$q_1 = \sqrt{q} \qquad q_2 = \eta q_1 h_1/h_2 = \eta\sqrt{q}h_M/h_Y \tag{5.63}$$

It's easy to check that with this transformation, the effect size variance of shared SNPs and SNPs acting only on trait 1 are the same:

$$\frac{q_1^2}{p_\pi} = \frac{q}{p_M q} = \frac{1}{p_M} \qquad \frac{1-q_1^2}{p_{\gamma_1}} = \frac{1-q}{p_M(1-q)} = \frac{1}{p_M} \tag{5.64}$$

- Null simulation with uncorrelated effects: 50,000 independent SNPs. Some SNPs may be shared by two traits, but genetic effects are uncorrelated. LCV and all other methods control type I error well (Figure 2a).
- Null simulation with correlated effects: certain percent of SNPs are shared between two traits, and $q_1 = q_2$ (but non-zero). Also vary power and polygenecity of the two traits. All MR methods fail in this case, but GCP performs well (Figure 2bc).
- Causal simulation: Choose $N_1 = N_2 = 25,000$, 5% SNPs are causal to trait 1, $q_1 = 1, q_2 = 0.2$. The power of LCV is much higher than other MR methods (Figure 3).
- Application to real data: 52 traits. 32% of trait pairs show nominal significance $P < 0.05$, and 59 pairs at FDR < 0.01, including 30 with GCP > 0.6 (Table 1). Some of the 30 pairs include: lipid and TG on MI, LDL on bone marrow density.
  - Some pairs in Table 1 are likely not causal: Triglycerides and Platelet distribution width. High cholesterol and Red blood cell count. Triglycerides and Reticulocyte count, Eosinophil count. Balding and number of children. MI and breast cancer.
  - Autism and education attainment: gcp = 0.13.
  - MR: MR reported significant causal relationships (1% FDR) for 271 of 429 trait pairs, including 155 reciprocal pairs of traits.
- Analysis: problems of GCP. GCP is only vaguely connected to causality. It tests asymmetry $q_1 > q_2$. It's possible that GCP is large even when q_1 is significantly smaller than 1. Ex. $q_1 = 0.5, q_2 = 0.05$, gcp = 0.62. Also, when q_2 is very small (regardless of q_1), conceptually L has no effect on Y_2, but GCP will converge to 1, e.g. $q_1 = 0.5, q_2 = 0.00001$, GCP = 0.88.
- Analysis: why the LCV estimator is robust to unequal polygeneicity (Table S2)? Simulation setting: 1% of variants affect L, 2% affect Y_1 only and 8% affect Y_2 only, with equal effect size distributions. If we interpret q_1 and q_2 as percent heritability explained by L (the auxilary test), we would find GCP > 0. However, the LCV estimators works well as long as $q_1 = q_2$, which are effect sizes of L on Y_1 and Y_2. To see this, let π_L, π_1, π_2 be the percent variants affecting L, Y_1 only and Y_2 only. Then:

$$\mathrm{E}\left(\alpha_1^3\alpha_2\right) = \pi_L \mathrm{E}\left(\alpha_1^3\alpha_2|L\right) + \pi_1 \mathrm{E}\left(\alpha_1^3\alpha_2|Y_1\right) + \pi_2 \mathrm{E}\left(\alpha_1^3\alpha_2|Y_2\right) \tag{5.65}$$

  Note that the last two terms are 0 because $\alpha_1^3\alpha_2 \neq 0$ if and only both α_1 and α_2 are non-zero. The same is true for $\mathrm{E}\left(\alpha_1\alpha_2^3\right)$. This explains that only variants acting on L contribute directly to the test statistic.
- Remark: other problems of LCV
  - LCV does not distinguish exposure and outcome, instead the two traits are "interchangable labels".
  - LCV does not infer specific SNPs contributing to the partial causality.

Causal associations between risk factors and common diseases inferred from GWAS summary data [Zhu and Yang, NC, 2018]

- Generalized SMR (GSMR) model: multiple independent SNPs, we expect similar estimate of b_{xy}. So the method does Generalized Least Square (GLS), regression of $\hat{b}_{zy}$ vs. $\hat{b}_{zx}$, accounting for different standard errors and LD between SNPs. This is equivalent to a weighted mean of $\hat{b}_{xy(i)}$ over all SNPs.
- HEIDI-outlier test: choose the SNP with the strongest effect one exposure among all SNPs based on $\hat{b}_{xy}$ - choose the ones in the third quantile. Then test if another SNP has a different $\hat{b}_{xy(i)}$, and remove outlier SNPs with $p < 0.01$.
- Application: 7 risk factors and a large number of phenotypes in UKBB and another dataset. Confirm: (1) BMI on T2D, CVD and hypternsion (2) LDL to CAD and dyslipidemia. (3) Blood pressure and hypertensive diseases and CVD. Note that: LDL vs. LOAD is not significant after removing the outlier using HEIDI.
- Adjusting for covariates (other risk factors) with GWAS summary statistics. Not change results much.
- Remark: the step of removing pleiotropic SNPs with HEIDI is arbitrary. A non-causal risk factor may share a fraction of SNPs with outcome. This may inflate the causality claim.

Bayesian variable selection with a pleiotropic loss function in Mendelian randomization (JAM-MR) [Gkatzionis and Newcombe, review for PLG, 2019]

- Background: JAM for fine-mapping. G-prior for β in linear regression. Inference of γ by MCMC.
- Background: extension of Bayesian inference without likelihood. Let θ be parameter of interest, we infer θ not via likelihood but with loss function:

$$P_l(\theta|D) \propto \pi(\theta) \exp\left(-wl(D,\theta)\right) \tag{5.66}$$

  where $\pi(\theta)$ is the prior, $l(D,\theta)$ is the loss function and w a weight parameter.

- JAM-MR model: our idea is to do variant (IV) selection on data of X (mediator). We penalize the variants with different effects (outliers). Introduce a loss function in JAM that favors variants with similar θ_j, where θ_j is the ratio estimator from SNP i. The loss function:

$$l_1(\hat{\theta}, \gamma) = \frac{1}{P_\gamma - 1} \sum_{j:\gamma_j=1} (\hat{\theta}_j - \hat{\theta}_\gamma)^2 \tag{5.67}$$

  where $\hat{\theta}_\gamma$ is the mean of univariate causal effect estimate in model γ. The final posterior combines: prior of γ (Beta), summary statistics based likelihood of X, and loss function.

- Causal effect estimation: (1) For each model γ, use meta-analysis, inverse variance weighting, to combine $\hat{\theta}_j$ for all $j \in \gamma$. (2) Combined estimate of all γ's: weighted average of all models, where the weights are determined by the posterior $p(\gamma|D)$.
- Analysis: how variable selection for X and consistency of causal effect estimates are combined?
  - When penalty w in the loss function is low: the model chooses most SNPs, including some "invalid" SNPs that have pleiotropic effects.
  - When penalty w increases: the model chooses good SNPs, but also remove SNPs with pleiotropic effects. The optimal range is w about 0.5-5 of sample size N_1.
  - When penalty w is large: even valid SNPs are not chosen by the model, leading to large standard errors of estimated effects.

When the number of pleiotropic SNPs with consistent effects gets large, it's possible that the method chooses these SNPs, rather than valid SNPs. These models will have high posterior probabilities, along with true models.

- Simulation setting: Equations (1) - (3), $\beta_{X,j}$ are IV effects sizes, always positive. δ_j is the direct effects (horizontal pleiotropy), and α_j's are effects on confounder. Use 50 IVs of trait 1, and 15 of them are pleiotropic. Simulate two values of θ (causal effect) 0 and 0.5. Four settings (Table 1) for pleiotropic variants: (1) Balanced pleiotropy: δ_js are symmetric. (2)-(4) Directional: δ_j are positive/constant or α_j non-zero.
- Note: in directional settings of simulation, δ_j are not proportional to $\beta_{X,j}$'s. Similarly, α_j are random across variants. So the simulations are easier than correlated pleiotropic effects.
- Results of scenario 1: all methods perform well except MR-Egger. MR-Egger has large standard errors, so it has high MSE, but low type I error. However, JAM-MR has considerable type I errors (12-22%).
- Results of scenario 2-4: IVW, MR-Egger have large standard error and MSE. Weighted median and mode estimator, MR-pressor generally perform well. JAM-MR seems to have lowest errors.
- Application in real data: two blood pressure and CHD. Causal effects by most methods, except mode estimator.

An integrative analysis of GWAS and intermediate molecular trait data reveals common molecular mechanisms supporting genetic similarity between seemingly unrelated complex traits (Sherlock-II) [Gu and Hao Li, 2019]

- Scoring a gene: ascertain eQTL with $p < 10^{-5}$, then the score defined as: $s = -\sum_i \log_{10}(p_i)$ where p_i is GWAS p-value, $1 \leq i \leq n$. LD pruning: $r^2 < 0.2$, and min. of 100kb between two SNPs. Note: truncate GWAS p-value at 10^{-9} makes the results less dominant by single SNPs.
- Computation null distribution: (1) Single SNP case: basically, the percent of SNPs with p-values falling into an interval. (2) General case: sum of independent RVs, so we can analytically obtain the null distribution.
- Correcting for pleiotropic SNPs: SNPs with pleiotropic effects (genes) are more likely to have small p-values in GWASs. To adjust for this, for any SNP (single-SNP case), its null distribution is obtained by getting the percent of SNPs with p-values in a range, only those SNPs matching number of genes. In the equation, C_i is number of genes SNP i is assigned to.
- Dealing with multiple tissues in eQTL: choose the tissue with the strongest eQTL.
- Results of gene-phenotype and metabolite-phenotype associations: 2000 gene-phenotype associations in 74 phenotypes and 400 metabolite-phenotype associations. (1) RA: 3 out of 5 genes are all in HLA regions, and the other two are plausible (however, not previously known). (2) several metabolite-trait association, often driven by multiple SNPs (Figure S1). Some justification of metabolites.
- Phenotype similarity based on gene-phenotype relations: some expected, e.g. same phenotype, or closely related (e.g. LDL and HDL, BP and SCZ). However, some unexpected, e.g. breast cancer and insulin.
- Genetic mechanisms linking seemingly unrelated phenotypes (Figure 3): (1) Age-at -menarche and BMI and childhood obesity: TGF-beta or MAP signaling in top genes. (2) Pathway enrichment test: find pathways contribute most to phenotype similarity, Pearson correlation of phenotypes using only genes in a pathway. RA and CD: antigen presentation, inflammatory cell death. (3) Insulin and breast cancer: cAMP/cGMP signaling. (4) T1D and T2D: apoptosis.

- Biclustering analysis: a module of 10 genes or so and 7 diverse phenotypes (Figure 4), likely driven by insulin signaling.

Mapping robust trans-associations via cross-condition mediation analysis and validating trans-associations of trans-genes for GWAS SNPs (CCmed) [Yang and Lin Chen, 2019]

- Multi-tissue mediation analysis (CCmed): first need to establish association of locus L_i with cis-gene C_i. To do this, association test of α_C using all SNPs in L_i (using variance component test). Results expressed as F_{ik} for locus-cis-gene i at tissue k. Next do the mediation analysis: estimating C_i to T_j effect β_1 while adjusting for L_i. The results are expressed as Z_{ijk} for cis-trans pair ij in tissue k. Then estimating probability of mediation for pair ij in at least K_1 tissues:

$$P_{\text{med},ij} \geq Pr(\alpha_C \neq 0 \text{ in all K tissues}) \times Pr(\beta_1 \neq 0 \text{ in at least } K_1 \text{ tissues}) \tag{5.68}$$

  The two probabilities are computed from Primo.

- Finding trans-genes targeted by GWAS SNPs using mediation: one can use GWAS SNPs for L_i above and find the trans-genes targeted by GWAS SNPs. The problem is that GWAS SNP may be in LD with a cis-eQTL of C_i, but C_i is not the true target. Let G_i be GWAS SNP at locus i, and L_i the eQTL. Comparing with CCmed, the first step test association of G_i with C_i, but adjusting for L_i. However if L_i and G_i in close LD, $r^2 > 0.5$, then not adjust for L_i. The second step also adjust for L_i. Finally, estimation of the probability of mediation is similar, but find the best configuration (over K tissues).

- Validating trans-genes using MR: MR-Robin. Model: $L_i \rightarrow X \rightarrow Y$. Let β_{yi} and β_{xi} be the effect of SNP i on X (expression) and Y (GWAS), then:

$$\beta_{yi} = (\gamma + \gamma_i)\beta_{xi} \tag{5.69}$$

  where γ is the causal effect and γ_i reflects the deviation due to LD and horizontal pleiotropy. Use a random effect for γ_i. It is difficult to do this for multiple tissues, so use reverse regression of tissue-specific eQTL vs. GWAS effects. For tissue k,

$$\beta_{xik} = (\theta + \theta_i)\beta_{yi} + \epsilon_{xik} \tag{5.70}$$

- Results of CCmed-GWAS: in SCZ, use all SNPs in 108 loci, and eQTL from multiple brain tissues. Found 1400 genes.

- Validation of trans-genes using cis-signals: (1) Cis-eQTLs of 1400 genes are enriched with low GWAS p-values. (2) The genes are enriched with PrediXcan predictions.

- Validation of trans-genes by MR-Robin: 40 genes.

- Remark: issues with CCmed include, the estimation of mediation probability is based on a conservative lower bound: it requires L_i to be cis-eQTL of C_i in all tissues. This limits the applicability of the method. Could use enumeration strategy like CCmed-GWAS.

- Remark: possible issues with CCmed-GWAS. Not enough to address LD: in general, it requires G_i to have independent effect beyond L_i. This reduces the power of detecting G_i effects. Having a LD condition (don't adjust for L_i if in close LD) helps, but it has the risk of failing to adjust for LD. In practice, a large number of SCZ genes found, so it's OK to be somewhat conservative.

- Remark: MR-Robin. Assumptions are: same effect of gene on trait across all tissues (same θ). A strong assumption.

MR accounting for weak effects and pleiotropy using profile likelihood [Jingshu Wang, NHS, 2019]

- Motivation: visual inspection of effect size correlation, often see over-dispersion, better explained through pleiotropic effects of SNPs.
- Profile likelihood approach: let Γ_j be the GWAS effect of variant j to outcome, and γ_j the effect to exposure, and α_j the pleiotropic effect, then:

$$\Gamma_j = \beta\gamma_j + \alpha_j \tag{5.71}$$

  We assume $\alpha_j \sim N(0, \tau^2)$. This allows one to have $P(D|\beta, \gamma_1, \cdots, \gamma_p, \tau^2)$. Profile likelihood: replace γ_j in likelihood with its MLE, and still obtain unbiased estimator of β.
- Weak IVs and selection bias: include variants with $\gamma_j = 0$ will reduce power, but does not bias the estimate. Need to use independent data to estimate effect size γ_j. Show the effects of selection bias: use BMI male > BMI female, expect effect size 1, however, MR methods may obtain < 1 estimate because of winner's curse (not account for standard errors of effects). Note: current methods may show < 1 estimate even with $p < 1E-8$ selection of IVs - due to winners' curse.
- Experiments that show the benefit of weak IVs: reduce CI of estimates. Show that with weak IVs, obtain similar estimates to strong IVs (with larger CI).
- Detecting multiple modes in profile-likelihood function: e.g. reverse causality, mode at 0 and $1/\beta$.

Phenome-scale causal network discovery with bidirectionalmediated Mendelian randomization [Brown and Knowles, review for NG, 2020]

- Causal model: D traits, and data of N samples. Let X be the genotype matrix of $N \times M$ dim. We have a causal graph G (sparse), with G_{ij} the direct causal effect of Y_i on Y_j. Let β, $M \times D$ matrix be the direct effects of SNPs on traits, and γ be unexplained effects: we have

$$Y = YG + X\beta + \gamma \tag{5.72}$$

  Let $\hat{\beta}$ be the estimated genetic effects, $\hat{\beta} \propto X^T Y$. One can show that $\hat{\beta}$ is related to G and β by:

$$Y(I-G) = X\beta \Rightarrow X^T Y = X^T X\beta(I-G)^{-1} \Rightarrow (X^T X)^{-1} X^T Y = \beta(I-G)^{-1} \Rightarrow \mathrm{E}\,(\hat{\beta}) = \beta(I-G)^{-1} \tag{5.73}$$

  If we denote $\hat{R}$ be the estimated total effect of one trait on another - estimates from 2SLS MR, then we can also show how $\hat{R}$ is related to G:

$$\hat{R} = \frac{1}{N}(X\hat{\beta})^T Y = \frac{1}{N}(X\hat{\beta})^T YG + \frac{1}{N}(X\hat{\beta})^T X\beta + \frac{1}{N}(X\hat{\beta})^T \gamma \tag{5.74}$$

  Taking expectation on both sides. The first in the RHS is $\mathrm{E}\,(\hat{R})G$. Claims that the second term is diagonal, and the last term has mean 0. This leaves:

$$\mathrm{E}\,(\hat{R}) = \mathrm{E}\,(\hat{R})G + D[\beta(I-G)^{-1}] \tag{5.75}$$

  where D is diagonal matrix. This leads to the equation for estimating G: $G = I - R^{-1}D[1/R^{-1}]$.
- Analysis: is $(X\hat{\beta})^T X\beta$ diagonal? We note that $X\hat{\beta}$ is the PRS of traits, so we denote as $\hat{Y}$. We are interested in if $(X\beta)_i$ for trait i $\hat{Y}_j$ for trait j are independent. Diagonal means that if SNP k has a direct effect on Y_i, i.e. $\beta_{ki} \neq 0$, then SNP k will not contribute to the PRS of another trait Y_j. This seems to be what the model assumes. However, this is not true because $\hat{\beta}$ captures the total effect: so if there is a causal path from Y_i to Y_j, then SNP k will have an effect on Y_j.
- Estimation of R: (1) variant weighting - similar to Steiger filtering. Then use Eggar regression to estimate R. For trait i and j: first determine the putative IVs, all SNPs associated with i at a given threshold. Then for these SNPs, compare their effect sizes for i and j using Welch test. The weight is determined by the test statistic: SNPs with larger effect difference will have higher weights. (2) Use two sets of statistics: one set to do SNP selection, and the other for estimating effects. In UKBB, use male and female associations.

- Estimation of G by sparse inverse regression (inspre): the goal is to estimate an approximate R^{-1}, which can be plugged in to the estimation equation for G. To estimate R^{-1}, we find U, and V s.t. $UV = I$, U is close to $\hat{R}$ and V is sparse. This is done by solving the optimization problem in Equation (9).

- Simulation using pairs of traits: three settings, no pleiotropy, independent pleiotropy and correlated pleiotropy, among 5000 causal SNPs, 1000 are pleiotropic, with $\rho_g = 0.2$. Show that the weighted Eggar regression is better than Eggar regression: both FP and power. The FP rate at setting 3 seems relatively low, 0.085 at targeted type I error rate 0.05.

- Simulation of causal graph: random graph, graphs with hubs, and graphs with high in-degrees. F1 scores of bimmer much higher than elasic net on Eggar.

- UKBB results summary: 400 traits. Found 8K pairs from the matrix R, at FDR 0.05. Then using the sparse graph (inspre): limit to 2K pairs.

- Some examples of UKBB results: some causal effects can be explained by mediations. Ex. Time spent on TV (TSWT) has causal effect on other traits, walking pace, wheezing in the chest, father's death age. These can be explained by TSWT on BMI. Another example: age of first sex (AFSI) on knee pain, wheezing in the chest, etc. which can be explained by BMI. Possible mechanism: age of puberty.

- Heart disease sub-network. Effect of WBC on heart disease via direct effect, cholesteroal and BP. Note: found BP $\rightarrow$ cholesterol.

- Hemoglobin sub-network: Hb effect on bleed gum and cardic arrhythmia. For bleeding gum: likely by platlet. For the latter, likely anemia.

- The highest trait by out-degree: WBC. Effect on neuroticism, suffer for nerves, anxiety. Remark: likely by stress/cortisone, a confounder.

- BMI is the second highest trait by out-degree: BMI found to have an effect on vegetable intake, and other dietary behavior.

- Q: in weighted Eggar regression, the weights can be negative? Presumably truncated at 0.

Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics [Darrous and Kutalik, review for NC, 2020]

- Model notations: effects of G are denoted as $\gamma_x, \gamma_y, \gamma_u$. Causal effects: $\alpha_{x\rightarrow y}$ and $\alpha_{y\rightarrow x}$. Confounder effects: q_x, q_y. Observed summary statistics $\hat{\beta}_x, \hat{\beta}_y$, with variant index k.

- Prior distributions for genetic effects: proportion of non-zeroeffects denoted as $\lambda_x, \lambda_y, \lambda_u$. However, because of LD, the proportions are higher, denoted as π_x, π_y, π_u.

- Model of true effect sizes: assume the genetic effects are given. Let ρ_k be the LD of SNP k with all other SNPs (vector), and $\gamma_u, \gamma_x, \gamma_y$ be genetic effects of all SNPs (vectors). Then the true effect β_k^y would have three parts: from variants acting on U, on X and on Y directly:

$$\beta_k^y = (\alpha_{x\rightarrow y} q_x + q_y)(\rho_k \cdot \gamma_u) + \alpha_{x\rightarrow y}(\rho_k \cdot \gamma_x) + (\rho_k \cdot \gamma_y) \tag{5.76}$$

  Similarly, we can obtain the model of β_k^x, where it has contribution from reverse causal effects.

- Likelihood model of summary statistics: marginalizing genetic effects. The variance of genetic effect now has LD score of a variant in it. And the joint distribution of the $\hat{\beta}_x$ and $\hat{\beta}_y$ now follows mixture of 8 normal distributions. The 8 components correspond to how a variants is associated with 3 traits. The terms of the covariance matrix captures the variance and correlation of genetic effects.

- Remark: treatment of LD. The model considers only marginal statistics, but the effect sizes would be correlated of SNPs in LD, so it is still not a model fully account for LD.
- Inference: MLE of model parameters. Reparameterize some parameters, $t_x = h_u q_x, t_y = h_u q_y$ as scaled confoudner effects. Also i_x, i_y similar to LD score intercept term. Limit the parameter ranges t_x and h2g in [0,1], t_y, α are [-1,1].
- Identifiability analysis. What matters is the ratio of effects size: t_y/t_x under confounder effect and $\alpha_{x\to y}$ under causal effect. Under the model, expect two lines (Fig. S2), one from causal effect, and the other from confounder. When confounding is severe (large effect, or very heritable), or when two confounders with opposite signs: bimodal likelihood.
- Obtaining SE of the parameters: use MLE or jackknife. In simulations, jackknife is slightly more conservative. In real data: bimodal likelihood occurs more often, and jackknife leads to large SE, so use LRT instead.
- Remark: comparison with CAUSE. The value of q is roughly $\pi_u/(\pi_u+\pi_x)$, the proportion of all variants of X that acts through U.
- Remark: the model has identifiability problem under broad scenarios: (1) Simple case, causal effect, no confounder. This leads to full correlation of effect sizes. This can be explained by a causal effect, or a fully shared confounder, $q = 1$. (2) More generally: a causal effect and a confounder with opposite effect (proportion q). This leads to the effect of Y be close to 0 for variants acting on the confounder. We could explain this pattern by: no causal effect, and a confounder effect with proportion $1 - q$.
- Results: see review file. Main findings: lower SE of parameters comparing with standard MR. General agreement with standard MR in real data, difference largely due to the use of genomewide markers by LHC.

Chapter 6

Epigenetics

6.1 Overview of Epigenetics

Epigenetics flipping the genetic switch [Neil Lamb] `http://hudsonalpha.org/wp-content/uploads/2014/04/epigenetics.pdf`

- Figure 1. mechanism. Agouti gene: normally ON briefly. But if consitintutely on, agouti mouse. This phenotype is caused by demethylation of the promoter of the Agouti gene.
- Figure 2. transgenerational inheritance. Agouti phenotype passed to next generation. Dietary supplement with vitamin (methyl groups) during pregenency and nursing in Aguoti mice: offsprings show normal phenotype.

A New Kind of Inheritance [Skinner, Scientific American, Aug, 2014]

- Epigenetic actors: DNA methylation (28M sites), histone modification, ncRNA (interacting with epigenetic marks). Epigenetic marks get copied during replication.
- Epigenetic inheritance from pesticide: endocrine disruptors (pesticide) on pregnant rats, the offsprings have smaller testeis and few sperms. Passes to next generation: more than 90% show abnormalities. Evidence of epimutation that interfere with gonad development in male embryos.
- Epigenetic inheritance from famine: using data of 300 people born in 1890, 1905, 1920, women whose paternal grandmothers experienced one of these feast-famine swings as young children had markedly higher rates of fatal cardiovascular disease. Similar observations have been made in descendants of a Dutch population that experienced famine during World War II.
- Mechanism of epigenetic inheritance: (1) First wave of reprogramming, removal of all methyl marks except imprinted gene, after conception. (2) Second wave during primordial germ cell development: essentially complete, including imprinted genes. Epimutation may happen in (2) so that they are protected, similar to imprinting.

6.2 Imprinting and Maternal Effect

Background [Fangyuan Zhang talk, Jan, 2015]

- Imprinting: About 1% of human genome is imprinted, only 90 genes have been detected.
- Maternal effect: the effect of a variant depends on the genotype of the mother.

- Experimental techniques: for imprinting, can use mice to study (control mating); for maternal effects, use assisted reproduction that use surrogate mothers.

Methods for Detection of Parent-of-Origin Effects in Genetic Studies of Case-Parents Triads [Weinberg, AJHG, 1999]

- Motivation: suppose we have a disease susceptibility locus, where a is the risk allele. The effect of a may depend on its parent of origin: e.g. it may increase the risk only when it is inherited from father, this is called imprinting. Meanwhile, there may be maternal effect (prenatal effect).

- Intuition of detecting imprinting: suppose we have children of genotype A/a, where A is from mother and a father, and of genotype a/A. If there is imprinting, the risk is different in the two genotypes (or the fraction of cases). The problem now is that we cannot directly test the different risks since we are conditioned on the affected children.

- Extension of TDT - transmission assymetry: in TDT, the distorion of transmitted vs. non-transmitted measures the RR of the disease allele. So we simply stratify the data by parent-of-origin of the alleles. Then we have a 2 by 2 table: T vs. DT where T is from father or from mother, and the differential distortion can be tested by Fisher's exact test. This is called TDT_{MvsF}. The test however is not strictly valid when both parents are heterozygous. To deal with that, we remove these parents, and the test is called "transmission asymmetry test (TAT)".

- Model of trios incorporting imprinting and maternal effect: we only consider the most informative cases where one of the parent is heterozygous and the other homozygous. There are four scenariors (similar to TAT): we write down the genotype combinations (mother then father), and the relative risk of the child.

  - AA $\times$ Aa $\to$ A/A (010): 1
  - AA $\times$ Aa $\to$ A/a (011): R_p
  - Aa $\times$ AA $\to$ A/A (100): S_1
  - Aa $\times$ AA $\to$ a/A (101): $I_m R_p S_1$

  where R_p is the RR of a in father, and I_m is the imprinting effect, S_1 is the maternal effect.

- Parental Assymetry Test (PAT) and Parent-of-origin LRT (PO-LRT): the most informative statistic is given a certain mating type, the relative ratio of A/a vs. a/A in the child. In LRT, we consider the relative frequency of the two as $P(M > F|\text{mating type}, C)/P(MCF|\text{mating type}, C)$, which depend on the parameter I_m, and we test using LRT (PO-LRT). A simple approarch is: suppose there is no maternal effect, then given the mating type combination, under $H_0 : I_m = 1$, the relative ratio of A/a is 0.5. And we test if the counts of A/a and a/A are equal - this is PAT.

- Comparison of tests:

  - Both TAT and PAT are valid only when there is no maternal effect.
  - PAT is more powerful than TAT.
  - LRT is more robust, but significantly loses power when there is no maternal effect.

Joint detection of association, imprinting and maternal effects using all children and their parents: LIME [Han & Lin, EJHG, 2013]

- Background:

  - Different designs such as case-parent trios, and may include multiple children (including unaffected), and general pedigrees.

– Different tests: (1) Nonparametric test: assumption of no maternal effect. (2) Parameteric test: stringent assumptions such as mating symmetry and parental allelic exchangeability (e.g. PAT is valid only when two assumptions hold).

- Motivation: testing both maternal effect and imprinting using likelihood; incorporate additional siblings. The study uses both case famlies and control families.

- Partial likelihood model: e.g. consider a family of two children, where one of them is affected. Let M, F, C_1, C_2 be the genotypes, and $D_1 = 1, D_2$ be the disease trait of the children. We are interested in the condtional prob.

$$P(M, F, C_1, C_2, D_2|D_1) = P(M, F, C_1, D_1)P(C_2|M, F)P(D_2|M, F, C_2)/P(D_1) \tag{6.1}$$

  This likelihood can be factorized s.t. only part of them is dependent on the imprinting parameters, and the other nuisance parameters. So our inference can be based on the partial likelihood containing imprinting effect.

- Intuition: the main nuisance parameters are frequencies of genotypes (mating types). To avoid them in testing, we use the idea similar to two sample Poisson test: given a mating type, we compare the frequency of cases vs controls, the number of cases conditioned on the total number of events follows Binomial distribution whose parameter depends only on the relative risk parameters, but not genotype frequencies.

Identifying Heterogeneous Transgenerational DNA Methylation Sites via Clustering in Beta Regression [Shengtong Han talk, 2014]

- Problem: given the methylation data of many CpG sites in n trios, we want to identify different inheritance patterns: e.g. some sites are average of parents, some are new methylation sites, some follow from one of the parent (imprinting). The problem is to find such patterns in an unsupervised fashion.

- Idea: we care about the inheritance patterns, not the absolute levels, so we model the relation of offspring and parents (using a linear model), then cluster the sites by their relations (coefficients). One issue is that statistically, we shouldn't simply average methylation levels of multiple observations of one site (not normally distributed); so instead we model the distribution of methylation level.

- Model: let O_j, M_j, F_j be the average methylation levels of offspring, mother and father at the j-th site, they are related by:

$$O_j = \gamma_{0j} + \gamma_{1j} M_j + \gamma_{2j} F_j \tag{6.2}$$

  Then the cofficients form clusters. To simplify, instead of creating a model for γ_j's, we simply assume there are K clusters, and each cluster has a set of values of γ's. To refine the model, instead of averaging observed methylation level, we directly model the methylation data. Let y_{ij} be the observed methyl. level of offspring i in the j-th site, and Z_{1ij}, Z_{2ij} be that of parents i. We model them as Beta distribution:

$$y_{ij} \sim \text{Beta}(\alpha_j^o, \beta_j^o), \quad Z_{1ij} \sim \text{Beta}(\alpha_j^M, \beta_j^M), \quad Z_{2ij} \sim \text{Beta}(\alpha_j^F, \beta_j^F) \tag{6.3}$$

  The average methylation level O_j, M_j, F_j are thus simple functions of α's and β's. The model is fit by EM.

- Remark: it may make more sense to model data of individual families, in other words, we directly model y_{ij} as function of γ_{0j}, γ_{1j}, γ_{2j} and Z_{1ij}, Z_{2ij} (γ's are defined wrt. indiviudal families, not average).

- Lesson: clustering based on similar relationships among variables.

6.3 Epigenetics in Human Diseases

Epigenome-wide association studies for common human diseases [Rakyan, NRG, 2011]:

- Goal: for any human complex disease, we remain unaware of the proportion of phenotypic variation that is attributable to inter-individual epigenomic variation. This problem can only be elucidated by large-scale, epigenome-wide association studies (EWASs).
- Epigenetic information can be transtmitted via: DNA methylation, hmC, histone modification, ncRNA (miRNA, piRNA, lncRNA).
- Types of DNAm variations:
  - Methylation variable position (MVP). A CpG site that shows differential methylation between different disease states.
  - Differentially methylated region (DMR). A region of the genome at which multiple adjacent CpG sites show differential methylation. they are typically < 1 kb, but they can exceed 1 Mb.
  - Variably methylated region (VMR). These regions are defined by increased variability rather than gain or loss of DNAm.
  - Allele-specific methylation (ASM). These are positions or regions that vary in DNAm depending on the parent-of-origin, the presence of a polymorphism or as a result of a stochastic event
- Evidence of epigenetic compoent in complex disease:
  - Monozygotic twin concordance for any complex disease is almost never 100%.
  - The incidence of several complex diseases - such as T1D - rising in the general population and is frequently altered in migrant populations, suggesting a role for non-genetic factors.
  - Epidemiological evidence suggests that a suboptimal in utero or early childhood environment can have an impact on disease outcomes (such as type 2 diabetes) in adulthood.
  - In cancer: gain of methylation in CGI, loss-of-imprinting, loss of DNAm at repeat, esp. satellite DNA (main structural component of heterochromatin).
- Causality problem: epigenetic variation can be causal for disease or can arise as a consequence of disease.
  - A key step towards achieving this goal is to determine whether the variation is present prior to any overt signs of disease. However, this does not guarantee causality.
  - It is also possible that the underlying genotype influences epigenetic variation, e.g. from methQTL stuides. Some evidence of trans-meQTL, but not as prevalent as cis-effects.
- Sources of DNAm variation: (1) inherited from parents (transgenerational inheritance); (2) environmentally-induced, including in utero effect: developmental reprogramming; (3) stochastic, could be present in many tissue if early in development; (4) genetic variation. Disease state could affect DNAm (source 2): e.g. changes of immune cell DNAm from autoimmune diseases.
- EWAS study designs:
  - Retrospective (case-control). However, a retrospective study cannot determine whether the identified epigenetic variants are due to disease-associated genetic differences, post-disease processes.
  - Parent-offspring pairs. These could be useful in EWASs that aim to identify transgenerational transmission of epigenetic marks. The genetic information could then be used to eliminate the possibility that genetic modifiers are causing the epigenetic variation.

  - Monozygotic twins. Monozygotic twins who are discordant for a disease of interest represent a useful resource for EWASs. However, these studies cannot be used to distinguish between cause and consequence.
  - Longitudinal cohorts: can be invaluable for establishing the temporal origins and stability of disease-associated epigenetic variation, thereby helping to distinguish causal epigenetic variants from consequential ones.
  - Example design: Start with genome-wide DNAm analysis of monozygotic twins who are discordant for the disease to identify disease-associated MVPs in immune-effector cells. Then take these MVPs and assay them in the same type of immune-effector cells from a prospective cohort to look at DNAm at these sites in unrelated individuals who were sampled both before and after disease onset. Any MVPs that can be validated prior to disease onset are then candidate causal variations.

- Environmental effects: unlike GWASs, environmental factors can also directly confound an EWAS by affecting both epigenotype and phenotype. Indeed, if GWAS data are also available on the EWAS individuals, it may be appropriate to adjust for leading principle coordinates of both genetic and epigenetic states.

Epigenome-wide Association Studies and the Interpretation of Disease-Omics [Birney, PLG, 2016]

- Specific challenges of EWAS:
  - Cell subtype heterogeneity. Even present in purified cell types (subtypes exist).
  - Cellular mosaicism: typically DNAm in most CpGs are 0 or 100%, so proportional changes in EWAS represent change of proportions.
  - DNAm variation can result from transcriptional variation?
  - Genetic variation between individuals: powerful influence, about 20-80% of DNAm variability. Typically in EWAS, not do things like population stratification.
- Advices on EWAS: (1) longitudal cohort if possible. (2) Address cellular heterogeneity: pure cell types, better methods (e.g. CellMix). (3) Correcting for transcriptional and genetic difference.

Epigenome-wide association study of body mass index and the adverse outcomes of aiposity [Nature, 2017]

- Background: methylation array: based on BiS conversion, or each CpG site has two bead types, recognizing different methylation status.
- Method: causal vs. consequential analysis of relationship among SNP, CpG and BMI. Causal model: SNP $\rightarrow$ CpG $\rightarrow$ BMI, use the strongest cis-SNP of CpG as IV, and the predicted SNP to BMI effect is the product of SNP to CpG effect and CpG-BMI correlation. Consequential model: similar, but use the polygenic score of BMI as IV.
- Data: EWAS on 5K samples. DNA methylation in blood (450K array). EWAS: 187 significant CpG loci associated with BMI. Covariates in EWAS: technical factors, SNP PCs, methylation PCs.
- Analysis of epigenetic heterogeneity across cell types: methylation in isolated CD4 and CD8 T-cells. Compare methylation levels of 187 loci in blood vs. fat, liver, muscle etc (21 tissues): $R = 0.37 - 0.93$.
- Causal and consequential analysis (Figure 2): most CpG sites are consequence of BMI, rather than the causes. Likely explanation: changes in lipid and glucose metabolism associated with BMI.
- Methylation as marks/predictors of T2D: independent of BMI.
- **Lesson**: DNA methylation of strong loci (associated with trait) may not be highly cell-type specific. However, most of them may not be causal loci.

Chapter 7

Systems Genetics

7.1 Methods for Molecular-QTL Analysis

Unique problems/opportunites of eQTL analysis:

- A large number of traits are analyzed simultaneously: this raises challenges of analysis (multiple testing correction); also makes it possible to develop new methods that exploits the correlation among traits.
- Mechanisms of eQTL: easier to study than complex traits. What the studies reveal about gene regulation.
- Road to phenotype: natural bridge between genotypes and compelx phenotypes.

Methods for detecting eQTLs [Personal notes]:

- Experiment design: linkage mapping in families or from crosses between two parental strains; association mapping using samples from unrelated individuals in a population. Association mapping has lower statistical power, but will be the method of choice in the future because of more genetic variations.
- Correction of confounders: the challenge is latent confounders for expression data: PCA, SVA, PEER. Also may need to correct for population ancestry (PC).
- eQTL method: the general idea is: divide the samples according to the marker alleles and test if the groups differ signficantly in expression. Could be enhanced by modeling genetic interactions among multiple loci. Common methods: linear model, non-parametric test (e.g. Spearman correlation).
- Dimensionality reduction: combine multiple transcripts that behave similarly into single traits.
- Meta-analysis: specific eQTLs are not generally replicated across studies. In model organisms, this may be due to the difference of strains used in different studies. In human populations, this could be due to various artifacts such as: the use of different p-value cutoffs, of different distance threshold of defining distal eQTLs. A meta-analysis using data from multiple studies, using a single consistent method is necessary for meaningful comparison.

Significance and multiple testing correction in QTL studies [personal notes]:

- Background: BH correction vs. Storey's q-value. The former is more conservative since it implicitly assumes that $\pi = 0$, while q-value method estimate π_0 from data (though the estimate is conservative).
- Background: is FDR correction valid when the tests are correlated? On average the FDR is correct, but at a specific study, FDR may be over- or under-estimated. If correlations are only local (ie. a test is correlated with only a small subset of tests), then FDR is OK.

- Challenge: why cannot we pool results of all genes? If we pool all tests, and do single multiple testing correction, we may not be able to control FDR for each phenotypes. Ex. one gene has many associated SNPs (tight LD): adding this signal gene will introduce many strongly associated SNPs. This leads to inflation of false positives in other genes because FDR measures global false discoveries.

- The importance of adjusting for each gene: if we know the effective number of independent tests per gene, we should adjust for significant using different thresholds for different genes.

- Calibration problem: FDR correction methods and Beta approximation (FastQTL) all require p-values to be calibrated. If not, need to calibrate e.g. by permutations. This happens when we test top SNPs, and min-p is not uniform, so we use permutation to get empirical p-values for min-p.

- Permutation-based strategy to control QTL discovery by eGenes: a general strategy. We obtain min-p (lead SNP) per gene; then control FDR at the level of all genes (but each gene contributes only one SNP). Typically, the min-p is not calibrated, so performe permutations to obtain null distribution of min-p for each gene. This adjusts for local LD.

- Adjusting for covariates in permutation: suppose we have data G_i (genotype) and Y_i (phenotype). We should keep Y_i, while permuting G_i, i.e. replace G_i by G_j for some j. When our causal diagram is $Z \to Y$, but no arrow from Z to X, then this permutation preserves the Z to Y effect. However, when Z is a confounder, is this permutation effective?

  - Remark: in general, permutation of a variable will destroy all the edges in the causal diagram of that variable.

- Analysis: when pooling is effective? If each gene has a similar number of tests, or we have already adjusted for the difference across genes, then it's OK to pool. Examples in literature:

  - McVicker, Degner: tests are mostly local, so a gene/peak has a small number of tests. McVicker: only 2kb around each peak.
  - Rasqual and Battle et al: adjust for the number of SNPs/tests per gene, by first performing Bonferroni correction for top SNP at each gene, then FDR on the p-values of top SNPs of all genes.

- Power problem: different genes/tests may have different power. Using a uniform cutoff thus may not be optimal. [Degner, Nature, 2012] do FDR correction, using different thresholds, for different bins of peaks.

How to detect QTL in the presence of hidden confounders? [M6A QTL meeting with M. Stephens, 2018]

- PCA: R has several versions of PCA. They may be different in normalizing/standardizing variables. E.g. prcomp() center, but does not standardize the variables. In general, recommend to standardize s.t. the variables contribute equally to PCA (otherwise, variable with large variance will dominate the reconstruction error).

- Null distribution in regression analysis (Beta-Binomial regression in our case): typically use asymptotical distribution to get p-values. This may break down, e.g. when the number of samples is not large relative to the degree of freedom of test (number of covariates in regression).

- How dependency changes null distribution? When the test statistics are correlated, it will change the null distribution: even though FDR is still correct on average, it can be inflated or deflated in a specific study. What matters is the global pairwise correlation. In the QTL case, even if Ys are correlated, Xs are most not, so Z-scores are largely independent. The null distribution should be OK.

- Permutation: one way to obtain true null distribution. Or we can use it to test if this poses a problem.

- Quantile normalization: when the outcome in regression may not follow normal, and there are outliers, we can quantile normalize the outcome.

- Correct for confounders: PCA vs. PEER vs. SVA. PEER differs from PCA in that it has normal prior to shrink coefficients. SVA solves the problem, where the variable of interest correlates with the confounder(s), e.g. PCs. If we regress out PC(s), we remove some of the signal. So ideally, we will only regress out the part of PCs that are uncorrelated with the variable of interest.

- Stabilization: sometimes, we may not have a lot of data for the parameters (interested ones or nuisance). Stabilizing the estimate via hierarchical Bayes or ASH would help. Ex. in paired peak calling problem, the read count in controls may be low - we can stabilize the estimate of background rate, e.g. by using a spatially smoothed prior.

Model of QTL mapping with read counts [personal notes; WASP; RASQUAL]

- Basic model: let Y_{ij} be the read count of feature (expression, peak, etc.) j of sample i, and G_i the genotype, we have:
$$Y_{ij}|G_i \sim NB(\lambda_{ij}K_i, \theta_j) \qquad \log \lambda_{ij} = \beta_0 + G_i\beta \tag{7.1}$$
where λ_{ij} is the relative mean (expression level expressed as fraction of reads), K_i the library size (or size factor) of sample i, and θ_j the overdispersion parameter.

- Normalization: the expected read count depends also on other factors, e.g. GC content and IP efficiency of a sample. So we should correct for these factors by replacing K_i with K_{ij}, which is the expected read counts in feature j if there is no such difference: library size, GC content, IP efficiency, PCs.

Research directions and statistical challenges [Gilad & Pritchard, TIG08; Kendziorski & Wang, Mamm Genome, 2006]; Goring, Tissue specificity of genetic regulation of gene expression, [NG, 2012]:

- Multiple hypothesis testing correction: especially important for eQTL studies. Ex. with multiple expression traits, often only the strongest eQTLs are considered, and the p-values are transformed to q values for correction.

- Rare eQTL: sequence information, perhaps coupled with different study designs, such as those based on large families, will be required to detect rare eQTLs of strong effect.

- Mapping the functional eQTL sites.

- Identifying eQTL hotspots: the simple method: count eQTLs. A better alternative is to sum the evidence of all transcripts (weighted) in a region, and this shows improvement over simple counting [Kendziorski & Attie, Biometrics, 2006].

- Reconstruction of GRN: e.g. identify local eQTLs for genes that also mapped as distal eQTLs for other genes. Correlation of transcripts is used for identifying functional modules.

- eQTL across different tissues: how are they shared. One strategy is to have large sample eQTL on a few accessible tissues and use the results as a reference for other tissues.

- Use eQTLs for linking genotypes to complex diseases, and cell line phenotypes (e.g. sensitivity to chemotherapeutic agents).

Reference: [Rockman & Kruglyak, Nature, 2006], [Revealing the architecture of gene regulation: the promise of eQTL studies, Gilad & Pritchard, TIG, 2008], [Mapping complex disease traits with global gene expression, Cookson & Lathrop, NRG, 2009], [From expression QTLs to personalized transcriptomics, Montgomery & Dermitzakis, NRG, 2011].

High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation [Veyrieras & Pritchard, PG, 2008]:

- Problem: find the causal variant (eQTN) of the expression traits.
- Methods:
  - Data: 210 unreleated individuals from HapMap project (immortalied B cells). A core dataset of 11,446 genes. The analysis of eQTLs is limited to cis-eQTLs, defined as 500 kb upstream or 500 kb downstream of the genes.
  - Method: hierarchical model: let Z_{jk} be the indicator variable of whether SNP j is an eQTN of the gene k, the prior of Z_{jk} is modeled as logistic regression of the functional annotations/features of SNP j: its distance to TSS or TES, its conservation, etc. The expression of individual genes is given by:
    $$P(E_k) = \sum_j P(Z_{jk} = 1)P(E_k|Z_{jk} = 1) \tag{7.2}$$
    The coefficient of the second term is expression-trait specific and the regression coefficients of the first term (prior) are shared by all genes.
- Results:
  - Distribution of eQTNs: strongly enriched near TSS and TES. Most of the background signals in simple eQTL analysis (linear regression) were removed. And more eQTNs were found: 1586 vs 744 (from simple analysis), from higher sensitivity of the hierarchical model to signals in locations that are likely a priori.
  - Functional annotation of eQTNs: the TSS and TES peaks tend to be highly conserved across mammals. Internal introns have a deficit of eQTNs compared to exons and the first introns.

Epistasis in yeast eQTL data [Hannum & Ideker, PG, 2009]:

- Motivation: in eQTL studies, one could identify genetic interactions among markers, what are the interpretations of such interactions? Are they enriched with PPIs?
- Methods:
  - Data: [Brem05]
  - Natural genetic interaction network: first identify markers that genetically interact [Storey, PB, 2005]; then bicluster markers: an exhaustive genome-wide scan is performed to identify interacting interval pairs, i.e. those that are enriched for marker-marker interactions.

  Results:
  - A network of 2,023 interactions between 1,977 genomic intervals.
  - Interacting intervals are enriched with protein complexes.

Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses [Stegle & Durbin, Nature Protocol, 2012]

- PCA: factors that explain total variation of expression. Compliexity is controlled by specificying the number of PCs.
- PCASig: similar to PCA, but the selection of PCs is determined by significance analysis.
- Surrogate variable analysis (SVA): accounting for fixed effect, correction for latent factors, using orthogonal vectors. Similar to PCAsig, but allow sparse non-orthogonal components.
- VBQTL (Stegle, PLCB, 2010): ARD prior to provide shrinkage. Interpretation of hidden factors: eg. cell growth (highly correlated, $r^2 = 0.96$). The global factors identified can be further analysed for biological signals, looking for GO term over-representation in the genes that they affect.

- PEER: hidden confounders can be explained as TFs and hot-spots. Use of PEER: Figure 1. A-B, eQTL mapping correcting for hidden confounders. C. association with hidden factors.

A statistical framework for joint eQTL analysis in multiple tissues [Flutre & Stephens, PLG, 2012]

- Background: weighted Z score method for multi-tissue analysis - combine p-values of multiple tissues. It does not so easily allow for investigation of heterogeneity. Problem: heterogeneity means the effect may be different in different tissues, thus combine them using weighted-Z is not optimal.

- Model: given a SNP and gene pair, suppose there are S tissues, let the effect size (linear model) in the s-th tissue be β_s. The problem is to define a prior on β_s so that information can be shared across multiple tissues. Define γ, a binary vector, as the "configuration" of the SNP across S tissues: 1 if active. Several models of the prior on configurations:

  - Default choice (BMA model): uniform prior on the number of active tissues (from 1 to S), then within each value, uniform prior on the configurations.
  - BMA.lite model: only two configurations, a single active tissue or all 1's with weight 0.5 each.
  - Hierarchical model (HM): over all genes, estimate the weight of each configuration by borrowing information across all genes.

  The prior of the effect sizes given the configurations: obviously if $\gamma_s = 0$, then $\beta_s = 0$. In general, we have:

  $$\beta_s|\bar{\beta}, \gamma_s = 1 \sim N(\bar{\beta}, \phi^2) \tag{7.3}$$

  Furthermore, a prior of mean effect size (across genes or SNPs?): $\bar{\beta} \sim N(0, \omega^2)$.

  - Why the model works? The prior of γ favors multi-tissue eQTL, in fact the prior of tissue-specific eQTL is only $1/S$.

- Statistical test: (1) Combining SNP information of a gene: for our Bayesian approach the test statistic is the average value of BFs over all SNPs in the cis candidate region of that gene; (2) Using BF as test statistic, and compute p-values by permutation (permutation of individual labels).

- Data: Dimas et al, Science, 2009, LCL, T-cells and fibroblast in 75 individuals. A subset of 5012 genes robustly expressed in all three cell-types, and cis-eSNPs only (within 1Mb of TSS) Joint mapping increases power: at FDR < 0.1, 1321 genes vs. 811 genes (tissue-by-tissue analysis)

- Application of hierarchical model: an estimated 88% of eQTLs being common to all three tissues. Caution: the estimates necessarily reflect patterns of sharing only for moderately strong eQTLs (strong enough to be detected): patterns of sharing could be different among weaker eQTLs.

- Remark:

  - RNA-seq data: count data, better to use a different model than normal.
  - When S is large, need a better model of prior. In particular, having a separate parameter for each possible configuration is unattractive (both statistically and computationally) for large S. Idea: use a tree model (similar to phylogenetic model of lineage-specific events); or an Ising model that favors related tissues.
  - Questions: what are biological influences of general or tissue-specific eQTL? eQTL location, type of genes (housekeeping vs. specific)?

WaveQTL [Heejung Shim & Stephens, AoAS, 2014]

- Background: wavelet transform method.

- Fourier transform: a function is decomposed into a sum of multiple basis functions (trigonometric). This leads to a compact representation: e.g. a function that looks complex may become a simple sum of a few basis functions. In practice, this often means removing noises in a function (the higher order terms).
- Wavelet transform: the idea of Fourier transform can be generalized. Instead of having basis functions of specific forms, we focus on certain "properties" of the function (mean, spatial asymmetry at different scales, etc.), and a function is represented as a set of properties. This can acheive similar goal of denoising by focusing on the top few components.

- Discrete Wavelet Transform (DWT) applied to spatial genomic data: suppose we have data $d = (d_1, \cdots, d_B)$ where d_i is the value at the i-th position in a region of size B. We can extract these features from the data, such as:

$$y_{01} = \sum_b d_b \tag{7.4}$$

the total count,

$$y_{11} = \sum_{b \leq B/2} d_b - \sum_{b > B/2} d_b \tag{7.5}$$

the difference of counts between the first and second halves; and similarly, y_{21} and y_{22}, the difference between the first and second quarters; and the third and fourth quarters; and so on. So the vector d can be transformed to a vector y (linear transform), which capture the same data. The advantage is that y has the denoising property s.t. we can focus on the first few components.

- Model of waveQTL: consider the sequencing data (depth of coverage) at a region, let it be d. We use DWT to get y. Let y_{sl} be the value at the scale s and location l. We perform association of genotype g and y_{sl} separately under a linear model - let γ_{sl} be the underlying binary indicator. To increase the power, we combine information across all y_{sl} of the same scale s by assuming a model:

$$P(\gamma_{sl} = 1 | \pi) = \pi_s \tag{7.6}$$

And we test if $\pi_s = 0, \forall s$ using LRT.

- Extenstion: multi-seq. The idea is similar, but we model the count data directly. Suppose the data is generated from multiple Poisson processes, let p_{sl} be the rate of the process of scale s and location l, and the observed count is the sum of the counts from all these processes. In other words, if we know the total count in a region, the difference of counts between the first and second halves, the difference of counts among quarters, and so on, we can recover the original count at each position. Then we perform the association analysis of genotype and p_{sl}.

- Question: the wavelet coefficients y_{sl} are not independent across scales, for example, if y_{sl} is large at some high resolution (e.g. at 3rd quarter vs. 4th quarter - a peak at 3rd quarter), then y_{sl} is likely non-zero at a lower resolution (1st half vs. 2nd half, since there is a peak in the second half). Can the model combine information across multiple scales?

- Lesson: for spatial/functional data, extract features/properties that summarizes the data, and this can be achieved through Fourier transforms, wavelet transforms, etc. More generally, we may work only on the extracted features even if we cannot completely recover the data.

Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals [Battle et al, GR, 2014]:

- Expression quantification: use HTSeq for gene expression, BEDTools for exon expression and cufflinks for isoform expression.

- Correction for latent confounders: HCP method [Mostafavi, 2013], correct for technical and biological factors, including blood cell-type frequencies and the time of the blood draw. Background: correct for hidden confounders greatly increase power of cis-eQTL. HCP: similar to factor analysis. 35 known confounders (Table S1): sequencing depth, cell type frequency, etc.
- Testing eQTL: Spearman rank correlation. For cis-eQTL, only SNPs within 1Mb, and Bonferroni correction within a gene. eQTL using gene-level significance at FDR 0.05.

GTEx [Science, 2015]:

- eQTL mapping: using Matrix eQTL. (1) Expression data: RPKM. Quantile normalization across genes in a given tissue. The expression values for each gene were transformed into a standard normal based on rank. (2) Correction of confounders: 15 PEER factors, gender, first three genotype PCs. (3) FDR control: based on Matrix eQTL, which uses BH corrections. Separate p-value thresholds and FDR calculation in cis and trans- analysis.
- Identifying eGene (eQTL containing gene): use minP as test statistic for each gene. Then permutation: swap sample label and expression data. For each gene, then obtain empirical p-value (each gene has its own null distribution from permutation), and do FDR correction, using Storey approach.

Fast and efficient QTL mapper for thousands of molecular phenotypes (fastQTL) [Ongen and Delaneau, Bioinfo, 2016]

- Background: permutation of a large number of times for each molecular phenotype. However, to get accurate statistical significance of the most associated QTLs may require a large number of permutations.
- Direct permutation: permutation that leaves genotypes unchanged (preserving LD). See Equation (1), usually add pseudocount of 1 in calculation of p-values (otherwise, p-value may be equal to 0). Do this for each gene separately.
- Adaptive permutation: for each gene, suppose the strongest p-value among all its SNPs is p. Permute a certain number of times s.t. at least B (e.g. 100) null p-values are smaller than p.
- Background: suppose we draw from uniform distribution n times, then the k-th smallest value follows $U \sim \text{Beta}(k, n)$.
- Beta approximation: model the null distribution of the smallest p-value in a gene as Beta distribution. In one phenotype, permute R times, and let $p_i, 1 \leq i \leq R$ be the smallest p-value in the i-th permutation. We can then fit $\text{Beta}(k, n)$ to p_i's. This allows us to obtain adjusted p-value for the min. p-value of a gene, and hence FDR control at the gene level.
- Results: the parameter k is close to 1 for most genes, and n is between 1000-4000, a few times lower than the number of tested SNPs.
- Beta approximation is much faster than direct permutation to get accurate p-values (Figure 2a): it requires 100-1000 permutations vs. 100K for direct permutations.
- Remark: the Beta approximation approach controls FDR at the gene level (eGenes).

Multi-tissue eQTL (MASH) [Sarah Urbut, May, 2015; Apr, 2017]

- Motivation: in the published muti-tissue eQTL paper, the effects are discretized (on or off in a tissue). However, we want the effects to be contiuous: e.g. possible that an eQTL is active in two tissues, but effect sizes very different.

- Idea: a prior covariance matrix of β (effect size in each tissue), the covariance terms encode both the effect sizes and the correlation of effect sizes of an eQTL across tissues. Use a mixture model for the prior covariance: for some eQTL/gene, it is correlated in one set of tissues; for another eQTL/gene, correlated in another set of tissues.
- Model: the effect sizes are standardized, i.e. Z-scores, defined as $Z = \beta/se(\beta)$. The likelihood is:

$$\hat{\beta}_j|\beta_j \sim N(\beta_j, \hat{V}_j) \tag{7.7}$$

  where β_j is the effect size of the j-th SNP-gene pair. The prior distribution:

$$\beta_j|\pi, U \sim \sum_{k,l} \pi_{k,l} N_k(0, \omega_l U_k) \tag{7.8}$$

  where U_k is the K components of the MVN and π the mixture proportions. ω_l are the stretch factors (to scale effect sizes). We assume that U and ω are pre-specified, and the problem is to estimate $\pi_{k,l}$.
  - Remark: U_k represents both correlations and effect sizes. Ex. large effects in both tissues and (large, small) in the two tissues (both are independent) are represented by two different U_k's.
  - Remark: (Dan's comments) the model is based on sharing of Z-scores, instead of actual effect sizes. But Z scores are affected by sample size (s.e. of β), which may differ substantially across studies.
  - Remark: we need ω_l because for each specific variant j, even if the effect size pattern is given, we still need to know/model the actual effect size. So ω_l captures the actual effect size for each SNP.
- Inference of π: use the fact that mixture of normal prior with normal likelihood leads to mixture of normal posterior. The reason of using fixed U is that EM algorithm on mixture of normal does not work if the normal covariance matrix is different.
- Specifying ω and U: we simply use a large number of ω_l's (the large ω_l's will be discarded by the data). For U, the idea is to approximate them using the observed covariance of effect sizes. Suppose we obtain the t-statistic (measured effect size) of SNP-gene pairs - for each gene, we choose the strongest SNP. Then the sample covariance matrix of the t-statistic is our starting point for U. Other choices of U_k are obtained through Sparse Factor Analysis (SFA): do SFA on the covariance matrix:

$$X = \Lambda F + E \tag{7.9}$$

  where X is the observed effect sizes, F the factors ($K \times R$ matrix, where R is the number of tissues), and Λ the loading of X on F. Choose the top q latent factors in U_k: $[(\Lambda F)^T(\Lambda F)]_q$. In GTEx analysis, use dimension reduction to learn 6 different U's, plus 44 rank-1 matrices (one for each tissue).
  - Remark: U_k represents the covariance structure given by the k-th latent factor: it is covariance of the reconstructed X using only the k-th factor. Note: use only the samples where the factor is non-zero?
  - Remark: factor analysis is based on the idea that effects can be deconvulated into sum of effects. How would this reconcile with mixture model (discrete structure)? Intuition: when latent variables are sparse, then for each sample, often only one latent variable is non-zero, so this reduces to mixture of normal.
- Opposite signs of effects in different tissues: observed in the data. However, this is likely due to SNPs in LD (two different SNPs have two different signs in two tissues). The evidence: two-SNP model provides a better fit of the data.
- Results of applying the analysis to GTEx data: observe the cases where additional tissues may (or may not) change the estimated effect in one tissue. In the negative example, the effect in brain is not changed. Overall tissue similarlity: brain is different from the rest.

- Effect sample size gain: relatively large from a few hundred or even dozen to > 1000. Q: vary across SNPs/genes? Also loss of power for tissue-specific eQTL?
- Remark: this is a generl statistical problem of inferring MVN. Need to specify the prior covariance matrix.
- Remark: how does the sparse factor approach encodes subtle configurations, e.g. a small set of eQTL have effects in a particular configuration $(1, 0, 1, 0)$ (suppose $R = 4$)? The intuition is that the model will learn a sparse factor that is active only in the tissues 1 and 3; then this subset of eQTL have higher loading in this factor.
- Lesson: factor model can explain the mixture scenario: if a subset of variables display certain covariance, then we create a factor that explains this covariance.
- Questions:
  - A simple strategy: for the prior of β, using binary configurations (could use mixed membership model so that only a small number of configurations will be actually used), and for each tissue, add a scaling parameter - e.g. a mixture of Gaussian. What's the disadvantage of this method?
  - Learning the covariance matrix: use the strongest SNP per gene. Does this create some kind of bias? Ex. when testing any gene-SNP pairs, most often the effect is small across all tissues.
  - Inference of U: sparse factor analysis, what assumptions? Ex. sparsity of loading.
  - The assumption of one eQTL per gene: how do we relax this assumption?
- Mash-Common-Baseline: Gene expression data across time series: time 0 vs. 1 to t. Let C be true expression (vector across all conditions): assume C follows MASH prior, and the error term is correlated (across conditions) because of shared control (time 0). Use MASH prior for true effects. Without correcting for shared control: much higher type I error. Model: let C_j be expression of j (R-dim. vector where R is number of conditions), we have:

  $$\hat{C}_j | C_j \sim N(C_j, \hat{V}) \tag{7.10}$$

  We consider the difference of expression vs. common control, denoted as $\delta_j = LC_j$, where L is $(R-1)\times R$ matrix (subtracting C_j for the common control). Then the observed differential expression, relative to common control, is:

  $$\hat{\delta}_j = L\hat{C}_j = LC_j + LE = \delta_j + E^* \tag{7.11}$$

  where $E^* \sim N(0, L\hat{V}L^T)$. We then model δ_j using MASH.
- Application of MASH to GWAS: 16 traits - summary statistics. Use top 1000 to initialize covariance matrices. Training: use 100K SNPs to learn the covariance matrices U_k. Then use MLE to estimate the parameters π with 50,000 random SNPs. Found 300K associations. Q: independent samples. Q: LD leads to shared effects.
- Questions/Remark: how to apply it to eQTL data, joint mapping across genes.

Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions (MASH) [Urbut and Stephens, NG, 2019]

- Learning U_k matrices: input is $J \times R$ matrix, where J is independent SNP-gene pairs (best SNP per gene). Two sources of U_k's: (1) Data driven covariance matrix: first obtain z-score matrix. Then use correlation matrix $Z^T Z$; PCA and SFA, and obtain low-rank (3-5) matrix approximations; add rank-1 matrices capturing the effect of a single factor. (2) Canonical correlation matrices: single effect, shared effect, etc. After obtaining initial U_k's, fit a mixture model (similar to MASH), that refines the estimates of U_k's. Note: U needs to be standardized before fitting MASH.

- Accounting for correlation due to sample overlap: for row j, $V_j = S_j C S_j$, where S_j is the diagonal matrix of standard errors, and C is the average correlation matrix of null SNPs.
- Model fitting and posterior inference: only performed on the input matrix data. Estimation of π's. Results are summarized: (1) lfsr for each row; (2) posterior effect estimates; (3) BF testing global null.
- Application to GTEx data: 16K genes, for each gene, choose the top SNP, defined by maximum Z scores across all tissues.
- Pattern of sharing and visualization (Figure 3): to visualize U_k (44×44), obtain the first few eigenvectors, which shows the loading to each tissue. In GTEx, the largest pattern (34% weight) show shared effects, with stronger effect sharing in brain.
- Examples (Figure 4): MASH can shrink the effect when there is little signal - shown by smaller posterior intervals.
- Gain of power by MASH: comparison with MASH-bmalite (a simpler version models only effect sharing - Flute et al.) and ASH. ASH: 13% with lfdr < 0.05 and MASH 47%.
- Sharing by sign and sharing by magnitude: sharing by sign is common, about 85%; but sharing by magnitude (less than 2 fold difference of effects) is only 30%.

7.1.1 Context-Specific eQTLs

Determining the loci of responses (response-QTL) [personal notes]:

- Problem: suppose we measure gene expression in two conditions, one untreated, the other some treatment. And we also have genotypes, we want to determine the loci that modify how cells respond to treatment (in terms of transcriptional change).
- Three strategies:
  - Strategy 1: differential eQTL. eQTL in one condition, but not the other; or different effect sizes.
  - Strategy 2: response eQTL. Treat the change of expression as new trait, and do QTL.
  - Strategy 3: gene-environment interactions, where environment is the treatment.
- Equivalence of the three approaches. We start with differential eQTL:

$$y_1 = \beta_1 G + \epsilon \quad y_2 = \beta_2 G + \epsilon \tag{7.12}$$

  So the difference $\Delta y = y_2 - y_1 = (\beta_2 - \beta_1)G + \epsilon$, a locus is differential eQTL ($\beta_1 \neq \beta_2$) if and only if it is associated with the response Δy. For strategy 3, we write it as:

$$y_1 = \beta G + \epsilon \quad y_2 = \beta G + \gamma G \times E + \epsilon \tag{7.13}$$

  So G is a differential eQTL if and only if $\gamma \neq 0$.
- The difference of these strategies lie in: (1) whether samples need to be matched: for the response eQTL approach, the samples need to be matched (no treatment vs. treatment). When this is the case, response eQTL is preferred, since it removes other sources of variations. (2) Whether include treatment as a covarite in the regression model.

Accounting for sample relatedness in response eQTL mapping [personal notes]:

- Ref: [Knowles and Gilad, eLife, 2018]

- Model: it is common in such experiments that the same individual is sampled multiple times (over treatment/time points). This leads to individual effects, which should be treated as random effects, and accounted for. Let y_{ij} be the expression (of a gene) of individual i, $1 \leq i \leq m$ in condition j, $1 \leq j \leq J$. It has fixed effects from genotypes (ignoring other fixed effects), which may vary across conditions, random effects from genetic background (which vary across conditions), and from individual non-genetic effect (a single effect shared across conditions). Our model is:

$$y_{ij} = v_j + \beta_j x_i + u_i + \xi_{ij} + \epsilon_{ij} \tag{7.14}$$

  where v_j is the mean expression across all individuals in condition j, x_i is genotype and β genetic effect. $u_i \sim N(0, \sigma_u^2)$ is the random, non-genetic effect from individual i, and $\xi \sim N(0, \sigma_\xi^2 K)$ is from genetic background (i.e. all other SNPs) with K being GRM, and $\epsilon_{ij} \sim N(0, \sigma_e^2)$. We can think of u_i as some special property of individual i, which is not genetically determined. Normally, in eQTL mapping, it would be treated as noised, but in experiments with repeated measurements of the same individual, it can be learned.

- Analysis: why it is important to model random effects u_i? Expression of gene in an individual i may be high and has nothing to do with genetics (e.g. epigenetic, stress, etc.). Not including this effect will not affect FP rates, since it is not correlated with genotypes. However, it adds substantial noise. With repeated experiments, it is possible to learn such individual effects by using multiple measurements. Accounting for such effects is similar to consider only responses in eQTL analysis.

- Model simplification: in practice, we are testing many SNPs for each gene. The u_i term is shared across all SNPs tested. So a practical strategy is to learn the random effects u_i, or σ_u^2 using all SNPs (adding genetic effects and kinship) only once, ignoring $\beta_j x_i$ term above. We can then write the model as $y_{ij} = v_j + \epsilon_{ij}^*$, with $\epsilon^* \sim N(0, V)$. Note V is $N \times N$ matrix, where $N = mJ$ is number of samples. Then in testing individual SNPs, account for the u_i and ξ_{ij} term:

$$y_{ij} = v_j + \beta_j x_i + \epsilon_{ij}^* \tag{7.15}$$

  where $\epsilon_{ij}^* \sim N(0, V)$.

A dynamic model for genome-wide association studies [Das & Wu, Hum Genet, 2011]:

- Motivation: suppose we are testing the time-series of a complex trait. The naive approach is, for each SNP, we test its association with the trait at each time point. However, this may lose power, as a SNP may influence one or more time points.

- Varying coefficient model: given a SNP to test, let $a(t)$ and $d(t)$ be the effect of the SNP (additive and dominant). The null hypothesis is $\forall t, a(t) = d(t) = 0$. Assuming $a(t)$ and $d(t)$ can be modeled as sum of polynomial functions (splines), fit the coefficients of the splines.

- Remark:
  - The idea of varying coefficient model: some coefficents of a complex model may be related in some way, in particular, some kind of continuity. Using non-parameteric model or splines to model the coefficients.
  - Comparison with Lasso: e.g. something like fusion penalty on the coefficients of the same SNP on the trait and different time points. Both approaches could acheive some type of "smoothness" (i.e. the coefficients of adjacent time points are similar). However, there are differences, e.g. Lasso forces sparsity but varying coefficient model not; Lasso, the "smoothness" is local, without any global trend; varying coefficient model with splines may not handle the case where the SNP affects a few unrelated time points, but not the others, etc.

Temporal Genetic Association and Temporal Genetic Causality Methods for Dissecting Complex Networks [Lin and Zhu, review for NC, 2017]

- Background: Granger causality, to infer $X \to Y$, show that prediction of Y_t can be improved by using X at earlier time points.
- Mapping dynamic eQTL (MPTGA): fit time-series expression data with a third-degree polynomial. A SNP is eQTL if the coefficients under the major allele is different from the coefficients under the minor allele. Do LRT: three coefficients (null model) vs. 6 coefficients (full model). Further extension: autocorrelation of expression across adjacent time points, ρ for the nearest time point, and ρ^2 for time difference of two, etc.
- Inferring causal model with Granger causality (TGCT): extend LCMS. To assess the model evidence of $M \to X \to Y$, use auto-correlation model for X and Granger model for Y:

$$X_{i,t} = \alpha_0 + \alpha_1 X_{i,t-1} \qquad Y_{i,t} = \beta_0 + \beta_1 Y_{i,t-1} + \beta_2 X_{i,t-1} \tag{7.16}$$

  The paper allows the coefficients α_0 and α_1 to be different under different genotypes. Apply the model to cis-trans gene pairs.
- Detecting cis-gene of teQTL hotspots: for a candidate gene, test only three models, all sharing $M \to X$. For these three models, the only difference lies in the distribution of Y, which depends on X on the previous time points. The candidate gene is assessed by the number of targets it causally regulates.
- Remark: MPTGA test suffers from potentially high d.o.f. TGCT: does not account for possible confounders (e.g. PCs).

Dynamic regulatory QTL mapping during differentiation [Ben Strober, NHS, 2017]

- Experiment: iPSC > cardiomycote, 14 lines, 16 time points (15 days). K-means clustering: 4 clusters of genes by expression dynamics.
- Per-time cis-eQTL mapping. WASP combined haplotype test. Empirical FDR by permutation (5) - variant-gene pair level. Replication in iPSC-eQTL: choose iPSC-eQTL from larger samples, and assess their p-values in differentiated cells.
- Correlation of eQTL effect sizes across time. This motivates a model with a small number of factors that vary over time.
- Factor analysis on eQTL effect sizes across time points: SNP on factors; factors vary over time (effect size of a SNP is linear combination of effect sizes of the SNP on factors). Possible interpretation: SNP effects on TFs, and expression is a linear combination of TFs (gene-specific).
- Mapping dynamic eQTL by GLM: Genotype, Time (encoded as continuous variable) and interaction term. Library size as covariates. Effect sizes may differ between different time points (interaction term). Problem: time effect is linear, and interaction is also linear.
- Cell line specific confounder: PCA on cell line x genes (and time) expression matrix.
- Permutation for controlling FDR: simulation under the null (no effect size difference across time). Concern: simulation does not capture non-linear time effects.
- Remark: possible explanation of dynamic eQTL, e.g. [TF] changes, their binding sites can be dynamic eQTL.
- HMM to infer differentiation states, then control for HMM states.Discussion: HMM not necessarily better than PC.
- Discussion: QTL mapping of shape, e.g. wavelet coefficients.
- Remark: cell heterogeneity. About 40% are cardiomyctes, the rest are other cell types.

- Lesson: challenges of context-specific QTL analysis in iPSC lines: (1) eQTL effects of a variant may change over time point, use interact terms. However, with many time points, this may lead to over-parameterization. (2) Different lines may differentiate at different rates; and cell type heterogeneity may differ between lines.Adjust for the difference using latent factors; using HMM states lead to similar results.

- Remark: Problem of dynamic eQTL mapping: one parameter for each time point - possible over-parameterization.May use factor-based eQTL to address the problem: learn eQTL effects on factors that represent a "pattern" of effects.

Response eQTL mapping and application in drug response [David Knowles, Stats seminar, 2017]; Determining the genetic basis of anthracycline-cardiotoxicity by molecular response QTL mapping in induced cardiomyocytes [Knowles and Gilad, eLife, 2018]

- Background: Anthracycline cardiotoxicity (ACT) of doxorubicin: common chemotherapy drug. GWAS of ACT response: <1000 subjects, one SNP.

- Experiment: 45 LCL-derived iPSCs, then cardiomycytes. Treatment with ACT at 5 different concentrations and do RNA-seq.

- Expression changes: 98% genes show DE. So fit a K-component mixture model for genes with Dirichlet prior. Learn dose-response curves of genes: different shapes.

- Likely model: initially drug induces (DNA) damage - damage response; then at high dosage, apoptosis.

- Response eQTL mapping: (1) Learn random effects: each sample is used multiple times (multiple concentrations), and this effect needs to be captured by random effects. Let y_{ncg} be expression of gene g of individual n with concentration c, it has several parts: treatment effect (fixed), latent factors, individual random effect, genetic effects (treated as random effects)

$$y_{ncg} = v_{cg} + \sum_k W_{kg} x_{nck} + u_{ng} + \xi_{ncg} + \epsilon_{ncg} \tag{7.17}$$

  where $u_{ng} \sim N(0, \sigma_u^2)$ is the individual random effect, and $\xi \sim N(0, \sigma_\xi^2 \Sigma)$ is the genetic effects and Σ is the kinship matrix, and $\epsilon \sim N(0, \sigma_e^2)$. To solve this model, integrate over W, u, ξ, ϵ and infer x and v and other parameters. This leads to MVN of $y_{:g}$ (data of a gene).
  (2) Testing individual SNP-gene pairs: test SNP effect on a gene while accounting for confounding by using the covariance matrix:

$$\Sigma_\pi = \Sigma_k \sigma_k^2 x_{:k} x_{k:}^T + \sigma_u^2 U + \sigma_\xi^2 \Sigma \tag{7.18}$$

  where U is the a matrix of which samples are for the same individual. Testing is done by comparing three models via LRT: $\mathrm{E}(y_{ncg}) = v_{cg}$ vs. $\mathrm{E}(y_{ncg}) = v_{cg} + \beta d_n$ vs. $\mathrm{E}(y_{ncg}) = v_{cg} + \beta_c d_n$, where d_n is the genotype of indivdiual n and β, β_c its effects.

- Inference: MOM, the covariance of y related to the parameters. Do EVD of covariance Σ_π, computationally efficient.

- Find 400 response eQTLs. Clustering dose-effect profiles of the SNPs.

- Higher enrichment of GWAS loci in response eQTL vs. eQTL only.

- Response ASE: ASE in resting and ASE at differenent concentrations.

- Model ASE: minor allele from Beta-Binomial, with mean logistic regression of β_c. Q: Whats β_c? Intuition: allele imbalance change with concentration. But why logistic regression, perhaps number of alternative alleles (in phased position)?

- Model of reQTL: TF binding? See enrichment of open chromatin.
- Remark: Testing response eQTL: different effects at different concentrations. High d.f. Alexis Battle: fit spline functions.

Shared Genetic Effects on Chromatin and Gene Expression Indicate a Role for Enhancer Priming in Immune Response [Alasoo and Gaffney, NG, 2018]

- Motivation: study response-eQTL, under their mechanisms. In particular, are response-eQTLs also response ca-QTLs? What are specific TFs driving response-eQTLs?
- Background: stimulation (IFN-gamma and pathogen) leads to signaling pathways and TFs: NF-kB, STAT2, IRF1.
- Models: how to explain stimulation-specific (response) eQTL? Figure 1a. chromatin effects are response-specific. Figure 1b. chromatin effects happen before stimulation, mediated by a pioneering factor, e.g. PU.1. In stimulation, TF such as IRF1, leads to eQTL.
- Experiment: macrophages, treated with pathogen, IFN-gamma and both. RNA-seq in 80 lines, and ATAC in 40 lines.
- Mapping eQTL and caQTL: on each of the four conditions separately. About 3K eQTLs and 20K caQTL regions. Enrichment of TF disruption: comparison with ASTB data of NFKB and STAT2 (Figure S5).
- Detecting condition-specific QTLs: $Y_i = G_i + E_i + G_i \times E_i$, where G_i is genotype and E_i treatment (four conditions). Also, for the same cell line, measured in four conditions, the errors are correlated across four conditions, so introduce a random effect term for each cell line. Results: 387 response eQTL in at least one condition. Cluster by eQTL patterns across four conditions (Figure 2a).
- Enhancer priming: focus on 145 caQTL-eQTL pairs that are likely driven by the same causal variants (lead variants LD > 0.8). For approximately half of the response eQTLs with a linked caQTL, the caQTL was present in naive cells before stimulation (Figure 2c). Most cases: same directions of effects. The reverse direction (eQTL before response caQTL) is much less, 15%. Enrichment of PU.1, CEBPa/b motif disruption in these cases.
- Regulation of multiple peaks by a single caQTL: use colocalization analysis, the fine-mapped SNPs same for multiple peaks. About 20% have master regions and dependent regions.
- Disease colocalization: 22 eQTLs, about half are in stimulated condition. 24 caQTL, most detected in naive cells. Most of these caQTLs are not eQTLs: explanation, they are often secondary eQTLs that are missed in the caQTL-eQTL colocalization analysis.
- Discussion: the implication of the study is chromatin + GWAS analysis may not find relevant cell states/types.
- Remark: under the priming model, gene expression is not activated, so the response eQTL is largely driven by expression changes (rather than change of genetic effects across conditions). This is not tested.
- Remark: the roles of stimulation-specific TFs, e.g. NF-kB and STAT2, are not investigated.
- Remark: not much joint analysis of caQTL and eQTL. In particular, the co-localization analysis (matching to the same SNP) considers only the strongest eQTL. This could miss other caQTLs that are weaker eQTLs.

7.1.2 Gene Networks and Trans-eQTL

Summary: methods and strategy for finding trans-regulatory relationship

- Joint learning of modules and genetic regulators: use regulators/genetic markers to guide the search of modules. Geronomo [Lee and Kohler, PNAS, 2006]. Bayesian Partition: expression of a gene is determined only by the genetic markers associated with that module.
- Association of genetic markers with factors: ICA paper [PLG, 2011].
- Searching for eQTL hot-spots: HESS.

Regulator finding in yeast [Bing & Hoeschele, Genetics, 2005]:

- Methods:
  - Data: [Brem, Science02].
  - eQTL confidence interval (CI): for each eQTL, need to determine the interval flanking the eQTL. Bootstrap resampling method.
  - To narrow down the genes in the eQTL CI: (1) fine mapping: if there are multiple markers in the CI, narrow down to significant pairs of markers; (2) gene selection using co-expression with the target gene.
- Results:
  - eQTL and CI: total of 570 (from [Brem02]) and 11 additional QTLs. The length of CIs: median 93kb, with 49 genes (from 0 to 717).
  - Candidate genes: in 65% eQTL regions, a single gene was retained as the candidate gene, for other regions, 0 - 6 genes (zero in less than 10% eQTLs). 45% are cis-regulation, and 55% are trans-regulation.
  - Transcriptional network: (Figure 4) overrepresented genes include: protein synthesis, aerobic respiration, transporter activity, lipid metabolism (Ura3, enzyme), pheromene response (MAT-α), cytokinesis during cell separation (AMN1, protein). Only 26 TFs in the network, and most of them have a regulatory role in just one other gene.

Causal genes and pathways [Tu & Sun, Bioinfo, 2006]:

- Problem: given eQTL data, find the causal genes (in eQTL region), and the pathways linking the causal genes to the target through TFs.
- Method:
  - Network: available from PPI, protein phosphorylation and TF-DNA interaction.
  - Search for the pathway from causal gene (unknown) to some TF (through intermediates) and to the target gene (TF-gene link), according to: 1) the causal gene should be in eQTL regions (i.e. a candidate list of causal genes is available); 2) the genes within the pathway should have correlated expression with each other.
  - Search algorithm: stochastic search, start from one TF, and at each step, move to an edge, according to the correlation coefficient at that edge. The algorithm stops if one candidate gene is reached.

Structural model analysis of multiple quantitative traits [Li & Churchill, PG, 2006]:

- Motivation: in QTL mapping, the traits are often correlated, and want to infer the genetic architecture of all traits: the QTLs common to all traits, and QTLs specific to individual traits, while taking into account the trait correlations.

- Background: SEM is a hierarchy of regression relationship among variables, similar to Bayesian networks.
- Methods:
  - Model selection problem: suppose Q is one locus, A and B are two traits, the possible models include, e.g. $Q \to A \to B$; $Q \to A, Q \to B$; etc. The goal is to choose the best model.
  - Initial model construction: first do single locus analysis on each individual traits:

    $$Y = \beta_0 + \beta_1 Q + \epsilon \tag{7.19}$$

    where Y is the trait vector, Q genotype vector, β_0 population mean, β_1 the effect of Q. Then identify pleotropic QTLs: suppose Q is a QTL of X, and we want to know if Q is also a QTL of Y, aware that Y may be correlated to X. The idea is: conditioned on X, can the rest of variations of Y explained by Q. This is done through a regression with additional variable X:

    $$Y = \beta_0 + \beta_1 Q + \beta_2 X + \epsilon \tag{7.20}$$

    If this results in large change of LOD (the effect of Q on Y), then X is causally connected to Q and Y (we need X to explain Y variation).
  - Model refinement: testing significance of coefficients.
- Question:
  - The effect of multiple QTLs on the same trait?
  - The relationship between two traits will not change when analyzing different loci, how is this modeled?

eQED: interpreting eQTL associations using protein networks [Suthram & Ideker, MSB, 2008]:

- Motivation: from eQTL data, find the related genes (the loci often contain multiple genes), and identify the regulatory pathways. The method by [Tu & Sun, Bioinfo06] suffers from the "dead ends", i.e. the search often cannot reach the candidate genes.
- Model: from a physical network, find the path from the source gene (reside in eQTL) to the target gene. Formulate as the electric circuit problem.
  - Sinlge locus model: find the causal genes within a eQTL region (muliple ones). Apply voltage at the eQTL, determined by the strength of that eQTL, (connected to all candidate genes) (the voltage at the target is supposed to be 0), and then currents flow along the network according to the conductance of edges (reliability). The problem is to determine the flow across each candidate gene, and choose the maximal one. Equivalent to random walk model, where source is one candidate and sink is the target.
  - Multi-locus model: apply voltages at multiple eQTLs, and calculate currents flowing across each candidate gene.
- Results:
  - Assess the prediction of casual-target associations: using genetic perturbation data to create a "gold standard".
  - Assess the direction of the edges: whether the current predominantly flows in one direction.
  - Prediction of regulatory pathways: the shortest route with highest total sum of currents across its interactions.

Mouse liver eQTL analysis with modules [Lan & Attie, PG, 2006]:

- Methods:
  - Data and analysis: 60 F_2 mice segregating for obesity and diabetes. 45,000 expression traits in liver were measured. Analysis with the standard interval mapping. The significance threshold is chosen as LOD of 3.4 or higher, which corresponds to FDR of 0.48 (weak threshold used).
  - Module identification: through both correlation across 60 individuals (in genetic dimension) and GO enrichment. From the list of genes with significant eQTLs (seed genes), find transcripts with Pearson CC of 0.7 or higher, then test for GO enrichment. A list of genes, starting from some seed transcripts, with strong GO enrichment will be chosen for analysis.
- Results:
  - QTL detection: with LOD of 3.4 or higher, found 6,016 transcripts with at least one eQTL. Among these, only 723 (best eQTL) were classified as cis- and the rest trans-. There are 15 regions (hotspots). When all eQTLs that have maximum LOD positions (significant or not) were considered, the regions clearly have GO enrichment.
  - Module analysis: from 6,016 seed transcripts, found 1,341 lists enriched for at least one GO term. The lists are combined to form 862 unique non-redundant lists each corresponding to one GO term.
  - Electron transport chain: none of 24 expression traits of ETC exceed LOD of 3.4, but many of them shared the linkage peak on chr. 2.
  - GPCR module: using 38 seeds identifies 174 genes correlated with these seeds, and they are all related to GPCR signaling. The co-regulation of these 174 genes can be verified. The eQTLs of these genes are clustered in 3 regions: Chr. 2, 10 or 13. Especially in the region of 10cm in Chr. 2: 50 genes of the module had major eQTL peak, and 81 had a sceondary peak.
  - Scd1 module: genes correlated with Scd1 (choose top 20), highly enriched for lipid metablism genes, and map to the same locations.

Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification (Geronemo) [Lee and Koller, PNAS, 2006]:

- Motivation: two ideas that could improve the single locus-single gene eQTL analysis:
  - Grouping genes to enhance the signal, assuming a gene group is affected by common SNPs.
  - Correlations among gene expressions, in particular, correlations between regulators and other genes.
- Geronemo begins by partitioning genes into modules with similar expression profiles. It then iterates over two steps: (1) learning a regulatory program (trans-G and trans-E regulators) for each module and (2) reassigning each gene to the module whose regulation program provides the best prediction for the gene's expression profile. The expression level (average of the model) is a function of all regulators (regression tree).
- Methods:
  - Data: [Brem, PNAS05]. Candidate regulators: TFs, kinases and phosphatases, chromatin modification factors, RNA factors (degradation and RNA processing).
  - Extensions of Module Networks: the number of modules is not fixed, and a single gene can be "broken off" from the module; a FDR permutation test when determining a split in learning the regression tree, a prior distribution that impose sparsity on the number of regulators and the number of targets per regulator.
- Results:

  - 79 modules containing at least 3 genes, spanning both trans-E (71 of 79) and trans-G (45 of 79) regulation. Adding trans-E regulation significant improves the proportation of genetic variance (PGV) of expression explained: explaining $> 50\%$ PGV for 828 genes, vs 238 in the original paper.
  - Zap1 module: a regulatory program of 10 genes. The regulators include: Zap1-E, Zap1-G (SNP), Gcr1-E and Puf4-E. With the standard eQTL analysis: only two genes of the module were found to be regulated by Zap1-G.
  - A large number of modules are probably controlled by chromatin regulation: (1) consective genes along the chromosomes; (2) in particular chromatin domains, e.g. telomeres; (3) enriched for targets of chromatin modifiying proteins. Ex. a telomere module (42 genes of which 40 are in the telomeres): the top regulator is a locus containing RIF2, and the transcriptional regulator Swi.

- Remark: the method is not designed for identifying trans- eQTLs. The trans- regulators may dominate the signals, and the effects of cis-eQTL on individual genes were ignored (which may be important for mapping trans- eQTLs, as they tend to have much smaller effect than cis-eQTLs). Also, the statistical test of the significance of individual regulators/eQTLs is not provided.

eQTL module mediated by TF activities [Sun & Li, Bioinfo, 2007]:

- Motivation: what is the mechanism of eQTL hotspots that are linked to multiple genes? One hypothesis is that some genes in the hotspot affects activities of TFs, which affect the target genes.
- Model: let M denote the DNA polymorphism, GC be the expression of gene cis-linked to the eQTL hotspot, GT be the expression of other genes linked to the hotspot, and TA be activity of some TF (estimated from the expression level of the known target genes of this TF). The goal is to compare two hypothesis:
  - Causal model: $(M \to GC) \to TA \to GT$.
  - Reactive model: $(M \to GC) \to GT \to TA$.

Harnessing naturally randomized transcription to infer regulatory relationships among genes [Chen & Storey, GB, 2007]:

- Motivation:
  - Biologically, we desire causal networks, where changing of a regulator leads to changing of its target. However, in the case of environmental perturbations (or sinlge gene knockouts): the expression of regulators and targets are often correlated, instead of causal. E.g. all genes in a module will be co-expressed under stimulations.
  - Possible to resolve the causality in eQTL data: genetic perturbations may randomize the expression level of one gene, and the effect on another gene can be tested.
- Methods:
  - Idea: view expression of a transcript as a random variable, determine if the transcript T_i causally influences T_j through random perturbation at the eQTL of T_i. Basically, if $T_i \to T_j$, then the locus of T_i should also be linked to T_j, and the causal influence of T_i on T_j can be tested.
  - Theorem: the causal relationship $L \to T_i \to T_j$ exists and there are no hidden variables causal for both T_i and T_j if and only if the following three conditions hold: $L \to T_i$, $L \to T_j$, and $L \perp T_j | T_i$. The last condition expresses: the causal effect from L on T_j can entirely be captured by T_i, thus no hidden variable.
  - Probability computation: apply the three conditions

$$P(L \to T_i \to T_j) = P(L \to T_i) P(L \to T_j | L \to T_i) P(L_i \perp T_j | L \to T_i \text{ and } L \to T_j) \tag{7.21}$$

    Each probability is computed by: LRT on the corresponding hypothesis, and then apply FDR to make the null statistics probabilities.

- Data: [Brem05] data, and for simplicity, only consider cis-linkage in regulators, i.e. $L_i \to T_i \to T_j$, where T_i is cis-linked to L_i.

- Results:
  - At probability threshold 90%, found 4,394 significant regulatory relationship among 2,145 genes where 127 are regulators.
  - Four regulators: (1) NAM9 (a mitochordrial robosomal component) → same or similar pathway; (2) CNS1 (co-chaperon) → transferase, and ribosome biogenesis; (3) ILV6 (enzyme in AA biosynthesis) → AA biosynthesis pathways; (4) SAL1 (mitochondrial transporter) → mitochondrial and member genes.

- Remark:
  - The regulatory links inferred in this manner reflects influences, not necessarily correspond to regulatory mechanism. E.g. changing expression of one gene, X, involved in RNA metabolism, the transcripts of many genes will be changed, but we would not think X as a regulator of the affected genes. In general, the transcriptional networks, signaling networks and metabolic networks are integrated, thus metabolites, signal sensing etc. all could have influences on gene expression: regulatory networks function to serve the stability of the metabolic networks, thus changing metabolites/external signals will induce regulatory networks, which change expression of relevant genes.
  - The QTL structure already implies causality: if T_i is cis-linked, and $L_i \to T_j$, then it must be $T_i \to T_j$.
  - The assumption that one locus control the transcript level is unrealistic: (1) there could be multiple loci controlling T_i, and (2) the dependence among loci may be important. E.g. suppose $L_i \to T_i \to T_j$, and $L_k \to T_j$, but L_k does not control T_i, then T_j will not be independent from L_i conditioned on T_i (as L_i will have information of L_k and L_k may influence T_j).
  - Two extra causal influences overlooked in this study: (1) $L_i \to T_j$: direct influence of L_i on T_j, e.x. through other genes (especially if L_i is an eQTL hotspot). With this link, the CI test does not hold. (2) $L_j \to T_j$: the QTL of T_j itself, modeling this should increase the power (instead of treating T_j as completely random once conditioned on T_i).

Using genetic markers to orient the edges in quantitative trait networks: the NEO software [Aten & Horvath, BMC Sys Biol, 2008]

- Common pleiotropic causal anchor (CPA) model: we want to test the causal relation between two correlated traits A and B. Suppose M_A represents the QTL of the trait A, five possible models: M_1 : $M_A \to A \to B$, $M_2 : M_A \to B \to A$, $M_3 : A \leftarrow M_A \to B$, $M_4 : M_A \to A \leftarrow B$, $M_5 : M_A \to B \leftarrow A$. The model M_1 has the following characteristics:
  - $M_A \to B$: i.e. M_A are common pleiotropic markers of both A and B. This is different from models M_2, M_4 and M_5.
  - $M_A \perp B|A$: this is different from M_2 and M_3.
  - $A \perp B|M_A$: this is different from M_3.

- Orthogonal causal anchor (OCA) model: utilizing the markers of B, M_B. Four possible models: $M_1 : M_A \to A \to B \leftarrow M_B$, $M_2 : M_A \to A \leftarrow B \leftarrow M_B$, $M_3 : M_A \to (A, B), M_B \to (A, B)$, $M_4 : M_A \to A \leftarrow C \to B \leftarrow M_B$. The model M_1 has the following characteristics:
  - Each of M_A has a pleiotropic effect on both A and B. Different from M_2 and M_4.
  - $A \perp M_B$: thus the name orthogonal causal anchor. Different from M_2 and M_3.

- $M_A \perp B|A$: different from M_3.

- Correlation-based tests: formulate the intuitions above using correlation and partial correlation. Ex. $M_A \to B$ is equivalent to $\text{Cor}(M_A, B) \neq 0$, $A \perp B|M_A$ is equivalent to $\text{Cor}(A, B|M_A) = 0$. To test if correlation is equal to 0, convert to Z scores, which should follow normal distribution if the correlation is indeed 0. In general, any causal model implies some relations of correlations/partial correlations, e.g. for $M_A \to A \to B$, we have:

$$\text{Cor}(M_A, B) = \text{Cor}(M_A, A)\text{Cor}(A, B) \tag{7.22}$$

  Thus if M_A is correlated to A, and A to B, then M_A to B; furthermore, $\text{Cor}(M_A, B)$ would be smaller than $\text{Cor}(M_A, A)$.

- SEM tests: combine all the evidence of a model in a SEM framework. Formalize the idea that a causal implies a correlation structure.

  - For each model, test a goodness-of-fit between the sample covariance matrix S, and the predicted covariance matrix based on the model, $\Sigma(\theta)$, where θ is model parameters. Since θ is unknown, need to replace with its MLE.
  - To establish a model, say M_1, compute the p-value of M_1 and all alternative models, the LEO score (local edge orienting) is based on the ratio of the p-value of M_1 and the p-value of the next best model.

- Remark: the selection of markers M_A and M_B may be important. Ex. under OCA, $M_A \to B$ if $A \to B$, but not under $B \to A$. However, if the later is true, $M_B \to A$, and one may select a marker of B as a marker of A, and it will influence both A and B.

Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks [Zhu and Schadt, NG, 2008]:

- Methods:

  - Use eQTL data to identify casual regulatory relationship: consider two expression traits, T_1 and T_2, if both map to the same QTL L, then determine if $L \to T_1 \to T_2$ or $L \to T_2 \to T_1$, similar to [Schadt and Lusis, NG, 2005].
  - Integrative reconstruction: use TFBS data (ChIP-chip plus conservation), eQTL data and PPI data (if at least half of gene in the complex contains a given TFBS, then all genes will be considered under control of that TF) as prior to construct Bayesian Networks (using expression data, similar to [Friedman, JCB, 2000]).

Lirnet: Learning a prior on regulatory potential from eQTL data [Lee & Koller, PG, 2009]:

- Methods:

  - Regression: (module network) suppose there are n regulators (both markers, or G-regulators, and trans- regulatory proteins, or E-regulators), and for any module m, the expression of its member gene g is:

$$y_{mg} = w_{m1}x_1 + w_{m2}x_2 + \cdots + w_{mn}x_n + \epsilon \tag{7.23}$$

  - Prior: (1) the G-regulators, its regulatory potential depends on the chromosome region, distance to gene, conservation, etc.; (2) E-regulators: the function of genes, etc. The prior is a sigmoid function of all features.
  - Inference: Lasso-type of regression, with SSE (for any module, sum over all genes: since the parameters only depend on the module and regulators, thus all genes would have the same values), L_2 regularization, prior distribution.

- Remark:
  - The model multiplicity problem is addressed by: (1) prior of the features: regulatory potential; (2) clustering of genes (response variables); (3) Lasso regression to reduce number of parameters.
  - Module-level regression: all genes share the same regulators (and parameters), thus there is only one predicted expression of any gene in a module (understood as module average). Thus SSE effectively measures the total distance from all genes to the module mean, summing over all modules.
  - Question: with prior distribution already specified, what is the interpretation of L_2 regularization (which is normally a prior)?

GFlasso: association analysis of a quantitative trait network [Kim & Xing, Bioinfo, 2009]:

- Motivation: some loci may affect multiple phenotypic traits simultaneously (e.g. a eQTL hotspot). Thus we would expect the same loci controlling multiple correlated traits.
- Background: the methods for multiple trait analysis:
  - Find loci that influence all phenotypes jointly.
  - Dimensionality reduction (e.g. PCA) on the traits, and apply linkage/association analysis on the new trait.
- Methods:
  - Data: N individuals, each individual has J SNPs, with value 0, 1 or 2 (the number of minor allels), denoted as $x_{ij}, 1 \le i \le N; 1 \le j \le J$; and has K phenotypes, denoted as $y_{ik}, 1 \le k \le K$. Also assume the correlation graph of the phenotypes is available, where r_{ml} is the correlation coefficient of two nodes (phenotypes) m and l.
  - Model: let β_{jk} be the regression coeffcient of the j-th predictor (SNP) on the k-th trait. The idea is: β_{jk} should be sparse; and the correlated traits should have similar values of β. The first term is encoded as a lasso penalty, and the second a fusion penalty for not having the same weights (β) on two correlated traits. Formally, the penalties are:

$$\lambda \sum_k \sum_j |\beta_{jk}| + \gamma \sum_{(m,l)\in E} f(r_{ml}) \sum_j |\beta_{jm} - \text{sign}(r_{ml})\beta_{jl}| \tag{7.24}$$

    where (m, l) denotes an edge in the correlation graph, and r_{ml} is the correlation and $f(r_{ml})$ is some monotic function of r_{ml}, e.g. $f(r) = 1$ (unweighted) or $f(r) = |r|$.
  - Model fitting: the objective function and constrains are convex, could use quadratic programming.

QTLnet: Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes [Chaibub-Neto & Yandell, Annals of Applied Stat, 2009]

- Model selection by conditional independence (CI): establish the model $Q \to X \to Y$ through the score:

$$\text{LOD}(y, q|x) = \text{LOD}(y, q, x) - \text{LOD}(y, x) = \log_{10} \frac{f(y|q, x)}{f(y)} - \log_{10} \frac{f(y|x)}{f(y)} \tag{7.25}$$

  i.e. the additional explanatory power of Q on Y, beyond that of X.
- Bayesian network: search for models that explain the trait network. Model averaging to obtain the posterior probability of each edge.

ReL (Regulatory Linkage) analysis [Gat-Viks & Shamir, PG, 2010]:

- Motivation: identify groups of genes regulated by the same eQTLs. These groups should have similar expression patterns (in other conditions).
- Idea: the groups of gene should: (1) have similar expression profiles; and (2) linked to the same eQTL interval. Multiplying the conditions (1) and (2), we have the eQTLs should be linked to conditions where the group show characteristic patterns. (Thus avoiding hard-grouping of genes before the analysis).
- Methods:
  - Data: 112 segregants, gene expression profiles and a compendium of expression profiles from perturbations of regulators (including TFs).
  - ReL score: linkage between a eQTL and a regulatory signature (the expression profile under one condition: up- and down- patterns). A high ReL score if the genes linked to the eQTL have different expression than the rest of genes.
  - Grouping regulatory signatures, corresponding to eQTL intervals: biclustering of eQTL-signature matrix. Then identify: (1) the target genes, those linked to the eQTL interval; (2) causal regulators in the eQTL interval by other information: PPI, protein binding to promoter and same process. In addition, the regulatory proteins of the ReL modules can be identified as the source of perturbations.
- Results:
  - Found 13 high-scoring ReL modules, covering 281 genes, 311 genetic markers and 82 regulatory proteins.
  - Uracil biosynthesis module: UAR3 (causal regulator), Ppr1 (regulatory protein). Possible mechanism: UAR3 mutation affects the uracil production rate, which negatively regulates the TF Ppr1.
  - Middle sporulation module: RFM1 (causal regulator), Hst1/Sum1 (regulatory proteins). Possible mechanism: RFM1 is a specificity factor that directs Hst1 (HDAC) to some of the promoters regulated by Sum1p.
  - Oxidative phosphorylation module: Cat5 and Crd1 (causal regulators), Swi3 (regulatory protein). Cat5 is required for ubiquinono biosynthesis and Crd1 for some lipid in the mitochondrial memberane; Swi3 is a subunit of the SWI/SNF chromatin remodeling complex.

A Bayesian Partition Method for Detecting Pleiotropic and Epistatic eQTL Modules [Zhang & Liu, PLCB, 2010]:

- Idea: partition genes into modules s.t. expression of each module can be explained by a small set of markers.
- Methods:
  - Model: partition of the genes and the markers, given the partition, the expression of a gene, g, in the i-th sample, in the module d, is given by:

$$y_{ig} = \delta_d(x_i) + r_i + \alpha_g + \epsilon_{ig} \tag{7.26}$$

    where r_i and α_g are sample and gene effects, respectively, and $\delta_d(x_i)$ is the function of the genotype of the i-th individual of the markers associated with d. This term is the genotype combination of the markers, e.g., if d is associated with 1 markers, it has three possible values; and if d is associated with 2 markers, it has 9 possible values; etc (possible epistasis is thus modeled).

- Results: identified 29 modules, 20 are linked to a single marker, and the rest to two markers, three of which have significant epistatic interactions. The modules are generally reasonable, most enriched with relatively specific GO terms.
- Remark:
  - An example of unsupervised learning, where the partition is helped by the demand that genes in the same module should be explained by the corresponding markers. Similar to Module Networks, where partition is helped by: genes in the same module explained by the regulators. The genes are clustered not according to their expression pattern directly, but by whether they are correlated with a common set of markers.
  - Limitations: the QTL effect on individual genes are not modeled (in particular, cis-eQTLs).

Cancer TRN by copy number and expression variation [Nordlander & Nelander, MSB-manuscript, 2010]:

- Idea: gene copy number variation provides natural variations of genotypes, and can be used to infer TRN.
- Model: gene-dosage model: the expression of a gene is determined by both the copy number of itself and of its regulators. Let ΔU be the copy numbers of a patient (for each gene), relative to some reference patient, and similarly, ΔY be the gene expression vector, relative the reference patient, we have:
$$S\Delta Y + \Delta U = 0 \tag{7.27}$$
where $S = W - V$, and $W = (w_{ij})$ is the gene-gene ineraction in mRNA synthesis and $V = (v_{ij})$ is the gene-gene interaction in mRNA degradation.
- Methods:
  - Model inference: the matrix S encodes the gene-gene interaction. It can be inferred by solving the optimization problem:
$$\min ||\Delta Y + S^{-1}\Delta U||^2_{Frobenius} + \lambda||S|| \tag{7.28}$$
The first term is a quadratic error term, and the second is the number of non-zero, non-diagonal elements in S (diagoal terms reflect the direct effect - more gene copy, higher expression, so will not be penalized), and λ is a tuning parameter favoring compact solutions.
  - Statistical confidence: 100 boostrapping - randomly select patients and infer S. Interactions present in 95% of simulations are reported as consensus network structure. (Because of λ, most terms in S will be zero, thus any non-zero term is considered a link in the network.)
- Results:
  - Construction of TRN in glioblastoma: first use correlation threshold (between copy number and expression) to filter genes, leading to 191 genes. Next construct the TRN, found 122 genes in the final network. A few pleiotropic regulators, EGFP and PDGFRA (well-known), and NDN, etc.
  - Validation of the subnetwork controlled by PDGFRA and NDN (role in glioblastoma not clearly established): effect of NDN and PDGFRA perturbation in U343 glioblastoma-derived cell line.
    * NDN over-expression reduce the grow rate of U343 cells.
    * Testing specific predictions of the model: CPNE8 induction by NDN, KCNH8 induction by PDGFRA, FGF9 induction by PDGFRA (but not in the presence of over-expression of NDN).
- Remark:

– The main equation relies on the assumption that (on average) patients do not have specific transcription or degradation parameters, i.e. no other sources of variation except gene copy number difference (if $\Delta U = 0$ - no copy number variation, ΔY must be 0).

Bayesian detection of expression quantitative trait loci hot spots (HESS) [Bottolo and Richardson, Genetics, 2011]

- MOM (mixture over markers): each response is associated with 0 or 1 marker with probability p_j for marker j.
- BAYES: for response k we have $\mathrm{E}(y_k) = G\beta_k$, where β_{kj} for marker j follows spike-and-slab:

$$\beta_{kj} \sim (1-\omega_j)\delta_0 + \omega_j N(0, \sigma_j^2) \tag{7.29}$$

  Different markers may have different ω_j and σ_j. The ones with large ω_j are "hot-spots".

- Regression model: k-th response y_k, and genotype X, we have: $y_k = X\beta_k + \epsilon_k$, where $\epsilon_k \sim N(0, \sigma_k^2 I)$. Let $\Gamma = (\gamma_{kj})$ be the indicator of whether marker j affects response k. $\beta_{kj} = 0$ if $\gamma_{kj} = 0$. Once Γ is given, non-zero coefficients β_k follows G-prior:

$$\beta_k|\gamma_k, \tau, \sigma_k^2 \sim N(0, \sigma_k^2\tau(X_{\gamma_k}^T X_{\gamma_k})^{-1}) \tag{7.30}$$

  Background: G-prior is the conjugate prior of Bayesian regression. So the variance of the effect is scaled by error variance σ_k^2 and covariance structure determined by data, and depends on one parameter τ.

  – Remark: the prior is shared among SNPs acting on the same response k. In contrast, in PEER, the prior is shared for the impact of the same factor (reflecting importance of factor).

- Prior of indicators/configuration:

$$\gamma_{kj} \sim \mathrm{Bern}(\omega_{jk}) \qquad \omega_{jk} = \omega_k \times \rho_j \tag{7.31}$$

  So ρ_j is an indication of whether j is a hotspot, and ω_k is an indicator of heritability of gene k.

- Inference: the joint likelihood:

$$P(Y, \Gamma, \Omega, \tau|X) = P(Y|X, \Gamma, \tau)P(\Gamma|\Omega)P(\Omega)P(\tau) \tag{7.32}$$

  where $\Omega = \{\omega_{kj}\}$. Inference by Gibbs sampling, with MCMC on each part. MCMC on Γ (configurations): use evolutionary stochastic search: run with several chains at different temperature. So moves across chains: global move; and within chains: local moves to explore neighborhood. MCMC on parameters: adaptive proposal (tune step size) and fixed proposal.

- Application to human monocyte eQTL data (n =1400): IDIN subnetwork of 600 genes, run HESS with 200 SNPs spanning 1Mb in a chromosome region.
- Lesson: Bayesian hotspot model, G-prior for effect sizes; factorized prior for inclusion probability.
- **Remark**: the model relies on a small input set of genes. In applications, one first finds a gene group whose factor(s) is associated with a SNP (locus), then use HESS on this group of genes and SNPs in the locus.

Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. [Rotival and Blankenberg, PLG, 2011]

- Data: monocyte gene expression in 1,400 subjects.

- Preprocessing (Figure 1): (1) MDS to filter outlier samples. (2) SVD: determine the number of components. Plot PVE vs. PCs, also do the same plot of randomized data. Find the number of components beyond which PCs explain only noise (Figure S5).
- ICA: 112 modules. Filtering: 21 components attributable to single individuals; 27 components with kurtosis < 3. Final results: 64 components.
- Annotating signatures: define genes associated with each signature as those in extreme distribution, control FDR. $> 60\%$ modules show enriched GO terms.
- Association test of patterns and SNPs: do association test at $P < 10^{-7}$, then do enrichment test: SNPs need to be associated with a significant number of genes (in trans) at relaxed threshold $p < 10^{-5}$.
- Results: 11 associations. 4/11 are driven by cell type composition difference across samples. Some locus: associated with AIDs.
- Issues: cross-hybridization can lead to false trans-eQTL, cell type composition variation across subjects is a confounder (some are genetic).
- Comparison with WGCNA: (1) WGCNA founds about 20 modules, most are correlated with ICA signatures, however only 31% ICA signatures are correlated with WGCNA. (2) ICA much better in association tests (Figure 5).

Pathway-Based Factor Analysis of Gene Expression Data Produces Highly Heritable Phenotypes That Associate with Age [Brown and Durbin, G3, 2015]

- Data: skin eQTL data of MuTHER (twins, $n = 657$).
- Define pathway phenotypes: (1) Regress out global covariates using PEER: they explain 37% of variation. (2) For each of 186 KEGG pathways, do PEER analysis on residual expression, and select 5 factors. They explain a median of 17% of variance.
- Association of pathway phenotypes with age: 62 significant associations, comparing with 7 if using single gene association followed by DAVID enrichment test. Comparison with single-gene test: much more significant results.
- Top pathways: insulin signaling, fatty acid metabolism, xenibiotic metabolism, cancer-related pathways.
- Contribution of genetics and environment: single gene h2 about 0.13, pathway phenotypes 0.18, age-associated pathways 0.32, global factor 0.18. However, not found significant SNP association with pathway phenotypes.
- Discussion: high heritability of pathways, averaging out noise at single gene level. For global factors: genetics unlikely to have coordinated effects on a large number of genes, so global factors mostly like represent technical factors.

Tensor decomposition for multiple-tissue gene expression experiments [Hore and Marchini, NG, 2016]

- Idea: explain the variation of gene expression across individuals and tissues using hidden factors. Then GWAS of hidden factors to identify eQTL.
- Basic model for one tissue: let Y_{nl} be the expression of gene l on sample n. Suppose we have two components, e.g. [TF] or growth rate. Let A_{n1}, A_{n2} be the activities of these two component in sample n, we have:

$$Y_{nl} = A_{n1}X_{1l} + A_{n2}X_{2l} + \epsilon_{nl} \tag{7.33}$$

where X_{1l}, X_{2l} are the effects of the two components on expression of l. In general, we define X_{cl} be the *gene loading matrix*, which describes the effect of a component c on genes. The model assumes a sparse prior for loading matrix:

$$X_{cl} \sim p_{cl} N(0, \beta_c^{-1}) + (1 - p_{cl})\delta_0 \tag{7.34}$$

where p_{cl} is the prior probability of gene l loading in C. We further use a Beat prior for p_{cl}.

- Multi-tissue model: more generally, we have A_{ntc} as the activity of c in tissue t of sample n. Then we can express the expression of gene l in tissue t of sample n as:

$$Y_{nlt} = \sum_c A_{ntc} X_{cl} + \epsilon_{nlt} \tag{7.35}$$

  We further assume that the activity tensor can be factorized as: $A_{ntc} = A_{nc} B_{tc}$. The intuition is that: the activities of components are mostly determined by tissue profiles, but modified by indivduals. The multi-tissue model is better than analysis of all tissues separately because X_{cl} is shared across tissues.

- Finding eQTL: we apply the model to all tissue expression data. And then we have A_{nc} for each sample, and can do GWAS on each component. If a SNP is associated with c, and the gene loading for c contains two genes, then the SNP is eQTL of the two genes.

- Data: MuTHER eQTL data, LCL, skip and adipose, 845 related individuals.

- Results: > 200 components in total, most active in single tissues. 26 components have genetic basis. But 20 have very sparse gene loading and correspond to mainly cis-eQTL. The six trans-component identifies some trans-eQTL: none was found by ICA or PCA (Figure 2-6).

- Example, MHC Class II (Figure 2): factor is strongly associated with a locus, CIITA ($p < 10^{-10}$). Gene loading: top genes are HLA. Highly active in LCL.

- Other examples: MHC Class I (Figure 3), Histone genes (Figure 4), Type I interferon response (Figure 5), Zinc-finger gene network (Figure 6). Overall: most cases have 1 strong locus ($P < 10^{-8}$) except Zinc-finger, and usually no other loci $P < 10^{-6}$.

- Lessons:
  - The advantage of tensor formulation: if one tissue only, the components can be arbitrarily rotated (identifiability issue), but this can be solved with tensor.
  - Sparse gene loading allows easy interpretation of eQTL.
  - Most expression variations have no genetic basis.

- Remark/Questions:
  - The key assumption is that gene loading is shared across tissues. This is certainly not true, but somewhat reasonable: e.g. we can think about the TF-gene network topology is fixed (gene loading), but [TF] may vary across tissues (activity).
  - The trans-eQTL discovered due to shared information across genes or across tissues? Probably both. But a general question is that if trans-eQTL are not often shared across tissues, does it have any advantage of using multi-tissue model?
  - Some components are mostly cis, affecting a very small number of genes. How are they chosen by the model?
  - Can we integrate multiple components to discover eQTL of a gene? If a gene appears in multiple components, we may gain power in this way.

Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis (GMAC) [Fan Yang and Lin Chen, GR, 2016]

- Background: mediation analysis. Suppose we want to show that M is a mediator, $X \to M \to Y$. We need establish: (1) $X \to Y$; (2) $X \to M$; (3) $Y = \beta_1 X + \beta_2 M$, test if $\beta_2 = 0$. Note that conditions 1 and 2 are necessary, otherwise, we may have the model, $X \to Y \leftarrow M$. Also if we already establishes (2), then Sobel test means we only need to test if $\beta_2 = 0$.

- Model: suppose we test a trio $L_i \to C_i \to T_i$, where C_i and T_i are cis- and trans-genes. We may have confounders of C_i and T_i, and it would be better to adjust for them. The key idea is to adjust for confounders for each trio. However, common children and/or intermediate variables between C_i and T_i are correlated with C_i and T_i, and including them as confounders would create FPs (collider) or reduce power. Note that common children and intermediate variables are both associated with L_i, while confounders not.

- Procedure: (1) For each trio L_i, C_i, T_i, first filter those associated with L_i. (2) Test mediation: Wald test of the regression coefficient of C_i vs T_i conditioned on L_i. (3) Control for FDR by permutation: permute cis-expression within genotype groups.

- Application to adipose: start with 8K cis-eQTLs, then test trans-association of 8K SNPs vs. 27K transcripts. Found 3300 trans-associations at $p < 10^{-5}$ (expect about 1600). Then do mediation analysis on these 3300 trios: found 300 significant ones at FDR < 0.05.

- Results: increase power to detect mediations, from 3K to 6K, comparing with PEER factor adjustment. Also, 20% cis-genes mediate two or more trans-genes.

- Cis-hub analysis: 615 cis-genes that mediate multiple trans-genes. Total of 20 cis-genes mediate ≥ 4 trans-genes (Table 2).

- Lesson: adaptive adjustment of possible confounders. Distinguish confounders, children (colliders) and intermediate variables.

Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation [Brynedal and Costapas, AJHG, 2017]

- Data: About 300 LCL RNA-seq, in 3 populations.

- CMPA: cross-phenotype meta-analysis, applied to one SNP and all transcripts. Test on $-log10(p)$ (9000 p-values, one for each gene. Under H0: SNP not associated with any, it follows exponential distribution. Under H1: fit the exponential distribution with a free parameter λ. Then do MLE.

- Account for correlation using Z-scores: obtain the correlation of Z-scores of every two genes; then simulate Z-scores for all genes. Obtain empirical p-values for each SNP.

- CPMA power analysis: even though we test the distribution of all p-values, the power is OK if there are 50 genes associated with the SNP.

- Found no genome-wide significant SNPs. But show that the number of SNPs with low FDR is higher than expected.

- Test trans-eQTL associate to the same probes across populations. Top SNPs from CPMA: (1) Replication across populations. (2) Consistency of effect direction: across popluations.

- 62 SNPs have significant overlaps between all three samples. 1000 SNPs consistent in directions. 8 in common: super trans-eQTL.

- Functional enrichment: of genes of the same trans-eQTL. The 8 top trans-eQTL: each regulates hundreds of transcripts, 77-800 (Table 2). Histogram of p-values of top eQTL. Enrichment of TF targets (ENCODE ChIP-seq) in about 4/8 trans-eQTL.

- Trans-eQTL: the 8 trans-eQTL are correlated with PEER factors. If using SVA to remove confounding variables, some signal can be removed.

An independent component analysis confounding factor correction framework for identifying broad impact expression quantitative trait loci (CONFETI) [Ju and Mezey, PLCB, 2017]

- Background: main challenge is that some confounding factors are heritable. Current strategies: (1) Joint estimation of SNP effects on gene expression and confounding factors [PANAMA, Fusi, PLCB, 2012]. (2) Use a subset of genes to estimate non-genetic confounding factors [Joo and Eskin, GB, 2014].

- Identifying heritable ICs: $Y = AS$ where A is IC values of n samples and S is $k \times g$ matrix, weighting matrix. Once ICs are identified, do association analysis of ICs, and find genetic ICs with Bonf. correction.

- Analysis: factors identified by ICA may be more distinct, and thus facilitate downstream analysis, e.g. genetic association. Show an example: best factors associated with sex, ICA much better than PCA (Figure S1).

- Correcting for non-heritable ICs in testing eQTLs: LMM approach. Use the remaining ICs to compute sample correlation, then fit LMM (Figure 1).

- Analysis: why use LMM instead of fixed covariates? Number of parameters may be large using fixed covariates.

- Simulation: real genotype data from yeast. Broad-impact eQTL: SNP affecting 10% of genes. Covariates/ICs: sparse (30%) or dense (all).

- Results from simulation: For eQTL mapping, CONFETI similar to PANAMA, which estimate genetic effects (eQTL) and factors simultaneously. For broad-effect QTL: CONFETI significantly better.

- Human data: (1) Cis-eQTLs: CONFETI modest gain. (2) Trans-eQTLs: CONFETI modest gain. PCA and PEER do very poorly, even worse than linear model.

- Replication of broad-impact eQTLs: only in MuTHER LCL (twin pairs), but not between blood dataset (DGN and NESDA). Figure 5: a small number of such eQTLs, only a few targets.

- Remark: CONFETI has some advantage over factor models assuming normality: the effects of factors on genes are likely non-normal. However, this may be captured by sparse models.

Moving beyond cis-regulation [Xuanyao Liu, NHS, 2019]

- Ref: GBAT: a gene-based association method for robust trans-gene regulation detection [2018].

- I. Omni-genic model. Background: Serum urate levels: 180 loci. top 28 loci explain 5% h2. 183 explain 7.7%. Common variants: 23%. Pedigree h2: 38%.

- Variance partition for cis- and trans- effects: (1) core genes: cis-QTL. (2) core genes: trans-QTL. (3) Covariance of core genes. Estimates of contribution of each part depends on genetic correlation of gene expression: if it is high for cis-genes, trans-genes can explain $> 70\%$ heritability.

- II. Trans-eQTL mapping. Ex. in NTR data, average cis-heritability is 0.02 and trans is 0.045. Across studies, trans-eQTL explain 2/3 heritability.

- Challenges of trans-eQTL: (1) correct for confounders in trans-eQTL can lead to reduced power and FPs due to collider effects (both cis and trans- genes affect the confounder). (2) RNA-seq read mis-mapping to homologous regions: 75% FP rates.
- GBAT: correct bias in read mapping (30-40% reads). Predicting gene expression: BLUP, leave-one-sample-out cross validation (train weights in training and evaluate in leave-one-out sample).
- BLUP: BLUP predictions are correlated with errors. This lead to false correlation of predicted cis-gene and trans-gene. This is commonly addressed by cross-validated prediction model: i.e. train the BLUP estimator (BLUE) in all dataset except one, then make prediction in the remaining data. Use cvBLUP to implement a fast way of getting leave-one-out prediction.
- Association test: $Y \sim \tilde{X} + R$, where $\tilde{X}$ is the predicted cis-expression, and R is the surrogate variable (SV) residual (after removing X).
- Quantile normalization: $y = C+$ error, C is t-distribution. QN(y) leads to increase FP. Solution: regress out C first in y, then QN(y.residual).
- DGN: 400 trans associations in 157 unique regulators. Enrichment of TFs in regulators. Some control more than 3 genes, e.g. NFKBIA regulates 4 other genes, SRCAP regulates 88 genes.
- Disease-specific trans-eQTL: SCZ vs. control, trans- genes in SCZ are three times enriched with TWAS hits, not in control. Discussion: controls may have different cell compositions (consequence of disease states).
- III. Interaction testing. Crohns disease: common variants explain 20% heritability. Interaction can increase it to 60Epistasis: gene-based interactions (predicted gene expression).
- Remark: why need cross-validation? Unlike TWAS, here prediction model of gene expression, and its application are performed in the same dataset.
- Lesson: what variables to control for in testing trans-eQTL? Account for SVs can lead to collider bias, so obtain SVs, then regress out X (genotype or imputed expression), then adjust the residual.
- Lesson: when X is t-distribution, do quantile normalization can lead to increased FPs.
- Q: how to deal with LD and pleiotropy? The general challenge of TWAS.

Pathway-level information extractor (PLIER) for gene expression data [Mao and Chikina, NM, 2019]

- Model: let Y be expression matrix, $Y = ZB$, where Z is the gene to factor loading. The idea is to choose Z to align with known pathways/gene sets, this is done by another decomposition $Z = CU$. This can be interpreted as a model: factor $\rightarrow$ gene set $\rightarrow$ genes.
- Inference: constraint $\text{rank}(Z) = k$ and $\text{rank}(B) = k$, where k is number of factors. Optimization: $\|Y - ZB\|_F^2 + \lambda_1\|Z - CU\|^2 + \lambda_2\|B\|^2 + \lambda_3\|U\|_1^2$, with L_1 penalty for U.
- Hyperparameters: (1) k: significant number of PCs. (2) λ_1: cannot use cross-validation as reconstruction error is always minimized at $\lambda_1 = 0$. Default setting of λ_1 and λ_2. (3) λ_3 controls sparsity of U (how much LV is associated with given gene sets), optimize λ s.t. 70% of LVs are associated with gene sets.
- Evaluation of LVs: compute AUC and FDR. A LV is considered high confidence, if AUC > 0.7 and FDR < 0.05. AUC: by cross-validation, for a given pathway, remove a certain subset, and use the rest of genes for training. Then test if the resulting Z matrix (gene-LV loading) can recover the left-out genes vs. random genes.

- Results in cell type decomposition analysis: works as well as CIBERSORT, which is based on reference transcriptome data.
- Trans-eQTL analysis: all LVs vs. all SNPs, do BH FDR corrections. For the remaining ones at FDR < 0.05, Then do gene level association test, and filter pathways with low gene-level support.
- Results in DGN: found 86 LVs assocated with 300 pathways. 12 LVs associated with 10 unique SNPs. Some SNP is associated with LV enriched in genes related to megakaryocyte/platelet lineage. However, the SNP is not found in GWAS catalog.
- Q: how FDR of LVs is controlled?

7.1.3 Allele-Specific Expression and Epigenomics

Background: allelic imbalance and allele-specific expression (ASE) [personal notes]

- ASE due to cis-regulatory polymorphism: suppose we have a SNP in a regulatory sequence (called test SNP), and the allele A is associated with higher TF binding and the allele a lower binding. Now consider a region targeted by this regulatory element (called linked region)in one heterozygous individual (Aa), the haplotype of the linked region lined to A is expressed highly and the haplotype linked to a is expressed at a lower level. This leads to ASE in this individual.
- Individual-level ASE and cohort-level ASE:
  - At the individual level, any cis-regulatory polymorphism always leads to some kind of ASE in the gene it controls (it may not be detectable, e.g. when there is no exonic SNP). Note that at the individual level, the rQTL and the gene can be far, but the haplotype can potentially be resolved (for an individual, a haplotype is an entire chromosome).
  - At the cohort level, to have a consistent ASE (e.g. exonic allele B always expressed highly), the exonic locus should be in good LD with the cis-eQTL.
- Relating ASE to regulatory effects of SNPs: suppose we have one individual heterozygous of the regulatory SNP of interest, Aa. This SNP creates differential expression of the linked transcript. All exonic SNPs in this transcript would thus show ASE. Consider SNP j, suppose we know the exonic SNP linked to A (phasing, which is possible because of LD), the reads mapped to this allele x_j follows $x_j \sim \text{Bin}(n_j, p)$, where n_j is the read depth at j, and p the extent of allelic imbalance. Let β be the effect size of A: defined as log fold change, then we have:

  $$p = e^{\beta}/(e^{\beta} + 1) \tag{7.36}$$

  Note that: the ASE of the transcripts is shared across the entire length, thus all exonic SNPs. SNPs are generally sparse, and much larger than read length, thus the read depth of each SNP can be considered independent "reading" of the underlying ASE.
- Using ASE for mapping cis-eQTL:
  - Even though ASE can be detected with a single individual, we cannot resolve cis-eQTL in one individual because of many possible SNPs in the cis-region of a gene. We need to study multiple indidivuals to find the cis-SNP(s) with consistent ASE.
  - We don't have to require the LD between cis-eQTL and ASE: we only need to test, if one allele in cis-eQTL is always associated with higher expression in the linked region (it doesn't matter if the highly expressed haplotype is consistent across multiple ndividuals) [McVicker & Pritchard, Science, 2013].
  - Alternatively, we treat ASE as a quantitative trait, and test association of an allele in the test SNP with ASE level. It is called aseQTL [Battle et al, GR, 2013]

  - Advantage: the allelic imbalance is entirely due to difference in cis (thus we control all the other covariates that would be required to compare expression level across individuals). Furthermore, allele imbalance test can be combined with the usual way of mapping cis-rQTL.

- Using ASE or ASB/ASM/ASHM (allelic-specific binding/methylation/histone modification) for disease studies: study the effect of one SNP on the regulatory activity/gene expression.
  - Utilizing ASE at the individual level: A disease locus may be a cis-rQTL, thus creating ASE in the target gene. We can use the ASE to map the target gene of the cis-rQTL: test in heterozygous individual, if a test SNP (potential cis-rQTL) is always associated with higher expression of the test gene (at the individual level).
  - Utilizing ASE at the cohort level: a disease locus may be a cis-rQTL, test if there is a cohort-level ASE in LD with this rQTL (if so, likely the gene is the causal gene behind the cis-rQTL).
  - A cis-rQTL may create ASB/ASHM/ASM (or allele-specific chromatin accessibility) in that site: this effect can be mapped using allelic imbalance in the site. Ex. in ChIP-seq experiments, if one allele is associated with a higher level of histone modification, then this allele will be enriched in all reads mapped to this region.
  - Some example studies: open chromatin in islet cells [Gaulton & Ferrer, NG, 2010]

ASE lecture: `https://www.ebi.ac.uk/training/online/course/embo-practical-course-analysis-high-throughp allele-specific-expression-and-eqtl`

- Basic model: binomial test, $x \sim \text{Bin}(n, p)$, where $p = 0.5$ if no ASE, and n the number of reads, and x the reads mapped to the reference allele.
- Read mapping bias: a study [Li et al, GR, 2008], QQ plot, most loci show ASE comes from reference alleles. Summary: 90 ASE, 61 show over-representation in reference alleles.
- Addresing read mapping bias: do simulation under null (no ASE). Found 1% of SNPs show ASE (75% reads mapped to reference allele). Simple solution: masking the alleles during read mapping.
- eQTL mapping using RNA-seq [Li et al, GR, 2008]: Negative correlation between GC content and expression. Solution: regress out GC content. Procedure for eQTL: adjusting for GC (regression out), then quantile normalization. Then correct for latent confounders using fist 16 PCs

Research progress in allele-specific expression and its regulatory mechanisms [Gaur & Liu, J Applied Genet, 2013]

- ASE studies in human: how broad ASE is
  - In [Vidal, 2011] estimate that at least 25% of the human genes display ASE.
  - In [Serre, 2008], 4.6% of heterozygous single-nucleotide polymorphism (SNP)sample pairs have evidence of ASE.
  - In [Lee, 2013]: using exonic SNP chip in colecteral cell lines, found two monoallelically expressed genes (ERAP2 and MYLK4), 32 genes with an allelic imbalance in their expression, and 13 genes showing allele substitution by RNA editing. Among a total of 34 allelically expressed genes, 15 genes (44%) were associated with cis-acting eQTL.
- cis-regulatory mechanism of ASE:
  - Cis-regulatory polymorphism (e.g. change TF binding) in strong linkage disequilibrium with variants within a gene.
  - Possible interaction between cis- and trans-regulatory polymorphism: the exonic SNP allele expressed highly might be different in different individuals.

- Epigenetic regulation of ASE: more general than imprinting.
  - Allele-specific methylation (ASM): ASM is relatively widespread across the mammalian genome, both cis- and parent of origin (POD) in nature, and often heterogeneous across tissues and individuals.
  - Allele-specific histone modification (ASHM): ASHM are associated with various disorders, such as diabetes. The loci that display ASE are substantially enriched with ASHM. Important findings: allele-specific differences exist in TF binding and open chromatin, they have consequences on downstream events such as expression, and at least some proportion of these differences is due to heritable genetic variation.
- Non-coding RNA regulation of ASE: 3-UTR-SNPs, targeted by miRNA and associate with mRNA stability.
- Method for detecting ASE:
  - Allelic bias in read mapping: the reference genome contains only one of the possible alleles. Idea: use enhanced reference genome that has information of alternative alleles.
  - A new method: construction of two haplotypes and remapping of the reads against the diploid transcriptome.

A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data [Skelly & Akey, GR, 2011]

- Background: the standard test of ASE, at a SNP, suppose there are n reads, and x is mapped to one allele, then ASE is detected by the binomial test of $p = .5$.
- Idea: combine information of all exonic SNPs of a gene to test if the gene has ASE or not.
- Model: for the gene i and SNP j, let x_{ij} be the number of reads mapped to the allele and p_{ij} be a measure of imbalance. Then x_{ij} follows binomial distribution with parameter p_{ij}. Furthermore, p_{ij} follows Beta distribution α_i, β_i. Let $p_i = \alpha_i/(\alpha_i + \beta_i)$ be the prior mean, then if the gene has no ASE, $p_i = .5$; and if it has ASE, p_i is different. We use a mixture distribution, π_0 genes have no ASE and the rest have ASE.

Identification of Genetic Variants That Affect Histone Modifications in Human Cells [McVicker and Pritchard, Science, 2013]

- Background: Beta Negative-binomial model. If $X|p \sim \text{NB}(r, p)$, and $p \sim \text{Beta}(\alpha, \beta)$, then $X \sim BNB(r, \alpha, \beta)$.
- Method: uses both read depth and allelic imbalance to map cis-quantitative trait loci (QTLs) with small sample sizes. Consider a test SNP, and a linked region (the target of the SNP, not necessarily in LD because we are using individual level ASE data). To see if the test SNP is a rQTL, we can use: (1) Variation of expression across individuals: the genotype of the rQTL (one of three possible genotypes) is associated with the expression level (total) of the linked region - this is measured by the total read depth. (2) Variation of expression between two alleles within a heterozygous individual (at the rQTL): it will lead to ASE in the linked region (allele imbalance in reads). Ex. the alternative allele is always associated with higher reads in the linked region (does not matter what are the genotypes in the hetero SNPs in the linked region).
- Total read depth (total level of the linked region): Let G_i be the genotype at the test SNP of individual i, let X_i be the total read count in the linked region, then X_i depends on G_i. Specifically X_i follows Poisson distribution with rate λ_i, and λ_i depends on the total read depth T_i of individual i, and the genotype G_i: $\lambda_i = 2\alpha T_i$ if $G_i = 0$; $\lambda_i = (\alpha + \beta)T_i$ if $G_i = 1$; $\lambda_i = 2\beta T_i$ if $G_i = 2$.

- Overdispersion problem for read depth: compare the mean and variance of read counts, find overdispersion across individuals. Even if Negative binomial cannot correct for it. Use Beta Negative Binomial distribution. The assumption is that $\lambda_i \sim T_i\beta(G_i)u_i$ where $\beta(G_i)$ is the expected effect given genotype G_i, and u_i an individual-specific factor. The model amounts to a distribution of u_i (instead of constant).

- Allelic imblance: if the test SNP is rQTL and is heterozygous in an individual, it will create allelic imbalance in the test region. Suppose the alternative allele increases the binding/expression, then in the test region of this individual linked to this allele, we expect more reads in a heterozygous SNP (whether the genotype of that SNP is a reference or alternative allele for that SNP is irrelevant). Let Y_{ij} be the reads of individual i at SNP j (in linked region to the test SNP). Then Y_{ij} follows binomial distribution with probability p, which is .5 under the null model, but not under allelic imbalance:

$$Y_{ij} \sim \text{Binom}(n_{ij}, p) \tag{7.37}$$

  To test allelic imbalance, we use all SNPs of the linked region:

$$L(D|p) = \prod_{i,j} P(y_{ij}|n_{ij}, p) \tag{7.38}$$

  and do LRT.

- Overdispersion problem for allele imbalance: permutation shows that the model cannot fully account for individual variation. Not clear how permutation test is performed: likely permute the test SNPs vs. linked regions. Use Beta-Binomial distribution: the model is $Y_i \sim \text{BetaBin}(N_i, p, \theta)$, where θ is overdispersion parameter (to be estimated) and $p = \alpha/(\alpha + \beta)$. We test if $\alpha/\beta = 1$.

- Combined haplotype test: To relate to the total read depth model, we assume $p = \alpha/(\alpha + \beta)$, and have a combined likelihood to test.

- Motif break analysis: take SNPs within TFBSs (Centipede, Posterior > 0.99), then correlation of PWM changes with allelic imbalance in various histone marks. Merge similar motifs: clustering analysis, distance is defined by overlap of predicted TFBSs from Centipede. Testing association of TF binding with allele imbalance: LRT, the allele imbalance (2kb ChIP-seq reads) is a function of change of PWM scores.

- Results of motif break analysis (Figure 3): Overall positive correlation with H3K27ac, and negative with H3K4me1. Specific TFs: Found 11 TFs at FDR < 0.1. Most are activators, but NRSF is a repressor. Validation: using allele-specific TF binding/ChIP-seq data (LCL). Confirm most of the 11 TFs (Figure S11).

- Remark: it is unclear how the method models multiple linked SNPs in the allelic imbalance test. The likelihood function suggests that they are independent (but later dropped index j). However, many of these SNPs must be in close LD, and thus not independent. Ex. any SNPs in the same haplotype of an individual will have the same reads (or similar), so the information is completely redundant.

- **Analysis**: Why individual variation in allele imbalance? Even if the test SNP has no effect, we still need to model individual variation of p (expected value is .5). Possible confounder: in an individual, the allele ratio may be affected by other SNPs (not the test SNP) or other sources (e.g. imprinting).

QuASAR: Quantitative Allele Specific Analysis of Reads, [Harvey,Bioinformatics, 2014]

- Motivation: detecting ASE using only RNA-seq data (no genotype calls). Joint inference of genotypes and ASE.

- Notations: we have RNA-seq data of one individual over many SNPs/sites, but multiple samples (e.g. different conditions). Note that in different samples, genotypes are shared but ASE may be different. Our data are R_{sl} the reads of sample s and SNP l, mapped to reference allele, and A_{sl} those to alternative allele. The data depends on the underlying genotype. When it is heterozygous, it also depends on ASE, let ρ_{sl} be the measure of ASE (proportion of reference allele).

- Allele-specific model: let g_l be the genotype of SNP l (unknown). When it is homozygous, the read counts follow simple binomial distribution, allowing sequencing errors. When it is heterzygous, we use Beta-Binomial distribution. Let $N_{sl} = R_{sl} + A_{sl}$ be the total read counts

$$R_{sl}|N_{sl} \sim \text{Beta-Bin}(N_{sl}, \psi(\rho_{sl}, \epsilon_s), M_s) \tag{7.39}$$

  where ϵ_s is the error rate, $\psi(\rho, \epsilon) = \rho(1-\epsilon) + (1-\rho)\epsilon$, and M_s is the overdispersion parameter. Note that: ϵ_s and M_s are shared across all SNPs for a sample. The likelihood function marginalize g_l, with a given prior (default: 1KG, but can be provided by users).

- Inference: EM to estimate genotype for each SNP l and error rates ϵ_s for each sample, using all data. Genotyping: fit the mixture model (3 possible genotypes) to infer g_l and estimate ϵ_s. In this step, assume $\rho = 0.5$.

- ASE inference: Consider only heterozygous sites with large MAP of genotype. (1) Estimate overdispersion parameter M_s using all sites of sample s: fix $\rho = 0.5$, and use grid search. (2) For each site l, test if $\rho_{sl} = 0.5$ using LRT, with fixed $\hat{M}_s$. Also determine confidence interval using asymptotic distribution.

- Analysis: problem occurs to decide between i) an heterozygous genotype call with extreme allelic imbalance, or ii) an homozygous genotype call with base call errors; our model will favor the latter ii) because it allows genotype uncertainty and base calling errors. Had genotypes been given, the only way to explain alt. allele reads is ASE.

- Comparison with other tests: calibration of p-values (Figure 4). Binomial test: inflated. Beta-Binomial test: better, still many small p-values (FPs), possibly because homozygote genotypes called falsely as heterozygotes. Comparison of p-values: Beta-Binomial test similar p-values, but "SNPs with more uncertainty on being heterozygous are corrected in a higher degree towards a less significant p-value".

- Q: why genotype uncertainty can influence p-values? If a SNP passes the MAP threshold, p-value of the test should no longer be affected by genotype uncertainty.

WASP: allele-specific software for robust molecular quantitative trait locus discovery [van de Geijn and Pritchard, NM, 2015]

- Addressing mapping bias (Figure 1): mapping to personalized genome does not addressing ref. bias, as the uniquely mappable reads of two alleles may differ. WASP: for reads that map to a polymorphic site, replace with the alt. allele and remap the reads - if mapped not to the exactly same location, discard the reads. Problem: underestimation of expression level.

- Filtering duplicate reads: this step can introduce ref. bias, as the highest scoring reads will usually map to ref. allele. WASP: filter duplicate reads randomly.

- Adjusting for total read depth (Figure SN1): samples could differ in their amplification efficiency, and hence the number of reads mapped to peaks vs. background. This changes the relationship of read depth vs. true expression, e.g. in low efficiency experiment, reads are not concentrated on high expression features. To adjust for this, we need: for a particular sample, given the expression level of a gene, how many reads (fraction of reads) we expect. So for each sample, we fit feature read count vs. total read count across all samples (proxy of true expression), and obtain sample-specific quadratic model. This adjust for sample difference in efficiency: e.g. a high-expression gene will have a lower fraction of reads (over the library) in low efficiency samples, comparing with high-efficiency samples.

- Adjusting for GC correction: similar to RASQUAL.

High-throughput allele-specific expression across 250 environmental conditions [Moyerbrailean & Luca, GR, 2016]

- Principle: GWAS SNPs may manifest as condition-specific eQTL (or ASE).
- Experiment: 50 treatments, 5 cell types. Two pass: first shallow sequencing < 10M, to find DE genes and only follow up experiments with large DE; then deep sequencing (130M).
- Detecting ASE: QuASAR on each condition separately.
- Detecting conditional ASE (cASE): two approaches
  - MeSH [Wen & Stephens]: test four models, not ASE in treatment and control; ASE only in treatment or control; ASE in both. Obtain BF for each of the models. Note: not capture quantitative changes of ASE magnitude across conditions.
  - ΔAST (differential allele-specific test): Z score from the difference of the estimated effect sizes in treatment vs. control.
- Detecting induced ASE: ASE in one condition (e.g. treatment) and gene not expressed in the other condition.
- Validation of ASE genes using GWAS: gene set enrichment. Per SNP heritability enriched in ASE, cASE and iASE genes. Example: genes of T2D show differential ASE in the caffeine treatment condition.
- Remark: detection of ASE jointly across multiple conditions may increase the power. Could use CorMotif or MASH.

Fine-mapping cellular QTLs with RASQUAL and ATAC-seq [Kumasaka and Gaffney, NG, 2016]

- Idea: we test association of a putative QTL (rSNP) with a feature (gene expression or peak). If the rSNP is real, it will also generate allele imbalance of the SNPs in the feature (fSNP), assuming rSNP and fSNPs are linked.
- Notations: Figures S26. We study one rSNP a time, and let i be index of sample. The genotype of sample i, G_i is linked to a feature (peak or expression), but there could be multiple variants in that feature. Ex. two fSNPs, our data are $Y_{i1} = (Y_{i1}^0, Y_{i1}^1)$ and $Y_{i2} = (Y_{i2}^0, Y_{i2}^1)$. Note: QTL can be located in the feature and QTL can be one of the feature variants. Denote G_{il} as the genotype of the l-th fSNP in sample i. We can consider the "diplotype", i.e. two haplotypes, at i and l: e.g. $D_{il} = 00/11$. The uncertainty of G_i and $D_{il}|G_i$ is modeled (below). For read counts, let Y_i be the total read counts in the feature, and $Y_{il} = (Y_{il}^{(0)}, Y_{il}^{(1)})$ the count of ref. and alt. allele of the fSNP l. Also denote Y_{i0} as the count of all reads not overlapping with any fSNP.
- Model overview: let $1-\pi, \pi$ be the allelic effect of QTL (allele bias). Then the expected rate (of read counts) would be $(1-\pi)\lambda$ for ref. allele and $\pi\lambda$ for the alt. allele, where λ is the relative expression level (proportion of reads of the feature in the library). (1) Between individual variation: Y_i depends on G_i using NB model, with parameters π, λ and θ (overdispersion). (2) Allele-specific (AS) model: Beta-Binomial model with the same overdispersion parameter as between-individual model, and error parameters δ (mapping and sequencing error) and ϕ (reference bias). So the overall model can be written as:

$$P(Y|G, \pi, \lambda, \theta, \delta, \phi) = \prod_i P(Y_i|G_i, \pi, \lambda, \theta) \prod_i \prod_l P(Y_{il}^{(1)}|Y_{il}, D_{il}, \pi, \delta, \phi, \theta) \tag{7.40}$$

  The validity of this can be seen in "Probability Decomposition" in Supplement.

- Total count model: let cis-effect of the QTL G_i as $f(G_i)$, which is $2(1-\pi)$ if $G_i = 0$, and 1 if $G_i = 1$ and 2π if $G_i = 2$. Then our model is:

$$Y_i|G_i \sim NB(\lambda K_i f(G_i), \theta K_i f(G_i)) \tag{7.41}$$

  where K_i is the size factor of sample i, λ is the average level of feature and θ overdispersion parameter. Under this model, Y_i depends only on the genotype of rSNP.

- NB and Beta-Binomial distributions: see the section "Negative binomial and beta binomial distributions" in Suppl. If we have two NB distributions with the same index of overdispersion, then the conditional distribution is Beta-Binomial and the sum is still NB. Let

$$X \sim NB(\lambda, \beta) \tag{7.42}$$

  where we parameter it s.t. λ is mean and $\text{Var}(Y) = \lambda + \frac{1}{\beta}\lambda^2$. Suppose

$$K \sim NB\left(\frac{\alpha\lambda}{\beta}, \alpha\right) \tag{7.43}$$

  It is easy to check X and K have the same index of overdispersion. Then the conditional distribution is:

$$Y|N = Y + K \sim \text{BetaBinom}\left(N; \frac{\beta}{\alpha+\beta}, \alpha+\beta\right) \tag{7.44}$$

- AS Model: see "Fitting specific over-dispersed count distributions". This leads to the conditional distribution of alt. allele reads for each feature variant:

$$Y_{il}^{(1)} \sim \text{Beta-Binom}\left(Y_{il}, \frac{f_1(D_{il})}{f_1(D_{il}) + f_0(D_{il})}, \theta(f_1(D_{il}) + f_0(D_{il}))\right) \tag{7.45}$$

  where $f_1(D_{il})$ and $f_0(D_{il})$ describe the relative level of ref. and alt. alleles at fSNP l when the diplotype is D_{il}, and is given by Table S4. Ex. when $D_{il} = 11/00$, we have $(f_0(D_{il}), f_1(D_{il})) = (1-\pi, \pi)$.

- Remark: the derivation of AS model comes from the general property of NB distributions. However, in our case, the overdispersion of NB results from individual random effects, which should be controlled in two alleles of the same individual. To understand why we still need BetaBinomial to account for overdispersion, we can think of "haplotype random effects": the means of the two haplotypes are equal, but they each have an additional random effect.

- Modeling errors and bias: the fraction of alleles change from above due to mapping errors (e.g. repeats) and reference bias. Let δ be the error of a read (allele switch), and ϕ be the reference bias ($\phi = 0.5$ no bias), then we modify $f_1(D_{il})$ and $f_0(D_{il})$ as:

$$\begin{aligned} \tilde{f}_0(D_{il}) &= 2(1-\phi)[(1-\delta)f_0(D_{il}) + \delta f_1(D_{il})] \\ \tilde{f}_1(D_{il}) &= 2\phi[\delta f_0(D_{il}) + (1-\delta)f_1(D_{il})] \end{aligned} \tag{7.46}$$

- Genotype and haplotype uncertainty: consider errors in imputation and phasing. Also haplotype switching between i and l can happen. Then G_i and D_{il} need to be marginalized in the likelihood (main equation in text). Use EM to estimate parameters, and marginalize genotypes/diplotypes.

- Allowing total counts to depend on ϕ and δ (Page S48): the current total count model does not consider reference bias. Ex. if many fSNPs in a sample are alt. alleles, then the observed total count may be lower than expected. To address this, we consider two types of reads: those not overlapping fSNPs, and those overlapping fSNPs. For the former, its expected read count depends only on rSNP, but for

the latter, it depends on both rSNP (effect) and fSNP (mapping bias), as described in the AS model. We have:

$$\mathrm{E}\,(Y_i) = \lambda \left[(1 - hL) K_i f(G_i) + hK_i \sum_l (\tilde{f}_0(D_{il}) + \tilde{f}_1(D_{il})) \right] \tag{7.47}$$

where h is the fraction of reads contributed by one fSNP. Inference of this model is difficult, since now Y_i model depends on a large number of hidden genotypes D_{il}. The model approximates by, for each G_i, averaging over all possible D_{il}'s. It is shown that the model makes relatively little difference comparing with the simpler model.

- Normalization problems and strategy:
  - GC content: amplification efficiency depends on GC content, and the relationship can vary across samples. This leads to noises: e.g. in one sample, amplification is high for high GC peaks, then those peaks will have more reads, not due to genotypes, but due to amplification bias.
  - Hidden confounders: e.g. some samples are subject to higher stress, which activates stress-related peaks. Then in these samples, the stress peaks have higher read counts, not due to genotypes.

  In general, GC and confounders reduces power, but do not create FPs. Two general strategies to address this are: (1) normalize RPKM/FPKM by multiplying a correction factor so that peaks with different GCs or confounders are comparable. Let Y_{ij} be the read count of peak j in sample i, then the log2-RPKM/FPKM is defined as:

$$y_{ij} = \log_2 \frac{Y_{ij} + 1}{l_j Y_i} \tag{7.48}$$

  where l_j is the length of feature j and $Y_i = \sum_j Y_{ij}/10^6$ is the library size. We will then adjust for y_{ij}. (2) Normalize the expected rates in a model (instead of the data): this is the size factor K_i in the total count model, and is simply defined as (for now, and to be adjusted later): $K_i = Y_{i\cdot}/Y_{\cdot\cdot}$, where $Y_{i\cdot} = \sum_j Y_{ij}/J$ and $Y_{\cdot\cdot} = \sum_{i,j} Y_{ij}/(NJ)$.

- Normalization by GC content ("GC correction for fragment counts and FPKMs"): we obtain a sample-specific profile of read enrichment vs. GC content. We first bin all features by their GC contents. Let $S_{il} = \sum_{j \in B_l} Y_{ij}$ be the total reads in bin l of sample i. We can then define enrichment of reads in bin l for sample i as:

$$F_{il} = \log_2 \frac{S_{il}/S_{\cdot l}}{S_{i\cdot}/S_{\cdot\cdot}} \tag{7.49}$$

  We can then fit a spline function of F_{il} vs. GC content of bin l, and let $c_{ij} = \hat{F}_{ij}$ be the expected log2 enrichment of feature j in sample i. This can be used to normalize RPKM/FPKM or the size factor:

$$\tilde{y}_{ij} = y_{ij} - c_{ij} \qquad K_{ij} = K_i e^{c_{ij}} \tag{7.50}$$

- Normalization by PCs ("Principal component correction"): the strategy is to first learn hidden confounders (e.g. stress level) for each sample, then for each feature, we adjust for the effect of confounders (the effects may differ for different features). To learn hidden confounders, we do PCA on y_{ij}'s, and choose PCs based on permutation (only PCs whose contributions are greater than permutation). To adjust for PCs with RPKM/FPKM: we regress y_{ij} vs. x_i (PCs), and obtain $\tilde{y}_{ij}$ as residuals. For the count data, we fit NB regression model (ignoring genetic effects), and assume the mean for peak j of sample i depends on PCs:

$$Y_{ij} \sim NB(\lambda_{ij} K_i, \theta_j) \qquad \log \lambda_{ij} = \alpha_j + x_i^T \beta_j \tag{7.51}$$

  The estimated parameters can be then used to adjust for size factor as: $K_{ij} = K_i \exp(x_i^T \beta_j)$.

- Multiple testing correction: (1) Correct for number of SNPs per feature: from the p-value of lead SNP, multiply by the number of SNPs tested (Bonferroni correction). (2) Calibration of genomewide threshold: permutation (swap genotypes and read counts of multiple samples), and run the same procedure on permuted data. From the list of p-values from permuted data, do FDR estimation: expected number of features passing threshold under null vs. observed number.
  - Remark: the null distribution is obtained from all features in permutations. However, we expect the null to be difference. To address this, first use Bonferroni correction as the first step to adjust for number of tested SNPs per features.
- Remark/Question: only consider lead SNPs in multiple testing? How many permutations are required?
- Comparison with Combined Haplotype Test (CHT): the last part of Suppl. and Table S2. The main differences:
  - Overdispersion: shared with between-individual variation and is feature-specific under RASQUAL. Sample-specific under CHT.
  - Reference bias: estimate for each feature under RASQUAL. Not estimated, but addressed by randomization under CHT.
  - Additional sequencing/mapping error: estimated for each feature under RASQUAL. Fixed error rate under CHT.
  - Genotype uncertainty and haplotype switching: not modeled by CHT.
- Assessing performance: use QTL of CTCF peaks, and enrichment of true causal variants (CTCF motif changing SNPs).
- ATAC-seq of 24 LCL: 2000 caQTL, and 800 of them are in the peaks, and some in perfect LD. Among 900 lead SNPs (peaks or in perfect LD), more than 600 change a motif. Also examples of a QTL affecting multiple peaks: particularly, a SNP is caQTL of an enhancer and caQTL of nearby promoter. Among 173 multi-peak caQTL: most often the effect of caQTL on master and dependent peaks are consistent. However, not confident for interactions more than 100kb away.
- Lesson: if we can detect true causal variants for regulatory SNPs, many of them act by changing TF binding motifs.
- Lesson: regulatory SNPs can be used as IVs to study relationship among molecular traits.

Leveraging allelic imbalance to refine fine-mapping for eQTL studies [Zou and Eskin, PLG, 2019]

- Limitations (from PLASMA paper): treat AS as binary. No phasing.

Allele-Specific QTL Fine Mapping with PLASMA [Wang and Gusev, AJHG, 2020]

- Background: QTL effect size vs. AS effect size, nonlinear relationship. To understand why: QTL effect sizes depend on AF, while AS effect size not.
- Model: let y_i be normalized expression, $y_i = x_i\beta + \epsilon_i$, where x_i is genotype. For AS signal, let w_i be the log. allelic ratio, and v_i be AS phasing, coded as 1 if alt. haplotype and -1 if ref. haplotype and 0 if heterozygotes. We have $w_i = v_i\phi + \xi_i$, where ξ_i models the error. Note that ξ_i is not the same for all, derived it from Beta-Binomial distribution to capture over-dispersion. The model is equivalent to weighted linear regression.
- Modeling assumption and fine-mapping: allow correlation of QTL and AS effect sizes. However, found that in practice, this is not helpful, so the default is 0. Shared signal: same causal variants for QTL and AS signals.

- Methods compared: RASQUAL+, convert χ^2 to Z scores, then do FINEMAP. Also AS-Meta [Zou and Eskin, PLG, 2019]. QTL only. PLASMA-J (use both QTL and AS) and PLASMA-AS (only use AS signal).
- Simulation: for QTL signal, simulation of data similar to quantile normalization. Simulation under low AS variance and high AS variance setting. PLASMA-J and PLASMA-AS best. RASQUAL+ works well in high AS variance setting, but poorly under low AS variance.
- Results in TCGA: PLASMA-J slightly better than PLASMA-AS, much better than QTL only and FINEMAP (Figure 5): median CS size of 32 vs. 167.
- Lesson: AS signal is much stronger than QTL signal. Quantile normalization procedure in QTL study probably significantly reduces the power.
- Q: How is the method applied to real data? QTL discovery step? Only applied to significant QTLs?
- Q: how is QTL model used in real data? Does it adjust for covariates/hidden factors?
- Remark: credible set size is very large for FINEMAP, etc. in real data. Ex. in tumor data, CS is bigger than 100 most of time.

7.1.4 Single-cell QTL

eQTL mapping in single cell experiments: [personal notes]

- Consider an scRNA-seq experiment where we collect data across multiple time points/conditions. Additionally, samples are multiplexed, so each batch may contain a mix of multiple samples.
- Creating pseudo-bulk data: first need to define cell types/populations. This is usually done by joint clustering analysis of all cells from all conditions. Downstreaming analysis is usually done on each cell type and each condition, separately. Sometimes, it may be better to use cell clusters defined from the joint clustering analysis, e.g. time point may not be a good marker of cell state [Cuomo, NC, 2020].
- Mapping eQTL by Linear mixed model (LMM): each sample is defined by genetic background (donor) and experimental batch (in multiplexed design). Both of which may affect gene expression. For a given gene, let y_{ij} be its expression at donor i in batch j. Our model will included fixed effects from eQTL X_i and latent variables (e.g. PCs) Z_{ij}, and random effects from genetic background (which is shared across multiple batches for the same donor) and batches.

  $$y_{ij} = X_i\beta + Z_{ij}\gamma + g_i + b_j + \epsilon_{ij} \tag{7.52}$$

  where g_i and b_j are random effects. Their correlation structure is given by GRM and sample-repeat structure.
- Mapping response/condition-specific eQTL: [Cuomo, NC, 2020] uses ASE interaction test on cis-eQTL already mapped. This removes the batch effects.

Science Forum: The single-cell eQTLGen consortium (sc-eQTLGen) [eLife, 2020]

- Advantages of single-cell eQTL mapping vs. bulk eQTL: mapping genetic loci of cell types and states; trajectory; variability of genes across cells; coexpression network. Possible cost advantage: easy to multiplex to collect data of multiple samples in one experiment.
- Mapping cell-type specific eQTL: (1) define cell types, better to use supervised approach, e.g. use HCA as reference. Linear model is found to work as well as more complex models (e.g. DNN). Important to have a rejection option: i.e. a cell not belong to any known cell type/state. (2) Unsupervised clustering still important: to define un-annotated cells.

- Personalized GRN: to define co-expression relationship/network for an individual: technical challenge because of read sparsity. (1) gene expression imputation (currently limited) (2) Use prior GRN (e.g. TF to target), and combine with imputation. (3) Use temporal information and/or pseudo-time, RNA velocity.
- Challenge with trans-eQTL and co-expression QTL: sharing summary statistics is difficult because of large size.
- GRNs from scRNA-seq and association study: personalized GRNs can be used in multiple ways. (1) Association test of GRN topology. (2) Improve eQTL detection: use use network to define prior.
- Application to phenotypes: assess convergence of effects on genes or processes.
- Remark: problems/challenges: how to perform association analysis on trajectory?
- Remark: association of genotypes with features from GRN (e.g. co-expression or regulatory strength of TF-gene). Opportunities for method development: (1) Use all samples together to do GRN reconstruction for each sample. (2) Model uncertainty of the GRN features.
- Remark: sc-eQTL data would support MR/mediation analysis of gene ¿ GRN features/cell composition/trajectory, because gene and cellular level traits (e.g. cell composition) are measured in the same samples.

Single-cell RNA sequencing identifies cell type specificcis-eQTLs and co-expression QTLs [van der Wijst and Franke, NG, 2018]

- Background: Controlling for batch effects in scRNA-seq using PCA. The top PCs capture mostly technical artifacts, they correlate to the proportion of non-zero reads in cells. This could create problem for clustering: e.g. PC1 correlated with droput rate (batch effect) and the rates differ in different batches, so the samples are clustered by batches.
- Background: Imputation for scRNA-seq. In Fludigm - borrow information from cells, good results. Drop-seq/10x: no current imputation method works.
- Experiment: 45 individuals in 8 batches. 10x Genomics for scRNA-seq, total of 25K cells. In pooled individuals in one batch: identify individuals using SNPs from reads.
- Mapping eQTL with pseudo-bulk data (pool all cells of one sample): log-transformed read counts and quantile normalization.
- Validation: BIOS 2K samples and DeepSAGE, bulk-RNA seq. Found 8% and 1% of the cis-eQTL from these two studies using pseudo-bulk.
- Create clusters of cell types: cluster cells by PCs, obtain 11 cell types. Then manually annotate the cell types.
- Cell-type specific eQTL: union of all cell types: 379 cis-eQTL, most are found in bulk-like analysis. Among 48 new cis-eQTL, 29 are found in bulk of bigger samples. For 19 new cis-eQTL: 3 are validated using cis-eQTL of purified cells.
- Possible explanation of cell-type specific eQTL that are missed by bulk study: (1) Rare cell types. (2) Could also happen in common cell types: eg. Figure 1b, cis-eQTL in CD4 T cell (common), but the expression level of the gene is low, comparing with other common cells. This masks the signal.
- Mapping co-expression QTL: choose CD4 T cells, 100 genes with cis-eQTL. Consider pairs these genes with other genes that are also significantly expressed. Estimate correlation of gene pairs in each sample, then co-expression coefficient vs. genotypes (limited to cis-eQTL). Found 2 cis-genes with co-expression QTL, Figure 2.

- Q: whats the mechanism of co-expression QTL? Ex. could cis-eQTL of one gene lead to co-expression QTL?
- Question: Whats the nature of the problem with PCA? If we properly normalize our data, do we get PCs that are uncorrelated with technical variables?
- Remark: dropout is also a problem for computing correlation.

Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression [Cuomo and Stegle, NC, 2020]

- Experiment: endoderm differentiation in 125 iPSC lines, SMART-seq and FACS for each cell. Pooled differentiation assay: 4-6 lines per experiment, then differentiation. Sample at Day 0, 1, 2, 3. Total: 36K cells, and 11K genes.
- Understanding sources of variation: LMM, time point as main source of variation, then cell line, then batch.
- PCA: using 500 most variable genes, PC1 (19%) corresponds to time points (Figure 1c). Assign psuedotime: PC1, scaled to [0,1]. Alternative: diffusion map, principal curve form top 2 PCs.
- Assign cells to stages: use canonical marks of stages, which correlate with psuedotime (Figure 1d). About 20K cells can be assigned, and 7K not.
- Stage-specific cis-eQTL mapping procedure: group samples by individuals, days, and experiment (batch). Use LMM to map: treat individual genotype and batch (sample repeat structure) as random effects. Let y_{ij} be expression of line i from batch j:

$$y_{ij} = G_i\beta + g_i + b_j + \epsilon_i \tag{7.53}$$

  where g_i is the random effect due to genetic background and b_j is the random effect due to batch. The random effect terms are correlated across samples, given by GRM and sample-repeat structure (which donors occur in the same batch). Also adjust for 10 PCs of expression data. Response: log2-CPM. Note: because the same sample is used in multiple experiments, its important to correct for individuals as random effects.
- Results of cis-eQTL mapping: (1) about 1K genes in each stage. Importantly, the power is substantially higher than using days. (2) Stage-speicficity: abou 30% are stage-specific. (3) 155 switching events: top eQTL change across stages. In 22 cases, see corresponding chromatin changes.
- Quantification of ASE (for dynamic and interaction eQTL): only performed for cis-eQTL. For a given eQTL, ASE was only quantified across cells from donors heterozygous for that eQTL variant. Donors that are not heterozygous at the eQTL variant of interest were not used for quantificatio. For each cell, quantification of ASE: (1) for each heterozygous exonic SNP, obtain ref. and alt. allele reads. (2) Aggregate SNP level to gene level alt. vs. ref. alleles, using phasing information. Then conversion to allelic fractions.
- Mapping of dynamic and interaction eQTL using ASE: Figure 4a. Identify 60 clusters of genes. Use 4 clusters to represent cell states, respiration and G2/M transition. Use ASE, which is free from batch effects. (1) Dynamic eQTL: ASE effect is a function of pseudo time, modeled as: ASE = $\text{pseudo} + \text{pseudo}^2 + \epsilon$. (2) Cellular factors: ASE = $\text{pseudo} + \text{pseudo}^2 + \text{factor} + \epsilon$. (3) Pseudotime-factor interaction test: ASE = $\text{pseudo} + \text{pseudo}^2 + \text{factor} + (\text{pseudo} \times \text{factor}) + \epsilon$. Tests were only performed for eQTL for which ASE was quantified in at least 50 cells.
- Dynamic eQTL: use sliding window (25% cells), and compute eQTL and ASE in each window. Then do regression of effects vs. pseudotime (linear or square). About 800 eQTLs show dynamic effects. Show several clusters of dynamic eQTL patterns.

- Possible mechanisms of dynamic eQTL: only partially correlated with gene expression changes over pseudotime.
- Cellular states modify eQTL effects: 668 eQTL that had an interaction effect with at least one factor (Fig. 4b). Ex 1: interaction of eQTL of RNASET2 with cellular respiration. Ex. 2: interaction of eQTL of SNRPC with G2/M transition. The patters are different: in SNRPC, G2/M has not much effect on expression on ref. allele. In RNASET2, respiration has a large effect on expression on both ref. and alt. alleles.
- Biomarkers of differentiation: genetic effects not associated later trajectory, some genes are.
- Lesson: use LMM to study variance components of gene expression in such experimental design.
- Lesson: using data-derived stages may be better than actual experimental time in cell differentiation.
- Remark: interaction test is based on quantification of ASE in each cell. This is difficult to generalize to caQTL and may miss many interactions.
- Remark: interaction eQTLs can result from different patterns. Ex. it may be driven by change of expression by conditions; or change of genetic effects.

Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation [Gaffney, biorxiv, 2020]

- Experiment: 215 iPSC lines, differentiation, D11, D30 and D52, then oxidative stress. ScRNA-seq. Multiplexed design: 7 and 24 cell lines per experiment.
- Clustering: joint analysis of all cells from all conditions. 14 cell populations, including 6 dominant ones. Some are non-neuronal cells.
- iPSC line differentiation efficiencies differ: large variations of cell type composition at D52. Some lines are more efficient, and more importantly, the difference is reproducible and extend to other neuron differentiation protocols.
- What determines differentiation efficiency? Not chr. X inactivation, genetic background (multiple lines of the same individual show large difference). Gene expression signatures: explain difference, and in particular a cluster of genes are predictive of poor differentiation.
- cis-eQTL mapping: on 14 populations and conditions separately. LMM, include a random effect term to capture differentiation efficiency difference across donors ("variance term $1/n$'), which influence the number of cells in a population (and accuracy of estimating expression).
- Results of cis-eQTL mapping: $>$ 1500 eQTLs for D11 and D30 in some common cell types, and 500-1000 for D52 and stress in common cells. For astrocytes, $<$ 100 eQTLs across conditions. Some examples of time point and condition-specific eQTLs (Figure 4b): effects much larger in one time point/condition than another (though some effects). Comparison with GTEx: about half are new.
- Colocalization with GWAS: coloc with 25 neurological traits. About half are detected in D52 and stress conditions. Two examples: (1) With SCZ: eQTL in stress, but not found in GTEx. (2) With SCZ: eQTL only in D11, but not in GTEx.
- Q: How does LMM work?
- Remark: (1) Gene expression signature of differentiation efficiency: enriched with neuronal lineage genes? (2) Differentiation efficiency may be explained by epigenetic states - the developmental competency model [Wang and Sander, CSC, 2015]. Hypothesis: if neuronal lineage enhancers are primed, then more likely to differentiate to neurons.

7.2 eQTL and regulatory QTL (rQTL) Studies

Research problems of eQTL and regulatory QTL:

- Genetic architecture of eQTL: how much is cis vs. trans? Spatial distributions? Tissue-specificity? Do trans-eQTL often cluster in co-regulated genes?
- Mechanisms of eQTL: TF-binding, chromatin structure, etc.
- Evolutionary explanation of patterns/genetic architecture of eQTL. Ex. why some genes have a lot more eQTL than other genes.
- Connection to complex diseases: phentypic consequences of eQTL.

Challenges of eQTL and regulatory genetics:

- Variation of gene expression: the relative role of pioneering TFs and secondary TFs? How to identify pioneering TFs?
- Why variation of TF binding often do not lead to change of gene expression?

Genetics of human gene expression: mapping DNA variants that influence gene expression. [Cheung & Spielman, NRG, 2009]:

- Tissue types: most are lymphoblastoid cell lines, or LCL (B cell lineage, or immortalized B cells), the other ones include: blood and adipose tissues, brain, lymphocytes, liver samples from surgical or cadavers.
- Comparison of primary cells vs cell lines: cell lines are relatively free of environmental influences. Ex. in human, the blood cell count, diet, medication, etc. can all influence gene expression in blood cells.

The Genetic and Mechanistic Basis for Variation in Gene Regulation [Pai and Gilad, PLG, 2015]

- eQTL mapping studies:
  - In [Battle et al], eQTL in 962 individuals, 78% of genes were linked to at least one eQTL. Most of them are in 5' end, suggesting transcriptional regulation (rather than RNA decay) might be more important.
  - Possible to map trans-eQTL reliably. Estimate that more 50% of heritability is from trans-eQTL.
- Approach: regQTL, find mechanisms underying eQTL by studying which part of the regulatory process is affected.
  - In LCLs, an estimated 10-20% of eQTLs are also classified as methylation QTLs (meQTLs), suggesting a small proportion of loci that affect gene expression do so by perturbing DNA methylation levels.
  - The problem of causality: meQTL may not explain the effect of eQTL, in other words, we may have two models, A) eQTL $\rightarrow$ DNA methylation $\rightarrow$ expression, or B) eQTL $\rightarrow$ DNA methylation and eQTL $\rightarrow$ expression. The two models can be distinguished by partial correlation analysis: under model B), DNA methylation and expression are uncorrelated if regress out eQTL. The data supports model B).
- Findings from regQTL studies: transcriptional regulation
  - Chromatin accessibility (TF binding) seem to play a major role: 55% of eQTL are found to be dsQTL.

  - Histone modification-QTL [Kasowski & Synder, Science, 2013]: QTL of enhancer states (defined primarily by H3K27ac and H3K4me1) are often not eQTL. Explanatios: many enhancers defined in this way are not functional, compensatory change (including enhancer redundancy) or other means of buffering.
  - Histone modification-QTL [Kilpinen, Science, 2013; McVicker, Science, 2013]: changes in sequence-based affinity for TF

- binding underlie a subset of observed histone-mark QTL (hmQTL). Also the effect is stronger on nascent RNA.
  - Additional support of TF binding as the driving event: [White & Cohen, PNAS, 2013], enhancer assay to test activity of a large number of enhancers. The test sequences include: 1,300 sequences bound by Crx from ChIP-seq and 3,000 control sequences containing the motif, but not bound. Results: the Crx-bound sequences drive expression, while unbound sequences do not. In the experiment, the chromatin context has been removed, so the difference of activities must be due to local sequence features.

- Findings from regQTL: post-transcriptional regulation
  - Splicing QTL (sQTL): many sQTLs fall directly within primary splice sites. However, also show an enrichment near TSS and 5'-UTR and TFBS.
  - RNA decay QTL (rdQTL): almost half of the nearly 200 rdQTLs identified showing counterintuitive associations with mRNA expression levels (higher rate of decay, higher expression). Overall, it was estimated that as many as 19% of eQTLs might be driven by differences in mRNA decay.
  - protein QTL (pQTL): [Wu & Synder, Nature, 2013] Only about half of these pQTLs were also found to be affecting transcript levels, suggesting many affecting regulation of translation or protein stability. Most of (80%) the eQTL SNPs found in these

- LCLs were not also associated with variation in protein levels: suggesting possible buffering mechanism.

- Predict variation in gene expression levels based on the DNA sequence:
  - Challenge: many changes in TF binding do not seem to result in measurable changes in gene expression levels. [Cusanovich & Gilad, PLG, 2014] knockdown expression of many TFs, only a small subset of genes inferred to be bound by a TF (using DHS or ChIP-seq data) were differentially expressed.

- Model of transcriptional variation:
  - The model that histone modifications regulate chromatin state, which in turn determine whether factors can bind to different sites need to be revised. Instead, TF binding seems to be the central event.
  - Possible model: pioneering TF binding concerted changes in histone marks, DNA methylation and nucleosome binding. Then chromatin areas that are accessible because of pioneer activity are available for binding by secondary factors.

- Lessons:
  - regQTL paradigm to understanding mechanism of transcriptional variation, and the difficulty of establishing causality.
  - TF binding may be the driving event; however variation of TF binding often do not correspond to expression change.

- Remark:
  - The issue of power: only a small fraction of eQTL are meQTL, but the power of detecting meQTL might be low, and this should be accounted for.

7.2.1 eQTL Studies in Human

Genetic architecture of gene expression:

- Heritability of gene expression:
  - Estimation from families [Goring, NG, 2007]: 1240 individuals in 30 large families. Most genes (86%) show significant heritability, but the level of heritability is moderate: 41% had heriability > 0.3 but just 5% had heritability > 0.5 . Thus non-genetic factors are also important.
  - Mean expression level difference across human populations: 17-29% genes have significant difference in mean expression levels between pairs of HapMap populations [Stranger, NG, 2007]. It is likely that this is due to environmental factors: few SNPs have large frequency difference across populations, and the comparison of genetically similar groups in different environments show 37% of expressed genes show significant difference [Idaghdour Y, PG, 2008].
- Detection rates of eQTL: in a mouse study, QTLs were detected for only 27% of genes with significant genetic differences in expression.
- Effect size distributions:
  - Number of QTLs: most studies detected only a single locus for most expression traits. However, most expression traits should involve multiple QTLs because no single QTLs can explain most of the genetic variation. Ex. it is estimated that in yeast, only 3% of expression traits are consistent with single-locus inheritance.
  - Effect size: consider only the most significant QTLs for the expression traits. In yeast, the median phenotypic effect of a detected QTL was 27% of genetic (heritable) variance of expression; in mice, average 25%, in human, 27-29%.
- Type of complex inheritance: a large fraction - transgressive segregation (segragants fall outside parent means); and a small faction - directional genetics (segregants fall between parent means).
- Cell-type dependent expression: a special form of gene-environment interactions. In mouses and rats, the genetic basis of variation of a gene's expression is sometimes shared between different tissues, but is often unique to each tissue [Cotsapas, Mamm Genome, 2006].

Local and distal eQTLs:

- Mechanisms of local eQTLs: polymorphism at (1) cis-regulatory sequences, (2) neighboring genes that control the expression, (3) coding sequences of auto-regulatory genes.
- Mechanisms of distal eQTLs: polymorphism at (1) coding or cis-regulatory sequences of regulators, including TFs and indirect regulators; (2) distal enhancers.
- Local vs. distal eQTLs:
  - Importance of local eQTLs: (1) In human, proximal eQTLs (within 2Mb) are much more common and most proximal eQTLs are close to the gene (within 100 kb), and distal eQTLs have much smaller effect sizes [Dixon, NG, 2007]. (2) Similar pattern in model organisms, eg. as many as 25% of all expression traits in yeast are affected by local QTL.
  - Bias of detection: often difficult to detect distal eQTLs (especially in low-power studies) because of multiple hypothesis testing. Evidence: (1) in studies with large sample size, most transcripts are linked to distal eQTLs. Ex. [Yvert, NG03], among 2,294 expression traits, 578 show local linkage while 1,716 show distal linkage. (2) In human studies, no eQTL hot-spots (distal) are found, suggesting many distal eQTLs may be missing.
  - What are genes in distal eQTLs? From the existing studies, TFs were not overrepresented near the distal QTLs for 1,716 linkages in yeast [Yvert, NG03].

  - A consistent result from most studies is that trans-eQTLs have weaker effects than cis-eQTLs. This view is contested by more recent studies that suggest that, despite lower effect-sizes, trans-eQTLs cumulatively explain more of the heritability of expression [Montgomery & Dermitzakis, NRG, 2011].
  - Within species, cis-eQTL explains about 35%; across species, about 64% [Regulatory changes underlying expression differences within and between Drosophila species, Wittkopp & Clark, NG, 2008]

- eQTL hotspots: a loci that affects expression of many genes. Not necessarily "master regulators", some of them have pleiotropic effects: e.g. a structural gene that significantly affects that phenotype.

Tissue specificity of eQTL:

- Reference: [Tissue specificity of genetic regulation of gene expression, Goring, NG, 2012]
- MuTHER eQTL data: skin, subcutaneous fat and peripheral blood (LCL).
  - cis-eQTL: Overall, 50-80% of the loci identified in one tissue were estimated to have gene regulatory effects in a comparison tissue. Effect sizes were often also comparable between tissues.
  - Trans-eQTL: largely tissue specific. However, this inference is less convincing than those made for cis eQTLs, because the power of detecting trans-eQTL is low in the first place.
- Proxy tissue problem: the question of whether or not eQTL information from one tissue is relevant for another is hotly debated. When eQTLs have not been reliably cataloged for the tissue affected by a disease, it seems entirely rational to try to use a proxy tissue.

Methods of identifying gene regulatory networks in eQTL data:

- Module identification and module QTL (mQTL): similar to the usual analysis, modules can be found by clustering of expression patterns across different genetic pertubrations/backgrounds. The module QTLs can be identified via linkage/association of genetic markers with module expression.
- Regulator finding of transcripts or modules: treating the transcripts or modules as complex traits, and apply similar strategy for integrating complex trait and eQTL data (the section "Systems Genetics").

Small-scale study of human eQTL [Cheung & Burdick, Nature, 2005]:

- Data: 57 CEPH individuals, lymphoblastoid cells (immortalized B cells), expression of genes that show significant linkage in previous experiment.
- Analysis: $\log_2$ transformed gene expression, regression on SNP genotypes (coded 0, 1, 2). The effect size is measured by R^2.
- Results:
  - 374 expression traits: with evidence of previous linkage, association test with markers in the region with linkage in the previous experiment. There are 17% expression traits with at least one marker that show evidience of association at nominal $P < 0.001$.
  - 27 expression traits: also with evidence of previous linkage, genome-wide test ($700,000$ markers). Out of 14 traits, significant association at nomimal $P < 6.7 \times 10^{-8}$. Most significant associations occur in cis- (within 50kb of 5' and 3'), only one trait has significant association in both cis- and in trans-. For non-significant traits, most have trans- association.
  - Experimental validation of one SNP (probably causal): about 2-fold didfference of expression with alternative alleles.

- Discussion: sample size estimation. In the ideal case where a single causal variant determines expression variation, to achieve a probability of 0.8 of detecting effect size R^2 of 0.1, the sample size of 500 would be needed.

Human eQTL in asthma dataset [Dixon & Cookson, NG, 2007]:

- Methods:
  - Data: 400 children from families recruited through a proband with asthma. 400,000 SNPs, measurement of 54,675 transcript (20,599 genes) in lymphoblastoid cells.
  - Analysis: FASTASSOC component of MERLIN, including sex (probably include family structure). LOD score 6.0 as threshold for significance, corresponding to FDR 0.05.
- Results:
  - Heritability and eQTL: 14,819 traits have $H^2 > 0.3$, the peak LOD score for association: 3.7 to 59.1. 1,989 traits have peak SNP LOD score > 6.0, and about 33% of H^2 in these traits can be explained by the peak SNPs. The GO category of most heritable genes: UPR, genes regulating cell cycle DNP repair, immune response, etc. Only 88 genes areassociated with three or more SNPs
  - Trans- and cis- (> 100 kb) associations: (1) 13 SNPs showed association with ten or more heritable expression traits with lod scores > 6; however, if limited to $H^2 > 0.3$ and remove MHC, only 3 SNPs are associated with five or more transcripts. (2) Trans-effects are weaker than cis-effects, and most LOD > 9 were in cis-. (3) However, numerous distant associations were found: the peak of association for 698 transcripts was on the same chromosome but > 100 kb away, and for 10,382 transcripts, the peak was on a different chromosome.
  - Application: a SNP is eQTL of the gene ORMDL3, and also a locus of childhood asthma, suggesting ORMDL3 is a candidate gene of the disease.

Population genomics of gene expression [Stranger & Dermitzakis, NG, 2007]

- Methods:
  - Data: expression of Epstein-Barr virus-transformed lymphoblastoid cell lines of 270 HapMap individuals: 30 Caucasian trios, 45 unrelated Chineses, 45 unrelated Japanese, and 30 Yoruba trios. Choose 13,643 distinct genes for final analysis.
  - Association analysis across populations: either linear regression (LR) or Spearman rank correlation (SRC) at FDR 4-5% - no significantly different results found. Either do association test within each population, or do pooled test in the whole with conditional permutation for correcting P values (P values will be inflated under conditional permutation).
  - Separate cis- and trans- test: For cis- analysis: permutation test within a region of 1 Mb. For trans- analysis: permutation test is too expensive, limit to 4 cateogires of SNPs: shown cis- effect, nonsynonymous SNPs, SNPs influencing splicing and SNPs within microRNAs, leading to about 25,000 SNPs.
- Results:
  - Population difference: 17 - 29% expression traits show significant difference across pairs of populations. However, caution: may be due to different ages of cell lines from different populations.
  - cis- association: a total of 831 genes show association in at least one population, the pooled association test gives similar results.
  - The functional analysis of cis-associated SNPs: many SNPs may actually be causal, rather than markers because of the high density of HapMap markers. They are very close to TSS (within a few hundred kb), and symmetric in 5' and 3'.

  - Trans- associations: only 108 genes show significant association in at least one population.

eQTL of human liver [Schadt & Ulrich, PLoS Biol, 2008]:

- Methods:
  - Data: human liver sample from 427 Caucasian subjects.
  - eQTL identification: association test using Kruskal-Wallis test. P value threshold is determined by FDR: at any given P value cutoff, do permutation (of sample labels) and count the number of predicted eQTLs, and compare this with the actual number of eQTLs. The P value thresholds corresponding to FDR threshold 10% are $5.0 \cdot 10^{-5}$ and $1.0 \cdot 10^{-8}$ for cis- and trans-eQTLs respectively. cis-eQTL defined as 1Mb of TSS or TES.
- Results:
  - cis-eQTL: about 3,000 genes significantly associated with at least one cis-eQTL. More than 30% of all cis-eQTL are more than 100 kb away from TSS or TES of the corresponding gene (suggesting nearest genes may not always be the true target of a SNP). Comparison btween blood, adipose and liver cis-eQTL: about 30% overlap, but the majority of cis-eQTLs may be tissue-specific.
  - trans-eQTL: 474 genes were found to have at least one trans-eQTL.
  - A more extensive eQTL set: use all significant $\sim 3,700$ SNPs from the two previous steps, and identified additional expression traits at FDR 10%: 3,053 more expression traits or 2,838 genes (a total of about 6,000 genes). A number of eQTL hot spots (defined as more than 20 expression traits) emerged in this full set: highly significant. The total results are found in Table S2.

Human cortical eQTL [Myers & Hardy, NG, 2007]:

- Methods:
  - Data: 193 neuropathologically normal human brain samples (postmortem). Correlations among 366,140 SNPs on the Affymetrix platform and the expression of the 14,078 detected transcripts (detected in at least 5% of 193 samples).
  - Statistical analysis: linear regression with additive model. Outliers due to genetic relatedness and ethnic bias were excluded. Corrected for several biological covariates (gender, age at death and cortical region) and several methodological covariates (day of expression hybridization, institute source of sample, postmortem interval and a covariate based on the total number of transcripts detected in each sample)
  - Statistical significance: report both uncorrected Wald P values and empirical P values from 1,000 permutations.
- Results:
  - Sigificant associations: at empirical P-value 0.05, 433 SNP-transcript pairs (99 transcripts) show cis-association (defined as 1Mb from either end of the gene) and 16,701 SNP-transcript pairs (2,876 transcripts) show trans association.
  - Strigent associations: no variation located within the transcript probe, gene expression detection rate in samples greater than 99%. 26 cis-associations (8 transcripts) and 336 trans-pairs (161 transcripts).
  - Positive control: MAPT expression and MAPT haplotype.
  - Comparison with LCL results: very few, two common cis-associations, and for stringent trans-associations, only one common transcript (but different SNPs).

Human eQTL in LCL of HapMap individuals [Duan & Dolan, AJHG, 2008]:

- Methods:
  - Data: HapMap lymphoblastoid cell lines from 30 CEU trios and 30 YRI trios, 12,747 transcript clusters (TCs) covering all exonic regions. 2 million SNPs.
  - Analysis: QTDT. FDR threshold 0.1 or p value 2×10^{-8}.
  - eQTN blocks and eQTN hotspots: eQTL blocks defined as a region ocontaining one or more eQTNs associated with the same TC, where betwen eQTN interval < 500 kb. eQTN hotspots: a region asscoiated with more than one non-redundant TCs, where the bin size is 500 kb.
- Results:
  - eQTL results: 4,677 significant TC-eQTN associations in CEU and 5,125 in YRI. In terms of TC and eQTN blocks (CEU): 741 unique TCs, out of which 67 TCs associated with 67 local (4Mb) eQTN blocks, and 691 TCs associated with 1,074 ditant eQTN blocks. 23 local TCs share eQTN blocks across populations (CEU and YRI), but none of 143 distnct TCs share the same eQTN blocks.
  - eQTN hotspots: 14 (CEU) and 38 (YRI) distant eQTN hotspots. eQTN harboring genes are enriched with nucleosome assembly genes, and membrane signaling.

Polymorphic cis- and trans- regulation of human gene expression [Cheung & Spielman, PB, 2010]:

- Methods:
  - Data: 45 CEPH families (3-generation, the family size could be big, e.g. 13 members in 13291), LCL.
  - Linkage analysis: regression of the phoneoytpe difference between siblings on the estimated proportion of marker alleles shared IBD between siblings. Only children data are used.
  - Association analysis: QTDT, using data of all members of CEPH families.
- Results:
  - Linkage results: 70 expression traits have proximal regulators, 1,574 have distal regulators, and 37 have both. 94% of distal regulators are in a different chromosome.
  - Association analysis of cis- linkage peaks: 63 of 100 expression traits with local eQTL show significant evidence - differential allelic expression by RNA-Seq.
  - Association analysis of trans- linkage peaks: for 1,611 (1,574 + 37) expression traits, define the trans-linkage peaks, then for each peak region, choose the SNPs of the candidate trans-regulators (coding and up/down-stream 5kb), and do association. at FDR 0.08, 917 expression traits have significant association with 742 trans-regulators, out of which 161 influence the expression of two or more genes.
  - Molecular validation of regulatory relationship: some pairs are known in literature (e.g. myoctye enhancing factor 2A, MEF2A and myosine regulatory light chain 2, MRLC2). Choose 25 regulators with modest evdience of linkage and association (QTDT P-value from 10^{-5} to 10^{-2}), and do knockdown (only successful in 18 regulators). 13 target genes show significant change of expression after knockdown of regulators. Also verify INSR as the regulator of expression of 4 other genes in primary fibroblast.
  - Trans-regulator analysis: among 742 trans-regulators, 15% are TFs, 19% play signaling roles, the rest in metabolism, protein transport or modification in ER, etc. Also the target genes and trans-regulators are often in the same functional pathway/GO.

Monocyte eQTL [Zeller & Cambien, PLos ONE, 2010]:

- Methods: monocytes of 1,490 unrelated individuals in GHS, 12,808 expression traits. 675K SNPs.

- Results:
  - eQTLs: at $P < 5.78 \times 10^{-12}$, 37,403 associtions between SNPs and expression, invovling 29K SNPs and 2,745 expression traits. The number of cis- and trans- regulated expression traits are 2,477 and 349, respectively. Changing the threshold has a strong effect on cis/trans ratio (Figure 1).
  - Comparison with previous eQTLs: $>$ 50% of previousely identified cis-eQTLs ([Stranger07, Dixon07, Schadt08]) were replicated in GHS. However, trans-eQTLs were hardly replicated.
  - Application to GWAS results: check if GWAS SNPs are associated with any expression traits in GHS and if yes, test if the expression traits are associated with risk factors (BMI, blood pressure, etc.). Very few GWAS results were compatibile with an effect mediated by gene expression at the locus. It's possible that monocyte is not the relevant tissue.

QTL of DNA methylation and expression in brain [Gibbs & Singleton, PG, 2010]:

- Methods:
  - Data: cerebellum, frontal cortex, temporal cortex, and pons regions of 150 individuals (600 tissue samples). Test associations of 22,184 genes and 1,629,853 SNPs (after imputation and quality filtering).
  - Statistical analysis: linear regression with additive model (regression of allele dosage and trait). Report both nominal P-values and empirical P-values (1000 permutations).
- Results:
  - Comparison of expression across brain regions: broadly similar (Figure 1D). Measures within frontal and temporal cortices were consistently the most alike and cerebellar tissue provided the most distinct profile of the four regions.
  - Significant associations: using FDR threshold based on empirical P-values (perhaps 0.05, a conservative threshold), about 5,000 associations per brain region, out of which about 75% are cis-associations (Table S4). Number of transcripts from this analysis: ranging from 280 (3.2%) in the pons to 391 (4.2%) in the temporal cortex.
  - Comparison of eQTL across regions: The majority of large effect and many moderate effect QTLs were shared across the four brain regions (comparing the R^2 value across tissues).

The architecture of gene regulatory variation across multiple human tissues: the MuTHER study [Nica & Spector, PG, 2011]:

- Methods:
  - Data: female Caucasian twins aged between 40 and 87 years old (mean 62 years) from the UK Adult Twin registry, 156 LCL, 160 skin, 166 fat samples. 865,544 SNPs with MAF $>$ 1% passed quality check (QC), on 18,170 genes.
  - Statistical analysis: Spearman rank correlation (SRC) on cis-associations (1Mb) for each tissue separately, P-value based on 10K permutations (permutation threshold, or PT, at 10^{-3}).
- Results:
  - Validation of cis-eQTLs in twins: use the twin data for validation (one analysis would use only one of a twin), the estimated proportion of true associations is very high: 0.93 in skin and 0.98 in LCL and fat. Even with eQTLs not replicated in twins, the estimated proportion of true associations is high: 0.84 for skin and 0.94 for LCL and fat.

- Validation of cis-eQTLs with earlier datasets: 40% of the genes for which we detect LCL eQTLs overlap with eQTLs detected in HapMap individuals. Likewise, 36% of the cis associations detected by Gibson et. al. in LCL derived from 194 southern Moroccan individuals overlap with genes reported in our study.
- Significant associations and tissue specificity: at PT 0.0001, 106 genes (12.35%) are shared across all tissues, 139 (16.2%) are shared in at least two tissues and 613 genes (71.44%) are detected in only one tissue. SRC-FA (factor analysis) results confirm the estimated about 30% of eQTLs to be shared in at least two tissues based on threshold eQTL discovery. Tissue-specific effects are largely not due to tissue-specific expression of the underlying transcripts, but the specific effects of SNPs. Even statistically tissue-shared eQTLs have additional dimensions of tissue-specificity and their mere discovery in multiple tissues does not guarantee similar magnitude of consequences.
- Genetic architecture: 7% of the genes tested are regulated by more than one independent cis eQTL.

- Remark: the replication of LCL cis-eQTL from earlier studies: only genes overlap or the actual associations?

Gene Expression in Skin and Lymphoblastoid Cells: Refined Statistical Method Reveals Extensive Overlap in cis-eQTL Signals [Ding & Abecasis, AJHG, 2010]

- Background: (1) Psoriasis, an immune-mediated, inflammatory disease of the skin and joints. (2) Dimas09 estimated 69 to 80% of cis-eQTLs are cell-type-specific (LCL, T cell, fibroblast). But the overlap may be underestimated because of the low power of detecting eQTL.

- eQTL mapping: SNP-gene expression associations separately in normal skin ($n = 57$), in uninvolved skin of patients ($n = 53$), and in lesional skin of patients ($n = 53$). The score test in Merlin (fastassoc option), limited to cis (1M) and one best eSNP per gene. p-value threshold $9E-7$, with FDR 0.01 in the three skin types.

- eQTL overlap across three skin types: identified 331, 275, and 235 independent cis-associations in normal, uninvolved, and lesional skin, respectively. 95.1%, 96.7%, and 98.7% of the significant cis-eQTLs in normal, uninvolved, and lesional skin, respectively, were detected in the other two skin types (at $p < 0.05$).

- Method of adjusting for eQTL overlap in two studies: given two studies, suppose in Study 1, at p-value threshold α_1, the FDR is FDR_1. Among all eQTL passing the threshold in Study 1, the fraction π_{raw} is replicated in Study 2 at p-value α_2. Since the Study 2 has relatively low power, clearly, some of the eQTL in Study 1 may be true, but may fail to be replicated, leading to an underestimate of eQTL overlap. To correct for that, note in all significant eQTL in Study 1:

  - Fraction FDR_1: are false positives. Among these eQTL, α_2 will be replicated in Study 2.
  - Fraction $1 - FDR_1$: are true ones. Among them, fraction π are also eQTL in Study 2 with probability $power_2$ to be replicated; and fraction $1 - \pi$ are not eQTL in Study with probability α_2 to be replicated.

  Thus we have:

$$FDR_1 \cdot \alpha_2 + (1 - FDR_1)[\pi \cdot power_2 + (1 - \pi) \cdot \alpha_2] = \pi_{\text{raw}} \tag{7.54}$$

  This allows one to compute π. Note that to estimate $power_2$, we first obtain $power_{2\text{raw}}$: suppose we estimate the effect sizes of all significant eQTL in Study 1, and estimate the power under this effect size distribution and the sample size of Study 2. Two corrections are needed:

  - False positives: among all significant hits in Study 1, not all of them are true. We have this equation:

$$power_{2\text{raw}} = (1 - FDR_1) \cdot power_{2\text{approx}} + FDR_1 \cdot \alpha_2 \tag{7.55}$$

    Solving this to obtain $power_{2\text{approx}}$.

  - Winner's curse: the effect sizes of the top hits in Study 1 are probably overestimated, so need to correct it too. The idea is to split the Study 1 data into two parts, one for identifying the significant hits, the other for estimating effect sizes.

- eQTL overlap across LCL and skin: at various FDR_1 and α_2 thresholds, about 70% LCL eQTL are also eQTL of skin (raw estimate without correction is about 30 to 40%).

- Comparison of GWAS p-values of eQTL:

  - Extract independent eSNPs: from 9462 eQTL at $p < 9E-7$ (FDR 0.01), find 389 independent eSNPs, using linkage disequilibrium (LD) while favoring SNPs with stronger cis-association p values.
  - Compare the distribution of disease-association p values of eSNPs and non-eSNPs (randomly sampled SNPs, excluding those within 1 Mb of regions known to be associated with psoriasis): QQ plot shows a trend for eQTL SNPs to be more strongly associated with psoriasis than non-eQTL SNPs.
  - The overlap between eQTL signals and psoriasis associations: top eight in the QQ plot (Most Significant Psoriasis Association).

- Additional eQTL overlap analysis: the LCL cis-eQTLs in our analysis with cis-eQTLs identified in fibroblasts and T cells generated by Dimas. 65%-70% of significant LCL cis-eQTLs were also present in fibroblasts and T cells.

- Lessons:

  - Measuring overlap of associations (or in general true hypothesis) across multiple studies: must take power into account.
  - eQTL similarlity in normal and controls: the disease state may not significantly alter the eQTL.

Single-Tissue and Cross-Tissue Heritability of Gene Expression Via Identity-by-Descent in Related or Unrelated Individuals [Price, PLG, 2011]

- Estimation of h^2: use Icelandic data, about 700. LMM using IBD as relationship matrix.

- Total heritability h^2 and contribution from cis π_{cis}: in blood, 42% of genes have $h^2 > 0$ and in adipose, 63%. In blood, π_{cis} about 37%, and in adipose, 24%. In previous literature, 12% in LCL.

- Small s.e. of average h^2 in cis and large s.e. of average h^2 in trans (many genes, expression are affected by common factors, e.g. environment or biological networks, simultaneously).

- Contribution of transgenerational epigenetic inheritance: compare h^2 in cis from IBD and unrelated individuals, very similar, suggesting that transgenerational inheritance makes a small contribution.

- Cross-tissue comparison: 50% shared cis-heritability, and close to 0 for trans.

- Lessons:

  - Generally higher standard error in estimation of h^2 in trans, while h^2 in cis is relatively robust.
  - Cis-heritability tends to be shared across cell types, while trans not. Thus tissues with mixed cell types have lower contribution from trans and higher from cis (blood > adipose > LCL).

Dissecting the regulatory architecture of gene expression QTLs [Gaffney & Pritchard, Genome biology, 2012]

- Data: LCL eQTL, for cis-eQTL, defined as upstream 100kb or downstream 100kb. Imputation using 1000GP data, 13.6M SNPs per individual. eQTL mapping shows that imputation helps with finding the causal SNPs (higher significance from imputed SNPs).

- Model challange: the causal SNPs may not be known in each candidate region, so use Bayesian posterior prob. to account for the uncertainty.
- Bayesian hierarchical model:
  - For each SNP, we use BIMBAM to relate its genotype with expression. The prior of effect size (given that the SNP is causal) is a mixture of five normal distributions.
  - The prior of SNP (binary indicator, whether it is causal or not): logistic function of multiple annotaitons (DHS, histone marks, etc.). For each annotation, the parameter λ_i represents the enrichment of eSNPs in that annotaiton (interpretation is odds ratio of being an eSNP given that it has an annotation).
  - Empirical Bayes to estimate the parameters λ_i combining all genes.
  - One additional difficulty is that many annotations correlate with spatial distributions, and eQTL also tend to occur in certain distribution, e.g. close to TSS. So define a background model that accounts for this spatial distribution: compare the model using annotations with the background model (which favors close regions).
- Enrichment of annotations in eQTL: LCL eQTL data and ENCODE annotations
  - SNPs located within open chromatin are 4-times more likely to be an eQTL. Enrichment of other marks: approximately three-fold enrichment in H3K27ac; H3K4me1: about 2-fold. If limit to regions upstream more than 5kb of TSS: each of the three leads to 4-7 times enrichment (strongest for H3K27a). No enrichment of eQTL in regions with repressive marks.
  - Total: 20% of all eQTNs occur within DNaseI hypersensitive sites. Over 40% of all eQTNs occur within either a DNaseI hypersensitive site or within a histone-modified region: even though the regions cover only 4.5% of the genome.
  - Enrichment in TFBSs: strong enrichment in c-Jun and NK-kb binding sites (ChIP-seq)
  - A large fraction of eQTNs occur very close to the TSS, and presumably affect the core and distal promoter architecture.
- Sequence conservation: using PhastCons, PhyloP and conserved TFBS, surprisingly little enrichment (no enrichment at larger distance from TSS). Likely due to correction for distance in the background model.
- Combined model:
  - Include all annoations including DHS, histone marks, motifs, TF ChIP-seq. Use AIC to select model. 10-fold cross validation shows that the combined model is better than the single or background model (higher likelihood).
  - Selected annotations: in regions > 5 kb from TSS, H3K27ac is the dominate one, other features add relatively little.
  - The combined model aids the identification of causal eSNPs.
- Lessons:
  - When there is a significant uncertainty of assigning to categories, frame enrichment testing problem as inference of parameters.
  - For distal enhancers, H3K27ac captures most of the information (comparing with DHS and other histone marks).
- Remark:

- Lack of signal in conserved regions: could be due to evolutionary selection against variation of sequences in these regions, thus eQTL (which are common SNPs) are depleted.

Mapping cis- and trans-regulatory effects across multiple tissues in twins, [Grundberg, NG, 2012]

- Data: Adipose, LCL, skin. 856 twins (1/3 MZ, rest DZ)
- Distribution of h2: average 0.26 (adipose) 0.16 (skin). Figure 1A.
- Cis-eQTL mapping: linear mixed model. High replication rate > 0.7.
- Estimated degree of overlap between tissues: Table 1. Two approaches: threshold and estimated proportion. Estimate that $> 60\%$ of cis-eQTL have effects in multiple tissues.
- Estimated heritability explained by cis-eQTL: $< 15\%$ from common SNPs. Increase to 30% or so with linkage: rare variants.
- Trans-eQTL: $> 60\%$ likely due to trans-eQTL. Found 500-1000 trans-eQTL in three tissues at $P < 5E-8$. Many of these trans-eQTL are associated with multiple genes.

Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs, [Brown & Engelhardt, PLG, 2013]

- Data: eleven eQTL studies from seven unique cell types, LCL, brain, liver, blood fibroblast and T cells.
- eQTL mapping: uniform pipeline.
  - Gene expression array: uniformly processed. Imputation.
  - Control for the confounding effects of both known covariates and unknown factors by removing the effects of principal components. For each dataset, do the regression analysis, controlling for covariates and PCs. Then project residual expression variation to the quantiles of a standard normal distribution to control for outliers, and used these projected values as the quantitative traits for association mapping.
  - Results represented by BF using BIMBAM. Assess FDR by permutation.
- eQTL results and replication studies across cell types:
  - Across all 11 studies, 29% of eQTL associated genes are independently associated with at least two SNPs in at least one study. In one study, such fraction is 3-22%.
  - Replication experiment: create trios, e.g. two LCL studies plus a liver study. Replication frequency is higher between the same cell types than between different cell types, and depends on a number of factors: log-BF, distance to TSS, etc. At high BF (log10-BF > 10), high replication between the same cell types (50% or higher), but lower between different cell types (20-40%). This suggests that cell-type specific eQTL tend to have smaller effects.
- Intersection with CRE annotations to understand cell-type specificity.
  - Define activating CREs and repressive CREs, and intersect eQTL with annotations from LCL and liver.
  - cis-eQTL enriched for overlaps with several classes of CREs, including DHS sites, and depleted within regions in which a CTCF binding site lies between the eQTL SNP and the target gene TSS. Almost universally, QTL SNPs are enriched within regions of activating CREs and depleted within repressive CREs. eQTL-CRE enrichment peaks immediately adjacent to the TSS for several classes of activating CREs, including H3K4me3 and H2A.Z.

  - Intersected eSNPs tend to be cell-type specific: significantly more overlap between eQTL and CREs (DHSs) derived from the same type than from different cell types. Ex. SORT1 eQTL overlaps with a cluster of liver enhancers (but not in LCL).
  - The proportion of eQTL SNP - TSS pairs with intervening insulators is remarkably consistent across cell types, suggesting that CTCF binding sites do not substantially affect cell-specific eQTL function.
  - Random forest classifer to predict if an eQTL is likely to be active in second cell type, using CRE data from the second cell type.

- Lessons:
  - eQTL mapping across multiple datasets: control for confounding and hidden confounding variables.
  - Cell-type specificity: large-effect, proximal eQTL are more likely to be common across multiple cell types. A significant fraction of cell-type specific eQTL (at least 50%) and they tend to overlap with CREs.

Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression [Fairfax and Knight, Science, 2014]

- Hypothesis: many SNPs have effects only under appropriate stimulations. So if we map eQTL in stimulated conditions, we will find identify more eQTL.
- Data: 228 individuals, treatment with LPS and IFN-γ, a total of four conditions (treated or resting).
- DE gene analysis: DE of canonical pathways, consistent with expectation of LPS effect.
- Condition-specific eQTL: map cis-eQTL in each condition separately, and use surrogate variable analysis to increase the power of eQTL mapping. The majority of cis-eQTL were observed only after stimulation. And 54% of resting eQTL were absent in stimulated condition. Limit to genes expressed in most samples: 33% show eQTL after treatment. The stimulated eQTL include many immune-related genes such as TFs, key cytokines and receptors.
- Trans-eQTL: some cis-eQTL are also trans-eQTL, possibly at a later time point. Ex. after 2-hr LPS treatment, cis-eQTL of IFNB1 was associated with expression of several genes at 24-hr in trans.
- IRF2 cis-eQTL and direct targets: IRF2 cis-eQTL associated with 300 trans-genes and located in DHS. ChIP-seq targets of IRF2 enriched with trans-associated genes.
- Utility for GWAS: overlap of cis-eQTL with GWAS, enriched for phenotypes such as bacterial infection, inflammation. Many GWAS loci overlap with eQTL specific to induced cells.
- **Lesson**: eQTL and GWAS loci may become effective only under certain conditions/stimulations.

Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals [Battle & Koller, GR, 2014]

- Background: ASE is defined on individual data. It was found that imprinting or other non-genetic factors explain only a small fraction of ASE. Also ASE is generally not caused by the trans-acting factors, so an observed ASE is likely due to some genetic difference between alleles, or cis-mutations that cause different expression (cis-eQTL).
- Data: RNA-seq from whole blood in 922 genotyped individuals from the Depression Genes and Networks cohort.
- eQTL mapping: found eQTL in the large majority (78.8% at FDR 0.05) of genes with quantifiable expression. Nearly half of SNPs in GWAS catalog are associated with some expression.

- Proximal regulatory variation:
  - cis-eQTLs explaining a median of 3.3% of expression variance (median 7.7% among genes with an eQTL), compared to 0.7% explained by age and sex combined.
  - Comparison with earlier studies of LCL: high replication rate ranging from 51% to 89%.
  - sQTL: using isoform ratio as a quantitative trait, found 2851 transcripts from 1370 unique genes with sQTL at FDR 0.05. Example: a SNP associated with LOAD is a sQTL, much stronger than its effect as eQTL.
  - ASE: found a set of regulatory variants consistently associated with allelic imbalance in nearby genes. Most ASEs in individuals can be explained by eQTL: 74% of individual ASE events co-occurring with heterozygous status for the single best cis-eQTL SNP of the same gene. The remaining cases are candidates for rare regulatory variation.
  - aseQTLs: associations between heterozygous status at individual regulatory variants and allelic imbalance at nearby expressed coding loci. Confirm that 641 of our cis-eQTL SNPs are also associated with changes in allele-specific expression in the corresponding gene at FDR 0.05.
- Distal regulatory variation:
  - Intra-chromosome eQTL: 381 genes affected by SNPs $>$ 500 kb away from TSS, including 269 genes affected by SNPs $>$ 1 Mb away. Find modules of coregulated genes, with 803 eQTL variants affecting two or more genes and 106 variants affecting three or more. Ex. (Figure 2B) rs11644386 affects a discontinuous group of genes, with the farthest association (CYLD) being $>$ 400 kb away, and does not have significant associations with two intermediate genes SNX20 and NOD2.
  - Trans-associations: 138 genes whose expression is associated with a distant SNP; and 5 trans-sQTL. Evidence of modularity: 20% of associated SNPs affecting two or more genes. The largest module is a set of 57 genes, enriched for platelet aggregation function ($P < 10^{-7}$), all affected by the SNP rs1354034, previously associated with mean platelet volume.
  - The majority of trans-eQTLs SNPs (76 of 138) also have cis-regulatory effects. Example: rs10251980, is a cis-eQTL for IKZF1, whose loss of function has been linked to leukemia (Mullighan et al. 2009). The SNP affects eight distant genes, five of which are up-regulated in response to tretinoin treatment in leukemia.
  - For trans-eQTL, the expression level of nearby genes mediate the trans effect 85% of the time.
- Natural selection on eQTL:
  - Negative correlation between effect size of cis-eQTL and MAF.
  - Depletion of cis-eQTLs among genes with annotations suggesting critical roles in cellular functioning: highly conserved and hubs in PPI network, and TFs. May explain the scarcity of trans-eQTL.
- Genomic properties/mechanisms of eQTL:
  - Position: QTL enrichment near TSS. sQTLs are concentrated among exonic and intronic loci. Splice site, essential splice site, and stop gained functional annotations are particularly enriched for sQTLs, beyond the effects of position.
  - Regulatory annotations: both eQTL and sQTL are enriched in TFBS ChIP and DHS. Role of TF binding in splicing: could be regulation of expression of particular isoforms under different conditions (mechanisms such as cotranscriptional splicing)
- Using genomic annotations to predict regulatory SNPs: LRVM. Given a gene and its adjacant SNPs (20kb near TSS), we define a_i to be indicator of whether SNP i is associated with the gene. For each SNP, we also define d_i as its intristic regulatory potential (binary). The variable d_i is related to annotation/feature vector of i. We assume that a_i is related to d of all SNPs, accounting for LD and

MAF. Use the observed a_i for each gene to learn the importance of features for regulatory potential d_i.

- Remark/Questions:
  - For eQTL of co-regulated genes: control for correlated expression, or independent association of SNPs with expression of multiple genes?

Heritability and genomics of gene expression in peripheral blood [Wright, NG, 2014]

- Estimation of h^2: three strategies:
  - ACE model: let y_i be phenotype, x_i covariates, a_i, c_i, e_i as additive, common environment and unique environmental contributions:

$$y_i = \mu + x_i\beta + a_i + c_i + e_i \tag{7.56}$$

    The difference in MZ and DZ twins is the distribution of a_i: $a \sim N(0, \sigma_a^2 A)$, where A is the genetic relationship matrix using the expected value.
  - IBD model: similar to ACE, except that A is defined by the actual IBD between DZ twins.
  - GCTA model: use unrelated individuals only in the analysis.
- Data: Netherlands Twin Registry (NTR), 1,308 pairs including 690 MZ pairs and 618 DZ pairs.
- Distribution of h^2: mean 0.10, +/− 0.142. 777 genes with h^2 significantly greater than 0 at $q < 0.05$. h^2 correlated with expression mean and variance. Stabilizing estimation of h^2: use herirachical model, assume a Gamma prior. With this model, the proportion of genes with $h^2 > 0.3$ is 7.9%.
- Local genetic contribution: use GCTA to estimate contribution of cis-eQTL (use cis-SNPs for GRM). This ratio is only 0.04 (mean) or 0.09 (median). Use local IBD approach, 0.11 (median) or 0.3 (mean).
- Mapping and validation of trans-eQTL: at $q < 0.001$, found about 600 eQTL, then apply additional QC, 348 robust trans-eQTL.
  - Additional QC: SNP and genes in LD; SNP alone (likely genotype quality problem), probe cross-hybridization and adjusting for local SNP.
  - Comparison with Westra data: in the eQTL found in Westra, estimated true discovery rate in NTR is about 23%.
  - Pleiotropicity of trans-eQTL: on average, π_1 (fraction of associated transcripts) is about 0.001 to 0.008.
- Remark/Question:
  - Lack of correlation between h^2 from GCTA and from twins?
  - Replication of trans-eQTL: how is it calculated for Westra data?

The human transcriptome across tissues and individuals [Mele, GTEx, Science, 2015]

- Tissue specificity of genes: use RPKM > 0.1 as threshold (80M reads). 88% of protein-coding genes and 71% of lncRNAs are expressed in at least one sample. For many tissues, expression are dominated by < 100 genes, e.g. hemoglobin in blood. 200 genes are exclusively expressed in one tissue, most in testis.
- Variation of gene expression across tissues and individuals: use LMM, 47% variation from tissue and 4$ from individuals. Genes with large individual variation: many in sex chromosome.

- Correlation of expression with age: 2,000 genes correlate with aging. Top one: EDA2R, and EDA is associated with age-related phenotype.
- Sex-specific expression: mostly sex-chromosome (X), suggesting escape of X-inactivation.
- Alternative splicing: the variation of isoform levels across tissues, largely explained by the gene expression (total isoform).

The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans [GTEx, Science, 2015]

- Data: 237 donors, 28 tissues per sample, 6.8M SNPs, with MAF greater than 5%. RNA-seq: 76bp PE, 82.1M reads per sample, use RPKM > 0.1 as threshold. Tissues: 29 solid, 11 brain regions, whole blood, LCL, skin.
- Single tissue eQTL: cis-eQTL only, use Matrix eQTL. To correct for multiple testing, permutation for the most significant SNP per gene. The number of eGenes range from 900 in heart to 2200 in thyroid, with a total of 6K eGenes in 9 tissues. The majority of cis-eQTL cluster around TSS.
- Multi-tissue eQTL:
  - Pair wise analysis to assess eQTL sharing: first find significant eSNPs in one tissue, then use the distribution of p-values of these pairs in the second tissue to estimate π_1, the proportation of non-null pairs. π_1 ranges from 0.54 to 0.9.
  - More than 50% of all detected eQTL are common to all nine tissues.
  - UNC approach: use minP across tissues as test statistic, to permutation. Found 7425 eGenes with FDR < 0.05 - 3 fold increase relative to the number for single tissue.
  - Chicago approach: 10K genes show a signicant eQTL at the same FDR.
- ASE:
  - Fraction of ASE: median of 6K sites that are heterozygous on one sample. About 2-3% are ASE. Brain shows the lowest ASE.
  - What drives ASE (tissue vs. individual): average correlation of total reads (expression level) or allelic ratio between samples, either from different individuals, or from different tissues (three cases). For total count, higher correlation between samples than between individuals. For allelic ratio, higher correlation between tissues of the same individuals, suggesting ASE is primarily determined by the genome.
  - ASE validation of cis-eQTL: NDRG4 example, allelic ratio in the heterozygous (for the cis-eQTL) higher than 50%.
- Splicing QTL:
  - Methods: Altrans - association with expression levels of exon junction, both novel and annotated splicing. sQTLSeekerR - association with isoform ratios, only annotated splicing forms.
  - Average of 1900 genes with sQTL using Altrans, and 250 with sQTLseekerR. The sQTL detected by two methods follow different patterns, in particular not all sQTL involve different exon usage, e.g. complex 3' event. Most are not tissue-specific.
- Gene co-expression network: for any single tissue. The network overlap substantially across tissues: about 0.3 to 0.58 correlated genes are still correlated in a second tissue. Use WGCNA to find the modules: enriched with GO and ENCODE TF binding.

- Tissue-specific expression profiles: for any individual, we can group all genes by their tissue-expression profiles, then we merge such clusters from all individuals. The individual variation of clusters is small. But there are genes (21%) that change modules across individuals. Identify SNPs correlated with module membership scores: modQTL, 58% of which are not eQTL.
- GWAS and eQTL: enrichment of GWAS signal in eQTL from specific tissues. In 34% of cases, eQTL-gene are not the nearest gene to SNP.
- Remark/Questions:
  - Most eQTL are not tissue-specific, but eQTL-GWAS enrichment analysis tend to be quite tissue-specific. And similarly enhancers are often tissue-specific. How to reconcile? Probably enhancer-eQTL are more tissue-specific than promoter-eQTL, and are more relevant to diseases?
  - What drives ASE? Only from cis-eQTL? The contribution from exonic SNPs (ex. NMD)?

Effect of predicted protein-truncating genetic variants on the human transcriptome [Rivas & MacArthur, Science, 2015]

- Goal: protein-truncating variants (PTV) on transcription.
- Data: GTEx, Geuvadis, 462 individuals with WGS and LCL RNA-seq. Variant discovery: frameshift indels disagree with two datasets (100 vs. 16 per subject)
- Transcripts with PTVs are expressed at lower levels and more tissue-specific (more tolerant).
- Splice junctions are less often used.
- Nonsense-mediated decay (NMD) pathway: exon juction (EJC). If a transcript has a premature stop-codon before EC, then SURF complex recognizes the stop codon and trigger NMD. Frameshift indels can cause NMDs: will always hit a stop codon prematurally.
- Calling indels: create reference genome that contains indels (heterozygotes)
- Carole: easier to call deletions than insertions.
- mmPCR-seq: microfludic PCR-seq, targeted RNA-seq. Validation of allelic ratio.
- NMD-inducing variants tend to create more ASEs.
- 50bp rule trigger NMD: only termination codons located more than 55bp upstream of 3'-most exon-exon juction trigger NMD.
- Validation of the rule: normal ¡ NMD-escaping SNVs ¡ NMD-triggerin SNVs.
- Of the ones triggering NMD: about 30% show no ASE.
- Predictive model of NMD: 38 features, 50bp rule, distance to start/stop codon, etc, by Random Forest. Important features: number of downtream exons, distance to donor site. Implication: genes with large number of exons more likely to have NMD.
- Dosage compensation of PTVs: rare.
- Splicing-disrupting variants have signiature outside essential sites.
- Question: fraction of nonsense that trigger NMD? What explains tissue-specificity? 158 * 40

A systematic heritability analysis of the human whole blood transcriptome [Huan & Levy, Human Genetics, 2015]

- Data: FHS, known pedigrees, about 5000 in 700 families and 400 unrelated. Whole blood.
- Method: (1) Estimation of h^2: variance component using pedigree. (2) Mapping eQTL: LMM, adjusting for relatedness, PC, cell type, etc.
- Overall heritability: mean 0.07, 40% genes have $h^2 > 0$ and 10% genes have $h^2 > 0.2$.
- Cis- vs. trans-eQTL: at FDR < 0.001 threshold, (1) High h^2 genes have more cis-eQTL; (2) High h^2 genes have fewer trans-eQTL. In total, 3% genes have trans-eQTL, among genes with $h^2 > 0.2$, 21% has trans-eQTL.
- Overlap of eQTL with GWAS loci of metabolic trait: a number of examples of trans-eQTL that are also GWAS, highly significant in both. Figure 4: a SNP in a regulatory gene is trans-eQTL of three other genes, and the SNP has $p < 10^{-9}$.

Distant Regulatory effects of genetic variation in multiple tissues [Jo and Battle, bioRxiv, 2016]

- Trans-eQTL mapping in GTEx data:
  - Found about 100 eGenes in 44 tissues, most tissues have < 5 eGenes. Testis has 28.
  - If limit to cis-eQTL: add 14 new eGenes.
- Properties of trans-eQTL: (1)Tissue specificity: higher than cis-eQTL; (2) Higher enrichment in enhancers than cis-eQTL.
- Pleiotropic effects of trans-eQTL: estimate $(1 - \pi_0)^{27}$ (effect in at least one more tissue), about 3%, much higher than cis-eQTL (close to 0).
- Replication: a separate dataset, substantial enrichment of low p-values.
- Examples of trans-eQTL: (1) FOXE1 locus in thyroid, broad transcriptional effect, corrected by PEER factors. Found 1085 uqniue trans-eGenes (FDR < 0.1). (2) A SNP at KLF14 locus, also cis-eQTL, associated with many genes in trans in adipose.

Systematic evaluation of genetic correlations between expressed transcripts in peripheral blood [Lukowski, review for NC, 2016]

- Data: CAGE data, 1,748 unrelated individuals, peripheral blood. Limit the analysis to 2,469 transcripts with $h^2 > 0.25$.
- Predicting pairs of genetically correlated transcripts: GREML, estimate the correlation of genetic (random) effects between two traits from all SNPs. Let r_P, r_G be phenotypic and genetic correlation, h_i, h_j be the heritability of two transcripts and e_i, e_j be the environmental contribution, we have

$$r_P = h_i h_j r_G + e_i e_j r_E \tag{7.57}$$

  where r_E is the environmental correlation.
- Results of genetic correlation analysis: among 2M pairs tested, 556 pairs with Bonferroni threshold, and 15K pairs with FDR < 0.05. For the strongest pairs (former), about 1/2 are trans-pairs; and for the latter, 94% are trans.
- Identifying shared eSNPs: to show that the results are replicated, identify shared eSNPs in a different study (2000 unrelated individuals). Method: for each pair, find the strong SNP in one, then test if it is associated with the other. Found 934 eSNPs with a significant effect in the second at Bonf. threshold. Next show that among 934 eSNPs (half trans), 100% replicated in CAGE.

- General issue of trans-eQTL replication: use Westra trans-eQTL, found large effect eQTL are often replicated, and the percent of replication correlates with effect sizes. Ex. at $z > 10$, most eQTL are replicated in at least one eQTL dataset.
- Genetically correlated transcripts are enriched in chromatin-interaction regions. In Bonferroni pairs of transcripts, many are located in interacting loci from Hi-C, representing 25-90 times enrichment. Interpretation: these likely represent pairs of transcripts that are regulated by the same sequences (or eQTL).
- Connectivity of genes in genetic correlation network: for each transcript, count the number of connections with $r_G >$ a threshold (2 or 3 σ, where σ is the sd of r_G). At $r_G > 2\sigma$, expect a gene to have 113 connections, and 7 for 3σ. Found 2,317/2,468 transcripts with more than 113 connections and 2,397/2,468 with more than 7 connections.
- Example from network analysis: two transcripts of the same gene, have different connections: some shared ones, but most are unique.
- Hubs are enriched with TFs. Also evidence that pairs tend to share GO terms.
- Remark/Lessons:
  - Trans-eQTL can drive the discovery of correlated genes; and can be replicated.
  - Significant number of co-regulation are local (perhaps due to common CREs).
  - High level of co-regulation among genes.
  - Need more analysis on the biological significance of co-regulation.

Heritability and Sparse Architecture of Gene Expression Traits [Wheeler, PLG, 2016]

- Comparison of Polygenic and BVSR priors; also Elastic net.
- Fit each gene separately: adjust for cross-tissue term for each individual (shared across tissues of the same individual)
- Cis-h2 analysis: genes with high h2 tend to be less constrained (measured by pLI).
- Method to study sparsity: estimate PGE of sparse and polygenic components using BSLMM.
- Sparsity estimates using BSLMM in DGN data: (1) high h2 genes has high PGE (proportion of variance explained by sparse component); (2) low h2 genes: cannot estimate well the PGE.
- Lasso predicts better than Elastic Net: therefore more sparsity.
- Orthogonal tissue decomposition: identify cross-tissue components.
- Remark: the analysis using EN doesnt take power into account. With more samples, the polygenic component could benefit more, and higher PVE.
- Remark: the Cross-Tissue component captures shared expression, but could just reflect individual covarites (hidden). Not represent the shared eQTL effects.

Genetic mapping of the plasma proteome to inform drug development [Joseph Maranville, HG seminar, 2017]

- SomaLogic platform for measuring 3K proteins in 3K individuals. DNA-based aptamers: barcoded, bind to proteins. Find the aptomaer that bind to protein. Potential problem: protein complexes may bind to aptamers together. But this is likely rare.

- pQTL mapping: Found 2K pQTL in 1.5K proteins, both cis and trans. Caution: cis-pQTL in coding sequences can alter aptamer binding.
- Intersection with disease GWAS: 88 pQTL, 253 GWAS signals. Do coloalization, 61 pQTL, 11 cis and 50 trans.
- Example: cis-pQTL of IL23 and IBD GWAS. Cis-pQTL of MMP12: disprove the gene of CVD gene by MR.
- Drug target analysis: among 600 targets, 111 have cis-pQTL. Could use them as IVs to study their impact on GWAS.
- A cis-pQTL of MST1: the gene binds to RON receptor, and RON KO has IBD phenotype.
- The cis-pQTL is also trans-pQTL of 8 other proteins.Near 8 trans target genes: 3 independent GWAS signals.
- Model: MST1 affects protein, then ROS signaling, which affect several trans-targets, some of which influence IBD.
- Comparison with eQTL: 60% cis-pQTL has no cis-eQTL (could be lower power).
- Remark: in the case of colocalization with trans-pQTL, it is hard to find the causal gene.
- Remark: a better analysis about the relationship of cis-eQTL and cis-pQTL is: heritability of pQTLs mediated through eQTLs.

Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies [Joehanes and Munson, GB, 2017]

- Framingham Heart study (FHS): 5K samples, whole blood eQTL. QC: polymorphism-in-probe effect is likely minor.
- Mapping eQTL: (1) Adjusting for observed covariates and family relationship with LMM: obtain residuals. (2) Adjusting for hidden covariates using PEER factors, and obtain p-values. (3) Multiple testing correction: for all associations with $p < 10^{-4}$, do BH correction to obtain FDR. Also adjusting for p-value inflation by Genomic Control, however, the results are only slightly affected. Note: correction is done separately for cis- and trans-eQTLs.
- Obtaining independent eQTL: step-wise regression to find independent eQTL. Found 19K independent cis-eQTL and 6K trans-eQTL. Sample size is important for the power: it scales linearly with cis-eQTL, but more with trans-eQTL. Double the size from 2500 to 5000 increase the trans-eQTL by 3-4 fold.
- Validation of eQTL: (1) Internal: 75% cis and 41% trans-eQTL are validated. (2) With previous studies: 50-70% previous cis-eQTL and 30-60% trans-eQTL are replicated. The replication in the other way is low due to lower power in previous work and different sequencing platforms, etc; but still 90% of cases the directions are consistent.
- Distribution of eQTL: highly enriched in transcribed regions, especially first exons and 5 UTRs (45 fold). Modest enrichment (2-fold) of trans-eQTL in regulatory regions.
- Clusters of trans-eQTL: 59 clusters with 6-200 genes. Some are due to genetic effect on cell type composition. In some clusters, found enrichment of TF motifs in promoters and miRNA targets. The majority 90% of trans-eQTL are not in any of the clusters.
- GWAS analysis: with CAD/MI 58 loci, 21 loci or 36% are lead cis-eQTL. Also an example where a SNP is the trans-eQTL of a cluster of genes (SH2B3 locus).

Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues [Xuanyao Liu and Price, AJHG, 2017]

- Using S-LDSC for estimating enrichment and heritability: let C be an annotation (possibly overlapping), and τ_C be per SNP heritability of SNPs in C. Then h^2 of C is the sum of SNP heritability of all SNPs in C. This allows us to define: (1) Proportion of h^2 by C: it is h^2 in C divided by h^2 summing over all categories. (2) Enrichment of C: defined as (1) divided by the percent of SNPs in C. Consider an example with two categories C_1 and C_2: let M_1, M_2 be numbers of SNPs in the two categories, and M_{12} be the SNPs in both. Let τ_1, τ_2 be per SNP heritability of C_1, C_2, then we have:

$$h^2(C_1) = \tau_1 M_1 + \tau_2 M_{12} \qquad h^2(\text{total}) = \tau_1 M_1 + \tau_2 M_2 \tag{7.58}$$

  where the second term of $h^2(C_1)$ is from SNPs shared with C_2.

- Applying S-LDSC to eQTL: apply it to all genes with > 0 heritability. Then average h^2 over all genes for each category separately. To obtain SE of estimates, using block jackknife: 200 genomic blocks (each block preserve cis-eQTL dependency).

- Estimation of enrichment in cis-eQTLs: include 50 annotations for joint estimation. Strongest: 5'UTR, TSS, conservation (7-10), promoter, enhancer (5). Estimates are generally consistent across different tissues, and different sample sizes.

- Genetic correlation between tissues: generally high between tissues in cis-eQTL, but very low (around 0.1) in trans-eQTLs.

- Remark: heritability explained by a category C is somewhat inflated, as it includes contribution of all SNPs in C, but some SNPs obtain bigger effects from other annotations.

Genetic effects on gene expression across human tissues [GTEx, Nature, 2017]

- Procedure for multiple testing correction: control number of eGenes at a given FDR. Use fastQTL: fit the null distribution of min. p-value per gene using Beta distribution. Then control FDR using q-values.

- Defining eVariants: (1) Determine the global threshold for min. p-values p_t: find a gene closest to FDR 0.05, and the empirical p-value for that gene. (2) Need to determine the threshold for each gene: from the null distribution of min. p-values, determine the threshold for p_t (convert p_t to the threshold using CDF). Then any SNPs below this threshold will be selected as eVariants for that gene.

- Identification of additional eQTL per gene: using forward-backward stepwise regression.

- Results of cis-eQTL per tissue: generally a few thousand eGenes at FDR < 0.05 per tissue. Replication by ASE: among ascertained eVariants, how often they are replicated in ASE (at nominal $p < 0.01$). The replication rate drops with distance to TSS, but saturated at 1.3Mb.

- Enrichment of cis-eQTLs in functional annotations: Enrichment test: compare percent of eQTLs in an annotation vs. control SNPs to estimate log-OR. Do this for Roadmap annotations. Enrichment is higher (could be > 10) for matched tissues (Figure 3a). Also do this for multiple variants per gene: generally higher enrichment in promoters than enhancers; but for secondary SNPs, the gap is smaller (Figure 3c).

- Shared eQTL across tissues are more likely to share CREs (Figure 3b).

- Fine-mapping and validation: use CAVIAR to fine-SNPs, and obtain PIPs of all in the credible sets. Correlation of PIPs with the proportion of eQTLs localized in DHS (Figure 3d).

- Characterizing contributions of different functional elements to cis-eQTLs: compare effect sizes of eQTLs in 3 UTRs, exons, splice sites, noncoding, etc. Smaller effects in untrasnlated regions comparing with upstream regulation (Figure 3e), except canonical splice sites.
- Replication of effect sizes in ASE: highly correlated effect sizes (Extended Data Figure 6)
- Trans-eQTL: 673 trans-eQTLs found, 113 have cis-associations. Removing PEER factors (15-35 factors) have major effects on trans-eQTL detection: Figure S13, in testis and stomach, much fewer trans-eQTL after removal. Testis has a large number of trans-eQTLs: about 28% (12 independent trans-loci) overlap with piRNAs.
- Enrichment of GWAS variants in eQTL: (1) eGenes: tissue-shared eGenes are less likely to be in OMIM or LoF intolerant. (2) 50% of GWAS loci are associated some gene expression: at $p < 0.05/44$ (44 is number of tissues). (3) In these variants, even with very strong cutoff, 10-20% cases the top genes vary across tissues (Figure 5c)
- Co-localization analysis of GWAS and eQTL: Figure 5d, 21 traits, about 50% of GWAS loci show colocalization; about half of cases are in nearest genes.
- **Lesson**: strategies for validation of cis-eQTLs (1) Use an independent dataset, e.g. ASE. Higher pi1, or replication rate; effect size correlation. (2) Spatial distribution: highest near TSS, but also saturate as one move to large distance. (3) Enrichment of functional annotations. This analysis can use different tissue-specific annotations (irrelevant tissues serve as negative control).
- Lesson: validation of fine-mapped SNPs, correlation of PIPs with functional features.

Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression (DICE) [Schmiedel and Vijayanand, Cell, 2018]

- Data: 91 samples, three innate immune cells, classical and non-classical monocytes and NK cells. B cells, naive CD4, CD8 and T-reg cells. Also six memory and differentiated T cells and two activated T cells.
- Results of eQTL mapping: about 2k-3k eGenes per cell type.
- Cell type specificity of eQTLs: many cell type specific effects, plot Z-scores across cell types (large blocks); in contrast, cell type specificity of RNA levels are lower (Figure 3F). Some examples: Figure 3G, GAB2, eQTL has opposite effects in B and T cells. Many eGenes are found only in activated T cells.
- Integration with GWAS: some example, LACC1 is involved in controlling T cell response. Found eQTL of LACC1 (only in T cells), also GWAS SNP of AID. The eQTL reduces expression of LACC1 in activated T cells: confirm that K.D. of the gene reduces T cell production of cytokines after stimulation.
- Lesson: cell-type specificity of genetic effects are probably higher than cell-type specificity of gene expression (or any molecular traits).
- Remark: missing the global pattern of cell-type specificity of eQTLs.

7.2.2 Regulatory QTL studies

Variation in Transcription Factor Binding Among Humans [Kasowski & Snyder, Science, 2010]

- ChIP-seq data of NF-kappB and RNA Pol 2 in 10 human and 1 chimp.
- Calling peaks from ChIP-seq:
  - Calling peaks in each sample with PeakSeq at $p < .001$.

  - Cluster peaks from different replicates of the same individual into BRs. Peaks not replicable in at least two were discarded.
  - Join BRs across individuals by combining intersecting events

- Test differential binding:
  - Normalization: normalize numbers of reads in a BR across all replicates and samples using quantile normalization.
  - Comparison between individuals: ANOVA, using a generalized linear model with Poisson error distribution, with Bonferroni correction (N being the number of pair-wise comparisons made).

- Spearman correlation of binding across replicates: median 0.95; correlation of binding across individuals: median 0.90.

- Difference between individuals: 7.5% and 25% of the NFB and PolII binding regions, respectively, differed significantly between two individuals (Poisson ANOVA comparison). The signal intensity in "lost" peaks are similar to background level, suggesting that the peak is completely absent, instead of the threshold effect.

- Variability of BRs: BRs near TSS show less variability than intergenic ones.

- Human-Chimp comparison of Pol 2 binding: analysis of human BRs with syntentic regions in chimp (81%). Binding differences between the chimpanzee cells and each of the ten human samples were identified: on average 32% of BRs show significant differences in binding (corrected $P < .05$).

DNase I sensitivity QTLs are a major determinant of human expression variation [Degner and Pritchard, Nature, 2012]

- dsQTL mapping: use non-overlapping 100bp windows as test units, and top 5% of windows in DNase reads as phenotypes. For each peak, test association of normalized read counts with SNPs within 40kb windows.

- Read count normalization: correct for library size and GC content first. Then standardized and quantile normalization. Finally, adjust for PCs: varying the number of PCs.

- Allele specific analysis: for only significant dsQTLs. Only consider those with 90 reads or more, then estimate the proportion of major alleles. To compare with dsQTL: now use raw read counts as the dependent variable and do linear regression. Use the estimated parameters to estimate the proportion of major alleles.

- Calibration of null distribution of p-values: Figure S13. P-values are uniform under two types of permutations: for each test window w, either permute genotype labels of samples; or test association of w to the SNPs in a random genomic location. Note: in sample permutation, pheotypic relations (among peaks, and PCs to peaks) are preserved.

- Determining statistical significance: significant associations found at FDR 10% using q-values. Correct for windows of different intensities (10 bins) separately, using different significance threshold varies (because of power difference). Top windows: many more significant associations at the same FDR (Figure S12). Most associations are found for top 1% of windows.

- Validation of dsQTLs: QQ plot of dsQTLs 2kb (stronger signal) vs. 40kb (Figure 1a). Correlation of effect size of dsQTL and allele-specific effects, $r = 0.72$ (Figure 1b).

- Distribution of dsQTLs in functional sequences: 41%in predicted enhancers, 26%in promoters, and 10% in insulators, even though those chromatin states together cover only 6.7% of the genome overall.

- dsQTLs and TF binding: 3.6-fold enrichment of dsQTLs in TF footprints (controlling for overall enrichment in DHS). In dsQTLs, higher accessibility alleles show higher PWM scores and ChIP-seq reads. Also correlation of chr. accessibility and TF binding (Figure 2e): all positive including CTCF. Note: use ChIP-seq data from LCL. If using other tissues, correlations much weaker (Figure S14).

- Overlap of dsQTLs and eQTLs (Table S4): (1) for each dsQTL, consider a window and all expressed genes in that window. For all dsQTL-gene pairs, vary window size and estimate the proportion of eQTLs via π_1 analysis. Results: 41% at 10kb and 16% at 100kb. (2) Estimate the percent of eQTLs that are significantly associated with a gene (all pairs): for each window, use FDR correction for all pairs (p-value threshold differs between window sizes). Found 1027 significant dsQTL-eQTL pairs at FDR $< 10\%$ for 100kb, or 5.3%. This is the maximum number of pairs, so use 100kb for analysis.

- Estimating proportion of dsQTLs that are eQTL of at least one nearby gene (Suppl 19.1): consider all genes within 100kb of a dsQTL, obtain its minimum p-value, and adjust for multiple genes using permutation - so each dsQTL has a single p-value. Then do π_1 analysis: about 39% dsQTLs are eQTLs. Similarly, do FDR correction on p-values (one value per dsQTL), at FDR < 0.1, found 809, or 16% dsQTLs that are eQTLs of at least one nearby gene.

- Estimating proportion of eQTLs that are dsQTLs: (1) Calling eGenes: 1200. (2) For each eQTL (strongest per eGene): find strongest dsQTL p-value within 100kb, then use permutation to adjust p-values. Estimate percent of dsQTLs using Storey's method: about 55% eQTLs are dsQTLs.

- Example: dsQTL in a gene intron that disrupts a CRE, and show corresponding gene expression variation (Figure 3ab).

- Joint dsQTL-eQTL analysis: (1) Directional comparison: for 70%, directions are consistent. The signal is especially strong if limiting to DHS within 1kb (Figure 3b). (2) Motif enrichment in enhancers (same direction) and repressors (opposite direction): different motifs. (3) Spatial distribution: 23% are within 1kb of TSS of the gene, and 39% within 10kb (Figure 4a).

- Possible mechanisms that control whether dsQTL is also an eQTL (Figure 4b): distance to TSS, trascribed region, CTCF between DHS and TSS.

Identification of Genetic Variants That Affect Histone Modifications in Human Cells [McVicker & Pritchard, Science, 2013]

- Histone and Pol II-QTL data: 10 unrelated Yoruba, 4 histone marks (H3K4me1, H3K4me3, H3K27ac, and H3K27me3) and Pol II binding.

- QTL mapping and calibration: define testable SNPs (enough reads) and test association with reads in 2kb windows around the tested SNPs. Test results are calibrated (Figure S4): permutation of haplotypes and/or two alleles (flipping alleles randomly). Then control FDR by q values, and merge overlapping windows after FDR correction.

- At an FDR threshold of 10%, we identified 582 distinct histone mark and Pol II QTLs.

- Correlation between histone QTL, dsQTL and eQTL: the effect of QTL on multiple molecular phenotypes

  - Most histone mark and Pol II QTLs are within 1 kb of a DHS, but many are far from known dsQTLs.
  - At dsQTL or eQTL sites, the histone marks are often different at different alleles. Individuals who are homozygous for the high-expression genotype generally have higher levels of DNase I sensitivity, H3K4me3, H3K27ac, and Pol II occupancy at transcription start sites (TSSs).

- TF binding sites in histone-QTL and PolII-QTL:

  - Method: evaluate whether polymorphisms in TFBSs are associated with allelic imbalance in histone marks or Pol II. Map all TFBSs in 10 individuals containing polymorphism, than test if changes of PWM scores correlate with allelic imbalance in histone marks or Pol II.
  - Increased transcription factor occupancy generally increases levels of nearby activating histone marks and lowers the levels of H3K27me3.
  - Identification of specific transcription factors that direct histone marking: Out of the 39 clusters (similar TFs) that have a sufficient number of polymorphic TFBSs to be testable, 11 have a significant association with at least one histone mark.

- Lessons: Non-coding variation (often TFBS) can affect multiple molecular traits simultaneously: TF binding, histone marks, DNase HS, PolII occupancy and gene expression.

Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels [Banovich and Gilad, PLG, 2013]

- Mapping meQTLs: 450K CpGs, enriched in promoters. Quantile normalization. Cis-meQTL: within 6kb of SNPs to max. power. Found 13K CpG sites with QTL, mostly independent. Only 13% are outside 3kb.

- Overlap with eQTLs: given eQTL, for those within 3kb of CpG, 25% are also meQTLs. However, for those both meQTL and eQTL (150), half have positive correlation.

- Overlap with regulatory QTLs (DHS, histone): often see coordiated changes of multiple molecular traits (Figure 3). Also Table 1: 30-40% regulatory QTLs (K4me1/3, K27ac, DHS) are also meQTL (vs. 5-10% by chance). More common to have negative correlation, but a substantial fraction have positive correlation.

- TF binding may drive meQTLs: SNPs disrupting TF binding in DHS (using CENTIPEDE), more likely are meQTLs 10-15% vs. 2-3% by chance (Table 2).

- Lesson: effect direction of methylation on expression is uncertain.

RNA splicing is a primary link between genetic variation and disease [Yang Li and Pritchard, Science, 2016]

- Data: 8 molecular phenotypes on about 60 LCL lines. One new phenotype is 4sU-seq, which measures transcription (initiation) rate - how labeled uridine is incorporated into mRNA.

- Information flow from chromatin to protein: (1) Correlation between molecular traits: relatively low correlation (< 0.5) for H3k27ac and the rest; high correlation (above 0.8) between transcription-related traits (including RP); moderate correlation with protein. (2) comparison of effect sizes of QTL, H3K27ac-QTL effects correlated with other traits, r about 0.5, and the Txn rate, RNA and RP-QTL all highly correlated, r about 0.9. Transcription-QTL and protein-QTL: r about 0.7-0.8.

- Sharing of QTL across phenotypes: π_1 is about 0.25 to 0.5 for enhancer-QTL with eQTL, 60% promoter-QTL are eQTL. Other pairs highly shared: e.g. txp rate-QTL, more than 70% are eQTL.

- What explain eQTL? About 60% eQTL are chromatin-QTL. Unexplained QTL are enriched in txn elongation, exons, introns, etc.

- Splicing QTL (sQTL): detection by LeafCutter. Found nearly 3K new sQTL. Mostly independent of eQTL: (1) Different spatial distribution. (2) In cases where eQTL and sQTL are in the same genes: lead SNPs often > 10 kb apart. (3) sQTL are enriched in splice regions, coding, introns while eQTL enriched in chromatin features.

- Splicing can be affected by chromatin features: show that chr. annotations are enriched in sQTL (vs. control). An example CTCF binding site.

- Enrichment of sQTL in GWAS loci: using fgwas or PolyTest. In autoimmune diseases, "sQTLs appear to have effects of similar or even larger magnitude than eQTLs. Remark: this does not mean that sQTL explain more disease risk, since the number of sQTL vs. eQTL is not taken into account.
- Pipeline for peaking calling: for H3K27ac, MACS peak calling for each sample separately, then merge overlapping peaks. The MACS windows then split into 1kb segments (if longer).
- Pipeline for QTL mapping: (1) Use WASP to adjust for differences in sequencing depth and GC (2) standardize all measurements by gene and then quantile-normalize by individual (3) Regress out PCs, the number of PCs (0 to 15) is chosen to maximize the power.
- Estimation of sharing QTL between phenotypes: e.g. eQTL vs. pQTL. Choose top SNP-gene eQTL pairs, e.g. p < 1E-4, then obtain the p-values of these pairs in pQTL, estimation of π_1 using Storeys method, R qvalue(). Use bootstrap to compute confidence interval for π_1.
- Estimation of QTL sharing using Bayesian approach: joint model of QTL effect sizes across molecular phenotypes. For each SNP-gene pair, model the effect size (vector) as a mixture of 0 and MVN. Combine all pairs to estimate the MVN covariance matrix and mixture parameters. However, its not clear how it is used to estimate QTL sharing.
- Estimation of fraction of eQTL that are chromatin QTL (partition of eQTL): choose a set of genes, their top eQTL, and then ask how often they are chromatin QTL by p-values. Comparison with control SNPs.
- Testing enrichment of annotations: comparison of sQTL (or unexplained eQTL) vs. chromatin eQTL. For these eQTL, first obtain the posterior causal probability (PCP) for each SNP. Then for each annotation, sum PCP over all SNPs, which is the expected number of causal SNPs fallen into the annotation. Compute the fold difference between two types of eQTL, and obtain confidence interval by bootstrap.
- Q: how the Bayesian model of QTL sharing is used to estimate proportion (Figure S3)?
- Q: Computing PCP: not handle LD?

Disease variants alter transcription factor levels and methylation of their binding sites [Bonder and Teijmans, NG, 2017]

- Data: 3,800 whole blood, methylation and expression.
- cis-meQTL: most are within 10kb of CpG sites. Modest enrichment in active TSSs and enhancers.
- cis-eQTM: expression Quantitative Trait Methylation. Most negative effects (69%), but some positive effects. Show different enrichment patterns: e.g. negative enriched in TSS and active enriched in polycomb. A decision tree that classifiers the two using histone marks and distance to TSS.
- Trans-meQTL: using 6000 SNPs previously associated with disease traits. 1/3 of them have trans-meQTL effects.
- Possible mechanism of trans-meQTL: change of [TF] in cis. Figure 4a: often trans-meQTL of multiple CpGs show consistent effect directions. Ex. NFkB cis-eQTL, changes methylation of many CpG sites, enriched with NFkB targets. Similar pattern with CTCF.

Genetic determinants of co-accessible chromatin regions in activated T cells across humans [Gate and Regev, NG, 2018]

- Data: CD4 T cells from 100 healthy donors, stimulation (CD3, CD28) and ATAC-seq profiling. Also RNA-seq. Also in situ Hi-C: 3.5B reads for CD4+ T cells.

- Chr. accessible peaks: difference due to stimulation. To characterize peaks: enrichment of known annotations, e.g. Th1, Th2, Th17, T-reg. Change of enrichment patterns for different sets: T-rest, or T-stim or shared.
- Enrichment of GWAS variants: three groups of peaks, however, the enrichment pattern is similar, mostly immune phenotypes, but also fatty acid levels high enrichment. Remark: no comparison with whole blood, or other cell types.
- Analysis of differential TFs in cell-type specific and shared peaks (Figure 2): e.g. in peaks specific to stimulated T cells, find TFs (via footprint analysis) important for CD4+ T-cell activation or differentiation. In shared peaks, found CTCF. Show footprint difference in different conditions (footprints changed after stimulation).
- Co-accessibility of peaks across individuals: (1) both 1Mb and 100kb resolution (Figure 3c), matching Hi-C interactions. (2) Between peaks within 1.5Mb bins, find 2000 pairs, often involving enhancers. Co-accessible peaks are closer than expected: median distance 380kb. Discussion: perhaps peak calling bias, e.g. call one peak into two adjacent peaks (actually one peak).
- Local caQTL mapping: only analyze 60K SNP-containing peaks, and SNPs within peaks - total of 150K SNP-peak pairs tested using Rasqual. Use permutation: "-r" option of Rasqual, 10 permutations. Found 3K local caQTLs (SNPs within peaks) at FDR < 0.05. These QTL explain most of cis-heritability.
- Mechanism of local caQTL: enrichment near TSS and TES. (1) 900 local ATAC-QTLs often disrupt TF binding: use DeltaSVM to predict the effects of SNPs on TF binding (pre-trained model), about half local ATQC-QTLs disrupt six TFs (Figure 4d). (2) Local ATAC-peaks that overlap BATF, ETS1 and CTCF binding sites: different chr. accessibility along TFBS footprints (aggregate over all targets of the TFs, Figure 4f). However, only 5% of the corresponding local-ATAC-QTLs directly alter the core motif sequences.
- Prediction of ATAC-QTL effects with delta-SVM: train gkm-SVM using the peaks, then predict SNP effects using delta-SVM. Correlation of deltaSVM scores with ATAC-QTL effects: R = 0.6 (Figure 4e).
- Enriched heritability of local caQTL in GWAS: 5-8 fold enrichment of local caQTL in AID associated loci. LDSC analysis: 50-60 fold enrichment of AID heritability, however, not significant after adjusting for multiple testing (?)
- Example: a local ATAC-QTL, associated with several AIDs, Disruption of BATF motif.
- Genetic determinants of co-accessibility.
- Expression effects of caQTLs: eQTL mapping using RASQUAL. Found 424 genes to be associated with at least one local caQTLs within 500kb. Estimate 30% of caQTLs are eQTLs.

Impact of regulatory variation across human iPSCs and differentiated cells [Banovich and Gilad, GR, 2018]

- Experiment: RNA-seq and ATAC-seq in LCL, iPSC (from LCL), about 60 lines. And 14 lines of Cardiomycytes from iPSC (CMs).
- Calling caQTL and eQTLs in CMs: even with only 14 lines, WASP is able to call 500 eQTLs and 4000 caQTLs.
- Cell type specific caQTLs explain cell-type specific eQTLs: LCL caQTLs better explain LCL eQTLs (Figure 2A: QQ plot). Note: see Supplementary Materials, Section 10 about definition of cell-type specific QTLs: eQTL, FDR < 0.1 in one and p value > 0.05 in the other. caQTL, $p < 5 \times 10^{-4}$ in one and $p > 0.05$ in the other.

- Cell-type specific caQTLs tend to be more open in that cell type: Figure 2B: heat map, e.g. LCL-specific caQTL are much more open in LCL than iPSC and iPSC-CM. Note that the percent depends on cutoffs to define caQTLs.

- What drives cell type specific caQTLs? Comparison of caQTLs between iPSC vs. LCL. Largely driven by chromatin accessibility across cell types, but about 20% of iPSC-specific caQTLs are located in LCL accessible regions (Figure 2E: for iPSC caQTLs, plot of ATAC-seq signal in two cell types).

- Possible mechanism of cell-type specific caQTLs: build DNN to predict chromatin states, high AUC. Show that predicted effects correlate with measured caQTL effects (R = 0.5), Figure 3E. Example of SNP showing caQTL effect only in iPSC, and motif change.

- Remark: the proportion of cell-types specific caQTL that are due to chromatin accessible change, may vary depending on comparison. For distant cell types, most are likely driven by chromatin state changes.

Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms [Pierce and Ahsan, NC, 2018]

- Data: 900 subjects for eQTL and 300 for meQTL (300 with both). Find 5000 cis-SNP-expression-methylation.

- Coloc. analysis: results are sensitive to the prior of coloc, p_{12}. The number of pairs passing threshold vary greatly (nearly 10 times) with 10-fold change of the prior. Use internal empirical calibration (similar to EB) to choose the prior.

- Results of coloc: about 2700 triplets. Results strongly depend on LD between the top cis-eQTL and top cis-meQTL. Also depend on LD score.

- Partial correlation analysis: reject pleiotropy, if G affects M methylation and E expression independently, then regressing out G in M and E, the residuals should be uncorrelated. About 10-20% pass threshold.

- Mediation analysis: Sorbel test in either M to E, or E to M. About 10-20% pass threshold in either direction. Note that mediation analysis cannot distinguish causality.

- Inferring causal relation between methylation and expression: (1) Direction of effects: eQTL and meQTL often have opposite effect direction, about 58%. This increases to 70-80% in loci with some causal evidence found from partial correlation and mediation analysis.(2) Bayes network analysis: most often M > E is selected.

- Lesson: four types of analysis in the paper to learn the model of M and E. Each has limitations:

  - Partial correlation analysis: test $H_0 : G \to E, G \to M$. Test: regress out G, and correlate residuals of M and E. It does not allow confounder, so it is a weak null model. Most often, there is a causal effect or some confounders acting on M and E.
  - Mediation analysis: cannot distinguish if there is a causal effect or causal direction. Even if we reject null from partial correlation analysis, and find a significant mediation, we still cannot say if there is a causal effect: could be a shared confounder.
  - MR: similar to co-localization. Could be due to LD between eQTL and meQTL.
  - Bayes network: to explicit model the (G, M, E) data, need to assume that there is no confounder.

Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease [Huan and Levy, NC, 2019]

- Data: 415K CpG sites in 4000 whole blood samples (FHS). Samples: 456 unrelated, and the rest from 500 families.
- MeQTL mapping and genetic architecture: (1) 100K (25%) CpG have $h^2 > 0.1$. Most often, single best cis-eQTLs do not explain all h2: mean 0.07 vs. 0.18. (2) Total: 394K independent cis-meQTL (within 1Mb) and 21K trans-meQTLs, where independent are defined by $r^2 > 0.2$.
- Features of me-QTLs: enrichment in promoters, TSS, enhancers.
- Causal CpGs of CVD and related traits using MR: merge nearby CpGs with high LD (CpGs within 2Mb and highly correlated with each other, often share meQTLs). Find 14K CpGs with 3 independent cis-meQTLs, do MR with CVD traits: found 92 CpGs.
- Finding target genes of CpGs: cis-association of CpGs and genes, then do co-localization analysis of eQTL-meQTL. Found 8 genes: test their roles in traits using MR, with cis-eQTLs as IVs.
- Trans-meQTL hotspots: 22 hotspots, at least 30 CpGs. In 74 genes near these hotspots: enrichment of TFs (ZNF genes).
- Remark: CpG-gene causal test to link CpGs to genes.

Common DNA sequence variation influences 3-dimensional conformation of the human genome [Gorkin and Bing Ren, GB, 2019]

- Experiment: dilution Hi-C in 20 LCL samples.
- Derive features: 40kb contact matrix, compartmentalization (PC1 from Hi-C), directional index (DI), insulation scores (INS) - measure abundance of interactions spanning a region (strong insulation, low INS). TAD boundaries are associated with low INS and high DI, but high DI can also occur elsewhere. FIRE scores: frequency of contact of a region with nearby regions (15-200kb).
- Remark: is INS score properly normalized? Low INS could just mean that the region doesnt have many interactions, e.g. gene desert.
- Inter-individual variation of chromatin structure features: generally correlations within an individual (replicates) are higher than between individuals.
- Detecting variable regions across samples: use limma, find regions whose variations across individuals are higher than within individuals. Found several thousand in each features. An example in Fig. 2A: different DIs.
- Covariation of variable regions with other epigenomic features: FIRE regions correlate with H3K27ac and H3K4me1. DI and INS regions correlate with histone modification levels as well as CTCF and Cohesin binding.
- CTCF-changing SNPs and chromatin loops: (1) Define chromatin loops: from external data. (2) Consider SNPs that change CTCF motifs in chromatin loop anchors. Found overall association with strength of chromatin loops.
- Q: How does one infer the strength of individual chromatin loops from Hi-C data?
- Mapping Hi-C QTLs: use hi-C derived features, and use LMM to include biological replicates, use 11 YRI samples for discovery. Features: FIRE, DI, INS, and contact frequency, but not PC1. For DI and INS: use window size of 200 Kb up- stream and downstream of the target bin, instead of larger bins. Choose test bins (40kb), and test all SNPs within that bin and SNPs in perfect LD. Found a few hundred SNPs per feature at FDR < 0.2.

- Effects of HiC-QTLs: often have effects on other epigenomic features. Ex. FIRE-QTLs, the high-FIRE allele is also associated with higher levels of active histone modifications and chromatin accessibility.
- Enrichment of HiC-QTLs in GWAS variants of AIDs: 1.6 - 1.7 fold enriched of nominal GWAS variants. Enrichment higher than the test bins.
- Remark: HiC-QTL mapping only use derived features, but not individual contact pairs.
- Remark: no discussion of the mechanisms of HiC-QTLs, enriched with CTCF motifs/targets? Or enriched in TAD boundaries?
- Remark: consequence of HiC-QTLs, do they tend to be eQTLs of multiple genes?

Genetic drivers of m6A methylation in human brain, heart, muscle, and lung [Xiong and Kellis, review for NG, 2020]

- Data: 102 m6A MeRIP profiles, 4 tissues, 76 samples. Average 42K sites per sample.
- Tissue-specific patterns of m6A: Samples clustered by tissues: with brain most different, and heart/muscle cluster together, closer to lung. Overall, about 45% m6A peaks are shared. For most tissue-specific m6As: genes are universally expressed.
- M6A QTL discovery: m6A normalization: only m6A levels, not mRNAs. Use fastQTL, search for SNPs in promoters, introns and exons. Found 400 genes in brain, 400 in lung and 200 in heart/muscle (merged), at empirical p-value < 0.005. This empirical p-values are defined using FastQTL: for min-p, do permutation to obtain its empirical p-value. Then use beta-binomial approximation to get the p-values. Note: no FDR correction.
- M6A-QTL validation: use only two samples, check effect size direction, whether consistent with discovery cohort. 82% and 62% respectively. Effect direction in the two samples: compare two genotypes (only SNPs that have different genotypes are chosen), and see if difference is consistent. Remark: let π be the TP proportion, for TP, expect 100% agreement, and for FP: 50%. Then we have 1*pi + 0.5 * (1-pi) = 0.72, so we have pi = 0.44.
- Tissue-specificity of m6A-QTL: >90% are tissue-specific (Fig. 3A), only 1-2% are significant in multiple tissues. Within a tissue: very small percent shared m6A-QTL; 70% are m6A QTL in one tissue but m6A peaks are also in another tissue; 24% are m6A-QTL in one tissue and also peak in one tissue (Fig. 3B). Using a more relaxed threshold of p = 0.05 in a second tissue, still 94% are tissue-specific. Plotting: Fig. 3D, show m6A-QTLs in tissue 2, but different shapes for m6A-QTLs found in tissue 1 - show that most tissue 1 m6A-QTLs have insignificant p-values, and low effect sizes.
- Comparison with m6A-QTLs in LCL: little agreement in effect directions.
- M6A-QTLs are modestly enriched with eQTL GTEx (<2 fold), but effect directions not consistent for shared m6A-QTL and eQTLs (Fig. 4b).
- GWAS enrichment: 55 GWAS traits, m6A-QTLs at p-value 1E-3 or 1E-4 (Fig. 5A). Use S-LDSC, remove m6A-QTLs overlapping with enhancers. See enrichment in tissue-specific fashion: e.g. lung m6A-QTLs in lung-related traits; muscle/heart not in NPDs. 13 traits show enrichment in multiple tissues, including height, BMI, and blood pressure. Remark: some not very tissue-specific patterns, brain QTLs enriched most strongly in height and BP; brain QTLs in lung traits.
- Some examples of GWAS SNPs that are m6A-QTLs: Fig. 5 de. In both cases, close to top GWAS, and top in m6A-QTL, not eQTL.
- Putative m6A regulators: enrichment of RBPs sites in m6A-QTLs using GARFIELD. Found 27 regulators in total (at $p < 0.05$, no multiple testing adjustment), including known readers DF2 and FMR1. Show PPI with known M6A regulators, writers, erases (source: stringdb R package).

- Correlation of regulator expression with m6A targets (Fig. 6c): some additional genes at $p < 0.1$. Correlation of [RBP] with m6A targets (require m6A-QTLs to be inside RBP binding sites). Remark: correlation extremely weak.

7.2.3 QTL Studies in Model Organisms

eQTLs in yeast: [Brem & Kruglyak, Science, 2002]

- Methods:
  - Data: expression of 6,215 genes are measured in 40 segregants, genotyped at 3312 markers.
  - Test linkage between a marker and an expression trait: partition the segregants into two groups, and compare the expression levels between the groups with Mann-Whitney test.
- Results:
  - Detection rate: 1528 messages show differential expression in parental strains. Linkage analysis: 570 messages show linkage to at least one locus (308 among 1528 parental different messages).
  - Local eQTLs (within 10kb): 185 out of 570 messages show linkage in this category.
  - Distal eQTLs: eight hotspots (Figure 3) - 40% linkages fell into one of the eight groups of genes controlled by these hotspots. The group size: 7 to 94 genes. Examples: leucine biosynthesis - linked to Leu2, Ura3 (enzymes); fatty acid metabolism - linked to TF Hap1; daughter-cell specific genes - linked to AMN1 (protein required for daughter cell separation, not TF); Msn2/4-dependent genes - show Msn2/4 consensus sites.

Trans-acting regulatory variation in yeast [Yvert & Kruglyak, NG, 2003]:

- Methods:
  - Data: 86 segregants from a cross between BY and RM strains.
  - Module QTL identification: treat the mean expression of a module as a quantitative trait, and do QTL analysis.
- Results:
  - Gene clustering: by pairwise correlation of expression patterns exceeding 0.725 (threshold determined by permutation test). Found 593 clusters of at least two genes and 205 clusters of more than two. Clusters often are enriched with specific processes: e.g. hexose transport, pheromone response, daughter-cell specificity.
  - Trans-acting loci: a total of 304 clusters show linkage to at least one position. Overall, 75% of genes and 80% of clusters did not show self-linkage: most genetic variation is expression is due to trans-acting factors.
  - Case studies: GPA1 is a causal locus of pheromone response cluster with nonsynonymous substitutions in the protein sequence; and AMN1 a causal locus of daughter cell separation cluster.
  - Most trans-variations do not map to TFs: the regulatory TFs of the gene clusters (according to ChIP-chip data) are not overrepresented in the module-QTL of these clusters. In fact, most trans-eQTLs are not mapped to TFs. Many classes of genes were found in trans-eQTLs.

Genetic landscape of gene expression in yeast [Brem & Kruglyak, PNAS, 2005]:

- Problem: estimate the genetic models of transcripts, e.g. the number of loci.
- Methods:

- Data and analysis: similar to [Brem02], but with 112 segregants, $2,957$ markers.
- Power of study: suppose the true model is a main locus plus many with infinitesimal effects, then the study has $> 90\%$ power of detect genes with a main locus of effect size 25%.
- Estimating genetic models: e.g. the fraction of transcripts with single QTL. The idea is: e.g. suppose all genes are controlled by a single locus, then in simulation (with this genetic model for all transcripts, and apply the same QTL analysis and threshold), most genes will be found with one QTL with large effect. The actual distribution of effect size is different, and the ratio of large-effect transcripts can be used to estimate the fraction of genes with single QTL. The same idea can be applied to estimate the fraction of genes with n equal-effect QTLs, $n = 1, \cdots, 10$.
- Epistasis test: tests for a difference between the mean expression levels of segregants and parents, because the means are equal for any additive inheritance pattern.

- Results:
  - QTLs: 3,546 transcripts have strong heritability $H^2 > 0.69$. Among these, 2,091 (59%) showed linkage to at least one QTL (FDR at 0.05). Linkage results were robust to different normalization procedures and linkage tests.
  - Effect size distribution: (effect sizes are estimated from independent test set, half of the data) for all transcripts with at least one QTL, choose the most significant QTL, the median QTL effect is 0.27, with 23% of QTLs have effect > 0.5.
  - Genetic model estimation: only 3% of highly heritable transcripts are consistent with single-locus model; 17-18% can be explained by models with one or two loci, and half of the transcripts requre models with $n > 5$.
  - Epistasis test: for 3,546 highly heritable transcripts, 583 (16%) passed the epistasis test.

Mouse liver eQTL [Schadt & Friend, Nature, 2003]:

- Methods: liver tissues from 111 F_2 mice from two standard inbred strains, 23,574 genes. Standard interval mapping.

- Results:
  - Detection rates: 7,861 genes were differentially expressed in two parental strains, among these, 2,213 gnes have at least one eQTL with LOD score > 4.3 ($P < 0.00005$). Without filtering based on DE, found 4,339 eQTLs over 3,701 genes with LOD greater than 4.3.
  - Effect size and number of eQTLs: on average eQTLs with LOD greater than 4.3 explained 25% of the variance in F_2. 40% of genes with at least one eQTL with LOD greater than 3.0 had more than one eQTL, and close to 4% of these genes had more than 3 eQTLs.
  - Local and distal eQTLs: about 34% of eQTLs are local (LOD > 4.3), however, 71% of eQTLs are local if LOD threshold is 7.0. Thus eQTLs with high LOD tend to be cis-acting, while moderate eQTLs act in trans- in most cases. Also several eQTL hot-spots were found.

Mouse eQTL in hematopoietic stem cells (HSC) [Bystrykh & Haan, NG, 2005]:

- Data: HSCs isolated from the bone marrow of D2 and B6 mice, 30 strains.

- eQTL pattern: 478 transcripts were associated to a QTL within 20 Mb of the gene itself. Also identified multiple "vertical bands": the QTLs that modulate expression of a large number of transcripts.

- Co-regulated genes: for four strongly cis-regulated transcripts (TF, receptors, etc), find the co-regulated genes through correlation. Many of these genes are downstream targets of the four chosen genes.

Mouse eQTL in brain [Chesler & Williams, NG, 2005]:

- Data: 80 recombinant inbred (RI) strains from BXD.
- eQTL: at FDR 10%, found 88 significant transcripts, most (83) were cis-linked; at FDR 25%, found 101 significant transcripts. Seven trans-regulatory QTL bands were identified: one band regulates 1,650 transcripts.
- Tissue specificity of expression regulation: the comparison with HSC eQTL, most global regulators (trans-) are tissue-specific. The cis-regulatory QTLs are often shared between brain and HSC.
- Synaptic vesicle-related module: most of the trans-acting bands also regulate this module [Chesler05-Figure 6]. Some of the transcripts are also cis-regulated in these lcoi, making them candidate modifiers of the synapse-related module.

7.3 Systems Genetics Methods

Systems genetics paradigm:

- Challenge: the central goal is to characterize the candidate genes identified through linkage or association studies, but whose function/mechanism remain unknown.
- **Reduction of a complex phenotype to molecular traits**: identify the intermediate molecular traits between genetic variations and phenotypes, then one can study these molecular traits: how are they influenced by genetic variations and how they influence the phenotypes. Each problem can then be studied separately, e.g. through an animal model.
  Example. atherosclerosis involves many aspects, in particluar, the response of endothelial cells to oxidized lipids. Study this system in vitro: how the gene expression of endotehlial cells change in response to oxidzied lipids. [Lusis at UCLA lecture]
- **Understand the links from DNA to molecular traits**: for cis-eQTL, this is about understanding the gene regulation. For trans-eQTL, this involves indirect effects (regulatory genes, but also compensatory effect, etc.).
- **Understand the links from genes to phenotype through biological processes and gene networks**: generally a gene influences some biological processes, and cellular phenotypes, then organismic level phenotypes. The knowledge of processes a gene is involved and of gene networks can help understand the causal consequence of a gene (including its possible global effects).
  Example: PPAR-gamma: predict the effect of deletion using network, the change of expression of insulin-related genes (good) and lipid-metabolism related genes (bad). Similar analysis on GPR105 (p2ry14), good effects on both T2D and fat/heart disease. [Schadt, TED2011 talk]

Systems genetics approach to complex diseases [personal notes]

- Problem: given multiple types of data related to a disease, including GWAS, transcriptome (in patients and controls), eQTL (in patients or independent cohorts) and other network information, how do we better understand the genetics of disease?
- Summary: key considerations to study a candidate module are: (1) Connection with phenotypes: association with traits, eQTL as IV. (2) Upstream regulation: TFs, RBPs, mediator of trans-eQTL. (3) Downstream effects: functions of genes, disease relevance.
  - Example: KLF14 study [Small and McCarthy, NG, 2018]: (1) trans-gene network is a target of T2D associated locus. (2) KLF14 expression as mediator, also enrichment of KLF14 ChIP-seq and motif. (3) Some T2D related genes, e.g. GLUT4, IDN (insulin degradation).

- Cellular phenotypes: a key aspect of systems genetics is to use intermediate phenotypes, particularly cellular phenotypes. The key is to find some metrics of cells that reflect phenotypes, e.g. cell proliferation and cytokine production for immune diseases. In the absence of direct measurement, we can infer them from transcriptome signatures.
  - Remark: often factor analysis assumes independence of factors. In the case of cellular phenotypes, we may need to consider their interactions/dependence. Ex. in yeast, cell growth pathways and stress response are negatively correlated.
- SNP-centered analysis: understand the mechanism of disease SNPs in terms of their effects on molecular and cellular phenotypes. Ex. KLF14 study, the SNP is a trans-eQTL of multiple genes, enriched with T2D risk genes. The limitation is that trans-effects of disease SNPs are difficult to study. Possible solution: combine many disease SNPs with PRS.
- Gene and pathway-centered analysis: identifying disease genes and pathways/modules.
  - Differential expression and co-expression analysis: genes and modules discovered in this way may be candidates for causal genes/pathways; and may be also biomarkers of disease subtypes. For modules: use association of eigen-genes with traits.
  - Finding causal genes: MR analysis, colocalization.
  - Finding causal pathways/modules by enrichment: (1) Enrichment of disease associated variants in cis of the genes. Could use cis-eQTL of genes. (2) Enrichment of GWAS variants in eQTL (both cis and trans) of the genes.
  - Finding causal pathways/modules by IVs: use eQTLs of pathways as IVs to test causality. Use latent factors to represent pathway activity. Or mixture of MR: some genes are causal to phenotypes, some not.
- Identification of key drivers of phenotypes: manipulation of drivers may cause or restore diseases.
  - Find regulatory genes of disease modules: TFs, RBPs.
  - Find master genetic regulators: eQTLs of disease modules. Ex. SESN3 study: find epilepsy related genes first, then search for eQTL. Challenge is that this study may be under-powered. Hypothesis: genes within modules are candidates for master regulators (e.g. EndoG in T1D), so we can search for SNPs/eQTL of genes in modules.
- Disease subtypes: it is important to know that a disease may have multiple subtypes. Keep this in mind when developing specific analysis.
  - Differential expression/co-expression analysis may be more powerful if we can segregate by subtypes (e.g. through PRS).
  - Possible strategy for disease subtype discovery: identify disease modules, and then define PRS based on SNP effects on the modules. Segregate patients by these PRSs.
  - How to confirm subtypes? Use effect size dependency on subtypes.
- Context specificity of molecular and cellular phenotypes: contexts are tissue type or proper stimuli (e.g. LPS for immune cells) that are relevant to the phenotypes. This would be in part of experimental design: how to choose the relevant cell types and stimulate cells in relevant ways. Also the anaysis part when considering stimuli: either response QTL, or QTL whose effects depend on stimuli (similar to GxE interactions). It may be more powerful to combine the conditions as many eQTL would be shared, in a hierarchical way.

Reference: [Molecular networks as sensors and drivers of common human diseases, Shadt, Nature, 2009; Schadt & Shaywitz, NRDD, 2009; Cookson & Lathrop, NRG, 2009; Rockman, Nature, 2008]

Case studies of GWAS-eQTL:

- Asthma study: a series of SNPs in strong LD spanning 200 kb containing 19 genes, strong effect in cis on the expression of one gene, ORMDL3.
- Crohn's disease (CD): the associated SNPs are located in a gene desert. Examination of LCL eQTL database showed that one SNP acts as a long-range cis-acting factor influencing expression of PTGER4.
- Body mass index (BMI): a missense SNP in SH2B1 assocated with BMI, but also with expression of EIF3C and TUFM. Not clear the causal relationship.

Difficulties/challenges of GWAS-eQTL studies:

- Expression measurement: many other sources of variations of expression, including batch effect, systematic bias in sample preparation, etc. And also results from different microarray platforms cannot be compared.
- Detecting loci with small effects: particularly hard with small sample size. Generally trans-eQTLs are hard to found.
- Missing heritability of expression: other variants include: (1) CNV: estimated that SNP and CNV captured 84% and 16% genetic variation in gene expression, but the signals have little overlap; (2) epigenetic factors: DNA methylation and histone modification.
- Gene expression in tissues: the current eQTL data often from a few tissues. However, want expression in tissues involved in diseases. The tissues are generally hard to access, or if from post-mortem sample, often have changes accompy death or surgery. The ideas: "exercise the genome", or treating cells with different ways: pro-inflamatory stress, metabolic stress, response to signaling molecules, therapeutic agents, etc.

Leveraging gene co-expression networks to pinpoint the regulation of complex traits and disease, with a focus on cardiovascular traits [BIFG, 2013]

- Factor analysis: ICA, PCA and NMF. NMF is best for decomposition into cell types or states.
- Co-expression network reconstruction (Figure 2): the metric for defining the adjacency matrix, such as Spearman correlation and Mutual information (MI). Pruning edges: clustering (WGCNA), ARACne prune weakest for every three-edge triplets, Gaussian graphical model, GENIE3 that uses tree to regress expression with predictors (TFs).
- Analysis: why Gaussian graphical model is not good for network reconstruction? It tends to find edges with low degrees, and miss high-degree edges. Ex. suppose we have a cluster of genes, when y can be predicted from x, then any other $z \to y$ will be missed.
- Understanding the regulators and functional context of modules (Figure 2): TFBS enrichment, pathway enrichment.
- Connecting co-expression networks with phenotypes: define candidate modules by enrichment of disease genes (GWAS pathway analysis). Two additional analysis: (1) Master genetic regulators: genetic variants associated with expression of multiple genes in a module. Bayesian multivariate eQTL mapping (network QTL). (2) Differential co-expression network analysis: co-factorization analysis. Note that the modules correlated with phenotypes may be used to construct disease subtypes.
- Using co-expression networks to study cardiovascular traits: (1) EndoG (Figure 3): a WGCNA module in human heart, with ENDOG being a hub and many mito. genes. Confirm the function of ENDOG in heart function in mouse model. (2) Ebi2 (Figure 4): a module in rat (7 tissues), IRF7 regulated inflammation network (motif/ChIP-seq analysis). Found this module is enriched with T1D genes, and Ebi2 a genetic regulator of this module.

- Lesson: best network construction methods use prior knowledge in the form of TF-gene interactions.
- **Lesson**: the general strategy is to identify co-expression modules related to diseases, and identify TF or genetic regulators of the modules. Finding hub genes may help find key regulators of the modules.
- Remark: the factor analysis models are all based on linear effects. Can we expand to nonlinear cases such as boosting?

Systems Genetics as a Tool to Identify Master Genetic Regulators in Complex Disease [Chapter 16 of Systems Genetics book, 2016]

- Strategies of mapping genetic regulators of modules: two step strategy (Figure 2). (1) Dimensionality reduction (e.g. eigengenes of modules), and do QTL mapping; (2) multivariate genetic mapping of genes in the modules.
- Example: Asxl2. From GWAS of bone marrow density, find Asxl2 as a candidate gene. Do WGCNA on rat expression data: find a module with Asxl2. Use the module to annotate the functions/mechanisms of Asxl2.
- Example: **KLF14**. Start with KLF14 as a risk gene of T2D, find it is a trans-eQTL of multiple genes (MuTHER data). Furthermore, for these genes, their cis-eQTLs are enriched with T2D variants.
- Example: Trem2. Using eQTL of 200 rat samples, find Trem2 a trans-eQTL of 190 genes.
- Example: **SESN3**. From gene expression data of 129 hippocampus brain samples, find a module enriched with epilepsy genes. Then do eQTL hotspot mapping (two-step strategy), identify SENSE3 as a trans-regulator.
- Lesson: co-expression networks and elucidate functions of disease genes.
- Lesson: if we start with known trait variants, finding that it is a trans-eQTL of other genes is not enough, as this can be due to pleiotropic effect of the trait variants. So need evidence that trans-associated genes are also disease related.

Multi-omics approaches to disease [Hasin and Lusis, GB, 2017]

- Genotype-first approach: start with GWAS variants, identify causal genes and targets. FTO study [NEJM paper], allele-specific enhancer activity, then confirm by ASE and eQTL. Use trans-eQTL and gene correlation to find the functions of target genes of IRX3/5: adipocyte differentiation.
- Phenotype-first approach: start with gene networks of diseases, then find causal genes. AD study: gene sets changed during progression of AD, find immune genes. Then GWAS variants are enriched in enhancers of immune-related genes, but not neuronal function related enhancers and promoters.
- Problem: protein levels are not reflected by mRNA levels. Only a subset of genes show good correlation.
- Lesson: increase GWAS signals by stratification of variants related to certain biological functions/process.

Camelot: predict phenotypes from linkage and expression [Chen & Pe'er, MSB, 2009]:

- Motivation: linkage analysis can reveal candidate genes of a trait; in addition, expression profile is also predictive. Combine the two dataset to better identify the genes underlying a trait.
- Methods:
  - Input data: 104 individual strains with 94 drug responses each [Perlstein, NG07]. For each strain wrt. one drug, the data: D - drug response, $\mathbf{L}$ - the genotype, i.e. 526 markers. And $\mathbf{E}$ - the expression profile (under drug-free condition) of 6,189 genes [Brem, PNAS05].

  - Pre-processing: expression features are limited to regulators (TFs, signaling molecules, chromatin factors, RNA factors), and multi-drug resistence genes (endosome transport, etc.).
  - Predictive model of phenotype: linear regression of D on $\mathbf{L}$ and $\mathbf{E}$. Use elastic net, non-parameter bootstrap to select features for prediction.
  - Causality test: suppose one transcript is identified as predictive, it might be that its effect has already been incorporated by the genotype $\mathbf{L}$, thus design a triangle-test, does the effect of this gene go beyond what is explained by the genotype? The idea is to control $\mathbf{L}$.
  - Zoom-in score: for an important feature in $\mathbf{L}$, still need to choose genes (often tens of). The test is based on: the expression level is also correlated with the trait; plus two priors: 1) conservation (deviations from the consensus sequence); 2) cis-linkage.

- Results:
  - Evaluation by predictability: compare predictions of phenotypes with 1) no expression data used; 2) linkdage analysis (only QTLs, and predict trait by linear regression).
  - Verify the role of selected genes in the relevant phenotypes: e.g. DHH1 in H_2O_2 response. The model: region in Chr XIV $\rightarrow$ DHH1 expression $\rightarrow$ mitochondrial biogenesis $\rightarrow$ oxidative stress response.

Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis [Voight, NG, 2010]

- Method (Supplement, page 36).
- Motivation: find the causal genes of GWAS-SNPs. The idea is to find the true cis-eQTL effect of the GWAS-SNP, but the problem is: disease SNPs can have significant cis-eQTL effect, even if they are not coincident.
- Idea: find the candidate target gene of the GWAS-SNP of interest, and then assess if the GWAS-SNP explains expression variation of that gene (ie. see if GWAS-SNP explains the effect of the best cis-eQTL).
- Step 1: for lead T2D SNPs, find its cis-effect on all nearby genes (2Mb) at $p < 0.001$. In some examples, multiple associations were found. E.g. HNF1 locus, CAMKK2 is the strongest.
- Step 2: for that gene, find the strongest cis-eQTL. Compare the cis-effect of the two SNPs: GWAS lead SNP and best cis-eSNP. In CAMKK2 example, best cis-eSNP has much larger effect than GWAS SNP, thus likely not coincident.
- Step 3: Conditional analysis, test if GWAS-SNP is sufficient to explain the signal at cis-eQTL. Regression of expression on SNPs: use GWAS-SNP as explanatory variable, but conditioned on best cis-eSNP. If GWAS-SNP evaporates, it suggests that the signal is entirely driven by best cis-eSNP. Alternatively, we view best cis-eSNP as explanatory variable and conditioned on GWAS-SNP. If no effect, it suggests that GWAS-SNP does not make much contribution.

Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics (COLOC) [Giambartolomei & Plagnol, PLG, 2014]

- Problem: given a locus, test if GWAS and eQTL signal colocalize. Compare 5 hypothesis (Figure 1): H_0, H_1, H_2 for no association, and association with only one trait; H_3 association with two traits on two independent SNPs; H_4 association with two traits on a single SNP.

- Model: Define a configuration D as the true association status of all SNPs in a locus wrt. the two traits ($2 \times n$ matrix, where n is the number of SNPs). The BF of a hypothesis depends on the likelihood under any configuration that is consistent with the hypothesis:

$$\frac{P(H_h|D)}{P(H_0|D)} = \sum_{S \in S_h} \frac{P(D|S)}{P(D|S_0)} \times \frac{P(S)}{P(S_0)}. \tag{7.59}$$

  For H_0, only one S. For H_1, H_2, H_4, n possible S and for H_3, $\binom{n}{2} - n$ possible S. We define the approximate BF of a SNP j as ABF_j (two values for two traits). Also for each SNP, define prior probability of associated with trait 1, trait 2 and both traits as p_1, p_2 and p_{12}. Then we have:

$$\frac{P(H_1|D)}{P(H_0|D)} = p_1 \sum_j ABF_j^1 \tag{7.60}$$

  Similarly we can obtain the BFs for other hypothesis (see Supplements). Note that ABF of SNPs for trait 2 will appear in the BF of H_2, and so on. Choose $p_1 = p_2 = 10^{-4}$ and The prior p_{12} for H_4, $p_{12} = 10^{-6}$. The evidence of a hypothesis is summarized as posterior probability (PP), which sum to 1 over 5 hypothesis.

- Examples (Figure 2): (A-B) PP3 (posterior prob of H_3) is large (C-D) PP4 is 82%: the same top SNP in both traits.

- Remark: the model uses the fact that when S has a single causal variant, then $P(D|S)/P(D|S_0)$ would be just ABF of the causal variant. Also the model makes the assumption that data of two traits are independent.

A gene-based association method for mapping traits using reference transcriptome data (PrediXcan) [Gamazon, NG, 2015]

- Data: DGN (Depression Genes and Networks), 922 whole-blood samples eQTL. WTCCC GWAS data.

- Prediction of gene expression: top SNP, polygenic score and elastic net (Lasso is similar). Only on cis-heritable genes. The mean h^2 is 0.153, with the predicted $R^2 = 0.114$ for top eQTL, 0.099 for polygenic score and 0.137 for elastic net.

  - Application to other dataset: on LCL, $R^2 = 0.0197$, 0.0367 for adipose, 0.0359 for lung and 0.0458 for whole blood.
  - Including trans-eQTL: results worse.

- Application to WTCCC: most findings are in autoimmune disease genes, and are either reported or located near the reported genes. The only exception: PTPRE (BD) and KCNN4 (HT).

- Remark: prediction error is not modeled (attenuation bias).

- **Remark**: correlation, not causality. Consider an example of a non-causal gene with 2 eSNPs, one of them is associated with phenotype. Under an extremely simple model, we have: at genotype (0,0), expression and phenotype are 0 and 0; at genotype (0,1), expression and phenotype 1 and 0; at genotype (1,0), expression and phenotype 1 and 1; and at genotype (1,1), expression and phenotype 2 and 1. There is a good correlation of expression and phenotype, $r = 0.5$, and it could be highly significant with large samples.

Integrative approaches for large-scale transcriptome-wide association studies (TWAS) [Gusev and Pasaniuc, NG, 2016]

- eQTL data: 3000 individuals for blood adipose eQTL, cis heritability of gene expression from 0.01 to 0.07 and trans h^2 0.04 to 0.06. Focus on 6000 cis-heritable genes.

- Prediction of gene expression: compare top eQTL, BSLMM and BLUP. Found that BSLMM performs the best.

- Summary statistics based TWAS: let W be the effect size in eQTL and Z be the standardized effect size in GWAS. The test statistic is $\sum_i W_i Z_i = W^T Z = Z^T W$. Its variance is given by (treating Z as random and W fixed):

$$\text{Var}\,(W^T Z) = (Z^T W)^T (Z^T W) = W^T Z Z^T W = W^T \Sigma_{s,s} W \tag{7.61}$$

  where $\Sigma_{s,s}$ is the LD matrix of SNPs. We use here the fact that for standardized effect sizes, covariance between Z_i and Z_j is just the LD between two SNPs.

- Comparison with COLOC: similar when there is a single causal variant, but better when the causal variant untyped or allelic heterogenity. Probably due to LD modeling.

- Application to lipid GWAS: found 25 new associations, 19 of which replicated in later studies. In GWAS of obesity related trait, found 69 new genes. Note: in choosing what genes to test for TWAS, use p-value from heritability analysis, $p < 0.01$ by default.

- **Remark**: causality analysis, if an eSNP has no signal in GWAS, we have $z_i = 0$, thus $w_i z_i = 0$, so it does not contribute to the correlation. But if a gene is true causal gene, this SNP should penalize.

- **Remark**: summary statistics modeling not proper modeling of LD, because the statistics is based on observed effect sizes, not true effect sizes. Because of LD, they could differ significantly.

Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets (SMR) [Zhu and Yang, NG, 2016]

- Problem: detect "pleiotropic association" between genes and phenotype, which is either due to pleiotropy or causality. "The MR approach using a single genetic variant is unable to distinguish between pleiotropy and causality".

- SMR method: let $\hat{b}_{zx}$ be the effect of SNP (z) on x and $\hat{b}_{zy}$ be the effect on y, then the estimated causal effect is $\hat{b}_{xy} = \hat{b}_{zy}/\hat{b}_{zx}$. Its variance is given by Equation (2). The MR statistic is $\hat{b}_{xy}^2/\text{Var}\,(\hat{b}_{xy})$. With summary data, we can show that the SMR statistic is given by z-scores of Z to X and Z to Y:

$$T_{SMR} = \frac{z_{zy}^2 z_{zx}^2}{z_{zy}^2 + z_{zx}^2} \tag{7.62}$$

  When applying the method, we could obtain effect sizes $\hat{b}_{zx}$ and $\hat{b}_{zy}$ using z-scores and AF of SNPs.

- HEIDI: detect heterogeneity of effects. The intuition is if the signal is due to linkage, then there are two causal variants (one for each trait), then the SMR estimated effects would be different between two SNPs. Suppose we choose a top SNP, and we assess if the SMR estimator at another SNP is different, if so, there is evidence of linkage. H_0: same effects in all SNPs (co-localization). The test statistic is $d_i = \hat{b}_{xy(i)} - \hat{b}_{xy(top)}$, and we can account for LD in the distribution of d_i.

- Results of GWAS-eQTL analysis: Westra eQTL data vs. 5 phenotypes. 68 genes for height, 9 for BMI, 2 for WHRadjBMI, 9 for rheumatoid arthritis and 16 for SCZ.

- Casuality analysis of genes: most genes have no trans-eQTL. For the 4 genes with trans-eQTL, none of them have consistent effects in GWAS.

- Discussion: we use HEIDI to filter out heterogeneous effects, so generally we favor HEIDI results with *large* p-values. But when SNPs are in high LD (two causal variants), the p-values will be large (because we do not have power to detect heterogeneous effects), so HEIDI may miss many cases due to linkage, but treat them as causal or pleiotropy.

- Note: see Supplementary Note 2 of [Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits, NC, 2018] for discussion of HEIDI. (1) Analogy with HWE in QC: in HWE test, we want to filter out SNPs showing deviation from HWE (H_1). (2) Threshold: Using $p < 0.05$ without multiple testing correction may be too conservative: especially when this is used in multiple steps, e.g. lose 15% of true pleiotropic effects (3 steps). So the paper filters out all with $p < 0.01$. (3) Inclusion of weak cis-eQTL: inflation of p-values, and reduce the power of detecting true pleiotropic associations. Suggest to remove SNPs not or in weak LD with top cis-eQTL.

Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization (Enloc) [Wen, PLG, 2017]

- Motivation: to assess co-localization, we need to know the prior probability of an eQTL being a GWAS variant. This step would benefit from Empirical Bayes estimate, which is not done in previous methods, eCAVIAR and coloc.

- Method outline: let d be whether a SNP is eQTL, treat d as annotations of SNPs in GWAS data for fine-mapping. The issue is that d is not observed, so we need to marginalize the missing d.

- Estimation of α (enrichment of eQTL in GWAS loci, prior in DAP): to deal with missing d, use DAP to sample the posterior of d 20-30 times: $p(d|Y_{qtl}, G_{qtl})$. Let $\hat{\alpha_1}^i$ be the estimate of the i-th imputed data, then the final estimate of α_1 is the mean of all estimates, and the variance can also be determined (Text S1). Note: this procedure underestimates α_1, as d posterior is from only eQTL data, while it should be conditioned on both eQTL and GWAS data.

- Fine-mapping of GWAS loci: let γ be SNP configuration, we need $P(\gamma_i = 1|D)$. DAP-1 is found sufficient in most cases. DAP-1 plus conditional analysis can find additional loci.

- Assessing co-localization: SNP colocalization probability (SCP) $P(\gamma_i = 1, d_i = 1|D, D_{qtl})$ given by Equation (8), where D is GWAS data. Note in LHS of the equation, should be $d_i = 1$. To see this, we note:
$$P(\gamma_i, d_i|D, D_{qtl}) = \frac{P(\gamma_i, d_i, D, D_{qtl})}{P(D, D_{qtl})} = \frac{P(d_i)P(D_{qtl}|d_i)P(\gamma_i|d_i)P(D|\gamma_i)}{P(D, D_{qtl})} \tag{7.63}$$
Note that $P(d_i)P(D_{qtl}|d_i) \propto P(d_i|D_{qtl})$, the PIP of eQTN. Full derivation in Text S1: in Eq. (7), the ratio is the product of posterior ratio of eQTN for $P(d_i = 1|D_{qtl})$; and the prior ratio of $P(\gamma_i = 1|d_i = 1)$. Note: $P(D|\gamma_i)$ term cancels out. Regional colocalization probability (RCP): is the sum of SCP for all SNPs.

- Analysis: with large $\hat{\alpha}_1$ (enrichment parameter for eQTNs), it is possible that a SNP with low eQTN PIP has a high SCP. Intuitively, SCP depends on the product of eQTL PIP and $P(\gamma_i|d_i)$, so a large $\hat{\alpha}_1$ can overcome small eQTL PIP. In other words, we can think of PIP from eQTL as the prior of a SNP being eQTN, and GWAS result as the data of this SNP, then the posterior of this SNP can be large if the SNP has high PIP in GWAS.

- Impact of the enrichment parameter α_1 (Figure 1): e.g. two SNPs in perfect LD, RCP is 0.5 with $\alpha_1 = 0$. But it increases significantly with large α_1. When it is 4, RCP is very close to 1.

- Connection with eCAVIAR and coloc: (1) eCAVIAR: simply fine-mapping on both traits, so it's equivalent to $\alpha_1 = 0$. (2) coloc: correspond to relatively large values of α_1, Equation (10), for default values of coloc, this is $\alpha_1 = 4.6$, or 100 fold enrichment. Intuitively, two independent association by chance are very unlikely to hit the same SNP $p_1 \times p_2 = 10^{-8}$; however, colocalization has a relatively large prior chance $p_{12} = 10^{-6}$.

- Simulation: use 1000GP real genotype data. Choose α_0 and ϕ (standard deviation of causal SNP effect size) s.t. the GWAS Z-score follows real distribution. Using height GWAS, $\alpha_0 = -8.4$ (or 2×10^{-4}) and $\phi = 0.4$.

- Accuracy of estimation of α_1 (Table 1): when it is < 2, often have large confidence interval, so statistically not significant. Comparison with alternative approaches (Figure S2): (1) Best SNP per gene is a reasonably good approximation: it slightly underestimates α_1 when true value is large (4 or 5) (2) Mean imputation (using PIP of eQTL): at $\alpha_1 \leq 2$, it has large standard error; at $\alpha_1 \geq 3$, mean imputation has similar standard error as multiple imputation, and slightly overestiamtes α_1 (while best SNP and multiple imputation underestimate). This can be due to the scaling effect (Text S1). Also shrinkage has small effect when $\alpha_1 > 3$.

- Power of colocalization (Figure 3): when $\alpha_1 = 0$, power is low, < 0.1. With larger α_1, e.g. 4, could reach 40%.

- Blood eQTL vs. lipid GWAS: α_1 varies from 0.4 to 5 for several lipid traits (Figure 5). Estimated number of eQTL 8.9K. The number of GWAS loci is around 50. The number of co-localized loci: around 20.

- Analysis: Enloc α_1 estimation may be too conservative or lead to large confidence interval. Enloc performs eQTL fine-mapping only on selected eGenes (user-specified). Typically, the list is conservative, say, FDR < 0.1. When running Torus to estimate α_1, still use all genes, but for genes not selected as eGenes, the PIPs of their SNPs would be 0.

Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics (S-PrediXcan) [Barbeira and Im, NC, 2018]

- Test statistic: Let T_g be expression of gene g, we know $T_g = X^T W$, where X is genotype. Our problem is simple regression of $Y \sim T_g$. The estimated effect is:

$$\hat{\gamma}_g = \frac{\text{Cov}\,(T_g, Y)}{\text{Var}\,(T_g)} \tag{7.64}$$

  Use $T_g = X^T W$, it is easy to see that

$$\text{Var}\,(T_g) = \hat{\sigma}_g^2 = W^T \Gamma W \tag{7.65}$$

  where Γ is the sample covariance matrix of X. Next we consider the covariance term:

$$\text{Cov}\,(T_g, Y) = \text{Cov}\,(\sum_l w_{lg} X_l, Y) = \sum_l w_{lg} \text{Cov}\,(X_l, Y) = \sum_l w_{lg} \hat{\beta}_l \hat{\sigma}_l^2 \tag{7.66}$$

  where we use the relation of $\text{Cov}\,(X_l, Y)$ and the estimated effect size of X_l.

- Computing Z-score: next we need to compute the s.e. of the estimate $\hat{\gamma}_g$. From the simple regression model, is just $\sigma_Y^2/(n\hat{\sigma}_g^2)$. We relate $\hat{\sigma}_l^2$ with $\hat{\sigma}_Y^2$. Putting all these together, we have the Z-score:

$$Z_g = \sum_l w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{se(\hat{\beta}_l)} \tag{7.67}$$

Probabilistic fine-mapping of transcriptome-wide association studies (FOCUS) [Mancuso and Pasaniuc, review for NG, 2018]

- Background: credible set. Given a set of configurations C, define its posterior as: $P(C|D) = \sum_{\gamma \in C} P(\gamma|D)$ Choose a minimum C to reach $P(C|D) \geq \rho$ for some threshold ρ.

- Problem: in TWAS, z-scores of genes can be correlated. Ex. if each gene has a causal eQTL, but the eQTL of multiple genes are in LD, then an eQTL of one gene may be associated with another gene. More formally, let $\hat{G}_j = X\Omega_j$ be the predicted expression of gene j, and Ω_j be the effect size of SNPs on gene j, then we can analyze dependency of $\hat{G}_j$ via LD and Ω_j's.

- Method overview: derive the distribution of TWAS Z-scores of all genes in LD regions, accounting for dependency of Z-scores due to LD.

- Model: let X be genotype vector, and G be expression of m genes, we write the true model of phenotype as:

$$y = X\beta + G\alpha + \epsilon \tag{7.68}$$

  where α is the vector of gene effects. For gene expression, our model is: $G = XW + E$, where W is eQTL effects. The imputed expression in TWAS: $\hat{G} = X\Omega$, where Ω is the estimated effects (accounting for LD). The TWAS Z-scores for gene j is normalized $\hat{G}_j^T y$. Let V be the LD matrix. This allows us to show that:

$$Z_{\text{twas}}|\lambda_{\text{snp}}, \lambda_{pe}, \Omega, V \sim N(\Omega^T V \lambda_{\text{snp}} + \Gamma\lambda_{pe}, \Gamma) \tag{7.69}$$

  where $\Gamma = \Omega^T V \Omega$ is the covariance of predicted expression of genes, $\lambda_{pe} \propto \alpha$ is the vector of true gene effects and λ_{SNP} is the pleiotropic effect of SNPs (vector). Assume a simple model for λ_{SNP}: vector of common effect size (a parameter to be estimated). To fine-map: spike-and-slab prior of λ_{pe}.

- Analysis: consider a region with two genes, several scenarios. (1) A single causal SNP has effects on both genes: only one is causal. Then this SNP has large GWAS effects. It is hard to distinguish the two genes. (2) Two causal SNPs, one for each gene: let SNP 1 be causal for gene 1, and SNP 2 for gene 2, and gene 1 is causal for trait. Then SNP 1 should have large GWAS effect, and SNP 2 smaller due to LD with SNP 1 (not because of effect on gene 2). Possible to fine map using either SNP level info. or gene-level information.

- Simulation: 25 blocks, in each block, average of one causal gene. Each gene has 1 or 2 causal SNPs in 100kb regions for both eQTL and GWAS. Ω estimated using GBLUP.

- Results: in simulation, type I error is controlled (Figure 2). When the causal gene is missing, most often choose null model in the credible set (Figure 3). For power, comparison with TWAS z-scores, higher AUC (Figure 4).

- Application to lipid GWAS: reduce the size of TWAS genes: from 60 to 30-40.

- Remark: the method may lose power. Intuitively, we should fine-map at the SNP level, as co-localization of eQTL and GWAS provides a good signal of causal gene. Also, the if we know causal variants of gene expression, Ω, the problem is considerably easier, however, no effort is done to fine-map eQTL, instead, Ω is obtained from a polygenic model (GBLUP).

Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits [Hormozdiari and Price, NG, 2018]

- Creating PIP annotations: use CAVAIR to fine-map SNPs. Several options: MaxCPP - for each SNP, the top PIP among all cis-genes (SNPs not in any credible set assigned 0). All-cis-eQTL: all cis-eQTL passing threshold. Top-cis-eQTL: top eQTL per gene. Credible set: binary annotation of whether a SNP is in the credible set.

- Estimation of enrichment and effects of annotations in S-LDSC: the model is for SNP j, we have the prior $\text{Var}(\beta_j) = \sum_c \tau_c a_{cj}$, where τ_c is the effect of annotation c. To estimate enrichment: assuming c is binary, then the heritability explained by c is simply the sum of heritability of all SNPs belonging to c: $h_g^2(c) = \sum_j a_{jc}\text{Var}(\beta_j)$. This relationship is generalized to continuous case.

- Difference of effects and enrichment of an annotation: if two annotations are correlated, then one will learn independent effects (one annotation has effect 0), but both will have significant enrichment. Intuitively, the SNPs of the two annotations will overlap, thus enriched with heritability.

- Simulation: Top-cis-eQTL show 3-4 times over-estimation of effects. Other annotations fine.

- Fine-mapped SNPs are enriched with GWAS heritability (Figure 2): averaging over multiple GWAS traits, use blood eQTL or FE-meta analysis eQTL (similar results). Enrichment: all eQTL (6%) - 2 fold; credible set (2%) - 2.5 fold; MaxCPP (0.1%, defined as the average value of annotations) - 5 fold. Effect: larger difference among the three groups.

- Lesson: to better use eQTL in studying traits, use MaxCPP or Credible set as annotations.

Principled multi-omic analysis reveals gene regulatory mechanisms of phenotype variation (pGENMi) [Hanson and Sinha, GR, 2018]

- Goal: identify TF-phenotype associations. Intuition: if TF is important, then its targets are likely enriched with GWAS signal.

- Model: given a TF, consider all its targets. Suppose we have p-values of target association with phenotypes (e.g. from TWAS). Let Z_g be an indicator of whether gene g is associated with phenotype. When $Z_g = 0$, uniform distribution, and when $Z_g = 1$, Beta distribution. The prior of Z_g depends on regulatory evidence of g: r_{gm} for the m-th type of evidence of TF regulating g. Use eQTL or eQTM: whether the eQTL or eQTM of g is a target of TF in ChIP-seq.

- Analysis: difference from Torus/S-LDSC: test enrichment of h2 in TF targets (1) use gene level evidence for phenotype association. (2) Use cis-eQTL/eQTM to link TF to target genes.

Quantifying genetic effects on disease mediated by assayed gene expression levels [Yao and Gusev, NG, 2020]

- Model: let w_j be the GWAS effect of SNP j, it may act on trait via gene expression with β_{jk} the effect on gene k, or pleiotropic effect γ_j. This leads to the effect size equation:

$$w_j = \sum_k \beta_{jk}\alpha_k + \gamma_j \tag{7.70}$$

  With this, derive LDSC type of equation/MOM estimator. Let χ^2_k be the chi-square statistics of SNP k, and d be index of gene category:

$$\mathrm{E}\left(\chi^2_k\right) = N\sum_c \tau_c l_{k,c} + N\sum_d \pi_d L_{k,d} + 1 \tag{7.71}$$

  where c is the index of SNP category, and π_d is per gene contribution, and $L_{k,d}$ is the expression score of SNP k. It is defined as:

$$L_{k,d} = \sum_{i\in D}\sum_j r^2_{jk}\beta^2_{ij} \tag{7.72}$$

  So the expression score of a SNP wrt. genes in set D measures its total LD to eQTLs of all genes in D, weighted by eQTL effect sizes.

- Simulations under model assumptions: 10,000 GWAS samples, 100-1,000 eQTL samples. Only Lasso and REML correction gives unbiased estimates. BLUP and OLS: terrible results. Varying sparsity: e.g. 10% of genes or 10% of SNPs, overall robust, but at the setting where mediated h^2 is high, significant under-estimate with 10% genes.

- Simulations under violation of effect size assumption: effect size of genes depend on cis-heritability. Binning by cis-heritability is important.

- Results: use GTEx and 42 traits, median explained h2 is 0.11. Also inverse relationship between cis-heritability of expression and disease heritability mediated by expression.

- Remark: it may be important to consider prediction errors, as the results are sensitive to prediction methods used.

Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics [Luningham and Jingjing Yang, AJHG, 2020]

- Prediction model of expression: BVSR model, with different priors on cis-SNPs and trans-SNPs. Fit with EM-MCMC: parallel blocks, using summary statistics to speed up computation, pruning of blocks - only those with certain p value cutoff. Running time: a single gene, about 30 mins.
- BGW-TWAS: same as S-PrediXcan, the Z-score of association is sum of w_{lg}, the weight of SNP l on gene g (posterior mean from BVSR), times GWAS Z-score, times $\hat{\sigma}_l/\hat{\sigma}_g$, where $\hat{\sigma}_l$ is the s.d of SNP l, and $\hat{\sigma}_g$ is the s.d. of imputed gene expression.
- Genetic architecture of gene expression: Table 4, in brain eQTL data of 500 samples, generally the total PIP sums to 1-3, most in trans.
- Results of BGW-TWAS in AD: use AD with individual level data. One gene with 5 supporting SNPs, all in trans-. However, 4 SNPs are in LD (but selected by BVSR).
- Remark: no cross-validation of gene expression heritability. Sign of potential problem: genes with smallest R^2 from BVSR have highest PIP sum (Table 4).

7.4 System Genetics Studies

Mouse obesity study with liver eQTL [Schadt & Friend, Nature, 2003]

- Methods: 112 F_2 mice using liver sample, see notes before for details.
- Results:
  - Candidate genes of obesity through expression data: 280 genes significantly DE in two groups of mice: high FPM and low FPM (related to obesity). The expression pattern of these gene can be used to cluster mice and the results correspond to the two groups.
  - Heterogeneity of the FPM trait: gene expression patterns ientify two subgroups of high FPM mice. Also verified by eQTL: only by using one subgroup of FPM mice (plus the low FPM group), some eQTLs could be identified (the chromosome 2 region).
  - eQTL of the candidate genes: five regions containing eQTL os more than 50% of the genes in the FPM set. Analysis of the chromosome 2 region identifies two candidate genes: they are located within this region, and they have eQTLs in this region. One gene is a protein glycoylatransferase, and the other is a cation-transporting ATPase.
- Remark: in this case, if just use QTL of FPM trait, only four QTLs were found with LOD score greater than 2.0.

Candidate genes in hematopoietic stem cells (HSC) turnover [Bystrykh & Haan, NG, 2005]:

- Problem: HSC turnover is determined by the fraction of cells in the S phase in HSC. What genes may influence the HSC turnover? The strategy is to use genetic variations (of strains) to identify the causal genes.
- Data: 30 BXD mouse strains (see before).
- Co-mapping of eQTLs to trait QTL: a QTL of HSC turnover has been identified before, called stem-cell proliferation (Scp2). 8 cis-acting eQTL were mapped to this region.

Candidate genes of weight in mouse through liver eQTL [Ghazalpour & Horvath, PG, 2006]:

- Data: Liver expression data of 135 famale mice from a F2 inter-cross between two inbred strains.

- Gene module identification: weighted co-expression network approach: the link between two genes is the power of the correlation coefficient (the power function makes the results less sensitive wherease unweighted networks display sensitivity to the choice of cutoff). 12 distince modules were found.
- Association of modules and clinical traits (weight): defined through average correlation (absolute value) of the genes in the module and the trait. The most significantly associated module is enriched for genes in the EMC-receptor interaction and complement and coagulation cascades.
- moduleQTL (mQTL) analysis: the locus with a significant enrichment for eQTL of the genes within the module, assessed by Fisher's exact test.

Molecular networks underlying metabolic syndrome [Chen and Schadt, Nature, 2008]:

- Motivation: diseases are caused by change of states of molecular networks (a functional module), thus identyfy these disease-causing modules.
- Methods:
  - 334 F_2 mouse from crossing two strains. Each mouse is: genotyped with 1,300 SNPs, expression-profiled (liver and adipose), and phenotyped (metabolic syndrome).
  - eQTL and cQTL (metabolic traits) identification.
  - Assess whether a expression trait is supported as having a causal relationship with metabolic traits [Schadt & Lusis, NG, 2005].
  - In co-expression networks, first identify modules and then test the modules where the causal genes are enriched, by Fisher Exact Test.

Genetics of obesity through eQTL in adipose tissue [Emilsson & Stefansson, Nature, 2008]:

- Methods:
  - Data: 2 cohorts (1,002 and 673 respectively) with expression profiling (23,720 transcripts) in blood and adipose tissues, and clinical traits including body mass index (BMI), percentage body fat (PBF), etc.
  - Identification of obesity-related gene modules: (1) select all genes that show differential expression in adipose tissue, then construct pairwise coexpression network (weighted) in human and search for co-regulated module; (2) the coexpression network in mouse and conserved modules; (3) test if the conserved module(s) is enriched with obesity-associated genes.
- Results:
  - Gene-clinical trait association: in blood sample, less than 10% genes show association with clinical traits; while in adipose tissue, 60-70% genes show association. To narrow down genes, find conserved co-regulated modules between human and mouse. This reveals a single core module enriched with macrophage function and metabolic traits (MEMN). And 886 genes in MEME (98%) are significantly correlated with BMI.
  - Expression QTLs: two methods are used (1) eQTLs using family information, with 1,732 microsatellites; (2) eSNPs using 317,503 SNPs on 150 unrelated individuals. The results between the two are highly consistent. At FDR 0.05, detected cis eSNPs for 2,417 (15%) expression traits in blood and 3,048 (14.6%) traits in adipose, and the two sets of cis eSNP highly overlap (50%).
  - Trans-eQTLs and trans-eSNPs: at FDR 0.05, found 52 and 25 trans eQTL signals of genome-wide significance in blood and adipose tissues; and 67 trans eSNPs signals in blood and 59 in adipose tissue. Few eQTL hotspots were detected, beyond what was expected by chance (through simulation: note that due to correlation of transcripts, some false eQTL hotspots would be expected). One possible hotspot at 3q22.

  - eQTL of the MEMN genes: most MEMN genes have cis eSNPs, and confirmed that some cis eSNPs show association to BMI (genotyping of 768 cis eSNPs).

Application of liver eQTL to GWAS of T1D and CAD [Schadt & Ulrich, PLoS Biol, 2008]:

- Methods: see the section "Genetics of Gene Regulation" for data and eQTL finding. To identify disease genes using eQTL, use co-mapping, i.e. find genes whose expression are mapped to the same SNP as the GWAS of the complex trait. The functional evidence of a candidate gene: construct the causal network of eQTL, and check if the linked genes in the network are enriched with known genes of these traits.
- Application to T1D: from T1D-associated SNPs, find the genes in the vincinity of these SNPs whose expressions are correlated with them (Table 1). The results may include eQTLs not passing the threshold: in fact, only 3 genes: RPS26, CLECL1 and HLA-DRB1 have eSNPs with $P < 10^{-5}$. Out of 13 genes found in this way, 9 are known T1D risk genes.
  - Rps26: the candidate gene for the cis-SNP (rs2292239) was thought to be Erbb3 in WTCCC studies. But the SNP is not associated with expression of Esrrb3, but Rps26 in liver eQTL data. The evidence that Rps26 is the true candidate gene: the causal network from the eQTL using mouse expression data Rps26, but not Erbb3 is related to a number of genes known to be important in T1D.
  - CLECL1: the candidate gene of rs3764021 was thought to be CLEC2D in WTCCC studies. With eQTL data, it was found this SNP is associated with CLECL1, but not CLEC2D.
- Application to CAD: WTCCC identified 7 significant SNPs, check if any of them is a eQTL using threshold $P < 0.05/(7 \cdot 40000) = 1.79 \cdot 10^{-7}$ (test 7 SNPs for 40,000 expression traits). Only one of them (rs599839) is associated with any expression trait: PSRC1, CELSR2, and SORT1. This SNP is also associated with LDL in a GWAS of LDL. Relevant evidence of these genes:
  - Psrc1, Sort1 were significantly associated with plasma LDL level in mouse, and
  - Psrc1 and Sort1 were part of the MEMN network in the causal network constructed from mouse eQTL data.

Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations [Nica & Dermitzakis, PG, 2010]:

- Motivation: for GWAS-SNPs, find the expression trait that it influences (thus suggesting the mechanism of its action). The simplest method would be to test if a GWAS-SNP is co-localized with some eQTL, but it is possible that GWAS-SNP and an eQTL may tag different causal SNPs becuase of LD.
- Methods:
  - Data: 976 GWAS SNPs from NHGRI, into 784 intervals (defined by recombination hotspots). Genotype and expression data of lymphoblastoid cell lines (LDL) of 109 unrelated HapMap individuals.
  - RTC: measure the association of GWAS-SNP with an expression trait. Suppose there is an eQTL in the interval of GWAS-SNP, the idea is: if GWAS-SNP has a large effect on expression, then the eQTL should have weak association with the expression trait after removing the effect of GWAS-SNP (i.e. association of eQTL and the residual of GWAS-SNP and expression trait). Since an interval may have multiple SNPs, using the rank of GWAS-SNP relative to all other SNPs in the interval.
- Results:

- GWAS SNPs are enriched for regulatory variants: for these SNPs, find its cis- and trans- regulatory effects. The distribution of regulatory effects in GWAS SNPs is higher than that of the random SNPs.
- Comparison of RTC and correlation-based measures: RTC performs better than r^2 or D' between GWAS-SNP and eQTL, in both simulation data and in real data. (1) RTC outperforms r^2 because it is able to recover causal effect even for low correlated pairs. (2) A high D' is insufficient to predict causal effect because it is impossible to distinguish causal from coincidental effects given a strong historical correlation.
- GWAS-SNPs with significant cis-regulatory effects: out of 976 GWAS SNPs, 130 have at least one cis-eQTL (with $P < 0.05$ after permutation) in the GWAS interval. Among these, 28 have RTC > 0.9. Many related genes are immunity related (consistent with the use of LDL expression data).
- GWAS-SNPs with significant trans-regulatory effects (trans- is defined as 5 Mb from TSS or TES): top 50 GWAS intervals ordered by trans-eQTL significant. Among these, 24 have RTC > 0.9.

- Remark:
  - The problem of using r^2 and D' is lack of normalization wrt. the local LD patterns. Ex. in a region with low LD, even small value of r^2 may be very significant; alternatively, in a region with high historical correlation, even high D' may not suggest causal effect.
  - The RTC method is essentially a way of assessing the significance of the feature (GWAS-SNP) in the presence of a correlated feature (eQTL). Thus using residual from using the eQTL feature, and regression (How is this method compared with e.g. feature selection by testing if $\beta = 0$ for the feature of interest?). Meanwhile, RTC performs normalization by ranking the GWAS-SNP in all SNPs of the same interval (thus normalize wrt. the SNP density, the LD of the region, etc.).

Association of pathway expression and complex traits [Zhong & Schadt, AJHG, 2010]:

- Strategy: the hypothesis is: the perturbation of expression of a pathway leads to a complex disease. Thus, to test association of a pathway (its expression) and a trait, we first find the eQTLs of the genes belong to this pathway, then test the association of these eQTLs to the trait.

- Methods:
  - Data: gene expression profiling of 707 liver samples, 916 omental adipose samples and 870 subcutaneous adipose samples.
  - Pathway expression-association test: for each gene, among all eSNPs (at FDR 10%), choose the one with the strongest association with T2D (usually single independent eSNP). Then test the pathway association with T2D, similar to GSEA (the deviation of P value distribution from the null distribution). Only 110 KEGG pathways are tested, with size of 20 to 200 genes.
  - Determine the FDR: any pathway has an enrichment score NES (normalized s.t. it is comparable across pathways), meanwhile, obtain the NES null distribution by permutation: switch case and control label and computer the NES scores for every gene set.
  - Controls in pathway association test: (1) positive controls - one set of 18 genes replicated to be associated to T2D in WTCCC and another set consisting of these 18 and other random 22 genes; (2) negative controls - gene sets of size 20, 40, ..., 200, randomly selected from the set of genes associated with at least one eSNP.
  - Weight subtraction algorithm: to perform the analysis on a different dataset (WTCCC vs. non-WTCCC), need to get the association statistic of non-WTCCC, however, only P-values from DIAGRAM (meta-analysis of WTCCC and non-WTCCC) are available. Need to calculate P

value from non-WTCCC. This is done by computing based on Z-scores (DIAGAM used fixed-effect weighted average of the summary Z scores):

$$Z_{\text{DIAGRAM}} = w_{\text{WTCCC}} \cdot Z_{\text{WTCCC}} + w_{\text{non-WTCCC}} \cdot Z_{\text{non-WTCCC}} \tag{7.73}$$

where the weights are proportional to the square root of the effective sample sizes.

- Results:
  - eQTL results: identified a total of 20,563 eSNPs of 9,964 genes. On average, 30% eSNPs are tissue-specific, the rest common to all three tissues.
  - Significant pathways with T2D using WTCCC: 23 out of 110 KEGG pathways have nominal P value < 0.05 (16 with FDR 20%). Correlation among KEGG pathways (shared genes) may explain the excess of significant pathways. Two positive controls are highly significant (due to 6 genes with eSNP, where 3 of them have small P_{T2D} in WTCCC), and negative controls not. Note that individual eSNPs generally show subtle T2D association (most have P value in the range of 10^{-3} to 10^{-4} in WTCCC data), see Figure 1.
  - Replication in DIAGRAM dataset (Table 1): 9 out of 23 pathways were replicated ($P < 0.05$). In a different replication, removing WTCCC statistics from the DIAGRM statistics, and 4 pathways were replicated.
  - Biological evidence of the pathways: some have known function in T2D, e.g. calcium signaling, PPAR signaling, TGF-β signaling. The hematopoietic cell lineage pathway reflects mainly the immune response and the inflammatory pathway, which has been extensively linked to T2D. Novel pathways include: tight junction, adherent junction, complement and coagulation, and antigen processing and presentation.
  - False negatives: four known T2D pathways were missed. Two due to technical reasons (not selected for pathway association test), and the other two may be due to the incomplete coverage of eQTL:many genes may not affect expression; eQTL study power may be low; relevant tissue not profiled, etc.

A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk [Heinig and Cook, Nature, 2010]:

- Data: eQTL of 7 tissues (fat, kidney, heart, skeletal msucle, liver, etc.) in recombinant inbred rat strains.
- Identifying IRF7 as a putative master regulator: (1) Found 147 TFs with at least one eQTL in 7 tissues, most ($> 90\%$) under trans-regulation. (2) Co-localization analysis: for each TF eQTL, find all transcripts controlled by the eQTL. This gives a list of transcripts for each TF. (3) Enrichment of direct TF targets in the transcripts associated with a TF. Over all TFs: 13 show over-representation of TFBSs in target gene promoters. IRF7 is the strongest with 23 targets, all controlled by a single locus at the rat 15q25 locus.
- Define TF-driven subnetworks (IDIN): Co-expression analysis: expand to genes co-expressed with IRF7 targets. This identified 247 genes. The targets are enriched with anti-viral genes. All the genes are in trans
- Causal gene regulating Irf7 and its targets: from the trans-eQTL locus, sequencing identified the gene Ebi2 (the other genes have no functional SNPs).
- Association of IDIN in human: SNPs close to any IDIN genes were significantly more likely to associate with T1D in GWAS than SNPs close to genes not in the network (Figure 3).

- Role of Ebi2: the minor C allele of SNP rs9585056 was associated with T1D risk, lower EBI2 expression (cis-eSNP) in both GHS (monocyte eQTL data) and Cardiogenics Study cohorts, and, on average, increased expression levels of IDIN genes in the Cardiogenics Study cohort.

Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA [Fehrmann and Lude Franke, PLG, 2011]:

- eQTL data: a genome-wide eQTL analysis on 289,044 common SNPs, present on the Illumina HumanHap300 platform in peripheral blood expression data of 1,469 unrelated individuals. Permutation test, FDR < 0.05.
- eQTL results: non-parametric Spearman's rank correlation. (1) cis-eQTL $P < 1.73E-3$, number of unique eQTL genes $= 7,589$, number of unique eQTL SNPs $= 48,717$ (2) trans-eQTL $P < 3.6E-9$, number of unique eQTL genes $= 202$, number of unique eQTL SNPs $= 467$. HLA single-nucleotide polymorphisms (SNPs) were 10-fold enriched for trans-eQTLs: 48% of the trans-acting SNPs map within the HLA.
- Trans-effect of trait-associated SNPs: among 1,167 SNPs from GWAS catalog, 472 (40.4%) of these SNPs were cis-eQTLs, affecting 538 genes; 67 (5.7%) SNPs were trans-acting on 113 genes.
- Some examples of trans-eSNPs that also affect phenotypes:
  - UC: rs2395185 is the strongest risk locus, and also the strongest SNP, trans-acting on AOAH, an enzyme that modulates host inflammatory responses to gram-negative bacterial invasion. AOAH is significantly co-expressed with colony stimulating factor 1 receptor (CSFR1) and HLA-DRA. Hyperstimulation of CSFR1 is implicated in UC.
  - T1D: 59% (30/51) of the known and tested T1D associated SNPs are cis-acting (on in total 53 unique genes) and 17% (9/50) are trans-acting on 22 unique genes. Potentially interesting trans-genes include CCL2, CFB, CLN1, KRT19, OSR1 and RARRES1, all strongly co-expressed with each other. CCL2 and CFB are known immune response genes and have been implicated in T1D before.
  - Breast cancer: rs3803662 trans-acting on origin recognition complex subunit 6 (ORC6L), involved in DNA replication and has been frequently used as part of prognostic profiles for predicting the clinical outcome in breast cancer.
- 7 unique pairs of unlinked SNPs that are associated with the same phenotype and that also affect the same downstream genes in trans or cis (at FDR 0.05), the expected number of pairs is only 0.15 (permutation test). Similar results at FDR = 0.50, 18 pairs, 21 times enrichment vs. random.
- Examples of unlinked SNPs that affect both expression and phenptypes:
  - Hemoglobin protein level: three independent loci on hemoglobin gamma G (HBG2) expression
  - Mean corpuscular volume (MCV): two independent SNPs affect several genes, some of which are known blood coagulation genes
  - mean platelet volume (MPV) : two SNPs, include one cis-, on tropomyosin 1 (TMP1).
- Phenotypic buffering: the effect of trans-eSNPs on genes are higher than the effect of these SNPs on phenotypes. This is a sign of causal influence. Ex. MPV- and MCV-associated SNPs explain between 1.41% and 10.99% of trans-expression variation while only explain 0.24% and 1.12% of the MPV and MCV phenotype variation.
- Replication of trans-eQTL:
  - Replication in monocyte expression data (Zeller10): 46 out of the 130 different trans-eQTLs, each has $P < 1E-5$ in the monocyte data.

– Replicated 18 trait-associated trans-eQTLs in four different non-blood tissues (subcutaneous adipose, visceral adipose, liver and muscle).

Genetic Mapping with Multiple Levels of Phenotypic Information Reveals Determinants of Lymphocyte Glucocorticoid Sensitivity. [Maranville & Di Rienzo, AJHG, 2013]

- Concept: lymphocyte GC sensitivity (percent inhibition of cell proliferation due to GC) is correlated with clincal response to GC therapy in a wide range of diseases. So we profile GC sensivity as well transcriptome respose.
- Data: PBMC from 88 AAs. First treat with PHA (induce inflammatory response), then with GC. Do expression profiling and GC sensitivity of individuals.
- Finding candidate genes of GC sensitivity: genes whose expression correlated with GC sensitivity, found 27 genes (or 85 with a different metric).
- GWAS for GC sensitivity: one SNP in RBMS3 highly significant, explaining 26% of variation in phenotype. This SNP also associates with 14/27 genes correlated with GC sensitivity.
- Experimental evidence of RBMS3, tumor suppressor gene: ASE of RBMS3 in this SNP, only in stimulated condition. Knockdown of RBMS3 induces PHA-mediated cell proliferation.
- **Lesson**: lymphocyte in vitro sensitivity as a surrogate of clincal response. Trans-acting effects of SNPs.

Systematic identification of trans eQTLs as putative drivers of known disease associations [Westra and Franke, NG, 2013]

- Identifying trans-eQTL: (1) eQTL data: 5000 peripheral blood samples. (2) Trans-eQTL analysis: focuse on 4,500 SNPs in GWAS catalogy, found 1,513 significant trans-eQTL affecting 400 genes. Trans-eQTL show similar effect sizes across various cohorts. (3) Replication of trans-eQTL: $> 50\%$ replicated in independent studies of peripheral blood eQTL.
- Tissue-specificity of trans-eQTL: most cannot be replicated in adipose.
- Case study: trans-eQTL affecting expression of disease-related genes. IKZF1 locus: two SNPs, one associated with SLE, also trans-eQTL of 8 genes, including two groups, complement, and type I interferon response genes, both known to play a role in SLE. Also confirm the IKZF1 as likely cis-gene.
- Convergence of multiple GWAS SNPs on the same downstream genes from trans-eQTL (Table 2): in 21 different traits, most are blood related, but also blood pressure, immune diseases and T1D.

Integration of Transcriptomic Profiling, Genome-wide Association, and Network Biology Reveals Molecular Mechanisms Underlying Blood Pressure Regulation. [Huan and Levy, Molecular Systems Biology, 2015]

- Strategy: (1) BP (blood pressure) - transcriptome analysis to identify DE genes and subnetworks correlated with BP. (2) To identify possible causal subnetworks, map eQTL, and use BP GWAS to prioritize subnetworks. (3) Upstream regulator analysis of candidate subnetworks.
- BP subnetwork analysis: transcriptome data from 3,679 FHS subjects. DE gene analysis: 83 genes. Define co-expressed modules (coEMs) - WGCNA, then correlate eigen-genes with BP. Found 6 coEMs.
- Identifying causal modules using SNP set enrichment analysis (SSEA): for each BP-correlated gene set, retrieve their eSNPs, and test overlap of these eSNPs with GWAS loci of BP. Use KS test and Fisher's exact test for assessing if the GWAS p-values of these eSNPs depart from null distribution. Results: 4 coEMs likely causal.

- Identification of key drivers (KDs): use blood Bayesian networks (BNs) and PPI. KD is defined as a local network hub whose neighbors show enrichment for BP genes in the causal gene set. Found 545 KDs from PPI network and 131 from BN. KD ranking: association in BP GWAS, BP correlation in expression, their statistics in KD analysis.
- Candidate KD gene: SH2B3, a missense SNP associated with BP and hypertension, and a trans-eSNP for 16 genes. Sh3B3 knockout mice show normal baseline BP but elevated BP in response to angiotensis II.

Systems genetics identifies Sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus [Johnson and Petretto, NC, 2015]

- Data: 129 TLE (epilepsy) patients, brain eQTL.
- Discovery of modules: Gaussian Graphical model, 400 genes. Then refine with hier. clustering: two modules of size about 70 genes.
- Validation of modules: enrichment of PPI (DAPPLE), KEGG enrichment, conservation in mouse (highly co-expressed).
- Genetic regulator: two step strategy (Figure 3), genetic association analysis with PC1 of module 1 (Bayesian variable selection model for association test). Then refine the association analysis: using hotspot analysis algorithm (HESS), find several more SNPs. Most are associated with a large percent of genes in module 1.
- Discovery and validation of SESN3: among genes in the QTL hotspot, SESN3 expression shows the highest correlation with the module. Validation: K.D. of SESN3, and observe reduced expression of genes in the module in macrophages.
- Lesson: trans-eQTL hotspot may be best studied at the level of modules < 100 genes. To narrow down to exact regulators of module: use co-expression of the putative regulator with genes in the module. Usually, genes in a QTL region are functionally unrelated, so this test may be able to discriminate genes in QTL regions.

Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types [Chun and Costapas, NG, 2017]

- Data: AID GWAS with 272 loci. CD4 T cells, monocytes and LCL eQTL data.
- Only in about 25% of cases, GWAS and eQTL signals colocalize.

Shared Genetic Regulatory Networks for Cardiovascular Disease and Type 2 Diabetes in Multiple Populations of Diverse Ethnicities in the United States [Shu & Yang, review for PLG, 2017]

- Goal: infer the common biological processes disrupted in the two diseases, and the key driver (KD) genes.
- Background: Patients of T2D are 2-6 times more likely to have CVD.
- Construction of tissue-specific modules: use WGCNA on gene expression data in relevant tissues, including adipose, blood, liver, heart, islet, kidney, muscle and brain, from human and mouse studies. 2,600 modules and annotate the modules using Reactome and KEGG.
- Relating co-expression modules to diseases: MSEA (market set enrichment analysis). (1) SNP-gene mapping: either eQTL; or SNPs within 50kb of genes that have annotations in Regulome. (2) For a gene set, test the enrichment of association in the mapped SNP set. Use chi-square like test, and obtain null from "shuffling gene labels". Results: 79 modules associated with CVD and 54 with T2D. 2 modules with both.

- Annotating modules and assessing the relationship between functional categories and diseases: significant sharing of functional categories (KEGG, Reactome) between two diseases.
  - Associate a category with a disease: if it is annotated to at least one disease module.
  - Shared categories/pathway: to rank them, consider the pathway associated with both diseases (through some modules). A pathway that is associated with multiple modules in both diseases would be favored (Figure 3).

  Top shared pathways include well-established processes such as lipid metabolism, glucose metabolism, oxidation, cytokine signaling. One novel finding: BCAA pathway, the genes themselves show little association, but their network neighbors do.

- Identifying key drivers: use GIANT networks and Bayesian networks from 25 CVD and T2D relevant tissues. KDs: genes whose local neighborhood neighborhoods show enrichment of genes from disease-associated modules. Results: 226 KDs. To find KDs associated with both CVD and T2D: significant in both phenotypes, replicated by GIANT and Bayesian networks, order by the strength of association between KD subnetworks and CVD/T2D Found 15 KDs at FDR < 0.1.

- Connection of KDs with known CVD/T2D genes: overrepresentation of these genes in KD neighborhoods in the networks. Remark: this may be circular.

- Role of KDs in CVD and T2D:
  - Mouse transcriptome data: 100 mouse strains with both expression and relevant phenotypes such as lipid levels, fasting glucose. All 14 KDs have significant trait associations.
  - Transcriptome perturbation by KDs: CAV1 knockdown changes expression of neighbor genes in vitro and in mouse model.
  - Mouse GWAS data: association of cadiometabolic phenotypes with KD or KD subnetwork genes.

  Note that KDs themselves are not top signals in GWAS.

- Biological evidence of KDs. CAV1: KO mouse shows phenotype. HMGCR: target of statin. IGF1. Three ECM KDs: e.g. SPARC inhibits adipogenesis and is associated with insulin resistance.

- Remark: the evidence of KDs do to fully establish their causality: (1) transcriptome correlation; (2) in mouse GWAS, association with KD neighbors, not KD genes.

- **Lessons**:
  - Gene pathway analysis: using co-expression modules in disease-related tissues can be a better srategy than using pre-defined GO/KEGG pathways.
  - Key driver gene identification: many neighbors (or targets) that are disease related. The KD genes may not have strong association in GWAS themselves. Also note that in this analysis, there are relatively large number of KDs in one disease.
  - Validation strategy: using mouse data, both GWAS and transcriptome association.

Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis (eQTLGen) [Vosa and Franke, Biorxiv, 2018]

- Data set: 30K samples, meta-analysis.

- Cis-eQTL: most are within 100kb. Distal eQTL (more than 100kb): 37% are in Hi-C contact.

- Gene prioritization of GWAS using cis-eQTL: SMR and DEPIT results do not agree.

- Trans-eQTL mapping: 10K trait-associated SNPs, about 1/3 are trans-eQTL, 6000 genes. Possible mechanisms: (1) TF - targets: 2.2 fold enrichment. (2) Colocalization with cis-eQTL: COLOC estimates 52%. (3) Mediation by cis-genes: (4) Gene co-expression. Together, TF-targets, cis-mediation and co-expression explain 17% of trans-eQTL. Also found interactions of cis-SNPs and cis-gene expression in determining trans-gene eQTL: 5 fold enrichment.
- **Analysis**: how to use trans-eQTLs to understand genetics of complex traits? Suppose we have a disease-associated SNP, and it is a trans-eQTL of some genes. The problems are: (1) What is the cis-mediator? (2) What are the trans-genes that mediate the effects of the SNP? We need additional evidence that the trans-genes are disease-related.
- Examples of how trans-eQTLs inform disease genetics: (1) GWAS SNP of age of menarche: cis-eQTL of ZNF131, which is involved in ER signaling. Trans-genes are enriched with ZNF131 targets by K.D. (2) T1D SNP: no cis-eQTL, but coding variant, trans-genes enriched with IFN response. (3) Asthma SNP: cis-gene involved in B cell proliferation, many trans-genes (cell type composition change).
- Convergence of multiple GWAS hits to the same genes (Figure 5): SNPs of SLE converge to a cluster of IFN response genes. Effect sizes of the SNPs on genes also show correlation: if one SNP shows a large effect in a a gene, other SNPs also show large effects.
- eQTS: expression quantitative trait scores, correlation of expression and PRS. Justification: let G_s be SNP s, and X_j be expression of gene j, and Y phenotype. Our causal model for gene j is:

$$\{G_s\} \xrightarrow{\beta_{s,j}} \{X_j\} \xrightarrow{\gamma_j} Y \tag{7.74}$$

  The PRS of sample i is basically Y_i (expectation). So we have the correlation of expression and PRS:

$$\text{Cov}\left(X_j, \text{PRS}\right) = \gamma_j \sigma_j^2 + \sum_{k \neq j} \gamma_k C(k, j) \tag{7.75}$$

  where σ_j^2 is the genetic variance of gene j, and $C(j, k)$ genetic correlation of expression of gene j and k.
- eQTS results: analysis of 1,200 traits and 28K samples, found 18K eQTLs effects, representing 689 unique traits and 2,500 unique genes. Average 10-20 genes per trait (some traits have none). Ex. ASD: found 21 genes, 3 genes have known functions in synapse.
- **Lesson**: prioritization by cis-eQTL is NOT easy, as shown by the disagreement of SMR and DEPICT.
- Remark: for majority of trans-eQTLs, we do not know their mechanisms. One mechanism is cell type composition change: could lead to many trans-eQTLs.
- **Lesson**: linking trans-eQTL and GWAS: (1) enrichment of some disease-related pathways; (2) plausible cis-genes with disease functions; (3) SNP leading to compositional change of disease-related cell type(s).
- **Remark**: under what conditions, we can translate correlation of expression with PRS to causal relations? In general, if PRS and gene share genetic variants, we will see correlation.

7.4.1 TWAS and cis-QTL Assisted GWAS

Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. [Nicolae & Cox, PG, 2010]:

- Motivation: the hypothesis is the expression changes can lead to complex diseases. This has two consequences: (1) a significant fraction of GWAS SNPs influence gene expression; (2) some eQTL genes likely increase disease risk, if expression perturbed. Testing these two hypothesis using GWAS and eQTL data.

- eQTL data: HapMap LCL data in SCAN.
- GWAS SNPs (1,598 from the GWAS catalog) are more likely to be eQTLs than MAF-matched random set of SNPs: at eQTL threshold $P < 10^{-4}$, found 625 GWAS SNPs (expected number about 600 at 5%), at $P < 10^{-6}$, found 46 (expected about 30 at 5%), and 17 (expected about 2 to 3 at 5%).
- Crohn's disease from WTCCC: (1) top 10,000 eQTLs: 357 SNPs associated with the trait (Crohn's disease) at $P < 0.01$, while the expected number was 117-178; (2) top 1000 GWAS-SNPs with Crohn's disease: enriched with eQTLs, e.g. at eQTL threshold 2, there are 324 SNPs (143.5 expected by chance), and at eQTL threshold 3, there are 172 SNPs (only 18.6 by chance).
- Other WTCCC diseases: (1) top 10,000 eQTLs: T1D and RA have significantly more SNPs than expected with phenotype associations $P < 0.01$. (2) the enrichment of eSNPs (eQTL function scores larger than 3) among the SNPs with the strongest associations to CD, T1D, RA, hypertension and bipolar disorder.
- Lesson: (1) Weak or intermediate SNPs may be eQTLs too; (2) eQTL data of LCL may be useful for complex diseases not directly related.
- Remark: the possible bias in the study, MAF (it is easier to identify trait-associated SNPs with high MAF because of higher power), LD (it is easier to find SNPs with higher LD). In the study, MAF is controlled, but not local LD.

Liver and adipose eQTLs are enriched for association to T2D [Zhong & Schadt, PG, 2010]:

- Hypothesis: many GWAS SNPs may be (weak) causal variants that affect gene expression. If this hypothesis is true, then among all eSNPs, some may influence expression of genes important for T2D, thus the T2D-associated SNPs may be enriched.
- eQTL Data: 427 liver cohort [Schadt08]; about 900 individuals with liver, subcutaneous adipose and omental adipose tissues. eQTL identification follows [Schadt08]. High overlap of eSNPs identified in three tissues in one cohort: about 70-80% of eSNPs in one tissue are also found in the other two. Also among liver eSNPs from the first cohort, there was a 66% overlap in eSNPs indentified between the two studies.
- The whole set of eSNPs: the distribution of P_{T2D} (p value in GWAS of T2D), is significantly different from random distribution (random SNP sets matching MAF). Ex. in DGI study, 6.2% of the eSNPs (241 out of 3,888 total) had $P_{T2D} < 0.05$, compared to a mean of 5.2% (202 out of 3,888) in random SNPs. The enrichment of SNPs with $P_{T2D} < 0.05$ is low, however, only 1.19 fold.
- Increase of enrichment: (1) using eSNPs of adiptose tissue of the genes that show differential expression in mouse studies (top 25%, i.e. 9,9000 genes). (2) further enrichment if using eSNPs of the causal disease subnetwork (MEMN1) of 159 genes: 37% vs. 9% (random) with $P_{T2D} < 0.05$.
- T2D causal gene: of 158 genes in the adipose causal network, 117 have cis or trans- eQTL in human, however, only 8 were identified with strong adipose cis-eQTL. Among these, only ME1 was associated with at least one cis-eSNP that was also associated with T2D with $P_{T2D} = 0.002$. The function of ME1 was verified.

Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain [Richards & O'Donovan, Molecular Psychiatry, 2011]:

- Hypothesis: polymorphisms that are associated with schizophrenia are enriched among brain eSNPs. If this hypothesis is true, then we classify all putative risk alleles of schizophrenia as "top eQTLs" and "bottom eQTLs", we would expect that the "top eQTL" risk alleles are better predictors of schizophrenia than "bottom eQTL" risk alleles. (The reason for testing predictability is the schizophrenia risk alleles are very weak.)

- Data: GWAS data from ISC and MGS. Brain eQTL data form [Myers07] and [Webster09]: analysis using linear regression controlling for many covirates, including a gene as a marker of number of neurons. Use cis-eQTLs in the analysis (trans-eQTLs did not show significant results). "Top eQTLs" are defined as top 5% ($P < 0.02$) and "bottom eQTLs" as bottom 5%.

- Comparison of "top eQTL" and "bottom eQTL" risk alleles: schizophrenia risk alleles are defined as $P < 0.5$, and the two classes of risk alleles are assessed by: applying the risk alleles in an independent test group, and compare the "polygenic score" (predictive score, summing all risk alleles weighted by log-odds ratio in the training data) of the cases and controls.

- The difference in the scores between the top and bottom cis-eQTLs was significantly greater in the cases than in the controls for all analyses. In other words, among the variants selected for marginal association to schizophrenia, those that additionally show evidence for being cis-eQTLs predict affection status better than those variants showing no evidence for being cis-eQTLs.

Loci nominally associated with autism from genome-wide analysis show enrichment of brain expression quantitative trait loci but not lymphoblastoid cell line expression quantitative trait loci. [Davis & Cox, Mol Autism, 2012]

- Background: pathway analysis of autism GWAS reveals some intriguing pathways, e.g. ubiquitination, synthesis and degradation of ketone bodies.

- Goal: enrichment of brain eQTL amongst top signals from the recent AGP GWAS.

- Data: AGP GWAS (4 SNP lists at $p < 0.001$), the SNP and CNV Annotation Database (SCAN) and genome-wide expression datasets in brain.

- eQTL mapping: Cerebellar (GSE35974) and parietal cortex (GSE35977) cis and trans eQTL were generated from 153 individuals of European ancestry. ComBat and Surrogate Variable Analysis were used to adjust for batch and both known and unknown covariate effects. Imputed genotype dosage data were analyzed for association with expression using PLINK. $P < 0.0001$ for cis-eSNP, and Bonferroni correction for trans-eSNP (23K probes).

- Enrichment analysis: among the four SNP list from ASP GWAS, count the number of significant eQTL. To test its significance, randomly sample the same number of SNPs as the GWAS SNP lists (matching MAF), and count significant eQTL.

- Significant enrichment of parietal ($P < 0.004$) and cerebellar ($P < 0.003$) eQTL, but not LCL eQTL ($P = 0.502$) among the top signals from the most broadly inclusive dataset of spectrum diagnosis including all ancestries. About 60 overlapped SNPs in parietal and cerebellar eQTL. SNP statistics in enrichment analysis:

  - A total number of 539 brain eQTL found in the primary AGP GWAS top signals (256 independent eSNPs).
  - Of the 214 SNPs that target only one gene, 124 (58%) act in cis and 88 (42%) act in trans.
  - 140 genes were uniquely implicated as eQTL targets. 62 (44%) were cis implicated and 78 (56%) implicated in trans. Only 18 (13%) of the 140 genes were also targeted by eQTL found in LCLs, and only 10 genes (7%) were found in both cerebellum and parietal tissues. In all pairwise comparisons of tissues, the overlap was not statistically significant.

- Implicated genes:

  - SLC25A12 was multiply-implicated by a unique set of 31 SNPs (in LD), all in cis (GWAS-p about 2E-5 to 1E-3).
  - PANX1 was targeted by nine cis eQTL and implicated in multiple tissues.

  - PANX2 was targeted by three trans eQTL, and the gene modulates the timing of neuroprogenitor commitment to a neuronal lineage in the hippocampus.
  - Comparison with autism DEX genes: four of the 140 brain eQTL target genes overlapped with the 1,153 differentially expressed genes identified by Voineagu et al, including SLC25A12 and PANX2.

- Lessons:
  - Trans-eQTL overlap with GWAS associations may be significant (more than half in this dataset), and the genes implicated may be highly plausible: e.g. for PANX2, both DEX in ASD vs. controls, and supported by other functional evidence.
  - eQTL overlap between multiple brain regions, and between tissues: may be small.

Targeted allelic expression profiling in human islets identifies cis-regulatory effects for multiple variants identified by type 2 diabetes genome-wide association studies. [Locke & Harries, Diabetes. 2014]

- Allele expression imbalance (AEI) measurement: amplification and measurement of mature (i.e. spliced) mRNA species and normalisation of allelic expression using genomic DNA from the same individual.
  - Genotyping of the genomic DNA: Sanger sequencing or TagMan genotyping assay to confirm that the lead SNP was heterozygous.
  - cDNA: treat cells with DNase to digest DNA, then reverse transcribed. PCR amplification.
  - Genomic DNA samples should show a 1:1 allelic ratio and thus any departure from 0 illustrates unequal amplification of alleles which must be corrected for.
  - Mean average allelic expression measurements: determined from two independent cDNAs reverse transcribed and amplified on different days.
  - Paired two-tailed T-tests, comparing genomic DNA and cDNA values from the same donor, were used to determine statistical significance for allelic expression.

- Validation of the robust measurement of the AEI:
  - Correlation between allelic expression measurements determined from independent cDNAs reverse transcribed and amplified on different days
  - Correlation between allelic expression measurements calculated from SNPs in high linkage disequilibrium with each other and residing within the same gene

- T2D candidate SNPs and genes: 65 lead SNPs from GWAS, 1525 proxy SNPs (r2¿0.8, CEU, 1000 Genomes Phase 1) were found. 45/1590 (2.8%) map to exons of 23 human RefSeq genes.

- AEI results: For 18 of the 23 genes, the TaqMan SNP assay can be designed to map entirely to exonic sequence. After filerting (too few heterozygous samples, too low gene expression level), allelic expression could be determined for 14 genes in samples from 36 white, non-diabetic donors. 7/14 genes show AEI (Figure 1) - results of cDNA show significant depature from 0 (genomic DNA for the purpose of normalization). And five genes are validated with another exonic proxy SNP. The AEI values of the five genes are small: < 1.25.

- Discussion:
  - The study is limited to genes with exonic SNPs, while cis-eQTL can study any candidate genes in a locus. Future studies may consider using intronic SNPs to measure allelic expression.
  - Comparison with targeted approach vs. RNA-seq: using RNA-seq, on average, each exonic SNP has coverage about 25. To detect an AEI $=1.25$, > 500 mapped reads (> 20 samples) would be needed.

7.4.2 Multi-omics QTL

Metabolic and transcriptional profiling of liver metabolism [Ferrara & Attie, PG, 2008]:

- Motivation: reconstruction the causal networks among transcripts and metabolites.
- Methods:
  - Data: 60 F2 mouse, genotyped (293 markers), liver gene expression is profiled, and concentration of 67 liver metabolites are determined through MS/MS, including 15 AAs and urea cycle intermediates, 45 acyl-carnitines and 7 organic acids (TCA and related intermediates).
  - QTL identification: both eQTL and mQTL, use interval mapping method.
  - Network inference: construct glx (glutamine) network, first select 250 transcripts most correlated with glx level and in the category "metabolism". Then test causality of locus, metabolite and transcript and generate the network [Chaibub, Genetics, 2008].
- Results:
  - Metabolites of common functional group are highly correlated, e.g. among all AAs, most pairs have CC > 0.5.
  - Correlation among metabolites and transcripts: metabolites generally correlated with related processes. E.g. 15 AAs correlated with transcripts in protein metabolism, glycolysis, TCA cycle and lipid metabolism.
  - Glx network: predict that Glx causally control expression of genes Agxt, Arg1 and Pck1. In experiment, adding 10 mM glx to culture cells, and found expression change of a number of transcripts, including the model predicted ones.

7.4.3 Systems Genetics of Model Organisms

Systems genetics of complex traits in Drosophila [Ayroles & Mackay, NG, 2009]:

- Motivation: popoulation variation of gene expression traits, and phenotypes; genetic basis of complex traits through association with gene expression.
- Methods:
  - Data: 40 fruit fly lines from natural popoulation, each is assayed with 18,800 transcripts (25 flys/sex/line, 3-5d young), and 6 traits: resistance to starvation stress, time to recover from a chill-induced coma, life span, a startle-induced locomotor response, mating speed and competitive fitness.
  - Analysis of expression variation: ANOVA to partition variation in expression betwen sexs, among lines, and the sex $\times$ line interactions.
  - Transcriptional modules: genes with correlated expression across population, first use ANOVA to identify the line terms, then cluster genes (graph-based method).
  - Transcripts associated with quantiative traits: let Y be transcript level, S be the influence of sex, and T be the influence of trait, then regression: $Y = \mu + S + T + S \times T + \epsilon$, to identify the genes significantly associated with phenotypic variations.
  - Modules in trait-associated genes: the goal is to find modules that are correlated (in addition to the common effect of the trait on all genes). Thus use residual terms after regression with traits to define correlation.
- Results:

- Genetic architecture of expression: nearly 80% transcripts are expressed in adult flies, about 90% of these have sex-biased expression. Two-thirds of the expressed trasncripts also show line variations. X-chromosomes is a hospitable spot for female-biased but not male-biased genes (mutations in X chromosomes tend to have large deleterious effect on males).
- Correlation of transcripts: the genome as whole is highly correlated at the transcript level. Cluster genes into 241 modules, the biggest two coorespond to sex-biased genes. Genes within a module tend to show similar tissue-specific expression (FlyAtlas), involved in the same pathway, or enriched with TFBSs.
- Association with phenotypes: a small amount of correlation among traits (trade-offs: e.g. resistence to starvation have longer life spans but reduced competitive fitness). Identify hundreds of genes for each trait and generally more than half are verified (mutation by P-element, and assess the trait).
- Transcriptional modules associated with traits: e.g. fitness trait: immune response, visual perception, function of nevous system, chemosensation and sex-specific transcripts.

Long-distance phenotypic logic chain enable precise inference of uncorrelated traits [Xionglei He, 2019]

- Yeast data: 815 haploid segregants, and 405 quantitative traits, such as morphology during each phase of cell cycle. Estimation of heritability of traits: h^2 median 0.15 and H^2 median 0.42.
- Uncorrelated traits can predict exemplar traits: e.g. Nuclear brightness in whole cell, 7 uncorrelated traits can predict well. Test these on all traits, find that the prediction performance using only uncorrelated traits can reach high levels, $\approx H^2$. In contrast, using SNPs or gene expression does much more poorly (Figure 2c,f).
- Explanation of the observation by latent dimensions: use autoencoder to learn 20 latent variables (linear transfer function). Then show that many traits are explained by these 20 latent variables: explain 85% variation. Example: nuclear brightness Y is explained by 5 latent variables. The seven uncorrelated traits that predict Y include 5 that are also explained by these 5 variables. Two other variables are needed to "cancel out" extra dimensions.
- Geometric view: Figure 3. suppose we have three variables, η, α, β. It's possible that η is in the hyperplane defined by α and β, but η is orthogonal to both. From latent variable perspective, it means that they all depend on the same set of latent variables (and 0 for other latent variables), thus can predict each other.
- Analysis [personal notes]: is it possible that Y is a linear function of other random variables, but Y and each of them marginally uncorrelated? Let $Y = \sum_j X_j \beta_j$, then for any i, we have

$$\mathrm{Cov}\left(X_i, Y\right) = \sum_j \beta_j \mathrm{Cov}\left(X_i, X_j\right) \tag{7.76}$$

  If we define the pairwise covariance matrix of X_i, X_j as: $\Sigma = \left[\mathrm{Cov}\left(X_i, X_j\right)\right]_{ij}$ and β is the vector of β_js, then the condition can be written in matrix form as:

$$\Sigma\beta = 0 \tag{7.77}$$

  When Σ is not full ranked, which would be the case if there are linear dependency of X_j's, then there are infinitely many solutions with $\beta \neq 0$.
- Analysis: probabilistic PCA perspective. Suppose z is latent random variable, with $z \sim N(0, I)$, and $x_1 = W_1 z$ and $x_2 = W_2 z$, where W_1 and W_2 are row vectors of the loading matrix. Then we have:

$$\mathrm{Cov}\left(x_1, x_2\right) = \mathrm{E}\left(W_1 z (W_2 z)^T\right) - \mathrm{E}\left(W_1 z\right)\mathrm{E}\left(W_2 z\right)^T = W_1 \mathrm{E}\left(z z^T\right) W_2^T - W_1 \mathrm{E}\left(z\right)\mathrm{E}\left(z\right)^T W_2^T = W_1 W_2^T \tag{7.78}$$

It is possible that this term is 0. If we have a relatively large number of latent variables, and their effects on observed variables are generally independent, then we may have this term often close to 0. However, even given marginally independent random variables, we can still predict one from others. Suppose Y can be expressed as $y = \beta z$, and x's can also be expressed as linear functions of z as in PPCA: $x = Wz$. Then we can solve $z = W^{-1}x$, where W is loading matrix. We then have: $y = \beta W^{-1}x$.

  - Remark: this analysis is based on PPCA. It is also possible to do the analysis with classical PCA.

7.5 Gene Regulation in Complex Traits

Experimental studies of disease loci and non-coding sequences [personal notes]

- Reference: [Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants, NG, 2014], [Obesity-associated variants within FTO form long-range functional connections with IRX3, Nature, 2014], [Leveraging cross-species transcription factor binding site patterns: From diabetes risk loci to disease mechanisms, Cell, 2014], [TCF7L2 loci from T2D GWAS, Marcola Nobrega lab, Dec, 2015]
- Transcriptional activity of non-coding sequences and variants:
  - To show the non-coding sequences and variants are functional: control, wt. sequence and variant in luciferase assay, show w.t. sequences can drive expression, but it is abolished by mutation [Islet paper, Figure 6C].
  - Allele-specific expression or eQTL: number of risk alleles associated with expression of target genes. Alternatively, ASE in heterozygous individuals. [FTO paper, Figure 2B]
- TF binding and regulation of enhancers:
  - EMSA to study protein binding of non-coding sequences. [Islet paper, Figure 6B].
  - To study the role of TF in regulation: knockdown of TF by RNAi [Islet paper, Figure 3C]
- In vivo expression pattern of enhancers:
  - In vivo transgenic reporter assays: zebrafish or mouse, establish the tissue-specific regulatory activities of enhancers. [FTO paper, Figure 1B]
- Enhancer-promoter interactions:
  - 3C or 4C: profile interactions between a test loci/enhancer with all regions within a certain distance (e.g. 1Mb). [FTO paper, Figure 1A].
- Tissues and endophenotypes affected by regulatory sequences: identify the relevant tissues and endophenotypes, experimentally study how they are affected by regulatory sequences or tissue-specific overexpression/deletion of the genes.
  - Relevant tissues: activity pattern of enhancers or expression pattern of genes, eQTL.
  - Endophenotype study: e.g. number and size of adipocytes in T2D, a target of TCF7L2 gene [TCF7L2 work from Marcelo lab]. Use tissue-specific expression.
  - Remark: the ability to detect relevant tissues (and stages) is a main advantage of studying non-coding sequences underlying diseases - similar to tissue-specific perturbation of gene expression.

Approaches for establishing the function of regulatory genetic variants involved in disease [Knight, Genom Med, 2014]

- Examples of regulatory variants important for diseases:
  - Variant in 3'UTR: Crohns-disease-associated variant in the 3 UTR of IRGM that alters binding by the microRNA mir-196, enhancing mRNA transcript stability and altering the efficacy of autophagy.
  - Alternative splicing: a variant of TNFRSF1A associated with multiple sclerosis, which encodes a novel form of TNFR1 that can block tumor necrosis factor
- Regulatory epigenomic data (Table 1: comprehensive resources)
  - FANTOM5: high-resolution context-specific maps of TSS and their usage for 432 different primary cell types, 135 tissues and 241 cell lines, enabling promoter-level characterization of gene expression. Also map active enhancers by eRNA.
  - UCSC Genome Browser: Variant Annotation Integrator.
  - Ensembl genome browser includes the Ensembl Variant Effect Predictor.
  - RegulomeDB: functional regions (ENCODE), eQTL, prediction of motif disruption.
  - Combined Annotation-Dependent Depletion method: 63 types of genomic annotation to establish deleteriousness for SNVs and small insertion-deletions [Kircher & Shendure, NG, 2014]
  - Conservation: 8.2% of the human genome is subject to negative selection and is likely to be functional [Rands & Lunter, PLG, 2014]
  - SNPnexus: coding SNPs, Regulatory elements (conserved TFBS in human/mice/rat, vista enhancers, microRNA sites), conserved sites from PhastCons and GERP
  - GWAS3D: [Junwen Wang lab] cell-type specific annotation, TFBS scanning, histone modifications, chromatin interactions
  - MAPPER2: TFBSs located in the upstream sequences of genes from the human, mouse and D.melanogaster genomes, combines TRANSFAC and JASPAR data with the search power of profile hidden Markov models (HMMs)
- Findings/insights from QTL studies:
  - Importance of trans-eQTL: may affect expression of other genes through transcriptional regulation or signaling (1) a cis-eQTL for the transcription factor KLF14: associated with T2D and HDL, act as a master trans regulator of adipose gene expression. (2) a cis-eQTL involving IFNB1: associated in trans with a downstream cytokine network, found in stimulated cells.
  - Importance of considering context/condition: eQTL analysis of the innate immune response transcriptome in monocytes defined associations involving canonical signaling pathways, key components of the inflammasome, downstream cytokines and receptors. Often disease-associated variants and were identified only in induced monocytes.
  - Coding SNVs: An estimated 15% of codons [Stergachis, Science, 2013] specify both amino acids and transcription factor binding sites.
- Methods for functionally studying regulatory variants
  - Allele-specific expression (ASE): early studies show that in addition to the small number of classical imprinted genes showing monoallelic expression, up to 15 to 20% of autosomal genes show heritable allele-specific differences. From [Lappalainen, Nature, 2013], LCLs from 462 individuals: almost all the identified ASE were driven by cis-regulatory variants rather than genotype-independent allele-specific epigenetic effects.
  - Allele-specific TF binding: use ChIP-seq, applied to heterozygous cell lines or individuals can provide direct evidence of relative occupancy by allele.

  - Chromatin interactions, in particular, Capture-C: cross-linking of chromatin interactions followed by capture of hundreds of target regions.
  - Genome editing by CRISPR/Cas9: (1) eQTL of SLFN5, used CRISPR-Cas9 to demonstrate loss of inducibility by IFN-beta on conversion from the heterozygous to homozygous state in a human embryonic kidney cell line. (2) T2D-associated variant in PPARG2: replaced the endogenous risk allele in a human pre-adipocyte cell strain with the non-risk allele and showed increased expression of the transcript.

A map of open chromatin in human pancreatic islets [Gaulton & Ferrer, NG, 2010]

- Motivation: use open chromatin (CRE map) to study T2D genetics.
- Open chromatin in pancreatic islets defined by FAIRE-seq.
- Overlap with T2D loci: of 350 SNPs in strong LD with a reported T2D locus, 38 SNPs at 10 loci overlapped islet FAIRE regions. Verification of rs7903146 in TCF7L2: test the hypothesis that rs7903146 variant changes chromatin state (accessibility) and the enhancer activity (hence likely a causal variant of T2D):
  - Allele imbalance at chromatin state: identify 9 individuals with heterozygous rs7903146, then find the allelic imbalance (T:C ratio) in the open chromatin (FAIRE-isolated DNA).
  - Enhancer activity: luciferase reporter assay in islet cells, compare the two enhancers differing in rs7903146. Only one allele shows enhancer activity. No difference in a control cell type.
- Lesson: use allele-specific chromatin state (similar to ASHM) to demonstrate that a SNP has an effect on chromatin and regulatory activity.

Systematic Localization of Common Disease-Associated Variation in Regulatory DNA [Maurano & Stamatoyannopoulos, Science, 2012]:

- DHS mapping: 349 cell and tissue samples, including 85 cell types studied under the ENCODE Project and 264 samples studied under the Roadmap Epigenomics Program. About 100 from cell lines, primary tissues, hematopoietic and differentiated cells, etc; and 233 diverse fetal tissue samples across days 60 to 160 after conception. Average of 200K DHS per cell type, and a total of 3M DHS (42.2% of the genome).
- Distribution of 5K non-coding SNPs in GWAS catalog (more than 600 studies/traits):
  - Hypothesis: non-coding SNPs involved in complex diseases are enriched in DHS.
  - Fully 76.6% of all noncoding GWAS SNPs either lie within a DHS (57.1%, 2931 SNPs) or are in complete linkage disequilibrium (LD) with SNPs in a nearby DHS (19.5%, 999 SNPs).
- Cell- and developmental stage specificity:
  - Hypothesis: DHS containing the non-coding SNPs are tissue/developmental stage-specific.
  - Examples: for a number of diseases, the variants are located in DHS specific to disease-relevant tissues. Ex. for cardiovascular diseases, find a SNP in the DHS specific to fetal heart.
  - Importance of early gestational exposures: 88.1% (2583) lie within DHSs active in fetal cells and tissues.
  - Enrichment or depletion of replicated disease-specific GWAS variants in fetal-stage DHSs: the most enriched traits are, menarche, cardiovascular disease, and body mass index (gestational exposures or growth trajectory known to play a role). Relative depletion in fetal DHSs of aging-related diseases, cancer, and inflammatory disorders with presumed (postnatal) environmental triggers.

- Identification of target genes of DHS:
  - Method: correlate DNase sensitivity of DHS with DNase I sensitivity patterns at cis-linked promoters. Use $r > 0.7$ as a cutoff. Verification via paired-end tag sequencing (ChIA-PET).
  - Identified 419 DHS-gene pairs (within 500kb). Fully 40.8% of correlated DHS-gene pairs span > 250 kb and 79% represent pairings with distant promoters versus those of the nearest gene.
  - Examples of target genes that play plausible roles in diseases: a SNP associated with platelet count, located in a DHS that physically interacts with the 222-kb distant promoter of JAK2.
- Alternation of TFBSs within DHS:
  - Hypothesis: many non-coding variants alter TFBS binding and chromatin states (thus affecting phenotypes).
  - Define TFBSs: scanning for known motif models at a stringency of $P < 1e - 4$.
  - Of GWAS SNPs in DHSs, 93.2% (2874) overlap a transcription factor recognition sequence.
  - Detection of altered chromatin structure in heterozygous SNPs (an imbalance in the fraction of reads obtained from each allele). Nearly 40% of GWAS variants in similarly sequenced DHSs would be expected to show allelic imbalance.
  - In general, TFBS binding removes nucleosome, making the site more accessible (sensitive). Figure 2C: the first and second examples (TFBS alleles have higher counts).
- Disease-associated variants in TFBSs of specific disease classes:
  - Hypothesis: for a TF related to a disease class, its TFBSs may be enriched in the disease-related DHS (i.e. the DHSs containing the noncoding variants associated with this disease).
  - Disease-related TFs: using known TFs (e.g. HNF1a for MODY), and interacting TFs of the known TFs (from Ingenuity)
  - Autoimmune diseases: IRF9 and 15 interacting TFs, 24.4% (64/262) of GWAS SNPs within DHSs of immune cells and associated with autoimmune disease alter one or more of the 15 transcription factor motifs from the IRF9-centered network.
  - Multiple related diseases often share the same TFs: TF-disease networks (Figure 4) for autoimmune diseases and cancer. Also six neuropsychiatric disorders with 23 transcription factors.
- Identification of disease-related cell types:
  - Hypothesis: in a disease-related cell type, the DHS will be enriched with disease-associated variants.
  - For Crohn, MS and QRS duration (heart): selective enrichment of SNPs associated with GWAS in relevant cell types. Furthermore, even at relatively low GWAS threshold (e.g. $p < 0.01$), the enrichment can be detected (higher enrichment at higher p threshold).
- Lessons:
  - Target genes of non-coding disease SNPs: some experimental data can reveal the long-range interactions between non-coding sequences and promoters (ChIA-PET, 5C). Also the correlation between epigenetic states of enhancers and those of promoters or gene expression.
  - TFs involved in complex diseases: enrichment of TFBSs in disease-associated non-coding SNPs
  - Disease-related tissues: enrichment of disease-associated SNPs in tissue-specific DHS.
- Remark:

- Target genes of DHS: through correlation of DHS and promoter sensitivity pattern across tissues. Problem: DHS may be tissue-specific (open in specific tissues) while promoters may be open in any tissue a gene may be expressed.
- Role of TFs in disease: through enrichment of TFBSs in disease-associated non-coding SNPs (located within DHS). Stringent p-value threshold.
- Common diseases, common networks: some TFs are found to be associated with multiple diseases. However, these diseases are known to share SNPs: suppose we have a single SNP common to multiple disorders, and this SNP is located in a DHS containing BS of some TF, then this TF would appear to be associated with mulitple disorders.

Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants [Pasquali & Ferrer, NG, 2014]

- Hypothesis: dysregulation of TF target enhancers in pancretic islet cells increases the T2D risk.
- Overview: ChIP-seq of pancretic islet TFs, and histone markers. Enrichment of GWAS-T2D SNPs in these CREs.
- Regulatory map of pancretic islets:
  - ChIP-seq of five TFs
  - Open chromatin states: FAIRE-seq and ChIP-seq of H2A.Z
  - Key histone modifications: H3K4me3, H3K4me1, H3K27ac and CTCF-binding. Clustering of these four states on open chromatin reveals five classes: promoters (C1), poised or inactive enhancers (C2), active enhancers (C3), CTCF-bound sites (C4) and the rest (C5).
- Pattern of cis-regulatory map:
  - Auto- and cross-regulatory relationship between five TFs.
  - Targets of the five TFs often overlap.
  - 92% of TFBS mapped to open chromatin, and they bind to distinct chromatin states.
- TF binding enhancers drive islet-specific transcription
  - TF-bound C3 sites (active enhancers) are associated with islet-specific gene expression, while non-C3 sites do not show association. Shown by luciferase assay in beta cells vs. fibroblast cells (Figure 3A).
  - Experimental validation of some of the TF-bound enhancers (TFBS clusters): knockdown of TF by RNAi reduces gene expression (Figure 3C). Also more likely to interact with promoters in 4C (Figure 3E).
  - Additional TF motifs were found in the enhancers.
- Sequence variation in islet enhancers is associated with T2D:
  - Enrichment of T2D and glycemia GWAS hits in clustered C3 sites, but not in orphan C3 sites.
  - Relaxing the threshold of GWAS p-values: fold enrichment of the sites in clustered C3 sites.
  - A catelog of causal cis-regulatory variants of T2D: intersect GWAS hits with clustered C3 sites and DNA-binding motif analysis.
  - Results of specific GWAS loci, T2D risk variant at ZFAND3: the variant disrupt TF binding, (EMSA experiment, Figure 6B) and change gene expression in vitro (luciferase in mouse MIN6 beta cells, Figure 6C).

- **Lessons**:
  - TF binding and chromatin states: TF binding are often associated with open chromatin. Functional targets of TFs are often active enhancers (in open chromatin).
  - Active enhancers, but not other states (promoters or inactive enhancers) are associated with tissue-specific expression.
- Remark: ideally, we will study how DNA variations correspond to change of epigenetic states/expression.

Leveraging Cross-Species Transcription Factor Binding Site Patterns: From Diabetes Risk Loci to Disease Mechanisms [Claussnitzer & Laumen, Cell, 2014]

- Motivation: the causal SNPs should be in CREs, which has the feature of cluster of conserved TFBSs.
- PMCA algorithm:
  - Start with a disease-related SNP, first find all non-coding SNPs (ncSNP) in high LD. Then find the region surrounding an nc SNP (60 bp) from the human genome, and the orthologous region in 15 vertebrate species.
  - Define the conserved TFBSs, conserved TFBS modules (oc-occuring TFBSs at the same order). The motifs are from Genomatix library (800 human TFs).
  - Scoring of the ncSNP (and the surround region): a significant enrichment of phylogenetically conserved TFBS modules. Basically counting the number of conserved TFBS, then evaluate its significance by randomization.
- Clusters of sites of homoebox TFs is a distinct feature of T2D loci: define positional bias as TFBS clustering relative to transcription start sites. From the T2D loci related to sequences (8 loci), find positional bias of Homeobox TFs, such as CART and PDX1.
- Remark: many functional TFBSs are not conserved, and this may significantly limit the power of PMCA.
- Lessons:
  - Cross-species conservation can be used to define CREs that are likely causal loci of diseases. The method may have high precision, even if the sensitivity is low.
  - A disease may be associated certain distinct TFs (or TF families), and the signature of the TFs can be found near the GWAS loci.

Genetic and epigenetic fine mapping of causal autoimmune disease variants [Farh and Bernstein, Nature, 2015]

- Fine-mapping method PICS: first show that χ^2 statistics decay with r^2 to causal SNP. Next, causal SNP may not be the strongest SNP due to statistical flucutations, infer the probability that a SNP is a lead SNP given that another SNP is causal, using permutation.
- Results of fine-mapping: (1) GWAS catelog index SNPs: **only 5% represent a causal SNP**. (2) Most GWAS signals cannot be resolved to a signal causal SNP.
- Causal SNPs and immune enhancers: PICS SNPs are enriched in stimuls-dependent enhancers. About 60% of SNPs are in immune-cell enhancers, many of which are induced by immune activation.
- Cell-type signatures of complex diseases: causal SNPs of GWAS loci from 21 AIDs, and enrichment in enhancers from different tissues. Results better than the expression pattern of genes targeted by coding GWAS hits. Almost all AIDs: CD4 T cells. Some such as SLE preferentially map to B cells. T1D: also enriched in pancreatic islet. UC: gastrointestinal tract elements.

- Causal SNPs and disruption of gene regulation: (1) TFBS (from ENCODE) enrichment: many TFs, top ones are NF-kB, IRF4 (Figure 5b). (2) 800 high confidence SNPs (average PIP 0.3): only 7% change motifs of over-represented TFs, this compares to about 1% by control SNPs. 13% change motifs of any TFs, similar to background. 26% residing within 100 bp of motifs.

The osteoarthritis and height GDF5 locus yields its secrets [NG, 2017]

- GDF5 locus found in GWAS of height. The gene GDF5 is a member of BMP family and a good candidate.
- Fine-mapping: transgenic mice carrying the upstream and downstream sequence of GDF5. Several experiments to identify the enhancer GROW1 (2.9kb): expression of reporter, rescue GDF5-null phenotype, deletion of the locus leads to phenotype.
- Positive selection of the GROW1 haplotype: long haplotype blocks in high LD, a signature of positive selection. Comparison of haplotypes between Euroasian vs. African: the allele is more common in Eurasian than African, also found in Nean and Denosovan.
- Unresolved issue: could be three independent SNPs in the locus.

7.6 Network Genetics

Problems of network genetics:

- **Reconstruction** of molecular networks.
- **Organizational principles** of networks: biological networks are far from random, and may possess certain features including master regulators, convergent nodes, coordinated modules, and so on. Identify and rationale such features. Note that the definition of organization here refers not only to topology, but also the influences.
- **Implications on phenotype**: what does the network say about the effect of changing one node on phenotypes?

Principles of network analysis of complex traits:

- **Association of gene modules with traits**: this is based on the principles of modularity and guilt-by-association. It can be done in multiple ways. Gene modules can be identified via clustering of expression patterns across genetic perturbations.
  - Dysregulation of modules in diseases: correlation of gene expression with traits; or change of co-expression pattern in diseases.
  - Enrichment of disease-related genes in modules.
- **Causal gene network** from eQTL: it is possible to construct causal networks from eQTL. This serves as a reference to interpret the link between genes and phenotypes.
- **Understanding the function from genes to phenotypes**: We can study the causal chain of events from genes to phenotypes. Suppose we represent the state of system as x (network state), and its disease state y, our goal is to learn the function $y = f(x)$. Without the network, we can only do linear modeling (sometimes allow interaction terms), but it is very limited. With the network, the function is more structured and greatly constrained.
- **Master regulators**: the structure of biological networks is that some nodes play unusually large roles than other nodes. Identifying such "master regulators/players" can be helpful.

- **Importance of biological contexts**: for many genes, their effect on phenotypes depend on biological contexts, both environmental (eg. smoking) and cellular conditions.
- **Multi-layered structure of networks**: the networks are organized into multiple layers, e.g. SNVs influence expression of RNA, which affects translation and protein levels. The enzyme levels control metabolic states. At even higher level, some part of networks are cell-specific, some are active across-tissues.
  Remark: simlarity to multi-layered ANN for deep learning, we have some features at each layer that determine the next layer.

Using networks to study candidate genes of diseases [personal notes]:

- Intra-connectivity of candidate genes: average degree, clustering coefficient of a subnetwork.
- Inter-connectivity with known genes:
  - Number of edges between two groups of genes.
  - Average distance between two groups of genes.
  - Some kind of weighted distance: a new gene highly connected to seed genes will receive higher score. Thus we can compare the scores of candidate genes with control genes.
- Subnetworks: in addition to statistical testing, often it is informative to visualize the subnetworks formed by the candidate and known genes. The subnetworks tend to be functionally coherent (enrichment analysis).
- Choice of networks: PPI network, gene co-expresison network. Ideally, use networks that are more relevant, e.g. synaptic PPI network for psychiatric diseases.
- Importance of control: to assess the significance, we need some control genes. This needs to be done carefully, e.g. candidate genes all have certain properties (e.g. brain expression), and the control genes should match these properties.

Network overview: Dan Nicolae's talk at Complex trait journal club

- Resources: Ingenuity, GeneWays (from text mining), STRING.
- Clustering coefficient: defined on a node, how related the neighbors of this node are.
- Network analysis for establishing the functional relationship among genes: when comparing two sets of genes and want to show one set is on average closer, need to be careful about the possible confounders.
  - Example: autism study, compare de novo genes vs. random genes. De novo genes are generally longer, and might have more PPIs.
- Building co-expression network (WGCNA): let s_{ij} be correlation between two nodes, then the edge weight $a_{ij} = s_{ij}^{\beta}$.

Network medicine: a network-based approach to human disease [Barabasi & Loscalzo, NRG, 2011]:

- Motivation: only about 10% of human genes have a known disease association. What are properties of disease networks, in terms of how disease genes are distributed?
- Gene topology: hub genes tend to be essential, more conserved, and pleiotropic (deletion leads to more phenotypes). However, not all essential genes are disease genes in humans. Mutations in genes that are essential in early development lead to spontaneous abortions. Essential genes that are not associated with disease show a strong tendency to be associated with hubs and are expressed in multiple tissues.

- Modularity: proteins involved in the same disease have an increased tendency to interact with each other. For example, one group observed 290 physical interactions between the products of genes associated with the same disorder, representing a tenfold increase relative to random expectation. Thus, each disease can be linked to a well-defined neighbourhood of the interactome, often referred to as a "disease module".

- The network parsimony principle: causal pathways are the shortest paths connecting the known disease components.

- Methods that use the modularity principle for predicting disease genes:

  - Linkage methods: the direct interaction partners of a disease protein are likely to be associated with the same disease phenotype. Example: severe combined immunodeficiency syndrome, the set of genes within the locus whose products interacted with a known disease protein were shown to be tenfold enriched in true disease-causing genes.

  - Disease module-based methods: constructing the interactome in the tissue and cell line of interest and identifying a subnetwork, or disease module, that contains most of the disease-associated genes. Variants of this methodology have been applied to a wide range of diseases and pathophenotypes, including several different types of cancer, neurological diseases, cardiovascular diseases, systemic inflammation, obesity, asthma, T2D, etc.

  - Diffusion-based methods: Proteins that interact with several disease proteins will gain a high probabilistic weight, as will those that may not directly interact with any disease proteins but are in close network proximity to them.

- Shared components hypothesis: Diseases that share disease-associated cellular components (genes, proteins, metabolites or microRNAs) show phenotypic similarity and comorbidity. In the human disease network (HDN) - two diseases are connected if sharing at least one gene, 867 of 1,284 diseases with an associated gene are connected to at least one other disease, and 516 of them belong to a single disease cluster. For example, cancers form a tightly interconnected and easily detectable cluster, which is held together by a small group of genes that are associated with multiple cancers.

- Same gene, different diseases: many disease pairs that share genes do not show significant comorbidity. One explanation is that different mutations in the same gene can have different effects on the gene product, and therefore different pathological consequences91 that are organ and context dependent. Such 'edgetic' alleles affect a specific subset of links in the interactome.

- Metabolic diseases: links that are induced by shared metabolic pathways are expected to be more relevant than are links based on shared genes. For example, purine metabolism consists of 62 reactions associated with 33 diseases. Comorbidity analysis confirms the functional relevance of metabolic coupling: disease pairs that are linked in the MDN have a 1.8-fold increased comorbidity compared to disease pairs that are not linked metabolically.

- Application: therapies that involve multiple targets, avoid side effects. Can one systematically identify multiple drug targets that have an optimal impact on the disease phenotype?

NEW: Network-Enabled Wisdom in Biology, Medicine, and Health Care [Schadt et al. Sci Transl Med, 2012]

- Using gene networks (co-expression, causal) for association analysis: basic strategy is to find modules/pathways linked to diseases.

  - Procedure: (1) identify modules; (2) eSNPs associating the expression of genes are extracted from eQTL data; (3) enrichment/pathway association analysis of the eSNPs.

  - Other ideas, e.g. (1) use correlation of expression data and phenotypes (e.g. diff. expression) to identify candidate modules; (2) use genotypes to distinguish causal and correlative/responsive modules.

- Gene-environment interactions and the importance of context: "increasing evidence suggests that most genetic risk variants are dependent on particular environmental contexts to effect risks for CCD". Use networks to identify the combined risk-enrichment for groups of functionally associated genes. Examples:
  - The effects of most DNA sequence variants linked to type 2 diabetes in Caucasians are manifested only in patients with a body mass index above 26.
  - DNA sequence variations linked to certain types of high blood pressure exert their negative effects only in the context of low physical activity.
- Decomposition of context: macroenvironment (smoking, exercise, toxicity, etc.) and microenvironment (cellular conditions). The context affects how DNA variants increase the disease risk. To understand DNA risk variants and context: need to use omics data to create molecular networks.
- Multi-layered molecular networks: with omics data of RNA, proteins, metabolites, possible to construct multi-layered networks, e.g. DNA $\rightarrow$ RNA $\rightarrow$ protein $\rightarrow$ metabolite $\rightarrow$ phenotype.
- Constructing tissue/organ specific networks are essential: they are likely more important in later phases of disease development, when pathological changes are spreading across the borders of individual organs.
- Implications for health-care:
  - DNA profiles: risk assessment early in life. Prevention through controlling environment, etc.
  - Activity profiles: biomarkers (CRP, liver-enzymes, and future ones) from OMICS data can provide a snapshot of network states, thus reflecting any signs of molecular pathology (such as tumor growth, atherosclerosis, inflammation, or immune responses).
  - Treatment: the DNA profiles and network states can guide the selection of treatment strategies. And the markers can monitor the disease progression/recovery.

Leveraging models of cell regulation and GWAS data in integrative network-based association studies [Califano & Schadt, NG, 2012]

- Pathway-wide association study (PWAS) strategy: the difficulty is pathways are often poorly characterized.
- Integrative network-based association studies (INAS): favored, the simultaneous reconstruction of context-specific gene regulatory networks.
- Why INAS approach? Linear pathways are a poor representation.
  - Complexity of GRNs: a TF may be regulate hundreds of genes, and combinatorial regulation of multiple TFs. Each TF is further regulated by many signal transduction proteins.
  - Cellular context and higher-level interactions among cells.
- Reverse enginerring of networks:
  - High throughput experiments of PPI, kinase, TF regulation, etc. An initial, albeit sparse, snapshot of regulatory networks, especially when integrated with other types of data that can help contextualize individual interactions.
  - Computational predictions: often rely on perturbations (internal or external) or temporal data, and often integrate multiple sources.

  - Functional interactions (GIs): another layer.

- Canonical pathway analysis: most successful examples in immunological pathways, NF-kappa B.
- Dysregulation of subnetworks in diseases:
  - Dysregulated gene set analysis via subnetworks (DEGAS) and interactome dysregulation enrichment analysis (IDEA). Examples: Parkinson's disease and B-cell lymphoma.
  - Search for subnetworks enriched in linkage or association of diseases. Or use networks (complexes) to increase the power of detecting epistasis.
- Molecular phenotypes/eQTL: construct networks from eQTL data.
- GRN analysis: identifying master regulators. Examples: human high-grade glioma, and normal physiological formation of germinal centers.
- Diseasome approaches: exploits previous biological knowledge of gene similarities and dissimilarities across diseases. Ex. G allele of the rs2076530 in BTNL2 is more frequent among individuals with T1D and RA than in healthy controls, whereas the A allele was more frequent in SLE than controls.
- Lessons:
  - A broad view of networks: physical interactions, functional, causal, correlated.
  - Importance of context-dependency of networks and dynamic nature. Diseases often involve multiple tissues.
  - Analysis of networks: master regulators, integrative analysis of multiple ones across diseases

The human disease network [Goh & Barbasi, PNAS, 2007]:

- Background:
  - Diseases are often caused by mutations of related genes: e.g. Zellweger syndrome is caused by mutations in any of at least 11 genes associated with peroxisome biogenesis.
  - Mutation of one gene can give rise to multiple disorders: e.g. mutations in TP53 have been linked to 11 cancer-related disorders.
- Data: gene-disease bipartite graph constructed from OMIM, which include data of both monogenetic disease and complex traits. 1,284 disorders and 1,777 disease genes.
- Human disease networks: link two diseases if they share a gene. The network is clustered into many modules of major disease classes. Cancer is a large cluster; metabolic diseases have low genetic heterogeneity and are not very connected; neurological disorders show high locus heteogeneity and also represent the most connected disease classes.
- Disease gene network (DGN): two genes are linked if they are associated with the same disease. several disease genes (e.g., TP53, PAX6) are involved in as many as 10 disorders, representing major hubs in the network.
- Disease associated gene modules: genes associated with the same disease tend to: interact with each other through PPI (10-fold enrichment); tend to be expressed in the same tissues; co-expressed; and in the same GO categories.

Diverse types of genetic variation converge on functional gene networks involved in schizophrenia [Gilman & Vitkup, Nature Neuro, 2012]

- Motivation: given a diverse set of genetic data (putative risk genes), find the subnetworks that are enriched with risk genes.

- Background network construction: first obtain gene features, which measures how related two genes are. These features include PPI, common annotations (a single feature based on GO), phylogenetic profile, and so on. To map these features to response (whether they have the same phenotype), use Naive Bayes. To train the model, use a known gene-disease network. The results are expressed as LR scores for two genes as edge weights (same phenotype vs. different phenotype).
- Genetic data: 159 de novo SNVs, 712 from de novo CNVs and 173 from 14 GWAS loci from SCZ data.
- Search for clusters: in a test cluster, each gene is from one of the three sources, and for CNVs and GWAS region, only one gene is allowed in a cluster.
  - Cluster scores: the sum of LR scores of all edges in the cluster. Obviously, the scores are not normalized by cluster sizes.
  - Cluster significance: random selection of clusters (matching the node connectivity), the obtain p-values.
- Biological processes and validation of the clusters: top cluster about 30 genes, $p < 0.001$. Cluster I: enriched in neurodevelopmental processes such as axon guidance, neuron projection development. Also evidence that the genes are expressed early in development.
- Relation to other disorders: the Cluster genes are significantly more connected to autism gene sets. Also find evidence that the genes of autism and SCZ may affect dendritic growth differently.
- Remark:
  - Search of clusters is challenging: esp. due to the constraints (one gene per CNV/GWAS region).
  - No validation of individual genes, e.g. the genes picked by the program from GWAS loci.

Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease [Bin Zhang, Cell, 2013]

- Background:
  - Progress in LOAD research is fundamentally limited by our reliance on mouse models of severe familial/early-onset Alzheimer's disease.
  - Genetics of LOAD: APOE accounts for 30% of genetic variance. GWAS implicates immunity (CLU, CR1, CD33, EPHA1, MS4A4A/MS4A6A), lipid processing (APOE, ABCA7), and endocytosis (PICALM, BIN1, CD2AP) as important causal biological processes. More recently, low-frequency missense variants in APP and TREM2 were found to confer strong protection or elevated risk of LOAD.
- Brain expression data: 1,647 autopsied tissues from dorsolateral prefrontal cortex (PFC), visual cortex (VC), and cerebellum (CB) in 549 brains of 376 LOAD patients and 173 nondemented healthy controls. Expression analysis: robust linear regression adjusting for covariates: age and sex, postmortem interval (PMI) in hours, and sample pH and RNA integrity number.
- Analysis pipeline: (1) Network construction: co-expression network; use genetic markers to anchor the causal network. (2) Module ranking: differential connectivity in LOAD and normal networks, enrichment of brain eSNPs.
- Module differential connectivity (MDC): comparison of the networks constructed from LOAD and normal people. Define MDC as the ratio of the average connectivity for any pair of module-sharing genes in LOAD compared to that of the same genes in the nondemented state. Detect a number of modules with high change of MDC (gain or loss of connectivity, GOC or LOC). This cannot be captured by the traditional diff. expression analysis.

- Functional categories of GOC or LOC modules: eg. the immune module shows the statistically most significant functional enrichment of all modules.
- Association of modules with LOAD pathophysiology: A covariance matrix of the average expression correlation (absolution value) between 49 modules (using PCA) and 25 LOAD-related traits is constructed. The immune/microglia showed correlation to the greatest number of LOAD-related neuropathology traits.
- Brain eSNP mapping and modules enriched with brain eSNPs: 10K eSNPs identified at FDR 10%. And cis-eSNPs are used to construct causal networks.
- The immune/microglia module is highlighted: (1) significant differential connectivity in LOAD; (2) the most significant enrichment of functional categories; (3) the highest degree of gene-expression correlation to LOAD neuropathology; (4) the PFC version of the module was highly enriched for brain eSNPs.
- Ranking causal regulators: based on (1) regulatory strength: the number of downstream nodes of the immune module in the causal network; (2) diff. expression in LOAD patients. TYROBP scored the highest.
- Convergent molecular pathway: TREM2 is known to associate and signal via TYROBP. The subnetwork also contains previous top GWAS risk loci including MS4A4A, MS4A6A, and CD33. TYROPB positions in several microglia activation-signaling cascades. Hypothesis: TYROBP may be associated with neuronal pruning activity of the complement system that may be reawakened in LOAD via amyloid-beta and tau aggregates.
- Experimental validation of TYROBP: in mouse microglia cells, perturbation of TYROBP, and measure diff. expression. The DE genes are highly enriched with the immune module. Also observe a strong correlation between pathway distance to TYROBP in the network and the fraction of DE genes.
- Lessons:
  - Change of connectivity in conditions: could capture more information than the standard DE. Example: suppose a pathway (complement system) is activated in pathological condition, a number of genes will be simultaneously up-regulated, creating significant coexpression.
  - Multiple measures to implicate disease-causing modules: change of expression/connectivity in disease; correlation with disease markers; functional coherence of the module. Could also add: enrichment of potential disease-related genes.
  - Identifying key regulators of disease module: by using the causal network.

Widespread macromocular interaction perturbations in human genetic disorders [Sahni & Vidal, Cell, 2015]

- Data: about 3,000 mutations in 1,140 genes from HGMD, Mendelian diseases. For control, use common variants from 1000 GP.
- Different types of mutatations when describing the change of interaction profiles: quasi-wild type (mutation does not change), quasi-nul (abolish most of the interactions), edgetic (specific edges).
- Impact of mutations on protein stability: if a mutation reduces stability, the mutated protein will require more interactions with chaperons and quality control factors (QCF), so using these interactions to profile the stability change. Overall, relatively small effect on protein stability, but a small fraction of mutations lead to increased interactions (reduced stability) - about 28% with at least one of the seven QCFs.

  - For those with increased binding, the mutations are often located in the core of the protein (than surface or disordered regions).

- Impact of mutations on PPI: use Y2H to profile PPI. Out of 1,300 PPIs, found 521 perturbed interactions. Among all mutations, about 2/3 change interactions (edgetic or quasi-null), and about half of them are edgetic.
  - Quasi-null mutations are often unstable or mis-folded and have lower expression level.
  - Comparing with controls (common variants): only 8% change interaction profile.
  - Polyphen and conservation analysis could distinguish changes vs. nonchanges (quasi-wildtype), but not between edgetic and quasi-null.
  - Edgetic mutations are enriched in structurally exposed residues compared to quasi-null mutations.
- Impact of mutations on protein-DNA interactions (PDI): enhanced Y1H on 70 TFs and 152 enhancers. 38% of mutations are quasi-null, 43% edgetic and 19% quasi-WT. Including both gain and loss of PDIs: likely due to loss of specificity from mutations. Quasi-null mutations are highly enriched in DBD regions.
- Remark/Questions:
  - The claim that using change of PPI can achieve precision of 96% and sensitivity 61% to distingush disease and non-disease alleles is misleading: the data are highly enriched with disease mutation.
  - The causal link from PPI changes to phenotypic consequences: not studied much in the paper.
  - Can we **extrapolate** from the current findings for unseen mutations?

Chapter 8

Genetics of Complex Traits

8.1 Overview of Genetics of Complex Traits

Mutations and functional changes of genes: including the change in the regulatory sequences (abolish expression of a gene is similar to the loss of activity of the protein product) [Human Molecular Genetics, 3rd Ed., Chapter 14]

- Two types of functional consequences of mutations:
  - Loss of function mutations: often heterogeneous, as many different mutations can lead to loss of function.
  - Gain of function mutations: generally rare. Mutational homogeneity is a indicator of gain of function.
- Loss of function mutations:
  - Small deletions and insertions, nonsense-mutations including premature termination, splicing mutations, frameshift, point mutation of essential AAs, etc.
  - Haploinsufficiency: some genes are dosage sensitive, thus 50% reduction of activity may cause abnormal phenotype. Dosage-sensitive genes are few, including: genes needed in large quantiy (e.g. elastin), genes that interact in certain proportions (e.g. in signaling/metabolic switches, in protein assembly such as α and β globins).
  - Dominant negative effect: a non-functional copy of the gene may interfere with the function of the normal copy. Ex. collagens - the assembly is disrupted by the non-functional copy; bHLH-ZIP family of TFs that bind in dimers - mutants can sequester functioning molecules into inactive dimers.
- Gain of function mutations: usually cause dominant phenotypes.
  - Chromosome rearrangements that induce exon shuffling/fussion: common in cancer.
  - Overexpression: e.g. by transposition of a gene to a active chromatin environment.
  - Mutations that make a protein insensitive to regulation: this would cause constitutively active proteins, e.g. constitutively active receptors in GPCR signaling.
  - Protein aggregation: often from unstable expanding repeats (which may leads to other effects such as reduction of transcription of nearby genes), notably CAG repeats that encode poly-Q tracts. Can be also due to chance events of protein misfolding. Particularly important in neurodegenerative diseases.

The pathways from genes to diseases: [Human Molecular Genetics, 3rd Ed., Chapter 16]

- Multiple paths to deficiency of one protein: may not be the protein itself, could be any step that leads to the production of this protein. Ex. immunodecificiency (lack of immmuoglobins) can be caused by failures in: immunoglobin gene processing, B-cell maturation, or other steps in the development of the immune system.
- Protein complexes and pathways: mutations of different members may lead to similar phenotypes. Ex. collagen of skin: mutations in COL1A1, COL1A2, type XI collagen all have similar phenotypes.
- Mutations in different members of a gene family: could cause related or overlapping syndromes. Ex. fibroblast growth factor receptors (FGFRs): 1-4, the mutants may affect the balance of different forms.
- Mutations often affect only a subset of the tissuss in which the gene is expressed: e.g. HD gene (Huntington disease) is widely expressed, but the mutation mainly affects brain.
- Dependence on genetic background and on environment: this would be expected for instance in dosage-sensitive genes. Ex. a common variant, R402Q, in tyrosinase gene (a key enzyme of melanoctyes) is normal, but can lead to ocular albinism in the presence of a mutation in MITF (a gene involved in differentiation of melanoctyes).

Genetic architecture of complex traits/diseases:

- Genetic architecture is ultimately determined by the complex, often opposing effects of selection, population history, migration and mutation rates.
- Diversity of genetic architecture: different between, for example, autims and intelligence, height. E.g. pervasive epistatic effects have been documented in autoimmune conditions, morphology and susceptibility to cancer, but fear-related phenotypes consists almost entirely of multiple small additive effects.
- A fundamental problem is the risk model: what genetic factors cause the disease? This may include for example, a single-hit risk model vs. multi-hit risk model (a major mutation causes the disease in one individual or multiple ones)? What role does genetic background (epistatic effect) play?
- Implication: understanding why genetic architecture differs for different traits could help when choosing the correct tools to find the underlying genes and deciding whether to look for common or rare variants.
- Reference: [Eichler & Nadeau, Nature, 2010]

Missing heritability of complex traits/diseases: [Eichler & Nadeau, Nature, 2010]

- Rare variants with possibly large effects: rare variants are individually rare, but collectively frequent,
- Many common variants with small effects: with many tests performed, there is a high false-negative rate in GWAS, as true associations are hidden in the fog of random associations.
- Genetic interactions (non-additive effects) such as dominance and epistasis: epistatis is found to be important in a number of diseases.
- Epigenetic effect: example, parent-of-origin effect (genetic interaction with parent). Epigenetic effects beyond imprinting that are sequence-independent and that might be environmentally induced but can be transmitted for one or more generations.
- Structural variations (deletions, duplications and inversions) of genomes: (1) individually rare but collectively common variations in the human population. An estimated 8% of the general population carry a large ($>$ 500 kb) deletion or duplication that occurs at an allele frequency of $< 0.05\%$. (2) Copy number variations (CNV): some genes are highly variable among individuals, are enriched in genes associated with drug detoxification, immunity and environmental interaction.

- Genotype-environment ($G \times E$) interaction: we are largely unable to identify what the relevant environments are.

Estimating genetic architecture:

- Motivation: even if we cannot detect all individual loci of a trait, we may still be able to estimate some key parameters of the genetic architecture such as the number of causal genes, the effect size distribution, the explained variance, etc.
- Strategy: model or simulate how the key parameters, π - the fraction of causal genes, and γ - effect size distribution, influence the key aspects of data, such as: the enrichment of low p-value genes relative to null distribution, the enrichment of de novo mutation events in probands vs. siblings, and other informative patterns: e.g. how often a gene has multiple de novo mutation events.
- Explained variance: suppose we want to estimate how much phenotypic variance can be explained by a type of genetic variance. For linear model, this is determined by the effect size and frequency of causal alleles; for binary trait, one can use liability threshold model.
- Reference: [Sanders, De novo mutations revealed by whole-exome sequencing are strongly associated with autism, Nature, 2012], [Iossifov, De Novo Gene Disruptions in Children on the Autistic Spectrum, Neuron, 2012]

Major diseases and traits studied by GWAS (`http://genome.gov/gwastudies/`):

- Metabolic diseases: obesity, T2D
- Heart and circulation system diseases: coronary heart disease, hypertension, sudden cardiac arrest, myocardial infarction
- Neurological and behavior diseases: Alzheimer's disease, Parkson's disease, schizophrenia, bipolar disorder, ADHD, autism
- Auto-immune diseases: T1D, multiple sclerosis, asthma, Crohn's disease, rheumatoid arthritis
- Cancer: breast cancer, prostate cancer, lung cancer, leukemia, colorectal cancer, ovarian cancer
- Other diseases: kidney stones, gallstones, AMD, osteoporosis
- Responses to treatment: radiation response, aromatase inhibitors (breast cancer), Warfarin (anticoagulant), anti-depressant, antipsychotic therapy, hepatitis C treatment, interferon beta therapy, treatment for acute lymphoblastic leukemia
- Phenotypic traits: height, body mass index, waist circumference, waist-hip ratio, longevity, eye color
- Behavior traits: alcohol dependence, smoking behavior, Nicotine dependence, personality dimensions, heroine addiction, cognitive ability
- Heart and blood function: pulmonary function, bone marrow density, blood pressure
- Metabolites: LDL, HDL, triglyceride, serum urate, HDL cholesterol, LDL cholesterol, serum calcium, vitamin D, insulin-related traits
- Proteins: tau (cerebrospinal fluid), adiponectin, glycated hemoglobin, immunoglobins, plasma level of liver enzymes
- Ceullar properties: telemere length, erythrocyte phenotypes

How do we understand genetics of complex traits from many small effects? [Personal notes]

- **Principle**: a complex trait may involve a set of interacting cell types, and disruption of normal functioning, maintenance of normal cellular states and differentiation, and cell interactions lead to phenotypic changes.
- Example: known risk genes of psychiatric diseases can affect:
  - Cellular functions: SCN2A and SHANK, synaptic functions.
  - Cellular states: mTOR pathway affects cell proliferation (and brain size). Many regulatory genes may change cell states: FMRP, MeCP2, CHD8.
  - Cell-cell interactions: a complement gene changes how immune cells prune neurons (synaptic pruning).
- **Principle**: each of these steps, especially changes of cellular states (e.g. differentiation or proliferation), involves coordinated actions of many genes. Major genes and pathways exist to control these processes. These may be why complex diseases genes are regulatory genes.
- Implications on genetic architecture: many cell types may be involved, and for each cell type, major genes regulating cellular functions, states and interactions (key processes) probably have large effects. Other genes with smaller effects:
  - The major genes/pathways are probably influenced by many other genes.
  - Specific genes in the key processes can affect disease risk. Ex. apoptosis: depends on regulatory genes, but also specific enzymes that act on apoptosis, e.g. enzymes involved in creating ROS.
- Do risk genes of a trait converge on some genes? Possible, but not necessarily true. Ex. SCN2A is important for synaptic building, and may be regulated by multiple genes and pathways, so we can imagine that some risk genes converge to SCN2A level. On the other hand, it is possible that a trait may depend on the balance of two cell types (e.g. white vs brown fat cells), and a number of genes are involved in cell fate determination without a single key node.
- Implications on methodology: we will need to understand: (1) How cell functions, states and interactions are disrupted in diseases? (2) What are major genes and pathways regulate these processes? For (1), we can use pathway analysis of GWAS, comparison of transcriptome (ideally single-cell) between patients and controls. For (2), genetic networks (trans-eQTL), GRN via TFs or RBPs.

Role of dynamics/stimulation-response CREs in human traits [personal notes]:

- Ref: The impact of proinflammatory cytokines on the -cell regulatory landscape provides insights into the genetics of type 1 diabetes [NG, 2019].
- Understand T1D genetics: Induced regulatory elements (IREs) are those are those that change upon relevant stimulation (inflammatory cytokines), in contrast to stable regulatory elements (SREs). IREs are found to be enriched in T1D variants while SREs in T2D. Possible model: SREs are involved in maintaining beta cell identify, housekeeping functions, etc., while IREs are involved in response to inflammatory cytokines. Inappropriate IRE responses (e.g. by genetic variations) may lead to over-reaction to cytokines, e.g. apoptosis, which leads to T1D.
- General model: the CREs of a cell have different functions, involved in different aspects of cellular functions: e.g. housekeeping, cellular differentiation and maturation, responses to cell function related stimuli. Variations of activities of different groups of CREs can lead to different phenotypes.
- Neurons and psychiatric diseases: some CREs may be involved in neuron differentiation, some in neuronal signal response (neuron maturation). So variations in the former may lead to defects in neuro-development, e.g. not enough differentiated cells; while variations in the latter may lead to inappropriate response to electric stimuli. These may correspond to different phenotypes: e.g. autism vs. epilepsy.

8.1.1 Genetic Stuides of Common Diseases

Breast cancer: [Human Molecular Genetics, Chapter 15]

- BRCA1 and BRCA2: identification in linkage studies on near-Mendelian families.
- BRCA1 account for 80-90% of families with both breast and ovarian cancer, but a much smaller proprotion of families with breast cancer alone. Male breast cancer was seen mainly in BRCA2 families.
- Risk: in affected families, BRCA1 mutation had an 85-90% chance of developing breast cancer; however the risk is much lower (36%) in broad families.

Alzheimer disease (AD): [Human Molecular Genetics, Chapter 15]

- Early onset: genes identified in near-Mendelian families - APP, presenilin-1/2. Mutations of these genes accout for 10% of early onset AD.
- Late onset: different genes, ApoE (E4 variant) is a strong candidate, accounting for 50% of suspectability of late-onset AD. ApoE is a class of apolipoprotein (lipid-binding protein), is essential for the normal catabolism of triglyceride-rich lipoprotein constituents. Apolipoprotein E enhances proteolytic breakdown of this peptide, and E4 variant is inefficient at catalyzing these reactions.

Type 1 diabetes (T1D):

- Disease: from autoimmune destruction of pancreatic β-cells, usually affecting young people.
- Loci: HLA-DQB and INS (insulin gene, the mutations affect expression level of insulin). Together explain 50% suspectability of T1D.

Type 2 diabetes (T2D):

- Disease: combination of impaired insulin secretion and decrease end-organ responsiveness. Known risk factors: age, obesity.
- Loci: calpain-10 (CAPN10), however, the evidence is limited.

WTCCC studies [WTCCC, Nature, 2007]:

- Study design:
  - 2,000 cases each for seven common diseases and 3,000 common controls (chosen from blood donors and from persons born in 1958) in British population. The cases are defined by clinical phenotypes, however, misclassification may happen in controls, as phenotypes of the controls are not collected, and some may develop diseases in the future. The misclassification rate is believed to be low. 153 individuals with non-Caucasian ancestry were excluded.
  - Affymetrix chip with about 500k SNPs, out of which 400k SNPs have MAFs $> 1\%$ (MAF: minor allele frequency).
- Comparison of groups for population structure: the general issue is to compare two groups using SNP frequencies. For each SNP, test the difference of its allele frequencies in two group using some form of χ^2-test (thus one test statistic for each SNP). Then the spectrum of the statistic can be compared with the null distribution (assume there is no difference between the two groups) using a Quantile-quantile plot.
- Association test:

- Power assessment: for case and control groups, simulate SNP data using population genetic models. Note that the MAFs of the causal SNPs in the case group is estimated from the effect sizes (using Bayes theorem). In this study, estimated to be (only for common variants, MAF $> 5\%$, much lower for rare variants): 43% for alleles with relative risk of 1.3, and 80% for a relative risk of 1.5, for a P-value threshold $5 \cdot 10^{-7}$.
- Trend test and genotyp test for any SNP: association of genotypes (three values per SNP) and phenotypes (case or control). A P value is computed for each SNP.
- Bayesian analysis: let M_0 be the null model (no association), and M_1 be the model of true association (additive effect for two alleles) and M_2 be a model where three genotypes could have different effects. A SNP is assessed by the posterior odds ratio:

$$\frac{P(M_1|D)}{P(M_0|D)} = \frac{P(M_1)}{P(M_0)} \frac{P(D|M_1)}{P(D|M_0)} \tag{8.1}$$

  The prior odds ratio can incorporate prior knowledge such as the distribution of SNPs (e.g. higher for nonsynonymous SNPs). A simple estimate would be, e.g. 10 causal SNPs in a region of 1Mb, and the prior odds ratio would be 10^{-5}. The Bayes factor is computed from the logistic regression: the log-odds is equal to μ for M_0, and $\mu + \gamma Z_i$ for M_1, where Z_i is the genotype of the i-th individual (0,1 or 2). The coefficient γ is the increase in log-odds of disease for every copy of the allele encoded as 1. The Bayes factor integrates over parameters θ under each model. For both models, a prior of $N(0, 0.2)$ is used for γ, and a prior of $N(0, 1)$ is used for μ.
- Significance threshold: use posterior odds ratio to select a P-value threshold, instead of multiple hypothesis correction (which is used for a single "global" hypothesis). The rationale is: whether a SNP is causal to the phenotype is one hypothesis to be tested. In the posterior odds ratio equation, the average $P(D|M_1)$ is the power of the study (given the hypothesis is true, M_1, what is the probability that we can detect it), assuming to be 0.5; and the average $P(D|M_0)$ is the P value threshold. Setting posterior odds ratio at $10:1$ leads to the P-value threshold of $5 \cdot 10^{-7}$.

- Imputation: use HapMap data of SNPs (about six times of the SNPs in this study) to impute the untyped SNPs with a HMM. Verification of the imputation shows 98.4% accuracy. Each imputed SNP will also be tested for association (with a somewhat stronger criteria).

- Results summary:
  - Population structure: 153 individuals were excluded by comparing their SNP data with those from HapMap. The SNP data from different geographic regions were compared, and 13 regions were shown strong geographical variations. Some of them are probably due to natural selection, e.g. lactase, and for infectious diseases.
  - Summary of findings: at $P = 5 \cdot 10^{-7}$, 1 in bipolar disorder (BD), 1 in coronrary artery disease (CAD), 9 in Crohn's disease (CD), 3 in rheumatoid arthritis (RA), 7 in type 1 diabetes (T1D) and 3 in type 2 diabetes (T2D). Also 58 loci with slightly less significant P values. 12 of the 25 strong signals represent known findings, and most of the rest are confirmed by replication studies.

- Individual diseases:
  - BD: PALB2 - stability of nuclear structures including chromatin; NDUFAB1 - mitochondrial respiratory chain; DCTN5 - intraceullar transport known to interact with DISC1 (known to be involved in BD).
  - CAD: CDKN2A/B - CDK inhibitors; CARD15 - caspase recruitment domain family member 15; IL23R; ATG16L1 - autophagy-related; MST1 - macrophage stimulating 1.
  - T2D: PPARG - peroxisomal proliferative activated receptor gamma; KCNJ11 - beta-cell KATP channel; TCF7L2 - transcription factor 7-like 2.

- Discussion:
  - The effects of the population structure in this study is small.
  - A common set of controls: not necessary to have controls that map the socio-demogrpahic variables of every disease case group.
  - The effects of most SNPs are small.

Large-scale whole-genome sequencing of the Icelandic population, [Gudbjartsson & Stefansson, NG, 2015]

- Motivation: infer from WGS data the *pattern of selection* in human genome (what might influence the strength of selection) and *association analysis*.
- Data: 2,636 Icelander WGS with median depth of 20x. Found about 20M SNPs and 1.5M indels (about 7%). Among indels, about twice are deletions, likely due to the challenge of calling insertions.
- Measuring selection: fraction of rare variants (FRV) at DAF $< 0.5\%$, and variant density. Higher FRV and lower density suggests negative selection. Note that these measures are affected by coverage, so limit the analysis only to regions with sufficient coverage (at least 15x).
- Pattern of genetic variants and selection:
  - A definite of indels of length 3.
  - LoF variants: 149 per individual, only 1.4 were seen in 1 or 2 out of all samples. Only 1 in 12 individuals have homozygous LoF variant with MAF $< 2\%$.
  - Selection at different functional classes of variants: higher selection at LoF, then missense mutations, then synonyous and UTR, then intronic and intergeic.
  - Selection at OMIM genes: strongest negative selection, in particular on variants that act through a dominant mode of inheritance.
  - Correlation of intra-species selection with mammalian conservation (GERP): when GERP score is positive (purifying selection), positive correlation with FRV, and negative correlation with density, as expected.
  - Selection at GO categories: high-density GOs related to communication of cells with environments, e.g. sensotry, defense. Low-density GOs: basic cellular processes.
  - Selection in different non-coding elements: higher selection in active promoters, strong enhancers, then weak enhancers, insulators, then heterochromatin.
  - *Ultra-sensitive regions: high FRV and high density*, unexpectedly.
- Imputation: accuracy is improved by long-range phasing of 104K Icelanders.
- Association: Test additive model (16M) and recessive model (10M) with regression model, Correction by Bonferroni (27M tests). One result: MYL4 (myosin light chain) and early-onset atrial fibrillation.
- Remark/lessons:
  - A key problem is to infer and understand pattern of selection in the genome. Simple measures: FRV and density. The methodological challenge is the local variation of mutation rates (thus exepected metrics of selection). *What if we use SFS and corresponding tests? Are they robust to mutation rate difference*?
  - Imputation at isolated population: possible, and can greatly increase the power of studies.

8.1.2 De Novo Mutations

De novo mutations in human genetic disease [Veltman & Brunner, NRG, 2012]:

- Basics of de novo mutations:
  - Rate of de novo mutations per genome: CNV - 0.03, indel - 3, SNV - 74. Per exome rate is about 1 per person.
  - Origin of de novo muations: an open question is whether these mutations occur mainly in the germline, during embryogenesis or somatically.
  - De novo mutations are more deleterious, on average, than inherited variation because they have been subjected to less stringent evolutionary selection.
- De novo mutations and their contribution to genetic diseases:
  - Contribution of de novo mutations: higher for neurodevelopmental diseases, e.g. ID and autism, because (1) large mutational target; (2) higher selection of affected individuals, thus inherited variants play a smaller role. Therefore, de novo mutations, although individually rare, may capture a significant part of the heritability for complex genetic diseases.
  - Recessive inherited alleles are unlikely to explain most cases of these diseases, as the empirical sibling recurrence is much less than 25% for intellectual disability, autism and schizophrenia.
- De novo mutations in rare sporadic genetic disease:
  - De novo CNVs: A well-known example is Down syndrome, which is caused by a de novo trisomy of chromosome 21. Recurrent de novo microdeletions and microduplications are now recognized as a common cause of clinically defined malformation syndromes.
  - De novo SNVs: application in Mendelian diseases. The first example: de novo SNV in SETBP1 in Schinzel-Giedion syndrome.
- De novo CNVs in common genetic diseases: De novo CNVs larger than 100 kb are infrequent in the normal population, occurring in approximately one in 50 individuals. By contrast, these large de novo CNVs occur in approximately 10% of all patients with sporadic ID, ASD and schizophrenia.
- De novo SNVs in cmmone genetic diseases:
  - Sporadic ID: nine non-synonymous de novo SNVs were validated in seven out of the ten individuals. In two patients, de novo nonsense mutations were found in known ID-associated genes: RAB39B and SYNGAP1. Four of the remaining mutations are likely to be detrimental to protein function, and they affect plausible candidate genes.
  - ASD: Note that rate between 0.63 and 1.00 per unaffected sibling: Different exome enrichment assays, sequencing methods and data-filtering steps.
- Challenges of interpreting de novo mutations: the detection of de novo mutations is no longer the limiting factor, the next pressing question is how to interpret any given de novo change in the context of a patient's phenotype.
  - Recurrently mutated genes: only 18 genes were found to be mutated de novo multiple times, a number that is not significantly different from simulated control data [Neale, 2012].
  - Adding functional information of genes to facilitate the interpretation of de novo mutations: gene group test, network methods, etc.
  - Mutant types: however, Neale et al did not observe a difference in the PolyPhen-2 classification of 101 non-synonymous de novo mutations identified in patients with an ASD as compared to random simulations.

- Future directions:
  - Development of algorithms that are targeted to the analysis of de novo mutations in the context of exome studies, possibly incorporating several of the elements outlined in Fig. 2 (gene function, impact of mutations on protein function, correlation with phenotypes).
  - Prenatal screening: the interpretation of these rare de novo events will be extremely challenging in a prenatal setting, especially because many of these mutations have variable penetrance and no phenotype information is available to guide interpretation.

Incorporating Functional Information in Tests of Excess De Novo Mutational Load (fitDNM) [Jiang & Allen, AJHG, 2015]

- Model: the number of de novo mutations at each locus follow multinomial distribution (of sample size n_l for locus l), with rate λ_{lk} for mutational type k (three mutations). To determine the rate, $P(X_l = k|A = 1)$, where X_l is the mutation at l and A is the disease status, use Bayes theorem to relate it to $P(A = 1|X_l = k)$. Suppose each mutation is associated with an indicator D, whether it disrupts the protein function, and the probability of disruption is ρ_{lk}. We have:

$$\lambda_{lk} \approx [1 + (\gamma - 1)\rho_{lk}]\pi_{lk} \tag{8.2}$$

  where π_{lk} is the mutation rate and γ the RR given that $D = 1$. The model assumes that ρ_{lk} are known, given by PolyPhen for missense mutations and 1 for LoF.

- Test: likelihood for all the sites, then the score test.

- Simulation studies: randomly sample π_{lk} and ρ_{lk} for all positions in three chosen genes, then sample mutations, and sample the disease status, which depends on the total number of disruping mutations in a gene (summing over D's for all positions) through a logistic regression model. The parameters of logistic regression are chosen to mimic full penetrance (0.8 to 0.95 if a gene has one LoF).

- Simulation results: fitDNM has twice of power of TADA-denovo. In an alternative scenario (similar to TADA): only LoF and highly probably missense mutations are causal, the power is similar to TADA.

- Results: similar to TADA, effectively one new gene when combining all four diseases (TRIO - 5 de novo mis3).

- Remark: the assumption that ρ_{lk} is known and given by PPH2 is highly problematic. In fact, most PPH2 mis3 mutations have probabilities close to 1 (> 0.95), but mis3 mutations are nowhere close to LoF mutations in terms of damaging effects. So the method places a much larger emphasis on missense mutations, which is not supported by data. In fact, even though the simulation finds that it doubles the power, in practice, it gave essentially identical results to TADA.

8.1.3 Somatic Mutations

Somatic mutation in single human neurons tracks developmental and transcriptional history [Lodato and Walsh, Science, 2015]

- Background: Ultra-deep sequencing for identifying somatic mutations. For most sites, germline genotypes are homozygous. Somatic mutations are present in a certain fraction of cells, detectable in sequencing reads. Problems of ultra-deep sequencing: (1) power is limited, if somatic mutations present in a small fraction of cells; (2) cannot know which mutations occur together.

- Fluroescence activated nuclear sorting (FANS), followed by WGS (40x) in 3 inividuals. Total of 36 neurons.

- Mutation rate and pattern: about 1,500 SNVs per neuron. Not correlate with replication timing, but show signatures of TCR (strand bias and expression). For neurons: replication-dependent mutations are less important.
- Polyclonal derivation of brain: 5 major clades. A cell is more related to cardiomycytes than 75% of its neighbors.

Somatic mutations in diseases and in development [Chris Walsh, HG seminar, 2017]

- Part I. Single gene mutations: change brain structure. Ex. microcephany (ASPM, COH1).
- Hemimegalencephaly (HME): enlargement of half of the cerbral cortex, caused by somatic mutations. Patient: removing half brain to remove seisure. Still largely normal. Mechanism: AKT3 mosaicism is responsible, a small percent of heterozygotes. ATK3: ser/thr kinase in mTOR pathway.
- A variety of mosaic overgrowth syndromes: mTOR gain-of-function mutations. Timing of mTOR mutation determines malformation type.
- Additionally, 5-15% of DNMs are somatic mosaic mutations.
- Part II. Single cell sequencing of neurons: nuclear sorting.
- Retrotransposons copy and paste via an RNA intermediate.
- Average <1 somatic LINE jump per neuronal genome. Most have no LINE.
- Somatic SNVs are extremely common in single neurons. About 1400 somatic SNV per neuronal genome, or 15-30 per exome. Most SNVs are found in just one neuron.
- Shared mutations can be used to order cells into lineage. BLocks/nested mutations of blocks. Found 4 major clades.
- Clonal mixing: common in human brain (not in mouse). Some neurons are more related to heart than neighbors.
- Single neuron SNVs: show signatures of transcriptional damage (ss DNA damage during transcription).
- DNA damage: increase somatic mutations, role in ageing.
- ASPM: microcephaly. ASPM dN/dS > 1 in human.
- Part III. Excess of recessive mutations in HARs (about 3000) in affected vs. unaffected.
- Mutations in a distant Cux1 enhancer, in a HAR. Cux1 is dosage-sensitive gene: modify neuron/synapse density.
- 2-3 SNVs per cell division. Higher rate of microsatelites. Mutation pattern is different in cancer: not enrichment of late-replicating genome; and opposite to TCR.
- Q: Sign of positive selection in somatic mutations?
- Q: Clonal mixing: same pattern in different individuals? Why? Functional advantage? Patterns consistent with known neuronal migrations? Can we use this to infer neuronal migrations?
- Q: estimation of somatic mutation rates? Number of cell divisions from fertilized eggs to neurons? Consistent across tissues? Ex. fast replicating cells have higher mutation rates?

8.1.4 Mendelian Diseases

FSS [Ng & Shendure, Nature, 2009]:

- Data: exome sequences of 8 HapMap individuals and 4 unrelated FSS patients (FSS: dominant Mendelian disease). 78% of genes have $> 95\%$ coding bases covered. Sequencing is done by hybridization capture (with short-gun libraries).
- Pattern of genetic variations:
  - SNPs: on average, 17,272 cSNPs were called per individual, with 92% in dbSNP. Comparing with the reference human genome, each individual has about 10K (Yoruba) and 8,489 (non-Africans) cSNPs.
  - Indels: on average 166 coding indels per individual.
- Causal gene of FSS: two criteria for the test gene (1) at least one nonsynonymous cSNP, splice-site disruption or coding indel is observed in the gene; (2) the mutations are not in dbSNP, nor in the eight HapMap exomes. Only MYH3 meets the two criteria (the known causal gene).
- Remark:
  - SNP pattern: most SNPs were accumulated during human evolution, which took about 3 million years. In one generation, the number of cSNPs per genome is about 0.1. This gives the number of SNPs per individual: $(3 \cdot 10^6/30) \cdot 0.1 = 10^4$, where 30 is the human life span. Also note that: the SNP distribution should be continuous (i.e. a continuous spectrum from rare to common) because the SNPs were accumulated over time, and should be biased towards less frequent SNPs because human popuation grows faster in recent times.
  - Causal gene identification: genetic heterogeneity is taken into accout here (allowing different mutations in 4 unrelated individuals).

8.1.5 Mitochondria

Genetics: Mitochondrial DNA in evolution and disease [NG, 2016]

- Two mouse strain differ in mtDNA: transfer mtDNA from one strain to the nuclear DNA background of another strain. The two strains have profound difference in disease, longevity, T2D, etc.
- Consequence of mtDNA variation: (1) Bioenergentic role of mt: permit accommodation to new diets or adjustment to thermal stress and activity demands. (2) Affecting Nuclear gene expression: modulating the levels of high-energy molecules (ATP, alpha-KG, etc.) generated through mitochondrial metabolism, which drive the modification of cytoplasmic signalling proteins and also add molecular modifications to nuclear proteins.
- Normal mtDNAs can be present in the cell in different proportions, a state known as heteroplasmy. Ex. 3243G mutation present in low proportion, T2D or autism; at 5090%, neurological, heart and muscle problems; 100% childhood disease or death.

8.2 Genetic Architecture of Complex Traits

Why are the alleles that increase the risk of diseaes not eliminated by natural selection? [Nesse & Williams, Why we get sick]

- Mutation-selection: high mutation rate (e.g. in large proteins), or recent mutations.

- Balacing selection: beneficial in some special circumstances, e.g. G6PD deficiency in areas with maleria (deficiency in glucose metabolism, protect against malarai parasite: kill the cell when the parasites use oxygen in red blood cells). A special form of balacing selection is heterozygote advantage: e.g. sickel-cell disease gene - heterozygotes protect against malaria.
- Late-onset diseases: the diseases occur only later in life, thus very weak selection, e.g. Alzheimer disease.
- Changing environment: genes that are normal or beneficial in ancestral environment may cause disease in modern environment, e.g. myopia. This involves gene-environment interaction.
- Selfish genes: e.g. meiotic drive (genes compete to enter a sperm or egg, even at the cost of the carriers).

Synthetic associations [Dickson & Goldstein, PLoS Biol, 2010]:

- Hypothesis: rare variants may create significant associations in common variant SNPs in GWAs (called synthetic associations).
- Background: the vast majorty of GWAS associations have never been tracked to causal sites, even though many surrounding regions have been resequenced.
- Intuition:
  - In the genealogy, any varirant "higher up in the genealoty" that partitions the parts of genealogy containing more causal disease variants will be identified as disease-associated (Figure 1).
  - Alternatively, if at time of occurrence of the disease allele at the causal site, the disease allele is associated with some varirant of one common SNP, and no/few recombinations occur afterwords, then the common SNP will appear disease-associated.
- Methods: simulation of genealogy of about 2000 cases and controls, the baseline risk is $(0.01, 01)$ and the genotypic relative risk (GRR) at 2 to 6.
- Significant associations caused by rare variants: in 30% of cases, a significant association with a common SNP was detected ($P < 10^{-8}$). The association with the causual variants is almost always stronger than with the synthetic associations. Also, sigifnicant synthetic associations depend on the associtions that occur within a single gene genealoty (no recombinations).
- Effect of recombinations: increasing recombination rates may acutally increase synthetic associations.
- Distance between causal variants and significant SNPs: in simulation, the median distance between the causal and synthetic varirants is 5 Mb. Also in GWAS of sickle cell anemia, synthetic associations in about 2.5 Mb region. Thus, the true associations can travel across multiple LD blocks to create synthetic associations.

The mystery of missing heritability: Genetic interactions create phantom heritability [Zuk & Lander, PNAS, 2012; NIH Lecture, From the 'Genetic Code' to the 'Genetic Code']:

- Explaining heritability: Let h^2_{known} be the heritability explained by the known loci, and h^2_{all} be the (true but unknown) heritability of the trait, then we have:

$$\Pi_{explained} = \frac{h^2_{known}}{h^2_{all}} \tag{8.3}$$

  is the fraction explained by the known loci. The heritability estimated from data (e.g. from twin-studies) is often different from the true heritability, and is denoted as h^2_{pop}. When

$$h^2_{pop} > h^2_{all} \tag{8.4}$$

  we overestimate h^2_{all}, thus $\Pi_{explained}$ is underestimated.

- Known heritability: suppose we have a linear model involving p loci:

$$Y = \sum_j X_j \beta_j + \epsilon \tag{8.5}$$

  Taking the variance, we have the variance explained by the p loci is:

$$h^2_{known} = \sum_j 2p_j(1-p_j)\beta_j^2 \tag{8.6}$$

  where p_j is the allele frequency of the j-th SNP.

- Overestimation of h^2_{all} under ACE model: we have:

$$h^2_{pop}(ACE) = 2(r_{MZ} - r_{DZ}) \tag{8.7}$$

  When there are epistatic interactions, the covariance between relatives is bigger (see the section "Covariance between relatives" in Notes Biology-BG), and the effect is the largest with monozygotic twins, so $h^2_{pop}(ACE) > h^2_{all}$.

- Limiting pathway (LP) model: suppose the phenotype is a function of k pathways: s.t. it is the maximum (or minimum) of k pathway variables. Denote the model as $LP(k, h^2_{pathway}, c_R)$, where $h^2_{pathway}$ is the heritability of each of the k pathway variables, and c_R is the extend of shared environment between relatives. Our goal is to determine h^2_{all} and $h^2_{pop}(ACE)$ under this model.

  - True heritability: let Z_i be the i-th pathway variable (Gaussian) and Z be the phenotype variable, then $Z = \max Z_1, \cdots, Z_k$. The heritability (assuming $V_P = 1$) is equal to the sum of variance explained by each pathway variable. The variance explained by the additive model of one pathway is equal to $h^2_{pathway}$, however, only part of this explains the variance of Z, and the fraction is $\text{Cov}(Z_i, Z)$), which can be determined from maximum of k Gaussian RVs. So we have:

$$h^2_{all} = k\text{Cov}(Z_i, Z)h^2_{pathway} \tag{8.8}$$

  - Apparent heritability: this involves computing r_{MZ} and r_{DZ} under the LP model.

- Case study: a good amount of heritability is explained for some traits such as T1D. Even higher when smaller-effect SNPs are included.

- Difficulty of epistasis mapping: need very large sample size to detect epistasis.

- Conclusion:

  - The heritability estimation is based on additive models, which are obviously not true. Thus the true heritability is essentially unknown.
  - What matters is the biology, not the heritability explained.

Genetic interactions improve models of quantitative traits [Tyler & Carter, NG, 2017]

- Motivation: many model organism studies support the prevalence of epistasis, but they are not found to be important in human population studies, why?

- Yeast growth QTL study: many strains, growth phenotype. Show that the prediction of growth rate is greatly improved by epistasis. In particular, find a number of **genetic capacitors**, defined as key genetic loci that each masked the effects of many other loci. Depending on its genotype, the capacitor either locks the phenotype near the population mean or permits its interaction partners to influence the phenotype.

An Expanded View of Complex Traits: From Polygenic to Omnigenic, [Boyle & Pritchard, Cell, 2017]

- Distribution of GWAS signals across the genome: using height as an example, 3.8% of causal SNPs (if allowing LD, 60% of SNPs) based on ASH. This mean about 100K causal SNPs across the genome.
- Weak enrichment of GWAS signals in functional gene groups: e.g. 5-10 enrichment of immune genes in Crohn's disease and RA (still not explain the majority of signal). In contrast, in rare variant studies, genes tend to be functionally connected, e.g. ASD and SCZ.
- Omnigenic model: a small set of core genes. The cellular networks are highly interconnected (small-world hypothesis) s.t. regulatory variants in peripheral can affect the activity of core genes. Implications:
  - Even though peripheral genes have smaller effects, they outnumber core genes, so together explain most of heritability.
  - The cis-QTL are trans-acting QTL on core genes.
- Pleiotropy: if the omnigenic model is correct, we expect widespread pleiotropy, as a regulatory QTL can create small effects in different core genes at different traits (that share the same cell type).
- Evolutionary changes of complex traits: most adaptive changes within a species are polgenic adaptation. Propose that it is also true in cross-species comparison. Problem of how pleiotropy may share selection and adaptation.
- Remarks:
  - Small world model: does not support the omnigeic model, buffering, compensatory changes.
  - Alternative explanation of the data: many genes could affect cellular states: how fast they divide, how efficient they deal with stress, how quickly it responds to some stimuli, etc. These genes do not directly act on core genes (disease-specific processes).
  - Importance of trans-acting QTL: supported by Fehrmann, NG, 2012 and Sherlock.
  - Evolutionary implications: Different pictures in model organisms and in cross-species comparison. The impact of population size: how efficient deleterious variants are removed.

8.3 Metabolism and Metabolic Diseases

Review: genetic loci of plasma lipoproteins: [Hegele, NRG, 2009; Teslovich, Nature, 2010]

- Chylomicron and VLDL function: APOB; APOE - removing chylomicron and VLDL remnants through interaction with its receptor; LIPC - hepatic lipase, receptor-mediated lipoprotein uptake
- LDL function: LDL receptor (LDLR); LDL receptor accessory protein 1 (LDLRAP1) - a chaperone through the early phase of endocytosis; PCSK9 - convertase involved in intracellular receptor degradation; LRP1 and LRP4, members of the LDL receptor-related protein family
- HDL function: APOA1 - main protein component in HDL; APOA5 - component in HDL, important in in regulating the plasma triglyceride levels; ABCA1 - cholesteral efflux pump in peripheral cells, transporting cholesteral to HDL particles; LCAT - esterification of cholesterol; cholesteryl ester transfer protein (CETP) - collecting triglycerides from VLDL and LDL to exchange for cholesterol in HDL; SCARB1, a HDL receptor that mediates selective uptake of cholesteryl ester; PLTP - phospholipid transfer protein, transferring phospholipids from triglyceride-rich lipoproteins to HDL.
- Cholesterol metabolism: HMGCR - enzyme in the rate-limiting step of cholesterol synthesis; ABCG5, ABCG8 and NPC1L1 - sterol absorption in intestine; CYP7A1, cholesterol 7-alpha-hydroxylase; STARD3, a cholesterol transport gene.

- TG metabolism: lipoprotein lipase (LPL); APOC2 - cofactor of LPL, endothelia lipase (LIPG); LPA, lipoprotein(a); ANGPTL3 and ANGPTL4 - inhibit endothelial lipase; LMF1 - maturation and transport of lipoprotein lipase through the secretory pathway; PNPLA2 and ABHD5 (cofactor of PNPLA2) - hydrolysis of triglycerides in adipose tissue; LPIN1 - direct lipid to adipose storage sites; PLIN1 - protecting lipid droplets until they can be broken down by hormone-sensitive lipase in adipose tissue.
- Carbohydrate metabolism: (potentially related through the interactions of carbohydrate and lipid metabolism, e.g. precursors of lipids are from glycolysis and pyruvate) GALNT2 - function in the first step of O-linked oligosaccharide biosynthesis; PPP1R3B - may be involved in regulating glycogen synthesis in liver and skeleton muscle
- Misc: TTC39B - encoding tetratricopeptide repeat domain 39B

Plasma lipoproteins: genetic influences and clinical implications [Hegele, NRG, 2009]

- Earlier studies with small case-control or cohort-based association studies and linkage studies: quantitative lipoprotein traits (normal variation) with candidate genes or genome-wide marker sets. Convincing data are lacking for significant metabolic roles for USF1, WWOX or numerous other genes found this way, and most are not replicated in later GWAS.
- Monogenic disease: use phenotypes of extreme levels of TG or cholesterol (hyperlipoproteinaemia, HLP). Some examples:
  - HLP type 2A: very high level (95 percentile) of LDL-C. 10% of these subjects have a discrete monogenic syndrome with mutations in LDLR, LDLRAP1, APOB or PCSK9.
  - HDL cholesterol level below the fifth percentile: have extremely rare monogenic disorders, some due to mutation in ABCA1, ApoA1 or LCAT
  - HLP type 1: plasma TG levels above the ninety-fifth percentile have rare monogenic disorders with mutations in lipoprotein lipase (LPL), APOC2 and APOA5 genes.
  - Lesson: For instance, plasma LDL cholesterol levels depend crucially on LDL receptor function, which in turn requires proper binding of apolipoprotein B, the presence of LDL receptor accessory protein 1 (LDLRAP1) as a chaperone through the early phase of endocytosis, and regulated intracellular receptor degradation by the convertase PCSK9.
- GWAS:
  - LDL cholesterol: approximately half of the associated genes had been identified previously, for example APOE, LDL receptor (LDLR), APOB, PCSK9 and HMGCR. Novel loci inclulde those containing sortilin 1 (SORT1), cartilage intermediate layer protein 2 (CILP2), basal cell adhesion molecule (BCAM) or the translocase gene TOMM40.
  - HDL cholesterol: approximately three-quarters of significant SNPs were in loci harbouring known genes such as CETP, LIPC, LPL, ABCA1, endothelial lipase (LIPG) and LCAT. Only GALNT2 and the MVK-MMAB locus had no previous connection to HDL metabolism
  - TG: approximately one-third of genes in significantly associated loci were known, such as APOA5, LPL, LIPC, APOB and ANGPTL3, whereas loci harbouring CILP2, TRIB1, GCKR, CHREBP (also called MLXIPL) and GALNT2 had minimal prior connection to TG metabolism.
  - Some genes in newly identified loci link TG with carbohydrate metabolism: (this is relevant as the most common lipid disturbance observed in diabetes is elevated plasma TG). GCKR21 encodes glucokinase regulatory protein; CHREBP63 encodes a glucose-responsive transcription factor that is active in hepatic glycolysis, lipogenesis and VLDL secretion; and GALNT2 encodes UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase.
- Animal studies:

  - Lipase maturation factor 1 (LMF1) is essential for the processing and secretion of both LPL and hepatic lipase.
  - Two genes encoding ATP-binding cassette proteins, ABCG5 and ABCG8: role in sterol excretion from intestinal cells.
  - NPC1L1: the role in intestinal sterol transport and as the target for a new class of drugs that inhibit sterol absorption.
  - Several adipocyte-based genes: Atgl (also known as Pnpla2), encoding adipose triglyceride lipase, which is involved in the intracellular lipolysis of TG77; Abhd5, the cofactor for ATGL; Lpin1, directs lipids to adipose storage sites79; and Plin, provides a scaffold that coordinates access of enzymes to lipid droplets in adipocytes.

- Future work: Mendelian randomization has been proposed as an approach to assess whether genetically determined intermediate traits, such as lipoprotein levels, are causally related to end points, such as CVD. Association of CVD risk with plasma lipoproteins can be diluted by non-genetic factors that alter plasma lipoproteins, whereas the association of CVD with the genetic determinants of lipoprotein levels is more direct and less susceptible to confouding effects.

Biological, clinical and population relevance of 95 loci for blood lipids [Teslovich, Nature, 2010]:

- Meta-analysis: 46 lipid GWASs, more than 100,000 individuals of European descent. A total of 2.6 million genotyped or imputed SNPs were tested for association with each of the four lipid traits: TG, HDL, LDL and total cholesterol (TC), in each study and the results were combined with a fixed-effects meta-analysis.

- GWAS loci: identified 95 loci that showed genome-wide significant association ($P < 5 \cdot 10^{-8}$) with at least one of the four traits. The total set of mapped variants (both lead and significant secondary SNPs) explains about 10% of the total variance in each lipid trait in the Framingham Heart Study, corresponding to about 25% of the genetic variance for each trait.

- eQTL: data of $> 39,000$ transcripts in liver (960 samples), omental fat (741 samples) and subcutaneous fat (609 samples). Identified cis-eQTL (500 kb) at $P < 5 \cdot 10^{-8}$: among 95 loci, 38 SNP-to-gene eQTLs in liver, 28 in omental fat, and 19 in subcutaneous fat.

- Clinical significance: association testing of lead SNPs for CAD: A limited number of loci met $P < 0.001$, with most of them being associated with LDL-C. Four novel CAD-associated loci related specifically to HDL-C or TG, but not LDL-C: IRS1 (HDL-C, TG), C6orf106 (HDL-C), KLF14 (HDL-C) and NAT2 (TG). Not clear if they affect CAD risk through HDL or pleiotropic effect (e.g. IRS1 for insulin)

- Functional evidence of some novel genes: GALNT2, PPP1R3B and TTC39B, overexpression or knock-down of these genes in mouse liver significantly change the plasmid lipid level.

Non-coding variant associated with cholesterol through SORT1 [Musunuru & Rader, Nature, 2010]:

- Background: 1p13 locus is strongly associated with both LDL and Myocardial infraction (MI) diease.

- 1p13 SNPs are associated with SORT1 and PSRC1 expression in liver eQTL data, but none of the SNPs show association in adipose tissues and in lymphocytes.

- A causal 1p13 variant: from the strongest association with LDL level, identified 6 SNPs in a region of 6.1 kb between CELSR2 and PSRC1. Resequencing of this region identified 16 SNPs. Then test the functional difference of two haplotypes of this noncoding sequence: in terms of luciferase expression, or SORT1 expression, in human Hep3B cells. One SNP rs12740374 has a causal influence (through enumerating all SNPs). Also, the minor alleles creates a new binding site for CEBPA (verified by in vitro binding experiment).

- SORT1 expression and LDL: in mouse, overexpression or small interfering RNA-knockdown of SORT1 in liver changes significantly the level of LDL. Note that: the overexpression (using special viral vectors/promoters) and knockdown were limited to liver.
- Remark:
  - Key experiments to establish the mechanism of the 1p13 locus is: (1) the association with LDL and SORT1 expression; (2) fine-mapping of causal variants: need sequencing and functional assay of variants (TF binding, gene expression); (3) SORT1 and LDL in mouse.
  - Lesson about the noncoding variant: located quite far away from the target gene (two genes between the SNP and SORT1).

Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis (DIAGRAM+ study) [Voight, NG, 2010]

- Stage 1:
  - Data: genome-wide association data from 8,130 individuals with type 2 diabetes (T2D) and 38,987 controls of European descent, combining data from WTCCC, DGI, FUSION, deCODE genetics, the Diabetes Gene Discovery Group, the Cooperative Health Research in the Region of Augsburg group (KORAgen), and the Rotterdam study and the European Special Population Research Network (EUROSPAN).
  - Meta-analysis: 2,426,886 imputed and genotyped autosomal SNPs into a fixed-effects, additive-model meta-analysis using the inverse-variance method.
  - Modest genomic control inflation: $\lambda = 1.07$. After removing SNPs within established T2D loci (Supplementary Table 3), the resulting quantile-quantile plot was consistent with a modest excess of disease associations of relatively small effect.
- Stage 2: in Stage 1, 23 new autosomal regions showing the most compelling evidence for association, all $P < 10^{-5}$. Replication test.
  - 21 showed directional consistency of effect between stage 1 and 2, binomial test $P = 3.3E - 5$. For 15, the stage 2 P value was < 0.05
  - Total: 31 loci including 20 previous reported and 11 new loci.
- eQTL analysis: cis-eQTL of blood and adipose tissues [Emilsson, Nature08]. several are cis-eQTLs, the stronest is one of KLF14. Also use conditional analysis to test cauaslity.
- Pathway analysis:
  - Gene list: from a list of loci, generate the genes using the nearest recombination hotspots. Results: 31 confirmed loci $\rightarrow$ 82 genes; 110 expanded loci (no HLA, all $P < 1E - 4$ in Stage 1) $\rightarrow$ 320 genes.
  - GRAIL analysis: literature evidence of connection
  - Pathway enrichment: PANTHER (2 general categories significant after Bonferroni correction); REACTOME (some significant, at $FDR < 0.2$): e.g. cell cycle, Notch signaling, FOXA transcription from the confirmed list.
  - PPI networks: some known interactions
  - MAGENTA: test enrichment of known T2D pathways. Method: assign each gene a P-value - the strongest associated SNP, corrected for confounders; then test significance using GSEA. Overall, we observed that gene sets related to cell cycle, inflammatory response, and fatty acid oxidation were nominally enriched for genes association.

A genome-wide perspective of genetic variation in human metabolism [Illig & Suhre, NG, 2010]:

- Data: concentration of 163 metabolites in blood of 1,809 individuals. And replication study in another 422 individuals.
- Analysis: use metaboites or metabolite ratios ($163 \cdot 162$) as traits, and do linear regression on every SNP assuming additive model (i.e. each copy of the minor allele makes a contribution to the trait). The threshold: 10^{-7} for metabolites and 10^{-9} for metabolite ratios.
- 15 loci were identified in 1,809 individuals, and 9 were replicated ($P < 0.05$ after Bonferroni correction for 15 tests). The associated genes often have matching functions: 3 genes in β-oxidation, one in generating energy from β-oxidation, three genes in fatty acid biosynthesis, 2 in AA metabolism, and two are transporters.
- Some loci (genes) are associated to clinical parameters in previous GWAS: e.g. FADS1 is associated with LDL and HDL level in previous GWAS; a SNP in APOA1-APOC3-APOA4-APOA5 cluster was associated with blood triglyceride levels in previous GWAS (with phosphatidylcholines in this study).

Expression-based genome-wide association study links the receptor CD44 in adipose tissue with type 2 diabetes [Kodama & Butte, PNAS, 2012]

- Idea: genes differentially expressed in repeated experiments (cases vs. controls) are likely to be causal genes. Denote this as eGWAS.
- Data: 130 independent microarray experiments, totaling 1,175 samples collected from public repositories.
- Test: ranked all 24,898 genes by the likelihood that repeated differential expression for that gene was due to chance. For each experiment, compute a d-score measuring diff. expression, then (1) count the number of experiments where the gene is DEX, and compute the significance, or (2) weighted-Z or other method of combining p-values.
- Results: 127 genes after Bonf. correction. Top gene: CD44, markedly differentially expressed in experiments studying diabetes in adipose tissue compared with other tissues.
- Evidence of CD44:
  - One of the known ligands for CD44, SPP1, is also a top gene. SPP1 may serve as a link between adipose tissue inflammation and insulin resistence.
  - In mouse model, CD44 expression increases in obese adipose tissue
  - CD44 deficient mouse: less adipose tissue inflammation and insulin resistence.
  - CD44 Blockade Decreases Blood Glucose Levels and Adipose Macrophage Infiltration
- Remark: genes repeatedly show diff. expression may be causal genes. The idea is: there may be multiple symptoms of T2D, thus many disease reactive genes, but in different patients, the symptoms may be different (e.g. kidney, eyes, etc. for T2D), and the reactive genes tend to be different. The repeatedly DEX genes tend to lie in the common denominator of all T2D patients (core T2D development pathway).

Integration of disease-specific single nucleotide polymorphisms, expression quantitative trait loci and coexpression networks reveal novel candidate genes for type 2 diabetes [Kang & Butte, Diabetologia, 2012]

- T2D expression traits: (1) 32 replicated SNPs from T2D GWAS, and then (2) 21 of these 32 SNPs were associated with 62 different expression traits (eQTL data from liver and subcutaneous and omental adipose tissue) at the uncorrected threshold of $p = 0.05$.

- Weighted co-expression networks: from mouse expression data. A total of 2,326 tissue-specific coexpression modules.
- Scoring T2D expression traits: (1) Out of the 62 type 2 diabetes expression traits, 33 were present in one or more coexpression network modules, resulting in the implication of 526 type 2 diabetes coexpression network modules. (2) for each gene, its score is the number of modules it appears (weighted by the inverse of the module size).
- Evaluation of predictions: map the T2D expression traits to eSNPs using the eQTL data, then assess the p-value distribution of these eSNPs (enrichment of low p-values).
- Remark:
  - The downstream genes of T2D SNPs from eQTL data are good candidate genes.
  - How to use gene expression data to score T2D candidates? (1) If some genes are known T2D gene, then the modules or other genes in the module are likely T2D-related. (2) Correlation of modules with T2D phenotype (if expression and phenotype data are available in the same subjects).

Whole-genome sequence-based analysis of thyroid function [Taylor & UK10K, Nature Comm, 2015]

- Traits: circulating concentrations of free thyroxine (FT4) and the pituitary hormone thyrotropin (TSH).
- Single point analysis: WGS of 2,287 samples, 8M SNPs.
- eQTL analysis: using Genevar database and UK-Twins study.
- Rare variant analysis: candidate genes from GWAS, then SKAT on the rare variants in 50K regions near a gene.
- GCTA and polygenic score analysis: GCTA to estimate heritability. A genetic score based on 67 SNPs previously associated with thyroid function in GWAS shows strong evidence of association with TSH and FT4. Also evidence of shared genetic pathways with TSH associated with the FT4 gene score.
- Lessons: some standard analysis on WGS association data include single-point analysis, RV analysis of regions, eQTL/rQTL, heritability, shared heritability.

The Genetic Architecture of type 2 diabetes [Nature, 2016]

- WGS analysis (GoT2D): about 2,600 samples. Single variant analysis: found 4 loci, three in previous GWAS. Imputation: more findings: 14 loci.
- WES analysis (T2DGenes):
  - Single variant: imputation on 28K cases and 50K controls, found 18 coding variants (Table 1). Most are common, in previous GWAS. Help resolve GWAS candidate genes.
  - Gene based test: define variant groups, LoF, LoF + NS (strict) and LoF + NS (broad). SKAT-O test, no overall signal, limit to candidate genes (694 mapped to known GWAS regions), 1 gene reaches significance.
- Rare variant burden in gene sets: Mendelian diabetes genes. Significant burden in LoF (OR = 1.8) and LoF + NS.strict (OR = 1.5).
- Fine-mapping and annotations: enrichment in CDS, TFBS, enhancers in adipose and pancreatic islets.
- Genetic architecture of T2D:

  - Estimation of LVE: for each SNP, estimate its LVE, and add up LVEs. Ext Figure 7: 6.3% LVE from CVs, 2.9% LVE from AF 0.1-5%.
  - Simulation: under different scenarios, depending on how much h^2 is explained by CVs or RVs. The pattern (number of discovered loci at a particular p-value threshold) is consistent with polygenic model, RV explains 25%.

- Remark: The method for estimating LVE: accounting for LD?

A functional genomics pipeline identifies extensive allelic heterogeneity and cross-tissue effects within obesity-associated GWAS loci [Joslin and Nobrega, submitted to NG, 2020]

- Background: BMI genetics, 97 loci, enriched in brain, then adipose.
- Differentiation of pre-adipocytes to mature adipocytes, and iPSC into hypothalamic neurons: pcHiC, ATAC, RNA.
- PcHiC data: 600K-900K interactions per time point (30/gene), fragment size 422 bp median, and interaction: 100-200kb.
- Dynamics of RNA, OCR and interactions: (1) Gene: three main clusters, down (2000 genes), up (4000), and V shape (6000 genes). (2) HSV plots of temporal patterns. ATAC: mostly up. PcHiC: more show down-regulation; also seems to have a delay in activation.
- Mapping genes to OCRs: use pcHiC, 3-4 peaks per gene.
- MPRA: 97 loci, lead variants and LD (.8), a total of 2400 variants. 800 regions have enhancer activities in brain and 500 in adipose, with 400 shared. Validation of enhancer using Luciferase: 65% validation rate (Figure S3c).
- Among all variants: 94 have enhancer-modulating activities (EMVars). 61 brain EMVars and 70 adipose EMVars and at least one was identified in 40/97 (41%) of tested GWAS loci. 2/3 of loci contain more than 1 EMVars. 37/94 (39%) of these variants affected enhancer activity in both cell types.
- Prioritization of candidate genes by assigning EMVars to targets using pcHiC: about half do not have interactions, but for those with interactions, often more than 1, with median about 3. Also use adipose and brain eQTL to assign targets. Define classes of 150-200 candidate genes in either adipose or neuron: some hi-C and eQTL in the right cell type (class I), some only hi-C and eQTL in a different type (class II), some either evidence (class IV).
- rs4776984: brain EMVar, it is adipose and brain eQTL, hi-C interaction with MAP2K5 in both adipose and neuron. Another SNP about 50kb away: EMVar, and hi-C with MAP2K5 and eQTL in both adipose and neuron.
- 16p11.2 locus: 600kb, spanning two independent GWAS loci, 10 EMVars. Also pcHiC interaction between the two loci. (1) Region 1: has 3 EMVras, but two are not eQTLs nor PC-HiC interaction. The third SNP at 3 UTR of SBK1, EMVar in both adipose and brain, eQTL of 5 genes in adipose and 7 in brain, and pcHiC of 18 genes. (2) Region 2: 7 EMVars, 5 are in perfect LD. All are eQTLs of 9 genes in adipose and 5 in brain, and each contacts several promoters.
- CRISPR deletion of enhancers containing rs2650492 (from SBK1 region) and rs9972768 (from the second region): in iPSC, then differentiation into hypothalamic neurons. (1) SBK1 expression is reduced in one stage for both deletions. SBK1 homolog in zebrafish and mouse: neurodevelopment defects. (2) Another nearby gene with pcHiC support: NUPR1 showing DE in HEK293 cells.
- Lesson: functional variants often have pleiotropic effects in Hi-C and eQTLs, and can have effects across tissues. Additionally, in any single locus, there could be multiple functional variants.

- Remark: (1) Limitation of MPRA: in natural context, may have fewer EMVars and be more tissue-specific. (2) Does HT neurons capture brain enrichment? Could redo brain enhancer enrichment, but removing/conditioning HT neurons. (3) While there are multiple EMVars, it only proves that regulatory variants are relatively common, but it does not prove that there are multiple disease-causing variants.

8.4 Immune-Related Traits

Genetics of immune related disorders [personal notes]:

- Role of infection in AIDs: possible models support the association of infection with AIDs: (1) Antigen mimicry: some self-antigens are similar to antigens from pathogens. (2) Pre-existing responses against self-antigens: exacerbated by infections, which act as adjuvants. (3) Inappropriate/excessive immune response against infections lead to AIDs.
- Sero-positive vs. sero-negative AIDs: for some diseases, it is important to have self-antigens (first two models), and in other diseases (model 3), not necessary.
- Difference of AIDs vs. allergic diseases: because infection is an important part of AIDs, so these AIDs involve genes in innate immunity against infections (e.g. autophage). Allergy: reaction against other foreign antigens that are not from infections.
- Question: Why in developed countries, the decrease of bacterial infection is not associated with a decrease of AID?
- Question: Organ transplantation: rejection of foreign MHC, how CD8 T cells are activated? The TCRs of CD8 T cells do not match the foreign MHC I.

Overview of autoimmune diseases: [Marrack, Nature Medicine, 2001], [Rious, Nature, 2005], [Zhernakova & Wijmenga, NRG, 2009]

- Autoimmune diseases occur in up to 3-5% of the general population.
- Target tissues: There is an autoimmune disease specific for nearly every organ in the body, e.g. T1D (beta pancreatic cells); MS (brain/spinal cord). In other autoimmune diseases, such as systemic lupus erythematosus (SLE), RA, no particular cell type seems to be targeted.
- Antigen specificity: AIDs are antigen-specific. For AIDs that target many cell types, the antigens are probably expressed throughout the body. However, recognition of widely expressed antigens sometimes results unexpectedly in organ selective manifestations, e.g. in RA, antibodies against a widely expressed antigen (IgG, fibrin) can target destruction of the joints selectively.
- Evidence of environmental impact: Bacterial LPS and mycobacteria can induce various autoimmune diseases, sometimes in the absence of any additional antigen besides that provided by the host itself. Numerous anecdotal reports describe association between the onset of various autoimmune diseases and infections. The environment can also affect the immunoreactivity of the individual by shifting the balance of T cells within the individual between Th1 and Th2 cells.
- Antigen (molecular) mimicry or cross-reactivity: If the agent codes for a peptide that is closely related to a peptide of the host, a vigorous responses might powerfully induce T cells. Ex. (1) In RA, host Hsp60 in joint is similar to Mycobacteria/Hsp65; (2) T1D: host Pancreatic beta cells/GAD is similar to Coxsackie B/P2-C; (3) MS: host Brain/myelin basic protein is similar to Papillomavirus/L2.
- Epidemiological difficulties: in developed countries, bacterial infections have dropped in frequency, and asthma disease thought to be driven by Th2 cells has increased without a concomitant decrease in the incidence of autoimmune diseases.

Model of AID genetics: AIDs develop when self-reactive lymphocytes escape from tolerance and are activated. This is believed to result from a combination of genetic variants, acquired environmental triggers such as infections, and stochastic events [ibid]:

- Basic model: some genetic polymorphism changes part of the immune system, e.g. antigen over-expression or negative feedback of TCR signaling, and these changes make the immune system more likely to be activated by self-antigens. With certain environmental trigger (providing adjuvants), e.g. infection or tissue injury, these self-antigens lead to auto-immune responses.
- Heritability and clustering of diseases in families: this can be explained by the basic model (1) heritability: both genetics and shared environment; (2) clustering: different environment triggers may lead to different diseases for the same genetic variations.
- Specificity of auto-antigens: why the change of the general component of the immune system, e.g. CTLA4, may lead to response to specific auto-antigens? Possible explanations:
  - Most of these component are probably associated with specific cell types, e.g. different JAK, STAT family members may be used by different Th cell types.
  - Most auto-antigens are not harmful, and only those vulnerable to environmental triggers such as virus infection can lead to disease. Certain types of pathogens may be common.
- Pre-existing anto-antibodies: this is the evidence of the model. Auto-antibodies are produced early in some immune diseases before the clinical symptoms appear; however, auto-antibodies are also found in healthy individuals. Epidemiological studies could help to elucidate whether the presence of auto-antibodies predisposes to autoimmune disease, or whether auto-antibody production is a consequence of disease.

Common autoimmune/inflammatory diseases: the symptoms, auto-antigens and disease pathogenesis [ibid]:

- Asthma: Chronic condition in the respiratory system in which the airways occasionally constrict, become inflamed, and are lined with excessive amounts of mucus. Inflammation in response to exposure to an environmental stimulant, such as an allergen, smoke or perfume, which is mediated by a TH2-type immune response and includes mast cells, eosinophil infiltrates and IgE antibodies
- Crohn's disease: Chronic, episodic, inflammatory bowel disease, which primarily causes ulceration of the small and large intestines but can affect any region of the digestive system. Unknown; involves an inappropriate immune response to commensal bacteria.
- Multiple sclerosis: Autoimmune attack of the central nervous system, which leads to demyelination of neurons, causing potentially debilitating physical and mental symptoms. After infection in the brain, trapped T cells initiate an autoimmune response to foreign myelin, thereby triggering inflammatory processes, stimulating other immune cells, cytokines and antibodies.
- Rheumatoid arthritis: Chronic inflammation of synovial joints. Autoimmune reaction against connective tissue components. Presence of rheumatoid factor and ACPA.
- Systemic lupus erythematosus: Chronic inflammation, can affect any part of the body, but often the heart, joints, skin, lungs, blood vessels, liver, kidneys and nervous system. Autoimmune reaction against nuclear proteins, which leads to the formation of immune complexes.
- T1D: Destruction of pancreatic β-cells, which leads to insufficient release of insulin from the pancreas. T cell-mediated autoimmune response, and production of auto-antibodies against islet cells, insulin, glutamic acid decarboxylase and protein tyrosine phosphatase.

Genes implicated in AIDs from earlier studies (single gene diseases) [ibid]:

- MHC: has been associated with almost all autoimmune diseases. The MHC locus spans approximately 4 MB and contains about 250 genes, of which about 60% have immune-related functions. The MHC region is characterized by extended LD blocks (up to 3 MB), and by a strong and complicated LD pattern between the blocks.
- AIRE is mutated in APS-1, autoimmune attack against multiple endocrine organs, the skin and other tissues. Mechanism: Decreased expression of self antigens in the thymus, resulting in defective negative selection of self-reactive T cells.
- CTLA4 works by competitively blocking the engagement of the activating receptor CD28 (by CD80 or CD86). Several AIDs, including Graves's disease, type 1 diabetes and other endocrinopathies, show a striking association with a CTLA4 polymorphism that results in reduced production of a truncated splice variant.
- FOXP3 is a regulator of regulatory T cells (CD4 CD25 T cells). Induced knockout or spontaneous mutation of the mouse Foxp3 gene led to a systemic autoimmune disease associated with the absence of CD4+CD25+ regulatory T cells.
- The Fas death receptor contributes to the deletion of mature T and B cells that recognize self antigens.

Background: T cell differentiation

- TH1 cells: induced by IL12, IL18, IL27 and IFNα, β, γ. Signaling: STAT1,3, or 4 $\rightarrow$ TBX21. Proliferation: IL18, IFNγ. IFNγ inhibits development of TH2 and TH17 cells.
- TH17 cells: A subset of CD4+ T-helper cells that produce interleukin 17 (IL-17), now thought to be more important than TH1 cells as mediators in immune-related diseases. Induced by IL6 and TGFβ. Signaling: STAT3, RORγt. Proliferation: IL23, IL21, IL17.
- Treg cells: There is increasing evidence that Treg cells are less active in chronic immune-related diseases. IL-2 and its receptor (encoded by IL2RA, IL2RB and IL2RG) are crucial in the activation and function of Treg cells. Also induced by TGFβ. IL2 deficient mice develop autoimmunity.

Prevention and treatment of AIDs [ibid]:

- Vaccination: associations with bacterial and viral infections have been suggested for most of the diseases discussed, e.g. a positive correlation between MS or SLE and auto-antibodies against EB virus. Defining the profile of the genetic susceptibility pathway, together with knowledge of environmental triggers, might help prevent immune-related diseases through vaccination.

Genes associated with AIDs from GWAS: identified 23 genes that are shared by two or more diseases (among 11 diseases, from 22 GWAS) [Zhernakova09]

- Immune cell signaling:
  - Shared TCR signaling genes: CTLA4, the protein tyrosine phosphatases PTPN2 and PTPN22 and the adaptor protein SH2B3. PTPN2, plays an important part in the negative regulation of the inflammatory response in T cells.
  - TH1 cell signaling: association of IL18RAP (interleukin-18 receptor accessory protein), IL12, IL10, STAT3 and STAT4 with almost all the diseases analyzed.
  - TH17 cell signaling: Chronically inflamed tissues are infiltrated with highly differentiated TH17 cells. Genes associated with TH17 cells (for example, IL23R and IL21) are associated with nearly all immune-related diseases.
  - Treg cell signaling: these cells are less active in chronic immune-related diseases. Two genes in the Treg activation cascade have now been associated with multiple autoimmune diseases: IL2RA (also known as CD25) and the locus that includes IL2 and IL21.

  - PDCD1 (programmed cell death 1) gene: associated with SLE. It has been shown to regulate peripheral tolerance in T and B cells.

- Antigen representation and expression:
  - MHC class II: different alleles have different abilities to present peptides from target cells to autoreactive CD4+ T cells. E.g. the major genetic contribution to RA involves particular HLA-DR alleles. Certain HLA alleles might be particularly good at presenting glutamic acid decarboxyase (GAD)-65 or insulin peptides to T cells, thus contributing to recognition and ultimate destruction of pancreatic beta cells.
  - Genes affect the expression, or distribution of auto-antigens either in lymphoid tissues or in the target organ. Ex. polymorphism of upsream sequence of insulin gene influences transcription of the insulin gene within the thymus and might thus affect T-cell tolerance to this antigen.
  - CLEC16A is associated with T1D and multiple sclerosis and encodes a C-type lectin receptor, which might play a part in antigen sampling by dendritic cells.

- Innate immunity and TNF signaling:
  - Barrier function: the association of the NOD2 gene and the MUC19 (mucin 19)-containing locus to Crohn's disease. Deficiency in another component of the mucosal mucus layer, MUC2, leads to inflammatory bowel disease. Another barrier risk factor is copy number variation (CNV) in defensin genes.
  - Autophage is known to be involved in Crohns disease, underscoring the role of intracellular processing of bacteria in disease pathogenesis. Three Crohn's disease genes, ATG16L1 (autophagy 16 related-like 1), IRGM (immunity-related GTPase M) and LRRK2 (leucine-rich repeat kinase 2). SLE is associated to another key autophagy molecule, ATG5.
  - Type 1 interferons mediate the early innate immune response to viral infections. IRF5 and IFIH1 (interferon induced with helicase C domain; also known as MDA5), involved in the IFNα pathway, are associated with several autoimmune and inflammatory diseases.
  - TNF signaling: TNFSF15, which is associated with Crohn's disease, is activated via stimulation by LPS. TNFAIP3, which is associated with SLE, rheumatoid arthritis and coeliac disease, encodes the A20 protein, which is required for termination of the NFκB signal that is mediated by innate immune receptors.

- Tissue response: the gene variants may influence relative isolation of tissues from the immune system and inhibition of function of invading lymphocytes.
  - The eye, one of the beststudied examples of a protected site, has barriers to T-cell infiltration and produces immunosuppressive cytokines, such as transforming growth factor (TGF)-beta.
  - FcγRIIA and FcγRIII alleles coding for proteins with lower than normal activity are associated with susceptibility to SLE and lupus nephritis, probably because these alleles clear circulating immune complexes inefficiently and allow increased immune complex deposition in the kidney.

- Other genes:
  - Chemokines: their major role is to regulate the immune response and recruit effector immune cells to sites of inflammation. CCL21 (chemokine (C-C motif) ligand 21) is associated with rheumatoid arthritis, CCR6 (chemokine receptor 6) is associated with Crohn's disease. Chemokine genes usually form a cluster with strong LD between genes.
  - TSHR (thyroidstimulating hormone receptor): associated with autoimmune thyroid disease.
  - The ORMDL3 gene is associated with asthma and Crohns disease and was prioritized for study owing to the strong cis-correlation between ORMDL3 expression and its associated genotype.

Genetic insights into common pathways and complex relationships among immune-mediated diseases [NRG, 2013]

- Sero-positive and sero-negative: whether a patient has auto-antibodies. Positive: T1D, RA. Negative: AS, Coelic disease, IBD, psorasis.
- Allelic heterogenity and rare variants: CARD9 and NOD2 loci, rare variants have distinct effects from common variants.
- Shard genetics: correlated and concordant (increase risk for both disease), correlated and discordant (opposite effects), non-correlated (different risk haplotypes in the same locus).
- Patterns of shared loci: Box 2, for some traits (e.g. all seronegative), most correlated and concordant, but for some pairs, e.g. IBD and T1D, most are discordant and non-correlated. Over 400 pairs, about 40% concordant, 14% discordant and the rest non-correlated.
- Specific loci: (1) Th1 and IL-23 pathway: IL12R and IL23 pathway activation in Th1 and Th17 cells. (2) Other loci: MHC, NF-kappa B. IRF.
- Disease specific loci: NOD2 (autophagy) in CD: important for gut bacterial. HNF4A in UC: epithelia barrier. INS (insulin) in T1D: auto-antibody generation.

Genetics of allergy and allergic sensitization: common variants, rare mutations [Curr Opinion Immuno, 2015]

- Background: allergy is mediated by Th2 cells (B cell activation) and AIDs by Th1 cells (macrophages, neutrophils)
- Allergy vs. asthma: likely high sharing, in one study, 9/10 allergy loci are also associated with asthma.
- Biological pathways from GWAS (Figure 1): Epithelia barrier function: FLG. Innate immunity sensing: some TLR genes. T cell activation: MHC, IL-2, IL2RB. T cell response: especially Th2 pathway, IL13, IL33. T-reg and immune tolerance: LRRC32, TGF-beta, SMAD3, FOXP3.
- Allergy vs. IBD: many shared loci, 12/18 allergy loci also IBD. However, the effects may be in the same direction or not.

Crohn's disease: WTCCC [WTCCC, Nature, 2007]

- Background: The pathogenic mechanisms are poorly understood, but probably involve a dysregulated immune response to commensal intestinal bacteria and possibly defects in mucosal barrier function or bacterial clearance.
- Replicate the genes/regions previously reported: IL23R, NOD2, ATG16L1 (ATG16 autophagy related 16-like 1) gene, a noncoding intergenic SNP mapping 14-kb telomeric to gene ZNF365 and 55-kb centromeric to the pseudogene antiquitin-like 4, within a 1.2Mb gene desert on chromosome 5p13.1
- Four new strong associations: all successfully replicated
  - IRGM: a GTP-binding protein which induces autophagy and is involved in elimination of intracellular bacteria
  - MST1 (macrophage stimulating 1), which encodes a protein influencing motile activity and phagocytosis by resident peritoneal macrophage. (The strongest SNP is a synonymous coding SNP within the BSN gene, involved in neurotransmitter release.)
  - NKX2-3: Targeted disruption of the murine homologue of NKX2-3 results in defective development of the intestine and secondary lymphoid organs.

  - PTPN2: encodes the T cell protein tyrosine phosphatase TCPTP, a key negative regulator of inflammatory responses. The same locus also shows strong association with T1D susceptibility and a weak association with RA (P = 1.9E-2)

- Other putative genes with weaker evidence, based on biological candidacy.
  - HLA: $P = 8.7E-7$
  - TNFAIP3 (TNFa induced protein 3): same pathway as NOD2
  - TNFSF15 (tumour necrosis factor super family, member 15): $P = 9E-5$, previously reported associated with CD
  - STAT3: $P = 3.1E-5$
  - CD40LG: CD40 ligand, $P = 1.3E-7$

T1D: WTCCC [WTCCC, Nature, 2007]:

- Six genes/regions for which there is strong pre-existing statistical support for a role in T1D-susceptibility: MHC, insulin gene, CTLA4, PTPN22, (IL2RA/CD25), IFIH1/MDA5. Five of these previously identified associations were detected in this scan ($P \leq 0.001$), the exception being the INS gene.

- Three novel regions from single point analysis, all replicated (12q13, 12q24 and 16p13). Four regions from multipoint analysis or from analysis of all AIDs, one replicated (18p11).
  - 12q13 region: extensive LD of > 10 genes. Candidates: ERBB3 (receptor tyrosine-protein kinase erbB-3 precursor)
  - 12q24 region: extensive LD of > 10 genes. Candidates: SH2B3/LNK (SH2B adaptor protein 3), TRAFD1 (TRAF-type zinc finger domain containing 1) and PTPN11 (protein tyrosine phosphatase, non-receptor type 11). Of those listed, PTPN11 is a particularly attractive candidate given a major role in insulin and immune signalling, and also the same family as PTPN22.
  - 16p13 region: two genes of unknown function
  - 18p11 region: seems to confer susceptibility to all three autoimmune conditions. PTPN2.

Crohn's disease: meta-analysis [Barrett & Daly, NG, 2008]:

- Data: 3 studies including WTCCC. The combined sample has 74% power at an OR of 1.2. Total: about 4,000 cases and controls respectively, and about 2,000 cases and controls, respectively, in the replication data.

- Method: the meta-analysis used a test that combined the results from each study (rather than mixing the raw data and compromising the case-control matching of each study). We summarized the standard 1 d.f. allele-based test of association as a Z score within each scan and combined scores across studies to produce a single meta-statistic for each SNP across all three datasets.

- A marked excess of significant associations, well beyond what would be attributable to the modest overall distributional inflation ($\lambda_{GC} < 1.16$).

- 526 SNPs from 74 distinct genomic loci that were associated with $P < 5E-5$, which is more than seven times the number of SNPs expected by chance even after correction for the modest overall inflation detected.

- Eleven associations previously replicated were among the 74 regions represented.

- The signficance of 63 new regions: strong departure of null distribution (even removing 21 loci, below), suggesting an enrichement of true associations.

- 19 new associations were replicated, and 2 other regions, 19p13, and MHC. The total of 32 expanded associations. Some newly implicated genes:
  - CCR6: homing receptor (GPCR), expressed by immature dendritic cells and memory T cells and is important for B-cell differentiation and tissue-specific migration of dendritic and T cells
  - IL12B: this gene encodes the p40 subunit, which is a constituent of both heterodimeric interleukins IL-12 and IL-23
  - STAT3 and JAK2: the role of both genes in IL23R signaling and the central role of STAT3 in Th17 differentiation.
  - ICOSLG (inducible T-cell co-stimulator ligand): expressed on intestinal (and other) epithelial cells and may have a role in their antigen presentation to and regulation of mucosal T lymphocytes
  - ITLN1 (intelectin-1): expressed in human small bowel and colon, and encodes a 120-kDa homotrimeric lectin recognizing galactofuranosyl residues found in cell walls of various microorganisms
- Coding sequence variation: just 9 of the 32 genome-wide significant associations were correlated with a known nsSNP ($r^2 > 0.5$)
- The possible role of cis-regulatory variation: five correlations between expression of a nearby gene and a CD-associated variant in [Dixon07] data. Expected number = 0.001.
  - PTGER4
  - IBD5 region: the CD-associated SNPs were associated with decreased SLC22A5 mRNA expression
  - The most significant CD-associated eQTL reported here affects ORMDL3, SNPs in precisely the same region were recently shown to be strongly associated with childhood asthma
- Using a liability-threshold model, we estimate that the 32 loci identified to date explain about 10% of the overall variance in disease risk, which may be as much as a fifth of the genetic risk, given previous estimates of CD heritability of approximately 50%

T1D meta-analysis [Barrett & T1DGC, NG, 2009]:

- Data: The total sample set included 7,514 cases and 9,045 reference samples. WTCCC, GoKinD/NIMH and T1DGC. The two earlier studies (WTCCC, GoKinD/NIMH) and the current one (T1DGC) used different platforms. As only 9% of SNPs are shared between these platforms, we used imputation to combine results across studies.
- Meta-analysis: Mantel's extension to the 1 degree-of-freedom (1-d.f.) Cochran-Armitage trend test that combined comparisons over the three studies.
- Forty-one distinct genomic locations provided evidence for association with T1D in the meta-analysis ($P < 10^{-6}$)
- Replication: After excluding previously reported associations, we further tested 27 regions in an independent set of 4,267 cases, 4,463 controls and 2,319 affected sib-pair (ASP) families. Of these, 18 regions were replicated ($P < 0.01$).
- Several of the 18 regions identified here contain genes of possible functional relevance to T1D.
  - The region 1q32.1 contains the immunoregulatory cytokine genes IL10, IL19 and IL20.
  - The region of strong LD at 9p24.2 contains only a single gene, GLIS3.
  - The region on 12p13.31 harbors a number of immunoregulatory genes including CD69, and calcium-dependent (C-type) lectin (CLEC) domain family with immune functions

GWAS of host control of HIV-1: Euro-CHAVI study [Fellay & Goldstein, Science, 2007]:

- Data: 486 patients, genotyping of 555,352 SNPs. The phenotype is (1) the viral set point, i.e. the level of circulating virus in the plasma during the nonsymptomatic phase preceding the progression to AIDS; (2) progression.
- Signficiant associations with viral load: with Bonferoni correction, two loci were found significant, both in HLA region.
  - Polymorphism in HCP5 gene: explains 9.6% of the total variation in set point. Also in strong LD with HLA-B*5701 ($r^2 = 1$). Both are possible candidates. For HCP5: it encodes a human endogenous retrovirus with sequence homology to HIV-1 pol, it may act as antisense RNA interfering with HIV-1 replication.
  - Polymorphism in HLA-C gene: explains 6.5% of the variation in set point. In weak LD with HCP5, however, the effect cannot be explained the HCP5 SNP. cis-eSNP data shows that the SNP is associated with expression of HLA-C.
  - No other significant association: no overall inflation of P values (indicating little contribution from population stratification), but an excess of low P values, beginning with the 355th most associated SNP.
- Significant associations with progression: SNP near ZND1 (RNA Pol I) gene (1 Mb from the previous candidates), explain 5.8% variation. Also cis-eSNP of ZND1, and plausible functional evidence.
- Replication: 140 Caucasian patients. All three associations were confirmed at $P < 0.05$.

GWAS of AIDS progression: GRIV study [Limous & Zagury, Genomewide Association Study of an AIDS Nonprogression Cohort Emphasizes the Role Played by HLA Genes (ANRS Genomewide Association Study 02), J Infect Dis, 2009]

- GRIV cohort: 275 human immunodeficiency virus (HIV) type 1-seropositive nonprogressor patients in relation to a control group of 1352 seronegative individuals. Genotyping: HumanHap300 BeadChips
- Statistical analysis:
  - For each SNP, we performed a standard case-control analysis using Fishers exact test (with PLINK software) to compare allelic distributions between the nonprogression group and the control group. Bonferroni correction
  - Population stratification: genomic inflation factor: $\lambda = 1.064$.
  - Meta-analysis: 286,529 SNPs are common between GRIV and Euro-CHAVI [Fellay07]. Meta-analysis using Fisher's method.
- Significant associations: HCP5 has the only significant SNP after Bonferroni corrections, the same SNP identified in the Euro-CHAVI study.
  - HCP5-independent signals: Most of the signals from chromosome 6 disappeared because of the genetic linkage with the HCP5 rs2395029-G.
  - The strongest signals were still found in the HLA region, with 2 SNPs of the ZNRD1/RNF39 region. Unlike the HCP5 rs2395029 SNP, none of the ZNRD1/RNF39 SNPs alleles seemed to correlate with viral load (figure 4), suggesting that this locus influences disease progression.
- Results of meta-analysis: found other associations.
  - C6orf48 rs9368699 SNP, in LD with the HCP5 SNP.
  - The PSORS1C1 gene exhibited 2 significant SNPs, rs3823418 and rs3815087. PSORS1C1 is a psoriasis-susceptibility candidate gene.

  - HLA-C-related SNP rs10484554 (identified in Euro-CHAVI study)

- The list of best SNPs:
  - Of the 50 best signals found in the meta-analysis, 46 originated from the HLA locus, emphasizingthe massive role played by HLA in the nonprogression phenotype.
  - In this GWAS alone, 31 of the 50 best signals were not from chromosome 6 (table 1) and were not found in the meta-analysis (table 6), suggesting that positive signals outside the HLA locus may be associated with the nonprogression phenotype without influencing viral load.

- Remark:
  - Viral load and progression may be associated with different loci.
  - There may be significant number of loci outside HLA, especially when considering disease progression.
  - An important problem for HLA loci is to identify the causal gene(s). This is difficult because of the extensive LD in the HLA region.

GWAS of HIV control: CHAVI study [Fellay & Goldstein, Common Genetic Variation and the Control of HIV-1 in Humans, PLG, 2009]

- Data: 2,554 infected Caucasian subjects. The study was powered to detect the effects of common variants down to 1.3% of explained variability. From Euro-CHAVI Consortium ($N = 1,397$) and MACS cohort ($N = 1,157$). A total of 2362 individuals were included in the set point association analyses and 1071 seroconverters were eligible for the analysis of disease progression.

- Statistical analysis: each SNP passing the QC step were tested for association with HIV-1 viremia (quantitative trait) at set point in separate linear regression models that included gender, age, and the 12 significant PC axes as covariates.

- Associations with viral load at set point: $\lambda = 1.006$.
  - The 2 SNPs previously reported as genome-wide significant (HCP5 and HLA-C) were confirmed to be the strongest determinants of variation in HIV-1 viral load.
  - Further independent associations in the MHC region: using stepwise regression model, found 4 additional SNPs. Altogether, a model including 6 SNPs, 4 alleles and homozygosity status shows that MHC variation explains 12% of the set point variability in this cohort.

- Association with disease progression:
  - Top associations: SNPs at HCP5/B*5701 and HLA-C. If viral load at set point is added to the models, the association signals are much weaker, demonstrating that the HCP5/B*5701 and HLA-C effects on disease progression are mainly driven by their impact on early viral control.
  - Another set of variants reached genome-wide significance ($P < 1E-8$): in high-LD and located around the ZNDR1 and RNF39 genes, close to the HLA-A locus. This association is largely independent of viremia, suggesting that a different mechanism of action is here modulating HIV disease progression.

- Candidate genes: tested a total of 34 SNPs in 21 genes, representing 27 previously reported associations with HIV-1 control. Most of them are not significant ($P > 0.05$). Signifciant ones are variants in CCR5 (HIV-1 co-receptor) and CCR2 (minor receptor).

GWAS of HIV-1 control: International HIV Controller Study [Perayra & Zhao, The major genetic determinants of HIV-1 control affect HLA class I peptide presentation, Science, 2010]

- Data: 974 controllers (cases) and 2648 progressors (controls) from multiple populations, genotyped on 1M SNPs. HIV controllers: able to control viral replication without therapy, typically maintain stable CD4+ cell counts and do not develop clinical disease.
- Statistical analysis: logistic regression including the major principal components as covariates to correct for population substructure. Genome-wide significance defined as $P < 5E-8$ (Bonferroni correction).
- Associations:
  - In the largest group, comprising 1712 individuals of European ancestry: 313 SNPs with genome-wide significance, all in MHC region.
  - Candidate genes: Only variants in the CCR5-CCR2 locus replicate with nominal statistical significance in our study.
  - Independent markers with associations to host control: using stepwise regression, identified 4 SNPs in the MHC region, near or in genes: HLA-C (cis-eSNP), HLA-B*57:01, MICA and PSORS1C3. These four SNPs explain 19% of the observed variance of host control in the European sample
- Causal polymorphisms in MHC genes: Specific amino acids in the HLA-B peptide binding groove, as well as an independent HLA-C effect, explain the SNP associations.

A CD8+ T cell transcription signature predicts prognosis in autoimmune disease [McKinney & Smith, NM, 2010]

- Data: purified CD8+ T cells in 32 individuals with AAV (a disease).
- Clustering of transcriptome: identify two subgroups of patients, each has a characteristic expression pattern (one subgroup of genes). In fact, only three genes are needed to define the two groups of patients.
- Unsupervised gene clusters also predict prognosis of patients, one group has little replase. The signatures: IL-7 signaling and TCR signaling in one group.

Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway. [Lee & Smith, Cell, 2013]

- Motivation: what determines the prognosis of CD patients? Some are aggressive CD, some are indolent.
- GWAS of prognosis: aggressive CD (668) and indolent CD (389). Limit to 81 genes involved in IL-2 or IL-7 signaling (from [McKinney, NM, 2010]). Found one SNP in FOXO3, OR = 0.62, and p value highly significant in combined (with replication) dataset.
- Regulatory effect of the SNP: not in LD with any coding region of FOXO3, and use ASE to show that minor allele is associated with higher expression (in heterozygous individuals) in monocytes.
- The effect of SNP on cellcular phenotype: association of the SNP with cytokine production (TNFα and IL-10) in stimulated condition.
- Mechanism of the SNP on FOXO3A: FOXO3 nuclear/cytoplasmic ratio is higher in G (minor allele), and show that it is due to de novo synthesis of FOXO3A in nucleus.
- Mechanism of FOXO3: affect TGFβ dependent production of cytokines. Show that FOXO3 binds to TGFβ promoter using ChIP-qPCR.
- Phenotypic study of FOXO3: deletion of the gene in mice leads to higher disease severity.
- **Lessons**:

- Limit to candidate genes, e.g. from DE studies, to increase the power.
- Cellular phenotypes: cell proliferation, cytokine production, etc.
- Studying mechanism of SNP: regulatory effect (eQTL, ASE), downstream process, e.g. TF expression could affect downstream genes.

The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease [Soranzo, Cell, 2018]

- Study: 150K subjects, 36 blood trait indices. GWAS analysis using multiple regression. Conditional regression plus LD pruning. 2700 sentinel variants: each representing high LD group ($r^2 > 0.8$). About 20% have AF $< 5\%$.
- Heritiability analysis: PVE from common variants (LDSC) somewhat higher than the total explained by discovered variants (both common and rare).
- Distribution of disease variants: introns explain twice more than intergenic, which is more than coding. Near genes similar to coding. By epigenomic states: transcription about the same as enhancer, and more than promoter. Enrichment of cell-type specific active enhancers: 5-10 fold enriched in matched cell types.
- Colocalization with molecular QTL: use BLUEPRINT eQTL, splicing QTL (sQTL) and caQTL. Half of overlap are with hQTL (no effect on expression). In all cases of overlap: 25% can be attributed to colocalization (SMR). Overall, 10% of GWAS loci can be assigned to molecular traits.
- MR analysis with related phenotypes: use multiple IVs. Some very large effects between myeloid cells and AIDs (e.g. asthma). With SCZ, find lymphocyte counts, however, this is driven by MHC locus.

Atopic dermatitis (AD) is associated with an increased risk for rheumatoid arthritis and inflammatory bowel disease, and a decreased risk for type 1 diabetes. [Schmidt, J Allergy Clin Immunol. 2016]

- Background: AD genetics, 13 European loci and 10 Asian loci. FLG has skin barrier function, and others immune dysregulation. Most loci are also shared with other immune diseases that are Th1/Th17 mediated, including IBD, RA and T1D. However, the sharing patterns are complex, with loci often having opposing effects. Ex. IL6R allele increases the risk of AD and Asthma and protective of RA.
- Co-morbidities of AD and immune diseases: AD is associated with higher risk of IBD and RA (RR = 1.3 - 1.7), and lower risk of T1D (RR = 0.7).
- Association of immune risk alleles with AD risk: 10 passes significance, 7 have agnostic effects on AD. For those with suggestive associations with AD: $< 50\%$ show effect direction consistent with epidemiological results.
- Discussion: AD patients receive systemic steroids, which may mediate the effects on other immune diseases.
- Discussion: AD, IBD and RA: T-cell mediated inflammation. For RA and IBD: TH1 and TH17 responses promoting autoimmunity contribute. Autoreactivity in up to one third of the patients with AD. Hypothesis: the development of RA and IBD in subgroups of patients with AD is precipitated by a sustained skin inflammation with increased TH1/TH17 signaling and secretion of proinflammatory cytokines such as TNF-alpha.

Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology [Ferreira and Paternoster, NG, 2017]

- GWAS of three phenotypes (treated as a single trait): 130 loci. 18 loci have more than 2 signals.
- Candidate genes: use coding and eQTL, 132 candidate genes.

- Cell type analysis using S-LDSC: enrichment in multiple cell types, Th1, Th2, Th17, T-reg, CD4, CD8, NK, B cells.
- Pathway analysis using candidate genes: T cell activation, B cell activation, B cell proliferation, isotype switching, IL-2 and IL-4 production (IL-2: T-cell activation, IL-4: Th2 cells).
- Comparison with other traits (Table S23): (1) Strong genetic correlation with asthma, r = 0.7. (2) Significant with obesity: 0.2. (3) Correlation with IBD/CD, r = 0.13, p = 0.004. With UC, not significant, r = 0.07.

Interrogation of human hematopoiesis at single-cell and single-variant resolution [Ulirsch and Sankaran, NG, 2019]

- Data: UKBB, blood cell related traits. Fine-mapping: on 3Mb regions, FINEMAP. ATAC-seq data in 18 cell populations.
- Summary of fine-mapping results: 38K variants with PP > 1%, and 1000 regions with one variant with PP > 0.5 (Figure 1CD). Enrichment in enhancers, coding, but very modest in UTRs.
- Finding cell types: clustering of cell type specific chromatin accessibility of fine-mapped SNPs. Show several clusters.
- Molecular mechanisms of fine-mapped variants: (1) Coding variants: suggest genes of Mendelian diseases. (2) Motif disrupting variants in ChIP-seq targets (2000 ChIP-seq data): 145 instances. A small number of instances per TF, about 1-2-fold enrichment than chance. Top motifs: SPI1, SP1, EGR1, KLF1, MAZ, MYC, JUND.
- Target genes of fine-mapped variants: (1) Hi-C from monocytes. (2) ATAC-RNA correlation in 16 cell types.

A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases [Zhu and Liming Liang, NG, 2019]

- Data: allergy, asthma, eczema, hay fever of UKBB.
- Genetic correlation among immune traits: high correlation among these traits, but not with IBD.
- Shared loci by cross-trait meta-analysis: found 38 shared loci.
- Tissue enrichment: for shared genes, skin shows highest enrichment, then whole blood, vagina, esophagus and lung.
- Model of shared and distinct genetics of immune diseases: allergic diseases, IgE mediated hypersensitivity. RA: immune-complex mediated hypersensitivity. IBD: delayed cell mediated hypersensitivity.
- Why skin and other tissues are enriched of shared genes? (1) Why skin: e.g. FLG gene: function of barrier. Mutations of FLG: sensitive to external allergens and dry skin. This can activate allergic immune response to many organs via blood. (2) Why other epithelial tissues? Share similarity in functions, e.g. epithelia lining in lung and skin.
- **Lesson**: The effect of genetic variants may be very indirect: e.g. a variant acts on skin, which changes permeability and in turn affects immune reactions.
- **Lesson**: different types of autoimmune and allergic diseases, mediated by different molecules/cell types.

Landscape of stimulation-responsive chromatin across diverse human immune cells [Calderon and Pritchard, NG, 2019]

- Data: 25 primary immune cell types from 4 donors, in both resting and stimulated states, also thymus cells. T cells simulated by cross linking TCRs and co-stimulatory receptors, monocytes by LPS and NK cells with cytokines (IL2, etc).
- Stimulation response in terms of OCR changes: large changes in B and T cells, but not innate immune cels. Also stimulation mostly increases OCR and expression in T/B cells.
- Calling ASC variants: WASP for mapping bias, then do binomial test, and FDR ¡ 0.1. Found 607 ASC on average per sample.
- Contribution of TFs to ASC: motif enrichment on OCRs find some cell-type specific TFs, e.g. B-ATF. Motif break type of analysis of B-ATF targets: in resting states, motif breaks do not correlate with allele imbalance; in stimulated cells, show correlation (Figure 4b: comparison of allele imbalance across heterozygous SNPs).
- Cell type specificity of ASC variants: For ASC in one cell type, three cases in the second cell type: shared, inaccessible chromatin and accessible chromatin but different effect (Figure 4c). Estimate the proportion of three scenarios (Figure 4d).
- Enrichment of AID h2g in stimulated OCRs: control tissues, calf muscle, breast epithelium. S-LDSC enrichment: progenitor cells not much enriched comparing with controls, resting PBMC higher and stimulated even higher (Figure 5a). The enrichment is widespread across multiple cell types. Clustering peaks: clusters of shared and cell type-specific peaks show enrichment (Figure 5c).
- Stimulated ASCs vs. resting ASCs: similar enrichment in GWAS (suggesting both are important); however, in eQTLs, resting ASCs more enriched, because the current eQTL data have less stimulated cells (Figure 6).
- Case study (Figure 7): candidate risk variant of RA and UC, OCR and ASC in stimulated T cells, but not other cells. SNP changes motif of NF-kappaB1, and confirm the effect on TF binding (ChIP-seq).
- Lesson: it is sometime difficult to define cell-type specific features (expression, OCRs, etc). Instead of performing DE or DA analysis, use clustering analysis to group features, and identify groups with shared, or cell-type specific patterns.
- Lesson: (about presentation) emphasize the novelty of findings, stimulated states are important for genetics. Shown this in multiple ways: h2g analysis, lack of representation in current eQTL data, case study.

The impact of proinflammatory cytokines on the -cell regulatory landscape provides insights into the genetics of type 1 diabetes [NG, 2019]

- Background: T1D genetics, about 60 loci mapped. Enrichment: T- and B-cell enhancers, but not much in beta cells.
- Experiment: human islet cells (HI) and beta cell lines (EC), treated with IL-1beta and IFN-gamma. ATAC, H3K27ac, RNA, DNAm, UMI-4C.
- Identification of cytokine induced regulatory elements (IREs): 3,800 OCRs that gain H3K27ac upon cytokine - called IREs. Most IREs map to distal regions. IREs associate with expression and protein changes.
- Groups of IREs: opening IREs: 2.4K, both OCR and H2K27ac induced by cytokine (most are not detectable by ATAC). Primed IREs: 1.3K, already open (primed) before cytokine.
- TF occupancy of two groups of IREs: both enriched with INF-gamma response elements and targets of STAT and NF-kappaB (all involved in response). Primed IREs: also enriched with islet lineage specific TFs.

- DNAm of IREs: low CpGs, and not changed much by cytokines. Neo-IREs (newly activated): demethylation by cytokine treatment.
- Model of IRE changes (Figure 2f): Primed IREs: already open, occupied by islet-lineage TFs, and low CpG, and upon activation, binding by cytokine response TFs and activate gene expression. Neo-IREs: low accessibility and high DNAm before stimulation, and activated by cytokine response TFs.
- Changes of chromatin interactions by UMI-4C: 13 regions (IREs), show evidence of more chromatin interaction upon stimulation.
- Enrichment of GWAS loci: (1) T2D: using index SNPs in LD proxy, enriched in stable regulatory elements (SREs), but not IREs. (2) T1D: enriched in IREs, but not SREs.
- Specific GWAS variants: one SNP, likely to be causal based on GWAS, overlaps with IRE. Validated by allele-specific reporter stimulated by cytokine. Use 4C to map target gene, TNFSF18, whose expression is induced by cytokine.
- Lesson: stimulation induced changes of epigenome and transcriptome: (1) Different groups of CREs show different dynamics/epigenome changes: stable, primed and neo (newly activated). (2) Different TFs likely contribute to behavior of different groups.
- Remark: in the absence of hi-C, we can potentially use gene expression changes to identify targets of IREs.

8.5 Neuro-Psychiatric Traits

News Feature: Better models for brain disease [PNAS, 2016]

- Mouse model: (1) Criticism of mouse models that are based entirely on behavior signs. (2) Mouse model carrying disease mutations. SHANK3 study (Guoping Feng): autism-version of Shank3, weakened neuronal signaling in the striatum during early development (area involved in repetitive behavior); SCZ-version of Shank3, reduced signaling in the medial prefrontal cortex in later development.
- Monkey model: MECP2 study.
- iPSC neurons: (1) bipolar study: iPSC neurons from bipolar patients, hyperexcitability. (2) Autism study: from patients with large brain, show overproduction of inhibitory neurons. Limitations: only neurons of very early developmental lineages.

X-linked diseases [ELS: Chromosome X: General Features]:

- Rett Syndrome: one of the most common causes of mental retardation in females. Rett syndrome is a dominant X-linked disease, so affected females are heterozygous. Affected males are rare because it is usually lethal in early male development. Rett syndrome is caused by mutations in the gene MECP2, which is involved in X-chromosome inactivation.
- Fragile X syndrome is a common X-linked mental retardation condition (one in 4000 males) with an unusual inheritance pattern. The disease is characterized by moderate to severe mental retardation, prominent jaw, large ears and high-pitched jocular speech. It is caused by mutation of one region of the X chromosome - the amplification of a short triplet repeat of CGG. The amplified triplet repeat becomes highly methylated and disrupts expression of the nearby gene fragile X mental retardation 1 (FMR1), leading to the disease. The inheritance of fragile X syndrome is unusual and complex because the repeats are unstable and can be amplified in germ cells or tissue.

The inheritance of Tourette Disorder: A review [Pauls & Scharf, J Obcessive-Compulsive and Related Disorders, 2014]

- TD is a genetic disease:
  - Aggregation studies: the risk of TD in first degree relatives in 10-100 times higher than general population.
  - Twin studies: MZ concordance rate of 50-77%, while DZ has 10-23% rate.

  Comorbility of TD: with OCD and ADHD.
- Linkage analysis: only one site is confirmed, in the gene HDC, a rate-limiting enzyme in histamine (HA) biosynthesis.
- Genome rearrangement and rare CNVs: some regions have been identified from cytogeneti abnormalities, sometimes overlapping with autism regions/genes. Three CNV studies with a few hundred to 1,000-2,000 samples (cases and controls): overlapping regions with ASD, and HA signaling. In the largest CNV study, found 3.3 fold burden of large deletions in recurrent pathogenetic CNVs in subjects with other neurodevelopmental disorders.
- GWAS: first GWAS of 1285 cases and 4964 controls, found no genome-wide significant loci, but the top signals enriched for brain eQTL. Heritability explained by common SNPs: 0.58 out of total of 0.6-0.8. TS and OCD have estimated genetic correlation of 0.41 (using two GWAS data).
- Summary of findings: HA signaling in TD, supported by (1) LoF mutation in HDC in a dense pedigree; (2) CNV studies found overrepresentation of HA signaling pathway; (3) overtransmission of SNPs in HDC region in 520 nuclear families.

Brain Expression Genome-Wide Association Study (eGWAS) Identifies Human Disease-Associated Variants [Zou and Ertekin-Taner, PLG, 2012]

- Data: 197 subjects with Alzheimer's disease (AD) neuropathology and 177 with other pathologies (non-AD), in cerebellar tissues and in temporal cortex.
- eGWAS analysis: analyzed the ADs and nonCADs separately on cisSNPs. The direction and magnitude of associations in both groups demonstrate remarkable similarities (Pearson's correlation coefficient = 0.98, $p < 0.0001$). Found 10,281 total eQTL (1,875 unique genes) in the combined datasets.
- Effect size: We found that the "best" cisSNP explained a median of 3% of the expression variation. For the top 746 probes, the "best" cisSNPs accounted for a median of 18% of the expression variance.
- eQTL overlap across brain regions: the top 2,980 cerebellar eGWAS associations were followed up in the temporal cortex validation study. We found that 2,685 top cerebellar cisSNP/transcript associations could be tested in the temporal cortex. The effect sizes are also highly correlated, Pearson correlation 0.94.
- Among 2,596 top cerebellar eGWAS cisSNP: identified 47 cisSNPs that were also associated with 36 diseases/traits in GWAS catalog, 2.4-fold enrichment than expected by chance. The results include both brain disease (Parkison, ADHD) and non-CNS diseases such as SLE.
- Intersection with GWAS of AD (ADGC data): SNPs with suggestive AD risk association ($p_{meta} < 1e-3$).
  - Imputation of SNPs in the current data: 77,126 cerebellar (63,652 unique SNPs, 2,338 unique genes) and 68,172 temporal cortex (57,922 unique SNPs and 2,201 unique genes)
  - 380 cisSNPs that were significant for the cerebellar transcript associations and also had suggestive AD risk associations (2.9-fold enrichment), 432 such temporal cortex cisSNPs (3.3-fold enrichment) and 356 cisSNPs significant in both the cerebellum and temporal cortex (2.7-fold enrichment

  - Did not identify strong transcript associations for some of the top genes recently implicated in AD risk in large LOAD GWAS studies

- Remark:
  - Significant overlap of eQTL across different phenotypes and different brain regions
  - The top cis-eSNP usually different from the top genes found in GWAS.

Genome sequencing identifies major causes of severe intellectual disability [Gilissen and Veltman, Nature, 2014]

- Comparison of diagnostic yield of ID: array, 12%; WES, 27%; WGS, 42%.
- Data: WGS of 50 patients with severe ID and their unaffected parents, an average genome-wide coverage of 80 fold.
- De novo mutation rates: 82 high-confidence potential de novo SNVs per genome. A protein-coding de novo substitution rate of 1.58, higher than all previous estimate, and a total of 84 in coding regions.
  - Enrichment of LoF mutations among 84 de novo coding mutations: $p = 1.6 \times 10^{-5}$.
  - Enrichment of de novo coding mutations in ID gene sets: 528 known ID genes (from HGMD and other sources) and 628 candidate ID genes. 9 genes, $p = 0.04$.
- De novo CNVs: 8 were detected, 4 of which in known ID or candidate ID genes (significant).
- De novo non-coding mutations: 43 in promoter regions (1), introns (38), splice site (1) or untranslated regions (3) of all known ID genes. Found no potential pathogenic mutations using ENCODE annotations (only one promoter has ENCODE annotation).
  - Effect on splicing was determined using Alamut software that integrates a number of prediction methods for splice signal detection as well as exonic splicing enhancer (ESE) binding site detection.
  - ENCODE annotation was based on Chromatin state segments of nine human cell types (Broad ChromHMM) and transcription factor binding sites (Txn Factor ChIP).
- Inherited variants: a single proband with compound heterozyous deletions affecting the VPS13B gene.
- Lesson:
  - Measure the value of WGS by **diagnostic yield** (the percent of patients whose genetic causes can be determined).
  - Mutations in introns could affect splicing.

De novo mutations in schizophrenia implicate synaptic networks [Fromer & ODonovan, Nature, 2014]

- Data: 617 Scz trios.
- De novo mutation rates:
  - Overall, the de novo rates in LoF, missense are not higher than controls (731 controls from published data sets)
  - Loss-of-function de novo mutations were more common in patients with relatively poor school performance ($p = 0.018$). Note: psychiatrists were explicitly instructed to exclude people with known intellectual disability.
- Recurrent genes: 18 recurrent genes (both missense and LoF): more than expected, $p = 0.03$. A single double-hit (LoF) gene: TAF13.

- Inherited variants:
  - Transmission: excess of transmission of nonsyn. singletons in de novo genes, $p = 0.01$.
  - Case-control: increased case-control ratio of rare (MAF $< 0.1\%$) LoF in de novo genes, $p = 0.003$.
- Gene set enrichment:
  - Enrichment of de novo mutations in candidate gene sets: ARC, NMDAR, FMRP targets.
  - GO set enrichment: a single set, assembly of actin filament bundles, is significant.

A polygenic burden of rare disruptive mutations in schizophrenia [Purcell & Sklar, Nature, 2014]

- Data: WES of 2,536 schizophrenia cases and 2,543 controls. Focusing on 2,500 genes implicated by unbiased, large-scale genome-wide screens, including GWAS, CNV and de novo SNV studies.
- QC procedures:
  - Filter 11 subjects with low-quality data along with likely spurious sites and genotypes. Per individual, 93% of targeted bases were covered at$\geq$ 10-fold (81% at $\geq$ 30-fold).
  - Cases and controls had similar technical sequencing metrics, including total coverage, proportion of deeply covered targets, and overall proportion of non-reference alleles.
  - 635,944 coding and splice-site passing variants of which 56% were singletons. High specificity and sensivity were verified.
- Individual variant analysis: a known common SNP. Gene-best test: SKAT results no significant gene; genic burden test with LoF variants, one gene with 10 in cases and 0 in control, however, not significant ($p = 1.7 \times 10^{-3}$).
- Define variant groups: (1) by function: LoF, strict damaging missense (predicted by five algorithms), broad damaging missense (by only one algorithm). (2) By AF: private, rare ($< .1\%$) and up to .5%. Total of 9 groups.
- Rare variant burden analysis: in 2,500 genes (Table 1).
  - LoF variants: (1) rare: 1,547 in cases vs 1,383 in controls, OR = 1.12, $P = 10^{-4}$. (2) Singletons: enrichment, $P = 8E - 4$. (3) Up to .5%: $P = 2E - 4$.
  - Missense variants: enrichment of strictly defined missense variants $P = 1.5 \times 10^{-3}$, but not broadly defined ones.
- Gene set burden analysis: consider 12 sets from de novo, GWAS and CNV studies. Focus on LoF variants at 3 different thresholds. Eight out of 12 sets were nominally significant. Three smaller sets (synatpic genes) have OR > 5.
- Comparison with autism/ID genes: e.g. de novo LoF genes in ASD. Found no enrichment in these gene sets.
- Refined burden analysis on 1,796 genes comprising all members of the most prominently enriched sets (Figure 1).
  - LoF: signal largely driven by novel/singleton variants. Signal is stronger in high expression genes.
  - NS variants: strict definition clearly better than PPH. Among all annotations, PPH and SIFT seem to perform better.

Noncoding Variation in Schizophrenia [Roussos & Sklar, Cell Reports, 2014]

- Data:

  - Brain eQTL: combine 8 published datasets (Table S1).
  - Brain CREs: K3K27ac, H3K4me1, DHS from ENCODE and Roadmap, including adult brain, fetal brain, primary cell culture and iPS. Use all datasets to define five types of CREs: active promoter; active enhancer; poised promoter; repressed enhancer; and open chromatin regions.

- GWAS SNPs and functional annotations:
  - Choose all GWAS SNPs at $p < 10^{-3}$ (42K SNPs), 37% are eSNPs. Among this, 4.9% were in active promoters, 9.6% in active enhancers (H3K4me1 and [H3K9ac or H3K27ac]), 3.5% in DHSs, 1.0% in poised promoters, and 1.5% in repressed enhancers.
  - Enrichment: highest in eSNPs (3.68 fold), active enhancers (2.30) and active promoters (2.13). However, active enhancers and promoters together cover about 10% of 42K GWAS SNPs at $p < 10^{-3}$.
  - cre-SNPs: eSNPs in cis-regulatory elements (CREs). At $p < 10^{-3}$, we have about 2,000 creSNPs; At $p < 10^{-5}$, about 400 creSNPs. Enrichment is strongest in creSNPs: about 5 fold at $p < 10^{-3}$, and 29 fold at $p < 10^{-5}$.

- Finding causal loci in GWAS loci:
  - 22 significant SNPs. 10 out of 22 overlap with eSNP, and have high RTC scores (test if eSNP and GWAS SNP tag the same causal variant).
  - Example: SNP in CACNA1C region, also its eSNP. Find 4 SNPs in perfect LD with the GWAS SNP, and lie within two predicted enhancers. Confirm the interaction of enhancer and promoter through 3C in human dorsolateral prefrontal cortex and iPS-derived neuron.

- Lessons: High fraction of GWAS SNPs in eQTL, much higher proportion than in CREs. Also active enhancers (H3K27ac) appear to cover more functional sequences than DHS.

Schizophrenia genetics complements its mechanistic understanding on Sekar et al, Nature, 2016 [NN, 2016]

- GWAS evidence of C4: (1) strongest signal in MHC near C4; (2) another hit in CSMD1, which regulates C4.

- C4 expression model: CNV in C4, associated with C4 expression. Also influenced by genotype. Build a model to predict C4 expression from genotype and CNVs.

- Correlation of predicted C4 expression and SCZ risk.

- Functional evidence of C4: (1) Expression elevated in SCZ brain. (2) Complement system: critical for synatpic pruning.

Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders [Singh & Barrett, NN, 2016]

- Study design: (1) Case-control: 1,000 cases in UK10K and 2500 from published Swedish study. (2) Trios: 1000 from 7 published studies.

- Enrichment of DNMs: no burden in missense mutations, only in LoF. General burden trend: SCZ < ASD < DD.

- Case-control: (1) Use LoF, LoF + damaging missense at different AF, run burden test and SKAT. No signal. (2) Found enrichment in rare LoF.

- Combining DN and case-control, using only LoF mutations. TADA or Fisher' test. Single gene: SETD1A.

- Biology of SETD1A and other evidence: LoF clustered in H3K4methylation domain. Top 3% constrained gene in ExAC. LoF in DDD.

Sparse whole-genome sequencing identifies two loci for major depressive disorder. CONVERGE, Nature, 2016

- Background: no loci of MDD found previously (9k cases). Likely due to heterogeneity of the genetics.
- Data: 5000, low coverage WGS (1.7x), Chinese women and 5k controls. 6M SNPs.
- Reducing genetic heterogeneity: Chinese women, only recurrent cases (more severe). Known risk factors recorded. In China, the MDD cases tend to be more severe (reluctance of reporting).
- Two loci reaching genome-wide significance: SIRT1 and LHPP. Replicated in 3k samples.
- Comparison with PGC results: the top SNPs not replicated, but direction yes. PGC polygenic risk score is significantly associated, $p < 0.01$, but explains only 0.1% of MDD risk. The most strongly associated loci have low AF in European: > 40 in CONVERGE vs. 3 and 8% in European.

Comprehensive integrative analyses identify ALMS1, GLT8D1 and CSNK2B as schizophrenia risk genes [Yang & Luo, review for Biological Psychiatry, 2017]

- Sherlock analysis of brain eQTL (Meyers, NG, 2007) and PGC GWAS: 10 risk genes at Bonferroni threshold, 6 cis and 4 trans.
- Expression patterns: most of these genes are expressed in brain (not surprising, they are found by brain eQTL); and most (6) genes have higher expression in early developmental stages.
- Network analysis: DAPPLE analysis on PPI network (GeneMania) found modest evidence of ALMS1 and CSNK2B. Co-expression network: some genes are co-expressed with known SCZ risk genes.
- Additional evidence on three risk genes: dysregulation in SCZ, independent brain eQTL, association with hipocampal structure (ENIGMA data) and cognitive traits.
- Experimental validation of GLT8D1 using NSC: promotes the self-renew and proliferation abilities of neural stem cells (NSCs) and inhibits NSCs differentiation.
- **Lesson**: It is OK to find genes using computational tool, then focus on several candidate genes and support them with multiple lines of evidence - we do not have to show that every new predicted gene is real.

Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior [Doan and Walsh, Cell, 2016]

- Evidence that HARs are functional: (1) enrichment in CTCF binding. (2) Epigenomic marks suggest that 29% act as enhancers in brain, heart and limb development. (3) HAR regions show enrichment of loci associated with SCZ.
- HARs are depleted of variants in human population.
- HARs are enriched of neural CREs: using Roadmap data.
- Many HARs act as regulatory elements of dosage-sensitive neural genes.
- Connection between HARs and ASD: de novo CNVs, point mutations in the HARs are enriched in ASD probands vs. controls.
- Lesson: to demonstrate functional connection with cognition and social behavior (1) regulation of neuronal genes; (2) association or enrichment with neuropsychiatric phenotypes.

Model of complex social behavior (vocal learning) using songbird or zebra finch [Somayeh Ahmadiantehrani from Sarah London lab, 2017]

- Background: 50% of monogenic ASD cases result in the disruption of mTOR cascade, e.g. PTEN and TSC2.
- Background: mTOR protein complex: ser/thr kinase. Control translation, lipid biosynthesis, autophage, etc.mTOR disruption in brain: blocks LTP and learning/memory deficits.
- Background: Some known brain region are associated with song learning.
- Experiment: isolated bird, then song from the speaker. Measure phos-S6 (downstream effector) in CMM, NCM.
- Treatments with drugs disrupting mTOR pathway reduces the similarity of junavile song vs. tutor. This is only effective when treating before tutoring experience.
- In vivo eletroporation (instead of virus) to change gene expression in bird brains. Proposal: PTEN knockdown.
- Q: Why treatment of drugs during development has no effect?
- Q: Song learning and language learning: evolutionary conserved?

De novo mutations in regulatory elements in neurodevelopmental disorders [Short and Hurles, Nature, 2018]

- Background: Targeted sequencing: (1) Capture arrays; (2) PCR on a set of primers.
- Targeted sequencing of 4.2Mb noncoding sequences: conserved sequences (4000), heart enhancers, VISTA enhancers (600). Sample size: 8000, majority do not carry a diagnostic variant in a protein coding gene (exome-negative).
- Quantifying constraint in human population: mutability (tri-nucleotide model) adjusted proportion of singletons (MAPS, from ExAC paper [Lek, Nature, 2017]). Estimated from 8000 unaffected DDD parents.
- Pattern of negative selection: (1) Heart enhancers: very little conservation (across species) and not constrained in human population. (2) VISTA (experimental) enhancers: large range of cross-species conservation, constrained in human (similar to exonic sequences). (3) Comparison of DHS constraint (in human): most constrained are fetal brain, HSC, ESC-derived neurons. But difference across tissues not statistically significant.
- Burden analysis: using mutation models. (1) Small burden in conserved sequences, $p = 0.04$, no burden in heart or VISTA enhancers. (2) Conserved elements + fetal brain chromHMM, $p < 1E-3$, 238 vs. 194 (expected). Conserved + fetal brain DHS, $p = 0.002$. Even higher burden in patients with neurodevelopmental phenotypes. (3) Control: no burden in same elements in patients without ND phenotypes.
- No enrichment in enhancers of known DD genes or pLI genes. To assign enhancers to genes: fetal Hi-C, DHS expression correlation. Remark: likely due to small DNM counts.
- Motif analysis: 45 TFs whose motifs are enriched in DNMs (DNMs increase the affinity).
- Recurrently mutated elements: limit to fetal brain active CNEs and evolutionarily conserved enhancers. The number (30) is twice higher than expected. Create clusters of fetal brain CNEs, but found no enrichment of individual clusters.

- The proportion of highly penetrant mutations is small: simulation of highly penetrant mutations (RR = 120), and assess the number of recurrent elements that are genome-wide significant. Since we observe 0, the proportion of highly penetrant mutations must be low.

Genome-wide association studies of brain imaging phenotypes in UK Biobank [Elliott and Smith, Nature, 2018]

- Background: (1) structure MRI: structural volumess, e.g. hippocampus, and other biomarkers such as microbleeds, white matter microstucture. (2) Diffusion MRI: brain structural connectivity. (3) Functional MRI: resting state activities and functional connectivities.
- Data: 3000 Image-derived phenotypes (IDPs) in 8K individuals from UKBB.
- Heritability: about half show significant heritability, mostly in 0.1-0.5. Resting state fMRI features show lowest heritability.
- Significant associations: at stringent thresholds, about 300 associations. Some associations across multiple phenotypes. (1) Association of genes involved in iron deposition, associated with some dMRI features. (2) SNP in iron channel associated with grey matter volume effects across the brain. The SNP is also associated with IQ, blood pressure and SCZ. (3) EM and EGF signaling associated with dMRI IDPs.
- Genetic correlations: suggestive correlations with ALS, SCZ and stroke with dMRI features in white matter tracts.

Risk loci for ADHD [NRG, 2018]

- 12 loci for ADHD, including FOXP2 for development, DUSP6 regulating dopamine levels.
- Genetic correlation: positive with MDD, neuroticism, and negative with intelligence and education.

The genetic and epigenetic basis of pediatric epilepsy [Gemma Carvill seminar, 2019]

- Epilepsy: recurrent and unprovoked seizures. About 1/26, wide spectrum.
- I. CHD2 mutation patterns in patiens vs. gnomAD: controls no LoF, and depletion in ATPase and helicase domains.
- Mouse phenotypes: het. CHD2, learning deficits, leions on multiples organs (no brain). No seizure.
- CHD2 binding: enrichment with promoters of Epilepis genes. half are pormoters.
- CHD2 phenotype in NPC: 1000 up-regulated genes, NPC diff. much faster. Targets: axongenessis (FE = 7), other migration. However, no DEs in neurons.
- Conclusion: premature neuronal differentiation. Neurons fire more ferquently. Organoid model: comparison of w.t. and hets.
- Q: if chr. remoderler, deletion of the gene should lead to down-regulation, but this is opposite to what is observed. Why?
- II. Dravet syndrome (form of DEE). 90
- SCN1A: poison exon in an intron. Expression in astrocyte (to degrade SCN1A).
- Mechanism: intronic variants disrupts SRSF1 binding sites. Increased inclusion of poison exon. Motif: GGAGGA.
- Poison exons are presents at other epi. genes.

- Use ASO (anti-sense oligo) to target poison exon: stop poison exons to increase expression of genes.
- Lesson: chromatin genes may change the development/cellular phenotype of cells.

Genome-wide association analysis of Parkinsons disease and schizophrenia reveals shared genetic architecture and identifies novel risk loci [review for Biological Psychiatry, 2019]

- Background: it is possible that two traits share some loci, but because some of them have same direction, some opposite, overall, the genetic correlation is 0.
- CondFDR analysis of SCZ and PD: increase the power if discovery. 9 distinct loci jointly associated with both SCZ and PD. Five of them have consistent effect directions.
- Validation of shared loci: nearest genes. Some show highest expression in astrocytes, endothelia cells and neurons. PPIs are more connected in astrocytes than in neurons.

Genetics of Verbal communication [Mellissa DeMille]

- World-wide language: by phonemes or specifically number of consonants. Ex. two African: 18 vs. 46.
- Remark: fitness of phonemes is possible, e.g. separation of t vs. d.
- Three aspects of languages: Speak, hearing and processing/phonological processing (convert sounds to meaning). Genetics: FOXP2, hearing/deafness gene, DCDC2 (associated with reading disability). DCDC2: localized in relevant brain regions.
- Consonants: precision of AP firing. Vowels: number of AP firing. DCDC2 K/O mice: poor precision of AP.
- READ1: CRE of DCDC2. Repeat alleles, complex.
- READ1 variation (RU1.1) correlates with number of consonants across many populations. Use PCs to adjust for relatedness.
- Future directions: polygenic selection? When evolved?
- Remark: is LMM sufficient to address for confounding - genetics is now confounded with population. Populations also differ in culture.
- Q: genetic sharing with education attainment? Reading/language genes.
- **Lesson**: genetic basis of language is hard, however, we can break it down in different parts, from vocalization to hearing. For each component, it is easier to relate molecular functions of genes to phenotypes, e.g. detection of sound patterns.

8.5.1 Autism

Biology of autism:

- Biological basis of brain region specialization: two hypothesis:
  - Mainly reflect the difference of neuron behavior or history, i.e. the neurons in different regions are largely interchangable, but they have different history. Evidence: most genes do not show significant difference in expression across regions in neurocortex.
  - Different expression programs are active in different regions. Evidence: neurotransmitters (and corresponding signaling pathways) that are specific to functions, e.g. oxytocin and vasopressin for social behavior.

- Focal vs. diffusive disruption: does autism involve disruption of specific brain circuits? Yes to some extent (frontal lobe, temporal lobe, etc.). But then how to explain that the mutaiton of some broadly expressed genes can lead to disruption in specific areas?

Review of Autism-related genes [Holt & Monaco, Links between genetics and pathophysiology in the autism spectrum disorders, EMBO Mol Med, 2011]:

- Candidate gene studies: approximately one hundred are reported as showing at least nominal association with ASD.
  - The serotonin transporter gene SLC6A4: serotonin levels have been shown to be altered in autism cohorts. Multiple studies have also identified association to variants in this gene.
  - An important example of linkage data being utilized is CNTNAP2, with further evidence from a subsequent GWAS. CNTNAP2 is a neurexin, and there is strong evidence that this gene family is influential in ASD. It is expressed in brain regions related to ASD, and there is evidence from Copy number variations (CNVs) identified in schizophrenia. Brain changes in patients with CNTANP2 mutation, suggesting underconnectivity.
  - SNPs in promoters and intron of MET: affect TF binding (SP1 and other TF). MET is also associated with schizophrenia, and again, the associated SNPs appear to affect gene expression
  - Four LRR (leucine rich repeat) genes are enriched in the brain and two show significant associations with autism, LRRN3 and LRRTM3. There are 313 members in this class of genes, some of which have been implicated by GWAS, but do not reach genome-wide significance thresholds.
- GWAS:
  - Wang et al found strong association to a locus between Cadherin 9 (CDH9) and Cadherin 10 (CDH10). Both genes encode neuronal cell adhesion molecules, and CDH10 is expressed in the frontal cortex, a region of the brain associated with ASDs. The SNP is likely influence gene regulation.
  - Weiss et al found significant association to a SNP between SEMA5A and TAS2R1. SEMA5A has been implicated in axonal guidance and is expressed at lower levels in cell lines and brains from individuals with ASD. The associated SNPs were 80kb upstream of SEMA5A, consistent with a role in gene regulation
- CNVs:
  - Sebat et al demonstrated a significant difference in frequence of de novo CNVs between sporadic cases (10%), familial cases (3%) and controls (1%)
  - Pinto et al: PTCHD1, likely to be involved in development of the cerebellum, and SHANK2 (Noor et al, 2010; Pinto et al, 2010). SHANK2 is related to SHANK3 (Durand et al, 2007; Moessner et al, 2007), which encodes a scaffolding protein located at synapses in the brain, since implicated in other studies
- Sequencing and point mutations: in SHANK3, Durand et al identified mutations in individuals with ASD including a single base insertion resulting in a frameshift, rare non-coding mutations not present in controls, as well as CNVs of potential significance.
- Additional pathways implicated in ASD: e.g. genes involved in synthesis and degradation of ketone bodies. The latter could affect γ-aminobutyric acid (GABA) levels.

Focal vs diffuse circuit disruption in ASD [Autism: Many Genes, Common Pathways? Geschwind, Cell, 2008]:

- Challenge: ASD susceptibility genes must converge on the disruption of function in brain regions supporting language, social cognition, and behavioral flexibility. What kind of expression pattern underlies the relative specificity in brain regions?
- Scenarior 1: focal gene expression of the specific gene product during development; when the risk allele is expressed, there is disruption of the cortical and subcortical brain networks supporting social responsiveness or language. Example: CNTNAP2, enriched in anterior regions of the developing cerebral cortex that overlap with circuitry involved in the development of joint attention (related to language)
- Scenarior 2: most known ASD susceptibility genes do not demonstrate regionally restricted expression. The core areas affected in autism involve integration of information from multiple, higher-level areas. Such functions could be easily perturbed by minor, but widespread disruptions in neural transmission.
- Evidence of Scenarior 2: one would expect to find subtle, widespread differences in many brain systems, many of which may not be the direct cause of the core features of autism. Such abnormalities, may explain the differences in sensory processing, motor function, and sensory-motor integration, etc. that have been variably associated with ASD.
- Concepts of focal versus diffuse circuit disruption are not mutually exclusive and both may cause different forms of ASD.

Brain regions affected by ASD [Connecting genes to brain in the autism spectrum disorders. Arch Neurol, 2010]:

- Background: ASDs can be conceptualized in terms of multiple genetic etiologies that disrupt the development and function of brain circuits mediating social cognition and language.
- Lessons from autism-related Mendelian/genetic diseases:
  - Fragile X syndrome: accelerated early head growth. Structural imaging found abnormalities in the caudate (increased), lateral ventricles (increased), and posterior vermis of the cerebellum (reduced). Functional MRI: involvement of frontostriatal circuitry (connecting frontal lobe regions with the basal ganglia).
  - Rett syndrome: cerebellar atrophy. Similarly, frontal and temporal cortices, the caudate are subject to the greatest regional reductions in gray matter volume.
  - Mutation in CNTNAP2: cortical dysplasia and abnormalities in neuronal migration provide support for an early developmental insult, particularly in the frontal and temporal neocortex.
  - Tuberous sclerosis: a comparison of IQ-matched patients with tuberous sclerosis with and without a diagnosis of autism identified altered energy metabolism in temporal neocortex, caudate, and cerebellum among ASD cases.
  - Together, these results suggest that these different monogenic risk factors for autism share a common involvement of frontal and temporal neocortex, caudate, and cerebellum.
- Lessons from idiopathic ASD: accelerated postnatal growth in defined brain regions and preferential involvement of specific regions, including frontal and temporal cortex, cerebellum, and amygdala.
- Neuropathological findings in ASD:
  - The most consistent neuropathological finding among the ASDs is the observation of errors in neuronal migration, particularly in frontal and temporal lobes.
  - Minicolumns: distance is reduced in the frontal and temporal cortex of individuals with a spectrum condition. Minicolumn findings are predominant in the frontal and temporal cortex, consistent with expression patterns for known ASD genes CNTNAP2 and MET.

  - The amygdala: involved in the modulation of social behavior, has long been implicated in the ASDs. Neurons in the amygdala were abnormally small and showed elevated packing density.

- Conclusion: connectivity between frontal, temporal, and additional interconnected regions mediating language and social behavior is critical to understanding the ASDs

Underconnectivity hypothesis [Links between genetics and pathophysiology in the autism spectrum disorders, Holt & Monaco, EMBO Mol Med, 2011]:

- Brain changes of ASD patients in a gross scale: localized failure of cerebellar development, cerebral cortical abnormalities, altered amygdala development and decreased corpus callosum size. Excessive growth of white matter in the first 2 years of life, followed by undergrowth, leading to macrocephaly in 20% of children with ASD.

- Brain changes of ASD patients in a fine scale: within minicolumns, neurons smaller in size, but of increased density, and their organization within minicolumns being altered. Minicolumns are of decreased width and increased number, resulting in the increase in density of neurons reported. Changes in minicolumns and neuron size may promote shorter connecting fibres, increasing local connectivity at the expense of connections between different cortical regions.

- Underconnectivity theory: fewer long range connections, leading to decreased synchronization between regions of the brain and decreased global information processing.

- Underconnectivity is particularly likely to affect connections between frontal and temporal lobe regions, and what connectivity there is, is likely to be unorganized. Neuroimaging studies have demonstrated a lack of synchronization and decreased communication and connectivity between them.

- Additional evidence of underconnectivity: individuals with ASD show reduced capacity for joint attention. Joint attention utilizes the prefrontal and anterior regions of the brain.

- Genetic evidence: genes involved in neuronal development, migration, growth and maturation, axonal growth, synapse and synaptic complexity have been implicated.

Genes, circuits, and precision therapies for autism and related neurodevelopmental disorders [Sahin & Sur, Science, 2015]

- Genetics of autism:
  - Approaches for mapping autism genes: recessive inherited variants in consauguinity families; families with high risk among women.
  - Environmental factors that may affect autism: maternal infection (immune activation), perinatal injury: premature infants with cerbellar hemorrhage have 30-fold higher risk, GI symptoms (microbiome).

- Molecular pathways:
  - Synapse: calcium channel (SCN2A), GluR and scaffolding proteins (SHANK, SYNGAP1, PSD95).
  - Transcriptional and translational regulation: MeCP2, FMRP.
  - Signaling pathways: (1) PI3K/mTOR pathway: TSC1, TSC2, PTEN, IGF1 (for treatment). (2) Ras-MAPK pathway.

- Neural circuitry: challenge is to identify cell types, regions, circuits critical for autism.
  - Approaches: conditional knockout, imaging studies (often difficult with autism patients).
  - Cortical projection neurons (extending axons to distant targets) from human co-expression networks.

  - Neuron types from mouse model: e.g. loss of certain neuronal types from deletion of ASD genes. GABAergic neurons, interneurons, inhibitory neurons.
  - Regions: basal ganglia, cerebellum, non-neuronal cells such as astrocytes and microglia.
  - A gene may be expressed in different regions, and control multiple behaviors (Figure 2), e.g. cortex: soical interaction; cerebellum: repetitive behavior.
  - Two subtypes of epilepsy: each involving a different neuronal types, and responding to different treatments.

- Treatment of autism:
  - Mechanism-based: e.g. mGluR antagonist (FXS).
  - Biomarkers for subtypes: e.g. biochemical measure, functional feature from imaging.
  - Preclinical models: e.g. iPS-neurons.

Getting to the core of autism [Iakoucheva and Sebat, Cell, 2019]

- Autism genes that are DBPs and RBPs: bind to other ASD genes, e.g. TBR1, CHD8, FMRP.
- Trans- effects of ASD genes: CHD8 and FOXP1 in iPSCs, strongest targets are trans, but not direct targets. Mouse models of developing brain: SETD5, TBX1 and FOXP1.
- DE analysis from postmortum brains of patients: upregulation of immune-microglia and mitochondrial modules, and downregulation of neuronal and synaptic modules in ASD and schizophrenia.
- Core ASD genes? May be hard to define, because traits do not emerge from a small number of genes. Master TFs and key synaptic genes may be equally important/core.
- Cell proliferation phenotypes of ASD patients: macrocephaly for some genetic subtypes of ASD - e.g. CHD8, PTEN, whereas the opposite phenotype (microcephaly) is associated with others - e.g. DYRK1A. iPSC model: more proliferation, but fewer excitatory synapses and matured into defective neuronal networks with less bursting.
- Dendritic Arborization and Synapse Number: MECP2 mutant reductions in neurite outgrowth, dendritic arborization, and excitatory synapses.
- Change of Electrophysiological behavior of neurons: e.g. SHANK2, KCNQ2.
- Reverse-genetic approach: starting with the genotype and determining how genes influence clinical phenotypes.
- Lesson: ASD genes form regulatory networks. Measurable cellular phenotypes of ASD gene mutations: proliferation, synapse numbers and morphology, electrophysiology.

A genome-wide scan for common alleles affecting risk for autism [Anney, Hum Mol Genetics, 2010]:

- Background: only rare de novo and inherited variants are soundly established genetic risk factors for ASD, and thus far these only account for a small proportion of the total genetic risk. In contrast, common variants rarely have such an impact on risk for any disorder, especially one like ASD that is known to diminish reproductive success.
- GWAS method: 1558 ASD families (4712 subjects). A priori we planned and conducted four non-independent GWA analyses corresponding to data partitions along axes of diagnosis and ancestry: spectrum versus strict and European versus all ancestries.

- Largest associations: in a 300 kb intronic region of MACROD2, $P = 2.1E-8$, (below the threshold of Bonferroni correction). Recent genome-wide studies have highlighted copy number variation at MACROD2 in an individual with schizophrenia (27), brain infarct (28) and brain volume in multiple sclerosis. Also $\tilde{5}$00 kb from the association signal is FLRT3, a cell adhesion molecule with functions in neuronal development.
- Replication: under independent ASD families from the Autism Genetics Resource Exchange (AGRE) database. Most genes are not replicated, e.g. for MACROD2, its $P = 0.13$ in AGRE, and $4.7E-8$ in AGP and AGRE.
- Exploratory analysis: using traits such as verbal and IQ. We do observe signals that are close to the threshold ($P < 1E-7$) in the discovery sample in PLD5, POU6F2 and an intergenic region on 8p21.3.
- Discussion: unbiased estimates of odds ratios detected by GWA studies are typically in the range of $1.1-1.3$ to have good power to detect such effect sizes requires many thousands of samples.

Functional impact of global rare copy number variations in autism spectrum disorders [Pinto, Nature, 2010]:

- Background:
  - Although ASDs are known to be highly heritable ($\tilde{9}$0%), the underlying genetic determinants are still largely unknown.
  - CNV examples include de novo events observed in 5-10% of ASD cases.
- Data: 1,275 ASD cases and their parents using the Illumina Infinium 1M single SNP microarray. 1,981 controls. The array contains a total of 1,072,820 markers (50-mer probes) for SNP and CNV analyses.
- Defining the CNVs: (1) CNV present at $< 1\%$ frequency in the total sample (cases and controls); (2) CNV $\geq$ 30kb in size (because $> 95\%$ of these could be confirmed). This stringent data set of 5,478 rare CNVs in 996 cases and 1,287 controls of European ancestry. Among these CNVs at least 5.6% (49/876) of trio families carried at least sone de novo CNV (average of 1.1 verified de novo CNVs/sample).
- Association test using CNVs:
  - Measures for CNV burden analysis: (1) CNV rate: by the number of CNVs per sample and the proportion of samples with one or more CNVs; (2) CNV size was assessed as both the total genomic segment covered by CNVs, as well as the average CNV size; (3) Gene-count: the average number of genes intersected by CNVs.
  - CNV burden analysis: compare the three measures of CNV burden in cases vs. controls (i.e., hypothesizing that cases will show greater burden of rare CNVs than controls). Permutation procedure for statistical significance of one-sided tests.
- Results of CNV-based gene association test: examples of novel ASD loci include SHANK2, SYNGAP1, and DLGAP2 based on the observation that de novo CNV affects these genes in cases but not controls. Also, a combination of rare de novo and inherited CNVs affecting NRXN1, IL1RAPL1, DMD, and the DiGeorge 22q11.2 region in ASD.
- Results of CNVR burden analysis: CNVR defined by merging overlapping CNVs. CNVR at DDX53/PTCHD1 emerged as a significant ASD risk factor. Specifically, we observed 7 ASD male cases with overlapping deletions at DDX53/PTCHD1 (Xp22.1) and no CNVs were observed at this locus for the initial 1,287 controls
- Analysis of ASD candidate genes:

  - Curated list of ASD candidate genes: (1) ASD-implicated: 36 genes and 10 loci strongly implicated in ASD; (2) Intellectual disability (ID): 110 genes and 17 loci know to be implicated in ID but not yet in ASD; (3) ASD candidates: 103 genes drawn from previous studies of common and rare variants for ASD. They include case reports of cytogenetic abnormalities, allelic association and CNV studies.
  - A higher proportion of cases with rare CNVs overlapping ASD implicated disease genes compared to controls, corresponding to a significant enrichment for genes in this set.

- Gene set analysis:
  - Fisher's exact test to assess which gene sets were more frequently affected by rare CNV events in ASD cases compared to controls.
  - Results: 76 gene sets affected by deletions (2.18% of sets tested) were found to be enriched and used to construct a functional map. Gene sets involved in cell and neuronal development and function (including projection, motility and proliferation) previously reported in ASD were identified. Novel gene sets include GTPase/Ras signalling, with component Rho GTPases known to be involved in regulating dendrite and spine plasticity and associated with intellectual disability.

- Questions/remarks:
  - SNP markers of CNVs: cannot detect multi-allelic CNVs.
  - Testing association of rare CNVs: burden test is essentially pooling/collapsing of all (rare) CNVs within a gene.

Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting [Betancur, Brain Res, 2011]:

- Background:
  - Over 70% of individuals with autism have intellectual disability (ID), while epilepsy occurs in about 25%.
  - ASDs are identified in about 1% of children (Baird et al., 2006) and are four times more common in males than in females.
  - About 10% of individuals with an ASD have an identified genetic etiology.
  - The genetic architecture of autism resembles that of ID, with many genetic and genomic disorders involved, each accounting for a small fraction of cases.

- Methods: An extensive literature search about genetic disorders with autism, ASD, pervasive developmental disorder, Asperger syndrome, PDD-NOS, or autistic/autistic-like traits. Results from common variant studies were not included because of the absence of replications.

- Identified more than 100 loci for which there is evidence for a causal role in ASDs. The majority have not been explored in ASD.

- Conclusion: autism represents the final common pathway for numerous genetic brain disorders. It is also of interest to see that the genes implicated in ASD go beyond those involved in synaptic function and affect a wide range of cellular processes.

- Remark/questions:
  - For some of the genes in Table 1, only a single case with ASD/autistic features ($n = 21$) or a single family with 2-3 males ($n = 6$) were identified. How to confirm the gene with such a small sample?

Autism [Talk by Kathryn Roeder at Lane Center meeting, Nov, 2011]:

- GWAS of autism:
  - With more than 30,000 individuals, find no SNPs with OR > 1.05.
  - Relation to schizophrenia: a much larger study identifies some SNPs, and if taking these SNPs and test in autism, some are significant.
- Copy number variants in autism: [Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism, Neuron, 2011]
  - Simpsons simplex family: parents are non-autism, and one child is autistic (only one, or the other child non-autistic).
  - Association study of CNVs: map about 500 Simpson families. Focus on CNVs in children, but not in parents. Found some CNVs that greatly increase the autism risk, e.g. 4 copies of 7q.11 and 14 copies of 16p.11 [Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron, 2011]
- Case-constrol studies using exom sequencing data: no genes pass the threshold of 10^{-4}, using a variety of gene-based tests.

Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. [Sanders & State, Neuron, 2011]

- Background:
  - Overrepresentation of large (mean size of 2.3 Mb) rare de novo events were more frequent in ASD probands in simplex families, compared to controls, or versus probands from multiplex families.
  - Across all studies, the burden of rare de novo CNVs in simplex probands (i.e., the percentage of individuals carrying =1 rare de novo event) has ranged from 5.0% to 11%.
  - Incomplete penetrance and diversity of phenotypic outcomes: 16p11.2 deletions or duplications have been found in individuals with ASD and intellectual disability (ID), seizure disorder, obesity, macrocephaly, and schizophrenia.
- Data: 4457 individuals from 1174 families, 872 quartets and 252 families trios. At a threshold of greater than 20 Illumina probes mapping within a genomic interval a combined total of 58 rare de novo CNVs were identified across the two studies (and Nimblegen). The sensitivity for small de novo events was low for both arrays.
- Pattern of de novo CNVs in 872 probands vs. 872 siblings (Table 1):
  - Rate in siblings: 16 in 872 siblings, or 1.7% (of persons containing at least one de novo CNV).
  - Rate in probands: 54 in 872 probands, or 5.8%.
  - Deletions vs. duplications: about equally split in siblings, slightly more in probands.
  - Autosomal vs. X-chromosome CNVs: only 2 in chrX.
  - Size distribution: small (< 100kb) - about 1 gene, medium ($100 - 1000$kb) - about 4 to 10 genes, and large (> 1Mb) - about 10-25 genes. In siblings: 3:9:4, in probands: 5:26:23.
  - Single Occurrence De Novo CNVs: 14 in siblings, 37 in probands. Double occurrence: 2 in siblings, 8 in probands. More than two occurrence: 0 in siblings, 9 in probands.
- Rare recurrent de novo CNVs:
  - 23 probands carried recurrent de novo CNVs in six distinct regions of the genome. Each of these intervals contained from 2 to 11 de novo CNVs in unrelated probands.

  - Statistical sigificance: first estimate the number of possible CNV regions, using the "unseen species problem", $C = 242$. Next, estimate the probabililty of multiple hits in the same region (close to binom.test). Ex. $n = 4$, $p = 7e - 6$.
  - Rare transmitted CNVs in probands: 8 overlap with one of the 51 de novo CNVs. In siblings, no overlapped region was found.

- Lessons:
  - Rare de novo CNV rate: about half of LoF.
  - OR about 3 for de novo medium to large CNVs.
  - Substantial overlap of de novo and transmitted CNVs in probands: 8 in probands vs. 0 found in siblings

Autistic-like social behaviour in Shank2-mutant mice improved by restoring NMDA receptor function [Won, Nature, 2012]

- Shank2 microdeletion leads to ASD in human: loss of exons 6, 7 and frameshift, leading to loss of PDZ domain.
- Characterizing Shank2 deletion mouse:
  - Normal reproduction and brain structure
  - Reduced social interaction: social interaction with a stranger mouse.
  - impaired spatial learning and memory in the Morris water maze
  - impairments in social communication by ultrasonic vocalizations (USVs)
- Effect on synaptic transmission: use hippocampal neurons. Basal excitation normal, butlong-term potentiation (LTP) induced by high-frequency stimulation or theta-burst stimulation was severely impaired in Shank2 -/- mice.
- Marked decrease in NMDA glutamate receptor (NMDAR) function: reducedNMDA/AMPA ratio.NMDAR-mediated transmission is selectively decreased (those mediated by AMPAR is normal)
- Shank2 deletion impairs NMDAR-associated signaling:phosphorylation but not total levels of CaMKII-/ (T286), ERK1/2 (p42/44) and p38 were significantly reduced.
- Agonist of NMDAR improved social interaction in Shank2 -/- mice.
- Lesson: to characterize the KO of risk genes: (1) Organism level: brain structure, learning and memory, social interactions. (2) Cellular level: e.g. synaptic transmission, basal state and stimulation. (3) Molecular level: processes disrupted, here focusing on particular type of receptor important for synaptic transmission.

Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. [Luo & Geschwind, AJHG, 2012]:

- Motivation: genetic associations for most individual rare CNVs are not clear. Gene-expression data, might confirm the presence of functional alterations (change of gene expression) related to a particular CNV and would thus be of significant utility.
- Background: Lymphoblasts provide useful data for a significantly overlapping set of genes expressed in the CNS.
- Data: 221 probands, 188 siblings from SSC. Microarray expression of 11,150 genes. 330 samples characterized by both genotyping data and expression data.

- Outlier genes: defined for each individual (proband or sibling), more than 3 SDs (99.7% confidence interval) from the mean expression of that gene across all samples.
  - Probands and siblings had a similar number of outlier genes per individual (about 10 down-regulated and 16 up-regulated). For brain-expressed genes: 77% and 73% of outlier genes were expressed in the human fetal brain in probands and siblings, respectively (no enrichment for genes expressed in the adult brain).
  - GO enrichment analysis: proband outlier genes, but not sibling, show an enrichment in neuron-related pathways. Note: CNV plays a small role here, because $> 90\%$ of the dysregulated genes in GeneGo neural pathways are outside CNVs.
- CNVs and the impact on gene expression:
  - The proportion of dysregulated genes within a given CNV: defined by dividing the number of dysregulated genes by the number of expressed genes within CNVs. A significantly higher proportion of dysregulated genes in rare de novo CNVs than in rare transmitted CNVs and common CNVs in probands.
  - Note: this analysis is independent of ASD phenotype, thus only demonstrate that de novo and rare CNVs have larger effect on gene expression on average than common CNVs. Similar to SNVs: RVs on average have bigger impact on gene function than CVs.
- Twenty-seven out of 40 rare de novo CNVs identified in probands had significantly more dysregulated genes than did the genome background. Percent of dysregulated genes range from 50-100% (for the 27 CNVs). Small non-recurrent CNVs: the outlier gene in the region are good candidates. Ex. TMLHE in Xq28 deletion (Figure 4D).
- 16p11.2 deletion and duplication:
  - Most genes in the region show expression correlation with dosage (12 out of 19), Figure 5. The genes showing the best correlation with dosage: include potassium channel tetramerisation domain containing 13 (KCTD13), aldolase A, fructose-bisphosphate (ALDOA), and MYC-associated zinc finger protein (MAZ). All are plausible candidate genes.
  - Consequence of the events on gene expression (trans-regulation): 70 DEX genes in 16p11.2-deletion cases and 135 DEX genes in 16p11.2-duplication cases. Not much overlap between the two lists, providing a functional basis for the different phenotypes observed in these two conditions.
  - Correlation of gene expression in 16p11.2 region with head size. The changes in these genes' expression accounted for more than 50
- Question: for a rare CNV, many genes are outliers. Do these genes appear as outliers only in carriers of CNVs, or outliers even in other individuals?
- Remark/lessons:
  - The genetic, expression and phenotype data are collected in the same set of individuals, thus analysis can be performed at the invididual level.
  - Outlier genes can be defined based on their differential expression between probands and siblings, however, due to heterogeneity of ASD, we do not expect the same gene is often differentially expressed in different individuals (about 10 DEX genes per individual on average). So any gene that is differentially expressed in one individual is considered.
  - For a given CNV, not all the genes in the regions will show sigificant expression changes. This is similar to missense mutations. So the consequence on gene expression (both in cis and in trans) can be important.

Patterns and rates of exonic de novo mutations in autism spectrum disorders [Neale & Daly, Nature, 2012]:

- Data: 175 ASD probands and their parents across five centres.
- Mutation rate estimation:
  - Genome-wide average, $1.2E-8$ (from earlier estimate) and exome is higher because of higher (50% vs. 40% in whole genome) GC content. Per bp mutation rate in exome: 1.5E-8.
  - Mutation rate matrix: 64 by 3, as the rate at each nucleotide depends on its two neighbors.
  - The proportion of mutation rates: from 1000 Genome project or human-primate comparisons. The equlibrium frequencies (or conditional frequencies) depend on the rates.
- Overall pattern of de novo mutations:
  - 161 coding region point mutations (101 missense, 50 silent and 10 nonsense), with an additional two conserved splice site (CSS) SNVs and six frameshift insertions/deletions (indels).
  - The power of individual gene is low, so assess the enrichment of de novo mutations among all genes. Similar to QQ plot in GWAS (enrichment of high-significance SNPs).
  - Expect 0.87 per exome per family vs. 0.92 observed de novo mutations per exome per family.
  - Missense, nonsense and Synonymous mutations: the proportion of the three categories in the observed events are similar to what is expected. Nonsense mutation is about 2-fold higher (6.2% observed vs. 3.3% expected).
- Secondary phenotype analysis (covariates): arental age strongly predicts the number of de novo events per offspring.
- Genes with multiple hits:
  - Three genes with two de novo mutations: BRCA2 (two missense), FAT1 (two missense) and KCNMA1 (one missense, one silent).
  - Simulations show two hits not enough to define a gene as a conclusive risk factor (Table S7): expect about 1.5 genes by chance.
- Genetic architecture: the number of causal genes and the effect size distribution:
  - Strategy: simulation the data under the assumption of fraction of causal genes and the average effect size (RR), to match the distribution of number of de novo mutations in samples (number of families with 0, 1, 2, etc., events in the genome).
  - Model: suppose X is the number de novo events, X follows Poisson distribution in Autism trios, and we want to find the rate. The model is similar to our de novo model (at gene level), but models the case of $H = 0$ or $H > 0$ (H: number of bad hits in the genome). Roughly: $P(A|X = x) = P(H = 0|X = x)P(A|H = 0) + P(H > 0|X = x)P(A|H > 0)$.
  - Results: 1000 causal genes with average $\gamma = 200$, highly inconsistent with the observed count distribution.
  - De novo SNV events will probably explain $< 5\%$ of the overall variance in autism risk (Table S4). This second quantity is calculated assuming a liability threshold model and additive contributions from the many genes contributing to autism risk.
- Protein network analyis:
  - Motivation: since we have little have confidence on individual genes, we want to see if there is any pattern in the connection/relation of genes.

  - Higher connectivity among de novo genes: In the set of 113 genes with missense or LOF mutations, significant enrichment of PPIs using DAPPLE.
  - Higher connectivity with the known ASD/ID genes: Distance between previously known ASD/intellectual disability genes and the current list in the PPI network: significantly smaller than control (genes with de novo variants in unaffected siblings), but the difference is small (3.66 vs. 3.78).

- Combined analysis with three papers:
  - 18 genes with two functional de novo mutations are observed in the complete data.
  - Expect 11.92 genes by chance. Simulation: draw a random set of mutations (by mutation rates), and count the number of times a gene is hit multiple times (Table S7)

De novo mutations revealed by whole-exome sequencing are strongly associated with autism [Sanders & State, Nature, 2012]

- Data: 238 families from the Simons Simplex Collection (SSC), two unaffected parents, an affected proband, and, in 200 families, an unaffected sibling.
  - Only those bases showing greater than 20 independent reads in all family members were considered for de novo mutation detection.
  - Only consider SNV, given the uncertainty of indel detection

- De novo mutations: overall pattern:
  - Analysis strategy: compare probands vs. siblings, using case/control (2 by 2 table) analysis, for enrichment test and OR estimation.
  - Among 200 quartets, 125 non-synonymous de novo SNVs were present in probands and 87 in siblings: 15 of these were nonsense (10 in probands; 5 in siblings) and 5 altered a canonical splice site (5 in probands; 0 in siblings).
  - The total number of non-synonymous de novo SNVs was significantly greater in probands compared to their unaffected siblings. Similarly for the odds ratio (OR) of non-synonymous to silent mutations in probands versus siblings (2 by 2 table test), $OR = 1.93$.
  - Comparison of de novo LOF mutations in brain-expressed genes: $N = 13$ in probands and 3 in siblings (significant), and OR about 5.65.

- Covariate analysis: the rate of de novo SNVs indeed increases with paternal age, and that paternal and maternal ages are highly correlated. Re-analysis accounting for age did not substantively alter any of the significant results reported here.

- Simulation for estimating genetic architecture:
  - Simulation: assume there are a certain number of ASD and non-ASD genes, for each gene, sample de novo events according to the mutation rates; then sample phenotypes by whether a subject has de novo mutations in ASD genes: (1) if not, sample by baseline penetrance, K; (2) if yes, sample by the penetrance (γK). After this procedure, select a certain number of probands and siblings, and count the number of events in each group. Tune γ and number of ASD genes s.t. the simulated counts match the counts in the real data.
  - Results (Table S3, S4): use $K = 0.21\%$, average relative risk under different numbers of genes, in particular, assuming there are 1000 ASD genes, the RR is $1.82\%/0.21\% = 8.7$ for nonsyn. variants (Table S3, 1000 genes) and $11.5\%/0.21\% = 54.7$ (Table S4, 1000 genes) for nonsense.

- Multiple-hit genes:

- Procedure for estimating FDR at certain number of de novo mutations via simulation: Note that only 1 gene in real data pass the threshold (say 2 de novo LOF mutations), so need to simulate "real" data as well (not just non-ASD genes). First simulate the data (according to the model tuned before), then count the genes with certain number (k) of de novo events. FDR is defined as the number of non-ASD genes have k events, divided by the total number of genes having k events.
- Under all models, two or more nonsense and/or splice site de novo mutations were highly unlikely to occur by chance ($Q = 0.005$). Only a single gene in our cohort, SCN2A, met these thresholds.
- No evidence that PolyPhen, SIFT, GERP, PhyloP or Grantham Score, either alone or in combination, differentiated de novo non-synonymous SNVs in probands compared to siblings.
- In the SSC cohort at least three, and most often four or more, brain-expressed non-synonymous de novo SNVs in the same gene would be necessary to show a significant association. Unlike the case of nonsense and splice site mutations, these simulations were highly sensitive to both sample size and heterogeneity models

- Remark:
  - Estimating FDR via simulation: we want to know the FDR at 2 de novo events in a gene. In real data, only 1 gene is predicted, so insufficient to estimate FDR. Instead, we simulate data from both H_0 and H_1, count the de novo events in genes, and calculate FDR. In general, when we need to evaluate a statistical method in the context of multiple testing, we may need to estimate its FDR, and this can be done via simulation.

Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations [O'Roak & Eichler, Nature, 2012]

- Network analysis: on 126 events (turncating or severe missense mutations)
  - Building PPI network: GeneMANIA.
  - Intra-connectivity among 126 genes: 49 genes mapped to a highly connected network, significant more edges than expected $P < 0.0001$.
  - Connectivity with curated 103 ASD genes: degree-aware disease gene prioritization (DADA). Genes with severe mutations ranked significantly higher than all other genes (Mann-Whitney test). Signal overwhelmingly driven by 49 highly connected genes. Sibling events as control.

De Novo Gene Disruptions in Children on the Autistic Spectrum [Iossifov & Wigler, Neuron, 2012]:

- Data: 343 families, a subset of the Simons Simplex Collection. In each family, only one is affected and at least one unaffected sibling.
- Overall pattern in probands vs. siblings (Table 2):
  - Overall SNVs are very similar in two groups: all SNVs (380 versus 364), synonymous (79 versus 69), or missense (207 versus 207).
  - Nonsense mutations (19 versus 9) and point mutations that alter splice sites (6 versus 3)
  - Filter for genes expressed in brain, count missense mutations that cause nonconservative amino acid changes, or count missense mutations at positions conserved among vertebrates, no statistical evidence for contribution from this type of mutation.
  - Indels: 53 indels in probands and 32 in siblings. Of these, 32 in probands and 15 in siblings caused frame shifts.
- Origin of mutations:

  - From the original sequencing and validation of our data, we were able to ascertain the parental haplotype for some de novo mutations. We found that the father is more frequently the parent of origin than the mother: 50/17 for SNVs and 6/1 for indels
  - De novo SNVs in children with the youngest fathers has lower mutation rate than in those with the oldest (p value of 0.013).
  - The assumption is the mutations arise in germline, but somatic mutations may also be possible. Because of the filters used, the de novo events reported are largely and perhaps almost entirely germline in origin.

- Multiple-hit genes:
  - No recurrences among our 59 LGD (likely gene disruption) targets. There are two overlaps with the 72 most likely candidate genes from our previous CNV study: NRXN1 and PHF2.
  - Missense: a few overlaps of the LGD targets and targets of missense mutations, two in siblings and one in probands, but this is well within random expectation.
  - 14 of our 59 LGD targets and 13 of 72 CNV target genes, with one in common, overlap with the 842 FMRP-associated genes. No significant overlap in siblings.

- Effects due to inheritance: This study lacks the power to discover small effects due to inheritance. Consider only rare variants, and then examined transmission to children, by affected status. No statistically significant transmission bias of either missense or LGDs.

- Purifying selection on FMRP genes: onsense and splice site rare variants: the proportion in FMRP genes is one-fourth of that in all genes. Missense variants show a much less extreme depletion in the FMRP-associated genes.

- Genetic architecture:
  - The total contribution from LGD mutations can be estimated as 31 events in 343 families (59 events in probands minus 28 events in siblings), or roughly 10% of affected children.
  - A simple power calculation indicates that we cannot rule out confidently even a 20% contribution to autism from de novo missense mutation. Despite these caveats, it is worth considering that de novo mutation causing merely amino acid substitution may only rarely create a dominant allele of strong effect.
  - We project that nearly half of autism target genes will be among the list of FMRP-associated genes.

- Combined analysis of all four datasets:
  - LOF mutations (including nonsense, splice site and frameshift indels): with the 59 from this study, a total of 127 hits in probands have been found. Judging from our two-fold differential rate in probands and siblings, we expect that at least half of the 127 hits, about 65, are causal.
  - Five genes were hit twice. DYRK1A and POGZ are the new recurrences found by combining our data with theirs.
  - From our estimate of 65 causal gene disruptions and 5 recurrent gene targets, we project that the total number of dosage-sensitive targets for autism is about 370 genes. Recurrence analysis: For a target size T, and K picks with replacement,we can calculate analytically how many targets R are picked twice or more. Given $K = 65, R = 5$, the most likely T is 370.

- Remark/questions:

- Family data analysis: two basic strategies (1) de novo mutations as in this paper; (2) transmission disequilibrium: i.e. disease alleles are preferrably transmitted to affected children vs. unaffected ones. Not in multiplex cases (multiple affected cases in a family): transmission genetics plays a greater role.
- Issues with sequencing data processing: data quality is one major isssue because of uneven coverage across the genome. To ensure quality, one may apply coverage filter: e.g. only variants above a certain coverage will be analyzed.
- Indels: some are frameshift ones (most likely LOF), even in-frame indels are more disruptive to a peptide than a mere substitution.

Rare Complete Knockouts in Humans: Population Distribution and Significant Role in Autism Spectrum Disorders [Lim & Daly, Neuron, 2013]

- Background: there are relatively few homozygous or compound heterozygous LoF variants (i.e., complete gene knockouts) in healthy individuals. Most of these complete knockouts found are common (minor allele frequency [MAF] $> 5\%$) and are distributed across a very small number (100-200) of genes, such as the olfactory receptors.
- In the ASD dataset, an average individual has 5 complete knockouts, however, most of them are common LoFs. If limiting to rare LoF (less than 5%), fewer than 5% of individuals harbor even a single rare complete knockout.
- There are a total of 91 such rare complete knockouts in the case-control data sets, with 62 of these found in the cases compared to 29 in the controls, representing a roughly 2-fold enrichment (933 cases and 869 controls).
- Enrichment of Rare Complete Knockouts Observed on the X Chromosome: we examined LoF variants with population frequency (assessed in female control samples) of $\leq 0.25\%$ (to match 5% LoF frequency in autosomes - the same homozygosity).
  - A significant enrichment of rare hemizygous LoFs in male cases, with 88 such events observed - 60 of them were found in male cases (784) and 28 of them were found in male controls (417) (OR = 1.5).
  - We found 2 of 170 female cases bearing a rare complete knockout on the X chromosome and 0 of 452 female controls.
- We found that there was a higher rate of rare complete knockouts in females (5.4%) compared to males (4%). Although 16% of the cases sequenced were female, 25% of the cases harboring rare complete knockouts were female. This is consistent with the model that females need a higher dose of genetic risk to manifest a diagnosis of ASD.

Synaptic, transcriptional, and chromatin genes disrupted in autism [De Rubeis & Buxbaum, Nature, 2014]

- Main results of TADA: 22 genes at FDR < 0.05 and 107 genes at FDR < 0.3.
- Differential gender analysis: females have a higher liability threshold (ASD is less common in females), the consequence is that: if a variant has the same effect on autism liability in males as it does in females, the RR of the variant in the females will be higher than that in the males. As a result, the variant will be at higher frequency in female ASD cases compared to males.
  - Intuition: males already have high risk, so the effect of a very strong mutation may increase the penetrance from 0.1 to 0.15; for females, the baseline penetrance is low, say 0.01, and a strong mutation can move it to, say 0.04. So in general, the RR (the ratio of penetrance) is higher in females than in males.

  - Liability threshold model: the thresholds are $t_m = 1.98$ and $t_f = 2.56$. Suppose the effect of a mutation is ΔZ. Plug in the relationship between RR and threshold (Equation 1.3 in Genetics Notes), and we can explore the difference of RR between males and females.
  - Relation to AF difference between males and females: suppose the AF is the same between male and female controls, then the difference of OR means that the AF in male cases will be different from AF in female cases.

- Enrichment analysis on FDR $< .3$ genes: evolutionarly constraints, FMRP and RBFOX (splicing factor) targets, nominal enrichment of synaptic and PSD genes.

- DAWN results: using constraint scores as prior. Found 160 genes, including 97 not in the FDR < 0.3 gene list. Found four clusters (Figure 2). C2 - some genes for neurodegenerative diseases; C3 - known ASD genes; C4 - chromatin regulation.

- Molecular analysis: SCN2A, CACNA1D. The locations of variants/mutations - relative to known pathogenetic mutations of these genes in other diseases.

- Analysis of HMG genes: enrichment and interconnection of all HMG genes found by TADA. First define transcriptional network using ChIP-seq data (ChEA database), then map the connections of TADA genes and other HMGs.

- Common genes of ASD and other mental diseases, Schiz., cogential heart disease and metabolic disorders.

- **Lessons**: analysis following gene discoveries may include: molecular-level analysis (where variants are located), enrichment of gene groups, interconnection among the discovered genes (regulatory relations, PPIs, and so on) and relationship with other diseases (shared genes).

The contribution of de novo coding mutations to autism spectrum disorder [Iosiffov and Wiggler, Nature, 2014]

- Rate of de novo mutations: LGD: 0.12 in siblings and 0.21 in probands. Missense: 0.82 for unaffected siblings and 0.94 for affected proband. So about 43% LGD and 13% missense DNMs are causal.

- Estimation of contribution of DNMs: ascertainment differential (increased rate of DNM in probands per individual) is 0.21, adding missense and LOF mutations. Adding CNVs, make the proportion 0.27. " Including copy number variants, coding de novo mutations contribute to about 30% of all simplex and 45% of female diagnoses.".

Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder, [Uddin & Scherer, NG, 2014]

- Hypothesis: (1) brain-expressed exons are under stronger selection; (2) brain-expressed, highly constrained exons are candidates of ASD.

- Contingency index: (1) Burden of rare missense mutations: in ESP data, count the number of such mutations (MAF below 0.05) divided by exon length; (2) Brain expression level of exons. An exon is considered "critical exons" if it has low burden and high brain expression (both defined as 75% percentile across the entire dataset, using the multiple tissue microarray expression indices). Also brain critical exons are defined by removing all critical exons highly expressed in at least one non-brain tissue.

- Association of exon expression and constraint
  - On several known ASD genes, an inverse correlation between the burden of rare missense mutations and the brain expression levels for exons.

  - Test 11 tissues, brain cerebellum expression showed a strong association with the missense mutation burden.

- Comparing entire ASD de novo genes and sibling de novo genes (SSC data): no significant difference in conservation scores, the distribution of the burden of rare missense mutations and PPH2 scores.

- Critical exons are enriched in putative ASD genes:

  - Among exons affected by ASD de novo mutations, 45% are critical exons; the ratio is 28% in the exons affected by de novo mutations in unaffected siblings.
  - 3,955 'brain-critical exons' (from 1,744 genes) with high expression specific to the brain and a low burden of rare mutations: enriched for FMRP targets, genes associated with ASD risk (autosomal dominant or X linked).

- Remark:

  - Table S7 has all the 1,744 genes containing all brain critical exons. No exon-level data is available, however.
  - Brain critical exons: defined using all exons, "For each dataset exons were sorted using a threshold of the 75th percentile for expression". In the brain critical exon genes: KATNAL2, CHD8, TBR1 not found.

ExAC talk [ASC, March, 2015]

- $n = 3,982$ families

- ExAC browser: information of variants, including AF and number of homozygoties in each population (African, Asian, etc.)

- Using ExAC to interpret DN mutations: focus on constrained genes

  - Among all DN LoFs found in ASD cases, a large fraction are not found in ExAC (OR = 2.8).
  - Mis3 mutations not found in ExAC: ASD de novo rate = .049, Control de novo rate = .024 (OR = 2.1)

- Using ExAC to interpret inherited mutations: focus on constrained genes

  - LoFs not found in ExAC: 1666 trios, 169 families, 105:64 - transmission bias.

- Constraint analysis: 19 genes have 4 dn LoF (ASD + ID), all have constraint Z scores > 4.5.

Excess of rare, inherited truncating mutations in autism. [Krumm & Eichler, NG, 2015]

- Data: 2,377 SSC families. Recall the DNMs, found a total of 1,544 DNMs. 21 new recurrently mutated genes (including missense).

- SNV transmission disequilibrium: private LGD (similar to LoF) variants in genes with the lower 50% RVIS values (constraint scores), OR = 1.14. And the signal is stronger for more constrained genes (strongest 1%, OR = 1.4). The signal is strongest for private variants, but also persists for rare LGDs. A clear bias in transmission from mothers to sons.

- CNV transmission disequilibrium: all transmitted CNVs, OR = 1.10. Deletion only 1.11 and duplication 1.12.

- Integrating SNVs and CNVs: obtain p-values from de novo SNVs and the rest, then combine with Fisher's method. None of the convergent genes is significant, lowest $p = 0.01$.

Whole-genome sequencing of quartet families with autism spectrum disorder [Yuen & Scherer, Nature Med, 2015]

- Motivations of WGS: (1) nongenic, non-coding RNA, CNVs; (2) WGS provides more uniform coverage of the exomes than WES.
- Data: WGS (Complete Genomics) of 85 multiplex families with 2 affected children. Also collect clinical info such as ID, language, adaptive functioning, family history, and so on. Average coverage of genome: 96.8% with average depth 56; average coverage of exome: 99.6% with at least 5x coverage and 74.8% with at least 40x read depth.
- De novo mutation rate: 59.3 de novo SNVs / genome, and 13.2 de novo small indels (< 100 bp). From the validation rate, 90% for SNVs and 60% for indels, estimate that 62 de novo events per genome.
- Burden analysis (combine de novo and inherited): rare LoF and missense mutations (MAF $< .01$), significant difference between children and parents in two genes groups: all genes related to abnormal mental function (687 genes), and all neuronal and brain functionrelated genes from GO (309 genes).
- Comparison between two siblings: test heterogeneity of ASD causes.
  - Only 29.5% of the variants (rare variants in LoF and missense) within the two genesets are shared between two siblings.
  - Classification of ASD-relevant mutations: known ASD, candidate ASD, novel ASD, and neurodevelopmental disorder. Consider only de novo LoF and damaging missense, inherited LoF, CNVs (both de novo and inherited).
  - Identified ASD-relevant mutations in 36 of 85 (42.4%) families. In only 14 of these 36 families did both affected siblings carry ASD-relevant mutations.
  - Similar discordant rate when limit to highly-confident ASD gens and LoF mutations, and pathogenic CNVs.
- Affected siblings could carry the same DN mutations: an example of a family where both affected siblings have the same de novo CNV in SCN2A. Apparent de novo events shared between siblings are not uncommon in ASD-affected families, and mechanistically they can be attributed to gonadal or germline mosaicism or parental somatic mosaicism. Also identified 21 de novo SNV events shared between two siblings in 16 families, none in exonic regions.
- Different clinical features of siblings with different ASD-relevant mutations: In siblings with shared mutations (11 out of 36 families), autism symptoms related to social and communication domains are not significantly different. But different in other sibling pairs.
- The involvement of deafness-associated genes: THRA gene (thyroid hormone) in one family. Also higher burden of LoF mutations in deafness-associated genes in children vs. parents.
- Remark:
  - Hypothesis that explain the finding: if autism generally requires multiple mutations, and our ASD gene list is quite incomplete, then it is possible that the shared genes are not detected.
  - The benefits of WGS: higher coverage in exomes and CNVs. Not using non-coding sequences.
- Lesson:
  - Burden analysis for de novo mutations: in general, need to control for gender and father's age.
  - Multiplex families, so likely that de novo mutations play a less important role. Still, it is possible that both siblings have the same de novo mutations because of genetic mosaicism (not uncommon).

Loss of δ-catenin function in severe autism [Turner & Chakravarti, Nature, 2015]

- Idea of identifying of autism genes: females are more protected, thus need more deleterious variants to reach the threshold. So deleterious variants are more enriched in female-enriched multiplex families.
- Genetic finding of CTNND2: 13 urelated female cases from multiplex families with severe autism. Found CTNND2 with multiple severe variants (highly conserved missense or LOF). Burden is significant: 362 additional females cases vs. 1GP or EPS as controls.
- CNV evidence of CTNND2: among all CNVs overlapping CTNND2, significant enrichment of exon-disrupting CNVs vs. non-disrupting CNVs in cases than in controls.
- Functional studies of CTNND2: (1) Phenotypic relevance to ASD: use in vitro model, dentric spine density; mouse ermrbyo morphology. (2) Function affected by CTNND2: expression of Wnt signaling genes in CTNND2 deletion.
- Expression pattern of CTNND2: (1) Expression trajectory (higher in fetal brain) and tissue specificity. (2) Co-expression with 500 SFARI autism genes.
- Pathway analysis with GeneMania: the GO functions of genes related to CTNND2, enrichment of dentrite morphogenesis and chromatin modification.

Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA [Turner & Eichler, AJHG, 2016]

- Data: WGS on 40 Simons quads initially, then WGS of 13 trios (affected) and 3 control trios. The selection of 40 families: 39 have no previously found LOF or CNV mutations. Coverage: 31.5, sequenced at NYGC on Illumina Hiseq X Ten.
- Variant calling: GATK HaplpotypeCaller, FreeBayes and Platypus. Private SNVs and indels found (not in dbSNP) will then be called by DNMFilter or TrioDenovo for DNMs. Comparison of callers: validation rate = 89% for variants called by both GATK and FreeBayes, and 29% for GATK only and 10% FreeBayes only. DNMFilter outperforms TrioDenovo.
- Validation of 691 de novo SNVs and indels. Overall 75.2% validation of all DNMs.
- Defining putative regulatory variants: fetal CNS DHS sites, with PhyloP scores > 4. Defining distance: d kb away from upstream of TSS or downstream of TES (not including introns).
- Variant statistics: 70 DNMs (both SNVs and indels) per individual, and that increases by 28 or so from using TrioDenovo (lower validation rate, about 50%). Total: 7,936 SNVs and 42 indels in 40 WGS families. The most common cause of false positives was under-calling in the parent.
  - SNV concordance between GATK and FreeBayes: very high, close to 90%.
  - Indels: low overlap between GATK and FreeBayes, only 30% or so of all indels called.
- Burden in non-coding regulator sites (Table 1): total DNM counts in cases and controls: 3787 vs. 2997. Total DNMs in non-coding: 204 vs. 171 (not significant). Burden in potential ASD genes (57 genes): combine de novo SNVs and de novo & Priviate CNVs that overlap with DHS: (1) 10kb: 5 vs. 0; (2) 25kb - 50 kb: 6 vs. 0. (3) 100kb: 8 vs 1. (4) 500kb: 21 vs 9.

De novo synonymous mutations in regulatory elements contribute to the genetic etiology of autism and schizophrenia [Takata & Karayiorgou, Neuron, 2016]

- Enrichment of near-splice site DN syn. mutations: about 2-fold enrichment in both ASD and SCZ. No enrichment in distal splice sites. Also in the distance distribution of DN syn. mutations: ASD and SCZ mutations are closer.

- Enrichment of mutations that affect exonic splicing regulator (ESR): similar pattern, about 2-fold enrichment in cases. Not in mutations that do not change ESR.
- Enrichment of DN syn. mutations within DHS: in cerebrum and frontal cortex, but not cerebellum. Not in DHS compiled from 125 cell types.
- Lack of enrichment with other annotations: miRNA binding sites (in coding), codon optimality or RNA secondary structure.
- Adjusting for multiple testing: 87 hypothesis tested in total, adjust using BH. Significant results: near splice site mutations for ASD and mutations in frontal cortex-derived DHS for SCZ.
- Gene set enrichment: genes with functional DNSMs are more likely to be intolerant (RVIS), and relevant gene sets.

Genome-wide characteristics of de novo mutations in autism [Yuen & Scherer, NPJ Genomic Med, 2016]

- Data: 200 trios. On average, 50 SNVs, 3.9 indels and 0.05 CNVs per individual. High validation rates for all DNMs.
- Pattern of DNMs: 239 clustered DNMs ($\geq$ 2 DNMs within 20kb of the same individual). About half are within 200 bp. Most clustered DNMs are maternal in origin, and are close to dn CNVs.
- Method of testing DNM enrichment: comparing with controls, 258 Dutch trios. The coverages are very different: 32x vs. 13x. Difference in GC. To adjust for the GC difference, use logistic regression, where y is proband or control, and x the features, e.g. GC of the DNMs.
- Enrichment in splicing and UTR: (1) Splicing: use SPIDEX score, modest enrichment, p close to 0.05. (2) UTR3: use PhyloP, find enrichment, p about 0.05 to 0.01.
- Enrichment of DNMs in promoter/enhancer regions: DeepBind loss at DHS: p around 0.05. DeepBind loss and PCons (conservation): much larger OR, and $p < 10^{-4}$.
- Remark: the analysis on various gene sets in the paper do not distinguish coding and noncoding DNMs.

Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder [Krishnan & Troyanskaya, NN, 2016]

- Prediction of ASD risk genes: (1) Training genes: about 500 from SFARI, assigned to four evidence levels. (2) Brain-specific gene network. (3) For each gene, its features as connectivity to other genes (20k features per gene), then train SVM: weigh the genes by assigning partial labels on genes with lower evidence (weighting improve the results).
- Validation of predictions: focus on top 2,500 genes. Enriched with various sets of genes, e.g. 60% of all LGD genes are in the 2500 gene list.
- Spatial and temporal pattern of ASD: in 2,500 genes, study their expression patterns. Found that the expression of these genes are highly enriched in prenatal (early, mid and late fetal) stages. In contrast, spatially, it is not very specific, enrichment across most regions.

Post-zygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations [Dou & Wei, review for Human Mutations, 2017]

- Post-zygotic mutations: mutations occur after zygote formation. Imagine a starting embryo: at some point in development, a mutation occurs (PZM) in some progeny cells, this would lead to mosaicism. Mosaicism in somatic cells: somatic mutations, and cannot pass to offsprings; when it occurs in germline cells: there is a chance that the PZM can pass to offsprings (DNM).

- Mosaicism in severe/lethal genes could lead to mild phenotype, but not diagnosis. Particularly, parental mosaicism could lead to recurrent risks in offsprings.
- Detecting pSNM in children: excessive G>T mutations in Yale dataset (single strand). Most are FPs, probably due to oxidative damage during experiment.
- Detecting parental pSNM transmitted to children: heterozygous in child and mosaic in parents.
- pSNM detection: 1000 child pSNM and 300 transmitted pSNMs. Validation: 50%. Validated pSNMs per subject was 0.152 for child pSNMs and 0.030 for transmitted parental pSNMs.
- Burden analysis of child pSNM: compare distribution of MAFs of pSNMs between probands and siblings. In missense/LoF pSNMs: seems that the distribution is skewed towards higher MAF in probands than in siblings, and the trend is not observed in neutral mutations. At MAF > 0.2, OR about 2, and $p = 0.02$.
- Burden analysis of parental transmitted pSNM: Missense/LoF pSNM with low MAF (< 0.2) are enriched in probands vs. siblings. OR = 5.36 for LoF and 1.63 for missense.
- Combine pSNM and DNMs for gene discovery: genes carrying pSNMs tend to have some statistical evidence in TADA-Denovo analysis of WES.
- Plausible explanations of results:
  - For parental transmitted pSNM, there is generally selection against deleterious mutations, so the rare mutations tend to be more deleterious (simliar to RVs vs. CVs).
  - For child pSNM, the high MAF mutations are likely more deleterious than low MAF ones: negative selection here may not apply because the mutations happen randomly, and MAF may reflect only the timing of mutations.
- Why RR of child missense pSNMs much larger than de novo missense mutations? Germline cell competition in parents: the more deleterious mutations cannot reach high frequency in parental germlines, which limits the effect size of DNMs. In contrast, child pSNM with high MAF emerger early during develepment, and there is not a lot of competition to remove these mutations.

Genomic Patterns of De Novo Mutation in Simplex Autism [Turner and Eichler, Cell, 2017]

- Coding DNM analysis: LGD depletion (ascertainment of SSC data), but DN missense mutations at CADD > 30 show OR = 2.
- UTR: combine 5 and 3 UTRs, OR = 1.1, p = 0.03.
- Promoters: in fetal brain, defined by ChromHMM, within a TFBS, OR = 1.8, p = 0.03 (34 vs. 19).
- Putative non-coding regulatory regions (pNCR): conserved TFBSs mapped to fetal brain DHS. Conservation: GERP scores > 2. OR = 1.3, p = 0.02. A strong burden in ESC enhancers, 8 vs. 0 (p = 0.02).
- GO analysis of genes closest to DNMs in TFBSs: top 5 GO BPs are related to neuro-development.
- Remark: the non-coding analysis are only nominally significant and do not survive multiple testing.
- Enrichment analysis in autism genes: Use Turner57 genes. (1) all putative functional mutations: LoF, missense, UTR, pNCN and exonic deletions, show OR = 2.2, p = 1.7E-3. (2) Individuals with two or more DNMs near SFARI (800) genes: more enriched in probands.
- Q: definition of TFBS?

An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder (CWAS) [Werling and Sanders, NG, 2018]

- Linear regression test for mutation burden: let y_i be the number of DNMs in sample i, x_i is the sample label (proband or sibling), and covariates are paternal age and sample-level sequencing metrics (e.g. average coverage).
- The effect of sequencing metrics and paternal age: do step-wise linear regression, the selected features are paternal age, percent of genome with coverage 30x or higher and percent of total excluded reads. The effects are generally small, e.g. RR = 1.024 before removing paternal age effect and RR = 1.005 after.
- Use permutations to obtain p-values of variant set tests: 10,000 permutations on case-control status within each family.
- Estimate correlation structure of annotation categories: Perform random simulations of DNMs across the genome many times, then burden test of categories using binomial test. For each simulation, obtain p-values of all categories, and transform p-values to Z-scores. Then we obtain correlations of Z-scores of different categories.
- MOM to estimate the effective number of tests: given a matrix of simulated p-values for each category and each simulation, let M be the number of simulations, and p_i be the minimum p-value in i-th simulation, then we calculate $M/\sum_i p_i - 1$. When all categories are perfectly correlated, it is easy to check that this gives 1.
- To define effective number of tests using spectral clustering: R correlation of categories, and $A = D^{1/2}RD1/2$, where D is the degree. Use EVD of A, and the number of eigenvectors that explain most of the variations gives the number of tests. Remark: similar to finding the number of connected components in a weighted graph. Note: The effective number is higher as one increases the sample size.
- Results for noncoding annotations (51000 tests): adjusting for 4,000 effective tests, no signal in CWAS plot. Perform K-means clustering of categories.
- **Remark**: reduce the problem of number of independent tests to number of clusters. A related problem is the rank of a matrix. An alternative approach is to cast as a prediction problem (mutations in cases or controls), with variable selection.

Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder [An and Sanders, Science, 2018]

- CWAS: 50K annotation categories. 500 coding annotations significant, but none from noncoding.
- Enriched noncoding categories: build de novo risk scores, Lasso classification of DNM status (cases vs controls). Find non-coding annotations contributing to the score, then test differences in cases vs. controls. Almost all signals are found in promoters, 2kb upstream of TSS (Figure 2).
- Conserved promoters: most promoter signals are from conserved sequences (PhastCons and/or PhyloP scores). Using conserved promoters: RR = 1.28 (Figure 3D).

Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks [Ruzzo and Wall, Cell, 2019]

- Data: WGS of 439 multiplex families, total of 960 affected children, and some unaffected ones.
- Rare inherited coding mutations: no burden of number of transmitted variants (affected vs. unaffected). Also no burden in private promoter variants, even limited to known ASD genes.

- Define high-risk inherited variants: transmitted to all affected but no unaffected children, then focus on those disrupting (coding or promoters) of constrained genes pLI > 0.9. Results: 96 genes, with 40 SVs disrupting promoters and 62 PTVs. Significant burden in PTVs: five times depletions.
- SVs disrupting promoters of two genes: case study.
- Remark: multiplex families, the variants transmitted to all affected children, but not unaffected are more likely to be pathogenic.

Autism-like phenotype and risk gene mRNA deadenylation by CPEB4 mis-splicing [Nature, 2018]

- Background: CPEB4 is a risk gene of syndormic ASD.
- Model: (1) CPEB4 splicing dsyregulation in ASD: decreased inclusion of a neuron-specific microexon. (2) CPEB4 binds to many ASD risk genes, and due to CEPB4 dysregulation, the ASD risk genes show reduced polyA tails and expression.
- Lesson: mRNA stability regulation, possibly via poly-A, is important for diseases.

Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism [Satterstrom and Buxbaum, Cell, 2020]

- Data: 6430 trios with 2000 siblings. Additionally, 14,365 case- control samples (5,556 ASD cases, 8,809 controls).
- Variant categories: three ties of PTVs by pLI scores, and 3 tiers by MPC scores. One tier of syn. variants.
- Enrichment of PTVs: 3.5 in highest PTV category in DNMs, and 1.2 in transmitted and 1.8 in case-control. Modest in next PTV category: 1.3 in DNM and case-control.
- Enrichment of missense: RR $= 2.1$ in strongest missense category in DNMs, no burden in transmitted (require LOFTEE tag to be high confidence HC) and 1.2 in case-control.
- TADA: de novo PTV, missenes and case-control PTV. Found 102 genes at FDR < 0.1. (1) PTV: gamma is a continuous function of pLI. See Figure S2K, at the highest level, gamma about 20-30. (2) Missense: two categories, misA and misB with effects 4.18, 22.15.
- Simulations to assess FDR: (1) Use real data, but use syn. variants to randomly set as PTV and missense, in random genes. (2) Pure simulation using multinomial. FDR is calibrated (FDR vs. q-value - Figure S2).

8.6 Neurological Diseases

Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways, [Cirulli & Goldstein, Science, 2015]

- Background: ALS is a neurodegenerative disease, characterized by loss of motor neurons. Approximately 10% are familiar, abut 20 known genes explain only a minority 10% of all sporadic ALS cases.
- Data: WES of 2,869 cases and 6,405 controls.
- Statistical analysis: for each gene, run one of 6 models and find the one with lowest p-value. The models are burden test, where the burden is the number of qualifying variants in a sample. The qualifying variants are defined as:

  - Either dominant or recessive coding: for dominant, only consider RVs with MAF less than 0.05% and 0.005% in ExAC (1 in 20K, ExAC sample size is 60K). For recessive, requies two variants, MAF 1%.
  - Coding: NS and splice variants. Not-benign: NS and splice, but remove all benign ones by PPH2. LoF.

- Results of gene mapping:
  - A strong ALS gene (known) SOD1 (dominant coding model). Burden: 0.8% in cases, 0.08% in controls. Spatial pattern: concentrated in 3' portion, however, many are predicted to be benign by PPH2.
  - A new finding, TBK1, combined $P < 10^{-10}$ (dominant not-benign model). Burden: 1.0% in cases and 0.19% in controls, with LoF variants having even higher burden, 0.38% in cases vs. 0.034% in controls. Spatial pattern: diffused.

- Autophagy in ALS: TBK1, OPTN and SQSTMP1 (previous findings) are involved in autophagy and inflammation. Function in clearance of protein and protein-RNA aggregates.

- Lesson: for burden analysis, important to use very strigent MAF threshold. Spatial distribution of variants are variable: sometimes they are concentrated, but often not. PPH2 predictions may not be very informative (SOD1, most case variants are "benign").

8.7 Cardiovascular Diseases

De novo mutations in histone-modifying genes in congenital heart disease (CHD) [Zaidi & Lifton, Nature, 2013]

- Data: trios of 362 affected probands and 264 unaffected probands (siblings of autism cases). WES: 96.0% of bases read eight or more times.

- Specificity and sensitivity of de novo calls: de novos are called using Bayesian quality scores (QS). At QS > 50, 90 calls are confirmed with 100% accuracy. Consequently, de novo mutation calls with QS ≥ 50 were included in the study. Sensitivity is further diminished by about 5% owing to bases with very low read coverage.

- Defining genes expressing in developing heart: about 4,000 genes are expressed at high levels (top quantile) in developing mouse, defined as HHE (high heart expression) genes. The rest LHE genes.

- Rates of de novo mutations:
  - Overall: 0.88 de novo mutations per subject in CHD cases and 0.85 in controls. Cases and controls have similar maternal and paternal ages.
  - All missense in HHE genes: 81 in cases and 32 in controls, or 0.22/case vs. 0.11/control.
  - LoF in HHE genes: 15 in cases and 2 in controls, or 0.04/case vs. 0.01/control.
  - All the signals are contributed by HHE genes: there was no significant difference in mutation frequency in CHD cases versus controls among LHE genes in all comparisons.

- Pathway analysis: 8 de novo mutations in a pathway related to H3K4 methylation. Three genes in this pathway (MLL2, KDM6A, CHD7) have previously been implicated in rare syndromic CHD. Also the H3K4me pathway is the only enriched pathway in GO analysis.

- Biology of the pathway: The combination of both activating (H3K4 methylation) and inactivating (H3K27 methylation) chromatin marks characterizes poised promoters and enhancers, which regulate expression of key developmental genes

- Promising genes (Table 2):
  - A total of six double-hit genes (Table S11). SMAD2: double hit, splice site, conserved missense. NAA15: double hit, both LoF.
  - Among the 17 above genes (?), ten have no damaging variants and seven have one to five among $> 9,500$ exomes in National Heart, Lung, and Blood Exome Sequencing Project, 1000 Genomes and Yale exome databases.
- Other structural, neurodevelopmental and growth abnormalities were common in subjects carrying de novo mutations in interesting genes. Ex. CUL3, neurodev. abnormality.
- From the increased fraction of patients with protein-altering mutations in HHE genes in CHD patients (0.22) versus controls (0.12), we estimate that such mutations have a role in about 10% of these patients (about 40 extra de novo in cases, thus covering about 40 people, or 10% of cases).
- Remark: The odds ratio calculation is weird, using the number of silent mutations (happen to be the same, 21, for cases and controls), instead of sample sizes. Also the odds ratio is defined on de novo as a whole (viewed de novo as an "exposure" or risk factor), but not on individual causal mutations.
- Lesson: For some disease genes: a broader phenotypic spectrum resulting from mutations, e.g. CUL3, CHD8 (H3K4 pathway)

Systems biology with high-throughput sequencing reveals genetic mechanisms underlying the Metabolic Syndrome in the Lyon Hypertensive Rat [John Ma, Cardivascular Genetics, 2015]

- Data: F2 intercross of of LH and LN rats. The two strains are genetically similar, but one is selected for metS phenotype, while the other (LN) is normal. Genotype of 1536 SNPs, 23 physiological traits (blood pressure, lipid level, blood glucose, body weight, plasma leptin, etc.) and liver RNA-seq.
- Phenotypic QTL (pQTL) mapping:
  - 169 offsprings from F2 intercross. 453 SNPs tagged all haplotypes differing between LH and LN.
  - Data analysis: R/qtl, 5% FDR using permutation testing.
  - Results: 17 pQTL were identified. Two overlapping between traits.
- eQTL mapping: in liver and kidney.
- TFBS analysis: compare the two strains LH and LN, find genes whose promoters/enhancer (also in selected regions) have TFBS disruptions.
- Candidate genes from integrated analysis: intersect multiple sets, including expression correlated with phenotype, within pQTL, having cis-eQTL, and containing disrupted TFBSs between two strains.
- **Remark/Lesson**:
  - The TFBS analysis is not based on eQTL or pQTL (that falls into some enhancers), instead it is based on direct sequence comparison of strains with different phenotypes.
  - Intersecting multiple gene lists to priorize candidate genes.

8.8 Pharmacogenetics and Pharmacogenomics

Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes [GoDARTS, NG, 2011]:

- Backgroud: metaformin is one major front-line drug of T2D. Its target may be AMPK: by activating AMPK through inhibition of the mitochondrial respiratory chain, it signals lack of energy in the cells, and may push cells to absorb glucose more efficiently from blood.
- GWAS of metaformin response:
  - Phenotype: the ability to reduce HbA1c (the most widely used measure of medium-term glycemic control) in the first 18 months of therapy to below 7%.
  - About 700K SNPs in 1,024 Scottish individuals with type 2 diabetes. Replication in two cohorts including 1,783 Scottish individuals and 1,113 individuals from the UK Prospective Diabetes Study.
  - Association test: logistic regression with covariates of metaformin response, such as the baseline HbA1c level.
- Association findings: $\lambda = 1.003$.
  - Only one locus at 11q22 has $P < 1.0 \cdot 10^{-6}$, the strongest SNP rs11212617. Replicated in two independent cohorts, and the combined OR is about 1.35. Explain explains only 2.5% of the variance in metformin response.
  - rs11212617 is not associated with T2D, lipids, and other glucose-related traits.
- Candidate gene:
  - Within the LD region of rs11212617, only ATM is a possible candidate for (1) ATM mutation causes ataxia telangiectasia, some of the patients show insulin resistence; (2) activation or inhibition of ATM alters AMPK activation.
  - ATM inhibition in rat cell lines: affects whether AMPK activity responds to metaformin treatment.
  - ATM function: involved in cell-cycle arrest upon DNA damage and DNA repair. This study shows a link between cancer pathways (both ATM and metaformin are associated with cancer risk), type 2 diabetes and metformin activation of AMPK.

8.9 Misc. Phenotypes

GWAS of exceptional longevity (EL) in humans [Sebastiani & Perls, Science, 2010]:

- Data source: 801 Caucasians subjects in NECS study (age of 95 to 119), 243 NECS controls, and 683 genetically matched Illumina controls (see below). Also a smaller data set as replication data.
- Genetic matching: significant population stratification was observed in randomly chosen controls (GC = 1.3) probably because most of the case subjects immigrated to US in a certain period and may be biased towards certain ethnicity. To address this, the idea is to choose among a large number of controls whose genotypes match those in the cases:
  - Analysis of ancestry: PCA on the subjects (cases and potential controls), and use the first four PCs to cluster subjects into 20 clusters.
  - Selection of controls: for each cluster that is represented in the cases, select controls in that cluster s.t. the ratio of case/control is the same in all clusters.
- Prediction/classification:

– Naive Bayes classifier: for a given set of SNPs Σ_k, a basic classifier can be built:

$$P(EL|\Sigma_k) = \frac{P(EL)\prod_{i=1}^{k} P(SNP_i|EL)}{P(EL)\prod_{i=1}^{k} P(SNP_i|EL) + P(AL)\prod_{i=1}^{k} P(SNP_i|AL)} \tag{8.9}$$

where AL is the average longevity. The conditional probabilities of $P(SNP_i|EL)$ and $P(SNP_i|AL)$ are from the gentoype frequencies in cases and controls.

– Ensemble classifier: rank all SNPs by their posterior probability of association, and create the SNP sets: $\Sigma_1, \Sigma_2, \cdots, \Sigma_K$. The ensemble classifier:

$$P(EL|\Sigma_1, \cdots, \Sigma_K) = \sum_{i=1}^{K} P(EL|\Sigma_i)/K \tag{8.10}$$

In the paper, $K = 150$ is chosen by the performance (spec. and sens. under cross validation).

- Genetic risk profiles: the cases may have distinct genetic signatures that relate to the subtypes of EL, so the cases are clustered according to their risk profiles.
  – Risk profiles: instead of directly cluster of genotypes (150 features), first convert them into risk profiles, defined as $P(EL|\Sigma_1), P(EL|\Sigma_2), \cdots, P(EL|\Sigma_{150})$. Thus different genotypes will be manifested as different profiles/curves.
  – Clustering: the profiles are clustered according to a Bayesian clustering algorithm.
- Significant SNPs (longevity-associated variants, or LAVs): 70 are found and 33 were replicated.
  – Alzheimer's disease (AD): APOE, CTNNA3, STX8 (cell cycle regulator, elevated in AD patients), PLCB3 (enzyme phosphlipase C $\beta 3$ involved in extracellular signals)
  – Insulin signaling: GIP, RAPGEF4, both involved in regulating insuline secretion. However, FOXO1, FOXO3A and IGF-IR showed no significant associations.
  – Chromosomal stability: HJURP
  – Immune response: IL7.
- Comparison with common disease SNPs: only 5 out of 1389 common disease SNPs (from all published studies) show significant associations with EL. Also compare the allele frequencies of the known disease-associated SNPs: no significant difference between cases and controls for the majority of these SNPs.
- EL classification: 150 SNPs are chosen. They are uncorrelated, with the average distance of 8Mb. The ensemble classifier achieves 77% spec. and sens. in the replication set.
- EL subtypes:
  – 19 clusters of 8 or more centenarians in the discovery set; and 11 in the replication set, with 10 of these clusters in both sets.
  – The comparison of clusters: different age distribution, and prevalence/onset of common diseases. Ex. C1 cluster had a significant delay in the onset of cardiovascular disease, dementia and hypertension.
  – Cluster 19 of 17 centenarians: lack most of the LAVs, suggesting there may be many more modifiers of EL to be discovered.

Remark: the paper may have serious methodological flaws:

– Genotyping platforms are different in cases and in controls: this would commonly produce false associations. Need to replicate with the same platform on both cases and controls (not done). The platform on the cases is known to have problems.

- The results seem too good: some SNPs have large effects, and 77% accuracy is extremely high, comparing with other complex traits. Also, the recent meta-analysis of GWAS of longevity does not find any significant SNPs.
- Indication of genotyping problems: For the 70 genome-wide-associated SNPs, the median missing data rate in EL samples was 9%, compared to 3% in controls. The two SNPs with strongest evidence for association, rs1036819 and rs9576827, are far out of HWE.
- Diagnosis of Manhattan plots: the significant SNPs stand out alone, without evidence in the nearby SNPs (which is the usual pattern of the Manhattan plots).

8.10 Personalized Medicine & Clinical Genetics

What is personalized medicine?

- Definition: Personalized medicine is "a form of medicine that uses information about a person's genes, proteins, and environment to prevent, diagnose, and treat disease" (National Cancer Institute 2011).
- Personal data:
  - Genomic data: genotype, DNA sequences, transcriptomics, epigenomics, proteomics, metabolomics, meta-genomics/microbiome.
  - Clinical data: medical records, family health history (FHH), other diagnostic variables.
- Clinical applications:
  - Risk prediction/preventation: from genotype/sequencing data, from FHH, from risk factors (body weight, etc.)
  - Diagnosis: disease markers, especially of subtypes; identification of putative therapeutic targets
  - Prognosis and Treatment: types of treatments, dosage.

Reference: [Green & Guyer, Charting a course for genomic medicine from base pairs to bedside, Nature, 2011], [Chan & Ginsburg, Personalized Medicine: Progress and Promise, ARGHG, 2011], [Personalized medicine: new genomics, old lessons, Offit, Human Genetics, 2011]

Background: understand the biology of genomes:

- Catalog of genetic variations, especially those that are associated with phenotypes. These provide both markers and candidate causal variants of diseases.
- Omics data: DNA modifications (epigenomics), gene products such as RNAs (transcriptomics) and proteins (proteomics), and indirect products of the genome such as metabolites (metabolomics) and carbohydrates (glycomics).
- Functional elements: their functions, and especially the role of non-coding sequences in health and disease.
- Gene regulation: its spatially and temporally dynamic nature presents formidable challenge, as some critical regulatory processes only occur during brief developmental periods or in difficult-to-access tissues.
- Gene networks: network analysis will benefit from understanding the dynamics of gene expression, protein localization and modification, as well as protein-protein and protein-DNA associations. The ultimate challenge will be to decipher the ways that networked genes produce phenotypes.
- Evolutionary relationships: the use of model organisms in functional studies, and diverse data sets from unicellular organisms to mammals.

From genomics to the biology of diseases:

- The power of genomic approaches to elucidate the biology of disease: e.g Crohn's disease. Following GWAS of Crohn's disease, cellular models and animal models have been developed for the effects of causal variants and the knowledge of the relevant biological pathways. Chemical screens have been designed to identify new candidate therapeutic agents.
- GWAS and NGS mapping for complex diseases:
  - Successful examples of GWAS: the complement pathway in age-related macular degeneration, autophagy pathways in Crohn's disease, a number of pathways not evident from the somatic genetics of cancer.
  - Lesson: Human disease susceptibility is the result of rare genetic variants of high penetrance as well as common genomic variants of low penetrance [Offit].
- Non-Medelian inheritance: imprinting, de novo germline mutations, and epigenetic mechanisms of inheritance.
- Catalogues of somatic mutations that contribute to all aspects of tumour biology for each major cancer type are under development.
- Non-coding sequences: GWAS have implicated hundreds of non-coding genomic regions in the pathogenesis of complex diseases.
- Phenotyping: widely accessible databases containing extensive phenotypic information linked to genome sequence data (genotype) are needed (e.g. dbGaP). Such efforts will benefit greatly from the linkage of genomic information to electronic medical/health records.
- The integration of genomic information and environmental exposure data: help to understand the links between biological factors and extrinsic triggers.
- Metagenomics: offers unprecedented opportunities for understanding the role of endogenous microbes and microbial communities in human health and disease.

Personal and genomic data:

- Family health history (FHH): a simple yet invaluable tool for the delivery of personal health risk information. [Guttmacher et al 2004. The family history: more important than ever. N. Engl. J. Med.]
- Personal risk factors: Framingham coronary heart disease model, the Gail model breast-cancer risk assessment [Gail & Greene 2000. Gail model and breast cancer. Lancet]
- Genomewide variation of sequences:
  - Hepatitis C treatment: e.g. a polymorphism on chromosome 19, 3 kb upstream of IL-28B, encoding interferon-lambda-3 was found to be associated with a twofold change in treatment response. [Ge et al. 2009. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature]
  - Statin response: a polymorphism located on SLCO1B1, a gene regulating hepatic uptake of statins, was associated with an increased risk of myopathy following statin treatment (odds ratio 4.5) [Link et al. 2008. SLCO1B1 variants and statin-induced myopathy: a genomewide study. N. Engl. J. Med.]
  - Next-generation sequencing: being applied to understand cancer, rare genetic disease, and microbial infection, etc., with the goal of elucidating functional gene variants.

- Transcriptomics:
  - Cancer classification using gene expression data: microarray data have been used for diagnosis, prognosis, and response to therapy [Parker et al. 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol.] [Van't et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature]
  - Classification in other complex diseases: cardiovascular disease, rheumatic diseases, neurologic diseases such as multiple sclerosis, psychiatric disorders such as schizophrenia, bipolar disorder, and major depression [Goes et al, 2008. The genetics of psychotic bipolar disorder. Curr. Psychiatry Rep.] [Bray, 2008. Gene expression in the etiology of schizophrenia. Schizophr. Bull.]
  - Patterns of differentially expressed miRNAs can conceivably be used clinically in the diagnosis and prognosis of disease.
- Metabolomics:
  - Background: It is estimated that the human metabolome contains approximately 5,000 discrete small-molecule metabolites, and the identification of metabolic changes associated with disease immediately suggests enzymatic drug targets.
  - Application: chronic disease states, such as diabetes, obesity, cardiovascular disease, cancer, and mental disorders [Bain et al. 2009. Metabolomics applied to diabetes research: moving from information to knowledge. Diabetes] [Griffin et al. 2004. Metabolic profiles of cancer cells. Nat. Rev. Cancer] [Newgard et al. 2009. A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. Cell Metab]
  - Metabolomics profiling has also directly been used as a tool in assessing drug toxicity
- Epigenomics:
  - Cancer: hypomethylation in oncogenes and regional hypermethylation in genes that are tumor suppressors
  - Other diseases: Methylation changes have also been shown in other diseases such as type 2 diabetes, cardiovascular pathologies, and autoimmune diseases.

Genomics in personalized medicine:

- Risk Prediction:
  - Mendelian disorders: the variant genes responsible for most Mendelian disorders will be identified and an immediate benefit will be an accurate diagnosis.
  - Cancer: e.g. BRCA1 and BRCA2 and susceptibility to breast cancer [Trainer et al 2010. The role of BRCA mutation testing in determining breast cancer therapy. Nat. Rev. Clin. Oncol.]. Microsatellite instability in mismatch repair genes MLH1 and MSH2 and the early detection of colon cancer.
- Disease diagnosis and molecular characterization:
  - Disease subtypes: genomic and molecular analyses have revealed distinct subtypes of disease, which have been traditionally defined by broad clinical or descriptive phenotypes.
  - Cancer genomics: [Lee et al., 2010. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature] [Pleasance et al. 2010. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature]
  - Frequent novel mutations of the PI3 kinase regulatory subunit gene were found along with an association between MGMT methylation status and mismatch repair mutations in posttreatment GBM [Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 2008]

- Disease prognosis:
  - Tumor-gene-expression signature models: combined with clinically relevant data such as survival outcomes [Van't Veer et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature]
  - Example: Oncotype DX is a 21-gene signature used to predict distant recurrence over 10 years [Sparano et al, 2008. Development of the 21-gene assay and its application in clinical practice and clinical trials. J. Clin. Oncol]
- Treatment:
  - New drug targets: development of drugs based on genomic knowledge is becoming increasingly commonplace, particularly for cancer drug development, e.g. HER2, Bcr-Abl inhibitor.
  - Patient stratification: using genomic information can aid clinical trials (allow the use of smaller numbers of participants and increase statistical power). Correlation of genomic signatures with therapeutic response will enable the targeting of appropriate patients at appropriate stages of their illness.
- Pharmacogenomics:
  - Resistance to cancer treatment: Two studies using differential gene expression found that genes involved in chemoresistance were also associated with a worse prognosis [Holleman et al. 2004. Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. N. Engl. J. Med.]
  - Response to cancer treatment: human epithelial growth factor receptor (HER2), was shown to be amplified in 25%-30% of breast cancers, and its overexpression correlated with a worse prognosis. Trastuzumab, a monoclonal antibody that targets HER2, is effective in reducing tumor burden.
  - HIV: genetically guided prescription of the antiretroviral drug abacavir is now the standard of care for HIV-infected patients.
  - An important example of pharmacogenetics is in the management of warfarin therapy [Klein et al. 2009. Estimation of the warfarin dose with clinical and pharmacogenetic data. N. Engl. J. Med.]
- Monitoring Disease Response to Therapy:
  - Peripheral blood mononuclear cell (PBMC) gene expression profiling (a set of 11 genes) is now used routinely in some centers to monitor the status of grafts following solid organ transplantation [Deng et al. 2006. Noninvasive discrimination of rejection in cardiac allograft recipients using gene expression profiling. Am. J. Transplant.]

Gene-environment interactions and Microbiomics:

- Gene-environment interactions:
  - A host-based gene expression signature was recently identified that may someday be used for early detection of viral infection. The same data were also used to show that gene expression patterns distinguish between bacterial and viral infections
  - Gene expression patterns in host cells also have been shown to change given other environmental stressors such as smoking as well as during disease states like asthma and chronic obstructive pulmonary disease [Seibold & Schwartz 2011. The lung: the natural boundary between nature and nurture. Annu. Rev. Physiol.]
- Pathogen genomes:

- Virulence: Comparative proteomic analysis between virulent and avirulent phenotypes [Bechah et al. 2010. Genomic, proteomic, and transcriptomic analysis of virulent and avirulent Rickettsia prowazekii reveals its adaptive mutation capabilities. Genome Res.]
- Resistance: how pathogens develop resistance, track the genomic changes over time in a patient after exposure to antibiotic treatment.
- Several diseases have been associated with large-scale imbalances in the gut microbiome, including inflammatory bowel disease, antibiotic-resistant diarrhea, and obesity [Ley et al 2006. Microbial ecology: human gut microbes associated with obesity. Nature]
- The Human Microbiome Project: creating not just a database that contains disease-causing pathogens, but also a control set of normal flora [Peterson et al. 2009. The NIH human microbiome project. Genome Res]

Areas/Challenges of personalized medicine:

- Association mapping with sequencing data: rare variants.
- Role of gene regulation and non-coding sequences in complex diseases, including cancer.
- Integrative mapping of complex diseases: with multiple types of dataset, e.g. sequences, transcriptomics and epigenoimcs. This strategy has been used in cancer genomics (TCGA).
- Gene networks and diseases: how the changes of genes affect the gene networks, and how the network changes lead to phenotypic changes. This is related to the study of epistatis (gene interactions). Ex. the effect of regulatory genes may be mediated through (more direct) effector genes.
- Gene-environment interaction: environmental variables may be manifested as other genomic data (e.g. expression, or metagenomic).
- Risk prediction and diagnosis with multiple types of personal data: genotype/sequencing data, electronic record and family history.
- Functional impact prediction of sequence changes: important for risk prediction.
- Patient stratification and disease subtyping: use genomic features to identify subtypes and use them for treatment.
- Drug target development: prediction of what genes may serve as good therapeutic targets.

Superheroes of disease resistance [NBT, 2016], on [Chen, NBT, 2016]

- A large number of samples, 1/2 Million: find all candidates with mutations in severe Mendelian diseases.
- Filtering: by AF, health of individual carriers, Sanger sequencing, penetrance of the mutations. Found 13 individuals resilient to one of eight Mendelian diseases.
- Difficulty of identifying modifier variant: for each variant (highly penetrant), perhaps only a few individuals in 1/2M samples contain this variant, thus little power of finding modifier loci.

Phenome-Wide Association Studies as a Tool to Advance Precision Medicine [Denny and Roden, ARGHG, 2016]

- Resources: eMERGE, China Kadoorie Biobank, GERA cohort, MVP.
- PheWAS: association of a SNP with many phenotypes.
- Phenotype definition: ICD codes from bills, most commonly used. Exist algorithms to map ICD to phenotypes: phewascatalog.org. Other data: Laboratory data, medication records such as endophenotypes and drug response.

- Application of PheWAS: often replicate existing association results, also report associations with new phenotypes.
- Other uses of EHR: define disease comorbidities, e.g. periodental disease with T2D and hypertension. Define disease subtypes: e.g. T2D subtypes associated with distinct traits, also different genetic variants.
- Challenges: multiple testing burden. Separate true pleiotropic effects with shared clinical comorbidity.

8.11 Genetics of Model Organisms

Using model organisms to establish phenotypic role of a gene:

- Need both deletion phenotype and genetic rescue (introducing the original gene to see if it restores the phenotype). Need rescue because the deletion experiment may introduce other unwanted changes.
- Specificity of phenotype: e.g. a mutation may affect growth, and thus obesity phenotype.
- Genetic interactions to explore pathways: if $A \rightarrow B$, then B would dominate the effect of A.

The future of model organisms in human disease research [NRG, 2011]

- Using model organism to gain a better understanding of fundamental biological processes that are related to human health. Ex. epistasis map from yeast; study the pathway of blood vessle generation in human using yeast (orthologous system).
- Model organisms may have advantages of gene mapping including population structure and experiment design: balanced alleles to avoid the rare allele issue, etc.
- Model organisms also have advantages of functional studies: mainly access to tissues, e.g. eQTL on multiple tissues, integration with phenotypic data.
- Model organsisms as model system to study the function of genes/variants, especially in the context of complex behavior/phenotypes.

Variation of sporulation efficiency in natural yeast strains [Gerke & Cohen, Science, 2009]:

- Problem: sporulation efficiency in oak tree strains and vineyard strains are very different (nearly 100% vs 3.5%). What is the genetic basis?
- QTL analysis (by crossing the two strains) identified 5 QTLs, out of which 3 could explain most of the effect ($R^2 = 0.87$), allowing two- and three way interactions between loci.
- Mapping nucleotide changes: candidate genes are relatively easy to locate within about 50-100 kb confidence intervals of the QTLs. (1) Rme1 (sporulation regulator): non-coding region substiution; (2) Ime1 (sporulation regulator): one subsitution in coding sequences (involved in PPI with other sporulation regulators), and the other in non-coding region; (3) Rsf1 (activator of mitochondrial genes): coding sequence substitution. May promoter Ime1 expression (which is sensitive to respiratory signal).
- Discussion:
  - Selection of sporulation efficiency in woodland environment, but not in vineyard. Consistent with the protein polymorphism of Ime1 (relaxed selection in vineyard strain).
  - Loci of a complex trait: coding sequence of regulators; noncoding/regulatory DNA sequence of regulators; genes that may influence the expression of regulators.

Rsu1 regulates ethanol consumption in Drosophila and humans [Ojelade, PNAS, 2015]

- Gal4 system in fruit fly: use P element containing Gal4 to delete a gene. Then introduce UAS-transgene (UAS: upstream activation sequence): activation by Gal4. TG is only expressed in individuals with P-element insertion.
- Background: fMRI. Comparison between two conditions. Visualization of fMRI images: multiple views. ROI (region of interest) analysis: extract β for each pixels (voxel), then average. Need to control for covariates such as gender.
- Joint CV-RV analysis using kernel method.