Advanced Gene Mapping Course

The Rockefeller University

January 27-31, 2025

Lectures

Table of Contents

Lectures given by: [1]Heather Cordell, [2]Suzanne Leal, [3]Wayne Patterson, [4]Gao Wang, [5]Andrew DeWan, [6]Jurg Ott, and [7]Shamil Sunyaev

Genome-wide association studies (GWAS) - Part 1

Heather J. Cordell

Population Health Sciences Institute
Faculty of Medical Sciences
Newcastle University, UK

heather.cordell@ncl.ac.uk

Genome-wide association studies (GWAS)

- Popular (and highly successful) approach over past $\sim$ 18 years
- Enabled by advances in high-throughput (microarray-based) genotyping technologies
- Idea is to measure the genotype at a set of single nucleotide polymorphisms (SNPs) across the genome, in a large set of unrelated individuals
  - Cases and controls
  - Or population cohort measured for relevant quantitative phenotypes (height, weight, blood pressure etc)
  - Or related individuals (family data) – but need to analyse differently

Genome-wide association studies (GWAS)

Two individuals

Person 1 ACCTGTGTGCCCAATGGCGTCCCATACTATCGG
ACCTGTGCGCCCAATGGCGTCCCATACTATCGG

Person 2 ACCTGTGCGCCCAGTGGCGTCCCATACTATCGG
ACCTGTGCGCCCAGTGGCGTCCCATAGTATCGG

- Test each SNP for association/correlation with disease or quantitative phenotype

Association testing: case/control studies

- Collect sample of affected individuals (cases) and unaffected individuals (controls)
  - Or a else a sample of random "population" controls
    - Most of whom will not have the disease of interest
- Examine the association (correlation) between alleles present at a genetic locus and presence/absence of disease
  - By comparing the distribution of genotypes in affected individuals with that seen in controls

Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2|2	500 $(= a)$	200 $(= b)$
1|2	1100 $(= c)$	820 $(= d)$
1|1	400 $(= e)$	980 $(= f)$
Total	2000	2000

Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2|2	500 $(= a)$	200 $(= b)$
1|2	1100 $(= c)$	820 $(= d)$
1|1	400 $(= e)$	980 $(= f)$
Total	2000	2000

- Test for association (correlation) between genotype and presence/absence of disease using standard χ^2 test for independence on 2 df

Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2|2	500 $(= a)$	200 $(= b)$
1|2	1100 $(= c)$	820 $(= d)$
1|1	400 $(= e)$	980 $(= f)$
Total	2000	2000

- Test for association (correlation) between genotype and presence/absence of disease using standard χ^2 test for independence on 2 df
  - Defined as $\sum_{i=1,6} \frac{(O_i - E_i)^2}{E_i}$ where O_i and E_i are observed and expected counts (calculated from the row and column totals) respectively
  - Generates a p value indicating how significant the association/correlation appears to be

Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2|2	500 $(= a)$	200 $(= b)$
1|2	1100 $(= c)$	820 $(= d)$
1|1	400 $(= e)$	980 $(= f)$
Total	2000	2000

- Test for association (correlation) between genotype and presence/absence of disease using standard χ^2 test for independence on 2 df
  - Defined as $\sum_{i=1,6} \frac{(O_i - E_i)^2}{E_i}$ where O_i and E_i are observed and expected counts (calculated from the row and column totals) respectively
  - Generates a p value indicating how significant the association/correlation appears to be
- Two odds ratios can be estimated
  - $\mathrm{OR}(2|2 : 1|1) = \frac{af}{be}$
  - $\mathrm{OR}(1|2 : 1|1) = \frac{cf}{de}$

Odds ratios

- Odds of disease are defined as P(diseased)/P(not diseased)
  - Odds ratio $\mathrm{OR}(2|2 : 1|1)$ repesents the factor by which your odds of disease must be multiplied, if you have genotype 2|2 as opposed to 1|1
    - i.e. the 'effect' of genotype 2|2

Odds ratios

- Odds of disease are defined as P(diseased)/P(not diseased)
  - Odds ratio $\mathrm{OR}(2|2 : 1|1)$ repesents the factor by which your odds of disease must be multiplied, if you have genotype 2|2 as opposed to 1|1
    - i.e. the 'effect' of genotype 2|2
- Similarly, we can define the OR for 1|2 vs 1|1
  - As the factor by which your odds of disease must be multiplied, if you have genotype 1|2 as opposed to 1|1
    - i.e. the 'effect' of genotype 1|2

Odds ratios

- Odds of disease are defined as P(diseased)/P(not diseased)
  - Odds ratio $\mathrm{OR}(2|2 : 1|1)$ repesents the factor by which your odds of disease must be multiplied, if you have genotype 2|2 as opposed to 1|1
    - i.e. the 'effect' of genotype 2|2
- Similarly, we can define the OR for 1|2 vs 1|1
  - As the factor by which your odds of disease must be multiplied, if you have genotype 1|2 as opposed to 1|1
    - i.e. the 'effect' of genotype 1|2
- ORs are closely related (often $\approx$) genotype relative risks
  - The factor by which your probability of disease must be multiplied, if you have genotype 1|2 as opposed to 1|1 (say)
- If your genotype has no effect on your probability (and therefore on your odds) of disease, then the ORs=1.
  - So the association test can be thought of as a test of the null hypothess that the ORs=1

Genotype relative risks

- If a disease is reasonably rare, the odds ratio approximates the genotype relative risk (GRR, RR)

Genotype	Penetrance	GRR	Odds	OR
1/1	0.01	1.0	$0.01/0.99 = 0.0101$	1.00
1/2	0.02	2.0	$0.02/0.98 = 0.0204$	2.02
2/2	0.05	5.0	$0.05/0.95 = 0.0526$	5.21

- If your genotype has no effect on your probability (and therefore your RR) of disease, then both the ORs and the GRRs=1.

Dominant/recessive effects

Dominant:

Genotype	Cases	Controls	Total
2|2 and 1|2	500+1100	200+820	700+1920
1|1	400	980	1380
Total	2000	2000	4000

Recessive:

Genotype	Cases	Controls	Total
2|2	500	200	700
1|2 and 1|1	1100+400	820+980	1920+1380
Total	2000	2000	4000

- Can also rearrange table to examine effects of alleles (1 df tests):

Counting alleles

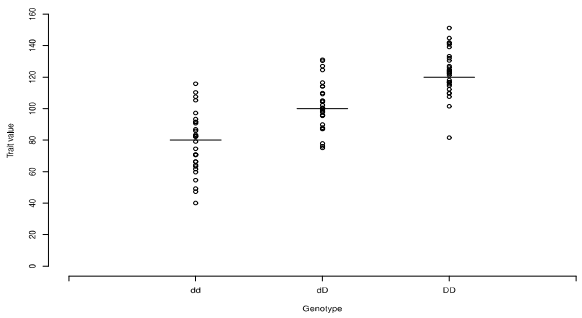
	Counts in	
Allele	Cases	Controls
2	2100 (=a)	1220 (=b)
1	1900 (=c)	2780 (=d)
Total	4000	4000

Allelic OR = ad/bc

- χ^2 test statistic on 1 df = $\sum_i (O_i - E_i)^2/E_i$ where O_i and E_i are the observed and expected values in cell i.
  - Assumes HWE under null and multiplicative allelic effects under alternative: considers chromosomes as independent units
  - Better approach: use counts in previous genotype table to perform a Cochran-Armitage trend test
  - Even better approach: use linear or logistic regression

Testing for association: quantitative traits

- Linear regression provides a natural test for quantitative traits
  - Fit a regression line $y = mx + c$
  - Test the null hypothesis that the slope $m = 0$



Logistic regression

- Used in case/control studies
  - Outcome is affected or unaffected
  - Model probability (and thus odds) of disease p as function of variable x (usually coded (0,1,2)) coding for genotype:

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x \quad \equiv c + mx$$

  - Use observed genotypes in cases and controls to estimate the values of regression coefficients β_0 and β_1
    - And to test whether $\beta_1 = 0$

Logistic regression

- Standard method used in standard epidemiological studies e.g. of risk factors such as smoking in lung cancer
- Main advantage is you can include more than one predictor in the regression equation e.g.

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where x_1, x_2, x_3 code for

- genotypes at 3 loci
- measured environmental covariates (e.g. age, sex, smoking etc),
- genetic principal component scores (to adjust for population substructure),
- interactions between loci etc. etc.

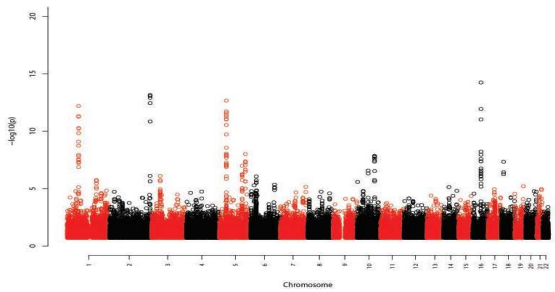
Testing for association

- All methods produce a test statistic and a p value at each SNP, indicating how significant the association/correlation observed appears to be
  - i.e. how likely it was to have occurred by chance
  - The threshold to declare 'genome-wide significance' is usually around $p = 5 \times 10^{-8}$
    - To account for multiple testing of many SNPs across the genome

Testing for association

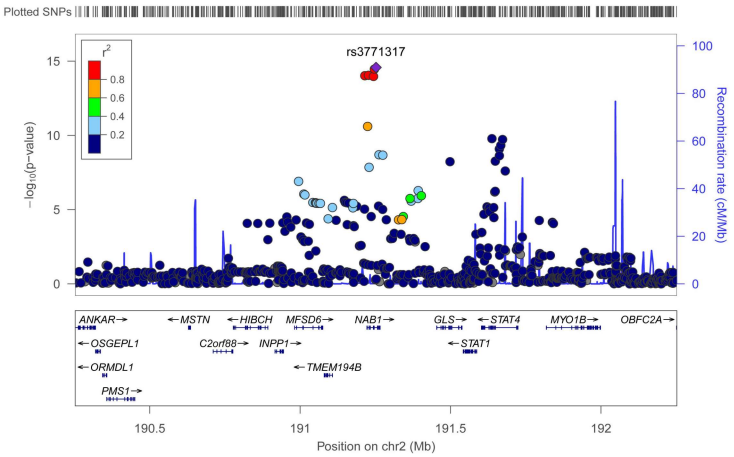
- All methods produce a test statistic and a p value at each SNP, indicating how significant the association/correlation observed appears to be
  - i.e. how likely it was to have occurred by chance
  - The threshold to declare 'genome-wide significance' is usually around $p = 5 \times 10^{-8}$
    - To account for multiple testing of many SNPs across the genome
- Alternative (Bayesian) methods produce a Bayes Factor
  - Indicates how likely the data is under the alternative hypothesis (of association between genotype and phenotype)
    - Compared to under the null hypothesis (of no association between genotype and phenotype)
  - Requires you to make some prior assumptions regarding the likely strength of associations (i.e. the value of the β's)
  - Choosing a sensible threshold (e.g. $\log_{10}$ BF> 4) requires you to make some prior assumptions regarding what proportion of SNPs in the genome are likely to be associated with the phenotype

Manhattan Plots



- At any location showing 'significant' association, we expect to see several SNPs in the same region showing association/correlation with phenotype
  - Due to the correlation or linkage disequilibrium (LD) between neighbouring SNPs

Close-up of hit region



Historical Perspective: Complement Factor H in AMD

- First (?) GWAS was by Klein et al. (2005) Science 308:385-389
- Typed 116,204 SNPs in 96 cases (with age-related macular degeneration, AMD) and 50 controls
  - Very small sample size – they were very lucky to find anything!
  - Luck was due to the fact the polymorphism has a very large effect (recessive OR=7.4)
- Klein et al. followed up on two SNPs passing threshold ($p < 4.8 \times 10^{-7}$)
  - Plus a third SNP that just failed to pass significance threshold, but lay in same region as first SNP

Complement Factor H in AMD

- Of the 3 SNPs followed up:
  - One appeared to be due to genotyping errors: significance disappeared on filling in some missing genotypes
  - First and third SNP lie in intron of Complement Factor H (*CFH*) gene
    - Lies in region previously implicated by family-based linkage studies
- Resequencing of the region identified a polymorphism of plausible functional effect
- Immunofluorescence experiments in the eyes of AMD patients supported the involvement of *CFH* in disease pathogenesis.

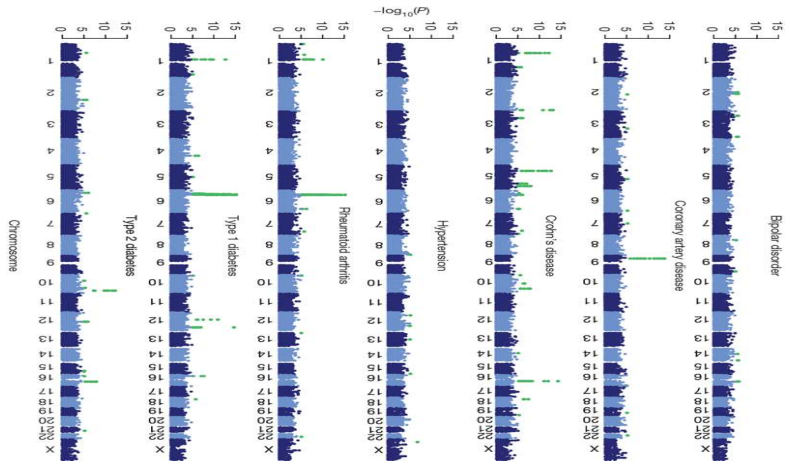
GWAS

- GWAS really got going in around 2007
  - Visscher et al. (2012) AJHG 90:7-24 "Five Years of GWAS Discovery"
  - Visscher et al. (2017) AJHG 101:5-22 "10 Years of GWAS Discovery: Biology, Function and Translation"
  - Abdellaoui et al. (2023) AJHG 110:179-194 "15 Years of GWAS Discovery: Realizing the promise"
- 2007/2008 saw a slew of high-profile GWAS publications
  - Breast cancer (Easton et al. 2007)
  - Rheumatoid Arthritis (Plenge et al. 2007)
  - Type 1 and Type 2 diabetes (Todd et al. 2007; Zeggini et al. 2008)
- Arguably the most influential was the Wellcome Trust Case Control Consortium (WTCCC) study of 7 different diseases
  - http://www.wtccc.org.uk/

WTCCC

- Nature 447: 661-678 (2007)
- Considered 2000 cases for each of the following diseases:
  - Bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, type 2 diabetes
- Compared each disease cohort to common control panel
  - 3000 population-based controls
  - From 1958 birth cohort and National Blood Service
- Highly successful
  - WTCCC found 24 separate association signals
  - Including highly convincing signals in 5 out of the 7 diseases studied
  - All were replicated in subsequent independent follow-up studies

Manhattan plots for 7 diseases



Lessons from WTCCC (and others)

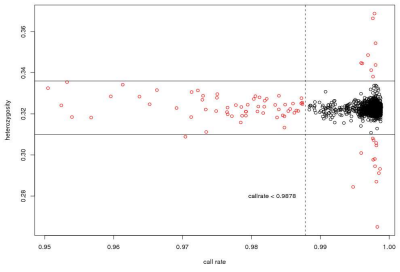
- Typically used rather standard statistical/epidemiological methods (χ^2 tests, t tests, logistic regression etc.)
- Success largely due to:
  - An appreciation of the importance of large sample size ($>$ 2000 cases, similar or greater number of controls)
  - Stringent quality control procedures for discarding low-quality SNPs and/or samples
  - Stringent significance thresholds ($p = 5 \times 10^{-8}$) to account for multiple testing and/or low prior prob of true effect
  - Importance of replication in an independent data set

Short break

Quality Control

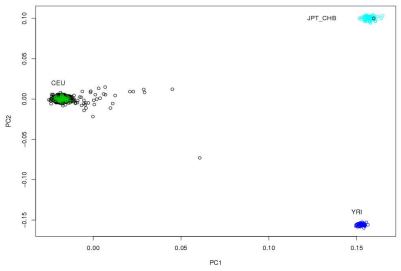
- Stringent QC checks are required for GWAS data
- Discard samples (people) deemed unreliable
  - Low genotype call rates, excess heterozygosity etc.
  - X chromosomal markers useful for checking gender
    - Males should 'appear' homozygous at all X markers
  - Genome-wide SNP data useful for checking relationships and ethnicity
- Discard data from SNPs deemed unreliable
  - On basis of genotype call rates, Mendelian misinheritances, Hardy-Weinberg disequilibrium
  - Exclude SNPs with low minor allele frequency (MAF)
- See tutorials at:
  - https://pubmed.ncbi.nlm.nih.gov/21085122/
  - https://pubmed.ncbi.nlm.nih.gov/29484742/

QC: call rates and heterozygosity



- 61 sample exclusions (low call-rate); 23 exclusions (heterozygosity)
- SNP exclusions also made based on call-rates, MAF and Hardy-Weinburg equilibrium (HWE)

QC: ethnicity tests



- Multidimensional scaling (with 210 HapMap individuals) identifies 33 samples with non-Caucasian ancestry
- MDS or similar multivariate methods can also be used to model more subtle population differences between samples...

Multivariate Analysis

- Several related multivariate analysis techniques have been proposed for detecting population structure in genome-wide association studies
  - Principal components analysis (PCA)
  - Principal coordinates analysis (PCoA)
  - Multidimensional scaling (MDS)

Multivariate Analysis

- Several related multivariate analysis techniques have been proposed for detecting population structure in genome-wide association studies
  - Principal components analysis (PCA)
  - Principal coordinates analysis (PCoA)
  - Multidimensional scaling (MDS)
- If population differences can be detected (and adjusted for) in association analysis, this offers a way to deal with the problem of population stratification
  - Population sampled actually consists of several 'sub-populations' that do not really intermix
  - Can lead to spurious false positives (type 1 errors) in case/control studies

Multivariate Analysis

- Several related multivariate analysis techniques have been proposed for detecting population structure in genome-wide association studies
  - Principal components analysis (PCA)
  - Principal coordinates analysis (PCoA)
  - Multidimensional scaling (MDS)
- If population differences can be detected (and adjusted for) in association analysis, this offers a way to deal with the problem of population stratification
  - Population sampled actually consists of several 'sub-populations' that do not really intermix
  - Can lead to spurious false positives (type 1 errors) in case/control studies
- These techniques can also be used in quality control (QC) procedures, to check for (and discard) gross population outliers

Multivariate Analysis

- Start with a normalised matrix of SNP genotype data corresponding to the genotypes at L loci (=rows) for n individuals (=columns)

$$\mathrm{X} = \begin{pmatrix} x_{11} & x_{12} & . & x_{1n} \\ x_{21} & x_{22} & . & x_{2n} \\ x_{31} & x_{32} & . & x_{3n} \\ . & . & . & . \\ . & . & . & . \\ x_{L1} & x_{L2} & . & x_{Ln} \end{pmatrix}$$

Multivariate Analysis

- Start with a normalised matrix of SNP genotype data corresponding to the genotypes at L loci (=rows) for n individuals (=columns)

$$\mathrm{X} = \begin{pmatrix} x_{11} & x_{12} & . & x_{1n} \\ x_{21} & x_{22} & . & x_{2n} \\ x_{31} & x_{32} & . & x_{3n} \\ . & . & . & . \\ . & . & . & . \\ x_{L1} & x_{L2} & . & x_{Ln} \end{pmatrix}$$

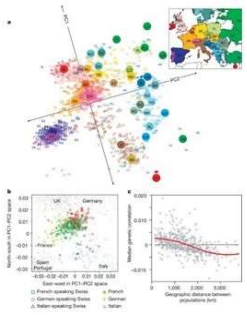
- Estimate covariance matrix $\Psi = \mathrm{X}^T\mathrm{X}$ between all pairs of individuals, with entries ψ_{ij} defined as the covariance (summing over SNPs) between column i and j of X
  - Represents average genome-wide identity by descent (IBD) (estimated from identity by state, IBS)
  - Compute the eigenvectors $\vec{v}_j$ and eigenvalues λ_j of matrix Ψ
    - Co-ordinate j of the kth eigenvector represents the ancestry of individual j along 'axis' k
- For technical details, see McVean (2009) PLoS Genetics 5;10:e1000686

Multivariate Analysis

- Many genetics packages e.g. (PLINK) will allow you to calculate the top 10 (or more) PCs
  - Different geographic populations can often be well separated by just the first two or three PCs
    - Useful for outlier detection
  - For more subtle differences, you may need to calculate more PCs
    - And include them as covariates in the regression equation
    - Post-GWAS QC can determine whether you have included 'enough'

Principal components analysis (PCA)

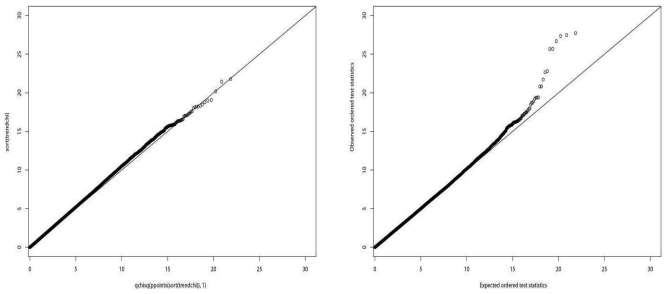
Genes mirror geography within Europe



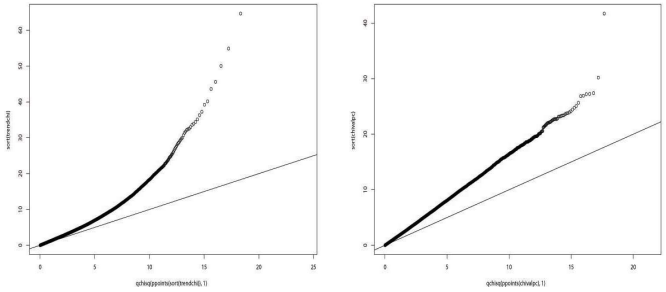
J Novembre *et al.* (2008) *Nature* **456(7218):98-101**, doi:10.1038/nature07331

Post GWAS QC: Q-Q Plots (good)

- Plot ordered test statistics (y axis) against their expected values under the null hypothesis (x axis)



Q-Q Plots (bad)



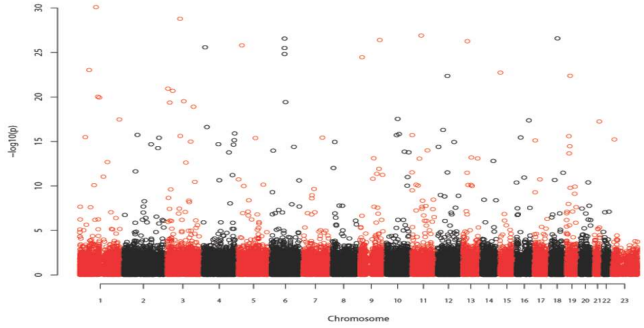
Population stratification

- A QQ plot showing constant inflation (straight line with slope > 1) can indicate population stratification/population substructure
- Simple solution: Genomic Control (Devlin and Roeder 1999)
  - Use your observed test statistics to estimate the slope (=inflation factor λ)
  - Divide each test statistic by λ to get an adjusted (deflated) test statistic
- More complicated solution: use PCA/MDS or similar
- Even more complicated solution: use linear mixed models

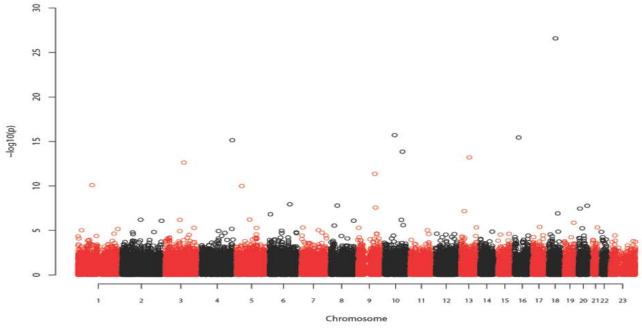
Relatedness

- With genome-wide data, can also infer relationships based on average identity by descent (IBD) $\Psi = X^T X$ or identity by state (IBS)
  - Using 'thinned' subset of markers with high minor allele frequency (MAF) and in approximate linkage equilibrium
  - Simple relationships (PO, FS, MZ/duplicates) can identified with only a few hundred markers
  - More complicated relationships require 10,000-50,000 SNPs
- Various software packages, including PLINK, KING and TRUFFLE

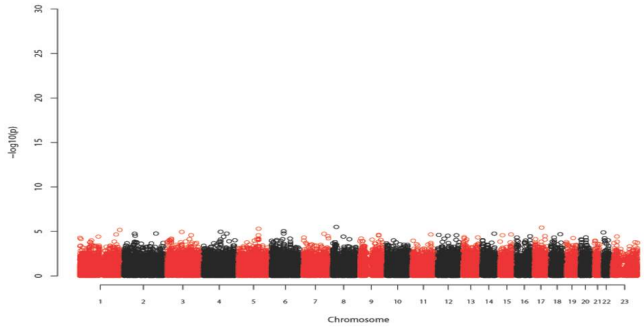
CHD GWAS results (low QC)



CHD GWAS results (better QC)



CHD GWAS results (final QC)



Genome-wide meta-analysis

- Puts together data (or results) from a number of different studies
  - Could analyse as one big study
  - But preferable to analyse using meta-analytic techniques
    - At each SNP construct an overall test based on the results (log ORs and standard errors) from the individual studies

Genome-wide meta-analysis

- Puts together data (or results) from a number of different studies
  - Could analyse as one big study
  - But preferable to analyse using meta-analytic techniques
    - At each SNP construct an overall test based on the results (log ORs and standard errors) from the individual studies
- Meta-analysis is often made easier by using *imputation*
  - Inferring (probabilistically) the genotypes at SNPs which have not actually been genotyped
    - On the basis of their known correlations with nearby SNPs that have been genotyped
    - Using a reference panel of people (e.g. 1000 Genomes) who have been genotyped at all SNPs

Genome-wide meta-analysis

- Puts together data (or results) from a number of different studies
  - Could analyse as one big study
  - But preferable to analyse using meta-analytic techniques
    - At each SNP construct an overall test based on the results (log ORs and standard errors) from the individual studies
- Meta-analysis is often made easier by using *imputation*
  - Inferring (probabilistically) the genotypes at SNPs which have not actually been genotyped
    - On the basis of their known correlations with nearby SNPs that have been genotyped
    - Using a reference panel of people (e.g. 1000 Genomes) who have been genotyped at all SNPs
- Enables meta-analysis of studies that used different genotyping platforms
  - By imputing to generate data at a common set of SNPs
    - Ideally while accounting for the imputation uncertainty in the downstream statistical analysis
    - In practice often don't bother - use post-imputation QC to remove poorly-imputed SNPS

Data Quality Control
Sequence and Genotype Array Data

Suzanne M. Leal, Ph.D.
sml3@cumc.columbia.edu

DNA Collection

- Blood samples
  - –When sampled most be frozen or extracted within several days
  - –For unlimited supply of DNA
    - Transformed cell lines
      - –Is expensive
- Buccal Swabs
  - –Small amounts of DNA
  - –DNA not stable
- Saliva (Origene collection kit)
  - –Can be stored at room temperature for two years before extraction

Measurement of DNA Concentrations

- Nanodrop
- Picogreen

Effect of Genotyping Error – Same Error Rates for Cases and Controls

- For family-based association studies - Trios
  - –Can increase both type I and II error
- Population based studies
  - –Increases type II error only

Quantitative Traits
If genotyping error is not correlated with trait values only type II errors will be increased

Effects of Genotype Errors

- Cases and controls are sequenced/genotyped
  - –At different times
  - –Different institutions
  - –One group, e.g., case or control, is predominately sequenced/genotyped in the same batch
- Can lead to different genotyping error rates in cases and controls
  - –In this situation both type I and II error can be increased
- If sequencing/genotyping cases and controls
  - –Randomize cases and controls so they are spread evenly across batches

Effects of Genotype Errors

- If genotyping error is correlated with quantitative trait values, it will also increase type I and II errors, e.g.,
  - –Individuals with elevated systolic blood pressure are genotyped in one batch
  - –Those with systolic blood pressure within the normotensive range in another batch
- If genotype errors are not correlated with trait values, i.e., random
  - –Will only increase type II errors

Solutions - Genotyping Errors

- To avoid batch effects increasing type I and II errors
  - In addition to QC
    - Include batch as a covariate
  - In cases of heterogeneity between batches
    - Analyze each batch separately and combine results via meta-analysis
      - First should test for homogeneity of the data
        - » When heterogeneity is high
          - » May not make sense to perform meta-analysis across all batches

Genotype SNPs (~20-96) before Exome or Whole Genome Sequencing

- Genotype markers which can be used as DNA fingerprint
- Allows for assessment of DNA quality
- Aids in determining the the genetic sex of study subjects
  - To aid in identification of potential sample swaps
- Detects cryptic duplicates
- For family data
  - Aids in determining close familial relationships
    - Non-paternity
    - Sample swaps
    - Cryptic relationships

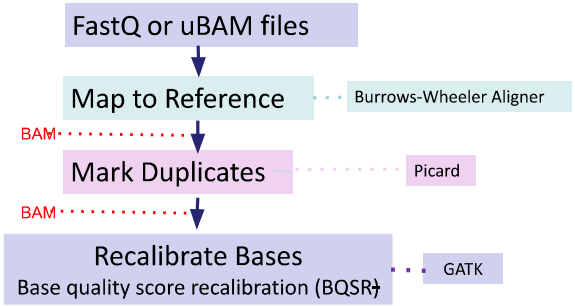
Detecting Genotyping Errors

- Duplicate samples genotyped using arrays to detect inconsistencies
  - Can use duplicate samples that are inconsistent to adjust clusters to improve allele calls
    - Will not detect systematic errors
- Usually generated only for genotype array data
  - Due to expense, duplicate samples are usually not generated for exome or whole genome sequencing studies

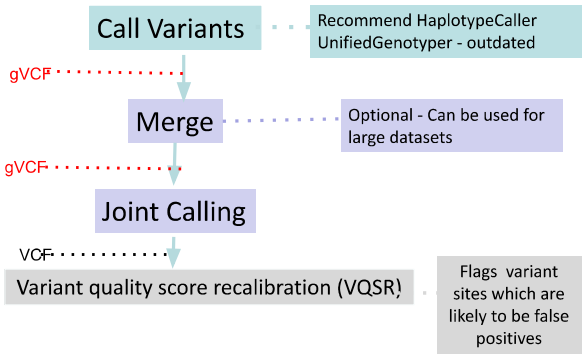
QC for Genotype Array vs. Sequence Data

- Same
  - Evaluation of genomic sex and comparing it to reported gender
  - Detection of cryptic duplicate samples
  - Determination of genetic relationships
  - Evaluation of ancestry and removal of outliers
  - Testing of variants for deviation from Hardy-Weinberg equilibrium (HWE)
  - Post analysis examine quantile-quantile plots (QQ)
- Different
  - Removal of genotypes due to low read-depth, genotype quality, etc.
  - Values used to remove samples and variants due to missing data

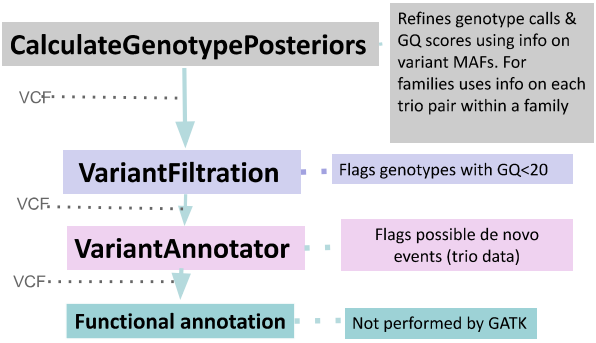
Variant Calling Pipeline -Step 1 Preprocessing



Variant Calling Pipeline-Step 2 Variant Discovery



Variant Calling Pipeline - Step 3 Call Set Refinement



Software to Detect Structural Variation

- Exome data copy number variation
  - –CoNIFER (Copy Number Inference From Exome Reads)
    - Krumm et al. 2012
  - –XHMM
    - Fromer et al. 2014
- WGS data structural variation
  - –MetaSV
    - Mohiyuddin et al. 2015
  - –LUMPY
    - Layer et al. 2014
  - –GATK-gCNV
    - Babadi et al. 2023
  - –GATK-SV

Variants with more than 2 Alleles

- Genetic analysis tools are usually developed to analyze variant sites that are diallelic
- Some variant sites may have >2 alleles
- The alleles at these sites need to be split
  - –New loci are made each retaining the same reference allele from the original sites
    - Each new locus only has 2 alleles
- Multiallelic sites can have higher error rates compared to diallelic sites

Insertions and Deletions

- For insertions and deletions
  - –Left normalization is performed
- These too also usually have more than two alleles
- Higher error rates than single nucleotide variants (SNVs)

Variant Calling

- BAM files are large and take considerable resources
  - –Storage is expensive
  - –One 30x whole genome is ~80-90 gigabytes
  - –A small study of 1,000 samples will consume 80 terabytes of disk space
- The cost of cloud computing to call variants
  - –(Souilmi et al. 2015)
  - –$5 per exome
  - –$50 per genome
    - For 1,000 samples
      - ~$5,000 exome
      - –$50,000 genome

Working with gVCF Files

- When combining data from different sources
  - –Best to obtain BAM files and realign and call variants
- If BAM files are unavailable
  - –Instead of obtaining VCF files
    - Obtain gVCF files
      - –To perform joint calling and complete the GATK pipeline
    - A whole genome gVCF
      - –~1 Gigabyte
        - »1/100th the size of a BAM file for one individual

Influences on Sequence Quality

- DNA quality
  - –Age of sample
  - –Extraction method
  - –Source of sample
    - e.g., blood, skin punch, buccal
- Read length
  - –Short read or long read technology
- Sequencing machines
- Median sequencing depth
- Alignment
- Variant calling method used and variant type
  - –Single nucleotide variants and insertion/deletions
  - –Structural variants

NGS Data Quality Control

- For exome/whole-genome sequence data QC is data specific
  - –Dependent on read depth
  - –Batch effects
  - –Availability of duplicate samples
  - –etc.
- Whole-genome sequence data
  - –Usually 30x
- Exome sequence data can have a wide-range of read depths
  - –e.g., 20x, 100x
- Note higher read depth is necessary for exome sequence data compared to whole genome sequence data
  - –Due to having uneven read depth coverage across the exome

NGS Data Quality Control

- GATK - Variant Quality Score Recalibration (VQSR)
  - –Used to determine variant sites of bad quality
    - Variant site is a false positive call
- Even after this step
  - –Concordance of duplicates (when available) and
  - –and Ti/Tv ratios are often low
- Additional QC needs to be performed

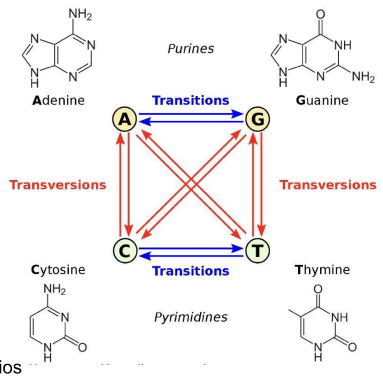
bcfTools

- Performs allele splitting and left normalization
- Remove genotypes, variants sites, and samples
- Calculate Ti/Tv ratios

NGS Data Quality Control

- Values which are used for DP, GQ, and missing data cut offs are based upon
  - –Concordance rates
    - If there are duplicate samples are available
  - –Ti/Tv ratios
    - By individual
    - By batch
    - Entire data set
  - –Amount of data removed
    - QC can remove substantial amounts of data which should be avoided
      - –e.g., >15% of variant sites

Transition/Transversion (Ti/TV) Ratios



AKA Ts/Tv ratios

Transition/Transversion (Ti/TV) Ratios

- Ti/Tv Ratios
  - Whole genome ~2.0
  - Exome novel ~2.7
  - Exome known ~3.5
- Ti/Tv ratios can be calculated for a
  - Sample or
  - Dataset
- Ti/Tv ratios can also be evaluated for subsets of data
  - e.g., by batch

Sequence QC
Removal of Genotype Calls and Samples

- Sequence depth of coverage
  - –DP_variant
    - High DP could be an indication of copy number variants
      - –Which can introduce false positive variant calls
        - »Due to down sampling in GATK maximum DP is 250
  - –DP_genotype
    - Concerned if depth is too low or too high
      - –Low insufficient reads to call a variant site
    - Remove genotypes with low read depth, e.g., DP≤8
- Genotype quality (GQ) score
  - –Removal genotypes with a low genotype quality core, e.g.,
    - GQ≤ 20

Sequence Data QC – Removal of Variant Sites

- Remove variants that fail VQSR
- Removal of sites with missing data
  - –e.g., missing > 10% of genotypes
- Removal of "novel" variant sites
  - –Not observed in databases
    - e.g., gnomAD v4 (Koenig et al. 2023)
  - –Occurs in one batch and the alternative allele is observed multiple times

Sequence Data QC – Removal of Variant Sites

- Removal of sites that deviate from Hardy-Weinberg Equilibrium (HWE)
  - –Must be evaluated by population, e.g., African American, European American
  - –Related individuals should be removed from the sample before testing for deviations from HWE
  - –Best to to test for deviation of HWE after determining ancestry and removal of outliers
    - Don't base on self-report of ancestry
  - –A variety of significant cut-offs to reject the null hypothesis of HWE are are used
  - –For biobank sized data a very stringent p-values is used
    - e.g., $.5.0 \times 10^{-15}$
    - To avoid removing too many variant sites

Sequence Data QC – Removal of Samples

- High levels of missing data
  - –e.g., >20% missing data
    - After taking the intersect of capture arrays
- Samples without phenotype information

QC – Assessing Genomic Sex

- When data is collected on study subjects, they are asked about their gender/sex and not their genetic sex
  - –Differences in gender/sex and genetic sex can be due to
    - Sample swaps
    - Study subjects who are not cisgender
- Some study subjects may have neither a XX nor XY karyotype
  - –Turner syndrome X0
  - –Klinefelter syndrome XXY
  - –Or another aneuploidy

QC – Assessing Sex Chromosomes

- Study subjects labeled as females with an excess of homozygous genotypes on the X chromosome can denote
  - –That their genetic sex is male
  - –Turner Syndrome
- If Y chromosome data available
  - Sequence or genotype array data
  - –Can distinguish between XY (male) from X0 (Turner syndrome)

QC – Assessing Sex Chromosomes

- Study subjects labeled as males with an excess of heterozygous SNPs* on the X chromosome can denote
  - –That their genetic sex is female
  - –Klinefelter syndrome
- Individuals who are XY (male) will be heterozygous for markers in the pseudoautosomal regions
- Availability of Y chromosome data
  - –Can greatly aid in determining genetic sex and if an individual has Turner or Klinefelter syndrome or another sex aneuploidies

*Both genetic males and females have two alleles for each locus on the X chromosome in the datafile, although genetic males are hemizygous

QC – Assessing Sex Chromosomes

- Individuals whose labeled gender/sex does not match their genetic sex are removed from the analysis
- This observation may be due to a sample swap
  - –When samples are swapped
    - Phenotype data will be incorrect
      - –e.g., a case may be labeled as a control

Cryptic Duplicate Individuals

- Duplicate samples are sometimes included in a study as part of QC to detect inconsistencies
  - –Will not detect systematic errors
  - –Usually not included in exome and whole genome sequencing studies
  - –Intentional duplicates can easily be removed before data quality control
- Cryptic duplicates (unintentional)
  - –DNA sample aliquoted more than once
  - –Individual ascertained more than once for a study
    - e.g. The same individual undergoes the same operation more than once and is ascertained each time
- For duplicate samples
  - –Only one should be retained
    - Can select the sample with the highest quality data

Related Individuals

- Individuals who are related to each other may participate in the same study
  - –Unknown to the investigator
  - –Or be part of the study design
- For related individuals
  - –PCA is performed first with unrelated individuals and related individuals are then projected onto the PCs of unrelated individuals
  - –Special analysis should be used
    - e.g., linear mixed models (LMM) and generalized LMM (GLMM)

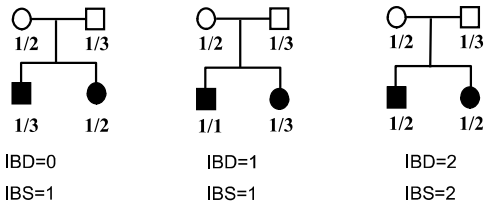
Related Individuals

- Mixed-models need to be used to analyze the data if related individuals are included
  - –Case-Control
    - Generalized linear mixed models (GLMM)
  - –Quantitative traits
    - Linear mixed models (LMM)
- If regression is performed ignoring that there are related individuals
  - –Type I error rates can be inflated

Identifying Cryptic Duplicate and Related Individuals

- Duplicate and related individuals can be detected
  - –By examining **Identity-by-State** (IBS) adjusted for allele frequencies (p-hat) between all pairs of individuals within a sample
  - –Identify-by-descent (IBD) sharing can be estimated

Identity by Descent (IBD)/Identity-by-State (IBS)



IBD Sharing Estimated Pairwise for all Individuals in a Samples

- PLINK (Purcell et al. 2007)
- Uses sequence (or genotype array) data to check IBD
  - –Prune markers to remove those in linkage disequilibrium (LD)
    - e.g., remove variants with r^2>0.1
  - –Use a MAF threshold
    - e.g., MAF>0.01
- P-hat is calculated using the "population" allele frequency
- Used to approximates IBD sharing
- IBD is the number of alleles of alleles which are shared between a pair of individuals
  - –Can either share 0, 1, and 2 alleles

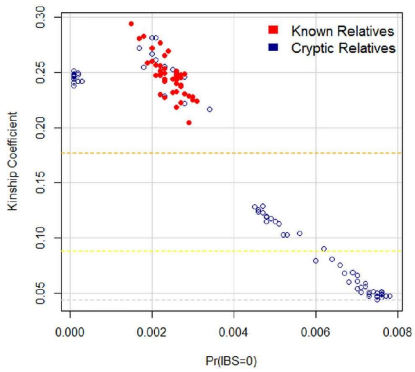
Identifying Cryptic Duplicate and Related Individuals

- Monozygote twins and duplicate samples will share 100% of their alleles IBD
  - –IBD=2 is 1.0 (can be lower due to genotyping error)
- Siblings and child-parent pairs will share 50% of their alleles IBD
  - –For parent-child IBD=1 is 1.0 (IBD=0 is 0 & IBD=2 is 0)
  - –For sibs IBD=1 is ~0.50 (IBD=0 is ~0.25 & IBD=2 is ~0.25)
    - For more distantly related individuals the IBD measure will be lower

Identifying Cryptic Duplicate and Related Individuals

- KING (Kinship-based INference for Gwas) can also be used to identify duplicate and related individuals
  - –Manichaikul et al. 2010
  - –KING is more robust to population substructure and admixture
    - Prune markers for LD (e.g., remove variants with r^2>0.1)
    - Use MAF threshold (e.g., MAF>0.01)
  - –Provides kinship coefficients
    - Duplicate samples
      - –Kinship coefficient equals 0.5
    - Siblings
      - –Kinship coefficient equals 0.25

King Graphical Output



Multiple Individuals Observed That are Distantly "Related"

- If individuals in sample come from different populations
  - –e.g., individuals from the same population within the sample will have inflated p-hat values due to incorrect allele frequencies
    - Incorrectly appear to be related to each other
- "Relatedness" amongst many individuals can also be observed when batches are combined and they have different genotyping error rates
  - –Individuals from the same batch appear to be related
- DNA contamination can cause "relatedness" between multiple individuals

UK Biobank Related Individuals > Kinship Coefficient 0.0625

White European

# of Relatives	# of relatives
1	86089
2	18491
3	3691
4	707
5	165
6	40
7	9
8	5
9	1
10	11
11	2
12	2
16	1
19	1
25	1
30	1
3985	1

African

# of relatives	# of individuals
1	715
2	153
3	26
4	10
5	3
6	5
7	5
8	4
9	1
10	4
11	2
13	3
17	2
19	3
20	2
21	1
23	1
.	.
.	.
.	.
390	1
391	1
393	1
396	1

Asian

# of relatives	# of individuals
1	743
2	115
3	33
4	4
5	4

Principal Components Analysis (PCA) / Multidimensional Scaling (MDS)

- Can be used to identify outliers
- Population substructure
  - –Individuals from different ancestry
    - e.g., African American samples included in samples of European Americans
- Batch effects
- Use a subset of markers which have been LD pruned
  - –Only very low levels of LD between marker loci
    - e.g., r^2<0.1
  - –MAF cutoff dependent on sample size
    - e.g.,MAF> 0.01
      - –Can use lower MAF for large sample sizes

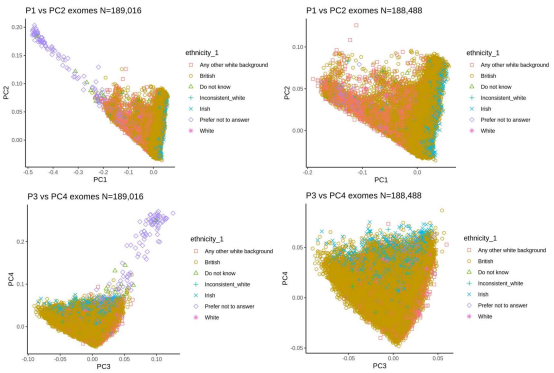
Principal Components Analysis (PCA) / Multidimensional Scaling (MDS)

- Unrelated individuals are used to generate PC plots
  - –Related individuals are projected onto to the PC plots
- Plot 1st component vs. 2nd component
  - –Additional PCs should also be plotted
    - e.g.. PCs 1-10
- Mahalanobis distance can be used to determine outliers
  - –e.g., <1
- Samples which are outliers are removed

PCA/MDS Can be Used to Identify Outliers

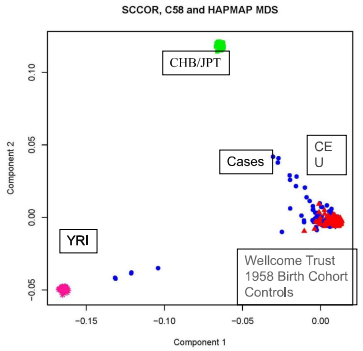
- Individuals of different ancestry
  - –e.g., African American samples included with European Americans samples
  - –Can use samples from HapMap/1000 genomes/Human Genome Diversity Project (HGDP) to aid in determining the ancestry for samples that are outliers
    - Should not include HapMap/1000 genomes samples when calculating components to control for population substructure/admixture in the analysis
- Batch effects

Principal Components Analysis Example

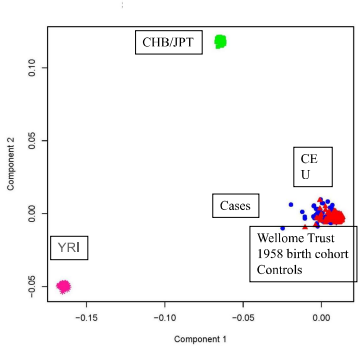


Mahalanobis distances were calculated and 0.3% of the samples with the most extreme values were removed.

Detecting Outliers Using PCA and HapMap Sample



Detecting Outliers Using PCA and HapMap Sample



Detection of Genotyping Error via Deviation from HWE

- Testing for deviations from HWE not very powerful to detect genotyping errors
- The power to detect deviations from HWE dependent on:
  - –Error rates
  - –Underlying error model
    - Random
    - Heterozygous genotypes -> homozygous genotypes
    - Homozygous genotypes ->Heterozygous genotype
  - –Minor allele frequencies (MAF)

Detection of Genotyping Error via Deviation from HWE

- Controls and Cases are evaluated separately
  - –Deviation found only in cases can be due to an association
- Test for deviation from HWE only in samples of the same ancestry
  - –Population substructure can introduce deviations from HWE
- Do not include related individuals when testing for deviations from HWE
  - –Can cause deviations from HWE

Detecting Genotyping Error – Examining HWE

- What criterion is used to remove variants due to a deviation from HWE
  - –GWAS studies have used 5.0×10^{-7} to 5.0×10^{-15}
- Quantitative Traits
  - –Caution should be used removing markers which deviate from HWE may be due to an association
    - Remove markers with extreme deviations from HWE and flag markers with less extreme deviations from HWE
- When performing imputation need to be more stringent in removing variants which deviate from HWE

Additional Useful Software - VCFtools

- Filter out specific variants
- Compare files
- Summarize variants
- Convert to different file types
  - –e.g., VCF to PLINK
- Merge files
- Create intersections and subsets of variants

Sequence Data QC Overview

- Remove variant sites that fail VQSR
- Remove genotypes with low DP, GQ scores, etc.
- Remove variant sites with large percent of missing data
- Remove samples with missing large percent of missing data
- Evaluate genetic sex of individuals based upon X and Y chromosomal data
  - –Sample mix-ups
  - –Individuals with Turner or Klinefelter Syndrome

Sequence Data QC Overview

- Evaluate samples for cryptically related individuals and duplicates
  - –Use variants which have been pruned for LD
    - e.g., r^2<0.1
  - –King or Plink algorithm
    - Always remove duplicate individuals
      - –Retaining only one in the sample
    - If sample includes related samples use linear mix models (LMM)/Generalized LMM (GLMM) to control for relatedness
      - –Best to perform even for data without related individuals
    - If only a few related individuals can retain only one individual of a relative group if not using LMM or GLMM
      - –May have to do if using a method which is not implemented in the LMM/GLMM framework

Sequence Data QC Overview

- Detection of sample outliers
  - –Perform principal components analysis (PCA) or multidimensional scaling (MDS) to detect outliers
    - Use variants pruned for LD
      - –e.g., r^2<0.1
    - Use unrelated individuals and then project related individuals onto the PCs
- Due to population substructure/admixture and batch effects
- Remove effects by
  - –Additional QC
  - –Removal of outliers (can be determined by Mahalanobis distance) and\or
  - –Inclusion of MDS or PCA components in the association analysis

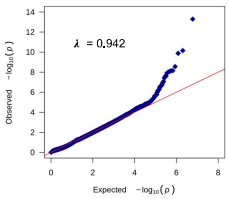
Sequence Data QC Overview

- Remove/flag variant sites that deviate from HWE in controls
  - –HWE should be only be tested in unrelated individuals from the same population
- Post Analysis - Quantile-Quantile (QQ) plots
  - –To evaluate uncontrolled batch effects and population substructure/admixture

QQ Plots - Genome Wide Association Diagnosis

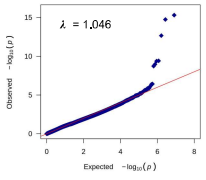
- Thousands of variants/genes are tested simultaneously
- The p-values of neutral markers follow the uniform distribution
- If there are systematic biases, e.g., population substructure, genotyping errors, there will be a deviation from the uniform distribution
- QQ plots offers an intuitive way to visually detect biases
- Observed p-values are ordered from largest to smallest and their -$\log_{10}$(p) values are plotted on the y axis and the expected -$\log_{10}$(p) values under the null (uniform distribution) on the x axis

QQ Plot of Exome Wide P-Values UK Biobank 200K



Hearing aid users

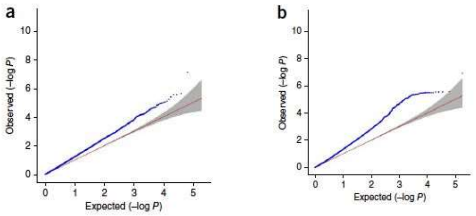
Case N= 6,436
Controls N= 96,601



Problem hearing with background noise

Cases N=65,660
Controls N= 96,601

QQ Plots show extreme inflation λ=1.32



Bulik-Sullivan et al. 2015

Genomic Inflation Factor to Evaluate Inflation of the Test Statistic

- Genomic Inflation Factor (GIF): ratio of the median of the test statistics to expected median and is usually represented as λ
  - –No inflation of the test statistic λ=1
  - –Inflation λ>1
  - –Deflation λ<1
    - Can be observed when a study is underpowered
- Problematic to examine the mean of the test statistic
  - –Can be large if many variants are associated
    - Particularly if they have very small p-values
    - Should not be used

Phenotype	Covariate	Mean Chi-Square	GIF (λ)
BP		1.23829	1.16932
BP	Age	1.24119	1.18025
BP	Age-EV1	1.09471	1
BP	Age-EV2	1.0881	1
BP	Age-EV4	1.08365	1
BP	Age-EV10	1.09592	1.00402
BPI		1.14931	1.08921
BPI	Age	1.16139	1.10113
BPI	Age-EV1	1.09079	1.01148
BPI	Age-EV2	1.0428	1
BPI	Age-EV4	1.04204	1
BPI	Age-EV10	1.06421	1.01724
BPII		1.17283	1.20664
BPII	Age	1.17583	1.20996
BPII	Age-EV1	1.09874	1.13065
BPII	Age-EV2	1.09904	1.13425
BPII	Age-EV4	1.09592	1.13609
BPII	Age-EV10	1.10046	1.1418
BPII	Sex Age-EV1	1.06956	1.06424
BPII	Sex Age-EV4	1.05817	1.05323
BPII	Sex Age-EV10	1.06338	1.05581

Evaluating Reason for Inflated λ

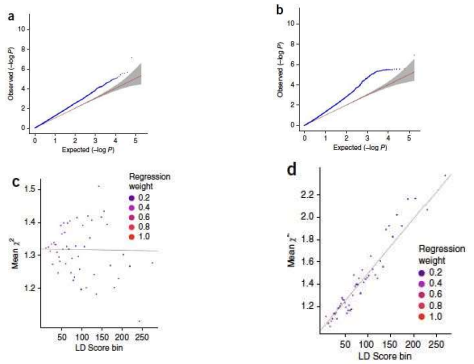
- LD score regression (LDSC) can be used to determined if the observed λ is inflated due to
  - –Problems in the data
    - Population substructure/admixture
    - Batch effects/genotyping errors
  - –Polygenicity
    - Many associated loci each with a very small effect size
- LDSC is performed and the intercept is examined
  - –If intercept is >1 than inflation is due to population substructure, etc.
  - –If intercept is ~1 than λ<1 is due to polygenicity

Performed Simulation Studies using LDSC Regression to Evaluate Polygenicity

Panels a & c data were simulated with population substructure. The λ=1.32 (a) & LDSC intercept = 1.30 (c)

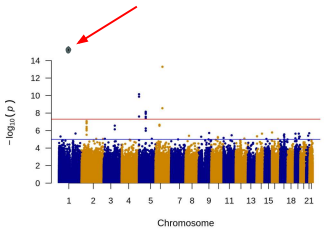
Panels b & d data were simulated with polygenicity with 0.1% of variants having a causal effect The λ=1.32 and LDSC intercept = 1.006

Panels c & d is shown the LDSC regression line



Bulik-Sullivan et al. (2015)

Post Analysis QC



Post Analysis QC

Post Analysis QC

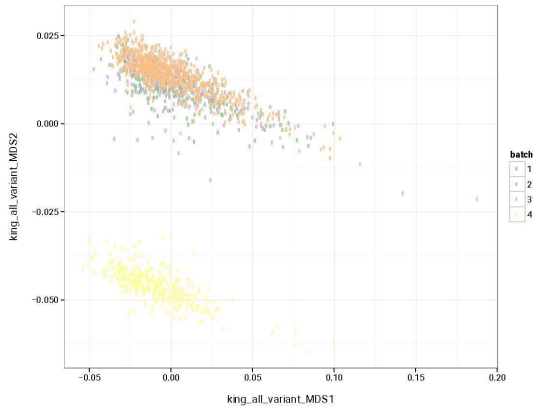
Example Project Description

- 1,667 Samples
- Seven cohorts
- Two sequencing centers
  - Center 1
    - Two capture arrays
      - NimbleGen V2Refseq 2010 (CA1): 1082
        - Batch 1 and 3
      - NimbleGen bigexome 2011 (CA2): 234
        - Batch 2
  - Center 2
    - One capture array
      - Agilent SureSelect
        - Batch 4
- Four batches
- No intentional duplicate samples

Example Project Description

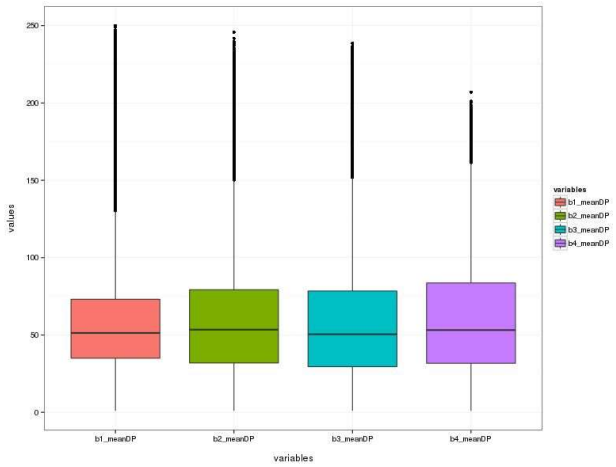
- Intersection of the three capture arrays used
  - NimbleGen V2Refseq 2010
    - Batch 1 and 3
  - NimbleGen bigexome 2011
    - Batch 2
  - Agilent Sure Select
    - Batch 4
- Sequencing machine
  - Illumina HiSeq
- Sequence alignment
  - BWA
- Multi-sample variant calling
  - GATK

MDS First 2 Components Before QC*

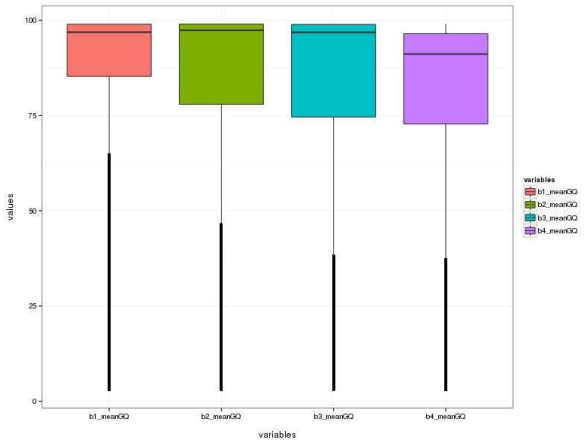


*After VQSR

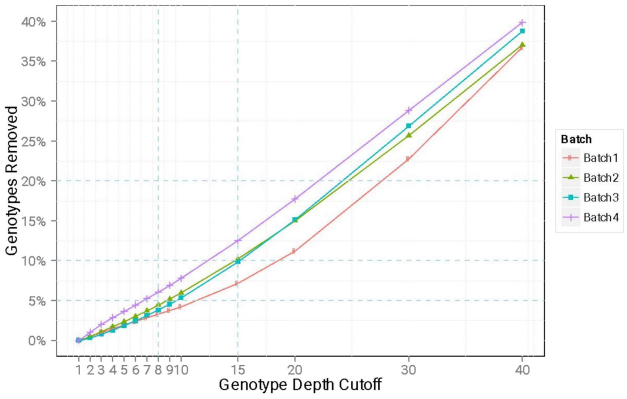
Mean GP (genotype) by Batch



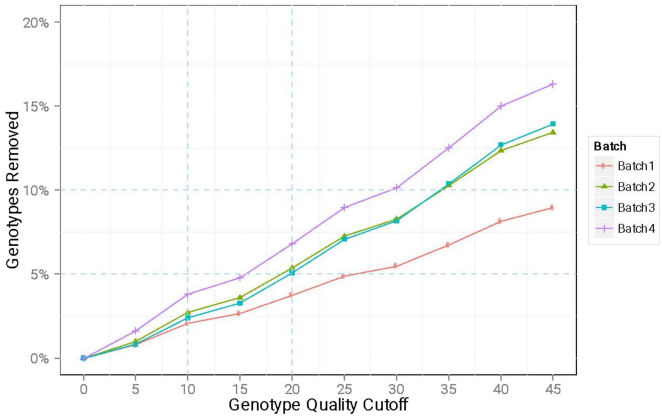
Mean GQ by Batch



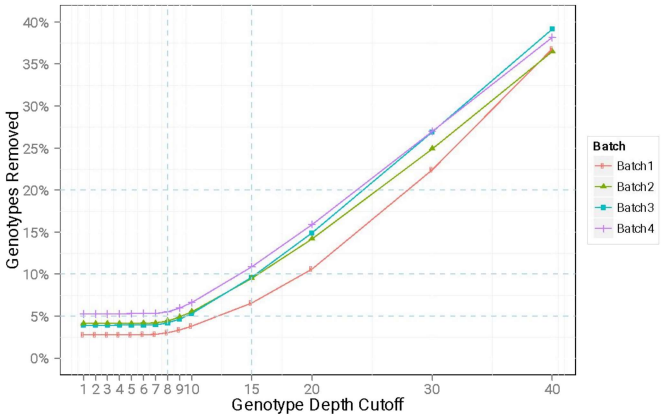
Genotypes Removed by DP (genotype) Cut-off by Batch



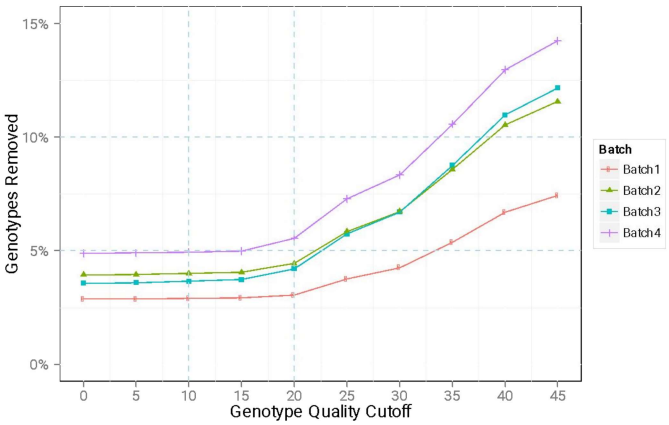
Genotypes Removed by GQ Cut-offs by Batch



Genotypes Removed by DP (genotype) Cut-off by Batch (First removing genotypes with GQ ≤ 20)



Genotypes Removed by GQ Cut-offs by Batch (First removing genotypes with a DP≤8)



Missing Rate Criteria & Sites Removed

	Variant sites removed if missing >10% of their genotypes	Variant sites removed if missing >5% of their genotypes
	Percent of genotype data removed	
Before QC*	2.5%	2.5%
After QC	12.9%	18.3%

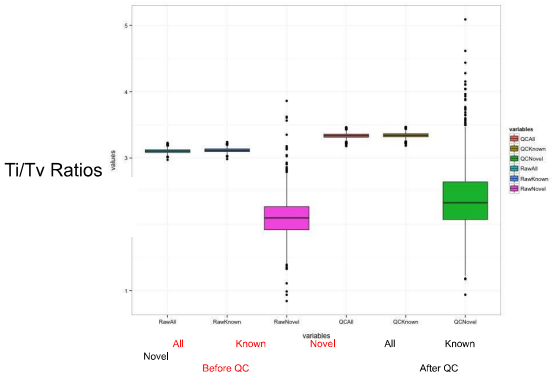
Variant sites missing >10% of their data were removed

*After VQSR

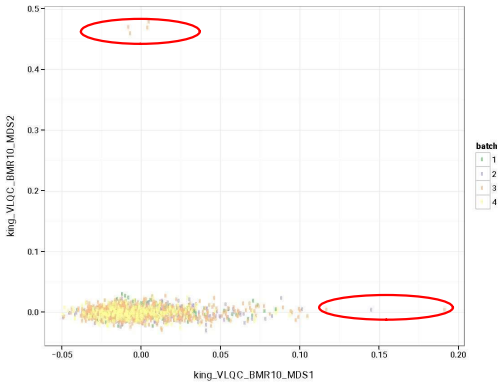
Ti/Tv Ratios during QC Process

	Known	Novel	All
Before VQSR	2.95 ± 0.05	1.18 ± 0.29	2.86 ± 0.07
Before additional QC	3.12 ± 0.03	2.01 ± 0.32	3.11 ± 0.03
Genotype QC DP≤8, GQ≤20	3.18 ± 0.04	2.10 ±0.32	3.16 ± 0.03
Remove sites missing >10% genotypes	3.39 ± 0.04	2.42 ± 0.52	3.39 ± 0.04
Remove batch specific novel sites ≥2 N=17,835	3.39 ± 0.04	2.41 ± 0.53	3.39 ± 0.04
Remove sites deviating from HWE $p \leq 5x10^{-8}$ N=4,414	3.41 ± 0.04	2.39 ± 0.54	3.40 ± 0.04

Ti/Tv Ratios by Individual Before and After QC



MDS First 2 Components After QC



Sequence Data QC

- Batch effects can sometimes be removed with additional QC
- Extreme outliers should be removed
- Additionally, MDS\PCA components can be included in the analysis to control for population substructure\admixture and batch effects
  - –Unless correlated with the outcome (phenotype)
  - –The MDS or PCA components should be recalculated after QC only including those samples included in the analysis
- Batch (dummy coding) may be included as a covariate in the analysis
  - –Unless correlated with the outcome (phenotype)

Convenience Controls

- Can reduce the cost of a study
- Genotype data
- Type I error can be increased
  - –Ascertainment from different population
  - –Differential genotyping error
    - Even if performed at the same facility
- Proper QC can reduce or remove biases

Convenience Controls–Sequence Data

- Obtain BAM files and recall cases and control together
  - –Can still have differential errors between cases and controls
  - –Check variant frequency by variant types in cases and control
    - Synonymous variants should have the same frequencies
    - Would not expect large differences in numbers of variants between cases and controls

Convenience Controls–Sequence Data

- For single variants can compare difference in frequencies with gnomAD but it is problematic
  - –Differences in frequencies can be due to differences in ancestry and/or sequencing errors
  - –Cannot adjust for confounders
    - e.g., sex, population substructure/admixture
- Don't perform an aggregate test using frequency information obtained from databases, e.g., gnomAD, TOPMed Bravo

Genotype Array Data
Genotype Data QC – Population Based Studies

- Initially remove DNA samples from individuals who are missing >10% or their genotype data
- For variant sites with a minor allele frequency (MAF)≥0.05
  - –Remove variants sites missing >5% of their genotype data
- For variant sites with a MAF<5%
  - –Remove variant sites missing > 1% of their genotype data
- Remove samples missing >3% genotype calls
- The genotypes for variant sites and samples with high rates of missing data
  - –May have higher genotype error rates

Order of Data Cleaning-Genotype Array Data

- Check genetic sex of individuals based on X-chromosome markers & Y chromosome marker data (if available)
  - –Remove individual whose reported gender/sex is inconsistent with genetic data
    - Could be due to a sample mix-up
- Check for cryptic duplicates and related individuals
  - –Used "trimmed data set of markers which are not in LD
    - e.g. r2<0.1
  - –Remove duplicate samples

Order of Data Cleaning-Genotype Array

- Perform PCA or MDS to check for outliers
  - –Use trimmed data set of markers which are not in LD
    - e.g., r2<0.1
  - –First with unrelated individuals and then project related individuals on the components
  - –Remove outliers from data
    - e.g., Mahalanobis distance
- Check for deviations from HWE
  - –Separately in cases and controls
  - –Only unrelated individuals
  - –If more than one ancestry group
    - Separately for each ancestry group
      - –As determined via PCA or MDS

Order of Data Cleaning-Genotype Array

- Examine QQ plots
  - –e.g., not controlling adequately for population admixture
    - Inflated test statistics Deflated p-values
- Examine Manhattan to detect associated variants which are not in LD with other variants
  - –Genotyping errors causing spurious associations

The Nuremberg Code (1947)

Ten Basic Principles, including:

"The voluntary consent of the human subject is absolutely essential…"

"The experiment should be conducted as to avoid all unnecessary physical and mental suffering and injury…"

"No experiment should be conducted where there is an a priori reason to believe that death or disabling injury will occur; except, perhaps, in those experiments where the experimental physicians also serve as subjects."

"During the course of the experiment, the human subject should be at liberty to bring the experiment to an end if he has reached the physical or mental state where continuation of the experiment seems to him to be impossible."

During the course of the experiment the scientist in charge must be prepared to terminate the experiment at any stage, if he has probable cause to believe…that a continuation of the experiment is likely to result in injury, disability, or death to the experimental subject.

Tuskegee Study of Untreated Syphilis in the Negro Male (1932-1972)

National Research Act (1974)

Required the creation of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

The Ethics of Conducting Research with Humans: The Belmont Report (1979)

- Beneficence
  - maximize benefits, minimize risks
- Justice
  - Who should bear the burdens of the research?
  - Who should benefit from results?
- Respect for Persons
  - Autonomy
  - Protect those with diminished autonomy

The Belmont Report was the basis for federal requirements of human research protections

Office for Human Research Protections

- 45 CFR 46 Subpart A ('Common Rule')
- Subpart B (Pregnant Women, Fetuses, and Nonviable/Questionable Viable Neonates),
- Subpart C (Prisoners),
- Subpart D (Minors)

Food & Drug Administration

(jurisdiction: clinical investigations of drugs, devices, biologics)

- 21 CFR 50: Protection of Human Subjects
- 21 CFR 56: Institutional Review Boards
- 21 CFR 312: Investigational Drugs
- 21 CFR 812: Investigational Devices

What is the Common Rule?

It is **the** Federal Policy for the Protection of Human Subjects

Originally promulgated in 1991, with no significant changes, until 1/21/19!

Rockefeller's Federal Wide Assurance (FWA) certifies compliance with this federal policy (for human research conducted or supported by Common Rule agencies…)

What's so Common about the Common Rule?

✔ 19 federal agencies follow the new Common Rule, e.g.,

- DHHS, including NIH (45 CFR 46, Subpart A)*
- DoD (32 CFR 219)
- NSF (45 CFR 690)
- Department of Energy (DoE) (10 CFR 745)
- Veterans Administration (38 CFR 16)
- Department of Education (DoEd) (34 CFR 97)

*FDA is within DHHS, but also has its own regulations

*DoJ has not signed on yet

First Question: Is your activity "human subjects research" (HSR)?

Specifically:

1. Is it HSR according to the Common Rule?
2. Is it HSR according to FDA?

(could be both!)

Start with the Common Rule

First assess:

Does the activity involve Research?

Common Rule Definition of Research:

"…**a systematic investigation**, including research development, testing and evaluation, **designed to develop or contribute to generalized knowledge**…"

(Both parts of the definition must be met)

Part I of the definition: What's a Systematic Investigation?

an activity that involves a prospective plan which incorporates data collection, either quantitative and/or qualitative, and data analysis to answer a question

Does a case study involve a systematic investigation?

Part II: What does 'designed to develop or contribute to generalizable knowledge' mean?

…designed to draw general conclusions:

✔ what we know about what is being tested is not yet firmly established or accepted;

and

✔ the activity is not dependent on the unique characteristics of the target population or system in which it will be implemented

An activity is not likely to be generalizable if the intent is:

The evaluation or improvement of a process, practice, or program at the site where the activity is being conducted

Results only to be applied to populations, or inform practice within the target population or within the site where the activity is being conducted

Implementation and evaluation of an evidence-based practice, process, or program (is it functioning as intended within the site where the activity is being conducted or with the local target population

If the activity IS research: Does the research involve human subjects, according to the Common Rule?

A living individual about whom an investigator conducting research:

(i) Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or

(ii) Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens.

Once you determine if the activity is or is not human subjects research according to the Common Rule…

You may still need to assess if the activity is human subjects research according to FDA

FDA Decisions

Does the activity evaluate an FDA-regulated test article (i.e., drug, biologic, device)?

Does the activity involve Human Subjects?

An individual who is, or becomes, a participant in research, either as a recipient of the test article or as a control. A subject may be either a healthy human or a patient. *Also included in the FDA human subject definition: The use of a biological specimen –even if de-identified-from an individual used to test an investigational device*

Does the activity involve research (clinical investigation)?

Any experiment that involves a test article and one or more human subjects...

If the activity IS human subjects research, next question: Is it exempt from the federal regulations? *

*this does not mean exempt from institutional review!

There are 6 HSR categories of research that are Exempt from IRB Review
Focus on: Exemption #4

Secondary research* for which consent is not required

***Secondary research only!** (i.e., re-using identifiable information and/or identifiable biospecimens that were, or will be, are collected for another reason, e.g., clinical or research)

Exemption #4: Secondary research uses of identifiable private information or identifiable biospecimens can be exempt under this category, if at least one of the following criteria is met:

Exemption 4(i)

The identifiable private information or identifiable biospecimens are publicly available;

Exemption 4(ii)

Identifiable private information…is **recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained** directly or through identifiers linked to the subject, the investigator does not contact the subjects, and the investigator will not re-identify subjects;

Exemption 4 (iii)

"The research involves only information collection and analysis involving the investigator's use of identifiable health information when that use is regulated under 45 CFR parts 160 AND 164, subparts A and E [HIPAA], for the purposes of "health care operations" or "research" as those terms are defined at 45 CFR 164.501 or "public health activities and purposes" as described under 45 CFR 164.512(b)"

Exemption 4 (iv)

The research is conducted by, or on behalf of, a Federal department or agency using government-generated or government-collected information obtained for non-research activities, if the research generates identifiable private information that is or will be maintained on information technology that is subject to and in compliance with section 208(b) of the E-Government Act of 2002, 44 U.S.C. 3501 note, if all of the identifiable private information collected, used, or generated as part of the activity will be maintained in systems of records subject to the Privacy Act of 1974, 5 U.S.C. 552a, and, if applicable, the information used in the research was collected subject to the Paperwork Reduction Act of 1995, 44 U.S.C. 3501 *et seq.*

What are the ethical standards that should be considered for all exempt studies?

Criteria	Yes	No	NA
The research holds out no more than minimal risk to participants	☐	☐	
Selection of participants is equitable	☐	☐	
If there is recording of identifiable information, there are adequate provisions to maintain the confidentiality of the data	☐	☐	☐
If there are interactions with participants, there are adequate provisions to protect the privacy interests of participants	☐	☐	☐
If there are interactions with participants, the consent process or information provided to potential subjects includes the following: ☐ N/A – there are no interactions and no other need for consent			
That the activity involves research	☐	☐	☐
A description of the procedures	☐	☐	☐
For Category 3 research that involves subject deception: A statement that subjects will be unaware of or misled regarding the nature or purposes of the research	☐	☐	☐
That participation is voluntary	☐	☐	☐
Name and contact information for the researcher	☐	☐	☐

If the activity IS human subjects research, but does not qualify for exemption, **it is** HSR that is not exempt, i.e., it is subject to federal regulations governing human research protection…

…including review by a federally mandated Institutional Review Board (IRB)

Two Types of Non-Exempt Review

1. Expedited Review
2. Full Board Review

For a non-exempt study to qualify for **Expedited (not full IRB Board) Review…**

…The research must be all of the following:

- no greater than minimal risk
- not involve prisoners (per OHRP guidance)
- not be classified
- not involve identifiable data that would place subjects at risk of criminal or civil liability or be damaging to the subjects financial standing, employability, insurability, reputation, or be stigmatizing. If it could, reasonable protections must be in place so that risks related to invasion of privacy and breach of confidentiality are no greater than minimal, **and**
- Fit into one or more of these categories: https://www.hhs.gov/ohrp/regulations-and-policy/guidance/categories-of-research-expedited-review-procedure-1998/index.html

If the nonexempt research doesn't qualify for expedited review, it must be reviewed at a convened IRB meeting.

Whether expedited or full board, a study must meet federally-defined criteria in order to be approved

i.e.,

"The .111 Criteria"

§ 46.111 Criteria for IRB approval of research.

(a) In order to approve research covered by this policy the IRB shall determine that all of the following requirements are satisfied:

1. Risks to subjects are minimized:

(i) By using procedures which are consistent with **sound research design** and which do not unnecessarily expose subjects to risk, and

(ii) Whenever appropriate, by using procedures already being performed on the subjects for diagnostic or treatment purposes

2. Risks to subjects are reasonable in relation to anticipated benefits, if any, to subjects, and the importance of the knowledge that may reasonably be expected to result

3. Selection of Subjects is Equitable

Consider:

- The setting in which the research will be conducted
- Who is included, who is excluded? Does it make scientific sense? Ethical sense?
- If applicable: Are children in a study involving a test article that hasn't first been tested in adults? Pregnant women before non-pregnant women?
- Costs or compensation that may impact 'fairness'
- Screening and recruitment?
- What about non-English speakers?

4. Informed consent will be sought from each prospective subject or the subject's legally authorized representative, in accordance with, and to the extent required by, §46.116

If not:

Are **ALL** the criteria for waiving informed consent or for altering/excluding specific elements of informed consent met?

5. Informed consent will be appropriately documented or appropriately waived in accordance with §46.117

If not:

Does the research meet one of the allowable criteria to waive documentation?

6. When appropriate, the research plan makes adequate provision for monitoring the data collected to ensure the safety of subjects

- **What** data will be monitored for safety purposes? **When? How?**
- **Who** will be responsible for evaluating safety data? Is a DSMB needed?
- Stopping Rules?
- Communication plan of findings to investigators and IRBs (from the IRB of Record or Sponsor)

7.When appropriate, there are adequate provisions to protect the privacy of subjects…

Consider:

- Settings where recruitment, consent, and research procedures and interactions will occur
- Provisions to ensure privacy for each of the above
- Provisions to ensure privacy when contacting or soliciting information from subjects

…and to protect the confidentiality of subject data

General:

- How will the data/biospecimens be stored?
- If identifiers will be removed or replaced, is there a possibility that such information/biospecimens could be re-identified?
- Will the data/biospecimens be shared/transmitted/ transferred to a third party or otherwise disclosed or released? How?
- Is there a potential risk of harm to individuals if the data/biospecimens are lost, stolen, compromised, or otherwise used in a way contrary to the parameters of the study?
- Plans for data retention and destruction?

A closer look at data security: minimize the risk of disclosure or breach of data

- Obtaining the data
  - What is the sensitivity of the data? Are all the data points that will be accessed or gathered for the research necessary to achieve the objectives of the research?
- Recording the data
  - What (if any) identifiers, including codes, will be recorded for the research?
- Storing the data
  - Where will paper research records, including signed consent forms, be stored? How will paper records be kept secure and restricted to authorized project personnel?
  - Where will the electronic research data be study be stored (University-provided database application like REDCap, IT file server, etc.)?
  - If there a key that links code numbers to identifiers, that list should be kept separate from the coded data, including copies of signed informed consent forms. Additionally, access to that list/key must be restricted to authorized research personnel.

Data security, continued

- Transporting or transmitting the data
  - If any research data will be collected on a mobile device, such as an electronic tablet, cell phone, or wireless activity tracker, details are needed regarding the physical security of the device, electronic security, and how the transfer of data from device to research storage location will be securely accomplished.
  - If any research data will be directly entered/sent by subjects over the internet or via email, will a University-provided database application (like REDCap) be used, or is there an encrypted tunnel to the site/application?
- Access to the data
  - How will the investigators ensure only approved research personnel have access to the stored research data? Password-protected files, role-based security, etc.?
- Sharing of the data
  - Will data be transferred or disclosed to or from the University? Is a contract or data transfer agreement necessary? What (if any) identifiers will be included? How will the data be securely transferred or disclosed (University-approved secure file transfer, etc.)?

Using Social Media in your research

Recruitment

- Seek to normalize social media recruitment to the extent possible, drawing analogies to traditional recruitment efforts
- Ensure that the proposed online recruitment strategy complies with all applicable federal and state laws, e.g.
  - Recruitment advertisements
  - Web site "Terms of Use"
  - Tell potential subjects that information shared via social media is not secure.

https://catalyst.harvard.edu/pdf/regulatory/Social_Media_Guidance.pdf

Using Social Media in your research

Recruitment

- Assure compliance between recruitment techniques and policies/terms of service of relevant websites.
  - If a proposed technique conflicts with website policies and terms of service, request a written exception from the site, OR
  - Depending on IRB policy, provide a statement explaining why the recruitment strategy warrants approval without an explicit exception, to be evaluated by the IRB with input from institutional legal counsel.

https://catalyst.harvard.edu/pdf/regulatory/Social_Media_Guidance.pdf

Using Social Media in your research

Recruitment

- Ensure that proposed social media recruitment strategies respect all relevant ethical norms, including:
  - Proposed recruitment does not involve deception or fabrication of online identities
  - Proposed recruitment does not involve members of research team 'lurking' or 'creeping' social media sites in ways members are unaware of
    - Strategy must be sensitive to the privacy of potential participants and respectful of the norms of the community being recruited
  - Recruitment will not involve advancements or contact that could embarrass or stigmatize potential subjects

https://catalyst.harvard.edu/pdf/regulatory/Social_Media_Guidance.pdf

Using Social Media in your research

Recruitment

- Enlist enrolled participants to facilitate introduction between members of their network and the research team. Ensure that consent will be obtained from current participants before they approach members of their online network for recruitment via their network or
- Ensure that a communication plan is in place for how the research team will handle online communication from enrolled participants that threatens the integrity of study

https://catalyst.harvard.edu/pdf/regulatory/Social_Media_Guidance.pdf

Using Social Media in your research

Data source

- A key issue in observational research using social media is whether the proposed project meets the criteria as human subjects research, and if so, what type of review is needed
  - Identifiable/de-identified data
  - Minimal risk/greater than minimal risk

Using Social Media in your research

Data source

- How is the data collected, transferred, etc.
  - Specify if research data will be collected as part of the recruitment process via social media. If so, describe what data will be collected. If that data is of a sensitive or confidential nature, describe how that data will be transferred to secure institutional servers and how will it be protected upon receipt.

And (111.b) When some or all of the subjects are likely to be vulnerable to coercion or undue influence, such as children, prisoners, individuals with impaired decision-making capacity, or economically or educationally disadvantaged persons, additional safeguards have been included in the study to protect the rights and welfare of these subjects.

(set aside issues with children, pregnant women/fetuses, prisoners, regulations for which are codified in the Common Rule subparts---more on that in a moment)

- What are some considerations when determining if additional safeguards are necessary and sufficient?
    - Examples:
        - For economically disadvantaged…is there payment? What is the amount? schedule?
        - For educationally disadvantaged…is the consent process particularly simplified? Should there be a witness to the consent process?

That's it for the .111 criteria… **but that's not all!**

Pregnant Women?
Subpart B of 45 CFR 46

Prisoners?
Subpart C of 45 CFR 46

Children?
Subpart D of 45 CFR 46

Department of Education (ED)?
Family Educational Rights and Privacy Act (FERPA) (34 CFR 99)
and the Protection of Pupil Rights Amendment (PPRA) (34 CFR 98)
See resources provided by ED when developing your research protocol

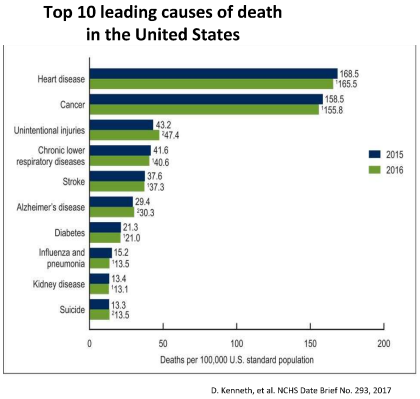
Investigational Drugs, biologics, devices?
FDA regulations at 21 CFR 50, 21 CFR 56, 21 CFR 312, 21 CFR 812

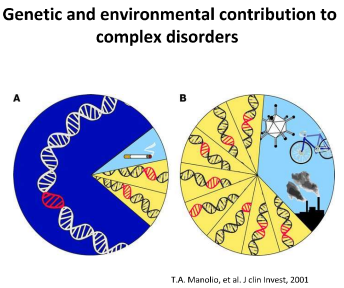
HIPAA?
45 CFR Part 160 and Subparts A and E of Part 164

Complex Trait Association Analysis of Rare Variants Obtained from Sequence Data

Suzanne M. Leal, Ph.D.
Sergievsky Family Professor of Neurological Sciences
Director of the Center for Statistical Genetics
Columbia University
sml3@Columbia.edu

Complex Diseases (Traits)





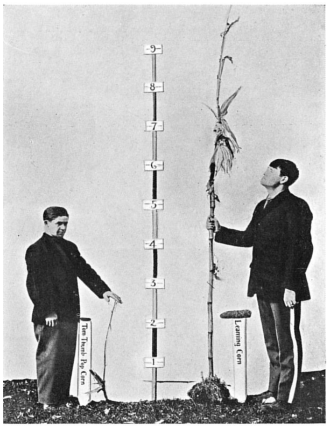
Heritability

- Broad-sense heritability
  - Considers all genetic factors
    - Phenotype = Genetics + Environmental Noise
      - Y = G + E
    - Var (Y) = Var (G) + Var (E)
  - H^2 = Var (G)/Var (Y)
- Narrow-sense heritability
  - Considers only additive contributions
    - Phenotype = Additive Genetics + Environmental Noise
      - Y = A + E
    - Var (Y) = Var (A) + Var (E)
  - h^2 = Var (A)/Var (Y)

Heritability for Common Traits

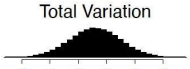
Human height heritability is ~80%

- Strongly associated common variation explain 21—29%
  - Statistically significant loci
- All common variation explains 60% of height heritability (h^2)

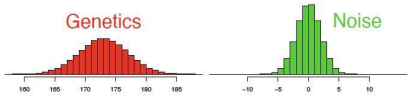


Height Heritability

- The variance of human height is about ~25 cm^2
  - Adjusted for sex
- Total Variation

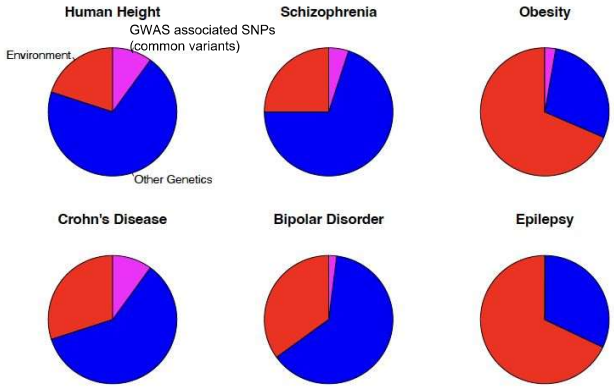


- ~20 cm^2 due to genetics
- ~5 cm^2 due to other factors (noise)



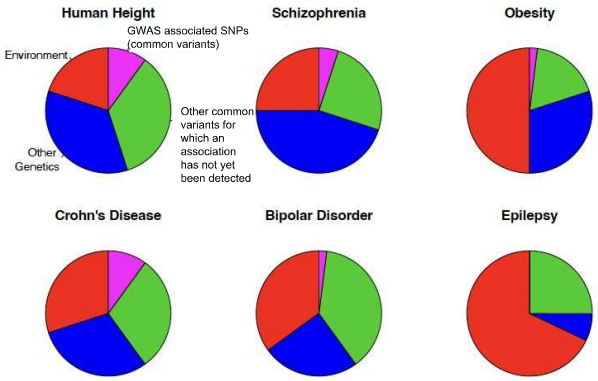
- The heritability of height is ~20/~25=~80%
- The heritability of height has been estimated using a variety of study types, e.g. twin, sibpairs

Heritability for Several Traits

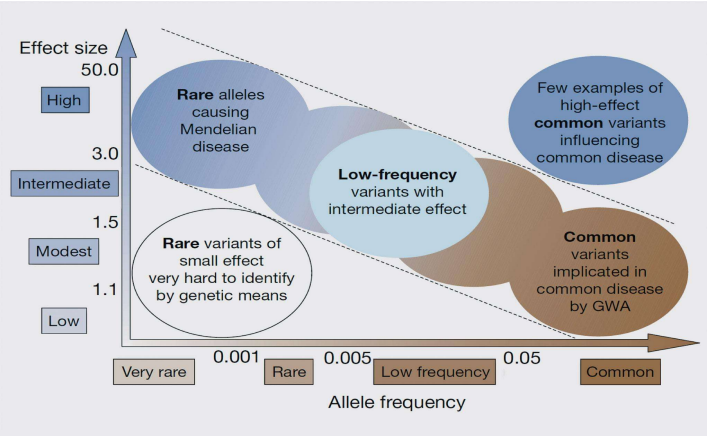


Area in blue is the "missing" heritability

Heritability for Several Traits



Allelic Architecture



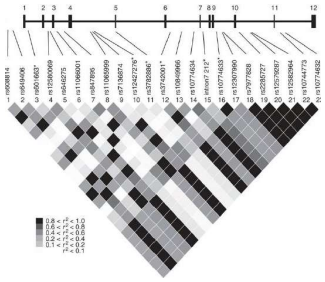
T. A. Manolio et al. Nature, 2009

Complex Disease – Common Variant Associations

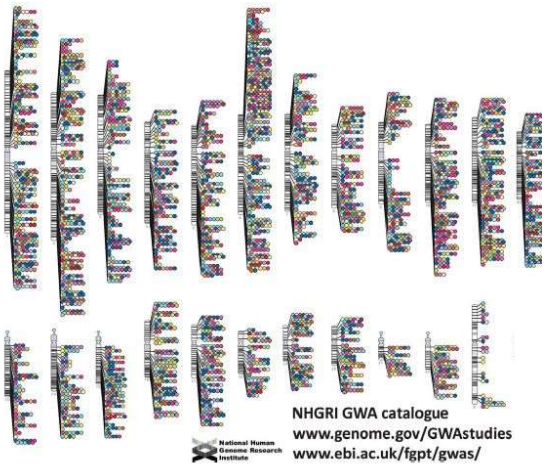
- Disease susceptibility is conferred by variants which are common within populations
  - Variants are old and widespread
- These variants have modest phenotypic effect
- This model is supported by many replicated examples
  - Age Related Macular Degeneration (Klein et al. 2005)
    - Complement factor H (CFH) gene

Studying Complex Traits – Common Variant Associations

- Hundreds of thousands of Single nucleotide polymorphism (SNPs) genotyped and analyzed
  - Indirect mapping
    - Markers usually had a minor allele frequency (MAF) > 0.05
    - Usually not pathogenic – tag SNPs
    - In linkage disequilibrium (LD) with disease susceptibility variant



Complex Trait – Common Variant Associations



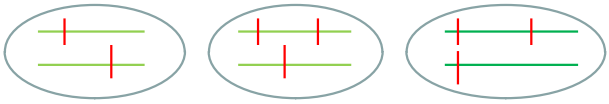
- Although highly successful in identifying thousands of complex trait loci
- Usually pathogenic susceptibility variant(s) not identified

Complex Disease – Rare Variant Associations

- Complex traits are the result of multiple rare variants
  - Although first thought that rare variants have large effects - their effect sizes are usually small
- Although these variants are rare, e.g., MAF<0.005
  - Collectively they may be quite common
    - Most variants found within a cohort are singletons
      - e.g. >50% of missense variants are singletons
- Direct tests of this hypothesis where first reported >20 years ago
  - Dallas Heart Study
    - Small sample ~1,200 individuals
      - Multi-ancestry
      - Used "extreme" sampling
    - Plasma low density lipoprotein levels (Cohen et al. 2004)
      - NPC1L1

Rationale for Rare Variant Aggregate Association Tests

- Testing individual variants with low effect sizes and MAFs
  - Underpowered to detect associations
- Testing variants in aggregate increases MAFs
  - Improving the power to detect associations



Determining MAF Cut-offs for Aggregate Rare Variant Association Tests

- MAF cut-offs are frequently used to determine which variants to analyze in aggregate rare variant association tests
- MAF from controls should not be used
  - Increases in type I error rates
- Determine variant frequency cut-offs from databases
  - e.g. gnomAD (Chen et al. 2024)
    - tp://gnomad.broadinstitute.org/)
- Using population frequencies for the population understudy
  - e.g., if studying Europeans use gnomAD frequencies for non-Finnish Europeans
    - If a variant is absent from the database, it is assumed to be rare
      - Most variants are singletons and not found in databases
- What is rare?
  - No set rule
  - Can use very low frequencies for biobank-scale studies
    - e.g., MAF <0.001, <0.005

Annotation of Variants

- Variants and regions must be annotated before performing rare variant aggregate tests
  - ANNOVAR (includes dbNSFP35a and dbscSNV1.1) (Kang et al. 2010)
  - VEP (McLaren et al. 2016)
- Determine region boundaries
  - Genes
  - Regulatory regions

Types of Annotation

- Variant frequencies
  - gnomAD
    - e.g., non-Finnish Europeans
- Variant type, e.g.,
  - Missense
  - Frameshift
  - Stop gain
  - Stop loss
  - Splice site
    - e.g., ± 2 bp from the splice site
- Functional Annotation, e.g.,
  - CADD scores (Schubach et al. 2023)
  - Eigen scores (Ionita-Laza et al. 2016)

Selecting Variants to Analyze in Aggregate

- Regions to analyze e.g.,
  - Genes
  - Transcripts
- Minor allele frequency threshold to include, e.g., ≤0.005
- Types of variants to include, e.g.,
  - Predicted loss of function variants (pLoF)
  - pLoF variants and missense and splice site variants with CADD score >20

Caveats - Aggregate Rare Variant Association Tests

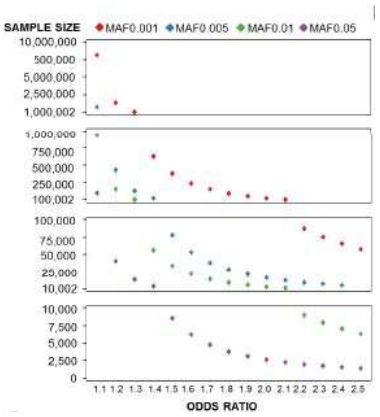
- Misclassification of variants can reduce power
  - Inclusion of non-causal variants
  - Exclusion of causal variants
- Analysis can be performed using region boundaries for
  - Genes
  - Genes within pathways
  - Regulatory regions
    - As determined for example by
      - FANTOM5 CAGE profiles to identify promoter regions (Noguchi et al. 2017)
      - STAAR pipeline that combines multiple *in silico* annotations (Li et al. 2020)
- Unlikely a sliding window approach will work
  - Size of window unknown and will differ across the genome

Analysis of Rare Variants

- For biobank sized datasets higher frequency rare variants, e.g., 0.5% can be analyzed individually
  - Using same same methods implemented for common variants

Example
$\alpha = 5 \times 10^{-8*}$
Disease prevalence 5%
1-β =0.80

*Note: a more stringent significance criterion may be necessary for genome-wide sequence data. Due to a larger number of effective tests compared to analysis of common variant GWAS panels



A Few Rare Variant Association Tests

- Combined Multivariate Collapsing (CMC)
  - Li and Leal AJHG 2008
- Burden of Rare Variants (BRV)
  - Auer, Wang, Leal Genet Epidemiol 2013
- Weighted Sum Statistic (WSS)
  - Madsen and Browning PloS Genet 2009
- Kernel based adaptive cluster (KBAC)
  - Liu and Leal PloS Genet 2010
- Variable Threshold (VT)
  - Price et al. AJHG 2010
- Sequence Kernel Association Test (SKAT)
  - Wu et al. AJHG 2011
- SKAT-0
  - Lee et al. AJHG 2012

Fixed Effect Tests
Random Effect Test
Optimal test

Types of Aggregate Analyses

- Frequency cut offs used to determine which variants to include in the analysis
  - Rare Variants (e.g., MAF<0.05% frequency)
  - Rare and low (MAF=0.05-5%) frequency variants
- Maximization approaches
- Tests developed to detection associations when variants effects are bidirectional
  - e.g., protective and detrimental
- Incorporate weights based upon annotation
  - Frequency
    - e.g., gnomAD
  - Functionality
    - CADD c-scores

Methods to Detect Rare Variant Associations Using Variant Frequency Cut-offs

- Combined multivariate & collapsing (CMC)
  - Li & Leal, AJHG 2008
- Collapsing scheme which can be used in the regression framework
  - Can use various criteria to determine which variants to collapse into subgroups
    - Variant frequency
    - Predicted functionality

CMC

- Define covariate Xj for individual j as

$$X_j = \begin{cases} 1 & \text{if rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

- Compute Fisher exact test for 2x2 table

Number of cases for which one or more rare variants are observed e.g., nonsynonymous variants freq. e.g.,<1%

Number of cases without a rare variants

	X=1	X=0
cases		
controls		

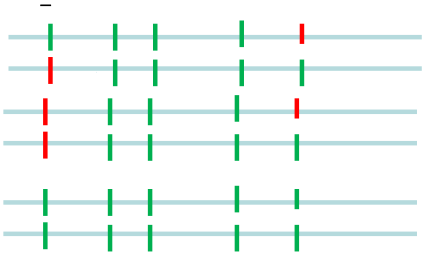
Number of controls for which one or more rare variants are observed

Number of controls without a rare variants

Can also use same coding in a regression framework

CMC

- Example of coding used in regression framework:
  - Binary coding $X_j = \begin{cases} 1 & \text{if rare variants present} \\ 0 & \text{otherwise} \end{cases}$
  - Gene region with 5 variant sites



Individual	Coding
1	1
2	1
3	0

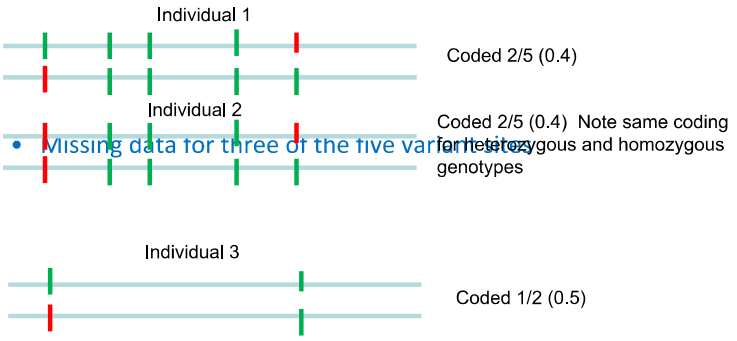
Rare Variant Sites
Green bars: Major allele is observed in the study subject
Red bars: Minor allele has been observed

Methods to Detect Rare Variant Associations Using Variant Frequency Cut-offs

- Gene-or Region-based Analysis of Variants of Intermediate and Low frequency (GRANVIL)
  - Aggregate number of rare variants used as regressors in a linear regression model
  - Can be extended to case-control studies
    - Morris & Zeggini 2010 Genet. Epidemiol
  - Test also referred to as MZ

GRANVIL

- Example of coding used in regression framework
  - Gene region with 5 variant sites – data available on all sites
  -
- Missing data for three of the five variant sites



Burden Rare Variant (BRV) extension (Auer et al. 2013 Genet Epidemiol)

Individual 1: Coded 2
Individual 2: Coded 3
Individual 3: Coded 1

Methods to Detect Rare Variant Associations Weighted Approaches

- Group-wise association test for rare variants using the Weighted Sum Statistic (WSS)
  - Variants are weighted inversely by their frequency in controls (rare variants are up-weighted)
    - Madsen & Browning, PLoS Genet 2009

Methods to Detect Rare Variant Associations Maximization Approaches

- Variable Threshold (VT) method
  - Uses variable allele frequency thresholds and maximizes the test statistic
  - Can also incorporate weighting based on functional information
    - Price et al. AJHG 2010
- RareCover
  - Maximizes the test statistic over all variants with a region using a greedy heuristic algorithm
    - Bhatia et al. 2010 PLoS Computational Biology

Methods to Detect Associations with Protective & Detrimental Variants within a Region

- C-alpha
  - Detects variants counts in cases and controls that deviate from the expected binomial distribution
    - For qualitative traits only
      - Neale et al. 2011 PLoS Genet
- Sequence Kernel Association Test (SKAT)
  - Variance components score test performed in a regression framework
    - Can also incorporate weighting
  - Wu et al. 2011 AJHG

Optimal Test

- SKAT-O
  - Maximizes power by adaptively using the data to combine a burden test and the sequence kernel association tests
    - Lee et al. 2012 AJHG

Significance Level for Rare Variant Association Tests

- For exome data where individual genes are analyzed usually a Bonferroni correction for the number of genes tested is used
  - There is very little to no LD between genes
- Bonferroni correction used
  - e.g., $p \leq 2.5 \times 10^{-6}$ (Correction for testing 20,000 genes)
- If non-coding regions are also analyzed
  - A more stringent criterion must be applied
    - Adjust for the total number of regions being tested

Problem of Missing Genotypes for Aggregate Rare Variant Association Tests

- Same frequency of missing variant calls in cases and controls
  - Decrease in power
- More variant calls missing for either cases or controls
  - Increase in Type I error
  - Decrease in power
- Remove variant sites which are missing genotypes, e.g., >10%
- Can impute missing genotypes using observed allele frequencies
  - For the entire sample
    - Not based on case or control status
- Analyze imputed data using dosages

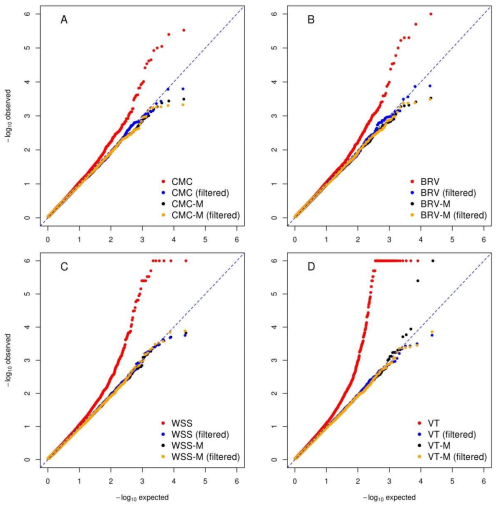
Dosages

- Genotypes are no longer assigned 0 (1/1), 1 (1/2) or 2 (2/2)
  - Due to uncertainty
- Each genotype is assigned a probability
  - Probabilities sum to 1
- For example
  - Probability of 0 (1/1) genotype is 0.98 and 1 (1/2) genotype is 0.015
- The dosage can be estimated for this example as follows

0 x 0.98 = 0
1 x 0.015 = 0.015
2 x 0.005 = 0.01
Dosage = 0.025

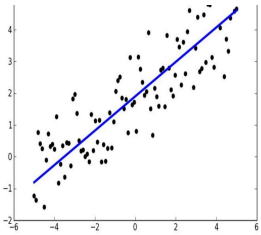
- Instead of using the most likely genotype the dosage is used

Results



Rare Variant Aggregate Methods

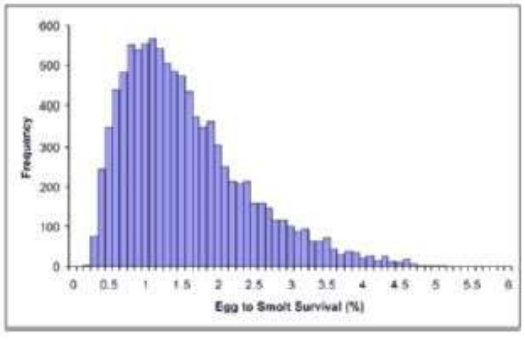
- Ideally should be performed in a regression framework to adjust for covariates
  - Logistic
  - Linear regression



- Almost all rare variant aggregate methods have been extended to be implemented within a regression framework
- Some have also been implemented in a linear mixed model (LMM)/generalized LMM (GLMM) framework

Analyzing Quantitative Variants

- Most rare variant aggregate analysis methods can be performed on quantitative traits
- If phenotype data includes outliers or deviates from normality
  - Can increase type I errors



Analyzing Quantitative Variants

- For data that deviates from normality
  - Quantile-quantile normalization
- For data that includes outliers
  - Winsorize
- Don't winsorize and then normalize
- Instead of analyzing quantitative trait values
  - Residual can be generated
    - Adjusting for confounders

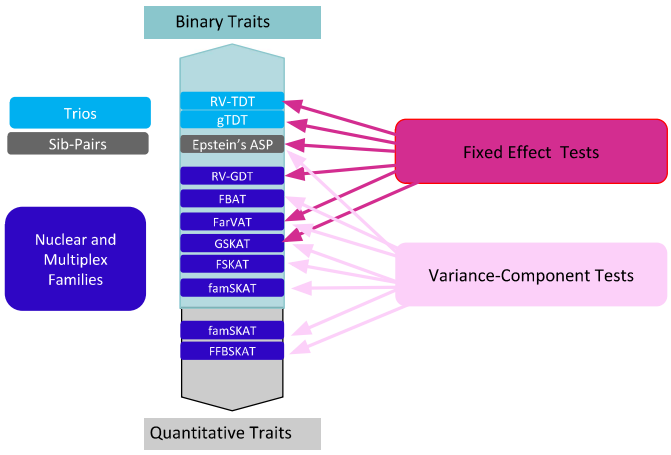
Extreme Quantitative Traits: Analyzing Individuals with the Highest and lowest trait values

- Selection of individuals with the highest and lowest trait values to generate genetic data and analyze
  - e.g., upper and lower 10%
- Analyze quantitative trait values
  - Linear regression
  - LMM
- Dichotomize data
  - Logistic regression
  - GLMM
    - Don't have problems of non-normality
    - Can lead to a loss of power

Extreme Quantitative Traits: Analyzing Individuals with the Highest and lowest trait values

- Can be more powerful than analyzing a randomly ascertained sample of the same size
  - Saving money in generating genetic data
    - e.g., whole-genome sequence data
- Power decreases with decreasing cohort size
  - Smaller sample size individuals in the upper and lower tails with not have very extreme quantitative traits values
  - For a set threshold sample size will decrease with decreasing cohort size
- Limits analysis to the phenotype which was selected for analysis
  - Most datasets are rich for quantitative traits
    - Secondary traits can be analyzed but their can be biases
      - Unless proper analysis is performed taking into account secondary trait status
- If possible, generate genetic data for a larger random sample
- Extreme sample design is currently not often used

Family-based Methods for Rare Variant Aggregate Association Analysis



Linear Mixed Model (LMM) & generalized LMM (GLMM) Analysis of Related & Unrelated Individuals

- LMM is an extension of the linear model to allow for both fixed & random effects and also allows for non-independence of samples
  - Early implementations calculated the kinship matrix Φ on the basis of known relationships, e.g., siblings
  - Amin et al. (2007) proposed to estimate kinships based on genome-wide variant data
    - The generalized relationship matrix (GRM) can be estimated for all individuals using for example identical-by-descent (IBD) sharing
- Extended to binary (case-control) traits - GLMM

LMM and GLMM: Analysis of Related & Unrelated Individuals

- Can be applied to analyze families, cryptically related, & unrelated individuals
  - e.g., UK Biobank
    - 500K study subjects of which 30.3% are ≤ 3rd degree relatives & 4.5% sib-pairs
- More recent implementation for large scale data using a variety of methods
  - BOLT-LMM (Loh et al. 2015)
  - FastGWA (Jiang et al. 2019)
  - SAIGE (Zhao et al. 2015)*
  - REGENIE (Mbatchou et al. 2020)
  - SMMAT (Chen et al. 2019)**
- *Can be used analyze data where case to control ratio is very unbalanced
  - e.g., 20 cases for every control
- **Cannot be used for UK Biobank Scale data

LMM and GLMM: Analysis of Related & Unrelated Individuals

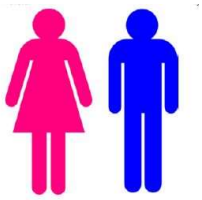
- REGENIE does not use the GRM to allow for biobank sized datasets
  - Uses ridge regression
    - Use 300,000-500,000 SNV
      - To few variants can cause the test statistics to be inflated
      - Too many variants will cause step 1 of Regenie to run extremely slowly
- This large-scale approximation may not control type I error for individuals that are closely related
  - e.g., when only families are being analyzed
    - Can use for example SMMAT
      - Which uses the GRM

LMM and GLMM: Analysis of Related & Unrelated Individuals

- A few programs which can perform rare variant aggregate analysis
  - REGENIE - Burden test, SKAT, & SKAT-O, etc
  - SMMAT - Burden, SKAT, & SKAT-O
  - rvtests (Zhan 2020) implements BOLT-LMM to perform burden association analysis

Inclusion of Covariates

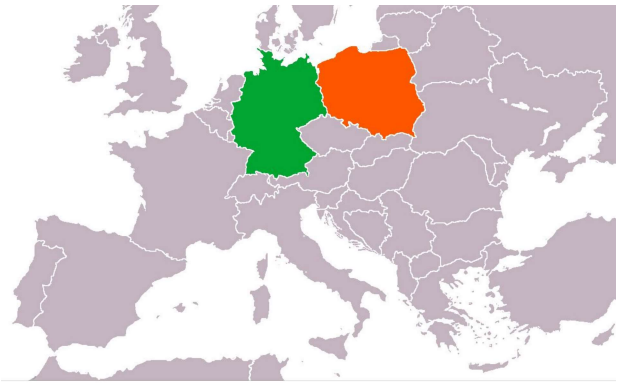
- Covariate can be included in the regression/LMM/GLMM model to control for potential confounders
  - Age
  - Sex
  - Body Mass Index (BMI)
  - Smoking status
  - Population substructure
- Covariates selection
  - Performed in a regression framework
    - Forward
    - Backward
    - Aike method



Missing Covariate Data

- If a sample is missing covariate information
  - It will be dropped from the analysis
    - Can greatly reduce power
- Solutions for missing covariates
  - Average value
    - Obtained from individuals with covariate data
  - Multiple imputation
    - The distribution of observed values for the covariate of interest are used to replace the missing value
      - Biases can occur when then there are large amounts of missing data
        - » >5-10% binary
        - » > 20-30% quantitative
      - Missing values replaced and performing
        - » Test performed multiple times, e.g., N=10
          - » For each test, the value is replaced based on the distribution
      - Can be very computationally intensive

Confounder -Population Substructure and Admixture



Population Substructure and Admixture

- If proportion of cases and controls sampled from each population is different
  - Can occur due to
    - Disease frequency is different between populations
    - Sloppy sampling
- Population substructure\admixture can cause detection of differences in variant frequencies within a gene which is due to sampling and not disease status
  - False positive findings can be increased

Example River People

Controlling for population substructure/admixture -

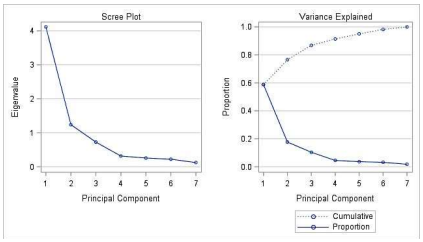
- Principal components (PCs) are also included in the model (regression /LMM/GLMM)
  - Obtained for the genotype array or sequence data
    - Markers which are not in LD are selected
      - Data is "pruned"
        - » e.g. r^2<0.1
    - Markers are selected that have a MAF above a specified cut-off
      - Dependent on sample size
        - » A lower MAF can be used for larger sample sizes, e.g.,
          - » MAF>0.001
        - » Using a lower MAF threshold will unable better control using fewer PCs

Controlling for population substructure/admixture -

- It is necessary to include PCs in the model to control for substructure/admixture when analysis is performed in an LMM/GLMM framework
  - LMM/GLMM is controlling for relatedness including cryptic relationships not population substructure/admixture
- Should not use the same PCs that were generated for quality control
  - Recalculate them using only the data which will be analyzed

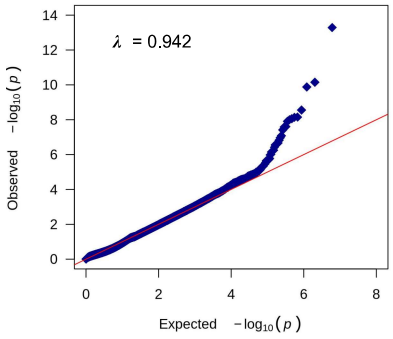
Selection of PC to Include in the Model

- Often a fixed number of PCs are selected to be included in the model, e.g., 5, 10, 20
  - Not advised
- The number of PCs to include can be based on
  - λ values
  - Percent of variance explained
  - Scree plots
    - Line graphs that shows the eigenvalues of the PCs. Plotting the eigenvalues against the number of PCs, with the eigenvalues representing how much each PC contributes to the total variance.



Controlling for population substructure/admixture

- Success of PCA\MDS in controlling for population substructure\admixture can be evaluated through lambda values and examining Quantile-Quantile (QQ) plots



Rare Variant Aggregate Association Analysis

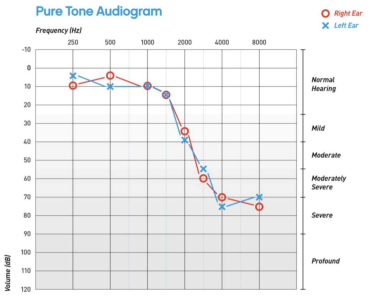
- When analyzing different populations, e.g.,
  - Africans
  - Europeans
- When analyzing data from different source, e.g.,
  - Data generated at different sequencing centers
- Analyze each group separately
- Meta-analysis can be used to combine the results from each group
  - First test for homogeneity before combining results via meta-analysis

Part II
Example of a Rare Variant Association Study

Analysis of UK Biobank Exome Data to Study the Etiology of Late-onset Hearing Loss

Age-related Hearing Loss (ARHL) (aka Presbycusis)

- ARHL can impact quality of life and daily functioning
- ARHL is one of the most common adult conditions
  - In the USA
    - ARHL affects 50% of individuals >75 years of age
    - It is estimated that 30-40 million will be affected with significant ARHL by 2030



Goals of the Study

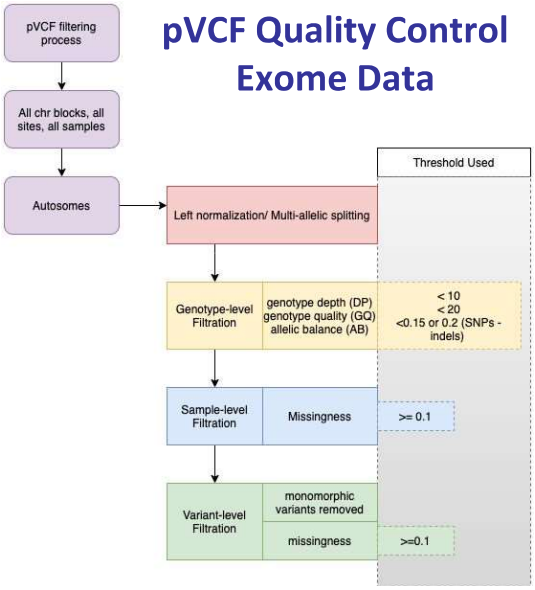
- Using data from the UK Biobank to detect associations between self-reported measures of ARHL and genetic variants
  - **H-aid** self-reported hearing aid use (f.3393: "Do you use a hearing aid most of the time?")
  - **H-diff** self-reported hearing difficulty (f.2247: "Do you have any difficulty with your hearing?")
  - **H-noise** self-reported hearing difficulty with background noise (f.2257: "Do you find it difficult to follow a conversation if there is background noise e.g., TV, radio, children playing)?
  - **H-both** individuals with both H-diff and H-noise
- With an emphasis of understanding the role that rare variation plays in ARHL

UK Biobank

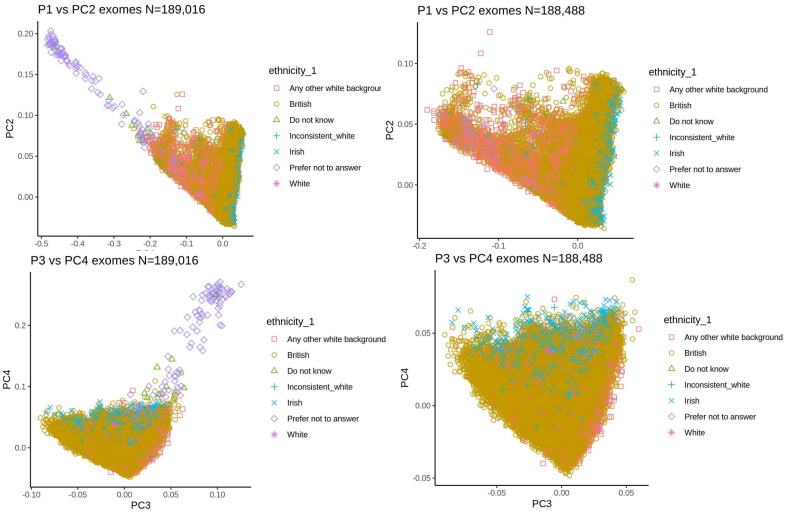
- 500,000 individuals randomly sampled
  - Aged 40-69 at time of enrollment
    - To be followed for at least 20 years
    - Predominantly white Europeans
      - Also includes South Asians and individuals of African Ancestry and smaller number of individuals of a few other ancestries
- Extensive phenotype data
  - Qualitative and quantitative traits
    - ICD-10 and ICD-9 codes
    - Self reports
    - Cognitive test
    - Brain MRIs
    - NMR-metabolomics data
- Genetic Data
  - Genotype and imputed data
  - Exome sequence data
  - Whole genome sequence data
  - Telomere length data

*Data showcase can be used to examine phenotypes and sample sizes available

pVCF Quality Control Exome Data



Principal Components Analysis and Exclusion of Outliers



Exclusion Criteria
Obtained from ICD10, ICD9, & Self Report

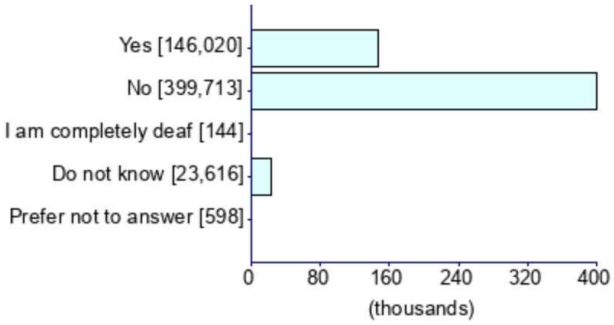
- Deafness
- Early-onset hearing impairment
- Otosclerosis
- Meniere's
- Labyrinthitis
- Disorders of acoustic nerve
- Bell's palsy
- History of chronic suppurative and nonsuppurative otitis media
- Meningitis
- Encephalitis, myelitis, and encephalomyelitis
- Etc.

Defining Cases and Controls

- Based on answers obtained from a touch screen
- Cases - self-reported hearing difficulty
  - f.2247: "Do you have any difficulty with your hearing?"
- Controls - did **not** have any self-reported HL or ID10/9 HL codes

Hearing difficulty/problems -Data field 2247

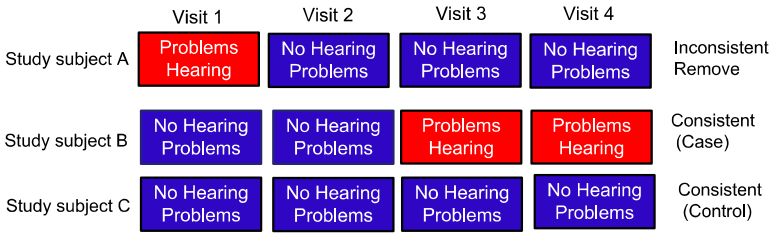
569,977* items of data are available, covering **498,704** participants



*Due to repeat visits

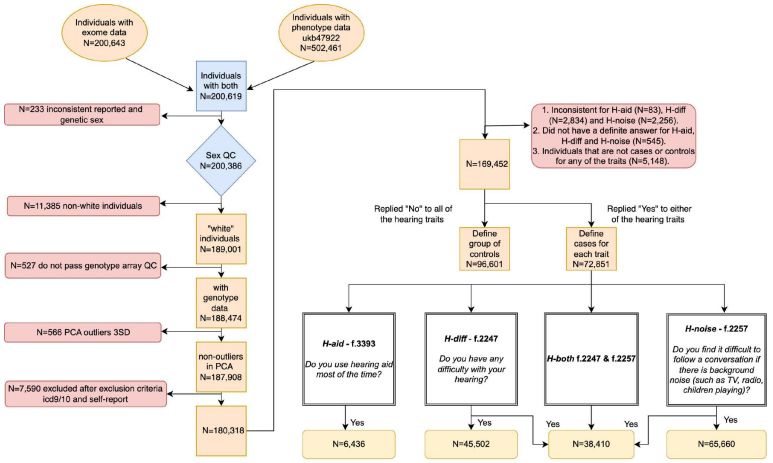
Repeat measures*

- Individuals with inconsistent answers removed

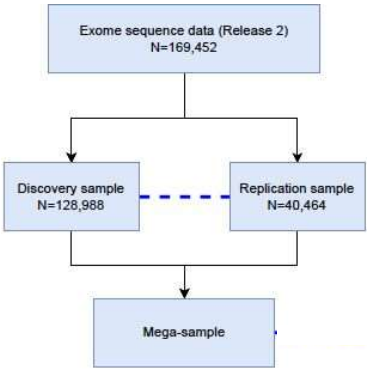


*Majority of study subjects currently have data from only one visit

Analysis of Exome Sequence Data for ARHL Releases 1 and 2 (N=200,000 exomes)



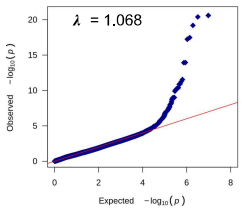
UK Biobank Discovery and Replication Samples of White Europeans with exome and ARHL Phenotype Data



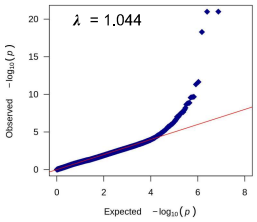
Analysis of Exome Data

- Analysis performed using generalized linear mixed models (GLMM) (REGENIE)
  - To control for inclusion of related individuals
    - For the UK Biobank data 30.3% of participants are ≤ 3rd degree relatives & 4.5% sib-pairs
  - Genotype array data (~800K) were used for the ridge regression
    - Data pruned to remove variants with a r^2>0.1 - to reduce the number of variant sites
      - Using exome data for the ridge regression led to an an inflated lambda value

QQ Plot using exome data for ridge regression



QQ Plot using genotype data for ridge regression



Analysis of Exome Data

- Analysis limited to individuals of White European Ancestry
- Sex, age, and two PCAs included as covariates
  - Age
    - cases first report of hearing difficulty
    - Controls age at last visit
  - PCAs recalculated for only individuals included in the analysis
    - Using LD pruned genotypes array data (r2<0.1)

Analysis of Exome data – Single Variant

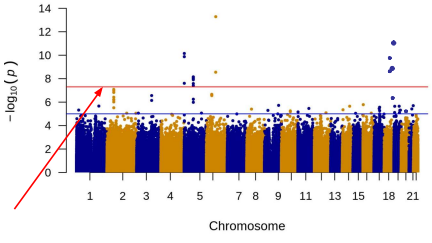
- Variants with four or more alternative alleles observed in the sample analyzed
  - A very low MAF was used since it was hypothesized some of the variants may have a large effect sizes

Significance Levels-Single Variant Analysis

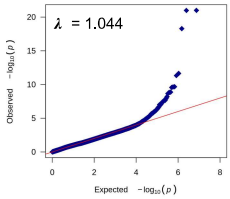
- Discovery sample
  - A genome-wide significance level (single variant analysis) was used to reject the null hypothesis of no association
    - $p \leq 5.0 \times 10^{-8}$
- Replication sample
  - Permutation was used to obtain empirical p-values
    - Adjusting for the phenotypes and variants brought to replication
      - $p \leq 0.05$

Hearing Difficulty - Data Field 2247 Single Variant Analysis

Manhattan Plot



QQ Plot



Genome-wide significance level 5 x 10^{-8} (red line)

Cases N=45,502
Controls N= 96,601

Analysis of the Discovery Sample & Replication Single Variant Analysis

Discovery sample single-variant associations analysis for age-related hearing loss traits

CHR	SNP	EA	EAF	Gene	H-aid			H-diff			H-noise			H-both		
					Beta(OR)	SE	P	Beta(OR)	SE	P	Beta(OR)	SE	P	Beta(OR)	SE	P
5	rs537688122	G	$6.65x10^{-4}$	*PDCD6*	1.99(7.3)	0.29	$2.25x10^{-10}$	1.32(3.7)	0.17	$1.12x10^{-15}$	1.04(2.8)	0.16	$5.50x10^{-11}$	1.27(3.6)	0.18	$1.02x10^{-12}$
5	rs549592074	C	$5.58x10^{-4}$	*PDCD6*	1.99(7.3)	0.32	$1.95x10^{-8}$	1.35(3.9)	0.18	$7.05x10^{-14}$	1.07(2.9)	0.18	$6.69x10^{-10}$	1.28(3.6)	0.19	$5.52x10^{-11}$
5	rs571370281	G	$7.04x10^{-4}$	*PDCD6*	1.92(6.8)	0.28	$6.02x10^{-10}$	1.33(3.8)	0.16	$1.14x10^{-14}$	1.03(2.8)	0.16	$2.26x10^{-11}$	1.29(3.6)	0.17	$9.66x10^{-14}$
6	rs1574430	C	$6.09x10^{-1}$	*SLC22A7*										0.06(1.1)	0.01	$2.77x10^{-8}$
6	rs2242416	G	$6.09x10^{-1}$	*CRIP3*										0.06(1.1)	0.01	$2.60x10^{-8}$
6	rs121912560	G	$7.63x10^{-5}$	*MYO6*	5.48(239.8)	1.12	$1.79x10^{-10}$	3.54 (34.5)	0.90	$3.41x10^{-9}$				3.73(41.7)	0.90	$3.76x10^{-10}$

Genome wide-significant variants ($p < 5x10^{-8}$) with hearing aid (H-aid), hearing difficulty (H-diff), hearing difficulty with background noise (H-noise) and the combined hearing trait (H-both) in the analysis of the discovery sample of White-European individuals from the UK Biobank. The p-values for replicated associations (empirical p-values <0.05 adjusting for variants and traits brought to replication) are shown in red; CHR - chromosome; EA - effect allele, EAF - effect allele frequency, OR - odds ratio, SE - standard error, P - p-value

Rare Variant Aggregate Analysis

- Genes with at least two variants were analyzed, e.g., pLoF variants
- Max coding was used
- Two masks were used
  - Mask 1 – pLoF variants
  - Mask 2 – pLoF and missense variants
- Minor allele frequency cut-off of <0.01 was used
  - The frequencies for each variant site were obtained from gnomAD (non-Finnish Europeans)

REGENIE Rare Variant Aggregate Analysis

- Three different codes can be used
  - Max
  - Sum
  - Comphet
    - This term is not correct because the phase is unknown
      - Variants may be on the same haplotype

Note: At the time of the study REGENIE could only perform fixed effect tests it now implements SKAT and SKAT-O

Single variant sites		max	sum	comphet
00000000000000	→	0	0	0
00000100010000	→	1	2	2
00201011010100	→	2	7	2

https://rgcgithub.github.io/regenie/options/

Selection of Variants to Include in Rare Variant Aggregate Association Tests

Annotation File		Mask File		AAF file
1:55039839:T:C PCSK9 LoF 1:55039842:G:A PCSK9 missense	+	Mask1 LoF Mask2 LoF,missense	+	1:55039839:T:C 1.53e-05 1:55039842:G:A 2.19e-06
1:55039839:T:C PCSK9 CADD30 1:55039842:G:A PCSK9 CADD20	+	Mask1 CADD score > 30 Mask2 CADD score > 20	+	1:55039839:T:C 1.53e-05 1:55039842:G:A 2.19e-06

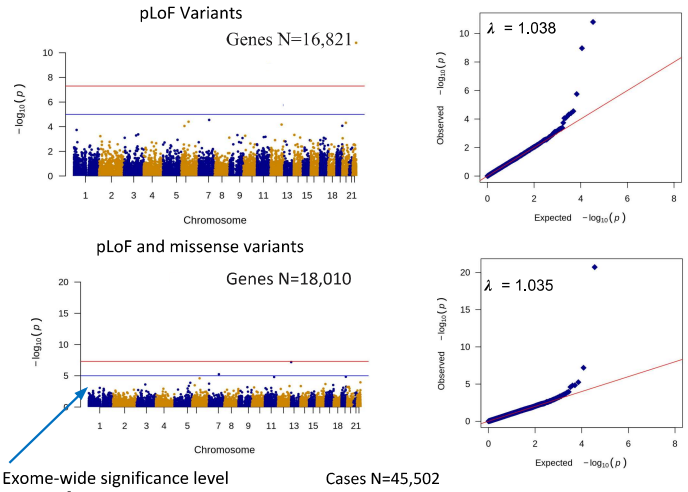
REGENIE will use information from the annotation and alternative allele frequency (AAF) files to build the Masks (variants to be included in the association testing)

Rare Variant Aggregate Analysis

- Exome sample was split
  - Second release of 150K exome were used as the discovery sample.
  - First release of 50K exome were used as the replication sample
- Entire exome sample (200K) was also analyzed*
- Discovery sample significance level
  - $p \leq 2.5 \times 10^{-6}$
    - 0.05/20.000 Bonferroni correction for testing 20,000 genes
- Replication sample significant level
  - $p \leq 0.05$
  - Empirical p-values generated
    - Permutation used to adjust for the number of phenotypes and genes brought to replication (pLoF and pLOF & missense)

*No replication sample available for these findings

Hearing Difficulty - Data Field 2247



Exome-wide significance level
2.5×10^{-6} (blue line)

Cases N=45,502
Controls N= 96,601

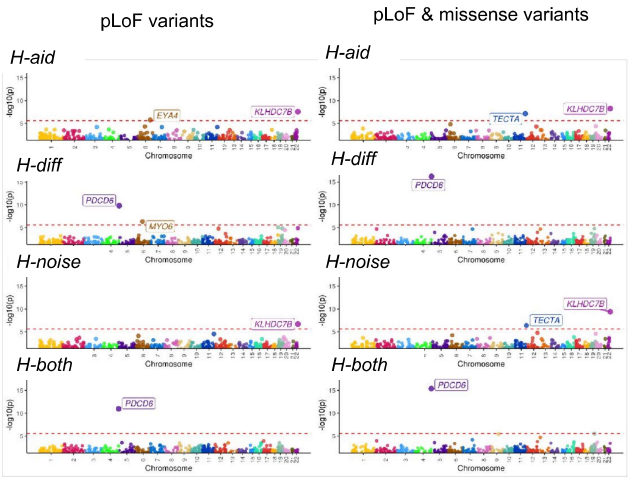
Rare Variant Aggregate Analysis – Discovery and Replication Samples

Discovery Sample Rare-variant aggregate association analysis with age-related hearing traits

Type of variation	Gene	H-aid Beta(OR)	H-aid SE	H-aid P	H-diff Beta(OR)	H-diff SE	H-diff P	H-noise Beta(OR)	H-noise SE	H-noise P	H-both Beta(OR)	H-both SE	H-both P
pLoF	*KLHDC7B*	1.29(3.6)	0.21	2.65×10^{-8}	0.69(1.9)	0.12	5.59×10^{-9}	0.56(1.8)	0.11	2.01×10^{-7}	0.77(2.2)	0.12	3.99×10^{-10}
	TECTA				0.84(2.3)	0.16	7.08×10^{-8}				0.84(2.3)	0.16	4.18×10^{-7}
	EYA4	3.30(27.1)	0.61	1.74×10^{-6}									
pLoF + missense	*PDCD6*	1.06(2.9)	0.15	1.57×10^{-10}	0.67(2.0)	0.08	6.22×10^{-17}	0.50(1.7)	0.07	1.08×10^{-11}	0.69(2.0)	0.08	4.07×10^{-16}
	PDCD6[a]	0.76(2.1)	0.19		0.45(1.6)	0.09	1.07×10^{-6}	0.34(1.4)	0.08		0.50(1.7)	0.10	2.26×10^{-7}
	MYO6	0.44(1.6)	0.08	4.54×10^{-7}									
	MYO6[b]	0.40(1.5)	0.08	7.30×10^{-6}									

Genes associated to an exome-wide significance level ($p<2.5 \times 10^{-6}$) with hearing aid (H-aid), hearing difficulty (H-diff), hearing difficulty with background noise (H-noise), and the combined trait (H-both). Using rare-variant aggregate association tests pLoF or missense + pLoF variants with a MAF<0.01 in gnomAD v2.1.1 were analyzed in the discovery and mega samples of white European individuals from the UK Biobank The p-values for replicated associations [empirical p-values <0.05 adjusting for genes (pLoF and missense & pLoF] and traits brought to replication] are shown in red

Manhattan Plot Rare Variant Aggregate Analysis – Discovery Sample



Analysis ARHL using 470,000 UK Biobank Exomes

- Analyzed White Europeans
  - With exome sequence and ARHL phenotype data
    - Cases with hearing difficulty (N=110,249)
    - Controls (N=231,960)
- Analysis the same as for the 200K exomes with a few exceptions
  - Rare variant aggregate analysis was performed using a MAF<0.005 instead of MAF<0.01
  - Both a burden test and SKAT-O was used to perform the rare variant aggregate analyses
  - Two variant classifications were used for the rare-variant aggregate analyses
    - pLoF – (includes variants within ± 2 bps from the splice site)
    - pLoF + with a CADD >20 [missense and splice region variants i.e., (> ± 2 bps and < ± 12 bps from the splice site)]

Single Variant Analysis MAF<0.005

Chr	Gene	Variant	rsID	EAF	H-diff β (SE)	H-diff P	HL Phenotypes [1,2]
4	*WFS1*	c.2470G>A:p.(E824K)	rs367547063	9.4×10^{-4}	1.29(0.26)	4.3×10^{-7}	Wolfram Syndrome 1, DFNA6/14/38
5	*POU4F3*	c.694G>C:p.(E232Q)	rs2126961780	1.1×10^{-4}	3.57(1.58)	1.6×10^{-4}	DFNA15
6	*COL11A2*	c.239T>C:p.(F80S)	rs1459651793	1.0×10^{-4}	2.25(0.30)	$1.2\times10^{-16**}$	DFNA13, DFNB53, ARHL[3]
6	*MYO6*	c.737A>G:p.(H246R)	rs121912560	5.4×10^{-3}	3.28(0.57)	$6.7\times10^{-17**}$	DFNA22, DFNB37, ARHL[3,4]
7	*SLC26A5*	g.103411631T>C	rs370598077	3.8×10^{-3}	0.25(0.04)	$8.2\times10^{-9**}$	DFNB61
7	*SLC26A5*	c.137T>C:p.(L46P)	rs141952919	5.3×10^{-3}	0.22(0.04)	$8.0\times10^{-8**}$	DFNB61
16	*TBC1D24*	c.920A>G:p.(N307S)	rs761934676	9.5×10^{-3}	1.46(0.26)	$4.8\times10^{-8**}$	DFNA65, ARHL[3]
17	*MYO15A*	c.346_355delinsG: p.(Y117_R119del)	rs876657898	1.2×10^{-3}	0.40(0.08)	1.9×10^{-7}	DFNB3
19	*CEACAM16*	c.95G>T:p.(S32I)	rs186687142	1.5×10^{-3}	0.32(0.07)	3.4×10^{-6}	DFNA4B, DFNB113
19	*CEACAM16*	c.96C>T:p.(S32S)	rs191552868	1.5×10^{-3}	0.32(0.07)	2.5×10^{-6}	DFNA4B, DFNB113
5	*PPWD1*	g.65572309C>T	rs113550012	4.0×10^{-3}	1.68(0.48)	1.9×10^{-4}	
16	*ZNF598*	c.443G>A:p.(R148H)	rs746426951	8.6×10^{-3}	1.45(0.27)	$4.2\times10^{-6**}$	
5	*PDCD6*	c.132A>G:p.(S44S)	rs537688122	5.4×10^{-4}	1.32(0.11)	$8.7\times10^{-34**}$	
5	*PDCD6*	c.139G>C:p.(E47Q)	rs549592074	1.9×10^{-4}	1.36(0.18)	$1.9\times10^{-14**}$	
5	*PDCD6*	c.146A>G:p.(Q49R)	rs571370281	4.2×10^{-4}	1.34(0.12)	$3.2\times10^{-29**}$	
6	*FILIP1*	g.75362956T>C	rs765264064	9.9×10^{-3}	1.38(0.26)	$3.1\times10^{-6**}$	
22	*NCAPH2*	g.50517688G>A	rs200126237	1.1×10^{-3}	0.45(0.08)	$1.8\times10^{-6**}$	
22	*KLHDC7B*	p.G943A&*34	rs749405486	5.8×10^{-4}	0.83(0.10)	$2.7\times10^{-15**}$	

Variants in red significant for problems hearing and variants in black were significant for hearing aids or hearing difficulty in background noise

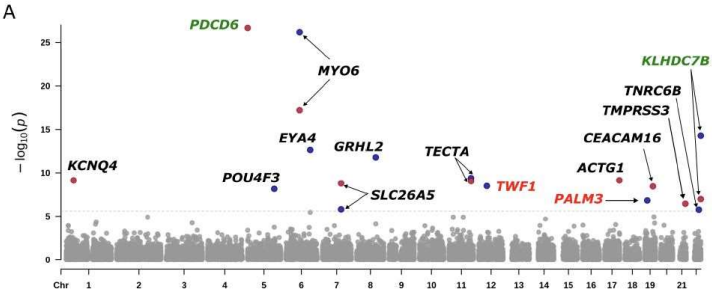
SKAT-O Rare Variant Aggregate Results – pLoF variants

pLoF variants						
Mendelian HL genes associated with ARHL						
Chr	Gene	H-aid p-value	H-diff p-value	H-noise p-value	H-both p-value	HL Phenotypes
3	*PLS1*	5.71×10^{-5}	$1.22\times10^{-7**}$	2.96×10^{-6}	$4.32\times10^{-9**}$	DFNA76
5	*POU4F3*	$6.97\times10^{-9**}$	1.43×10^{-4}	3.92×10^{-4}	5.30×10^{-5}	DFNA15, ARHL
6	*MYO6*	$6.74\times10^{-27**}$	$2.60\times10^{-17**}$	$1.93\times10^{-13**}$	$5.30\times10^{-18**}$	DFNA22, DFNB37, ARHL
6	*EYA4*	$2.37\times10^{-13**}$	3.59×10^{-4}	1.87×10^{-2}	2.94×10^{-4}	DFNA10, ARHL
7	*SLC26A5*	$1.55\times10^{-6*}$	$5.41\times10^{-12**}$	$4.35\times10^{-8**}$	$7.23\times10^{-12**}$	DFNB61, ARHL
8	*GRHL2*	$1.69\times10^{-12**}$	1.58×10^{-4}	7.25×10^{-3}	7.26×10^{-5}	DFNA28
11	*TECTA*	$4.03\times10^{-10**}$	$1.22\times10^{-13**}$	$8.80\times10^{-11**}$	$2.63\times10^{-13**}$	DFNA8/12, DFNB21, ARHL
14	*SIX1*	2.04×10^{-5}	$6.03\times10^{-8**}$	4.10×10^{-6}	$9.81\times10^{-8**}$	DFNA23, BORS, ARHL
22	*TNRC6B*	$1.71\times10^{-6*}$	7.52×10^{-3}	2.77×10^{-2}	1.67×10^{-3}	Developmental Delay (with HL)
Novel genes associated with ARHL						
12	*TWF1*	$3.20\times10^{-9**}$	7.3×10^{-6}	1.60×10^{-4}	$1.90\times10^{-6*}$	HL in Dalmatian dogs
19	*PALM3*	$1.49\times10^{-7**}$	1.78×10^{-4}	1.69×10^{-3}	1.96×10^{-5}	Mouse HL gene[8]
Genes previously associated with ARHL that are not know to be involved in Mendelian HL						
22	*KLHDC7B*[3,4]	$5.11\times10^{-15**}$	$6.65\times10^{-28**}$	$2.29\times10^{-19**}$	$3.23\times10^{-28**}$	

SKAT-O Rare Variant Aggregate Results pLoF + missense and splice region variants with CADD>20

pLoF, missense and splice-region variants with CADD≥20						
Mendelian HL genes associated with ARHL						
CHR	Gene	H-aid p-value	H-diff p-value	H-noise p-value	H-both p-value	HL Phenotypes
1	*KCNQ4*	$7.28\times10^{-10**}$	1.40×10^{-5}	1.30×10^{-4}	$1.41\times10^{-6*}$	DFNA2A
6	*MYO6*	$6.40\times10^{-18**}$	$7.87\times10^{-12**}$	$1.38\times10^{-6*}$	$3.99\times10^{-11**}$	DFNA22; DFNB37, ARHL
7	*SLC26A5*	$1.61\times10^{-9**}$	$8.60\times10^{-19**}$	$8.06\times10^{-14**}$	$9.09\times10^{-20**}$	DFNB61, ARHL
11	*MYO7A*	1.24×10^{-5}	3.46×10^{-6}	6.36×10^{-5}	$7.82\times10^{-7**}$	DFNA11; DFNB2; Usher syndrome
11	*TECTA*	$8.88\times10^{-10**}$	$1.07\times10^{-8**}$	6.55×10^{-6}	$8.12\times10^{-10**}$	DFNA8/12; DFNB21, ARHL
17	*ACTG1*	$7.17\times10^{-10**}$	8.50×10^{-5}	6.91×10^{-3}	2.67×10^{-5}	DFNA20/26
19	*CEACAM16*	$3.58\times10^{-9**}$	$1.31\times10^{-13**}$	$6.62\times10^{-8**}$	$2.94\times10^{-14**}$	DFNA4B; DFNB113, ARHL
21	*TMPRSS3*	$3.52\times10^{-7**}$	$1.19\times10^{-6**}$	3.42×10^{-4}	4.63×10^{-6}	DFNB8/10
Novel genes associated with ARHL						
1	*FBXO2*	1.20×10^{-4}	$1.46\times10^{-6*}$	1.98×10^{-5}	$6.79\times10^{-7**}$	ARHL in mice
17	*TXNDC17*	1.76×10^{-2}	$9.53\times10^{-7**}$	2.34×10^{-4}	4.14×10^{-6}	
Genes previously associated with ARHL that are not known to be involved in Mendelian HL						
5	*PDCD6*[4]	$2.21\times10^{-27**}$	$4.51\times10^{-41**}$	$2.09\times10^{-24**}$	$1.95\times10^{-41**}$	
22	*KLHDC7B*[3,4]	$1.03\times10^{-7**}$	$2.49\times10^{-16**}$	$6.58\times10^{-10**}$	$7.47\times10^{-12**}$	

Manhattan Plot Hearing Difficulty Rare Variant Aggregate



Results from SKAT-O

Overview

- For rare variant aggregate analysis results were consistent between SKAT-O and the burden test
  - Although more significant results were obtained from the SKAT-O
    - And some variants that were significant for SKAT-O were not significant for the burden test
- Mendelian genes (although not necessarily the same variants as for Mendelian hearing impairment) play an important role in ARHL
- For additional information see
  - Cornejo-Sanchez et al. (2023 and 2025)

Future Direction

- Replicate findings
  - for example, in All of Us
- Performing Mendelian Randomization and testing for pleiotropy (vertical & horizontal) to evaluate associations between ARHL and comorbidities
  - e.g., comorbidities dementia, depression
- Analysis of UK Biobank and All of Us WGS data including
  - structural variants
  - Performing rare variant aggregate tests outside of the coding regions

Newcastle University

wellcome

Genome-wide association studies (GWAS) - Part 2

More advanced topics: Linear Mixed Models and G×G or G×E interactions

Heather J. Cordell

Population Health Sciences Institute
Faculty of Medical Sciences
Newcastle University, UK

heather.cordell@ncl.ac.uk

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing linkage analysis and estimating heritability
  - Using family (pedigree) data i.e. related individuals

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing linkage analysis and estimating heritability
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing linkage analysis and estimating heritability
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure
  - Estimating the heritability accounted for various partitions of SNPs:
    - All SNPs typed on a GWAS panel
    - All typed SNPs and others in LD with them
    - Partitions of SNPs in various functional categories

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing linkage analysis and estimating heritability
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure
  - Estimating the heritability accounted for various partitions of SNPs:
    - All SNPs typed on a GWAS panel
    - All typed SNPs and others in LD with them
    - Partitions of SNPs in various functional categories
  - Investigating genetic correlations between different traits

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing linkage analysis and estimating heritability
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure
  - Estimating the heritability accounted for various partitions of SNPs:
    - All SNPs typed on a GWAS panel
    - All typed SNPs and others in LD with them
    - Partitions of SNPs in various functional categories
  - Investigating genetic correlations between different traits
  - Predicting trait values in a new individual

Linear Mixed Models (LMMs)

- A linear mixed model is a statistical model in which the dependent variable is a linear function of both fixed and random independent variables
  - Known respectively as fixed and random effects
  - Fixed effects are considered 'fixed' at their measured values
  - Random effects are considered to be sampled from a distribution

Linear Mixed Models (LMMs)

- A linear mixed model is a statistical model in which the dependent variable is a linear function of both fixed and random independent variables
  - Known respectively as fixed and random effects
  - Fixed effects are considered 'fixed' at their measured values
  - Random effects are considered to be sampled from a distribution
- Recall the usual linear regression model

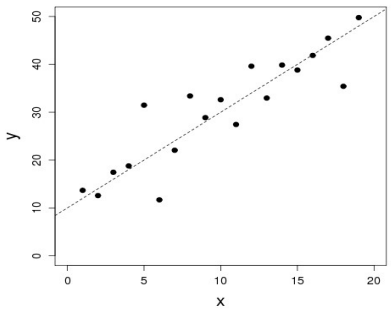
$$y = mx + c \qquad \text{or} \qquad y = \beta_0 + \beta_1 x$$

- This model may also be written

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  - y_i refers to the trait value of person i
  - x_i refers to the measured value of person i's predictor variable
  - ϵ_i refers to the displacement from the regression line
    - i.e. the discrepency between the observed and the predicted y value

Linear Regression



Linear Mixed Models (LMMs)

- In linear regression we have $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - Here β_0 and β_1 are fixed effects while ϵ_i is a random error
  - x_i is the 'loading' of fixed effect β_1 that someone has (based on their genotype)
  - β_0, β_1, ϵ_i are unknown (to be estimated given data on y_i and x_i $\forall_i$)

Linear Mixed Models (LMMs)

- In linear regression we have $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - Here β_0 and β_1 are fixed effects while ϵ_i is a random error
  - x_i is the 'loading' of fixed effect β_1 that someone has (based on their genotype)
  - β_0, β_1, ϵ_i are unknown (to be estimated given data on y_i and x_i $\forall_i$)
- In matrix notation we can write this model:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

  - or $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$

Linear Mixed Models (LMMs)

- In linear regression we have $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - Here β_0 and β_1 are fixed effects while ϵ_i is a random error
  - x_i is the 'loading' of fixed effect β_1 that someone has (based on their genotype)
  - β_0, β_1, ϵ_i are unknown (to be estimated given data on y_i and x_i $\forall_i$)
- In matrix notation we can write this model:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

  - or $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$
- A LMM takes the form $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$
  - where $\boldsymbol{u}$ corresponds to a vector of random effects
    - with loadings specified in $\boldsymbol{Z}$

Linear Mixed Models (LMMs)

- E.g. suppose 2 fixed effects β_1 and β_2, and 3 random effects (plus n random errors)
- Then $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$ corresponds to:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ . & . \\ . & . \\ x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ . & . & . \\ . & . & . \\ z_{n1} & z_{n2} & z_{n3} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

  - or $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + u_1 z_{i1} + u_2 z_{i2} + u_3 z_{i3} + \epsilon_i$

LMMs in genetics

- In genetics we generally work with two equivalent forms of LMM
- One is: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$
  - The (unknown) random effect u_l corresponds to a scaled additive effect of causal variant (locus) l
    - We assume there are many (m) such causal variants all across the genome
    - Considering it to be a random effect (within a population of interest) could be thought of as taking a Bayesian perspective

LMMs in genetics

- In genetics we generally work with two equivalent forms of LMM
- One is: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$
  - The (unknown) random effect u_l corresponds to a scaled additive effect of causal variant (locus) l
    - We assume there are many (m) such causal variants all across the genome
    - Considering it to be a random effect (within a population of interest) could be thought of as taking a Bayesian perspective
  - $\boldsymbol{Z}$ is usually a standardized genotype matrix i.e. z_{il} takes value

$$\left(\frac{-2f_l}{\sqrt{2f_l(1-f_l)}}, \frac{(1-2f_l)}{\sqrt{2f_l(1-f_l)}}, \frac{2(1-f_l)}{\sqrt{2f_l(1-f_l)}} \right)$$

    if individual i has genotype $(1/1, 1/2, 2/2)$
    - where f_l is the frequency of allele 2 at locus l

LMMs in genetics

- The other form is: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{g} + \boldsymbol{\epsilon}$
  - Where $g_i = \sum_{l=1}^{m} z_{il} u_l$ is the total genetic effect in individual i, summed over all the causal loci
- In this form, g_i **can be considered as a random effect operating in individual** i
  - The vector of random effects $\boldsymbol{g}$ takes distribution $\boldsymbol{g} \sim N(0, \boldsymbol{G}\sigma_a^2)$
    - Where $\boldsymbol{G}$ is the genetic relationship matrix (GRM) between individuals – i.e. their IBD sharing at the causal loci
    - $\sigma_a^2 = m\sigma_u^2$ is the total additive genetic variance
    - $\boldsymbol{G}$=$\boldsymbol{ZZ'}/m$

LMMs in genetics

- The other form is: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{g} + \boldsymbol{\epsilon}$
  - Where $g_i = \sum_{l=1}^{m} z_{il} u_l$ is the total genetic effect in individual i, summed over all the causal loci
- In this form, g_i **can be considered as a random effect operating in individual** i
  - The vector of random effects $\boldsymbol{g}$ takes distribution $\boldsymbol{g} \sim N(0, \boldsymbol{G}\sigma_a^2)$
    - Where $\boldsymbol{G}$ is the genetic relationship matrix (GRM) between individuals – i.e. their IBD sharing at the causal loci
    - $\sigma_a^2 = m\sigma_u^2$ is the total additive genetic variance
    - $\boldsymbol{G}$=$\boldsymbol{ZZ'}/m$
- For family data (close relatives), the expected values of the elements of $\boldsymbol{G}$ equal the expected IBD sharing
  - i.e. twice the kinship coefficients
  - Thus $\boldsymbol{G}$ is just equal to twice the kinship matrix
    - Models their expected relatedness at the causal loci (and elsewhere)

Use of LMMs in genetics

- The formulation $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{g} + \boldsymbol{\epsilon}$ is known as the **Animal Model** and has been used extensively in plant and animal breeding
  - Mostly to predict the *breeding values* g_i in order to inform breeding strategies
    - E.g. to increase milk yield, meat production etc. etc.
  - Similar approaches could be used for *prediction* of trait values given genotype data
- In the mid 1990s it became popular in human genetics as the backbone of **variance components linkage analysis**
- Now commonly used in association analysis (GWAS)
  - To correct for relatedness, when testing for association

Testing for association using LMMs

- Idea is to test a fixed SNP effect β_1
  - While including a random effect γ_i that models relatedness
- Fit regression model: $y_i = \beta_0 + \beta_1 x_i + \gamma_i$
  - y is the trait value
  - x is a variable coding for genotype at the test SNP (e.g. an allele count, coded 0, 1, 2 for genotypes 1/1, 1/2, 2/2)
  - $\gamma_i = g_i + \epsilon_i$

Testing for association using LMMs

- Idea is to test a fixed SNP effect β_1
  - While including a random effect γ_i that models relatedness
- Fit regression model: $y_i = \beta_0 + \beta_1 x_i + \gamma_i$
  - y is the trait value
  - x is a variable coding for genotype at the test SNP (e.g. an allele count, coded 0, 1, 2 for genotypes 1/1, 1/2, 2/2)
  - $\gamma_i = g_i + \epsilon_i$
  - We assume $\boldsymbol{\gamma} \sim MVN(0, \boldsymbol{V})$ where variance/covariance matrix $\boldsymbol{V}$ follows standard variance components model
    - Variance/covariance matrix structured as:

$$\begin{aligned} V_{ij} &= \sigma_a^2 + \sigma_e^2 \quad (i = j) \\ V_{ij} &= 2\Phi_{ij}\sigma_a^2 \quad (i \neq j) \end{aligned}$$

    - σ_a^2, σ_e^2 represent the additive polygenic variance (due to all loci) and the environmental (=error) variance, respectively

Testing for association using LMMs

- LMMs were first (?) applied in human genetics by Boerwinkle et al. (1986) and Abney et al. (2002)
- Chen and Abecasis (2007) implemented them via the "FAmily based Score Test Approximation" (FASTA) in the MERLIN software package
  - Closely related to earlier QTDT method (Abecasis et al. 2000a;b) which implements a slightly more general/complex model
  - FASTA was also implemented in GenABEL, along with a similar test called GRAMMAR (Aulchenko et al. 2007)

Estimating the genetic relationship matrix

- These early implementations calculated the kinship matrix Φ on the basis of known (theoretical) kinships constructed from known pedigree relationships
- Amin et al. (2007) proposed instead *estimating* the kinships based on genome-wide SNP data
  - Ideally we want to use $\boldsymbol{G}=\boldsymbol{ZZ'}/m$, the genetic relationship matrix (GRM) between individuals at the causal loci
  - Since we don't know the causal loci, we approximate $\boldsymbol{G}$ by $\boldsymbol{A}$, the overall GRM between individuals

Estimating the genetic relationship matrix

- Once you move to estimating the GRM, you are no longer limited to using family data
- Kang et al. (2010) and Zhang et al. (2010) suggested applying the approach to apparently unrelated individuals
  - As a way of accounting for population substructure/stratification
  - Also proposed applying to binary traits (case/control coded 1/0)
  - Implemented in EMMAX and TASSEL software, respectively

Estimating the genetic relationship matrix

- Once you move to estimating the GRM, you are no longer limited to using family data
- Kang et al. (2010) and Zhang et al. (2010) suggested applying the approach to apparently unrelated individuals
  - As a way of accounting for population substructure/stratification
  - Also proposed applying to binary traits (case/control coded 1/0)
  - Implemented in EMMAX and TASSEL software, respectively
- Subsequently a number of other publications/software packages have implemented essentially the same model
  - FaST-LMM (Lippert et al. 2011)
  - GEMMA (Zhou and Stephens 2012)
  - GenABEL (GRAMMAR-Gamma) (Svishcheva et al. 2012)
  - MMM (Pirinen et al. 2013)
  - MENDEL (Zhou et al. 2014)
  - RAREMETALWORKER
  - GCTA
  - DISSECT

Software implementations

- Main difference between them is the precise computational tricks used to speed up the calculations
  - And the convenience/ease of use
    - See comparison in Eu-Ahsunthornwattana et al. (2014) PLoS Genetics 10(7):e1004445

Software implementations

- Main difference between them is the precise computational tricks used to speed up the calculations
  - And the convenience/ease of use
    - See comparison in Eu-Ahsunthornwattana et al. (2014) PLoS Genetics 10(7):e1004445
- BOLT-LMM (Loh et al. 2016) uses a slightly different approach, based on a Bayesian implementation of LMM formulation 1:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$$

  - One of the first mixed model packages that worked for really large-scale (e.g. UK Biobank) datasets
  - Now potentially (?) superseded by fastGWA module in GCTA, which achieves similar ends by using a sparse GRM
  - And by REGENIE, which uses a conceptually similar formulation, but achieves savings in compute time and memory by analysing the residuals following a whole-genome blockwise ridge regression
  - See also LDAK-KVIK

Binary traits

- For binary traits, coding cases and controls as a 1/0 quantitative trait is not optimal
  - Though in practice it seems to work reasonably well
- LTMLM (Hayeck et al. 2015) and LEAP (Weissbrod et al. 2015) instead use an underlying *liability model* to improve power
  - Assuming known disease prevalence

Binary traits

- For binary traits, coding cases and controls as a 1/0 quantitative trait is not optimal
  - Though in practice it seems to work reasonably well
- LTMLM (Hayeck et al. 2015) and LEAP (Weissbrod et al. 2015) instead use an underlying *liability model* to improve power
  - Assuming known disease prevalence
- Chen et al. (2016) showed that high levels of population stratification can invalidate the analysis, when applied to a case/control sample
  - Resulting in a mixture of **inflated** and **deflated** test statistics
  - Developed GMMAT software to address this problem
  - Similar functionality is available in the CARAT software (Jiang et al. 2016, AJHG 98:243-55)

Case/control imbalance for binary traits

- SAIGE software (Zhou et al. 2018, AJHG 50(9):1335-1341) implements a mixed model test that deals with large case-control imbalance, as you might see (for example) in UK Biobank
- REGENIE also implements this same saddle point approximation (SPA) test
  - Along with an approximate Firth penalized likelihood-ratio test
- SPA also implemented in LDAK-KVIK

Estimating heritability using GCTA

- Seminal paper by Yang et al. (2010) [Nat Genet 42(7):565-9]
- Showed that by framing the relationship between height and genetic factors as an LMM, 45% of the trait variance could be explained by considering 294,831 SNPs simultaneously in the GRM
  - So-called 'SNP heritability' or 'chip heritability'
  - Modelling effects at all genotyped SNPs explained the 'known' heritability ($\approx$ 80%) much better than just the top SNPs from GWAS
- Moreover, if you estimate (and correct for) the deflation caused by fact that the genotyped SNPS are not in full LD with the causal SNPs, the variance explained goes up to 84% (s.e. 16%), consistent with 'known' heritability
- Subsequently many papers have shown similar results for a variety of complex traits

Estimating SNP heritability using GCTA

- Basic idea is to use formulation

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{g} + \boldsymbol{\epsilon}$$

with $\boldsymbol{g} \sim N(0, \boldsymbol{A}\sigma_a^2)$ and $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{I}\sigma_e^2)$ so $\boldsymbol{V} = \boldsymbol{A}\sigma_a^2 + \boldsymbol{I}\sigma_e^2$

  - $\boldsymbol{X\beta}$ are the loadings of fixed effects (e.g. covariates) to be (optionally) included
  - $\boldsymbol{A}$ is the GRM between individuals, estimated using all genotyped SNPs
  - σ_a^2 and σ_e^2 estimated using REML (or MLE)
  - Thus we can estimate the heritability accounted for by the genotyped SNPs as $\sigma_a^2/(\sigma_a^2 + \sigma_e^2)$

- Implemented in several software packages including GCTA (GREML) and DISSECT
  - ALBI software (Schweiger et al. 2016, AJHG 98:1181-1192) can then be used to construct accurate confidence intervals for the heritability

LD Score regression

- Some recent work has focussed on achieving similar ends
  - i.e. estimating
    - heritability explained by sets of SNPs
    - SNP-based genetic correlations across traits

  using summary statistics only

  - Bulik-Sullivan et al. (2015) [Nat Genet 47:291-295]
  - Bulik-Sullivan et al. (2015) [Nat Genet 47:1236-1241]
    - Clever idea that allows the variance component parameters to be estimated via a simple regression on 'LD Scores'
    - LDSC software (https://github.com/bulik/ldsc)

Bulik-Sullivan et al. (2015) [Nat Genet 47:291-295]

- "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies"
- Main idea was to come up with a "better" test statistic correction for GWAS than the inflation factor (λ) used in genomic control
  - Which they argued was often too stringent, based on the fact that polygenicity accounts for the majority of test statistic inflation in GWAS of large sample size
  - Variants in LD with causal variant(s) will show elevated test statistics in association analysis, proportional to the LD (measured by r^2) with the causal variant(s)
    - In contrast, inflation from cryptic relatedness or population stratification will not correlate with amount of LD

Bulik-Sullivan et al. (2015) [Nat Genet 47:291-295]

- Assume a polygenic model, in which effect sizes for variants are drawn independently from distributions with variance $\propto 1/(p(1-p))$
  - p is the minor allele frequency (MAF)

Bulik-Sullivan et al. (2015) [Nat Genet 47:291-295]

- Assume a polygenic model, in which effect sizes for variants are drawn independently from distributions with variance $\propto 1/(p(1-p))$
  - p is the minor allele frequency (MAF)
- Then the expected χ^2 statistic of variant j is:

$$E(\chi^2|l_j) = \frac{Nh^2}{M} l_j + Na + 1$$

  - N is the sample size
  - M is the number of SNPs
  - h^2/M is the average heritability explained per SNP
  - a measures the contribution of confounding biases, such as cryptic relatedness and population stratification
  - $l_j = \sum_k r_{jk}^2$ is the LD Score of variant j, which measures the amount of genetic variation tagged by j

Bulik-Sullivan et al. (2015) [Nat Genet 47:291-295]

- Assume a polygenic model, in which effect sizes for variants are drawn independently from distributions with variance $\propto 1/(p(1-p))$
  - p is the minor allele frequency (MAF)
- Then the expected χ^2 statistic of variant j is:

$$E(\chi^2|l_j) = \frac{Nh^2}{M} l_j + Na + 1$$

  - N is the sample size
  - M is the number of SNPs
  - h^2/M is the average heritability explained per SNP
  - a measures the contribution of confounding biases, such as cryptic relatedness and population stratification
  - $l_j = \sum_k r_{jk}^2$ is the LD Score of variant j, which measures the amount of genetic variation tagged by j
- Thus, if you regress the χ^2 statistics of variants from GWAS against their LD Scores, the intercept minus one is an estimator of the mean contribution of confounding bias to the inflation in the test statistics.
  - And the slope times M/N is an estimate of the overall SNP heritability

Bulik-Sullivan et al. (2015) [Nat Genet 47:1236-1241]

- "An atlas of genetic correlations across human diseases and traits"
- Main focus is introducing a technique—cross-trait LD Score regression—for estimating genetic correlation that requires only GWAS summary statistics and is not biased by sample overlap
  - Via a simple extension of single-trait LD Score regression

Bulik-Sullivan et al. (2015) [Nat Genet 47:1236-1241]

- "An atlas of genetic correlations across human diseases and traits"
- Main focus is introducing a technique—cross-trait LD Score regression—for estimating genetic correlation that requires only GWAS summary statistics and is not biased by sample overlap
  - Via a simple extension of single-trait LD Score regression
- Under a polygenic model, if z_{ij} is the z score for association for trait i and SNP j, then the expected value of $z_{1j}z_{2j}$ for a SNP j is

$$E[z_{1j}z_{2j}|l_j] = \frac{\sqrt{N_1N_2}\rho_g}{M} l_j + \frac{\rho N_S}{\sqrt{N_1N_2}}$$

  - N is the sample size for study i
  - ρ_g is the genetic covariance
  - l_j is the LD score for SNP j
  - N_s is the number of individuals included in both studies
  - ρ is the phenotypic correlation among the N_S overlapping samples

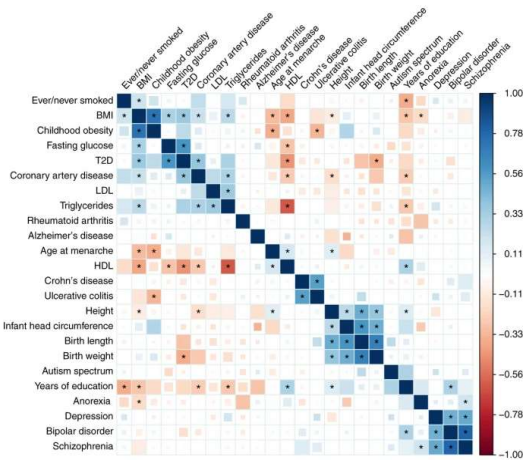
Bulik-Sullivan et al. (2015) [Nat Genet 47:1236-1241]

- We have

$$E[z_{1j}z_{2j}|l_j] = \frac{\sqrt{N_1N_2}\rho_g}{M} l_j + \frac{\rho N_S}{\sqrt{N_1N_2}}$$

- We can estimate genetic covariance ρ_g using the slope from the regression of $z_{1j}z_{2j}$ on LD Score $z_{1j}z_{2j}$
  - If study 1 and study 2 are the same study, then $N_1 = N_2 = N_S$ and $\rho_g = h_g^2$, so this reduces to LD Score regression for a single trait
- Normalizing genetic covariance by SNP heritabilities yields genetic correlation $r_g = \rho_g / \sqrt{h_1^2 h_2^2}$ where h_i^2 denotes the SNP heritability from study i.
- Bulik-Sullivan et al. point out that the heritability parameter estimated by LDSC is subtly different from the heritability parameter h_g^2 estimated by GCTA
  - And similarly for the genetic covariance parameter ρ_g

Genetic correlations among 24 traits



Short break

Gene-gene (and gene-environment) interactions

- GWAS have been extraordinarily successful at detecting genetic locations harboring genes associated with complex disease
  - But the SNPs identified do not account for the known (estimated) heritability for most disorders
  - Could G×G and G×E effects account for part of the 'missing heritability'?
    - Zuk et al. (2012) PNAS 109:1193-1198

Gene-gene (and gene-environment) interactions

- GWAS have been extraordinarily successful at detecting genetic locations harboring genes associated with complex disease
  - But the SNPs identified do not account for the known (estimated) heritability for most disorders
  - Could G×G and G×E effects account for part of the 'missing heritability'?
    - Zuk et al. (2012) PNAS 109:1193-1198
- Effects operating through interactions may not be visible unless you stratify by or take account of the interacting genetic (or environmental) factors
  - By modelling interactions, we hope to increase our power to detect loci with weak marginal effects

Gene-gene (and gene-environment) interactions

- GWAS have been extraordinarily successful at detecting genetic locations harboring genes associated with complex disease
  - But the SNPs identified do not account for the known (estimated) heritability for most disorders
  - Could G×G and G×E effects account for part of the 'missing heritability'?
    - Zuk et al. (2012) PNAS 109:1193-1198
- Effects operating through interactions may not be visible unless you stratify by or take account of the interacting genetic (or environmental) factors
  - By modelling interactions, we hope to increase our power to detect loci with weak marginal effects
- Phenomenon of biological interest?
  - Identifying genes that interact to cause disease could help us understand the mechanisms and pathways in disease progression

Heather Cordell (Newcastle) GWAS (Part 2) 26 / 41

Definition of (pairwise) interaction

- Statistical interaction most easily described in terms a of (logistic) regression framework
  - Supppose x_1 and x_2 are binary factors whose presence/absence (coded 1/0) may be associated with a disease outcome
  - Logistic regression models their effect on the log odds of disease as:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

Marginal effect of factor 1

$$\log \frac{p}{1-p} = \beta_0 + \beta_2 x_2$$

Marginal effect of factor 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Main effects of factors 1 and 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

Main effects and interaction term

- For quantitative traits, use linear regression (replace $\log \frac{p}{1-p}$ with y)
- For modelling as an LMM, add in a random effect γ

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

- β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

- β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
  - Having factor 1 adds β_1 to your trait value

Heather Cordell (Newcastle) GWAS (Part 2) 28 / 41

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

- β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
  - Having factor 1 adds β_1 to your trait value
  - Having factor 2 adds β_2 to your trait value

Heather Cordell (Newcastle) GWAS (Part 2) 28 / 41

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value
    - Having both factors adds an additional β_{12} to your trait value
      $\Rightarrow$ Implies that the overall effect of two variables is greater (or less) than the 'sum of the parts'
    - The 'effect' of factor 2 is different in the presence/absence of factor 1

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value
    - Having both factors adds an additional β_{12} to your trait value
      $\Rightarrow$ Implies that the overall effect of two variables is greater (or less) than the 'sum of the parts'
    - The 'effect' of factor 2 is different in the presence/absence of factor 1
- Suppose no main effects ($\beta_1 = \beta_2 = 0$)

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_{12}$	β_0
0	β_0	β_0

  - Trait value only differs from baseline if both factors present

Gene-gene interaction (epistasis)

- However SNPs are not binary, but rather take 3 levels according to the number of copies (0,1,2) of the susceptibility allele possessed
- Most general 'saturated' (9 parameter) genotype model allows all 9 penetrances to take different values
  - Via modelling log odds in terms of:
    - A baseline effect (β_0)
    - Main effects of locus G (β_{G_1}, β_{G_2})
    - Main effects of locus H (β_{H_1}, β_{H_2})
    - 4 interaction terms

	Locus H		
Locus G	2	1	0
2	$\beta_0+\beta_{G_2}+\beta_{H_2}+\beta_{22}$	$\beta_0+\beta_{G_2}+\beta_{H_1}+\beta_{21}$	$\beta_0+\beta_{G_2}$
1	$\beta_0+\beta_{G_1}+\beta_{H_2}+\beta_{12}$	$\beta_0+\beta_{G_1}+\beta_{H_1}+\beta_{11}$	$\beta_0+\beta_{G_1}$
0	$\beta_0+\beta_{H_2}$	$\beta_0+\beta_{H_1}$	β_0

  - Corresponds in statistical analysis packages to coding x_1, x_2 (0,1,2) as a "factor"

Gene-gene interaction

- Alternatively we can assume additive effects of each allele at each locus:
  - Corresponds to fitting

$$\log \frac{p}{1-p} = \beta_0 + \beta_G x_1 + \beta_H x_2 + \beta_{GH} x_1 x_2$$

  with x_1, x_2 coded (0,1,2)

	Locus H		
Locus G	2	1	0
2	$\beta_0 + 2\beta_G + 2\beta_H + 4\beta_{GH}$	$\beta_0 + 2\beta_G + \beta_H + 2\beta_{GH}$	$\beta_0 + 2\beta_G$
1	$\beta_0 + \beta_G + 2\beta_H + 2\beta_{GH}$	$\beta_0 + \beta_G + \beta_H + \beta_{GH}$	$\beta_0 + \beta_G$
0	$\beta_0 + 2\beta_H$	$\beta_0 + \beta_H$	β_0

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact
- However, existence of statistical interaction does imply both loci are "involved" in disease in some way

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact
- However, existence of statistical interaction does imply both loci are "involved" in disease in some way
  - Good starting point for further investigation of their (joint) action

Gene-environment (G×E) interactions

- The same regression model

$$\log \frac{p}{1-p} = \beta_0 + \beta_G x_1 + \beta_H x_2 + \beta_{GH} x_1 x_2$$

can be used to model interaction between a genetic factor G and an environmental factor H

- With the environmental variable x_2 coded in binary fashion (e.g. smoking) or quantitatively (e.g. age)

Gene-environment (G×E) interactions

- The same regression model

$$\log \frac{p}{1-p} = \beta_0 + \beta_G x_1 + \beta_H x_2 + \beta_{GH} x_1 x_2$$

can be used to model interaction between a genetic factor G and an environmental factor H

- With the environmental variable x_2 coded in binary fashion (e.g. smoking) or quantitatively (e.g. age)
- Focus of analysis is often risk estimation
  - Estimating genetic risks in particular environments
  - Estimating effect of environmental factor on particular genetic background
    - Important for treatment/screening strategies and public health interventions
- For G×G, focus of interest is more related to
  - Increasing power to detect an effect (by taking into account the effects of other genetic loci)
  - Modelling the biology, especially related to the joint action of the loci

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

- 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone
- Depending on circumstances, any of these tests may be a sensible option

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone
- Depending on circumstances, any of these tests may be a sensible option
- Most tests of interaction/joint action can be thought of as a version of one or other of these tests
  - Although different tests vary in their precise details
  - And their relationship to the logistic regression formulation not always clearly described
    - See Howey and Cordell (2017) https://pubmed.ncbi.nlm.nih.gov/28852712/

G×G versus G×E in the context of GWAS

- Typically GWAS measure thousands if not millions of genetic variants
  - But only a few (tens or at most 100s) of environmental factors
- Feasible to consider all G×E combinations
- All pairwise G×G combinations possible, but much more time consuming
  - And leads to greater multiplicity of tests
  - Also, why stop at 2-way interactions?
    - Could look at all 3 way, 4 way etc. combinations
    - Scale of problem quickly gets out of hand
    - Less obvious reason to do this for G×E...

G×G in the context of GWAS

- Many recent publications have focussed on finding clever computational tricks to speed up exhaustive search procedure
  - BOOST (Wan et al. (2010) AJHG 87:325-340)
  - SIXPAC (Prabhu and Pe'er (2012) Genome Res 22:2230-2240)
  - Kam-Thong et al. (2012) Hum Hered 73:220-236 (GPUs)
  - Fråanberg et al. (2015) PLOS Genetics 11(9):e1005502 "Discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests"

G×G in the context of GWAS

- Many recent publications have focussed on finding clever computational tricks to speed up exhaustive search procedure
  - BOOST (Wan et al. (2010) AJHG 87:325-340)
  - SIXPAC (Prabhu and Pe'er (2012) Genome Res 22:2230-2240)
  - Kam-Thong et al. (2012) Hum Hered 73:220-236 (GPUs)
  - Fråanberg et al. (2015) PLOS Genetics 11(9):e1005502 "Discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests"
- Or have proposed filtering based on single-locus significance or other (biological or statistical) considerations
  - Reduces multiple testing burden, improves interpretability

G×G in the context of GWAS

- Many recent publications have focussed on finding clever computational tricks to speed up exhaustive search procedure
  - BOOST (Wan et al. (2010) AJHG 87:325-340)
  - SIXPAC (Prabhu and Pe'er (2012) Genome Res 22:2230-2240)
  - Kam-Thong et al. (2012) Hum Hered 73:220-236 (GPUs)
  - Fråanberg et al. (2015) PLOS Genetics 11(9):e1005502 "Discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests"
- Or have proposed filtering based on single-locus significance or other (biological or statistical) considerations
  - Reduces multiple testing burden, improves interpretability
- Or have proposed testing at the gene level rather than the SNP level
  - Ma et al. (2013) PLoS Genet 9(2): e1003321
    - Compared 4 different tests that combine P values from pairwise (SNP x SNP) interaction tests
    - Showed that the truncated tests did best
    - Presented an application only considering gene pairs known to exhibit protein-protein interactions

Case-only analysis

- Piergorsh et al. 1994; Yang et al. 1999; Weinberg and Umbach 2000
- Several authors have shown that, for binary predictor variables, a test of the interaction term β_{12} in the logistic regresssion model

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  can be obtained by testing for correlation (association) between the genotypes at two separate loci, within the sample of cases
- Gains power from making assumption that genotypes (alleles) at the two loci are uncorrelated in the population
  - So only really suitable for unlinked or loosely linked loci (since closely linked loci are likely to be in LD)
- Alternatively contrast the genotype correlations in cases with those seen in controls (`--fast-epistasis` in PLINK)

Testing correlation between loci

- A similar idea is implemented in EPIBLASTER (Kam-Thong et al. 2011; EJHG 19:465-571)
- Wu et al. (2010) (PLoS Genet 6:e1001131) also proposed a similar approach – though needs adjustment to give correct type I error rates
- See also Joint Effects (JE) statistics (Ueki and Cordell 2012; PLoS Genetics 8(4):e1002625)
- All these methods test whether correlation exists (case-only) or is different in cases and controls (case/control)
  - Via testing a log OR for association between two loci
  - However, the log OR for association (λ) encapsulates a slightly different quantity between the different methods
- All implemented (along with standard logistic and linear regression) in CASSI
  - http://www.staff.ncl.ac.uk/richard.howey/cassi/

Empirical evidence for G×G interactions

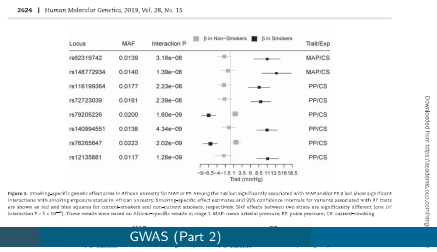
- Epistasis among *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* in multiple sclerosis (Lincoln et al. 2009 PNAS 106:7542-7547)
- *HLA-C* and *ERAP1* in psoriasis (Strange et al. 2010)
- *HLA-B27* and *ERAP1* in ankylosing spondylitis (Evans et al. 2011)
- *BANK1* and *BLK* in SLE (Castillejo-Lopez et al. 2012)
- Gusareva et al. (2014) found a reasonably convincing (partially replicating) interaction between SNPs on chromosome 6 (*KHDRBS2*) and 13 (*CRYL1*) in Alzheimer's disease
- Dai et al. (2016) [AJHG 99:352-365] identified 3 loci simultaneously interacting with established risk factors gastresophageal reflux, obesity and tobacco smoking, with respect to risk for Barrett's esophagus

Empirical evidence for G×G interactions

- Hemani et al. 2014 (Nature 508:249-253) found 501 instances of epistatic effects on gene expression, of which 30 could be replicated in two independent samples
  - Many SNPs are close together, could represent haplotype effects?
  - Or the effect of a single untyped variant?
  - See caveats in
    - Wood et al. (2014) Nature 514(7520):E3-5. PMID:25279928
    - Fish et al. (2016) Am J Hum Genet 99(4):817–830. PMID:27640306
- The Hemani et al. paper was subsequently retracted (https://www.nature.com/articles/s41586-021-03766-y)

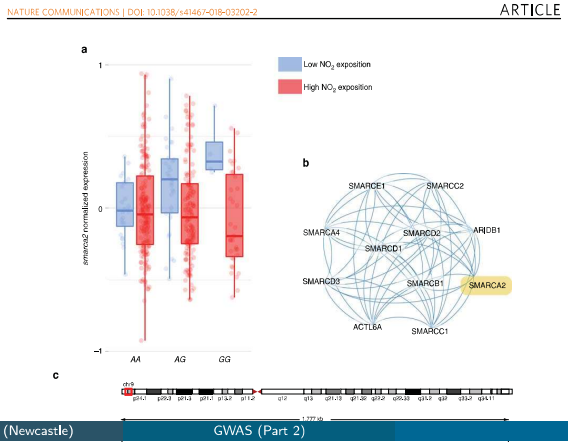
Empirical evidence for G×E interactions

- Myers et al. (2014) Hum Mol Genet 23(19): 5251-9 "Genome-wide Interaction Studies Reveal Sex-Specific Asthma Risk Alleles"
- Small et al. (2018) Nat Genet 50(4): 572-580 "Regulatory Variants at KLF14 Influence Type 2 Diabetes Risk via a Female-Specific Effect on Adipocyte Size and Body Composition"
- Sung et al. (2019) Hum Molec Genet 28(15): 2615-2633 "A multi-ancestry genome-wide study incorporating gene-smoking interactions identifies multiple new loci for pulse pressure and mean arterial pressure."



Empirical evidence for G×E interactions

- Favé et al. (2018) Nat Commun 9(1): 827 "Gene-by-environment Interactions in Urban Populations Modulate Risk Phenotypes"


NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-03202-2 ARTICLE

Power Analysis for Single and Rare Variant Aggregate Association Analyses

Suzanne M. Leal, Ph.D.
Sergievsky Family Professor of Neurological Sciences
Director of the Center for Statistical Genetics
Columbia University
sml3@Columbia.edu

Why Estimate Sample Sizes and/or Power? To avoid wasting time and money

- Does not make sense to perform an inadequately powered study
  - Unlikely to be able to correctly reject the null hypothesis
    - Due to inadequate sample size
- Collaborations and obtaining data from biobanks
  - Can aid in increasing sample sizes
    - Caveats
      - Disease definition may not be the same between studies
      - Study subjects may be drawn for different populations
      - Processing of genetic material maybe not be consistent

Why Estimate Sample Sizes and/or Power? Almost always necessary for grant proposals

- Can be denied funding if unable to demonstrate planned study has adequate power
  - Realistic disease models are necessary when performing power calculations
  - Use correct α
    - Which is corrected for multiple testing
      - » Which will be performed
    - e.g., use genome-wide significant level of 5×10^{-8} for GWAS studies

Power and Sample Size Estimation for Case-Control Data

- The correct α must be use for sample size estimation/power analysis
- Type I (α) the probability of rejecting the null hypothesis of no association when it is true
- Due to multiple testing a more stringent value than α=0.05 is used to control the Family Wise Error Rate

Power and Sample Size Estimation for Case-Control Data

- GWAS of common variants where each variant is test separately
  - $\alpha=5 \times 10^{-8}$ (Bonferroni Correction for testing 1,000,000 variant sites)
  - Shown to be a good approximation for the effective number of tests
    - Valid even when more than 1,000,000 variant sites tested
  - Effective number of tests is dependent of the linkage disequilibrium (LD) structure
- Single variant tests using whole genome sequence data
  - Many more rare variants than common variants
    - Lower levels of LD between rare variants than between common variants
  - The number of effective tests for rare variants is higher than for analysis limited to common variants
  - α is yet to be determined for association analysis of whole genome sequence data

An Example of Determining Genome-wide Significance Levels for Common Variants

- Using genotypes from the Wellcome Trust Case-Control Consortium
- Dudbridge and Gusnato, Genet Epidemiol 2008
- Estimated a genome-wide significance threshold for the UK European population
- By sub-sampling genotypes at increasing densities and using permutation to estimate the nominal p-value for a 5% family-wise error
- Then extrapolating to infinite density
- The genome wide significance threshold estimate ~7.2×10^{-8}
- Estimate is based on LD structure for Europeans
  - Not sufficiently stringent for populations of African Ancestry

Power and Sample Size Estimation for Aggregate Rare Variant Tests

- For gene-based rare variant aggregate methods a Bonferroni correction for the number of genes/regions tested is used
  - e.g., 20,000 genes significance level $\alpha=2.5 \times 10^{-6}$
    - Can use a less stringent criteria
      - Not all genes have two or more variants
        - » Divide 0.05 by number of genes tested
  - Very low levels of LD between variants in separate genes
    - Therefore, a Bonferroni correction is not overly stringent
    - The number of tests ≅ effective number tests
      - This would not be the case for variants in LD
  - If units other than genes (transcripts) or additional regions are used
    - A more stringent criteria is necessary
      - If there is no (or little correlations) between the regions it is not overly stringent to use a Bonferroni correction

Power and Sample Size Estimation for Replication Studies

- For replication studies can base the significance level (α)
- On the number of genes/variants being brought from the discovery (stage I) study
- To replication (stage II)
- For example, if it is hypothesized that 20 genes and 80 independent variants will be brought to stage II (replication)
  - A Bonferroni correct can be made for performing 100 tests
    - An $\alpha = 5.0 \times 10^{-3}$ can be used for a family wise error rate of 0.05

Estimating Power/Sample Sizes For Single Variant Tests

- Can be obtained analytically
- Information necessary
  - Prevalence
  - Risk allele frequency
  - Effect size (odds ratio-for case control data)
  - Genetic model for the susceptibility variant
    - Recessive ($\gamma_1=1$)
    - Dominant ($\gamma_2=\gamma_1$)
    - Additive ($\gamma_2=2\gamma_1-1$)
    - Multiplicative ($\gamma_2=\gamma_1^2$)

Estimating Power/Sample Sizes For Individual Variants

- Usually, information on disease prevalence is known from epidemiological data
- A range of risk allele allele frequencies and effect sizes are used
- A variety of genetic models can also used
  - Dominant
  - Additive
  - Multiplicative

Armitage Trend Test

- Power and Sample size
  - Calculated under different models
    - Where γ is the relative risk
      - Multiplicative
        - » $\gamma_2=\gamma_1^2$
      - Additive
        - » $\gamma_2=2\gamma_1-1$
      - Dominant
        - » $\gamma_2=\gamma_1$
      - Recessive
        - » $\gamma_1=1$

Gamma is the Relative Risk not the Odd Ratio

- Most software for power calculations/sample size estimation use the relative risk (ɣ) and not the odds ratio
- The relative risk only approximates the odds ratio when disease is rare (Prevalence ~< 0.1%)
  - The relative risk is not appropriate for common traits when a case-control design is used

Correspondence Between the Odds Ratio and Relative Risk

Dominant Model

Disease Prevalence	1/2* RR=1.5	2/2** RR=1.5
0.01	1.51	1.51
0.10	1.59	1.59
0.20	1.71	1.71

Multiplicative Model

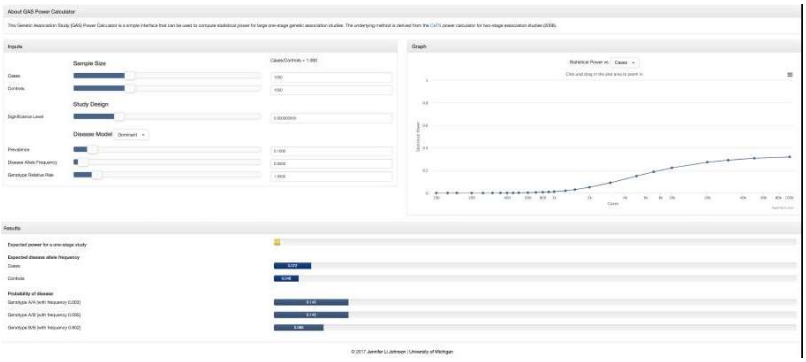
Disease Prevalence	1/2 RR=1.5	2/2 RR=2.25
0.01	1.51	2.28
0.10	1.59	2.61
0.20	1.71	3.25

Marker minor allele and disease allele frequency 0.01
D' and r^2=1
*1/2 genotype – heterozygous (one copy of the alternative allele)
**2/2 genotype - homozygous for the alternative allele

Genetic Association Study (GAS) Power Calculator

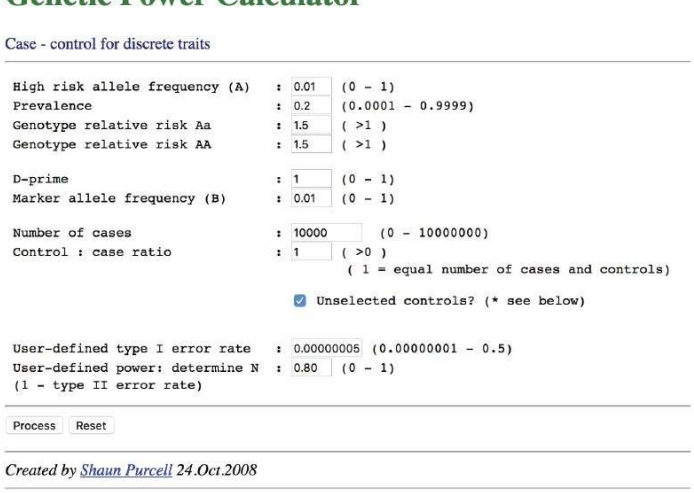
- http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/index.html
- A one-stage study power calculator
  - Which was derived from CaTs
    - Which is to perform two-stage genome wide association studies
      - Skol et al. 2006
- Cochran Armitage Trend Test
- Displays graphs of the results

GAS Power Calculator



Genetic Power Calculator

- http://zzz.bwh.harvard.edu/gpc/
- S Purcell & P Sham
- Uses the methods described in Sham PC et al. (2000)
  - VC QTL linkage for sibships
  - VC QTL association for sibships
  - VC QTL linkage for sibships conditional on the trait
  - TDT for discrete traits
  - Case-Control for discrete traits
  - TDT for quantitative traits
  - Case-Control quantitative traits
- Although input is the relative risk
  - Displays odds ratios



Power Association With Errors (PAWE)

- http://compgen.rutgers.edu/pawe/
  - Gordon et al. 2002, 2003
- Implements the linear trend test
- Four different error models can be used
  - See online documentation for complete explanation
- Can either perform:
  - Power calculations for a fixed sample size
  - Sample size calculations for a fixed power
- The genotype frequencies can be generated either using a:
  - Genetic model free method or
  - Genetic model-based method

Quanto

- Provides sample size and power calculations for
- Genetic and environmental main effects
- Interactions
  - Gene x gene
  - Gene x environment
- Sample & power calculations can be carried for:
  - Case-control
    - Unmatched
    - Matched
  - Case-sibling
  - Case-parent (trios)
    - Quantitative
    - Qualitative
  - Independent sample of individuals
    - Quantitative traits
      - Assumption sampled from a random population
- Can only be run under windows
  - https://quanto.software.informer.com/download/

Linkage Disequilibrium (LD)

- Power will be reduced if causal variant is not in perfect LD (r^2=1) with the tag SNP
- Can adjust sample size when r^2 <1 to increase power to the same level as when r^2=1
- Can estimate sample size when $r^2 \neq 1$
  - $N/r^2=N'$
  - Valid only for multiplicative model
  - (Pritchard and Przeworski, 2001)
- Power calculations almost always assume that r^2=1
- For whole genome sequence data this should be the true
  - since the causal variant should be included in the data

Power Analysis for Rare Variant Aggregate Association Tests

- Many unknown parameters must be modeled
  - Allelic architecture within a genetic region
    - Varied across genes and populations
  - Effects of variants within a region
    - Fixed or varied effect sizes of causal variants
    - Bidirectional effect of variants
    - Proportion of non-causal variants
- Power estimated empirically
- Simplified assumptions can be made to obtain analytical estimates
  - All variants have the same effect size
  - No non-causal variants within a region that is analyzed in aggregate

Simplistic Analytical Power Calculation for Rare-variant Aggregate Association Analysis

- Assumption
  - All rare variants are causal and have the same effect size
- Although usual not be correct
  - Provides a gestalt of the power for a given samples or sample size for a given power
- Use aggregate of allele frequencies
  - For example, assume a cumulative allele frequency of 0.025
  - Use an exome-wide significant level e.g., 2.5×10^{-6}
- Provide disease prevalence and penetrance model
- Perform calculations in the same manner as described for single variants

SKAT Power Calculator

- R Library
- Provides a haplotype matrix
  - 10,000 haplotypes over 200kb region
  - Simulated using a calibrated coalescent model (cosi)
  - Mimicking linkage disequilibrium structure of European ancestry
  - User can also provide haplotype data
- Power and sample size calculations for binary and quantitative traits
- User specify proportion of variants that increase or lower risk

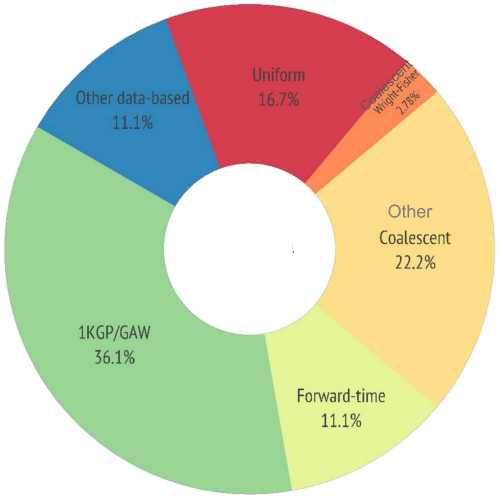
Empirical Power Calculations

- A variety of methods can be used to generate variant data to empirically estimate power
- Variant data is generated
  - Based upon a penetrance model samples of cases and controls are generated
  - Or a quantitative trait is generated based upon the genetic variance
- Multiple replicates are generated and analyzed
  - To determine the power

Empirical Power Calculations

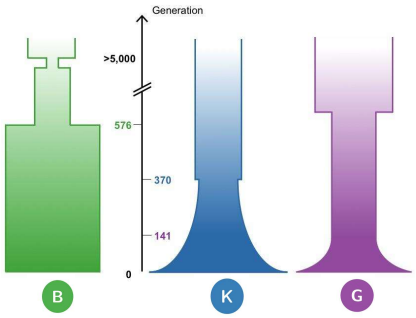
- Examples
  - 5,000 replicates are generated each with 20,000 cases and 20,000 controls
    - The power is the proportion of replicates with p-value less than the specified threshold, e.g., $5x10^{-8}$
  - For rare-variant aggregate tests all autosomal genes are generated and those genes with more than two rare variants (e.g., predicted loss of function) are analyzed
    - The power is the proportion of genes that were tested with p-value which is below a specified threshold, e.g., $2.5x10^{-6}$

Simulation Methods



Note: Not all methods give a realistic distribution of variants & in particular for rare variants

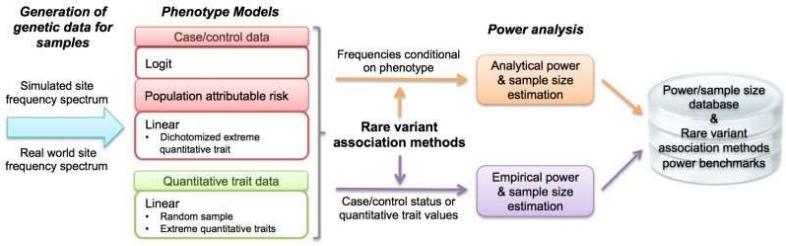
Generating Exome Sequence Data Sets Forward-time Simulation

Data	Haplotype Counts	Demographics
Boyko	105,814*	
Kyrukov	1,800,000*	
Gazave	1,308,000*	

*Selection coefficients used to define "variant type"
-"Missense" (1.0×10^{-5} – 1.8×10^{-2})
-"Nonsense, splice site and frameshift" (>1.8×10^{-2})

SEQPower

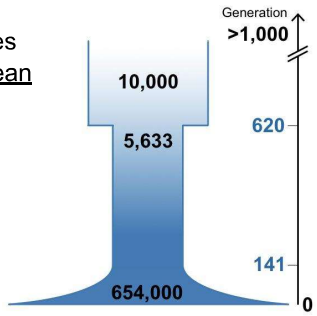
http://www.bioinformatics.org/spower/



Wang et al. 2014 Bioinformatics

Generating Variants: Using a European Demographic Model and Exome Sequence Data

- Variant data generated on 18,397 genes
- Variant data simulated using a European population demographic model
  - Gazave et al. 2013

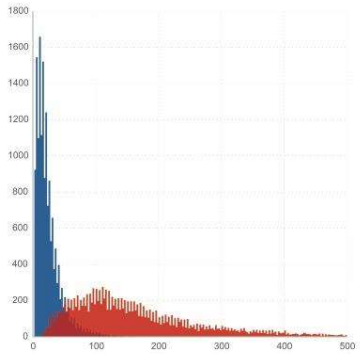


- Variants generated using exome sequence data
  - 4332 Exomes obtained from European American

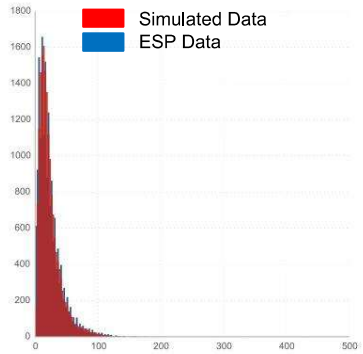
Which method performs better and why?

Does Generating Variant Data Using the European Population Demographic Model Perform Well?

Distribution of number of variants per gene

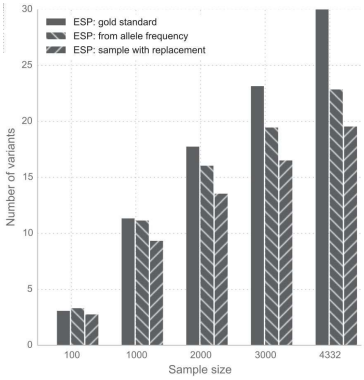


- Simulated variant counts based on the entire simulated population

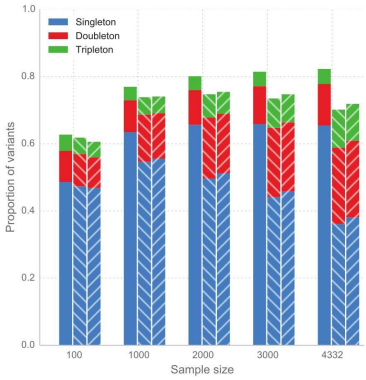


- Simulated variant counts based on haplotype pool down-sampled to ESP size

Simulating Data Using Sequence Data (ESP)

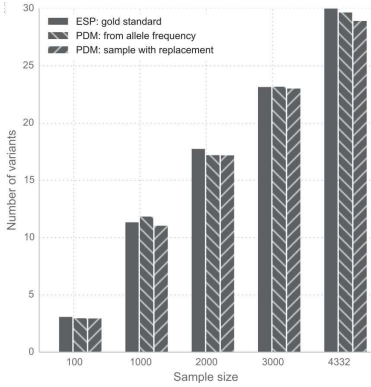


Number of Variant Sites

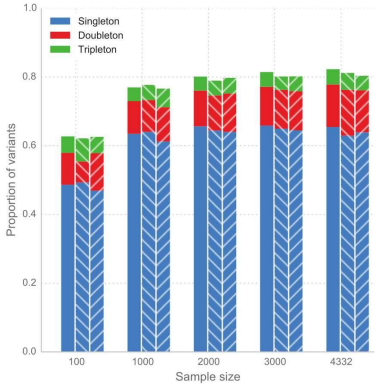


Proportion of Variant Sites that are Singletons, Doubletons and Tripletons

Simulating Data: Using Population Demographic Models (PDM)



Number of Variant Sites



Proportion of Variant Sites that are Singletons, Doubletons and Tripletons

Simulation Studies to Evaluate Power for Rare Variant Association Studies

- It is unknown which genes are important in disease etiology
  - Correct allelic architecture is unknown
- Can get a better understanding of power to detect associations by generating variants for the entire exome
- Use a variety of disease models
  - Odds ratios
  - Proportion of pathogenic variants
- Analyze those genes\regions with more than two variant sites
  - e.g., those with 2 or more variant sites
- Determine power as the proportion of genes that meet exome-wide significance (e.g., $\alpha=2.5x10^{-6}$)
  - If addition regions besides genes are analyzed
    - A more stringent α value should be used

Power Analysis

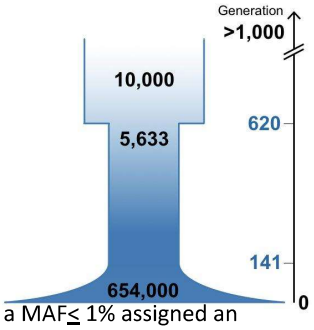
- For tests of individual variants
  - Power depended on sample size, disease prevalence, minor allele frequency, genetic model and variant effect size
- For rare variants (aggregate association tests)
  - Also dependent on the allelic architecture
    - Cumulative variant frequency within analyzed region
    - Proportion of causal variants
      - How much contamination from non-causal variants
    - Effect sizes the same the same or different across gene regions
      - Effects of variants in the same or different directions
        - » Protective and detrimental for binary traits
        - » Increase and decrease quantitative trait values

Power Analysis Rare Variants (Aggregate Association Tests)

- Power will not only vary between traits greatly
- The power to detect an association will also vary drastically between genes for the same complex trait
  - For some causal genes even with hundreds of thousands of samples power will be low
  - While for other causal genes a few thousand samples may be sufficient

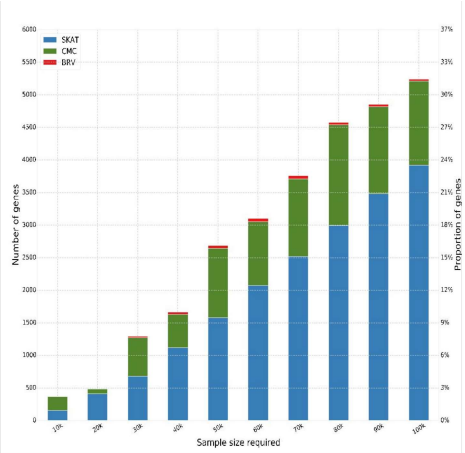
How Large of a Sample Size is Necessary to Detect Rare Variant Associations?

- Data generated on 18,397 genes
- Variant data simulated using a European population demographic model
  - Gazave et al. 2013



- Every missense, nonsense and splice with a MAF≤ 1% assigned an odds ratio of 1.5
- Sample sizes to detect X number of genes determined for
  - $\alpha = 2.5 \times 10^{-6}$
  - power=0.8

Sample Sizes Necessary to Detect an Association (Case-Control Data)



Imputing and Analyzing Imputed Genotype data

Suzanne M. Leal, Ph.D.
sml3@cumc.columbia.edu

Motivation for Imputation of Genotype Data

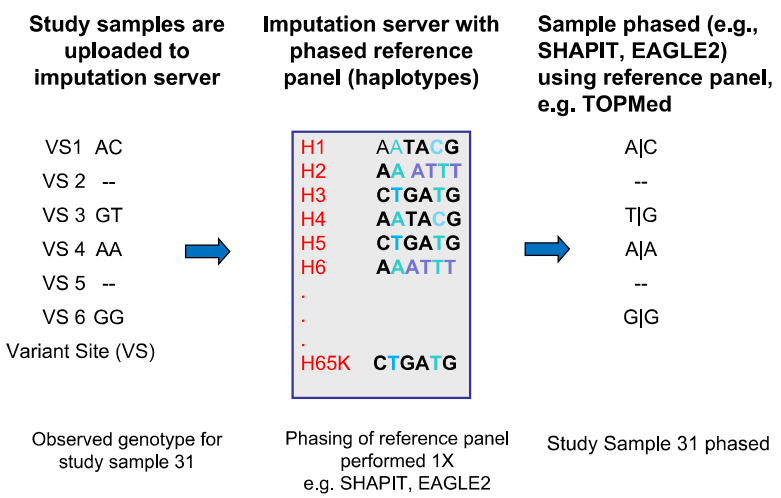
- Obtain genotypes for variant sites that are not genotyped
  - Additional variants can be tested for associations
    - Providing additional power to tag causal variant sites
    - Potential inclusion of causal variants that are unavailable on genotyping arrays
      - Aids in fine-mapping
- Considerably less expensive than generating whole genome sequence data
  - Does come at a cost of accuracy
    - In particularly for very rare variants
  - Imputed data will be available for rare variants if
    - For a variant site the alternative allele has to be observed at least ~8X in the reference panel in order for it to be imputed

Imputation of Variants

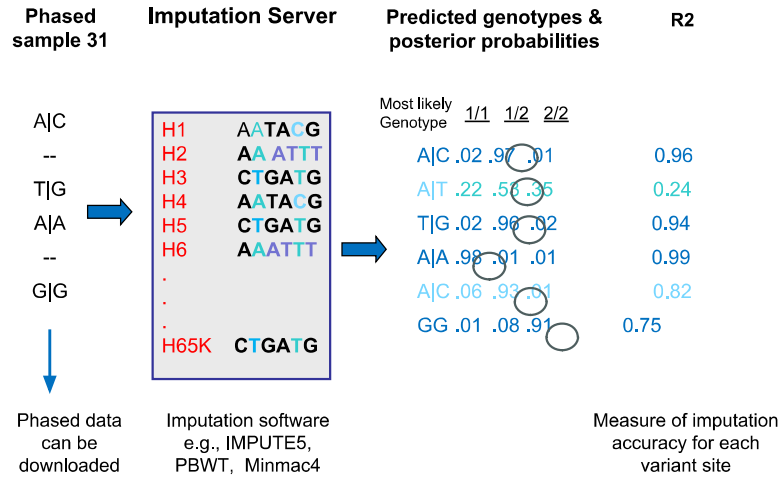
- Can be performed locally or on an imputation server
- Imputation locally has it limitation due to availability of a references panel
  - Internal data
  - 1000 genomes
  - Haplotype reference consortium (HRC)
    - Only part of this dataset is made publicly available
- Smaller imputation panels will impact the ability to impute lower frequency and rare variants
- Additionally, smaller sample sizes and less ancestral diversity will lead to a decrease in the imputation accuracy

Phasing and performing imputation using an Imputation Server

Imputation Step 1 Phasing



Step 2 Imputation



Measures of Imputation Accuracy

- R^2/INFO
  - Measures of imputation accuracy
    - Most programs report R2
      - Impute provides INFO scores
- r^2 is the correlation between the dosage and genotype obtained from sequence or genotype array data
  - Must have imputed data and sequence or genotype array data for the same person to estimate r^2

Step 3 Analysis of Imputed Data

- Variants are filtered according to R^2 values
  - e.g., analyze variants with an R^2>0.8
- Most likely genotypes are not analyzed instead dosages are analyzed
- The dosage can be estimated as follows for variant site 1 sample 12: A|C with prior probabilities 1/1= 0.02, 1/2=0.97, & 2/2=0.01 (R^2=0.96)

Genotype 1/1	0 x 0.02 = 0.0
Genotype 1/2	1 x 0.97 = 0.97
Genotype 2/2	2 x 0.01 = 0.02
Dosage	0.99

- The dosage for variant site 2 sample 12: A|T with prior probabilities 1/1= 0.22, 1/2=0.53, & 2/2=0.35 (R^2=0.23)

Genotype 1/1	0 x 0.22 = 0.0
Genotype 1/2	1 x 0.53 = 0.53
Genotype 2/2	2 x 0.35 = 0.70
Dosage	1.23

Imputation Panels

- 1000 Genomes Phase 3*
  - 2,504 unrelated reference samples
    - 26 populations from Africa, the Americas, Europe, East Asia, & South Asia
- African Genome Resource
- Asthma among African-ancestry Populations in the Americas (CAAPA)
- Genome Asia Pilot (GAsP)
- HAPMAP2
- Haplotype Reference Consortium (HRC) *
  - 32,470 references samples (39,635,008 variants)
    - Predominately European Ancestry

*Commonly used imputation panels

Imputation Reference Panels

- Multi-ethnic HLA
- Southeast Asian Reference Database (SEAD)
- The Trans-Omics for Precision Medicine (TOPMed)*
  - Version R3 133,597 reference samples (445,600,184 variants)
    - ~40% European, ~29% African/African American, ~19% Hispanic/Latino, ~8% Asian, & ~4% other/unknown)
- UK10K
- Westlake Biobank for Chinese (WBBC)

*Commonly used imputation panels

Imputation Servers

- Michigan (US)
  - Reference panels include, HRC, 1,000 Genomes, etc.
  - Phasing EAGLE2
  - Imputation Minmax4
  - https://imputationserver.sph.umich.edu/index.html#!
- NHLBI (US)
  - Reference panel TOPMed
  - Phasing EAGLE2
  - Imputation Minmax4
  - https://imputation.biodatacatalyst.nhlbi.nih.gov/#!

Imputation Servers

- Sanger (UK)
  - Reference panels include HRC, 1,000 Genomes, etc.
  - Phasing SHAPEIT or EAGLE2
  - Imputation PBWT
  - https://www.sanger.ac.uk/tool/sanger-imputation-service/
- Westlake (People's Republic of China)
  - Reference panels include 1000 Genomes, GAsP, SEAD, & WBBC
  - Phasing SHAPEIT2
  - Imputation Minmax4
  - https://imputationserver.westlake.edu.cn/index.html

What Impacts Imputation Quality?

- Reference (imputation) panel
  - Sample size
    - Larger samples
      - Increase imputation accuracy
      - Ability to impute rare variants
  - Ancestry diversity
- Target sample
  - Density of markers
  - Genotype quality
  - Ancestry and representation on the imputation panel
  - The population's linkage disequilibrium structure

Note: Since each target sample is phased and imputed separately using the pre-phased imputation panel on the imputation server, sample size of the target sample does not impact imputation accuracy

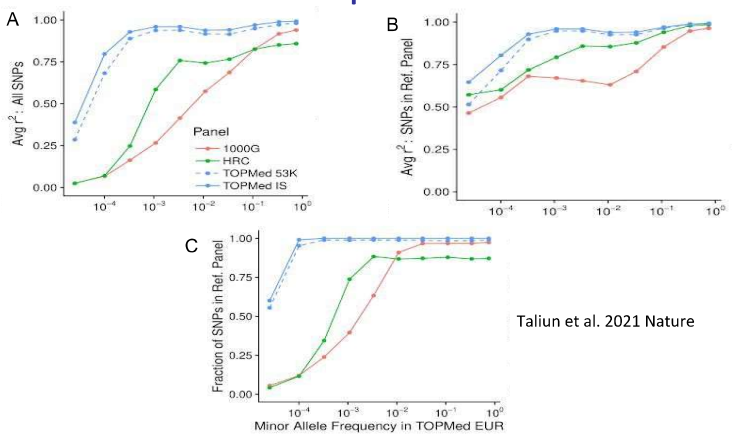
How Well do 1000 Genomes, HRC, and TOPMed Imputation Reference Panels Perform?

- Reference Panels
  - 1000 Genomes Phase 3
    - 2,504 reference samples
      - 26 populations from Africa, the Americas, Europe, East Asia, & South Asia
  - HRC v1.1 2016
    - 32,470 references samples (39,635,008 variants)
      - Predominately European Ancestry
  - TOPMed (Version r2)
    - 97,256 reference samples (308,107,085 variants)
      - Diverse population from the USA 48.49% European, 25.95% African/African American, 17.57% Hispanic/Latino/Admixed Americans, 1.22% East Asian, 0.66 South Asians, 6.11% other/unknown)
  - TOPMed (53K)
    - 53,831 reference samples

How Well do 1000 Genomes, HRC, and TOPMed Imputation Reference Panels Perform?

- Target Sample
  - 100 ancestry specific samples,
    - e.g. Europeans, African-Americans, & South Asians
  - Obtained from BioMe
    - Samples are not included in any of the reference panels

Europeans



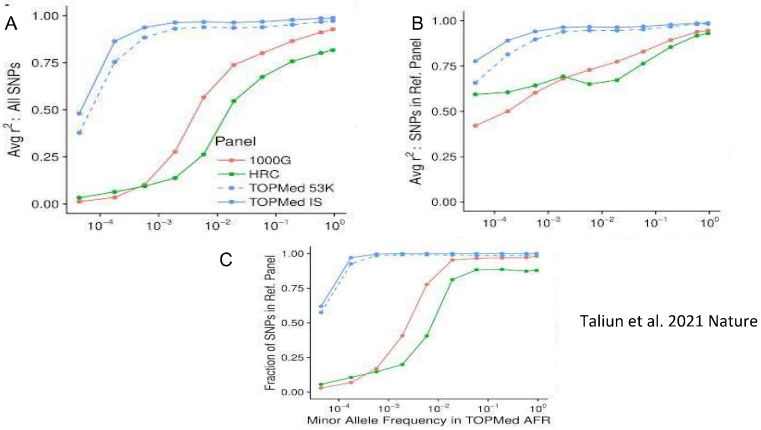
Taliun et al. 2021 Nature

Panel A: r^2 between the sequence-based genotypes and imputed dosages across all variants, assigning $r^2 = 0$ to variants absent from each reference panel

Panels B: average r2 with only the variants present from each reference panel

Panel C: The proportion of variants present in the reference panels

African Americans



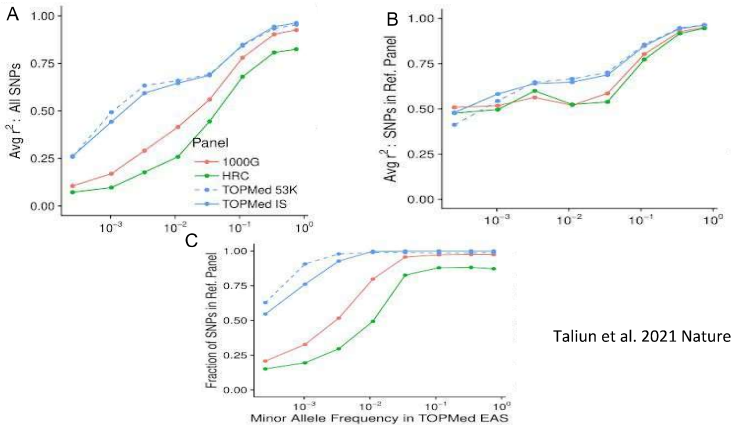
Taliun et al. 2021 Nature

Panel A: r^2 between the sequence-based genotypes and imputed dosages across all variants, assigning $r^2 = 0$ to variants absent from each reference panel

Panels B: Average r^2 with only the variants present from each reference panel

Panel C: The proportion of variants present in the reference panels

East Asians



Taliun et al. 2021 Nature

Panel A: r^2 between the sequence-based genotypes and imputed dosages across all variants, assigning $r^2 = 0$ to variants absent from each reference panel

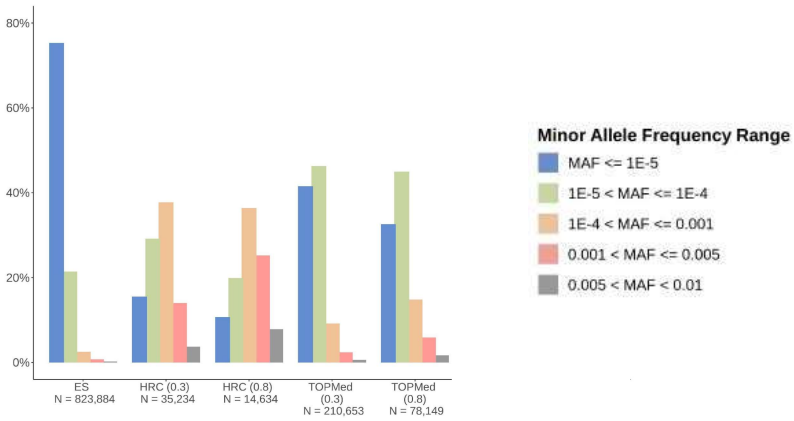
Panels B: Average r^2 with only the variants present from each reference panel

Panel C: The proportion of variants present in the reference panels

Comparison of Rare Variant Distributions

- Unrelated white European UK Biobank study participants (N=168,206) with
  - Release 2 exome sequence
  - Genotype array data available
- Imputed variants using both HRC and TOPMed (v2)
- Comparison of variant distributions
  - Exome sequence (ES) data
  - HRC imputed data r2>0.3 and r2>0.8
  - TOPMed imputed data r2>0.3 and r2>0.8

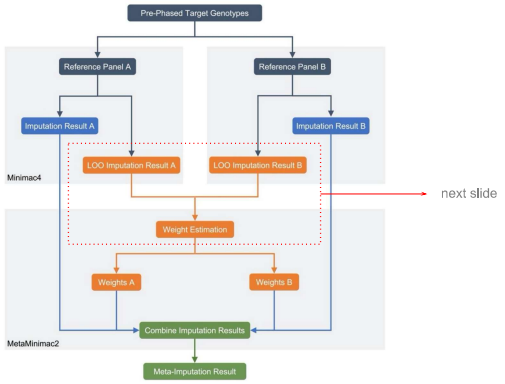
Distribution of Rare Variants



Variants for chromosomes 1 and 2 in coding regions

Meta-imputation (I)

Use, in turn, two or more reference panels*, then combine the results



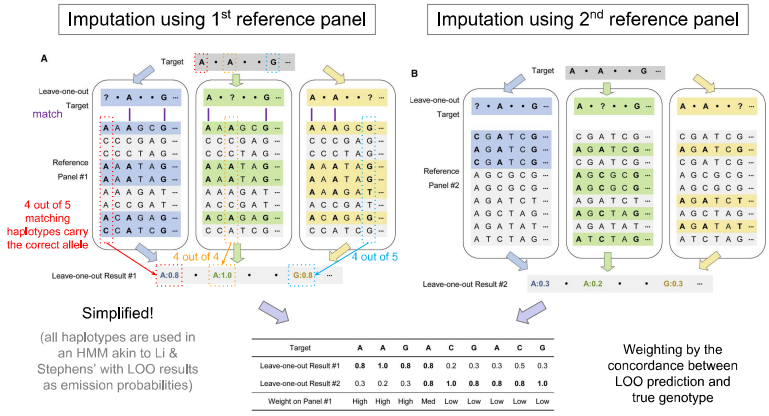
Requires pre-phased data

*The reference panels must use the same genome build

Yu et al. (2022) Am J Hum Genet

Meta-imputation (II)

Obtain region-specific weights via leave-one-out (LOO) in an HMM



Yu et al. (2022) Am J Hum Genet

Imputation of Variants without using an Imputation Server

- Imputation locally has it limitation due to availability of references panels
  - Internal data
  - 1000 genomes
  - HRC
    - Only part of this dataset is made publicly available to download to use locally
- Can be computationally intensive to phase and impute genotypes locally
- All haplotype phasing and imputation software used on imputation servers are publicly available
- Due to data sharing limitation in particular within the European Union
  - It may not be possible to use imputation servers which are located in the US, UK, or China

Commonly Used Phasing Software

- BEAGLE v5.4*
  - Browning & Browning 2007
- SHAPE-IT 4
  - Delaneau et al. 2012
- EAGLE2
  - Loh et al. 2016)

*Can also perform imputation

Commonly Used Imputation Software

- IMPUTE5
  - Rubinacci et al. 2020
- PBWT – Positional Burrows Wheeler Transform
  - Rubinacci et al. 2020
- Minmac4
  - Das et al. 2016
- GLIMPSE2*
  - Rubinacci et al. 2022

*For imputing into low coverage sequence data, e.g., sequence data obtained from ancient DNA

Using Imputation to Detect Genotyping Errors

- Can provide information on genotyping error by comparing the genotype of the imputed variant with genotypes obtained from array or sequence data
  - Would suggest there is genotype error if for the imputed data the R^2 (measure of imputation accuracy) is high
    - But the r^2 (correlation) between the imputed variant and the genotypes obtained from sequence or array data is low.
  - Association analysis results obtained for the imputed variant and the same variant obtained from genotyping array or sequencing vary greatly even though the R2 value is high for the imputed variant
    - Suggest that there is probably genotyping error for the variant obtained from genotyping array or sequence data
- The variant obtained from array or sequence data can be replaced with the imputed variant

Combining data obtained from different genotyping arrays

- Variants that don't overlap between arrays can be imputed
  - As well as variants not available on any of the arrays
- Caution should be used because the imputation quality can vary between datasets
  - Influenced by different error rates between datasets
  - Principal components analysis (PCA) can be used to determine if the potential problems
    - If additional quality control is necessary
- If there are more cases or controls for a particular dataset
  - Type I errors can be increased

COLUMBIA

#StatFunGen

Gene mapping in complex trait functional genomics

From single markers to systems biology

Gao Wang, PhD, Assistant Professor of Neurological Sciences

January 29, 2025

1. Overview of post-GWAS analysis
2. Statistical fine-mapping in GWAS and QTL studies
3. Exercise: fine-mapping
4. Overview of molecular quantitative trait loci (QTL) studies
5. QTL-GWAS loci: multi-trait analysis and colocalization
6. QTL-GWAS genes: transcriptome-wide association studies
7. Exercise: multivariate fine-mapping and TWAS
8. Connections: fine-mapping, colocalization, TWAS and MR

Overview of post-GWAS analysis

Follow-up questions on GWAS

Statistical

- How many independent association signals exist?
- Which variants are likely causal vs. tagging?

Biological

- Which genes mediate these associations?
- Through which biological pathways?
- In which tissues/cell-types do these variants act?

Translational

- Can we identify promising drug targets?
- Can we predict disease risk or trait values?
- How can we prioritize follow-up experiments?

Sources of association signals

Causal association — meaningful

- Tested genetic variations influence traits directly

Linkage disequilibrium (LD) — useful

- Tested genetic variations associated with other nearby variations that influence traits
- Meaningful or misleading, in different contexts

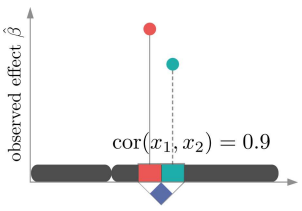
Population stratification — misleading

- Tested genetic variations is unrelated to traits, but is associated due to sampling differences
- eg, minor allele frequency, disease prevalence

Impact of LD on GWAS analysis

Oligogenic: trait influenced by a few genetic variants

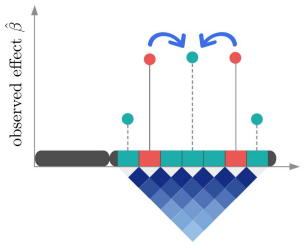
- Misleading: difficult to identify causal variants
- Useful: 'tag SNPs' in array based GWAS design



Impact of LD on GWAS analysis

Polygenic: trait influenced by numerous genetic variants

- Misleading: spurious strong association due to LD 'buddies'
- Useful: whole-genome prediction with sparse models



Post-GWAS analysis toolkit

Variant-level analysis

- Fine-mapping: identify likely causal variants
- Colocalization: shared genetic effects
- Mendelian randomization: causal inference

Aggregated analysis

- LD score regression: confounding, heritability, functional enrichment and genetic correlation
- Gene-based tests and pathway enrichment
- Transcriptome-wide association (TWAS)
- Polygenic prediction

Key challenge: Connecting statistical associations to biological mechanisms

Kontou and Bagos *BioData Mining* (2024) 17:31
https://doi.org/10.1186/s13040-024-00385-x

BioData Mining

RESEARCH Open Access

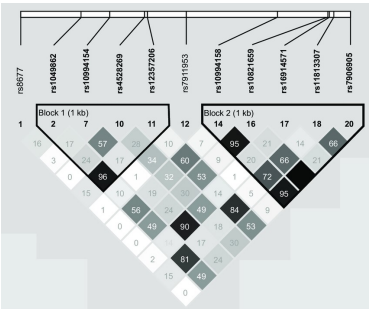
The goldmine of GWAS summary statistics: a systematic review of methods and tools

Panagiota I. Kontou[1] and Pantelis G. Bagos[2*]

Figure: "Our review identified a total of 305 functioning software tools and databases dedicated to GWAS summary statistics, each with unique strengths and limitations."

LD and LD score regression

How many independent associations?

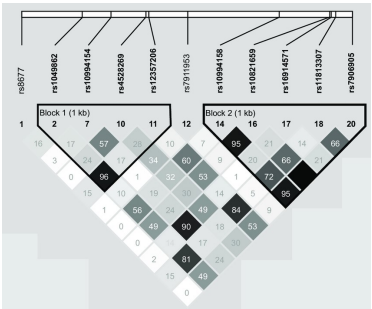


LD clumping

- Select lead SNP by p-value
- Remove SNPs in LD given fixed threshold (*e.g.* $r^2 > 0.1$)
- Within fixed window (*e.g.* 1Mb)
- Implementation: PLINK, bigsnpr

Note: LD pruning can be implemented as LD clumping on MAF.

How many independent associations?

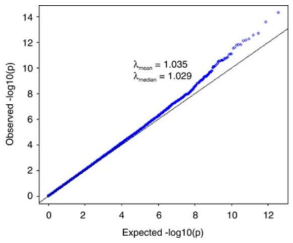


Limitations

- **Population-specific:** Window size and LD patterns vary by ancestry
- **Strong effects bias:**
  - Even very low LD SNPs near well-powered loci can show high significance
  - Example: APOE4 allele in Alzheimer's disease GWAS drives signals across chromosome 19

A second thought on genomic inflation

Population stratification? Or, polygenic inheritance + LD?



Suggested reading: Yang et al (2011) EJHG

LD score regression

Initial motivation: distinguish polygenicity from confounding

- Under pure genetic drift, LD is uncorrelated to magnitude of allele frequency differences between populations

LD score regression

Initial motivation: distinguish polygenicity from confounding

- Under pure genetic drift, LD is uncorrelated to magnitude of allele frequency differences between populations

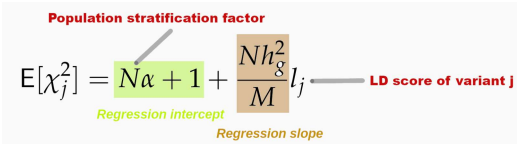
Assuming each SNP explains the same amount of trait variance,

$$E[\chi_j^2|\ell_j] = 1 + Na + \frac{Nh^2}{M}\ell_j$$

- **LD score of SNP j:** $\ell_j = \sum_{k=1}^{M} r_{jk}^2$
- N is the GWAS sample size
- $\frac{h^2}{M}$ is the average heritability explained per SNP
- $a \neq 0$ indicates **confounding** (*e.g.* population stratification).

LD score regression (LDSC)

Separating h_g^2 and population stratification



A more powerful and accurate correction factor for GWAS summary statistics compared to genomic control approach.

- Bulik-Sullivan et al (2015) Nature Genetics — the LDSC regression paper
- Zhu and Stephens (2017) AoAS — a neat, alternative LDSC regression model derivation in supplemental material

LDSC application: heritability estimation

Narrow sense heritibility

- Proportion of phenotypic variation explained by additive genetic factors

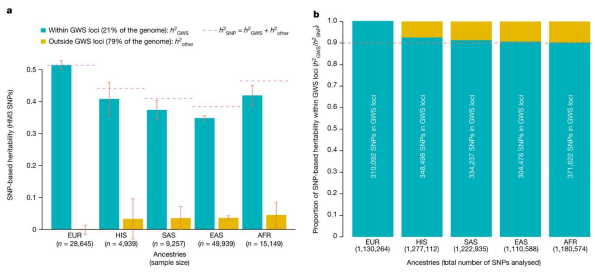
Estimation strategy

- Pedigree design: genetic covariance and IBD sharing
- Population design: linear mixed models

Population design, summary statistics

- LDSC to estimate SNP-based heritability
- Stratified LDSC (S-LDSC) to partition heritability by functional annotations

Variance of height explained in GWAS



Yengo *et al.* (2022) Nature

Functional annotation and enrichment analysis

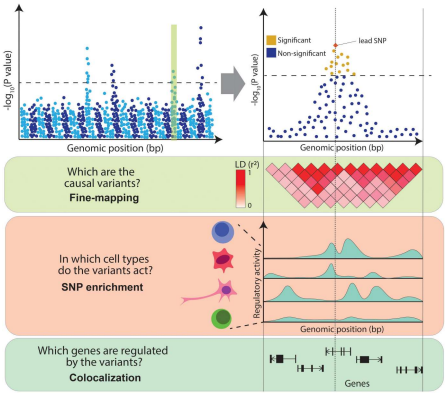
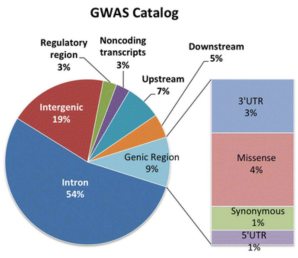


Figure: Cano-Gamez and Trynka (2020) Front. Genet.

GWAS variants catelog by functional annotations

Most GWAS associations are non-coding



Lee *et al.* (2018) *Human Genetics*

Enrichment of functional annotations in GWAS

Question: where in the genome is heritability concentrated?

- Integrate directly as range based binary annotations
  - Finucane et al (2015) Nature Genetics — Stratified LDSC paper
- Extension: variant specific continuous annotations
  - Gazal et al (2017) Nature Genetics
- Tissue specific variant level annotations independent of GWAS results
  - Deep Learning methods
  - Zhou et al (2015) Nature Genetics, Zhou et al (2018) Nature Genetics, Lai *et al.* (2022) PLoS Comp Bio
  - Avsec *et al.* (2021) Nature Methods

A Polygenic Model: Partitioned Heritability

Stratified LD Score Regression assumes heritability varies linearly with functional annotations:

$$E[\chi_j^2|\ell_{jA_1},\ldots,\ell_{jA_K}] = 1 + Na + N\sum_{k=1}^{K}\tau_k\ell_{jA_k}$$

- $\tau_k = \frac{h^2_{gk}}{M_k}$ quantifies heritability per SNP within annotation A_k.
- $\ell_{jA_k} = \sum w_k r^2_{jk}$ is weighted LD score for SNP j based on SNPs within annotation A_k.
- w_k can be binary (e.g., 0/1) or continuous (e.g., functional relevance scores).
- **Enrichment:** relative contribution of heritability explained by annotation A_k, $\frac{\tau_k}{\frac{h^2_g}{M}}$.

S-LDSC is a genome-wide approach

- A single locus may not have enough power to leverage annotation enrichment
- Genome-wide evaluation of thousands of annotations can increase power of fine-mapping
  - Lead to new loci to discover
- Functional enrichment can be done under the same framework as heritability estimation
  - Prioritize genomic features / tissues / cell-types
- **Enrichment coefficient may be transferrable cross population**
  - Weissbrod *et al.* (2022) Nat. Genet.

Cell-type enrichment in GWAS traits via S-LDSC

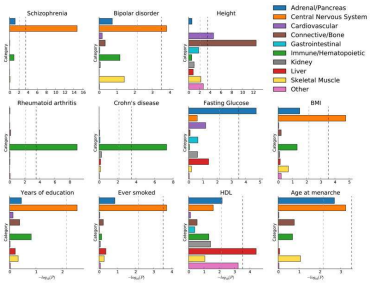


Figure: Finucane *et al.* (2015) Nature Genetics

A sparse model (a somewhat oligogenic view)

Generalized linear model for SNP effects given K annotations

$$\beta_j = (1-\pi_j)\delta_0 + \pi_j g(\Theta)$$
$$\pi_j := \Pr(\gamma_j = 1 | \boldsymbol{\alpha}, \boldsymbol{d})$$
$$\log\left[\frac{\pi_j}{1-\pi_j}\right] = \alpha_0 + \sum_{k=1}^{K} \alpha_k d_{kj}$$

α are **log fold enrichment** of functional genomic features

- Suggested reading: Wen (2016) AoAS

Enrichment of DNase I in GTEx eQTLs

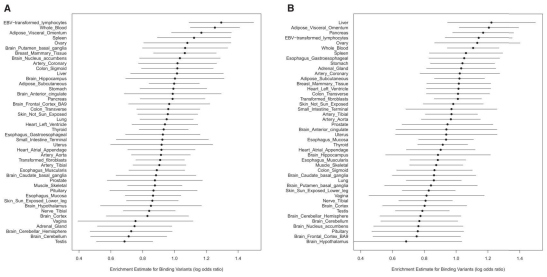


Figure: Wen *et al.* (2016) AJHG

Gene Sets / Pathway Databases

Gene Ontology (GO):

- Three principal categories: Biological Processes, Cellular Components and Molecular Functions.
- GO terms have hierarchical structure, from broad to narrow (e.g., metabolism → GTP biosynthesis).
- Each GO term is linked to a curated list of associated genes.

Gene Sets / Pathway Databases

Gene Ontology (GO):

- Three principal categories: Biological Processes, Cellular Components and Molecular Functions.
- GO terms have hierarchical structure, from broad to narrow (e.g., metabolism → GTP biosynthesis).
- Each GO term is linked to a curated list of associated genes.

KEGG:

- Focuses on metabolic, disease, and trait-related pathways.
- Smaller and more curated compared to GO, emphasizing enzymatic and biological networks compared to disease and trait related.

Over-Representation Analysis for Functional Enrichment

	In Pathway	Not in Pathway
Association Signal	a	b
No Association Signal	c	d

Statistical tests: Fisher's Exact Test, χ^2-test, Binomial proportion test.

Over-Representation Analysis for Functional Enrichment

	In Pathway	Not in Pathway
Association Signal	a	b
No Association Signal	c	d

Statistical tests: Fisher's Exact Test, χ^2-test, Binomial proportion test.

Jackknife standard error: $\text{SE}_{\text{jack}} = \sqrt{\frac{N-1}{N}\sum_{i=1}^{N}(\hat{\theta}_{-i} - \bar{\theta}_{(\cdot)})^2}$

- $\hat{\theta}_{-i}$ is estimate with ith chromosome removed, $\bar{\theta}_{(\cdot)}$ is mean of leave-one-out estimates, N is number of chromosomes
- A convenient way to account for dependent genetic variables (LD) and annotations (*e.g.* overlapping gene-sets)

Over-Representation Analysis for Functional Enrichment

	In Pathway	Not in Pathway
Association Signal	a	b
No Association Signal	c	d

Statistical tests: Fisher's Exact Test, χ^2-test, Binomial proportion test.

Remaining sources of bias:

- LD proxy / buddies (e.g., captured by LD score).
- Gene density (longer genes capture more SNPs).
- Distance to transcription start site (TSS).
- Minor allele frequency (MAF).

SNPs for ORA can be chosen to match on these factors to eliminate bias.

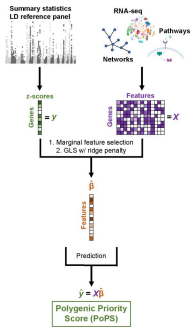
Other Enrichment Analysis Methods

Additional approaches:

- Gene Set Enrichment Analysis (GSEA)
- Stratified Linkage Disequilibrium Score Regression (sLDSC)

Polygenic Priority Score (PoPS):

- An extension to sLDSC and GSEA.
- Integrates polygenic signals across gene sets.
- **Reference:** Weeks et al. (2023) *Nature Genetics*.



Statistical fine-mapping in GWAS and QTL studies

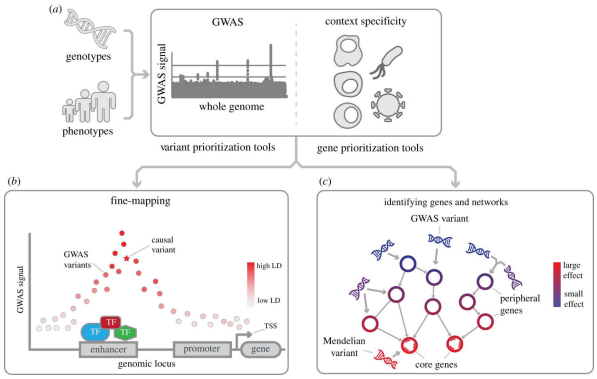
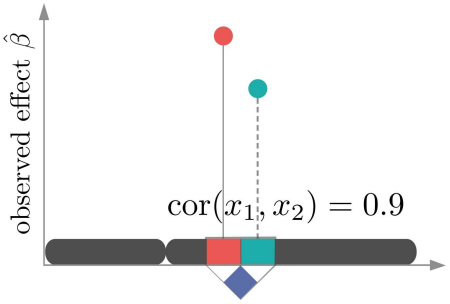


Figure: Broekema *et al.* (2020) Open Biol.

Correlated variables in association studies

Due to a phenomenon called **linkage disequilibrium** (LD)



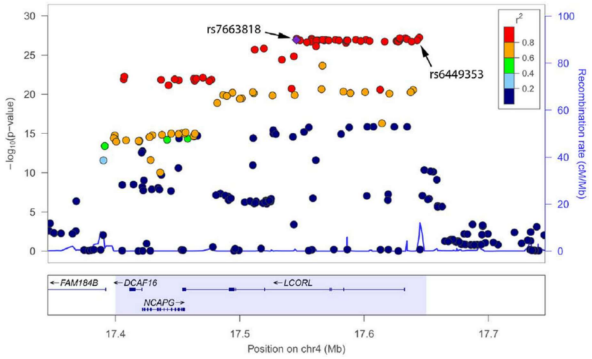


Figure: N'Diaye *et al.* (2011) PLoS Genet.

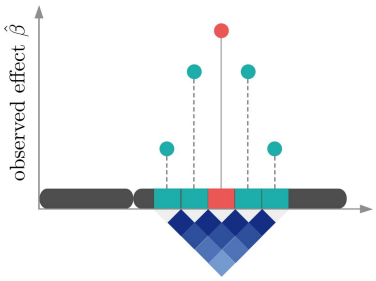
Objectives

Statistical fine-mapping aids in the identification of causal variants, in order to

- interpret association signals (pinpoint to genes)
- understand biological function of a variant
- elucidate genetic architecture of complex and molecular phenotypes

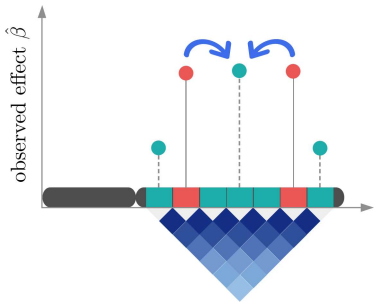
Identify non-zero effect ("causal") variables

Simply pick the top association in an LD block? Maybe?



Identify non-zero effect ("causal") variables

Simply pick the top association in an LD block? ... or not!



Architecture: sparse effects, polygenic background

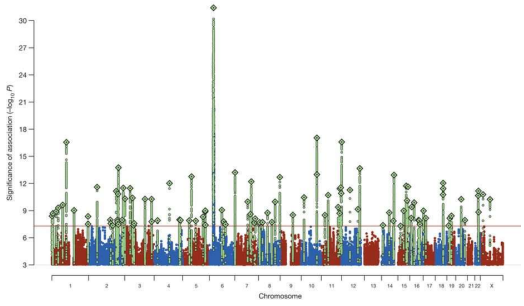


Figure: O'Donovan *et al.* (2014) Nature

Challenge: large biobank sample applications

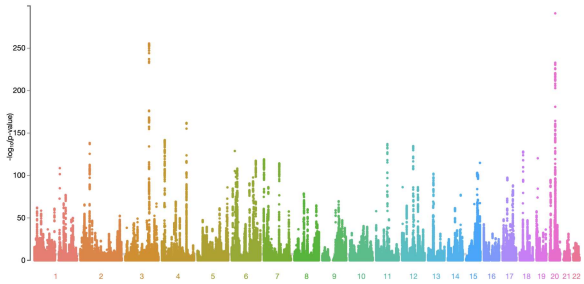


Figure: UK Biobank height GWAS, `http://nealelab.is/uk-biobank`

Toy example, concepts and conditional analysis

Single Causal variant model

Effect variable (red) correlated with non-effect variable (blue)

	Case		Control	
SNP	G1	G2	G1	G2
rs1	1200	800	1000	1000
rs2	1190	809	1000	1000

Single Causal variant model

Compute likelihood ratios (LR) H_1 vs H_0,

	Case		Control			
SNP	G1	G2	G1	G2	LR	p-value
rs1	1200	800	1000	1000	6.15×10^{8}	1.99×10^{-10}
rs2	1190	809	1000	1000	0.94×10^{8}	1.37×10^{-9}

Single Causal variant model

Compute likelihood ratios (LR) H_1 vs H_0,

	Case		Control			
SNP	G1	G2	G1	G2	LR	p-value
rs1	1200	800	1000	1000	6.15×10^{8}	1.99×10^{-10}
rs2	1190	809	1000	1000	0.94×10^{8}	1.37×10^{-9}

Probability of association assuming **one effect variable**,

$$\frac{\text{LR}_1}{\text{LR}_1 + \text{LR}_2} = \frac{6.15}{6.15 + 0.94} = 0.87 \quad \frac{\text{LR}_2}{\text{LR}_1 + \text{LR}_2} = \frac{0.94}{6.15 + 0.94} = 0.13$$

Per variable contingency table analysis, likelihood ratio test

For a 2×2 table with observed frequencies O_{ij} and expected frequencies E_{ij}:

$$G = 2 \sum_{i,j} O_{ij} \log \left(\frac{O_{ij}}{E_{ij}} \right)$$

Under H_0 (independence):

$$E_{ij} = \frac{(\text{row total}_i)(\text{column total}_j)}{n}$$

Likelihood Ratio:

$$LR = \exp(G/2)$$

The test statistic G follows a χ^2 distribution with 1 degree of freedom under H_0

Per variable contingency table analysis, R code

```r
# returns likelihood ratio and p-value of H_1 vs H_0
get_2x2_lrt = function(tbl) {
    tbl = as.table(matrix(tbl, 2,2,
      dimnames=list(status=c('case','control'),
      genotype=c('minor_allele','major_allele'))))
    test = MASS::loglm(~status+genotype,data=tbl)
    lr = exp(test$lrt / 2)
    pval = 1 - pchisq(test$lrt, df=1)
    return(list(lr=lr, pval=pval))
}
res1 = get_2x2_lrt(c(1200,800,1000,1000))
res2 = get_2x2_lrt(c(1190,809,1000,1000))
```

Single Causal variant model Bayesian variable selection

Compute Bayes factors (BF) H_1 vs H_0,

	Case		Control			
SNP	G1	G2	G1	G2	BF	$\log_{10}$BF
rs1	1200	800	1000	1000	2.40×10^7	7.38
rs2	1190	809	1000	1000	0.36×10^7	6.56

Single Causal variant model Bayesian variable selection

	Case		Control			
SNP	G1	G2	G1	G2	BF	$\log_{10}$BF
rs1	1200	800	1000	1000	2.40×10^7	7.38
rs2	1190	809	1000	1000	0.36×10^7	6.56

Posterior inclusion probability of association assuming **one effect variable**,

$$\text{PIP}_1 = \frac{\text{BF}_1}{\text{BF}_1 + \text{BF}_2} = \frac{2.40}{2.40 + 0.36} = 0.87 \quad \text{PIP}_2 = \frac{\text{BF}_2}{\text{BF}_1 + \text{BF}_2} = \frac{0.36}{2.40 + 0.36} = 0.13$$

Bayesian analysis for 2×2 tables, model specification

Data: $\mathbf{n} = (n_{11}, n_{12}, n_{21}, n_{22})$

Multinomial Likelihood:

$$\mathbf{n} \sim \text{Multinomial}(N, \boldsymbol{\theta})$$

$$P(\mathbf{n}|\boldsymbol{\theta}) = \frac{N!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \theta_{ij}^{n_{ij}}$$

Independence Model (H_0):

$\theta_{ij} = \pi_i \times \gamma_j$

(Requires row-column independence)

$\pi \sim \text{Dirichlet}(a, a); \gamma \sim \text{Dirichlet}(a, a)$

Association Model (H_1):

$\boldsymbol{\theta} \sim \text{Dirichlet}(a, a, a, a)$

(Allows for all possible associations)

Bayes Factor:

$$\text{BF} = \frac{P(\text{data}|H_1)}{P(\text{data}|H_0)}$$

$$= \frac{\int P(\mathbf{n}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|H_1) d\boldsymbol{\theta}}{\int P(\mathbf{n}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|H_0) d\boldsymbol{\theta}}$$

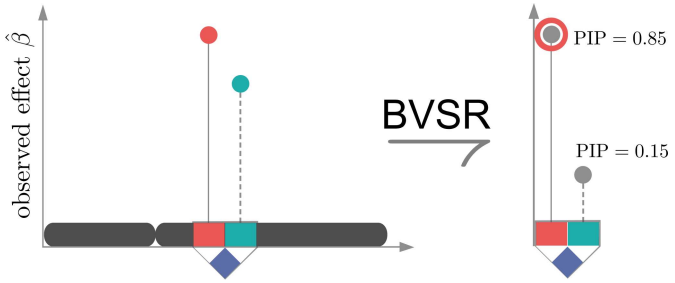
Common choice: $a = 1$ (uniform prior)

Bayesian contingency table analysis, R code

```
# returns Bayes factor of H_1 vs H_0
get_2x2_bf = function(tbl, prior_a = 1) {
    n = matrix(tbl, 2, 2, byrow=T)
    dimnames(n) = list(status=c('case','control'),
                       genotype=c('minor','major'))
    res = BayesFactor::contingencyTableBF(n, sampleType="indepMulti",
          fixedMargin="rows", priorConcentration=prior_a)
    bf = exp(res@bayesFactor$bf)
    return(list(bf = bf, log10bf = log(bf)/log(10)))
}
res1 = get_2x2_bf(c(1200,800,1000,1000))
res2 = get_2x2_bf(c(1190,809,1000,1000))
```

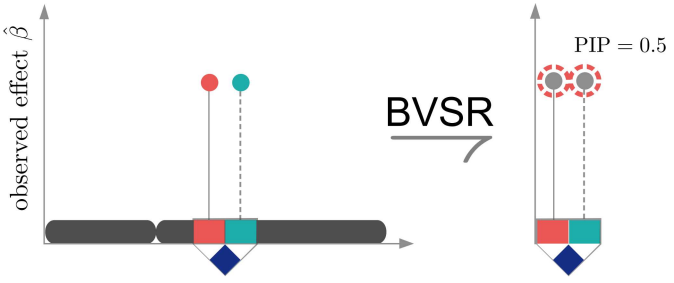

Bayesian variable selection: PIP

Computes **Posterior Inclusion Probability** (PIP)



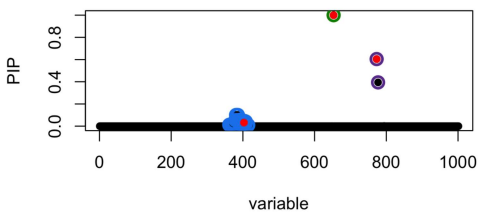
Bayesian variable selection: PIP

Computes **Posterior Inclusion Probability** (PIP)



Bayesian variable selection: Credible Sets

"Clusters" of signals to account for correlations between variables (eg LD)



Bayesian variable selection: Credible Sets

- 95% credible set S: Pr(*effect variable in* S) $\geq 95\%$
- *e.g.* , "Single effect" model:

$$\sum_{j \in S} PIP_{(j)} \geq 95\%$$

  where $PIP_{(j)}$'s are in descending order.
- Formal definition: Wang *et al.* (2020) J. R. Stat. Soc. B

Multiple causal effects: step-wise search

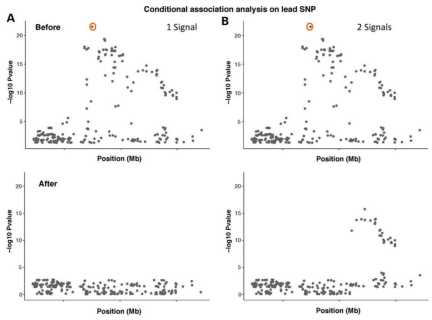


Figure: Spain and Barrett (2015) Hum. Mol. Genet.

A simple frequentist conditional analysis

Forward selection algorithm

1. For each SNP fit a simple linear regression model
2. Select the SNP j that has the largest model likelihood
3. Form residuals $\boldsymbol{y}' := \boldsymbol{y} - \boldsymbol{X}_j \hat{b}_j$, and repeat

A simple frequentist conditional analysis

Forward selection algorithm

1. For each SNP fit a simple linear regression model
2. Select the SNP j that has the largest model likelihood
3. Form residuals $\boldsymbol{y}' := \boldsymbol{y} - \boldsymbol{X}_j \hat{b}_j$, and repeat

A greedy algorithm to choose the "best" SNPs, but is incapable of capturing multiple SNPs in LD

(Approximate) COnditional and JOint Analysis

Method overview

- Frequentist framework for summary statistics
- Implements stepwise variable selection
- GCTA software implementation (Yang et al. 2011)

(Approximate) COnditional and JOint Analysis

Method overview

- Frequentist framework for summary statistics
- Implements stepwise variable selection
- GCTA software implementation (Yang et al. 2011)

Limitation

- May "over-regress" on lead variants when multiple causal variants exist in partial LD
- Example: if two SNPs in partial LD ($r^2 = 0.3$) are both causal:
  - First SNP captures most association
  - Second SNP effect underestimated after conditioning
- Result: Potentially missing independent functional variants

Need to sample from a distribution, not selecting the best signal

COJO algorithm, Yang et al. (2012) Nat. Genet.

1. Start with a model with the most significant SNP in the single-SNP meta-analysis across the whole genome with P value below a cutoff P value, such as 5×10^{-8}.
2. For the ith step, calculate the P values of all the remaining SNPs conditional on the SNP(s) that have already been selected in the model. To avoid problems due to collinearity, if the squared multiple correlation between a SNP to be tested and the selected SNP(s) is larger than a cutoff value, such as 0.9, the conditional P value for that SNP will be set to 1.
3. Select the SNP with minimum conditional P value that is lower than the cutoff P value. However, if adding the new SNP causes new collinearity problems between any of the selected SNPs and the others, we drop the new SNP and repeat this process.
4. Fit all the selected SNPs jointly in a model and drop the SNP with the largest P value that is greater than the cutoff P value.
5. Repeat (2), (3) and (4) until no SNPs can be added or removed from the model.

Quantify uncertainty: which to select among many correlated variables

Bayesian forward selection algorithm

1. For each SNP j, fit a simple Bayesian linear regression model to get Bayes Factor BF_j
2. Form weight for each SNP, $w_j \propto \text{BF}_j$
3. Form residuals $\boldsymbol{y}' := \boldsymbol{y} - \sum_j w_j \boldsymbol{X}_j \hat{b}_j$, and repeat

Quantify uncertainty: which to select among many correlated variables

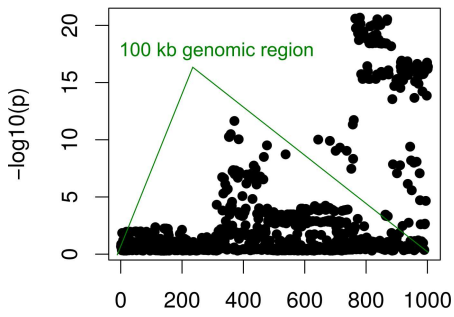
Bayesian forward selection algorithm

1. For each SNP j, fit a simple Bayesian linear regression model to get Bayes Factor BF_j
2. Form weight for each SNP, $w_j \propto \text{BF}_j$
3. Form residuals $\boldsymbol{y}' := \boldsymbol{y} - \sum_j w_j \boldsymbol{X}_j \hat{b}_j$, and repeat

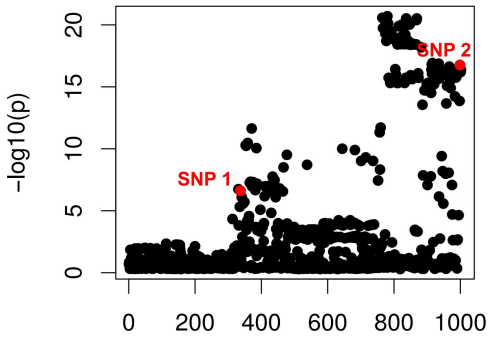
Still the same limitation: what if a "bad decision" was made early on?

A motivating example

Data available as `data(susieR::N2finemapping)`



A motivating example

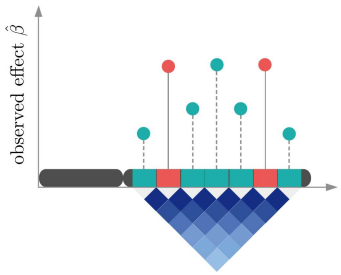


Fine-mapping using individual level data

Bayesian Variable Selection Regression and the SUm of SIngle Effects model

Detecting multiple effect variables

Intuition: A model involving the two effect variables should fit the data better than that involving only the top variable.



Software	Trait type[a]	Input covariates[b]	Uses summary statistics?	Maximum number of causal variants[c]	Input annotation?	Causal search	Main output
BIMBAM v1.0	qt and binary	No	No	Fixed	No	Exhaustive	Bayes factor
mvBIMBAM v1.0.0	mqt	No	Yes	1	No	Exhaustive	Bayes factor
SNPTEST v2.5.4-beta3	qt, binary, mqt and multinomial	No	No	1	No	Exhaustive	Bayes factor
piMASS v0.9	qt and binary	No	No	Computed	No	MCMC	Bayes factor and PIP
BVS v4.12.1	Binary	Yes	No	Computed	Yes	MCMC	Bayes factor and PIP
FM-QTL	qt	No	No	Computed	Yes	MCMC	Bayes factor and PIP
DAP v1.0.0	qt	Yes	Yes	1, fixed and computed	Yes	Exhaustive	Bayes factor and PIP
Fine-mapping	Multinomial	Yes	No	Computed	No	Greedy	PIP
Trinculo	Multinomial	Yes	No	Computed	No	Greedy	Bayes factor and PIP
BayesFM	Binary	Yes	No	20	No	MCMC	PIP
ABF	qt and binary[d]	Yes	Yes	1	No	Exhaustive	Bayes factor
fgwas v0.3.6	qt and binary[d]	No	Yes	1	Yes	Exhaustive	Bayes factor and PIP
CAVIAR/eCAVIAR	qt and binary[d]	No	Yes	Fixed	No	Exhaustive	ρ probability confidence set and PIP
PAINTOR v3.0	qt, binary[d] and mqt	No	Yes	Fixed and computed	Yes	Exhaustive and MCMC	Bayes factor and PIP
CAVIARBF v0.2.1	qt and binary[d]	No	Yes	Fixed	Yes	Exhaustive	Bayes factor and PIP
FINEMAP v1.1	qt and binary[d]	No	Yes	Fixed	No	Shotgun stochastic search	Bayes factor and PIP
JAM in R2BGLiMS v0.1	qt and binary[d]	No	Yes	Fixed and computed	No	Exhaustive and MCMC	Bayes factor and PIP

Figure: Schaid *et al.* (2018) Nat. Rev. Genet.

Bayesian Variable Selection Regression (BVSR) model

Fine-mapping is a particular multiple regression problem:

$$\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{b}_{p\times 1} + \boldsymbol{e}_{n\times 1}$$

- $\boldsymbol{b}$ is sparse: most of its elements are 0's
- Columns of $\boldsymbol{X}$ are very correlated

BVSR posterior

Assess **combinations** of variables

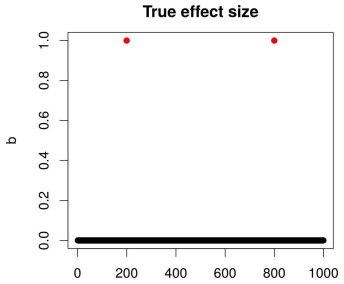
	SNPs	1	2	3	4	5	⋯	Probability
model configurations		1	0	1	0	0	⋯	0.25
		1	0	0	1	0	⋯	0.25
		0	1	1	0	0	⋯	0.25
		0	1	0	1	0	⋯	0.25

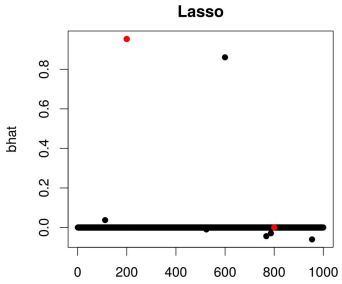
- $\text{PIP}_j := Pr(z_j \text{ is non-zero})$

$$\text{PIP} = (0.5, 0.5, 0.5, 0.5, 0, \cdots)$$

BVSR quantifies uncertainty in variable selection

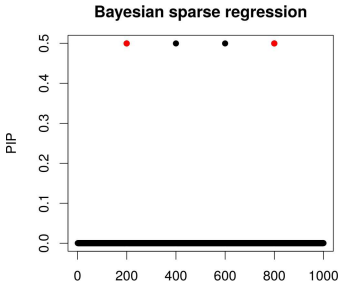
$b_1 \neq 0$ or $b_2 \neq 0$, and $b_3 \neq 0$ or $b_4 \neq 0$.

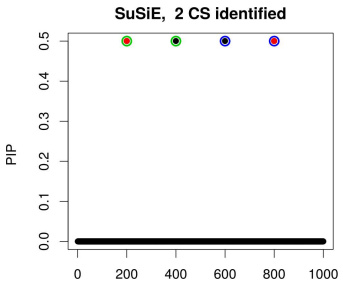




BVSR quantifies uncertainty in variable selection

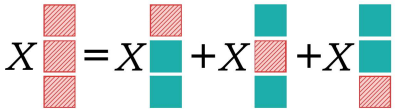
$b_1 \neq 0$ or $b_2 \neq 0$, and $b_3 \neq 0$ or $b_4 \neq 0$.





The Sum of Single Effects model (*SuSiE*)

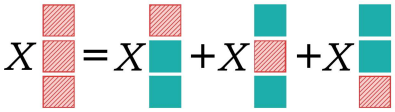
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$
$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l$$



Wang *et al.* (2020) J. R. Stat. Soc. B

The Sum of Single Effects model (*SuSiE*)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$
$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l$$

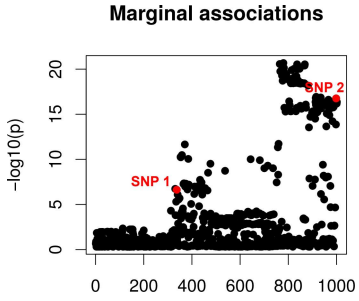


A variational approximation to posterior under *SuSiE*

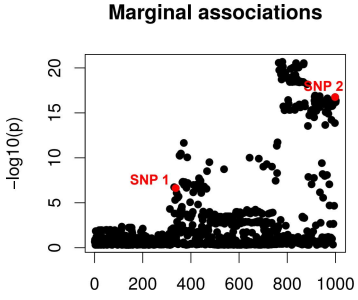
$$q(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L) = \prod_l q_l(\boldsymbol{b}_l)$$

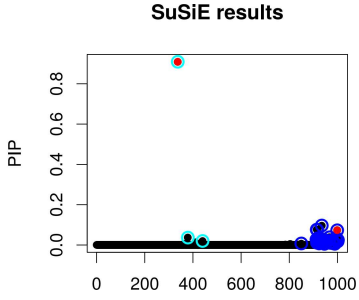
- $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L$ are treated as **independent** *a posteriori*.
- **Do not** assume q_l factorizes over the elements of $\boldsymbol{b}_l$.

Real-world example illustrated

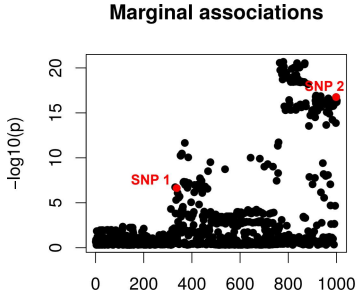


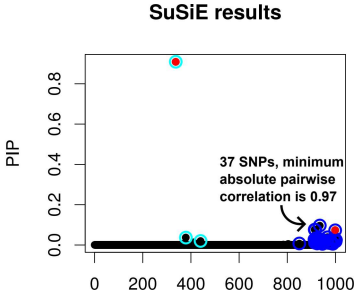
Real-world example illustrated



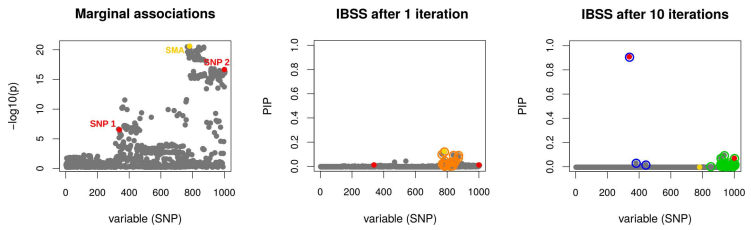


Real-world example illustrated

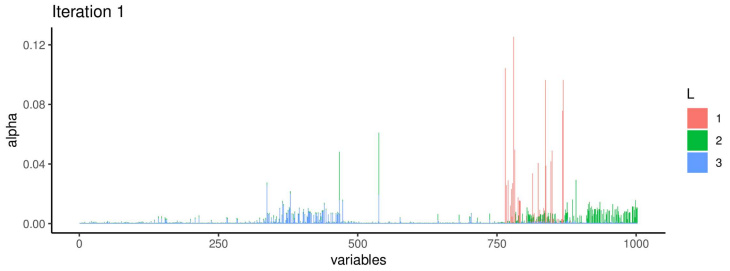




Real-world example illustrated



The IBSS algorithm iterations breakdown



Fine-mapping using association summary statistics

with practical considerations in real-world data

Summary statistics in association studies

- X, genotype matrix
- Y, phenotype matrix, can be multiple traits
- X^TY, association results — effect size estimate
- X^TX, LD matrix
- XX^T, genomic relatedness matrix, reflects kinship
- Y^TY, trait correlation, relevant in multi-trait analysis and integration

Reasons to work with summary statistics

Advantage over full data (genotypes and phenotypes):

- Easier to obtain and share with others
- Convenient to use: QC and data wrestling barely needed
- Computationally suitable for large-sample problems
  - $\mathcal{O}(p^2)$ (summary statistics) $\ll \mathcal{O}(np)$ (full data)
  - when sample size $n \gg$ variants in fine-mapped region p

Suggested reading: Pasaniuc and Price (2017) Nat. Rev. Genet.

Association testing marginal effects summary statistics

z-scores from univariate association studies:

$$\hat{z}_j := \hat{\beta}_j / s_j,$$

where

$$\hat{\beta}_j := (\boldsymbol{x}_j^\top \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{y} \quad s_j := \sqrt{\hat{\sigma}_j^2 (\boldsymbol{x}_j^\top \boldsymbol{x})^{-1}}$$

- **Sufficient** statistics: $\boldsymbol{x}^\top \boldsymbol{x}$, $\boldsymbol{x}^\top \boldsymbol{y}$, $\hat{\sigma}_j^2$
- $\hat{\sigma}_j^2$ is the estimated residual variance from regressing $\boldsymbol{y}$ on $\boldsymbol{x}_j$

Approximate multiple linear regression from summary statistics

Ordinary least-squares:

$$\beta_{\text{joint}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}) = (N^{-1}\mathbf{X}'\mathbf{X})^{-1}(N^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{R}_{\text{GWAS}})^{-1}\mathbf{D}^{-1}\beta_{\text{GWAS}}$$

where:

- $\mathbf{X}$ is the matrix a $N \times M$ matrix of genotypes (at the focal locus) in the GWAS sample
- $\mathbf{y}$ is the vector of phenotypes for the N individuals in the GWAS
- $\mathbf{R}_{\text{GWAS}}$ is the correlation matrix between SNPs in the GWAS sample
- β_{GWAS} is the vector marginal SNP effects
- $\mathbf{D}$ is a diagonal matrix: $D_j = \text{var}(X_j)$ is the variance of allele counts at SNP j in the GWAS sample, as a function of sample size and MAF

Approximate multiple linear regression from summary statistics

Ordinary least-squares:

$$\beta_{\text{joint}} = (\mathbf{R}_{\text{GWAS}})^{-1}\mathbf{D}^{-1}_{\text{GWAS}}\beta_{\text{GWAS}}$$

Issue: However, $\mathbf{R}_{\text{GWAS}}$ and $\mathbf{D}_{\text{GWAS}}$ may not always been available.

Solutions: approximate $\mathbf{R}_{\text{GWAS}}$ and $\mathbf{D}_{\text{GWAS}}$ from a sample of individuals with same genetic ancestries as the GWAS sample.

$$\beta_{\text{joint}} \approx (\mathbf{R}_{\text{REF}})^{-1}\mathbf{D}^{-1}_{\text{REF}}\beta_{\text{GWAS}}$$

Crucial assumption in many summary statistics-based methods.

Regression model for z-scores

$$\hat{\boldsymbol{z}} \sim N(\hat{\boldsymbol{R}}\boldsymbol{z}, \hat{\boldsymbol{R}})$$

Assumptions:

1. Heritability of any single SNP is small
2. $\hat{\boldsymbol{R}}$ is sample genotypic correlation for the same study
3. Genotypes used to computed $\boldsymbol{z}$ and $\hat{\boldsymbol{R}}$ are accurate

Regression for with $\hat{\beta}$ and SE$(\hat{\beta})$

The $\hat{z}$ model (exchangable z-scores):

$$\hat{\boldsymbol{z}} \sim N(\hat{\boldsymbol{R}}\boldsymbol{z}, \hat{\boldsymbol{R}})$$

The $\hat{b}, \hat{s}$ model (exchangable effects):

$$\hat{\boldsymbol{b}}|\hat{\mathbf{s}} \sim N(\hat{\mathbf{S}}\hat{\boldsymbol{R}}\hat{\mathbf{S}}^{-1}\boldsymbol{b}, \hat{\mathbf{S}}\hat{\boldsymbol{R}}\hat{\mathbf{S}})$$

- $\hat{z}$ model: lower MAF variants have larger effects
- $\hat{b}, \hat{s}$ model: effect sizes are the same regardless of MAF
- $\hat{b}, \hat{s}$ model takes sample size into consideration
  - No longer have to assume small effect per SNP

Suggested reading: Zhu and Stephens (2017) AoAS

Properties of per SNP z scores

- z-score for a SNP depends on effects of both itself and other correlated SNPs:

$$\text{E}(\hat{z}_j|\hat{\boldsymbol{R}}) = \sum_{i=1}^{p} r_{ij}z_j.$$

  GWAS marginal effects are biased due to LD!

- z-scores are correlated,

$$\text{Cor}(\hat{z}_j, \hat{z}_k) = r_{jk}, \forall j, k$$

- Application: LD score regression (LDSC)

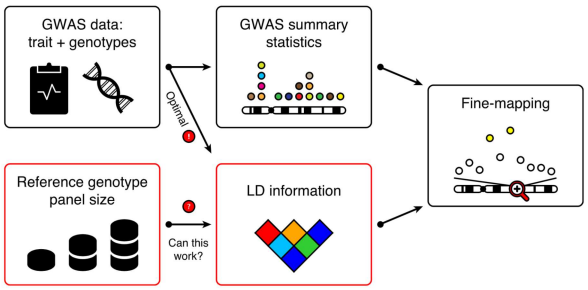


Figure: Benner *et al.* (2017) Am. J. Hum. Genet.

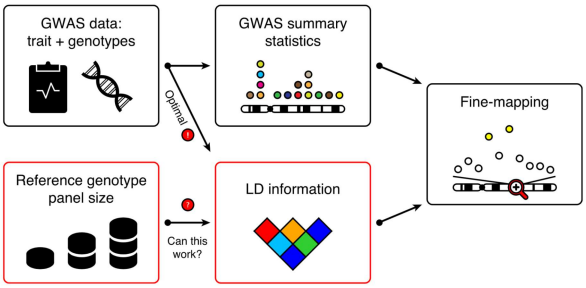


Figure: Benner *et al.* (2017) Am. J. Hum. Genet.

Posterior $\propto$ Likelihood $\times$ Prior

$f(\beta|\text{Summary data}) \propto f(\text{Summary data}|\beta) \times f(\beta)$

Single Causal variant model fine-mapping using summary statistics

- m SNPs in the region to fine-map
- Prior = each SNP has the same probability to be causal
- Posterior:

$$P(C_j|Z_1, ..., Z_m) = \frac{\exp(\frac{Z_j^2}{2})}{\sum_{k=1}^{m} \exp(\frac{Z_k^2}{2})}$$

SuSiE Regression with Summary Statistics (RSS)

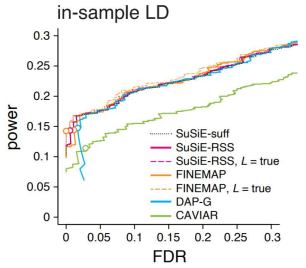
"Single effects": $\boldsymbol{z}_l$'s

$$\hat{\boldsymbol{z}} \sim N(\hat{\boldsymbol{R}}\boldsymbol{z}, \hat{\boldsymbol{R}})$$
$$\boldsymbol{z} = \sum_{l=1}^{L} \boldsymbol{z}_l$$
$$\boldsymbol{z}_l = \gamma_l z_l$$
$$z_l \sim N(0, \omega_l^2)$$
$$\gamma_l \sim \text{Mult}(1, \boldsymbol{\pi})$$

$\boldsymbol{z}$ = $\boldsymbol{z}_1$ + $\boldsymbol{z}_2$ + $\boldsymbol{z}_3$

Suggested reading: Zou et al (2022) PLoS Genet.

Summary statistics fine-mapping methods comparison



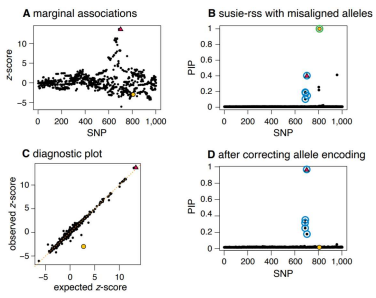
In practice people often use SuSiE RSS, Zou *et al.* (2022) PLoS Genet + FINEMAP, Benner *et al.* (2016) Bioinformatics

Impact of bioinformatics error: allele flips

What is allele flip?

- Different allele encoding between GWAS and LD reference
- *e.g.* AA=0, AC=1, CC=2 in GWAS; AA=2, AC=1, CC=0 in LD reference genotype
- A challenging problem coupled with strand flip, when merging sequence data from different platforms

Impact of allele flips



Zou *et al.* (2022) PLoS Genet.

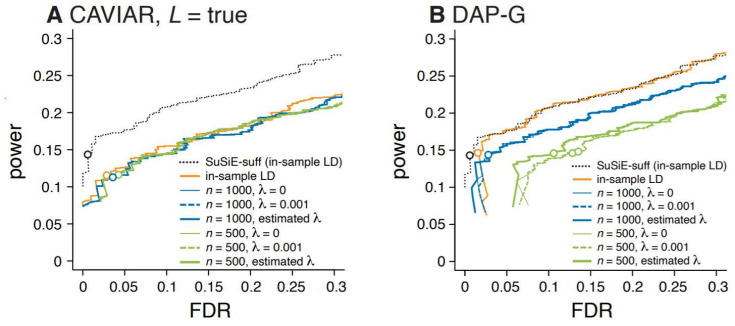
Addressing the allele flip challenge

R packages,

- `susieR::susie_rss()` function implements a diagnosis
- `bigsnpr::snp_match()` function implements a basic allele matching for two sets of summary statistics
- `MungeSumstats` R package provides a suite of tools to standardize association testing summary statistics.

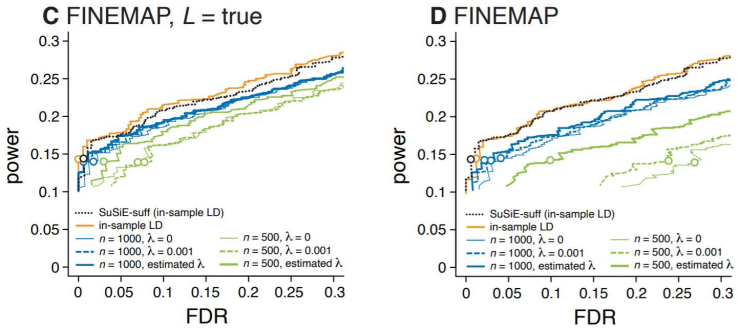
73

Impact of mis-matched LD reference: PIP



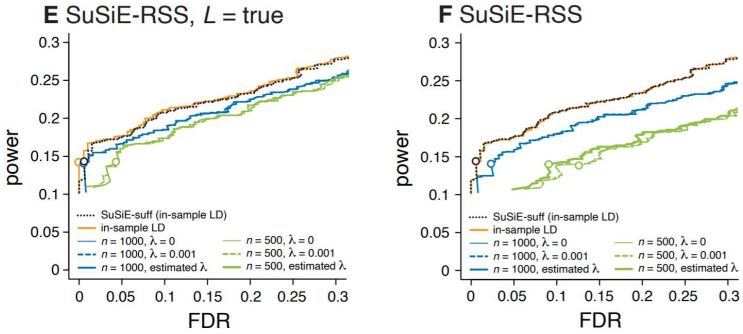
74

Impact of mis-matched LD reference: PIP



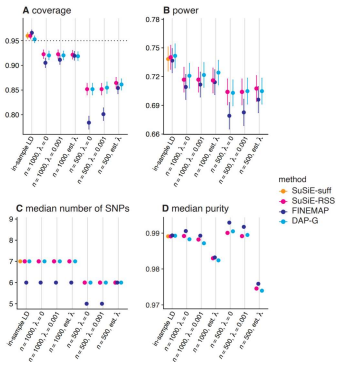
74

Impact of mis-matched LD reference: PIP



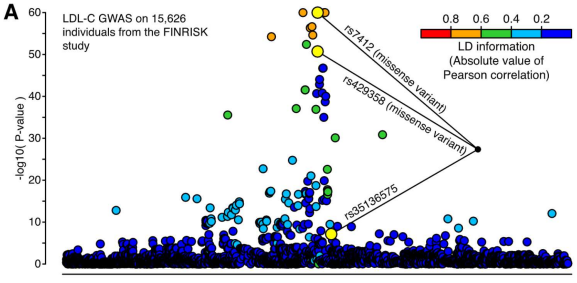
74

Impact of mis-matched LD reference: credible sets



75

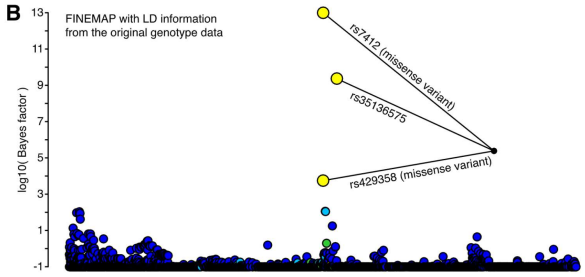
Impact of mis-matched LD reference: real data



Benner *et al.* (2017) Am. J. Hum. Genet.

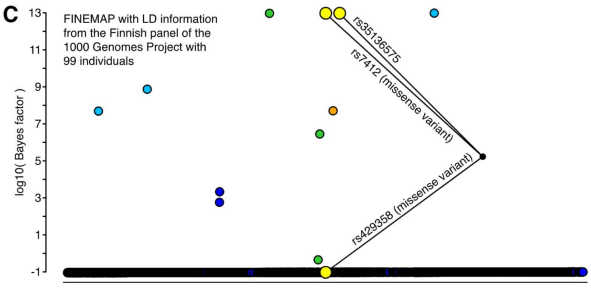
76

Impact of mis-matched LD reference: real data



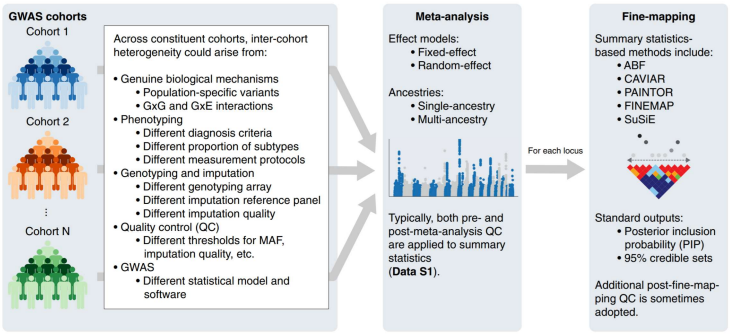
Benner *et al.* (2017) Am. J. Hum. Genet.

Impact of mis-matched LD reference: real data

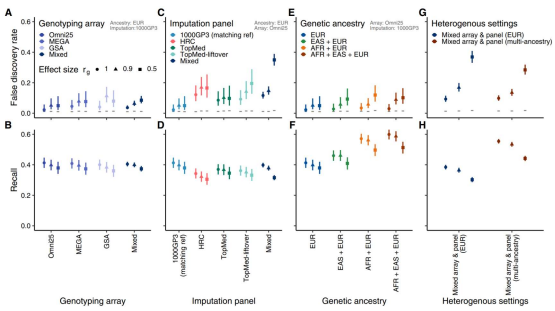


Benner *et al.* (2017) Am. J. Hum. Genet.

Fine-mapping in meta-analysis: overview



Key factors of reference data discrepancy in meta-analysis

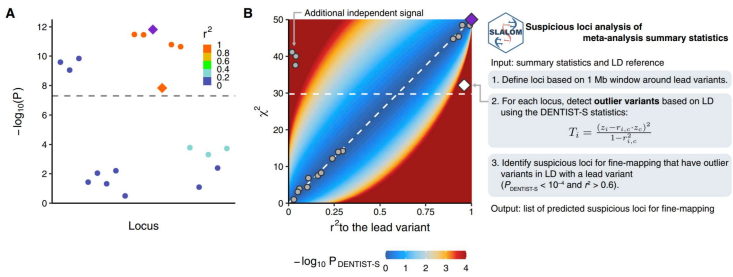


Kanai *et al.* (2022) Cell Genomics

Fine-mapping in meta-analysis: mismatch diagnosis

1. Compare allele frequencies between LD panel and GWAS, filter out SNPs with too large differences (GCTA: `--diff-freq`)
2. Filter per-SNP sample size outliers
3. Use methods like DENTIST and SLALOM to detect LD inconsistencies

Fine-mapping in meta-analysis: mismatch diagnosis



Chen *et al.* (2021) Nat. Comm. (DENTIST)
Kanai *et al.* (2022) Cell Genomics (SLALOM)

Covariate adjustment in LD reference

Consider two GWAS regression analysis:

1. Evaluate SNP effect in "Trait $\sim$ SNP + Age + Sex + PCs"
2. Fit model "Trait $\sim$ Age + Sex + PCs", compute residual of Trait (remove covariates), and evaluate SNP effect in model Residual_Trait $\sim$ SNP

Are these two analysis equivalent?

Covariate adjustment in LD reference

Consider two GWAS regression analysis:

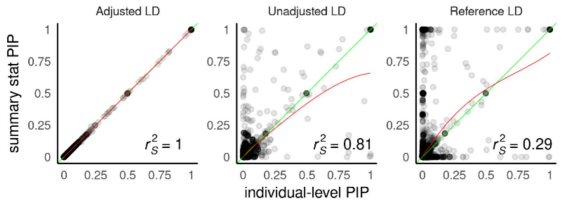
1. Evaluate SNP effect in "Trait $\sim$ SNP + Age + Sex + PCs"
2. Fit model "Trait $\sim$ Age + Sex + PCs", compute residual of Trait (remove covariates), and evaluate SNP effect in model Residual_Trait $\sim$ SNP

They are not equivalent because covariates should also be removed from SNP data: Residual_Trait $\sim$ Residual_SNP

More technical details see McCaw *et al.* (2020) Biometrics

Kinship (GRM) adjustment in LD reference

Kinship, represented as Genetic Relationship Matrix (GRM) should be removed from genotype data before computing LD reference for fine-mapping with summary statistics



Quick *et al.* (2020) biorxiv

Functional enrichment in fine-mapped variants

Signals concentrated in tissue / cell specific functional area

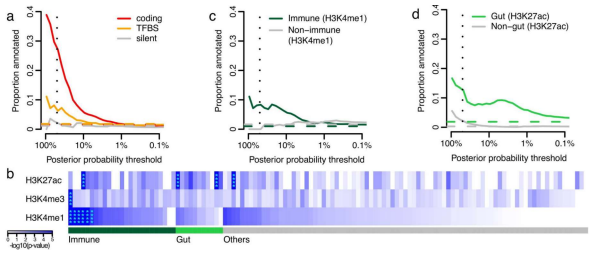
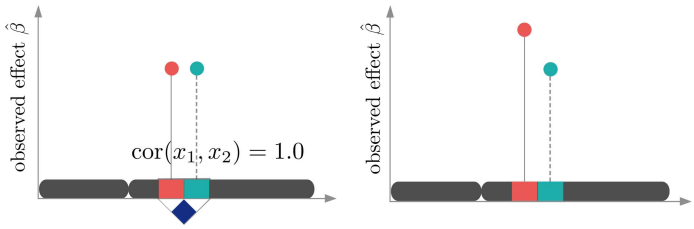


Figure: Huang *et al.* (2017) Nature

Annotations improves fine-mapping resolution



Integrating functional information prioritizes the left SNP.

Previous example: SNP 1 is causal, SNP 2 in high LD with it

	Case		Control			
SNP	G1	G2	G1	G2	BF	$\log_{10}$BF
rs1	1200	800	1000	1000	2.40×10^7	7.38
rs2	1190	809	1000	1000	0.36×10^7	6.56

When functional annotation is considered

Posterior inclusion probability of association assuming **one effect variable**,

$$\text{PIP}_1 = \frac{\text{BF}_1}{\text{BF}_1 + \text{BF}_2} = \frac{2.40}{2.40 + 0.36} = 0.87 \quad \text{PIP}_2 = \frac{\text{BF}_2}{\text{BF}_1 + \text{BF}_2} = \frac{0.36}{2.40 + 0.36} = 0.13$$

What if we determine *a priori* that SNP 1 is **twice as important** as SNP 2?

$$\frac{2 \times \text{BF}_1}{2 \times \text{BF}_1 + \text{BF}_2} = 0.93 \quad \frac{\text{BF}_2}{2 \times \text{BF}_1 + \text{BF}_2} = 0.07$$

86

Fine-mapping with functional annotations

Functional annotation in Bayesian Variable Selection Regression:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \\ \boldsymbol{e} &\sim N(0, \sigma^2 I_n) \\ \gamma_j &\sim \text{Bernoulli}(\pi) \\ \boldsymbol{b}_\gamma | \gamma &\sim g(\cdot) \\ \boldsymbol{b}_{-\gamma} | \gamma &\sim \delta_0 \end{aligned}$$

Key idea: π, prior inclusion probability, can be modelled by **enrichment** of functional annotations

87

Functionally informed fine-mapping in UK Biobank

In analyses of 49 UK Biobank traits, PolyFun + SuSiE identified >32% more fine-mapped variant–trait pairs compared to using SuSiE alone.

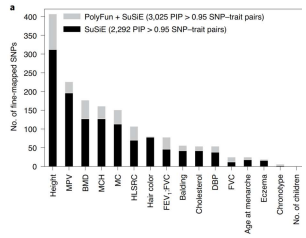


Figure: Weissbrod *et al.* (2020) Nat. Genet.

88

Example: *SuSiE* with functional informed prior

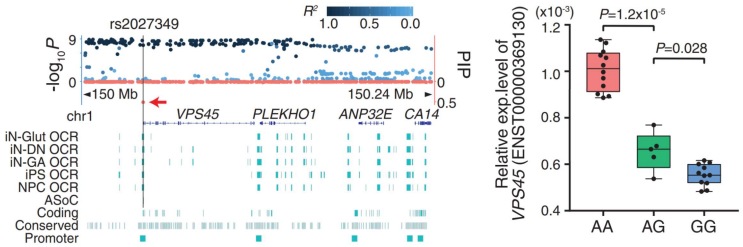


Figure: Zhang *et al.* (2020) Science

89

Caution: disease specific enrichment

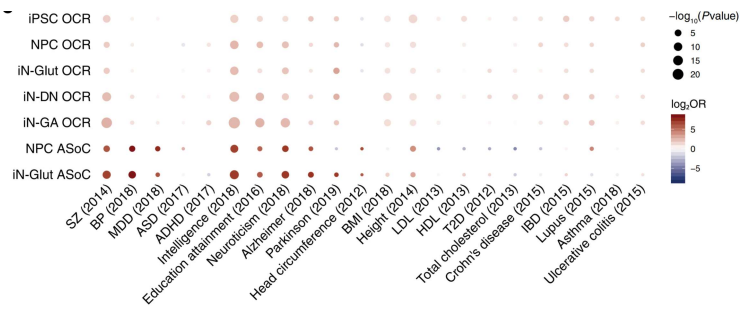


Figure: Zhang *et al.* (2020) Science

90

Exercise: fine-mapping

Overview of molecular quantitative trait loci (QTL) studies

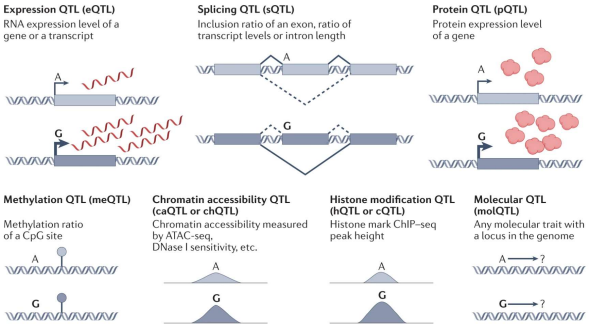


Figure: Aguet *et al.* (2023) Nature Reviews Methods Primers

Regulatory variation in complex traits and disease

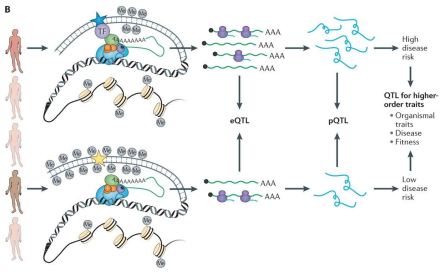


Figure: Albert & Kruglyak (2015) Nature Reviews Genetics

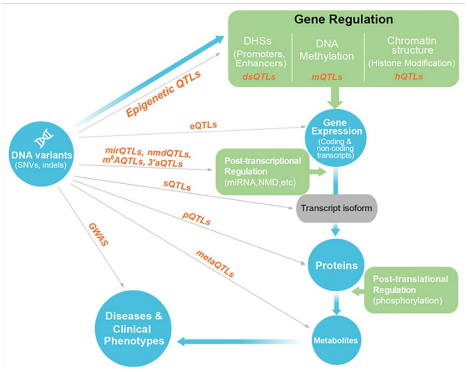


Figure: Olayinka *et al.* (2022) Curr Protoc.

Molecular QTL (xQTL) Categories

DNA and Chromatin Modifications	
DNA Methylation	**Chromatin State**
meQTL (methylation QTL)	haQTL (histone acetylation QTL)
dmrQTL (differentially methylated region QTL)	tfbsQTL (transcription factor binding site QTL)
	caQTL (chromatin accessibility QTL)
RNA Expression and Processing	
Gene Expression	**RNA Processing**
eQTL (expression QTL)	sQTL (splicing QTL: intron usage, exon skipping, splice site usage)
ct-eQTL (cell-type specific eQTL)	atsQTL (alternative transcript start QTL)
aseQTL/aimQTL (allele-specific expression QTL)	apaQTL (alternative polyadenylation QTL)
lncQTL (long non-coding RNA QTL)	nmdQTL (nonsense-mediated decay QTL)
mircQTL (microRNA QTL)	iso-eQTL (isoform-specific QTL)
circQTL (circular RNA QTL)	m6A-QTL (RNA modification QTL)
piQTL (pi RNA QTL)	
Protein and Other Molecular Phenotypes	
Proteomics | pQTL (protein QTL)	
Metabolomics & Lipidomics | metaQTL (metabolite QTL)	

Functional genomics context of xQTL

- Functional genomics features
  - cis-regulatory elements, histone marks, TF binding sites, chromatin accessibility, chromatin structure (Hi-C), etc.
  - Tissue / cell specific functional genomics contexts
- Other coding, conserved, MAF and LD-related annotations
- Machine learning derived

Other xQTL contexts

Multiple tissues

- Various brain tissues / cells
- CSF
- Other relevant tissues / cells (e.g. immune response)

Multi-ancestry & cohorts

- European
- African / African American
- Hispanic

Multiple, related complex traits

- Alzheimer's disease (AD), AD endophenotypes, related neurodegenerative disorders

96

cis- (local) and trans- (distant) xQTL

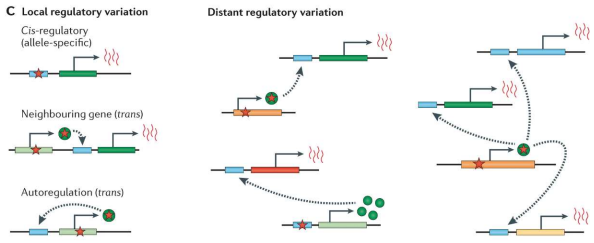


Figure: Albert & Kruglyak (2015) Nature Reviews Genetics

eQTL mapping workflow and example

Credit: eQTL mapping workflow inspired by Dr. Xin He's lecture notes
sn-eQTL example illustrates Dr. Masashi Fujita's work, Fujita *et al*. 2024.

98

RNA-seq measures gene expression

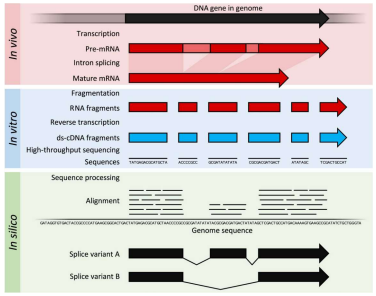


Figure: Lowe *et al*. (2017) PLoS Comp. Bio.

eQTL mapping: challenges compared to GWAS

Expression quantative trait loci (eQTL) mapping

- Identifies genetic variants associated with gene expression, revealing regulatory mechanisms.
- Challenges:
  - Noisy gene expression data influenced by sample quality and cell type composition.
  - Need for rigorous control of false discoveries in large datasets.
  - High dimensionality due to testing many genes against numerous variants.

100

Procedure of xQTL mapping: an overview

- **Quantification of molecular traits.** Such as expression levels, splicing patterns, methylation, and protein abundance, measured using technologies like RNA-seq, bisulfite sequencing, or proteomics.
- **QC and normalization of molecular traits.** Such as quantile normalization adjust for systematic biases, enabling comparisons across samples.
- **Identify hidden covariates.** Hidden factors such as cell-type composition and batch effects are addressed using techniques like PCA or factor analysis.
- **Genotype data QC and preprocessing.** Similar to GWAS.
- **Missing molecular phenotype data imputation.**
- **Association mapping for thousands of molecular features.** Identify genetic associations with molecular traits: simple regression with multiple testing correction or variable selection regression such as fine-mapping.

101

Quantifying gene expression levels from read counts

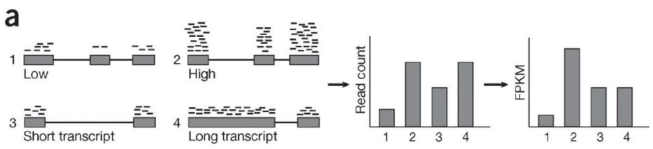


Figure: Carber *et al.* (2011) Nature Methods

- The number of reads mapping to a transcript reflects its expression level.
- However, read counts also depend on:
  - Gene length
  - Total number of reads in a sample (library size)

Common measures of gene expression from RNA-seq

- **RPKM** (Reads Per Kilo-base exon model per Million mappable reads) $R = \frac{10^9 \cdot C}{N \cdot L}$, where:
  - C: number of reads in a transcript
  - N: total number of reads in a sample
  - L: gene length
- Not suited for eQTL mapping: sum of RPKM reads in each sample can be different.
- **TPM** (Transcripts Per Million) Similar to RPKM, but total TPMs across genes sum to a constant (10^6).
- Log_2 transformation is often applied to RPKM/TPM values.

Gene expression data often needs additional normalization

- Goal: Adjust systematic biases for easier comparison across samples.
- Library size difference is one bias addressed by RPKM/TPM normalization.
- Additional sources of bias:
  - Sequencing bias toward GC-rich genes.
  - Dependency among genes due to RNA-seq protocol: higher expression of some genes reduces read counts of others.
- Other normalization methods mitigate these biases.
  - DESeq2 uses median ratios of $\frac{\text{gene count in sample}}{\text{geometric means across samples of the gene}}$ across samples to adjust both library size and composition biases.

Quantile normalization is a general strategy to adjust for sample differences

- Systematic differences in samples arise from multiple sources.
- Library size and composition biases adjustments may not suffice.
- Quantile normalization adjusts data so that *gene expression distributions are identical across samples*.
- Genes at **the same rank** in different samples are **assigned the same expression value** post-normalization.

Procedure of quantile normalization

- For a gene in sample j, let the quantile of its expression be q and its expression level be x_{qj}.
- Adjust expression to the mean expression of genes at rank q across samples:

$$\tilde{x}_q = \frac{1}{n}\sum_{j=1}^{n} x_{qj}$$

- Replace x_{qj} in sample j with $\tilde{x}_q$.

Association is typically tested by linear regression

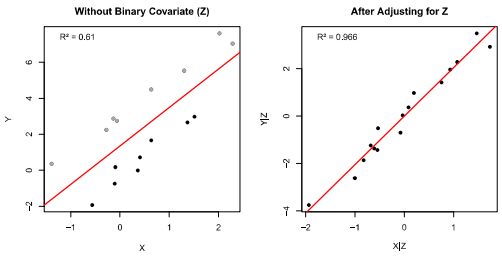
- Linear model tests the association between normalized expression Y and SNP genotypes X, adjusting for covariates Z:

$$Y = X\beta + Z\gamma + \varepsilon$$

- Covariates:
  - Known: age, gender, RNA quality, ancestry, batch effects.
  - Hidden: cell growth rate, unmeasured batch effects, cell type composition.

Adjusting for covariates in association testing increases power

Regressing out Z improves power for detecting the association between X and Y.



Adjusting for covariates in association testing example

```
n <- 16; X <- rnorm(n)
Z <- rep(c(0,1), each=n/2)  # Binary covariate
Y <- 2*X + 3*Z + rnorm(n, sd=0.3)  # Strong Z effect
mod1 <- lm(Y ~ X)  # Without covariate
Xresid <- residuals(lm(X ~ Z))  # Adjust X for Z
Yresid <- residuals(lm(Y ~ Z))  # Adjust Y for Z
mod2 <- lm(Yresid ~ Xresid)  # After adjustment
r2 <- c(raw = summary(mod1)$r.squared,
  adjusted = summary(mod2)$r.squared)
```


PCA is often used to adjust hidden covariates

Hidden covariates, such as cell type composition, often influence the expression of many genes. We can leverage gene expression correlations to identify and adjust for these hidden factors.

- **PCA (Principal Component Analysis)**:
  - Projects expression data into lower dimensions.
  - Each sample is represented by a small number of principal component (PC) scores.
  - These PC scores are used as covariates in QTL association tests.

PCA is often used to adjust hidden covariates

Hidden covariates, such as cell type composition, often influence the expression of many genes. We can leverage gene expression correlations to identify and adjust for these hidden factors.

- Number of PC to include can be estimated from data:

PCA outperforms popular hidden variable inference methods for molecular QTL mapping

Heather J. Zhou, Lei Li, Yumei Li, Wei Li & Jingyi Jessica Li

Genome Biology **23**, Article number: 210 (2022) | Cite this article

Multiple testing correction in cis-eQTL mapping

Statistical significance of xQTL:

- For cis-xQTL mapping, control the False Discovery Rate (FDR) at the gene level.
- SNPs in linkage disequilibrium (LD) often cluster near a gene, so reporting all significant SNPs is not efficient.

Identifying **eGenes:**

- eGenes are defined as genes with at least one significant cis-eQTL.
- Null hypothesis: No SNP near a gene is associated with its molecular trait.
- A gene-level statistic (e.g., minimum p-value of SNPs) is used for further analysis.

Multiple testing correction in cis-eQTL mapping

Step 1: permutation testing for null distribution **at one gene:**

- Generate the null distribution by permuting sample labels and recomputing gene-level statistics (e.g. minimum p-value).
- Calculate empirical p-values by comparing observed statistics to the null.

$$p_{\text{empirical}} = \frac{\#(\text{null test statistics} \geq \text{observed statistic})}{\text{total permutations}}$$

- Tools like `TensorQTL` perform permutation tests efficiently ("adaptive permutation").

Multiple testing correction in cis-eQTL mapping

Step 2: FDR for **genome-wide significance:**

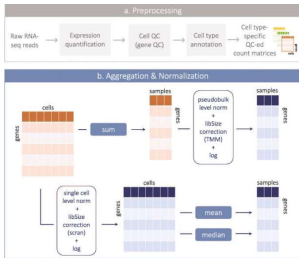
- Apply FDR correction to empirical p-values to identify significant eGenes genome-wide.
- BH procedure, or q-value

Multiple testing correction in cis-eQTL mapping

Bayesian fine-mapping provides an alternative to multiple testing control:

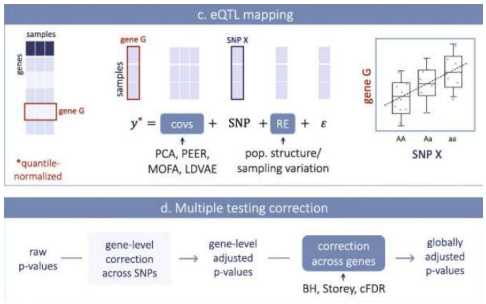
- Posterior inclusion probability (PIP) and credible sets (CS) are used to assess SNP significance.
- Multiple testing correction is unnecessary because only one test is performed per gene, modeling all SNPs together with PIP directly quantify SNP relevance.
- eGenes can be defined as those with at least one credible set of fine-mapped SNPs.

Cell type eQTL workflow: expression quantification



Cuomo *et al.* (2021) Genome Biol.

Cell type eQTL workflow: association testing



Cuomo *et al.* (2021) Genome Biol.

A case study: cell specific eQTL in brains

single-nuclei RNAseq (snRNAseq) from frozen human brain in the prefrontal cortex region

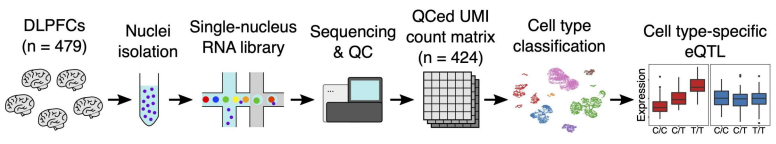
Article | Published: 21 March 2024

Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain

Masashi Fujita, Zongmei Gao, Lu Zeng, Cristin McCabe, Charles C. White, Bernard Ng, Gilad Sahar Green, Orit Rozenblatt-Rosen, Devan Phillips, Liat Amir-Zilberstein, Hyo Lee, Richard V. Pearse II, Atlas Khan, Badri N. Vardarajan, Krzysztof Kiryluk, Chun Jimmie Ye, Hans-Ulrich Klein, Gao Wang, Aviv Regev, Naomi Habib, Julie A. Schneider, Yanling Wang, Tracy Young-Pearse, Sara Mostafavi, ... Philip L. De Jager + Show authors

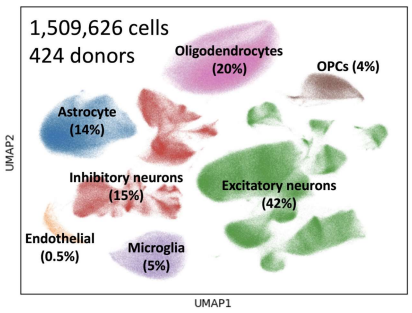
Nature Genetics **56**, 605–614 (2024) | Cite this article

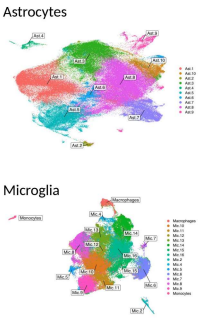
Study design overview



- Frozen dorsolateral prefrontal cortex (DLPFC) tissues of ROSMAP cohort were analyzed.
- ROSMAP participants have agreed to receive annual cognitive testing and to donote their brain upon their death. Donated brains were pathologically examined.
- Single cell gene expression libraries were constructed using 10x Genomics 3' kit.

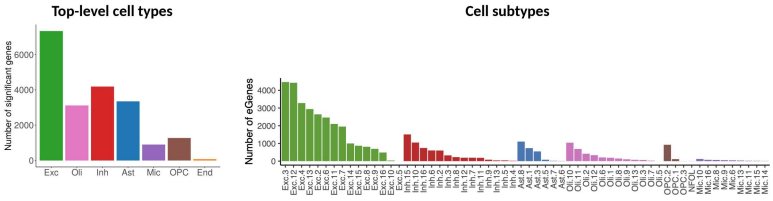
Atlas of brain cells in DLPFC





116

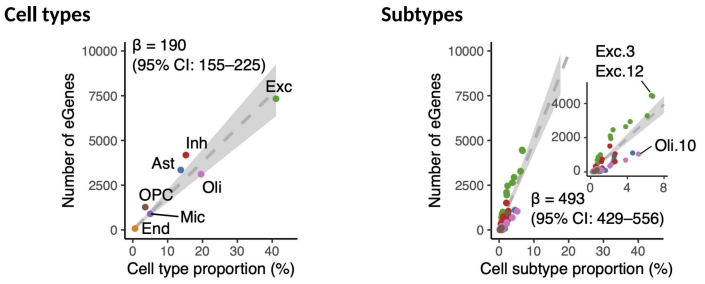
Cell specific gene and eQTL (ct-eQTL)



117

Limitation: rare cell types are under-represented

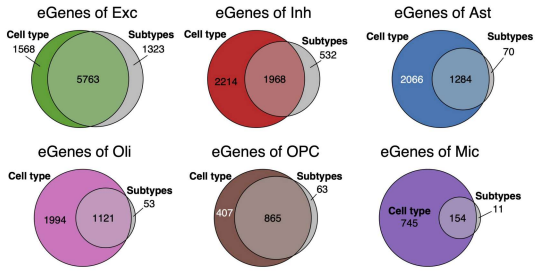
Cell type proportion is correlated with eQTL detection power



118

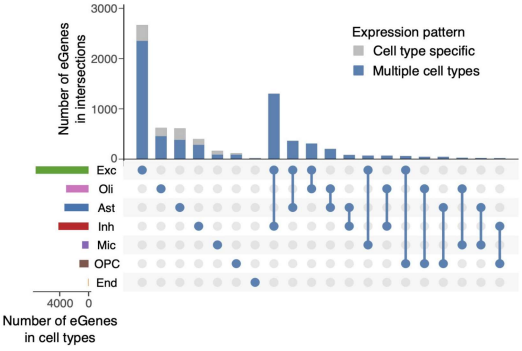
ct-eQTL discovery and comparison with eQTL

Subtype eQTL analysis boosted eQTL detection



119

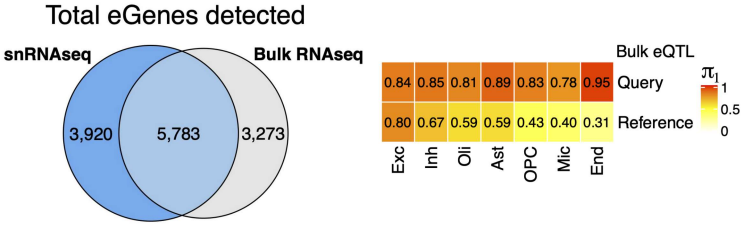
Summary of ct-eQTL specificity



120

Replication in bulk RNAseq eQTL

Low replication rate π_1 in rare cell types is consistent with the lack of detection power in those cell types

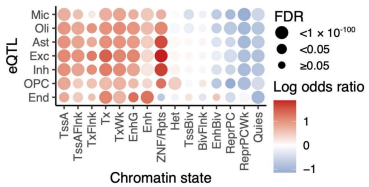


121

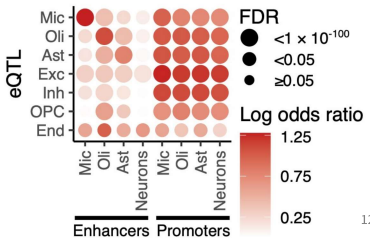
Chromatin states of ct-eQTL

eQTL are enriched in euchromatin (transcriptionally active regions and enhancers) and relatively depleted in heterochromatin

Chromatin states in bulk DLPFC

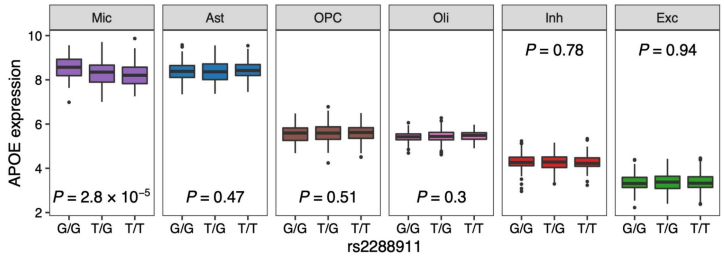


Brain cell type-specific enhancers and promoters (Nott et al., 2019)



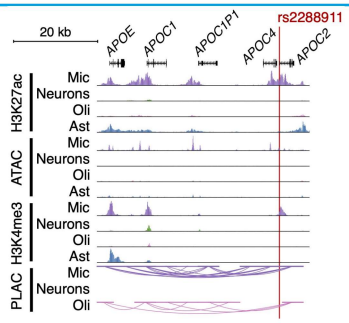
A discovery relevant to Alzheimer's disease

APOE (Apolipoprotein E) expression is specific to Mic and Ast; it's genetic regulation is specific to Mic.



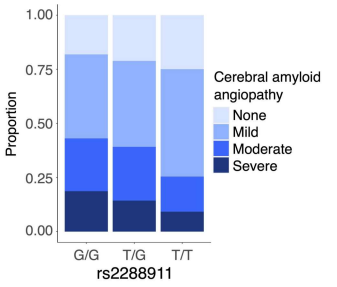
Follow up on Mic. specific eQTL rs2288911

rs2288911 is located at a microglia-specific enhancer

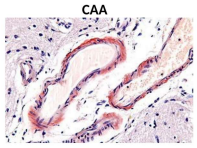


Follow up on Mic. specific eQTL rs2288911

rs2288911 is a risk SNP of AD and cerebral amyloid angiopathy (CAA)

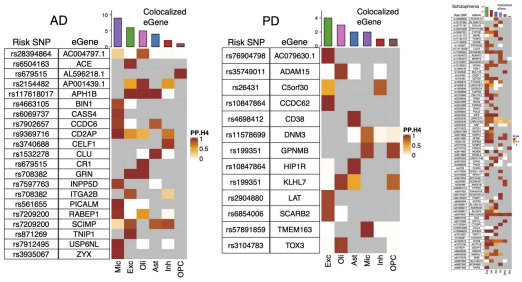


- rs2288911 is associated with the risk of AD in GWAS ($p = 1.0 \times 10^{-12}$).
- In ROSMAP, this SNP is associated with CAA ($p = 1.2 \times 10^{-7}$). This association persists even after adjusted for the *APOE* ε4 allele.



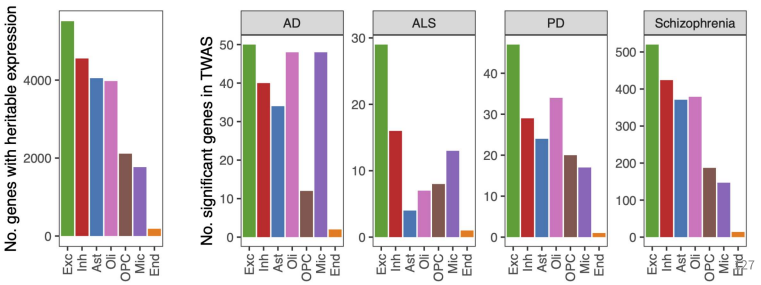
Integrative study with disease GWAS

Colocalization with GWAS signals of neurological diseases highlights the necessity of ct-eQTL studies



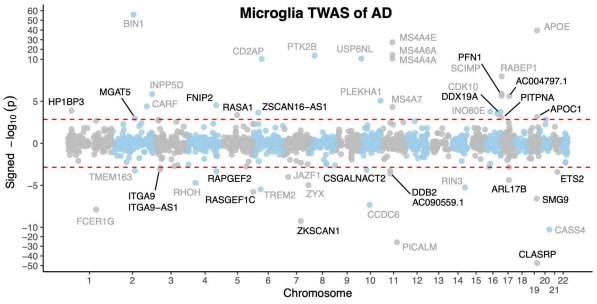
Transcriptome-wide association studies

Cell type-specific TWAS shows enrichment of microglia genes in Alzhemier's disease



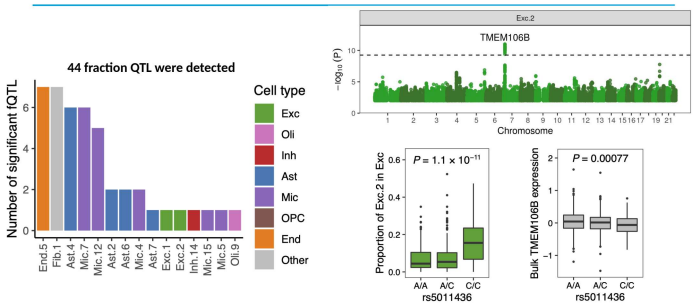
Transcriptome-wide association studies

Microglia-specific TWAS identifies novel genes associated with Alzhemier's disease



QTL of cell fraction may confer disease risk

Fraction QTL (fQTL) of Exc.2 colocalizes with AD risk SNP of gene *TMEM106B*



Rare xQTL can improve PRS for complex traits

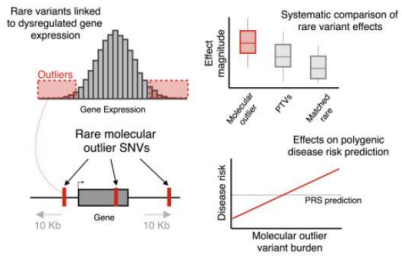


Figure: Smail *et al.* (2022) AJHG

Also see Li *et al.* (2017) Nature; Ferraro *et al.* (2020) Science

Missing regulation in eQTL and GWAS

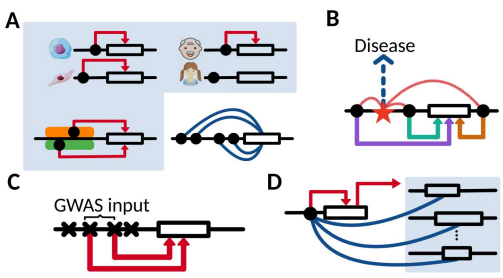
The missing link between genetic association and regulatory function

Noah J Connally, Sumaiya Nazeen, Daniel Lee, Huwenbo Shi, John Stamatoyannopoulos, Sung Chun, Chris Cotsapas, Christopher A Cassa, Shamil R Sunyaev

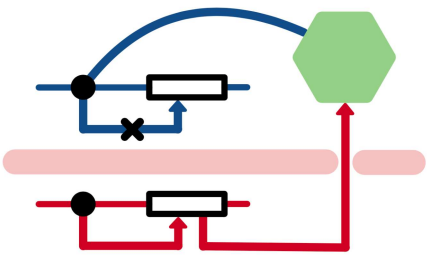
… by applying a gene-based approach we found limited evidence that the baseline expression of trait-related genes explains GWAS associations, whether using colocalization methods (8% of genes implicated), transcription-wide association (2% of genes implicated), or a combination of regulatory annotations and distance (4% of genes implicated). These results contradict the hypothesis that most complex trait-associated variants coincide with homeostatic expression QTLs, suggesting that better models are needed. The field must confront this deficit and pursue this 'missing regulation.'

Connally et al, December 2022, elife; also see Mostafavi et al + Prichard 2022

Fancier methods, or more targeted hypotheses?



Fancier methods, or more targeted hypotheses?



QTL-GWAS loci: multi-trait analysis and colocalization

More phenotypes, more complications

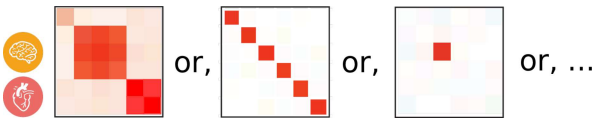


Figure: Plausible patterns of sharing

Major challenges

- **For a given variant**: the less assumption made on multivariate effects, the more parameters to estimate.
  - Fixed effect (FE) and Random Effects (RE) models are restrictive but easy to fit.
- **Different variants**: may fit in different multivariate effect models, **local patterns of genetic heterogeneity among traits**

A naive mixture model

"FE and RE are equally likely for any variant":

$$U_{mixed} = 0.5 \times \begin{bmatrix} \sigma_0^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 \end{bmatrix} + 0.5 \times \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_0^2 \end{bmatrix}$$

Prior allows for equal possibility of both; data will determine where posterior lands.

A data-adaptive mixture model

Instead of making assumptions, can we **learn from data**:

- What are the latent structures for multivariate effects?
- How often does each structure appear?

and use these to construct the mixture model?

What are shared, and what ain't?

Decomposing effect estimates (gene by tissue), $\widehat{\beta} = \boldsymbol{LF} + \boldsymbol{E}$

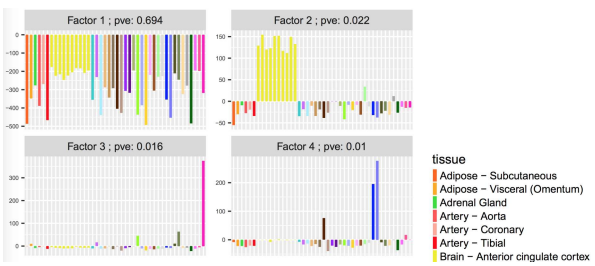


Figure: Sparse factor analysis of GTEx data

Incorporating all possible patterns

Multivariate effects of a variant follows the k-th pattern with probability π_k:

$$U_{mixed} = \pi_1 \times \begin{bmatrix} 2.4 & 0.3 \\ 0.3 & 1.5 \end{bmatrix} + \pi_2 \times \begin{bmatrix} 1.6 & 0.001 \\ 0.001 & 0.02 \end{bmatrix} + \pi_3 \times \cdots$$

How much are shared?

Multivariate **A**daptive **SH**rinkage model for effect estimates:

$$\widehat{\beta} \sim N_R(\beta, \boldsymbol{S})$$
$$\beta \sim \sum_k \pi_k N_R(\mathbf{0}, \boldsymbol{U}_k)$$

$\boldsymbol{U}_k$ was **learned from data**, π_k is **to be learned** next!

How much are shared?

Maximum likelihood for the finite mixture model:

$$\widehat{\pi} = \text{argmax} \sum_j \log\left(\sum_k \pi_k L_{k_j}\right)$$

This is a convex optimization problem $\Rightarrow$ **very fast to fit!**

How much are shared?

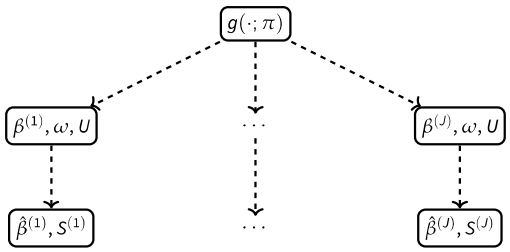
Maximum likelihood for the finite mixture model:

$$\widehat{\pi} = \text{argmax} \sum_j \log\left(\sum_k \pi_k L_{k_j}\right)$$

Multiple testing is **an opportunity to learn!**

- The more the tests, the better we learn.

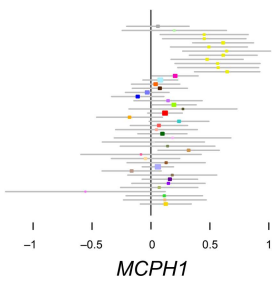
Multivariate Adaptive SHrinkage Illustrated



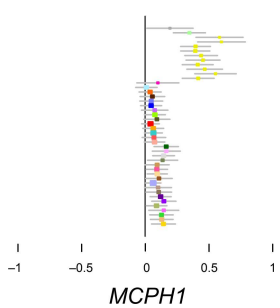
- Step 1: estimated π_k via EM algorithm using data across genome.
- Step 2: apply this prior to each variant in association mapping.

MASH improves power

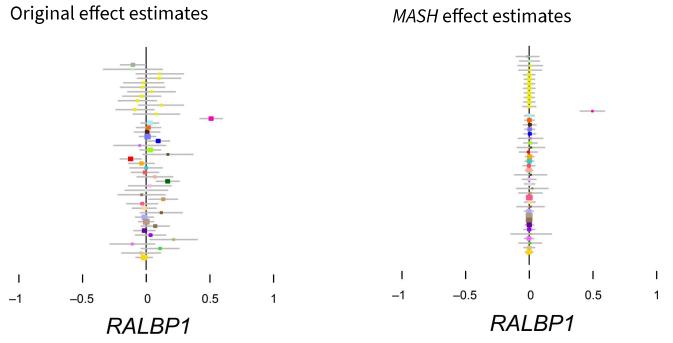
Original effect estimates



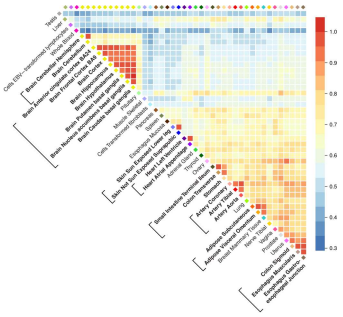
MASH effect estimates



MASH reduces noise



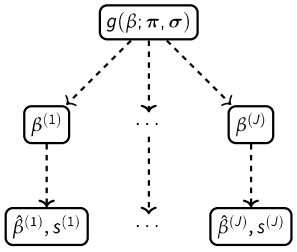
MASH reveals quantitative heterogeneity



A genome-wide *SuSiE* model

Learn shrinkage priors from genome-wide univariate tests

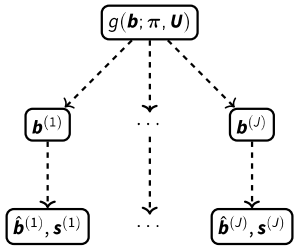
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$
$$\boldsymbol{e} \sim N(0, \sigma^2 I_n)$$
$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l$$
$$\boldsymbol{b}_l = \gamma_l \beta_l$$
$$\gamma_l \sim \text{Mult}(1, \boldsymbol{\alpha})$$
$$\beta_l \sim \sum_k \pi_k N(0, \sigma_k^2)$$



Multivariate Multiple Regression fine-mapping

The *mvSuSiE* model,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$$
$$\boldsymbol{E} \sim N_{N\times R}(\boldsymbol{0}, \boldsymbol{I} \otimes \boldsymbol{V})$$
$$\boldsymbol{B} = \sum_{l=1}^{L} \boldsymbol{B}_l$$
$$\boldsymbol{B}_l = \gamma_l \boldsymbol{b}_l^T$$
$$\gamma_l \sim \text{Mult}(1, \boldsymbol{\alpha})$$
$$\boldsymbol{b}_l \sim \sum_{k,m} \pi_{km} N_R(\boldsymbol{0}, \omega_m \boldsymbol{U}_k)$$



Application to multivariate fine-mapping

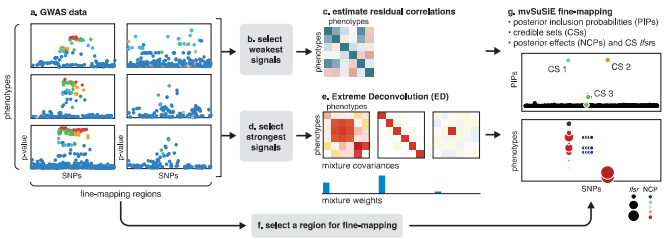


Figure: mvSuSiE fine-mapping with adaptive shrinkage model

Zou *et al.* (2023) biorxiv

Multi-ancestry fine-mapping as a special case

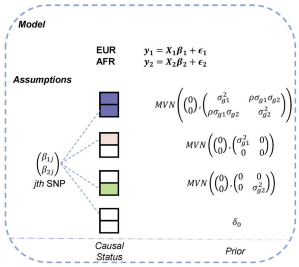
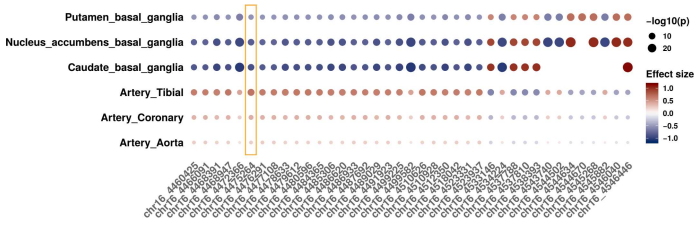


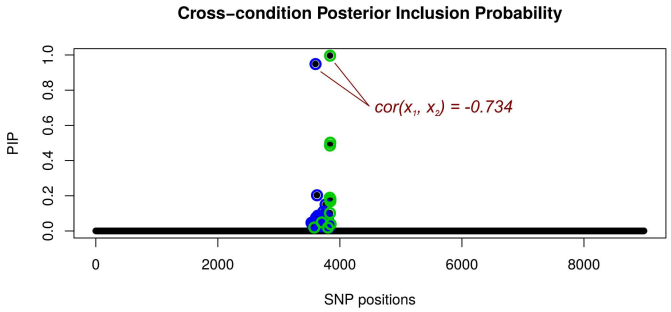
Figure: MESuSiE for cross-ancestry fine-mapping

Gao & Zhou *et al.* (2024) Nat. Genet.

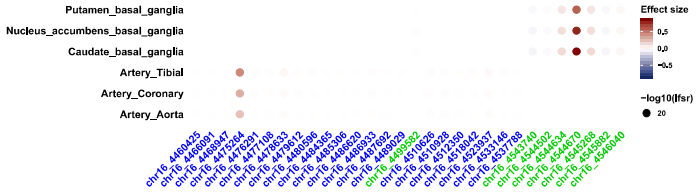
"one eQTL per gene" architecture



mvSuSiE fine-mapping architecture



mvSuSiE fine-mapping architecture

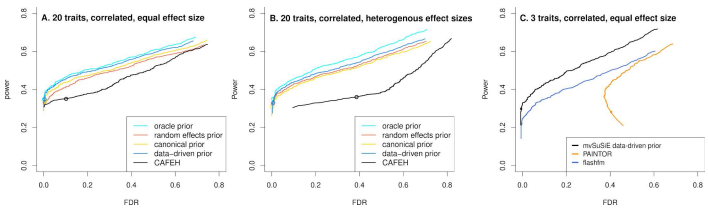


Multi-trait fine-mapping methods & challenges

	mvSuSiE	CAFEH	PAINTOR	MTHESS	BayesSUR	flashfm	msCaviar	HyPrColoc	moloc
>5 traits integrated	■	■						■	
>10 traits integrated	■	■						■	
Multiple causal signals	■	■		■	■				
Individual level data	■	■		■	■				
Summary statistics	■	■	■			■	■	■	■
Missing data						■			
Trait specific LD							■		
Correlated effects	■			■	■	■			
Trait specific effects	■	■				■		■	■
Arbitrary heterogeneous effects	■								
Arbitrary multi-trait colocalization		■				■		■	■
Correlated traits	■			■	■	■		■	
Partial sample overlap	■					■		■	
Functional annotation			■						
Trait specific functional annotation									
Genome-wide scalability	■	■						■	

Reference: CAFEH: Arvanitis et al (2022); PAINTOR: Kichaev et al (2017); MTHESS: Lewin et al (2016); BayesSUR: Zhao et al (2021); flashfm: Hernández et al (2021); msCaviar: LaPierre et al (2021); HyPrColoc: Foley et al (2021); moloc: Giambartolomei et al (2018).

Comparison to other methods



GWAS application: 16 blood traits in UK Biobank

Analysis overview

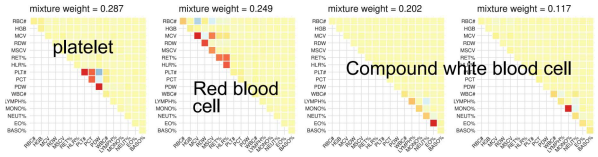
- Sample size 248,980; 975 candidate regions fine-mapped
- Average #SNPs per region 4,776; maximum 36,605

GWAS application: 16 blood traits in UK Biobank

Analysis overview

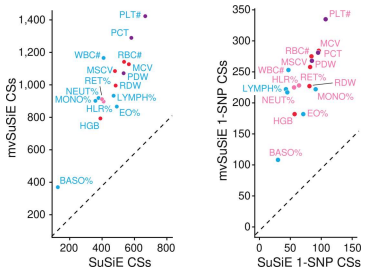
- Sample size 248,980; 975 candidate regions fine-mapped
- Average #SNPs per region 4,776; maximum 36,605

Top patterns of effect size sharing inferred from data:



GWAS application: 16 blood traits in UK Biobank

Many more signals identified compared to fine-mapping per each trait



Beyond per trait per variant association studies

Statistical fine-mapping (multiple regressors)

- Identify non-zero effect variables by accounting for LD

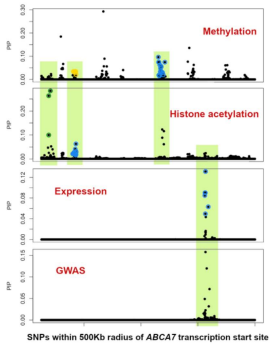
Meta-analysis (multiple responses)

- Integrate information across multiple conditions / studies

"Causal" variants across multiple conditions?

- Cross-population fine-mapping; colocalization; pleiotropy; mediation; …

The problem



The problem

For a genetic variable analyzed in two conditions:

$$P(\text{"causal" in trait 1 \& 2} \mid \text{association data for 1 \& 2})$$

The problem

For a genetic variable analyzed in two conditions:

$$P(\text{"causal" in trait 1 \& 2} \mid \text{association data for 1 \& 2})$$

Denote data as D_1 and D_2, and use indicator variables γ_1, γ_2 for variable having effects in 1 and 2, and hyperparameters Θ:

$$P(\gamma_1 = 1, \gamma_2 = 1 | D_1, D_2, \Theta)$$

Multivariate relationships?

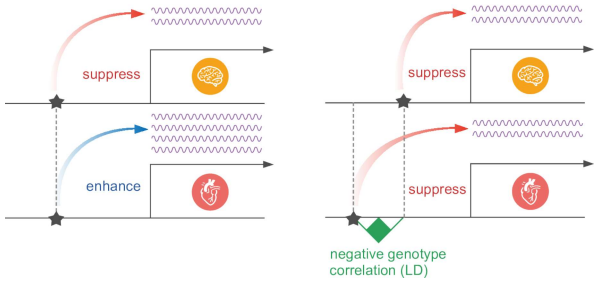


Figure: Pleiotropy or Linkage?

QTL-GWAS colocalization model and inference

- **Problem:** For variants analyzed in GWAS and eQTL studies, compute:

$$P(\gamma_g = 1, \gamma_e = 1 | D_g, D_e, \Theta)$$

QTL-GWAS colocalization model and inference

- **Problem:** For variants analyzed in GWAS and eQTL studies, compute:

$$P(\gamma_g = 1, \gamma_e = 1 | D_g, D_e, \Theta)$$

- **Model:** Compute posterior probabilities (PP_h) for:
  - H_0: No associations ($\gamma_g = 0, \gamma_e = 0$)
  - H_1: eQTL only ($\gamma_g = 0, \gamma_e = 1$)
  - H_2: GWAS only ($\gamma_g = 1, \gamma_e = 0$)
  - H_3: Two independent variants ($\gamma_g = 1, \gamma_e = 1$, independent)
  - H_4: Shared causal variant ($\gamma_g = 1, \gamma_e = 1$, colocalized)

QTL-GWAS colocalization model and inference

- **Problem:** For variants analyzed in GWAS and eQTL studies, compute:

$$P(\gamma_g = 1, \gamma_e = 1 | D_g, D_e, \Theta)$$

- **Model:** Compute posterior probabilities (PP_h) for:
  - H_0: No associations ($\gamma_g = 0, \gamma_e = 0$)
  - H_1: eQTL only ($\gamma_g = 0, \gamma_e = 1$)
  - H_2: GWAS only ($\gamma_g = 1, \gamma_e = 0$)
  - H_3: Two independent variants ($\gamma_g = 1, \gamma_e = 1$, independent)
  - H_4: Shared causal variant ($\gamma_g = 1, \gamma_e = 1$, colocalized)
- **Posterior computation:**

$$PP_h \propto \sum_{S \in S_h} P(S) \cdot P(D|S)$$

where S_h are configurations for model H_h, with **prior** $P(S)$ and likelihood $P(D|S)$

Colocalization method: *coloc*

coloc [Giambartolomei *et al.* (2014) PLoS Genet.]

- On X: "one causal" assumption
- On Y: the null + 4 combinations given "one causal"
  1. In 1 but not 2
  2. In 2 but not 1
  3. In 1 and 2 but not the same variable
  4. In 1 and 2 and the same variable (colocalization)
  5. No association in both data 1 and 2

Colocalization method: *eCAVIAR*

eCAVIAR [Hormozdiari *et al.* (2016) Am. J. Hum. Genet.]

- On X: multiple effect variables
- On Y: each effect variable can be
  1. In 1 but not 2
  2. In 2 but not 1
  3. In both 1 and 2
  4. No association in both data 1 and 2

eCAVIAR effects assumption

Effect sizes are independent,

$$U = \begin{bmatrix} \sigma_g^2 & 0 \\ 0 & \sigma_e^2 \end{bmatrix}$$

Colocalization method: *enloc*

enloc [Wen *et al.* (2017) PLoS Genet.]

- Key difference: cross-condition effects not independent
- **eQTL signals are enriched in GWAS**

Colocalization method: *enloc*

enloc [Wen *et al.* (2017) PLoS Genet.]

- Key difference: cross-condition effects not independent
- **eQTL signals are enriched in GWAS**

But how?

- Through a simple logistic link using eQTL as an annotation for j

$$\log\left[\frac{\pi}{1-\pi}\right] = \alpha_0 + \alpha\gamma_e$$

and in this context

$$\pi := P(\gamma_g = 1 | \gamma_e = 1)$$

enloc two step procedure

1. Obtain $P(\gamma_g = 1)$ and $P(\gamma_e = 1)$ using fine-mapping
2. Fit the enrichment model via multiple imputation

Connections between colocalization methods

- *eCAVIAR* is a special case of *enloc* with $\alpha = 0$.
- *coloc* is a special case of "one causal" fine-mapping based *enloc* with fixed, high(!) α value by default.
- Recent *coloc* extension: *coloc* version 5, aka *SuSiE-coloc* removed the "one causal" assumption.
  - Wallace (2021) PLoS Genetics
  - `https://chr1swallace.github.io/coloc/`

Connections between colocalization methods

- *eCAVIAR* is a special case of *enloc* with $\alpha = 0$.
- *coloc* is a special case of "one causal" fine-mapping based *enloc* with fixed, high(!) α value by default.
- Recent *coloc* extension: *coloc* version 5, aka *SuSiE-coloc* removed the "one causal" assumption.
  - Wallace (2021) PLoS Genetics
  - `https://chr1swallace.github.io/coloc/`

Summary: pattern and scale of effect size correlations, represented as different prior models.

Practical considerations

- Choice of prior
  - Best to **estimate enrichment α from data**
  - $\alpha \in [0, 5]$ suggested by $> 4,000$ GWAS + GTEx data
- LD reference mismatch: underestimate α, thus power loss

Hukku *et al.* (2021) Am. J. Hum. Genet.

Multi-trait colocalization model configurations

Assuming a single causal variant in the loci.

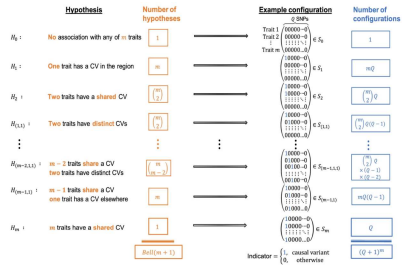
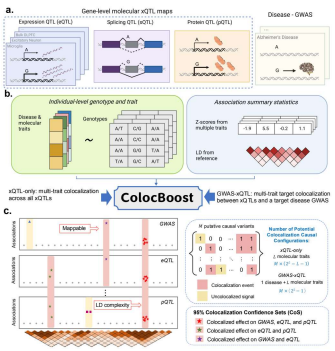
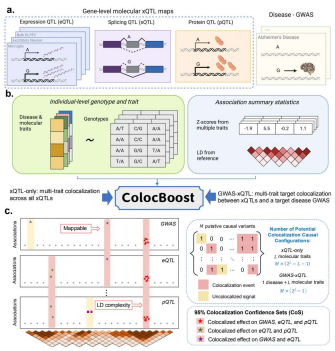


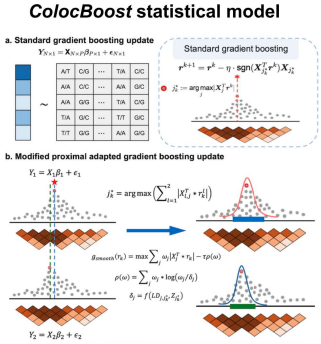
Figure: `HyPrColoc`, Foley *et al.* (2021) Nat. Comm.

ColocBoost, a new multi-trait colocalization method

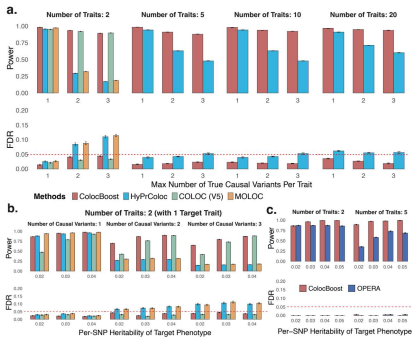


ColocBoost, a new multi-trait colocalization method

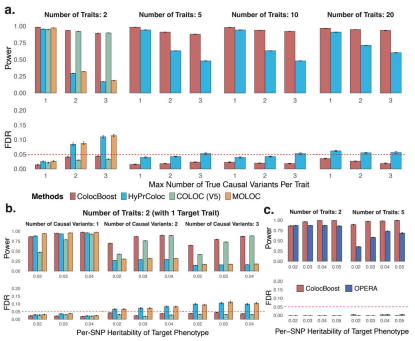


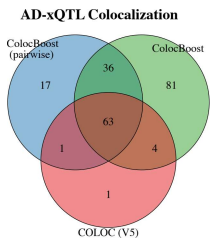


ColocBoost outperforms existing methods

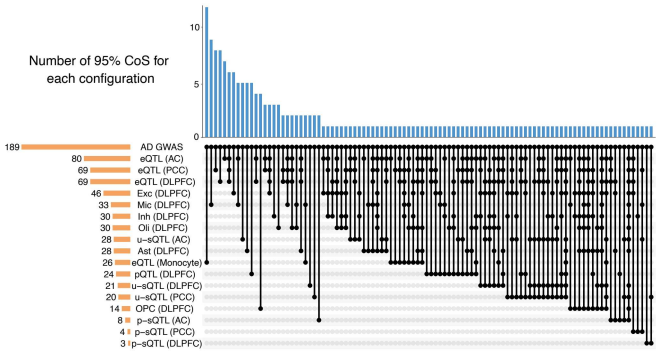


ColocBoost outperforms existing methods





Brain xQTL and Alzheimer's disease GWAS colocalization



QTL-GWAS genes: transcriptome-wide association studies

Motivation: eQTLs are enriched in GWAS signals

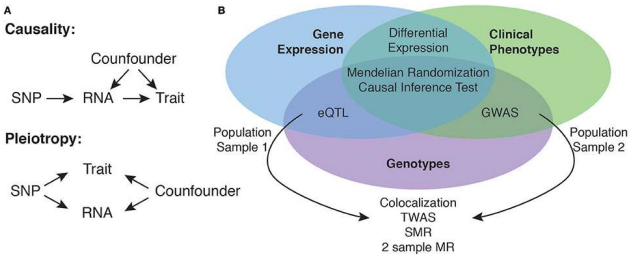


Figure: Heinig (2018) Front. Cardiovasc. Med.

Transcriptome-wide association study (TWAS)

Contributions of multiple genetic variants to complex traits through their impact on molecular phenotypes

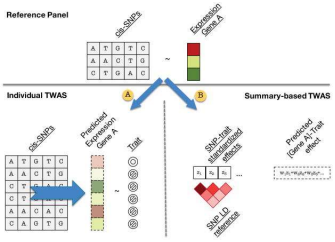
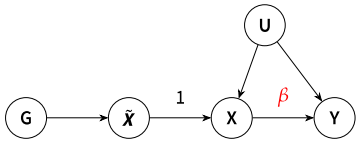


Figure: Gusev *et al.* (2016) Nat. Genet.

The underlying model of TWAS



- **G**: Genetic variants.
- **X**: Gene expression.
- $\tilde{\mathbf{X}}$: Genetic component of expression (**X** predicted from **G**).
- **Y**: Phenotype.
- **U**: Unobserved confounder.

$Y \sim \tilde{X}$ regression estimates effect β of expression on phenotype.

TWAS: a two-stage procedure (individual level data)

Stage 1: Build a prediction model for gene expression (***X***) from genotypes (***G***)

- **Model** for one gene: $X = G\boldsymbol{w} + \varepsilon_X$
- Predicted expression: $\widetilde{X} = G\hat{\boldsymbol{w}}$

TWAS: a two-stage procedure (individual level data)

Stage 1: Build a prediction model for gene expression ($\boldsymbol{X}$) from genotypes ($\boldsymbol{G}$)

- **Model** for one gene: $X = G\boldsymbol{w} + \varepsilon_X$
- Predicted expression: $\widetilde{X} = G\widehat{\boldsymbol{w}}$

Stage 2: Test for association of predicted expression ($\widetilde{\boldsymbol{X}}$) with phenotype ($\boldsymbol{Y}$)

- **Model** for one gene: $Y = \widetilde{X}\beta + \varepsilon_Y$
- Test the null hypothesis: $\beta = 0$

TWAS challenge: association vs causality

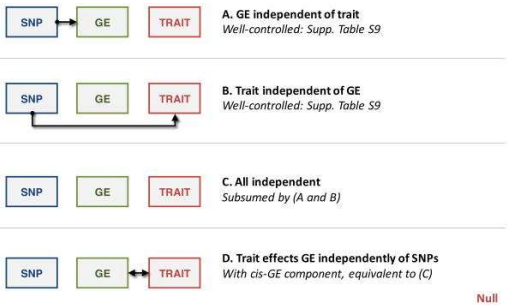


Figure: Gusev *et al.* (2016) Nat. Genet.

TWAS challenge: association vs causality

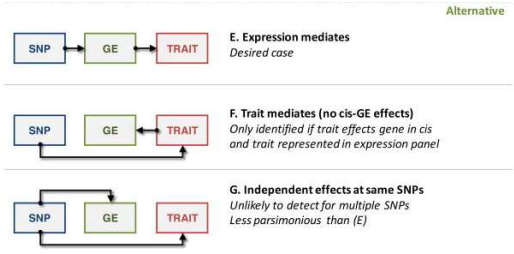


Figure: Gusev *et al.* (2016) Nat. Genet.

TWAS challenge: technical considerations

Ideal TWAS setup

- Homogenous population
- Tissue and cell-type specific
- Training data-set is large and complete ($N > 200$)

But in reality

- Cross population TWAS aplications
- Multiple tissue and cell-types
- Availability of individual level data vs summary statistics

TWAS methods review

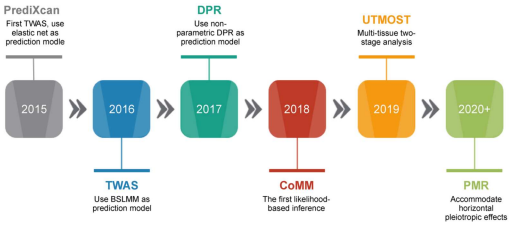


Figure: Zhu and Zhou *et al.* (2021) Quantitative Biology

A list of TWAS methods timeline created by Rui Dong (Leal lab, Columbia): `https://github.com/cumc/handson-tutorials/tree/main/contents/twas/20241231_TWAS_literature_review.ipynb`.

TWAS applications review

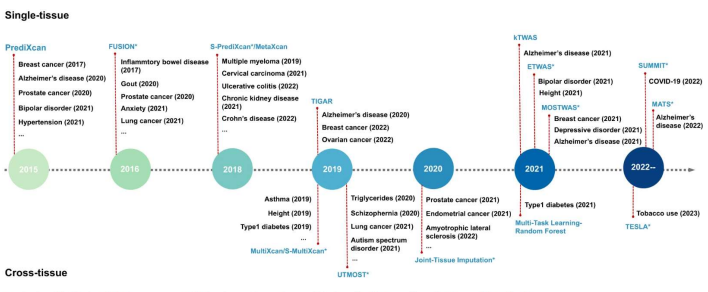


Figure: Mai *et al.* (2023) Communications Biology

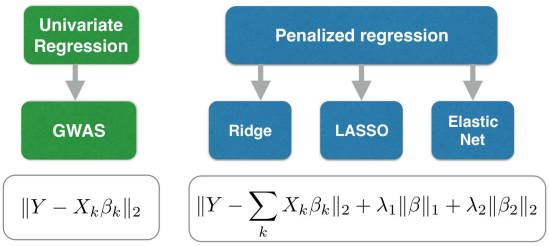
TWAS regression methods

Credit: slide contents on univariate TWAS methods from Dr. Haky Im
2018 HGEN 471, The University of Chicago

181

Univariate TWAS methods overview

$$Y = \sum_{k=1}^{M} \beta_k X_k + \epsilon$$



$\|Y - X_k\beta_k\|_2$

$\|Y - \sum_k X_k\beta_k\|_2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta_2\|_2$

182

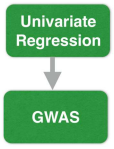
Simple regression method

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

$$Y = \sum_{k=1}^{M} \hat{\beta}_k^{\text{GWAS}} X_k$$



183

Ridge regression / BLUP

REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

AJHG 2011

$$Y = \sum_{k=1}^{M} \hat{\beta}_k^{\text{Ridge}} X_k$$

$\|Y - \sum_k X_k\beta_k\|_2 + \qquad \lambda_2\|\beta_2\|_2$

184

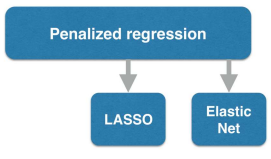
Other penalized regression

J. R. Statist. Soc. B (2005)
67, Part 2, *pp.* 301–320

Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie
Stanford University, USA

$$Y = \sum_{k=1}^{M} \hat{\beta}_k^{\text{E-N}} X_k$$



$\|Y - \sum_k X_k\beta_k\|_2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta_2\|_2$

185

Bayesian variable selection regression

OPEN ACCESS Freely available online — PLOS | GENETICS

Polygenic Modeling with Bayesian Sparse Linear Mixed Models

Xiang Zhou[1]*, Peter Carbonetto[1], Matthew Stephens[1,2]*

$$Y = \sum_{k=1}^{M} \beta_k^L X_k + \sum_{k=1}^{M} \beta_k^S X_k + \epsilon$$

$$\beta_k^L \sim N(0, \sigma_L^2)$$

$$\beta_k^S \sim N(0, \sigma_S^2)$$

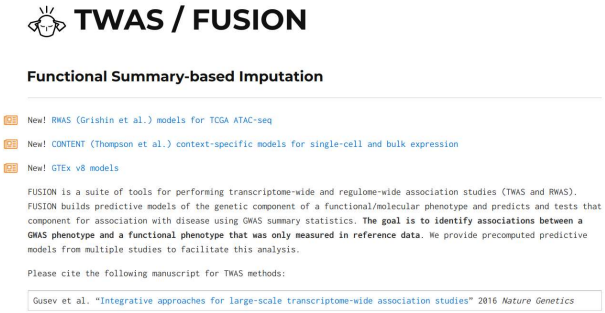
MultiBLUP: improved SNP-based prediction for complex traits

Doug Speed and David J Balding

Genome Res. published online June 24, 2014
Access the most recent version at doi:10.1101/gr.169375.113

186

Choice of methods: cross validation



Likelihood based approach

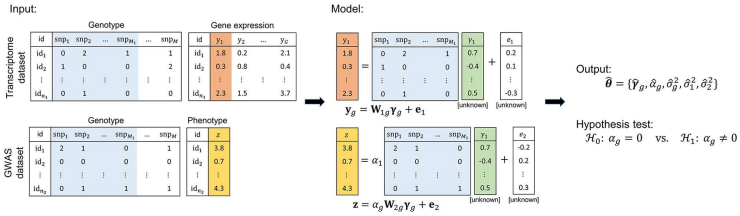
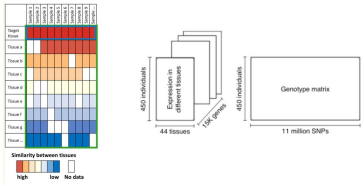


Figure: CoMM, Yeung *et al.* (2019)

Also see Yuan *et al.* (2022) likelihood based Mendelian Randomization

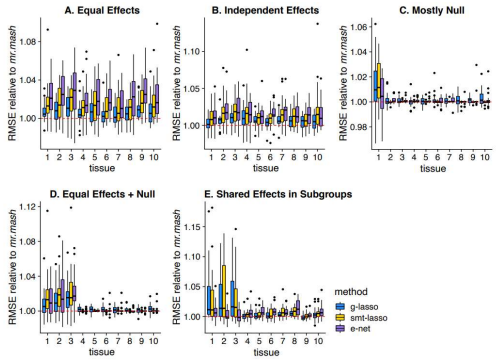
Multivariate TWAS methods

Leverage similarity between molecular phenotypes



- UTMOST, Yu *et al.* (2019) Nature Genetics
- MR-JTI, Zhou *et al.* (2020) Nature Genetics
- mr.mash, Morgante *et al.* (2024) PLoS Genetic

Multivariate TWAS method: mr.mash



Prioritize genes near GWAS hits

Step 1: Map SNPs to genes

- Define gene boundaries: often use TSS $\pm$ 100kb window
- Consider additional regulatory regions when available

Step 2: Aggregate evidence

- Common test statistic: $T = \sum_{j=1}^{m} Z_j^2$ for independent m SNPs
- Implementation: VEGAS (PLINK), fastBAT (GCTA), MAGMA

Note: Methods differ in how they calculate p-values for the aggregated statistics

Prioritize genes near GWAS hits

Step 2: Aggregate evidence

- Common test statistic: $T = \sum_{j=1}^{m} Z_j^2$ for independent m SNPs
- Implementation: VEGAS (PLINK), fastBAT (GCTA), MAGMA

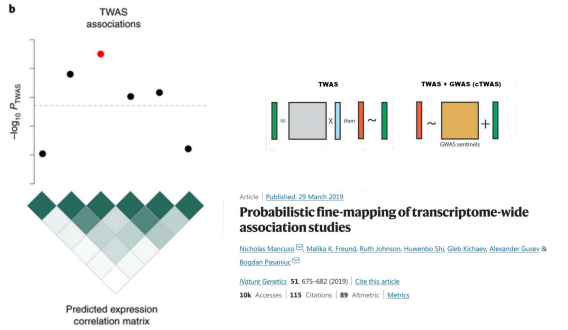
Connection to TWAS:

- Summary statistics based TWAS test: $Z_{\text{TWAS}} = \frac{\sum_{j=1}^{m} w_j Z_j}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} w_i w_j R_{ij}}}$
  - w_j: **weights precomputed from molecular QTL** (*e.g.* eQTL)
  - R_{ij}: accounts for LD between variants i and j
- Basic approach: Assumes $w_j = 1$ (unweighted), $R_{ij} = I$ (independent SNPs)
- Then $(Z_{\text{TWAS}})^2 = \frac{(\sum_{j=1}^{m} Z_j)^2}{\sum_{j=1}^{m} 1} \to \sum_{j=1}^{m} Z_j^2$

Exercise: multivariate fine-mapping and TWAS

Connections: fine-mapping, colocalization, TWAS and MR

TWAS and fine-mapping: variable selection



Article | Published: 29 March 2019

Probabilistic fine-mapping of transcriptome-wide association studies

Nicholas Mancuso, Malika K. Freund, Ruth Johnson, Huwenbo Shi, Gleb Kichaev, Alexander Gusev & Bogdan Pasaniuc

Nature Genetics 51, 675–682 (2019) | Cite this article

10k Accesses | 115 Citations | 89 Altmetric | Metrics

TWAS and fine-mapping: variable selection

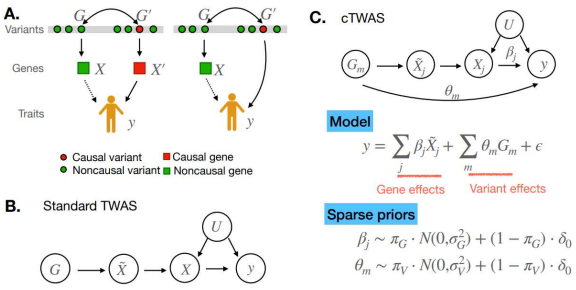


Figure: Zhao *et al.* (2024) Nat. Genet.

TWAS and colocalization: pleiotropy

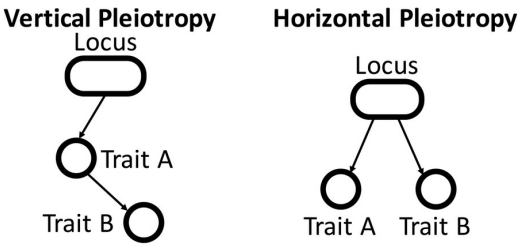
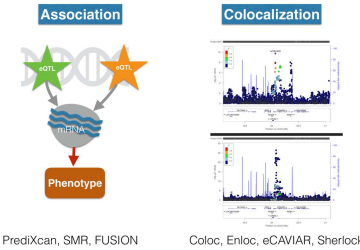


Figure: Jordan *et al.* (2019) Genome Biology

TWAS + colocalization: pleiotropy



PrediXcan, SMR, FUSION

Coloc, Enloc, eCAVIAR, Sherlock

- Image credit: Haky Im
- "Locus level" colocalization: Hukku *et al.* (2022) AJHG; Okamoto *et al.* (2023) AJHG.

TWAS and colocalization: statistical framework

$$\boldsymbol{M} = \mu_M \mathbf{1} + \boldsymbol{G}\boldsymbol{\beta}_E + \mathrm{e}_M, \mathrm{e}_M \sim \mathrm{N}\left(\boldsymbol{0}, \sigma_M^2 \boldsymbol{I}\right)$$
$$\boldsymbol{Y} = \mu_Y \mathbf{1} + \gamma \boldsymbol{M} + \boldsymbol{G}\boldsymbol{\beta}_Y + \mathrm{e}_Y, \mathrm{e}_Y \sim \mathrm{N}\left(\boldsymbol{0}, \sigma_Y^2 \boldsymbol{I}\right)$$

- "locus level", $Pr(\gamma \neq 0|\text{Data}) \propto Pr(\gamma \neq 0)Pr(\text{Data})$
- $\Pr(\gamma \neq 0) = Pr(coloc) \times Pr(twas)$
- Data: z-score from TWAS.
- Key idea: Test $\gamma = 0$, not to estimate γ which is Mendelian Randomization.

196

TWAS and Mendelian randomization

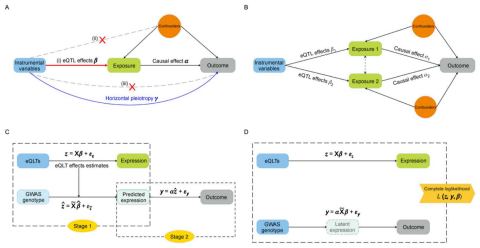


Figure: Zhu and Zhou (2021) Quantitative Biology

TWAS can be viewed as two-sample MR — using various IV selection methods.

197

Colocalization and Mendelian randomization

- **Colocalization** answers:
  "Are gene and trait expressions caused by the same variants?"
- **Mendelian Randomization (MR)** answers:
  "Does gene expression cause traits?"

Suggested reading:

Zuber V, Grinberg NF, Gill D, Manipur I, Slob EAW, Patel A, Wallace C, Burgess S (2022). Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *American Journal of Human Genetics*, 109(5):767-782. doi: 10.1016/j.ajhg.2022.04.001

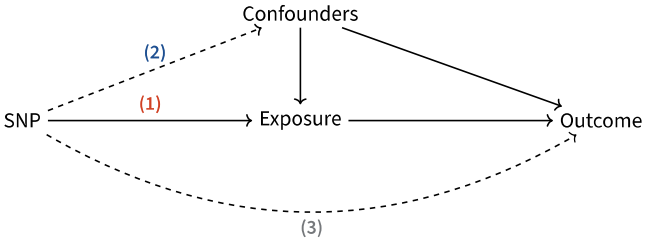
198

Appendix: Mendelian Randomization (MR)

Model and methods outlined for reference only
MR will be lectured in detail on Thursday

199

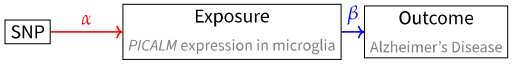
Mendelian Randomization: 3 Core Assumptions



1. **SNP is associated with the exposure.**
2. **SNP is NOT associated with confounding variables.**
3. **SNP ONLY associated with outcome through the exposure.**

200

Mendelian randomization with molecular traits as exposure



$$\text{Ratio Estimate} = \frac{\text{G-Y association}}{\text{G-X association}} = \hat{\beta}$$

201

Timeline of MR methods

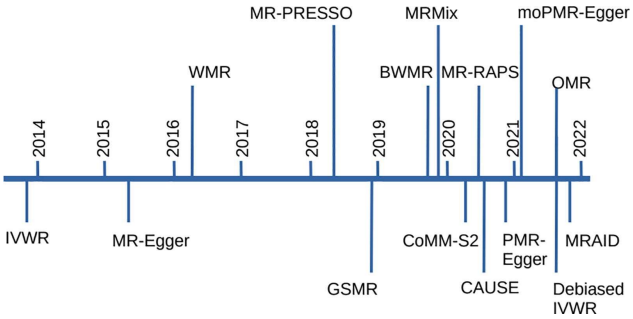


Figure: Boehm and Zhou (2022) Comp. Struct. Biotechnol. J.

MRAID — MR with automated instrument determination

MRAID for individual level data

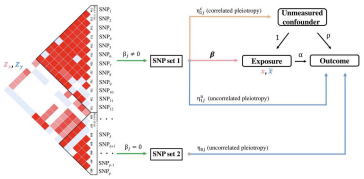


Figure: Yuan *et al.* (2022) Sci. Adv.

Exposure GWAS

$$\mathbf{x} = \mathbf{Z_x}\beta + \epsilon_\mathbf{x}$$

Outcome GWAS

$$\tilde{\mathbf{x}} = \mathbf{Z_y}\beta + \tilde{\epsilon}_\mathbf{x}$$

$$\mathbf{y} = \underbrace{\tilde{\mathbf{x}}\alpha}_{\substack{\text{SNP set 1}\\ \text{through}\\ \text{exposure}}} + \underbrace{\mathbf{Z_y}\eta_0}_{\substack{\text{SNP set 2}}} + \underbrace{\mathbf{Z_y}\eta_1}_{\substack{\text{SNP set 1}\\ \text{but not}\\ \text{through}\\ \text{exposure}}} + \epsilon_\mathbf{y}$$

MRAID — MR with automated instrument determination

MRAID for summary statistics

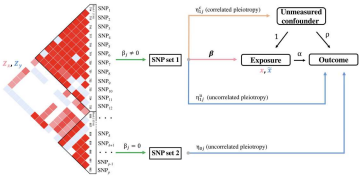


Figure: Yuan *et al.* (2022) Sci. Adv.

Exposure GWAS

$$\hat{\beta}_\mathbf{x} = \Sigma_1\beta + \mathbf{e_x}$$

Outcome GWAS

$$\hat{\beta}_\mathbf{y} = \underbrace{\alpha\Sigma_2\beta}_{\substack{\text{SNP set}\\ 1\\ \text{through}\\ \text{exposure}}} + \underbrace{\rho\Sigma_2(\beta \circ \mathbf{v})}_{\substack{\text{SNP set 1}\\ \text{through}\\ \text{confounder}\\ \text{Correlated}\\ \text{plieotropy}}} + \underbrace{\Sigma_2\eta_\mathbf{u}}_{\substack{\text{SNP set 1 and 2}\\ \text{but through}\\ \text{neither exposure}\\ \text{nor confounder}\\ \text{Uncorrelated}\\ \text{plieotropy}}} + \mathbf{e}_y$$

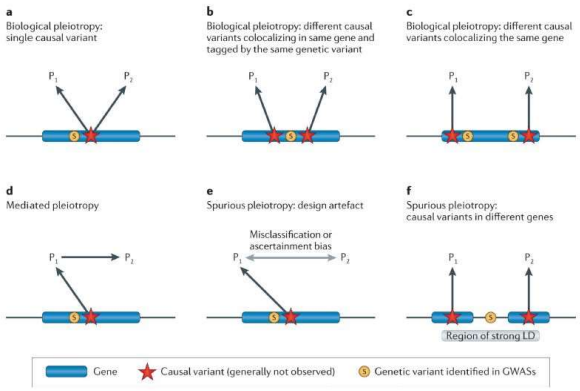
From cross-phenotype associations to pleiotropy in human genetic studies

Andrew DeWan, PhD, MPH
Associate Professor of Epidemiology
Co-Director, Yale Center for Perinatal, Pediatric and Environmental Epidemiology
Director of Graduate Studies
Yale School of Public Health

Yale SCHOOL OF PUBLIC HEALTH

Pleiotropy

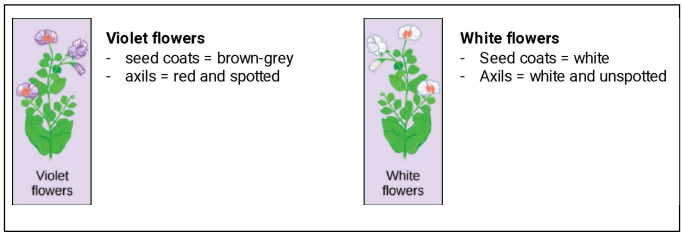
- Phenomenon in which a genetic locus affects more than one trait or disease
- Molecular level
  - Single gene with multiple physiological function
  - Two domains of a single gene product with different functions and affecting multiple phenotypes
  - Gene product with a single function that affects multiple phenotypes acting in multiple tissues
- Statistical level
  - A locus displaying cross-phenotype associations is often considered pleiotropic
  - Can be at the variant, gene or region level



Solovieff et al. Nat Rev Genet. 2013 July ; 14(7): 483–495. doi:10.1038/nrg3461.

Early example of "pleiotropy"

Gregor Mendel documented one of the earliest examples of pleiotropy in his pea plant experiments



Mendel, J. G., 1866 Experiments in plant hybridization. Verhandlungen des naturforschenden Vereines in Brunn 4: 3–47 (in German).

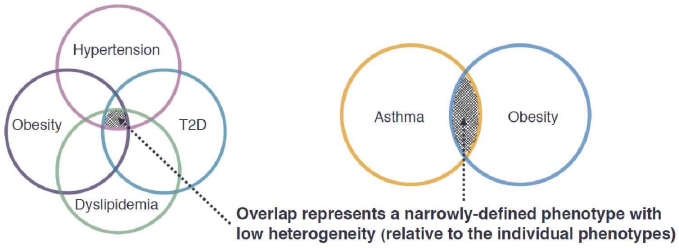
Examples in humans

- Marfan syndrome
  - FBN1 (fibrillin-1)
  - thinness, joint hypermobility, limb elongation, lens dislocation, and increased susceptibility to heart disease.
- Holt-Oram syndrome,
  - TBX5 (transcription factor)
  - cardiac and limb defects
- Nijmegen breakage syndrome
  - NBS1 (DNA damage repair protein)
  - microcephaly, immunodeficiency, and cancer predisposition

Pleiotropy and complex disease comorbidity

- Examples of correlated (comorbid) disease
  - Obesity, hypertension, dyslipidemia, type 2 diabetes (metabolic disorder)
  - Depression, anxiety, personality disorders (psychiatric disorder)
  - Asthma, obesity (pro-inflammatory conditions)
- Why do certain disease occur together
  - Causality
  - Shared environmental risk factors
  - Shared genetic risk factors

Pleiotropy and complex disease comorbidity



Pleiotropy and complex disease comorbidity

- Pleiotropy-informed analyses consider multiple phenotypes together and take into account the correlation between the phenotypes
  - Analyzing multiple correlated phenotype (e.g. comorbid diseases) is equivalent to analyzing a single narrowly-defined phenotype with low heterogeneity

Pleiotropy and complex disease comorbidity

- Detecting shared genetics and/or molecular pathways between comorbid diseases can help us understand exactly how the etiology of the diseases overlap
- Etiologic overlaps:
  - provide opportunities for novel interventions that prevent or treat the comorbidity, rather than preventing/treating each disease separately
  - facilitate drug repurposing (that is, known drugs targeting a pleiotropic locus may be repurposed to treat other diseases controlled by that locus, precluding the need for the development and testing of a brand-new drug)

Abundant Pleiotropy in Human Complex Diseases and Traits

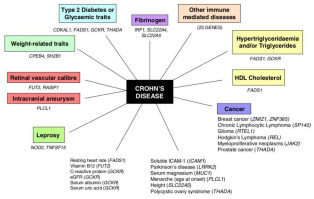
Shanya Sivakumaran,[1,6] Felix Agakov,[1,2,6] Evropi Theodoratou,[1,6] James G. Prendergast,[3] Lina Zgaga,[1,4] Teri Manolio,[5] Igor Rudan,[1] Paul McKeigue,[1] James F. Wilson,[1] and Harry Campbell[1,*]

The American Journal of Human Genetics *89*, 607–618, November 11, 2011

Table 6. Extent of Pleiotropy in Different Disease Classes

	Genes			SNPs		
Disease Class	**Pleiotropic (%)**	**Nonpleiotropic (%)**	**p Value[a]**	**Pleiotropic (%)**	**Nonpleiotropic (%)**	**p Value[a]**
All (comparison group)	233 (16.9)	1147 (83.1)	–	77 (4.6)	1610 (95.4)	–
Immune-mediated phenotypes	106 (37.7)	175 (62.3)	<0.0001	31 (8.3)	343 (91.7)	0.0066
Cancer	49 (34.8)	92 (65.2)	<0.0001	8 (4.8)	158 (95.2)	0.8456
Metabolic syndrome	79 (28.5)	198 (71.5)	<0.0001	30 (8.4)	327 (91.6)	0.0056

[a] Fisher's exact test p value.



Pleiotropy in gene mapping

- Mapping a single genotype to multiple phenotypes has the potential to uncover novel links between traits or diseases
- It can also offer insights into the mechanistic underpinnings of known comorbidities
- It can increase power to detect novel associations with one or more phenotypes

A practitioners' guide for studying pleiotropy in genetic epi studies

Am J Epidemiol. 2017 Aug 11. doi: 10.1093/aje/kwx296. [Epub ahead of print]

Statistical Analysis of Multiple Phenotypes in Genetic Epidemiological Studies:From Cross-Phenotype Associations to Pleiotropy.

Salinas YD, Wang Z, DeWan AT.

Abstract

In the context of genetics, pleiotropy refers to the phenomenon in which a single genetic locus affects more than one trait or disease. Genetic epidemiological studies have identified loci associated with multiple phenotypes, and these cross-phenotype associations are often incorrectly interpreted as examples of pleiotropy. Pleiotropy is only one possible explanation for cross-phenotype associations. Cross-phenotype associations may also arise due to issues related to study design, confounder bias, or non-genetic causal links between the phenotypes under analysis. Therefore, it is necessary to dissect cross-phenotype associations carefully to uncover true pleiotropic loci. In this review, we describe statistical methods that can be used to identify robust statistical evidence of pleiotropy. First, we provide an overview of univariate and multivariate methods for discovery of cross-phenotype associations and highlight important considerations for choosing among available methods. Then, we describe how to dissect cross-phenotype associations by using mediation analysis. Pleiotropic loci provide insights into the mechanistic underpinnings of disease comorbidity, and may serve as novel targets for interventions that simultaneously treat multiple diseases. Discerning between different types of cross-phenotype associations is necessary to realize the public health potential of pleiotropic loci.

KEYWORDS: genetic epidemiology; mediation analysis; pleiotropy

13

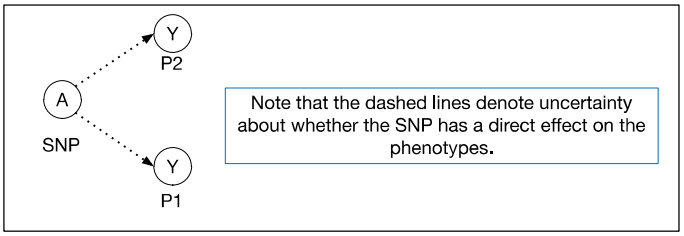
Guidelines for generating robust statistical evidence of pleiotropy

Discover CP associations ⇨ **Dissect** CP associations ⇨ **Classify** them as examples of biological, mediated, or spurious pleiotropy

14

Cross-phenotype (CP) associations

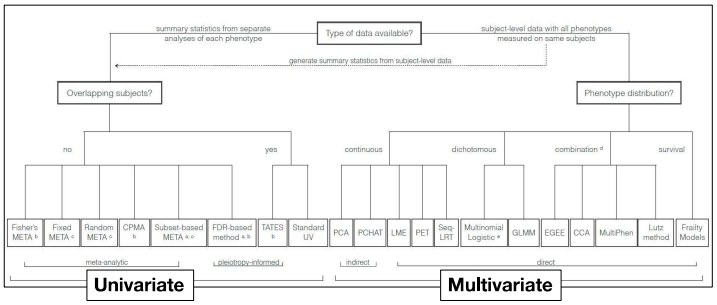
Statistical associations between a **single genetic locus** – a single gene or a single variant within a gene – and **multiple phenotypes**



Note that the dashed lines denote uncertainty about whether the SNP has a direct effect on the phenotypes.

15

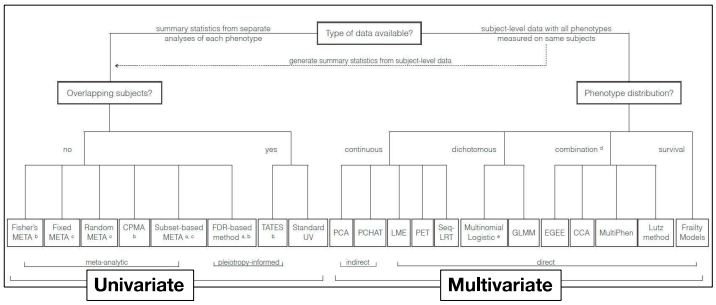
Analytic options for discovery of CP associations



Key distinction:
- Univariate methods examine the association between a given SNP and each trait *separately*
- Multivariate methods examine the association between a given SNP and each trait by modeling the traits *jointly*

16

Analytic options for discovery of CP associations



Choice between univariate and multivariate approaches depends on:
- Types of data available on our phenotypes of interest
  - Summary statistics vs. individual-level data?
  - Are the phenotypes measured on the same subjects?
- Distribution of the phenotypes (e.g., quantitative or disease trait)

Univariate methods are by far the most commonly used to detect CP associations

- Univariate methods include (but are not limited to) the methods you've discussed in class so far:
  - allelic Chi-Square test
  - genotypic Chi-Square test
  - regression-based methods
- The overall approach is to:
  - obtain univariate association p-values for each phenotype
  - declare CP associations at genetic loci that are statistically significantly associated with each phenotype

Hypothetical example: Discovery of CP associations for hypertension and heart disease by using logistic regression

Step 1. Fit two univariate regression models within PLINK

$$E[hypertension] = \beta_0 + \boldsymbol{\beta_1} * SNP$$
$$E[heart\ disease] = \beta_0 + \boldsymbol{\beta_1} * SNP$$

Word of caution: The univariate tests of association should be marginal tests (conducted irrespectively of the second phenotype) NOT conditional tests (conducted on a subset defined based on absence/presence of the second phenotype). In this example, what that means is that the regression for hypertension should be fit on all subjects *irrespectively* of their heart disease status; and the regression for heart disease should be fit on all subjects *irrespectively* of their hypertension status. More on this later!

Hypothetical example: Discovery of CP associations for hypertension and heart disease by using logistic regression

Step 1. Fit two univariate regression models within PLINK

$$E[hypertension] = \beta_0 + \boldsymbol{\beta_1} * SNP$$

$$E[heart\ disease] = \beta_0 + \boldsymbol{\beta_1} * SNP$$

Step 2. For a given SNP, examine p-values for $\boldsymbol{\beta_1}$ from <u>each</u> model.

- P-value for $\boldsymbol{\beta_1}$ in hypertension model = 1.03×10^{-12}
- P-value for $\boldsymbol{\beta_1}$ in heart disease model = 6.02×10^{-9}

Step 3. Declare CP associations at a given SNP, if the p-values for $\boldsymbol{\beta_1}$ in <u>each</u> model surpass the study significance threshold.

- Assuming the standard GWAS significance threshold (alpha=5×10^{-8}), there is a statistically significant association with both hypertension and heart disease at this particular SNP. Therefore, we have sufficient statistical evidence to declare a CP association at this SNP.

Using multivariate methods to increase the power to detect cross-phenotype associations

A Comparison of Multivariate Genome-Wide Association Methods

Tessel E. Galesloot[1], Kristel van Steen[2,3], Lambertus A. L. M. Kiemeney[1,4], Luc L. Janss[5]*, Sita H. Vermeulen[1,6]*

1 Department for Health Evidence, Radboud university medical center, Nijmegen, The Netherlands, **2** Systems and Modeling Unit, Montefiore Institute, University of Liège, Liège, Belgium, **3** Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium, **4** Department of Urology, Radboud university medical center, Nijmegen, The Netherlands, **5** Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark, **6** Department of Human Genetics, Radboud university medical center,

PLOS ONE | www.plosone.org 1 April 2014 | Volume 9 | Issue 4 | e95923

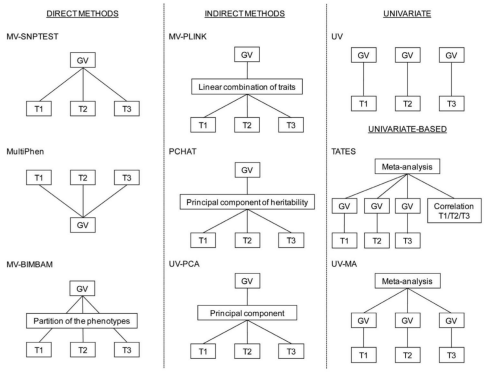
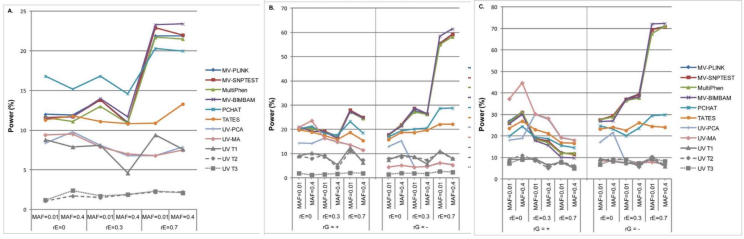


Table 1. Simulation scenarios.

# traits associated with QTL	Heritability (h^2_j)	Effect size (a_j)	rG	rE	MAF (q)
1	$h^2_1 = 0.1\%$, $h^2_2 = h^2_3 = 0$	$a_1 > 0$, $a_2 = a_3 = 0$	0	$3\times0/3\times0.3/3\times0.7$	0.01/0.4
2	$h^2_1 = h^2_2 = 0.1\%$, $h^2_3 = 0$	$a_1 = a_2$, $a_3 = 0$	+	$3\times0/3\times0.3/3\times0.7$	0.01/0.4
		$-a_1 = a_2$, $a_3 = 0$	−	$3\times0/3\times0.3/3\times0.7$	0.01/0.4
3	$h^2_1 = h^2_2 = h^2_3 = 0.1\%$	$a_1 = a_2 = a_3$	+	$3\times0/3\times0.3/3\times0.7$	0.01/0.4
	$h^2_1 = h^2_2 = h^2_3 = 0.1\%$	$-a_1 = a_2 = a_3$	−	$3\times0/3\times0.3/3\times0.7$	0.01/0.4

MAF indicates minor allele frequency; j, trait; QTL, quantitative trait locus; rE, residual correlation; rG, genetic correlation.
doi:10.1371/journal.pone.0095923.t001



Simulation scenarios

A comparison of univariate and multivariate GWAS methods for analysis of multiple dichotomous phenotypes

Yasmmyn D. Salinas[1], Andrew T. DeWan[1], and Zuoheng Wang[2]

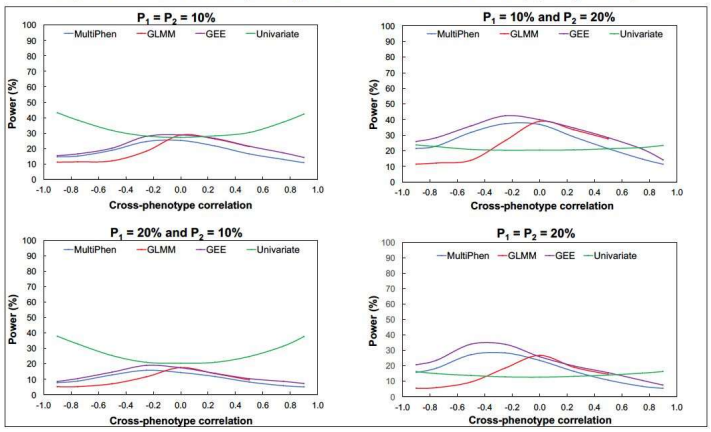
[1] Department of Chronic Disease Epidemiology; [2] Department of Biostatistics, Yale School of Public Health, Yale University, 60 College St, New Haven, Connecticut, USA

# traits associated	h_j^2	$r_{Y1,Y2}$	P_j
1	h_1^2=0.1%,h_2^2=0%	[-0.9,0.9]	P1 = P2 = 10%
			P1 = P2 = 20%
			P1 = 10%, P2 = 20%
			P1 = 20%, P2 = 10%
2	$h_1^2 = h_2^2$= 0.1%	[-0.9,0.9]	P1 = P2 = 10%
			P1 = P2 = 20%
			P1 = 10%, P2 = 20%
			P1 = 20%, P2 = 10%
2	h_1^2 = 0.1%,h_2^2 = 0.05%	[-0.9,0.9]	P1 = P2 = 10%
			P1 = P2 = 20%
			P1 = 10%, P2 = 20%
			P1 = 20%, P2 = 10%

Genet. Epidemiol. 41 (7), 689-689

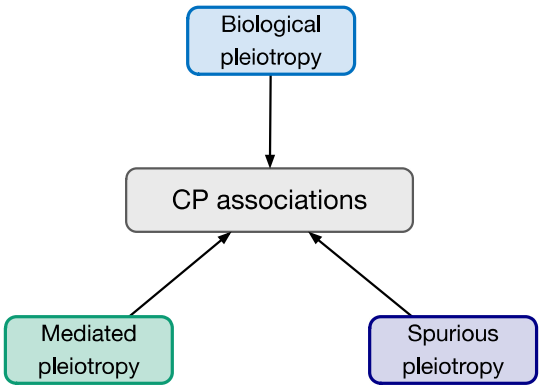
PLEIOTROPY PRESENT
equal effect sizes

Figure 2. Power when both phenotypes are associated with the SNP ($h_1^2 = h_2^2 = 0.1\%$) [a]



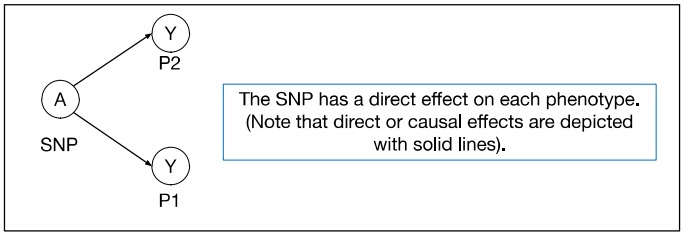
[a] Results for GLMMs are shown for $r_{Y1,Y2} \leq 0.5$ only, since the models experienced convergence issues for $r_{Y1,Y2} > 0.5$.

Problem: CP associations need not be indicative of pleiotropy



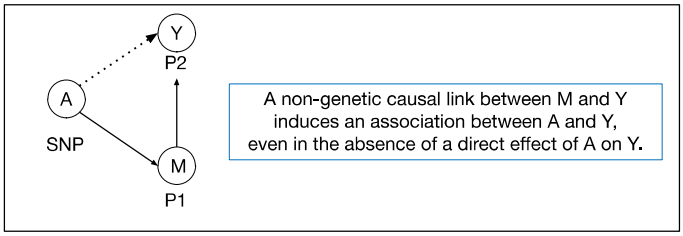
Biological pleiotropy

Independent associations between a genetic locus (A) and multiple phenotypic outcomes (Y)



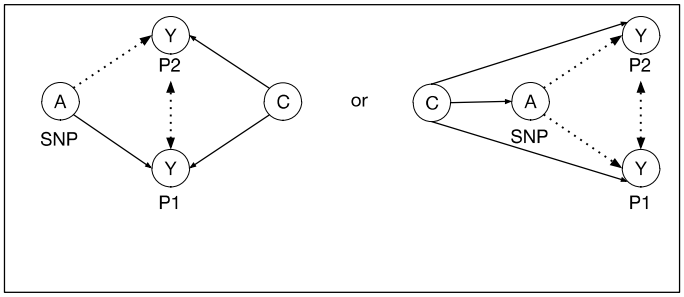
Mediated pleiotropy

Association between a genetic locus (A) and an intermediate phenotype (M) that causes a second phenotypic outcome (Y)



Spurious pleiotropy

Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes

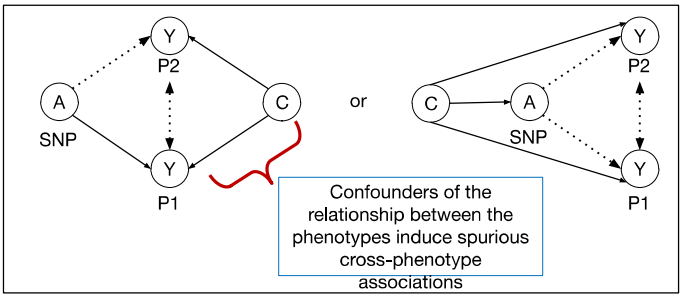


*Linkage disequilibrium is the non-random co-segregation of alleles.

31

Spurious pleiotropy

Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes

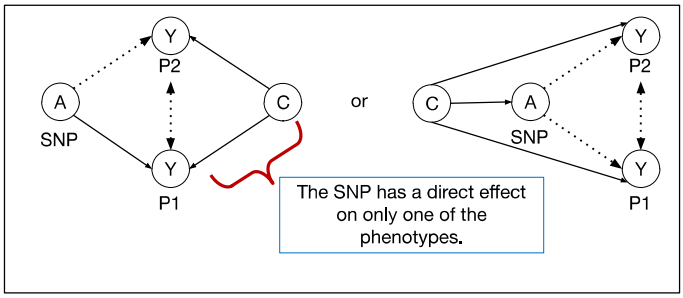


*Linkage disequilibrium is the non-random co-segregation of alleles.

32

Spurious pleiotropy

Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes

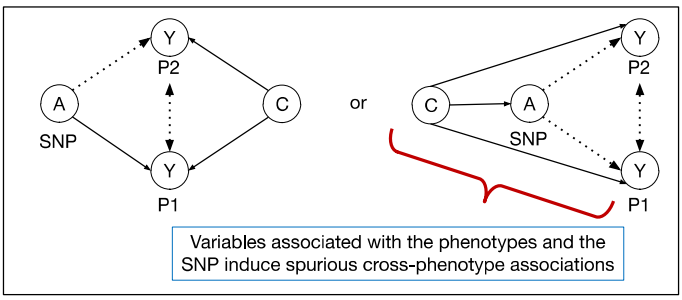


*Linkage disequilibrium is the non-random co-segregation of alleles.

33

Spurious pleiotropy

Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes

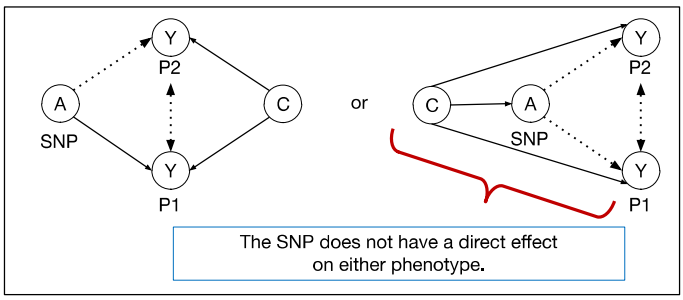


*Linkage disequilibrium is the non-random co-segregation of alleles.

34

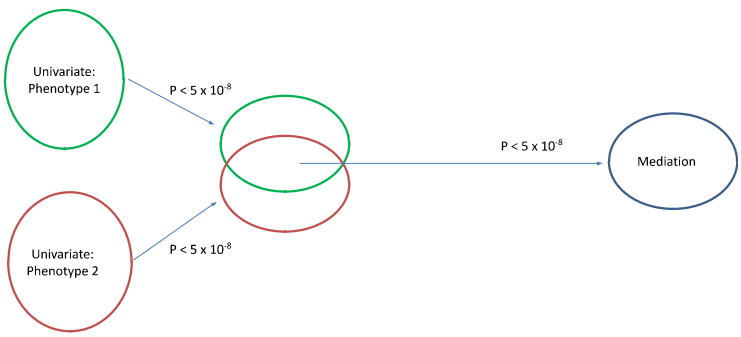
Spurious pleiotropy

Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes

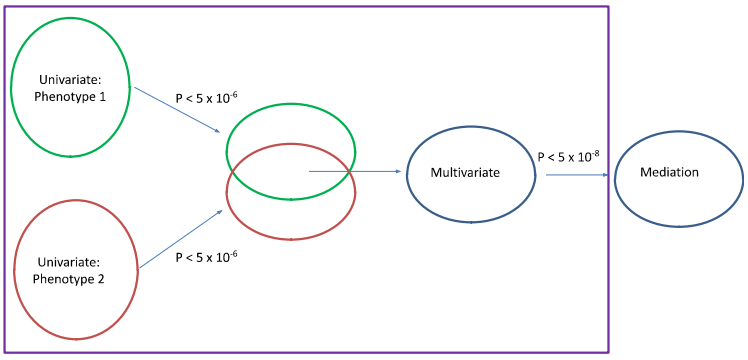


*Linkage disequilibrium is the non-random co-segregation of alleles.

Pleiotropy exercise (Parts 1 and 2)



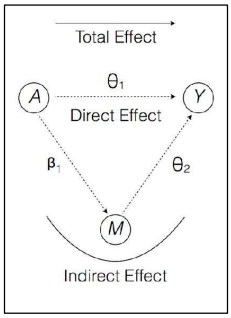
Pleiotropy exercise (Parts 1 and 2)



Guidelines for generating robust statistical evidence of pleiotropy

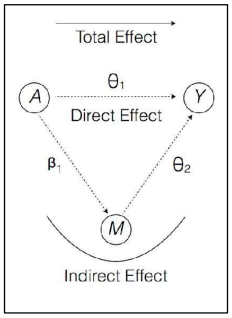
Mediation analysis provides a tool for dissecting CP associations

- Mediation analysis decomposes the **total effect** of the SNP (*A*) on a phenotypic outcome (*Y*) into:
  - **Direct effect:** effect of *A* on *Y* that occurs independently of an intermediate phenotype (*M*)
  - **Indirect effect:** effect of *A* on *Y* that occurs through the intermediate phenotype *M*



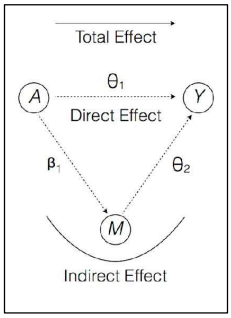
Mediation analysis: Data requirements

- All phenotypes must be measured on the same subjects
- Temporality must be ascertained
  - The occurrence of the intermediate variable *M* must precede that of the phenotypic outcome variable *Y*



Mediation analysis: Assumptions

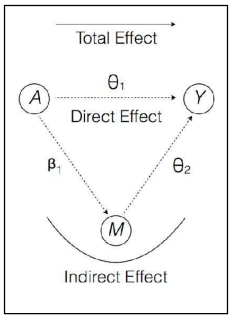
- There must be no unmeasured:
  - confounders of the total effect
  - confounders of the relationship between SNP *A* and the mediator *M*
  - confounders of the relationship between mediator *M* and phenotypic outcome *Y*



Mediation analysis: Assumptions

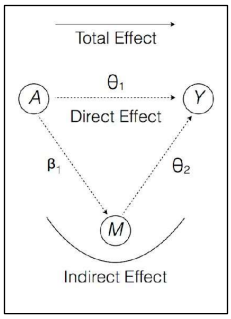
Typically met in genetic epi studies!

- There must be no unmeasured:
  - confounders of the total effect
  - confounders of the relationship between SNP *A* and the mediator *M*
  - confounders of the relationship between mediator *M* and phenotypic outcome *Y*



Mediation analysis: Assumptions

- There must be no unmeasured:
  - confounders of the total effect
  - confounders of the relationship between SNP *A* and the mediator *M*
  - confounders of the relationship between mediator *M* and phenotypic outcome *Y*

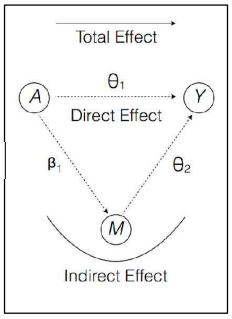


Requires adjustment for <u>known</u> confounders to prevent bias (<u>Note</u>: this effectively restricts the use of mediation analyses to datasets in which data on such variables have been collected)

Mediation analysis: Regression-based approach

- Requires fitting two regression models, one for mediator *M* and one for phenotypic outcome *Y:*
  - $E[M\,|a,c] = \beta_0 + \boldsymbol{\beta_1} a + \beta_2' c$
  - $E[Y\,|a,m,c\,] = \theta_0 + \boldsymbol{\theta_1} a + \boldsymbol{\theta_2} m + \theta_4' c$

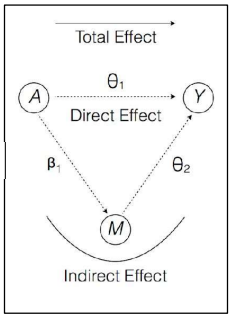
Assesses the effect of *A* on *M*, while controlling for measured confounders *(C)*



Mediation analysis: Regression-based approach

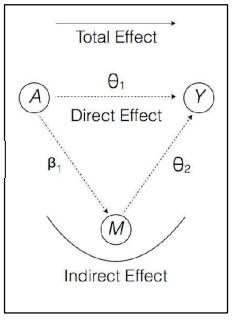
- Requires fitting two regression models, one for mediator *M* and one for phenotypic outcome *Y:*
  - $E[M\,|a,c] = \beta_0 + \boldsymbol{\beta_1} a + \beta_2' c$
  - $E[Y\,|a,m,c\,] = \theta_0 + \boldsymbol{\theta_1} a + \boldsymbol{\theta_2} m + \theta_4' c$

Assesses the effect of *A* on *Y*, while controlling for both *M* and *C*



Mediation analysis: Regression-based approach

- Requires fitting two regression models, one for mediator *M* and one for phenotypic outcome *Y:*
  - $E[M\,|a,c] = \beta_0 + \boldsymbol{\beta_1} a + \beta_2' c$
  - $E[Y\,|a,m,c\,] = \theta_0 + \boldsymbol{\theta_1} a + \boldsymbol{\theta_2} m + \theta_4' c$
- The parameter estimates from these models (**namely $\boldsymbol{\beta_1}$, $\boldsymbol{\theta_1}$, and $\boldsymbol{\theta_2}$**) are used to estimate the direct and indirect effects

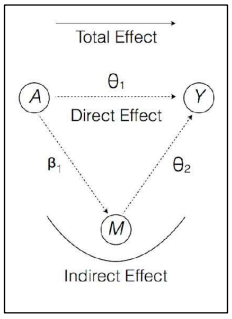


Guidelines for generating robust statistical evidence of pleiotropy

Discover CP associations → **Dissect** CP associations → **Classify** them as examples of biological, mediated, or spurious pleiotropy

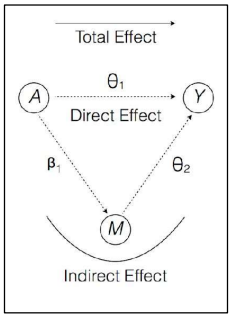
Mediation analysis: Interpretation

- **Mediated pleiotropy**
  - Complete mediation: SNP *A* is associated with mediator *M* and the total effect of *A* on phenotypic outcome *Y* is equal to its indirect effect (i.e., the direct effect is equal to 0).
  - Incomplete mediation: SNP *A* is associated with mediator *M* and *A* has both direct and indirect effects on phenotypic outcome *Y* (i.e., the total effect is equal to the sum of the direct and indirect effects)
- **Biological pleiotropy**
  - SNP A is associated with mediator M, and the total effect of SNP A on phenotypic outcome Y is equal to its direct effect (i.e., the indirect effect is equal to 0)



Mediation analysis: Interpretation

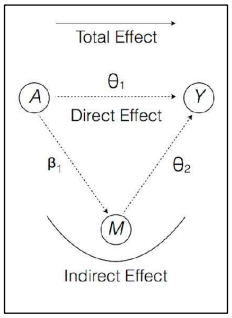
- **Mediated pleiotropy**
  - Complete mediation: SNP *A* is associated with mediator *M* and the total effect of *A* on phenotypic outcome *Y* is equal to its indirect effect (i.e., the direct effect is equal to 0).
- **Biological pleiotropy**
  - SNP A is associated with mediator M, and the total effect of SNP A on phenotypic outcome Y is equal to its direct effect (i.e., the indirect effect is equal to 0)
  - Incomplete mediation: SNP *A* is associated with mediator *M* and *A* has both direct and indirect effects on phenotypic outcome *Y* (i.e., the total effect is equal to the sum of the direct and indirect effects)



49

Mediation analysis: Interpretation

- **Spurious pleiotropy**
  - SNP A is not associated with mediator M after controlling for measured confounders



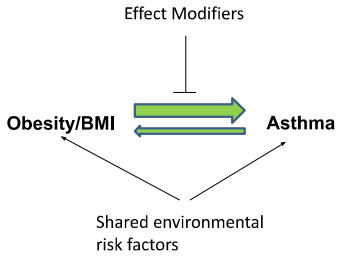
mediation R package

```
> med.fit<-glm(W1~rs1_2, data=combined, family=binomial("logit"))
> out.fit<-glm(W2~W1+rs1_2, data=combined, family=binomial("logit"))
> med.out<-mediate(med.fit,out.fit, treat="rs1_2", mediator="W1", boot=TRUE, boot.ci.type="bca", sims=1000)
> summary(med.out)

Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the BCa Method

                              Estimate 95% CI Lower 95% CI Upper p-value
ACME (control)                 0.02152      0.01823         0.03  <2e-16 ***
ACME (treated)                 0.02199      0.01868         0.03  <2e-16 ***
ADE (control)                  0.00723      0.00415         0.01  <2e-16 ***
ADE (treated)                  0.00771      0.00443         0.01  <2e-16 ***
Total Effect                   0.02922      0.02461         0.03  <2e-16 ***
Prop. Mediated (control)       0.73634      0.65429         0.84  <2e-16 ***
Prop. Mediated (treated)       0.75247      0.67272         0.85  <2e-16 ***
ACME (average)                 0.02175      0.01847         0.03  <2e-16 ***
ADE (average)                  0.00747      0.00426         0.01  <2e-16 ***
Prop. Mediated (average)       0.74441      0.66254         0.84  <2e-16 ***
```

51

Empirical searches for pleiotropic loci for asthma and comorbidities

Asthma-obesity comorbidity



Ford ES. The epidemiology of obesity and asthma. J Allergy Clin Immunol. 2005;115(5):897-909; quiz 10.
Stukus DR. Obesity and asthma: The chicken or the egg? J Allergy Clin Immunol. 2014.
Kim SH, Sutherland ER, Gelfand EW. Is there a link between obesity and asthma? Allergy Asthma Immunol Res. 2014;6(3):189-95.
Egan KB, Ettinger AS, DeWan AT, Holford TR, Holmen TL, Bracken MB. Longitudinal associations between asthma and general and abdominal weight status among Norwegian adolescents and young adults: the HUNT Study. Pediatric obesity. 2014.

53

Discovery and Mediation Analysis of Cross-Phenotype Associations Between Asthma and Body Mass Index in 12q13.2

Yasmmyn D. Salinas*, Zuoheng Wang, and Andrew T. DeWan

* Correspondence to Dr. Yasmmyn D. Salinas, Department of Chronic Disease Epidemiology, Yale School of Public Health, 60 College Street, New Haven, CT 06520 (e-mail: yasmmyn.salinas@yale.edu).

Am J Epidemiol. 2021;190(1):85–94

54

Study population

- N = 305,945 White, British subjects from the UK Biobank (a population-based prospective cohort study of > 500,000 subjects, aged 40-69 years at baseline)

Study design

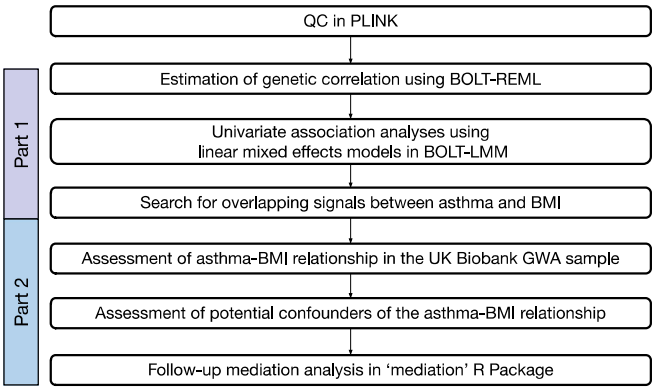
- Two parts:
  - Genome-wide search for cross-phenotype associations with asthma and body mass index
  - Follow-up mediation analysis to dissect genome-wide significant CP associations

biobank uk
Improving the health of future generations

Phenotype definitions

- BMI at baseline (kg/m^2):
  - calculated based on height and weight measurements collected by trained UK Biobank staff at the recruitment sites
- Asthma diagnosed prior to baseline (yes/no):
  - ascertained via the question "Has a doctor ever told you that you had asthma?"
  - Note: In mediation analyses, two subgroups were created based on age-at-diagnosis

Statistical Methods



Overlap in GWA signals

Association with BMI among the 1,457 SNPs with genome-wide significant p-values for asthma

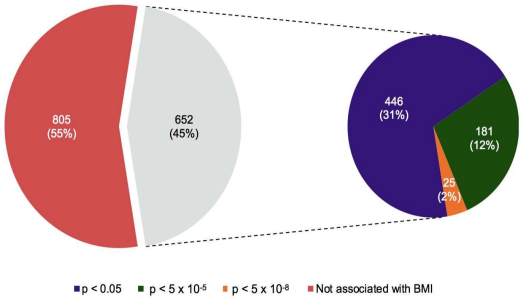


Figure 1. Overlap in GWA signals between asthma and BMI. Results for asthma are for the analysis of all asthmatic subjects (35,373 asthmatics vs. 270,572 non-asthmatics). Results for BMI are for the quantitative BMI analysis (n=305,945). Both analyses are sex- and age-adjusted. The threshold for genome-wide significance was alpha=5x10^{-8}.

Overlap in GWA signals

Association with asthma among the 1,699 SNPs with genome-wide significant p-values for BMI

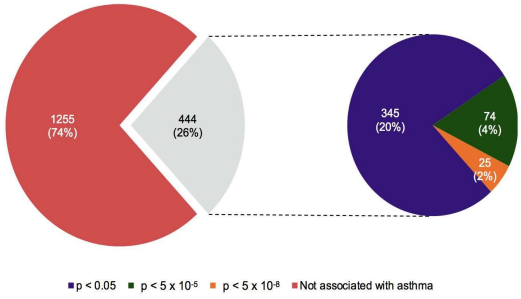
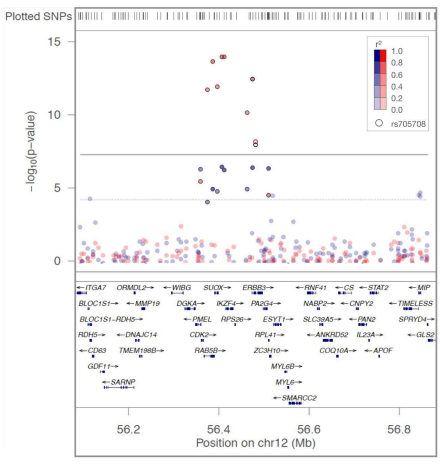


Figure 1. Overlap in GWA signals between asthma and BMI. Results for asthma are for the analysis of all asthmatic subjects (35,373 asthmatics vs. 270,572 non-asthmatics). Results for BMI are for the quantitative BMI analysis (n=305,945). Both analyses are sex- and age-adjusted. The threshold for genome-wide significance was alpha=5x10^{-8}.

Regional plot around rs705708 for BMI (blue) and asthma (red)



Cross-phenotype associations in 12q13.2

Table 2. Cross-phenotype associations in 12q13.2 [a]

					Asthma		BMI[c]	
SNP	Gene	BP	Effect/reference allele	EAF	OR (95% CI)	P[c]	beta (95% CI)	P[d]
rs2069408	*CDK2*	56,364,321	G/A	0.3388	1.04 (1.02, 1.06)	3.30×10^{-6}	-0.06 (-0.08, -0.04)	5.40×10^{-7}
rs1873914	*RAB5*	56,379,427	C/G	0.4237	1.06 (1.04, 1.08)	2.40×10^{-12}	-0.05 (-0.07, -0.02)	7.90×10^{-5}
rs705702 [b]	*SUOX*	56,390,636	G/A	0.3376	1.07 (1.05, 1.09)	3.10×10^{-14}	-0.05 (-0.08, -0.03)	1.10×10^{-5}
rs10876864 [b]	*SUOX*	56,401,085	G/A	0.4279	1.06 (1.04, 1.08)	1.50×10^{-12}	-0.05 (-0.07, -0.03)	1.60×10^{-5}
rs1701704	*IKZF4*	56,412,487	G/T	0.3433	1.07 (1.05, 1.09)	1.50×10^{-14}	-0.06 (-0.09, -0.04)	3.70×10^{-7}
rs2456973	*IKZF4*	56,416,928	C/A	0.3432	1.07 (1.05, 1.09)	1.50×10^{-14}	-0.06 (-0.08, -0.04)	6.00×10^{-7}
rs11171739 [b]	*ERBB3*	56,470,625	C/T	0.4337	1.06 (1.04, 1.07)	8.80×10^{-11}	-0.05 (-0.07, -0.03)	1.10×10^{-5}
rs2292239	*ERBB3*	56,482,180	T/G	0.3470	1.07 (1.05, 1.08)	4.50×10^{-13}	-0.06 (-0.08, -0.04)	4.20×10^{-7}
rs705708	*ERBB3*	56,488,913	A/G	0.4712	1.05 (1.03, 1.07)	7.20×10^{-9}	-0.06 (-0.09, -0.04)	1.30×10^{-8}
rs11171747 [b]	*ESYT1*	56518408	T/G	0.6180	1.04 (1.02, 1.05)	2.90×10^{-5}	-0.06 (-0.08, -0.04)	4.50×10^{-7}

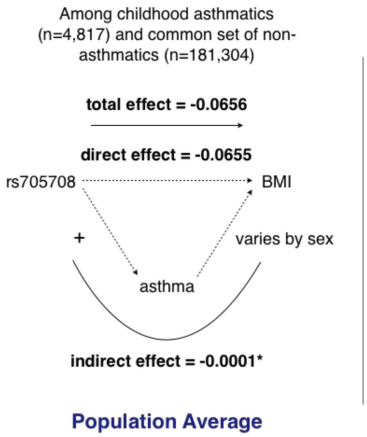
Abbreviations: BP = base-pair ; BMI = body mass index; CI = confidence interval; EAF = effect allele frequency; OR = odds ratio; SNP = single-nucleotide polymorphism

a. Results shown for SNPs with p < 5×10^{-8} for asthma and p < 0.05 for BMI.
b. For intergenic SNPs, the nearest gene is listed, with priority given to genes directly downstream of variant.
c. P-value from BOLT-LMM, derived using the standard "infinitesimal" mixed model.
d. P-value from BOLT-LMM, derived using the Gaussian mixture model.

61

Decomposing the effect of rs705708 on BMI via mediation analysis

62



Note: Effect estimates shown are adjusted for common determinants of asthma and BMI: age, sex, breast-feeding status, exposure to maternal smoking, and smoking status at asthma diagnosis (adult analyses only). Unless otherwise noted by an asterisk(*), all paths are significant at the 0.05 level.

63

Conclusions

- rs705708 has a positive direct effect on asthma
  - Stronger in magnitude for childhood asthma
- rs705708 has a negative direct effect on BMI
  - Consistent in magnitude and direction in analyses including childhood vs. adult asthmatics
- This suggests that locus 12q13.2, tagged by rs705708, has pleiotropic effects on asthma and BMI.

64

Conclusions

- 12q13.2 is multigenic and our CP associations span genes *CDK2, RAB5*, *SUOX*, *IZK4*, *RPS26*, *ERBB3*, and *ESYT1*.
  - rs705708 is the top regional BMI signal and resides in *ERBB3*.
  - The top regional asthma signal, rs2456973, resides in *IZKF4*.
  - While rs705708 and rs2456973 could be in LD with the same causative variant in either *ERBB3* or *IKZF4* or another gene in 12q13.2, it is also possible that each variant could tag a distinct, trait-specific causative variant in different genes.
- Therefore, locus 12q13.2 displays pleiotropic effects on asthma and BMI, but this may not be an example of pleiotropy at the gene level (biological pleiotropy).

What if we expand this investigation to look at more phenotypes correlated with asthma?

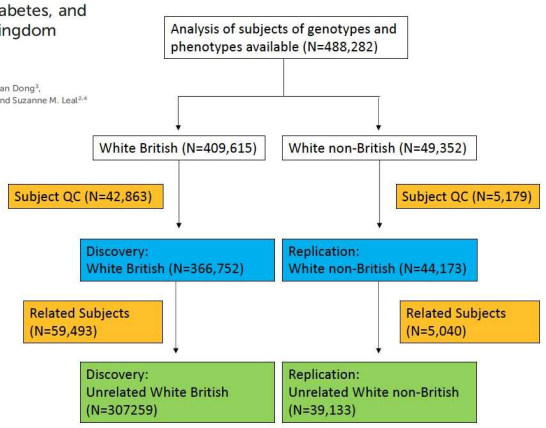
Asthma, T2D and anthropometric measures

- Obesity is a well-established risk factor for both asthma and T2D.
  - While highly correlated, waist circumference (WC) can provide distinct information on adiposity as it is a measure of visceral obesity, specifically WC adjusted for BMI. WC is often used in studies of chronic diseases.
  - Increased WC has been shown to be an additional risk factor for T2D and asthma even after adjusting for BMI
- Elevated blood glucose and T2D have been linked to increased risk of asthma in adults, and conversely, asthma has been associated with increased risk of developing T2D in adults.
- Height is a highly heritable polygenic trait; there is evidence that shorter individuals have an increased risk for developing T2D and individuals with childhood onset asthma have shorter stature as adults compared to non-asthmatics

Variants in *JAZF1* are associated with asthma, type 2 diabetes, and height in the United Kingdom biobank population

Andrew T. DeWan[1*†], Megan E. Cahill[1†], Diana M. Cornejo-Sanchez[2], Yining Li[1], Zihan Dong[3], Tabassum Fabiha[3], Hao Sun[3], Gao Wang[2] and Suzanne M. Leal[2,4]

10.3389/fgene.2023.1129389



biobank uk
Improving the health of future generations

Discovery

	# variants	# variants removed in this step	# subjects	# subjects removed in this step
Step 1: Initial variant QC				
Genotyped variants	805,426			
Autosomal variants	784,256	21,170		
Covered by both arrays	733,322	50,934		
Batch level qc	687,004	46,318		
SNPs only (indels removed)	674,489	12,515		
Step 2: Subject QC[1]				
Genotypes available			488,377	
Phenotypes available			488,282	95
Genetic and reported sex match			487,910	372
Sex chromosomes non-XX XY			487,440	470
Outliers in heterozygosity/missing rate			486,477	963
"Caucasian" (f.22006)			408,186	78,291
Individual call rate > 99%			366,752	41,434
Unrelated[2]			307,259	59,493
Step 3: Final variant QC2				
call rate > 99%	639,852	34,627		
HWE p<5x10-8	631,088	29,843		
MAF > 1%	550,062	80,995		
Total	**550,062**		**366,752**	

[1]Subject QC was performed using the 674,489 variants

[2]Unrelated subjects were used for final variant QC and then the set of variants were selected for the full set of subjects

Replication

	# variants	# variants removed in this step	# subjects	# subjects removed in this step
Step 1: Initial variant QC				
Genotyped variants	805,426			
Autosomal variants	784,256	21,170		
Covered by both arrays	733,322	50,934		
Batch level qc	687,004	46,318		
SNPs only (indels removed)	674,489	12,515		
Step 2: Subject QC[1]				
Mahalanobis defined non-British white			49,352	
Genetic and reported sex match			49,307	45
Sex chromosomes non-XX XY			49,265	42
Outliers in heterozygosity/missing rate			49,133	132
Individual call rate > 99%			44,173	4,960
Unrelated[2]			39,133	5,040
Step 3: Final variant QC2				
call rate > 99%	639,847	34,642		
HWE p<5x10-8	631,089	8758		
MAF > 1%	550,028	81,061		
Total	**550,028**		**44,173**	

[1]Subject QC was performed using the 674,489 variants

[2]Unrelated subjects were used for final variant QC and then the set of variants were selected for the full set of subjects

Genotypes were imputed by UK Biobank and used in the analysis; INFO > 0.8 and MAF > 0.001
N=13,407,279 (Discovery) N=13,793,916 (Replication)

Phenotypes

- Asthma: defined by either ICD-10 code (field 41270, J45 or J46) or self-reported diagnosis by a doctor (field 6152).
- T2D: defined by either ICD-10 code (UK Biobank field 41270, code E11) or self-reported diagnosis by a doctor at ≥ 30 years of age (fields 2443 and 2976). Individuals with type 1 diabetes [self-reported diabetes that occurred < 30 years of age or E10] or gestational diabetes [self-report (field 4011) or O24] were excluded from both cases and controls.
- Anthropometric measurements:
  - Waist circumference (WC), adjusted and unadjusted for BMI
  - Height
  - Weight
  - BMI
  - When used as outcomes, WC, BMI, height, and weight were transformed using rank-based inverse normal transformation as implemented in R

Discovery and replication sample demographics

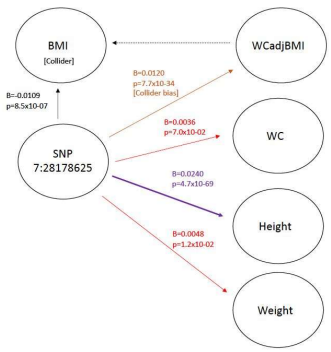
		Asthma Cases	Asthma Controls	T2D Cases	T2D Controls	WC (cm)	BMI (kg/m^2)	Height (cm)	Weight (kg)
Discovery	N/mean (sd)	48,623	290,722	22,670	313,404	90.33 (14.46)	27.42 (4.76)	168.71 (9.23)	78.26 (15.87)
	Sex (N [%] male)	20,483 (42.13)	135,525 (46.62)	14,202 (62.65)	141,147 (45.04)	167,655 (45.87)	167,654 (45.87)	167,854 (45.87)	167,748 (45.87)
	Age at recruitment (mean [sd])	56.48 (8.18)	56.89 (7.98)	60.47 (6.68)	56.60 (8.04)	56.91 (8.00)	56.92 (8.00)	56.92 (8.00)	56.92 (8.00)
Replication	N/mean (sd)	5,822	36,615	2,625	39,324	89.74 (13.75)	27.28 (4.92)	168.36 (9.21)	77.56 (16.25)
	Sex (N [%] male)	2,454 (42.15)	16,252 (44.39)	1,610 (61.33)	16,936 (43.07)	19,379 (44.09)	19,380 (44.08)	19,391 (44.06)	19,389 (44.09)
	Age at recruitment (mean [sd])	55.26 (8.30)	55.40 (8.16)	59.34 (7.06)	55.13 (8.17)	55.44 (8.17)	55.44 (8.17)	55.44 (8.17)	55.44 (8.18)

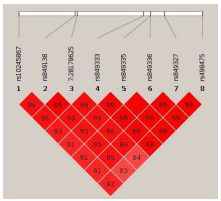
Variants in *JAZF1* with genome-wide significant associations with asthma, T2D and at least one anthropometric measure

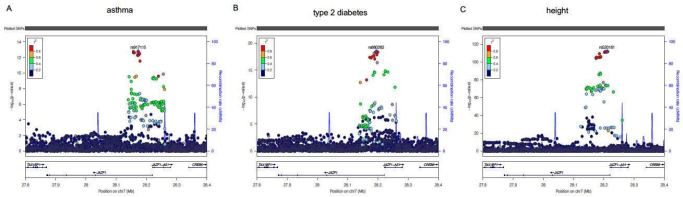
					Asthma				T2D				WCadjBMI				WC (unadjusted)			
					Discovery		Replication		Discovery		Replication		Discovery		Replication		Discovery		Replication	
Variant	BP[1]	EA[2]	EAF[3]	INFO[4]	Beta	P-value	Beta	P-value	Beta	P-value	Beta	P-value	Beta	P-value	Beta	P-value	Beta	P-value	Beta	P-value
rs10245867	28142186	T	0.329	0.995	0.0471	$2.8x10^{-10}$	0.0642	$2.7x10^{-03}$	0.0635	$1.9x10^{-06}$	0.096	$2.2x10^{-03}$	0.0058	$3.4x10^{-04}$	0.0122	$6.3x10^{-05}$	-0.0006	$7.8x10^{-01}$	0.0089	$1.5x10^{-01}$
rs849138	28177338	G	0.493	0.998	0.0387	$4.0x10^{-08}$	0.0398	$5.1x10^{-02}$	0.0863	$5.2x10^{-18}$	0.0906	$2.3x10^{-03}$	0.0119	$1.3x10^{-02}$	0.0192	$3.2x10^{-11}$	0.0036	$6.8x10^{-02}$	0.0157	$6.6x10^{-03}$
7:28178625	28178625	TG	0.493	0.997	0.0394	$2.3x10^{-08}$	0.0410	$4.4x10^{-02}$	0.0860	$6.7x10^{-18}$	0.092	$1.9x10^{-03}$	0.0120	$7.7x10^{-04}$	0.0188	$8.2x10^{-11}$	0.0036	$7.0x10^{-02}$	0.0151	$9.3x10^{-03}$
rs849333	28219812	A	0.342	0.996	0.0461	$4.5x10^{-10}$	0.0502	$1.9x10^{-02}$	0.0813	$7.2x10^{-15}$	0.077	$1.4x10^{-02}$	0.0067	$1.7x10^{-01}$	0.0115	$1.5x10^{-04}$	0.0000	$9.8x10^{-01}$	0.0081	$1.8x10^{-01}$
rs849335	28223990	T	0.341	N/A	0.0464	$3.3x10^{-10}$	0.0519	$1.5x10^{-02}$	0.0828	$2.0x10^{-15}$	0.076	$1.5x10^{-02}$	0.0069	$4.3x10^{-11}$	0.0105	$4.9x10^{-04}$	0.0004	$8.4x10^{-01}$	0.0064	$2.9x10^{-01}$
rs849336	28224053	A	0.341	1	0.0458	$5.5x10^{-10}$	0.0531	$1.3x10^{-02}$	0.0831	$1.6x10^{-15}$	0.076	$1.5x10^{-02}$	0.0069	$4.2x10^{-11}$	0.0105	$4.9x10^{-04}$	0.0005	$8.2x10^{-01}$	0.0064	$2.9x10^{-01}$
rs849327	28232457	A	0.341	0.997	0.0459	$5.4x10^{-10}$	0.0527	$1.4x10^{-02}$	0.0829	$2.0x10^{-15}$	0.078	$1.2x10^{-02}$	0.0071	$1.4x10^{-11}$	0.0107	$4.3x10^{-04}$	0.0007	$7.5x10^{-01}$	0.0069	$2.6x10^{-01}$
rs498475	28256240	G	0.366	0.989	0.0426	$5.7x10^{-08}$	0.0564	$7.3x10^{-03}$	0.0732	$1.4x10^{-12}$	0.069	$2.4x10^{-02}$	0.0059	$9.8x10^{-09}$	0.0107	$3.6x10^{-04}$	0.0012	$5.7x10^{-01}$	0.0082	$1.7x10^{-01}$

					BMI				Height				Weight			
					Discovery		Replication		Discovery		Replication		Discovery		Replication	
Variant	BP[1]	EA[2]	EAF[3]	INFO[4]	Beta	P-value	Beta	P-value	Beta	P-value	Beta	P-value	Beta	P-value	Beta	P-value
rs10245867	28142186	T	0.329	0.995	-0.0079	$7.6x10^{-04}$	######	$4.1x10^{-01}$	0.0069	$2.3x10^{-06}$	0.0146	$1.6x10^{-03}$	-0.0029	$1.5x10^{-01}$	0.0026	$6.7x10^{-01}$
rs849138	28177338	G	0.493	0.998	-0.0108	$1.1x10^{-06}$	######	$4.0x10^{-01}$	0.0241	$1.8x10^{-69}$	0.0361	$1.7x10^{-16}$	0.0049	$1.0x10^{-02}$	0.0134	$2.0x10^{-02}$
7:28178625	28178625	TG	0.493	0.997	-0.0109	$8.5x10^{-07}$	######	$3.8x10^{-01}$	0.0240	$4.7x10^{-69}$	0.0355	$4.9x10^{-16}$	0.0048	$1.2x10^{-02}$	0.0129	$2.5x10^{-02}$
rs849333	28219812	A	0.342	0.996	-0.0082	$4.1x10^{-04}$	######	$4.5x10^{-01}$	0.0065	$6.4x10^{-06}$	0.0163	$3.8x10^{-04}$	-0.0031	$1.2x10^{-01}$	0.0039	$5.1x10^{-01}$
rs849335	28223990	T	0.341	N/A	-0.0078	$7.5x10^{-04}$	######	$3.6x10^{-01}$	0.0073	$4.7x10^{-07}$	0.0152	$9.1x10^{-04}$	-0.0024	$2.4x10^{-01}$	0.0027	$6.5x10^{-01}$
rs849336	28224053	A	0.341	1	-0.0078	$8.2x10^{-04}$	######	$3.6x10^{-01}$	0.0073	$4.3x10^{-07}$	0.0151	$9.8x10^{-04}$	-0.0023	$2.5x10^{-01}$	0.0027	$6.6x10^{-01}$
rs849327	28232457	A	0.341	0.997	-0.0078	$8.7x10^{-04}$	######	$3.9x10^{-01}$	0.0074	$3.0x10^{-07}$	0.0149	$1.2x10^{-03}$	-0.0022	$2.7x10^{-01}$	0.0029	$6.3x10^{-01}$
rs498475	28256240	G	0.366	0.989	-0.0055	$1.7x10^{-02}$	######	$6.0x10^{-01}$	0.0051	$3.5x10^{-04}$	0.0127	$4.9x10^{-03}$	-0.0019	$3.5x10^{-01}$	0.0030	$6.1x10^{-01}$

Collider Bias







Univariate fine-mapping results in the *JAZF1* region for asthma, T2D and height

					Association p-values			Asthma		T2D		Height	
Variant	Position	A1[1]	AF[2]	INFO[3]	Asthma	T2D	Height	CS[4]	PIP[5]	CS[4]	PIP[5]	CS[4]	PIP[5]
7:27945054	27945054	CT	0.067	0.896	3.84E-01	6.24E-01	2.14E-16	None		0.0038	None	0.0034	HEIGHT_1	0.8047
rs773503722	27949022	CA	0.089	0.957	8.80E-01	3.00E-02	8.33E-13	None		0.0034	None	0.0047	HEIGHT_1	0.0248
rs12538288	27982312	C	0.059	0.984	9.47E-01	7.48E-02	1.12E-17	None		0.0034	None	0.0038	HEIGHT_1	0.0223
rs7783410	27982633	C	0.059	0.984	9.18E-01	6.25E-02	1.08E-17	None		0.0034	None	0.0039	HEIGHT_1	0.0233
rs4722746	27985368	C	0.059	0.983	9.08E-01	6.48E-02	2.00E-17	None		0.0034	None	0.0039	HEIGHT_1	0.0172
rs6970113	27991924	C	0.059	0.983	9.14E-01	7.68E-02	2.21E-17	None		0.0034	None	0.0038	HEIGHT_1	0.0167
rs1557779	27994758	T	0.072	0.984	9.68E-01	2.25E-02	1.18E-18	None		0.0034	None	0.0044	HEIGHT_1	0.0585
7:28154215	28154215	GT	0.792	0.979	3.03E-13	8.32E-05	4.57E-02	ASTHMA_1		0.0452	None	0.0034	None	0.0035
rs4722758	28156606	C	0.800	0.999	1.07E-13	6.68E-05	3.57E-02	ASTHMA_1		0.1135	None	0.0034	None	0.0033
rs6977955	28156887	C	0.800	0.999	1.61E-13	6.60E-05	3.79E-02	ASTHMA_1		0.0779	None	0.0034	None	0.0034
rs4719922	28156058	C	0.800	1.000	1.42E-13	6.44E-05	3.58E-02	ASTHMA_1		0.0877	None	0.0034	None	0.0033
7:28159795	28159795	GT	0.190	0.854	2.64E-01	1.13E-02	1.68E-57	None		0.0040	None	0.0033	HEIGHT_2	1.0000
rs9648346	28160113	C	0.800	1.000	1.50E-13	6.26E-05	3.58E-02	ASTHMA_1		0.0834	None	0.0034	None	0.0033
7:28168745	28168745	GTCTT	0.792	0.996	6.71E-14	2.66E-06	1.15E-02	ASTHMA_1		0.1410	None	0.0034	None	0.0028
rs2189965	28172014	C	0.800	1.000	3.01E-13	7.72E-05	3.06E-02	ASTHMA_1		0.0440	None	0.0034	None	0.0028
rs2189966	28172066	T	0.792	1.000	1.40E-13	4.59E-06	1.35E-02	ASTHMA_1		0.0711	None	0.0034	None	0.0034
rs4722760	28172183	A	0.800	0.999	3.07E-13	7.10E-05	2.93E-02	ASTHMA_1		0.0432	None	0.0034	None	0.0028
rs917115	28172586	T	0.792	0.999	1.36E-13	4.50E-06	1.44E-02	ASTHMA_1		0.0729	None	0.0034	None	0.0035
rs702814	28172732	C	0.494	0.997	8.97E-08	6.28E-18	1.25E-68	None		0.0052	T2D_1	0.0200	None	0.0017
rs917116	28172739	T	0.791	0.994	2.32E-13	4.30E-06	1.20E-02	ASTHMA_1		0.0454	None	0.0034	None	0.0033
rs67250450	28174986	T	0.800	0.999	3.06E-13	6.63E-05	3.05E-02	ASTHMA_1		0.0434	None	0.0034	None	0.0029
rs917117	28176305	G	0.801	0.998	1.94E-13	7.06E-05	3.67E-02	ASTHMA_1		0.0656	None	0.0034	None	0.0034
rs11495981	28177301	C	0.801	0.998	1.93E-13	7.09E-05	3.68E-02	ASTHMA_1		0.0660	None	0.0034	None	0.0034
rs864745	28180556	T	0.494	1.000	2.89E-07	3.93E-18	1.72E-70	None		0.0048	T2D_1	0.0286	None	0.0021
rs849142	28185891	T	0.494	1.000	4.75E-07	2.91E-18	7.77E-70	None		0.0046	T2D_1	0.0371	None	0.0018
rs11455969	28186775	A	0.494	0.999	3.79E-07	3.55E-18	1.12E-69	None		0.0047	T2D_1	0.0313	None	0.0018
rs10622246	28187111	A	0.498	0.994	5.96E-07	1.08E-18	3.08E-68	None		0.0046	T2D_1	0.0560	None	0.0020
rs849133	28192280	C	0.498	0.999	4.56E-07	7.26E-19	2.03E-71	None		0.0047	T2D_1	0.1330	None	0.0019
rs860262	28194397	C	0.498	0.999	5.75E-07	5.77E-19	1.46E-71	None		0.0046	T2D_1	0.1648	None	0.0020
rs849134	28196222	A	0.496	1.000	7.83E-07	1.84E-18	2.09E-70	None		0.0045	T2D_1	0.0575	None	0.0017
rs849135	28196413	G	0.499	1.000	6.09E-07	6.24E-19	1.61E-70	None		0.0045	T2D_1	0.1522	None	0.0016
rs1708302	28196677	C	0.499	0.999	6.07E-07	7.03E-19	6.77E-71	None		0.0045	T2D_1	0.1363	None	0.0017
rs1513272	28200097	C	0.499	0.999	5.92E-07	7.36E-19	6.69E-71	None		0.0046	T2D_1	0.1303	None	0.0017
rs537124	28203142	C	0.295	0.999	8.89E-01	1.49E-08	8.11E-112	None		0.0055	None	0.0033	HEIGHT_3	0.0309
rs552707	28205303	T	0.296	0.999	9.26E-01	1.39E-08	4.56E-112	None		0.0056	None	0.0033	HEIGHT_3	0.0493
rs543511926	28207300	G	0.999	0.828	6.85E-02	5.27E-01	4.23E-01	None		0.0056	None	0.0033	HEIGHT_3	0.0317
rs520161	28210660	T	0.299	1.000	8.67E-01	6.60E-09	1.91E-112	None		0.0052	None	0.0034	HEIGHT_3	0.5385
rs508347	28212824	T	0.299	0.999	8.14E-01	1.11E-08	3.53E-112	None		0.0050	None	0.0033	HEIGHT_3	0.3277
rs10226758	28214614	A	0.656	0.803	1.84E-05	5.52E-02	6.01E-24	None		0.0037	None	0.0037	HEIGHT_4	1.0000

Mediation results for the two variants in *JAZF1* with cross-phenotype associations for asthma, T2D and height

	Total Effect				Direct Effect				Indirect Effect				Proportion Mediated			
	Beta	Lower CI[1]	Upper CI[1]	p-value	Beta	Lower CI[1]	Upper CI[1]	p-value	Beta	Lower CI[1]	Upper CI[1]	p-value	Estimate	Lower CI[1]	Upper CI[1]	p-value
Variant	asthma-T2D															
rs849138	0.0048	0.0037	0.0059	<2e-16	0.0047	0.0036	0.0058	<2e-16	0.0001	0.0001	0.0002	<2e-16	0.0247	0.0156	0.0400	<2e-16
7:28178625	0.0046	0.0035	0.0057	<2e-16	0.0047	0.0037	0.0058	<2e-16	0.0001	0.0001	0.0002	<2e-16	0.0252	0.0133	0.0471	<2e-16
	Height-T2D															
rs849138	0.0048	0.0037	0.0059	<2e-16	0.0051	0.0039	0.0061	<2e-16	0.0003	0.0003	-0.0002	<2e-16	-0.0523	-0.0759	-0.0400	<2e-16
7:28178625	0.0048	0.0037	0.0059	<2e-16	0.0045	0.0035	0.0056	<2e-16	0.0002	0.0002	0.0003	<2e-16	0.0488	0.0363	0.0700	<2e-16

JAZF1

- JAZF1 encodes a protein with three zinc fingers and acts as a transcriptional repressor.
- It is member of a chaperone complex that orchestrates acetylation at regulatory regions controlling the expression of many genes involved in ribosome biogenesis.
- Work on the Jazf1 knockout mouse induced pluripotent stem cells suggests JAZF1 is involved in differentiation of β-cells and glucose homeostasis.
- JAZF1 appears to limit inflammation in adipose tissue and mice overexpressing JAZF1 have lower body and fat weight.
- In mouse airway epithelial cultures, JAZF1 expression was shown to be necessary for multiciliated cell differentiation, which is important for removing contaminants from the airway.
- These functional studies suggest the plausibility of the role of JAZF1 in asthma and T2D, but do not suggest a genetic link between these phenotypes.

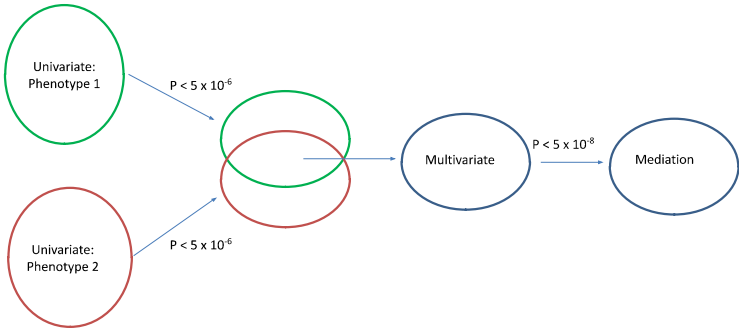
Previous associations with *JAZF1*

- Previous studies have found variants within JAZF1 to be associated separately with T2D, obesity phenotypes, as well as, height.
- These findings include at least one study that reports a significant association with SNPs in JAZF1 with WC adjusted for BMI.
  - Our findings also suggest that previous associations with SNPs in JAZF1 with WC adjusted for BMI are likely due to the same collider bias we observed, and the variants are associated with height, not adiposity.
- There is evidence JAZF1 is associated with child-onset and possibly adult-onset asthma

Conclusions

- While previous studies have identified associations with variants in JAZF1 associated with some aspect of all three phenotypes, this is the first time that asthma, T2D, and anthropometric measurements have been analyzed simultaneously in the same dataset and the first attempt at dissecting whether there are overlapping causal variants and/or biological pathways for these phenotypes.
- This study provides the strongest evidence for an association of variants in *JAZF1* with asthma compared to previous studies.
- Variants in *JAZF1* are associated with asthma, type 2 diabetes and height which provides a promising link between these three phenotypes, but the fine-mapped variant(s) for asthma, type 2 diabetes and height are unique.
- These results are consistent with biological pleiotropy at the gene-level for all three phenotypes.
- Mounting evidence that pleiotropy is more common at the gene-level (different causal variants) rather than at the variant level (shared causal variants)

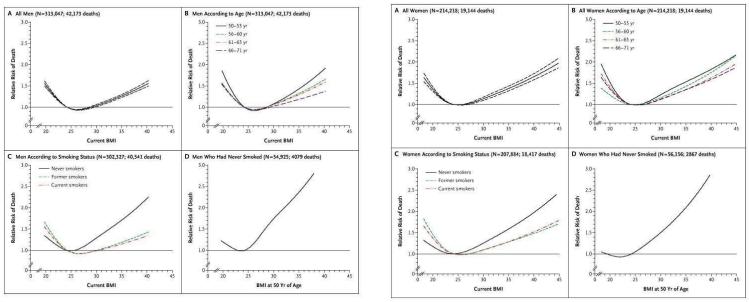
Pleiotropy exercise (Part 3)



Mendelian randomization: An Introduction

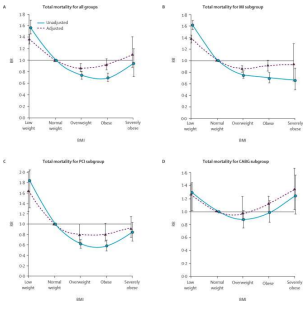
Andrew DeWan, PhD, MPH
Associate Professor of Epidemiology
Co-Director, Yale Center for Perinatal, Pediatric and Environmental Epidemiology
Director of Graduate Studies
Yale School of Public Health

Yale SCHOOL OF PUBLIC HEALTH

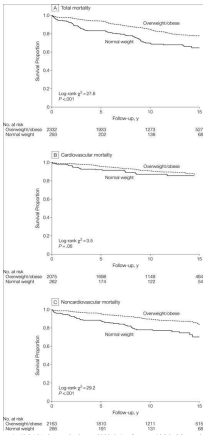


Adams et al. (2006) Overweight, Obesity and Mortality in a Large Prospective Cohort of Persons 50 to 71 Years Old, N Engl J Med 255:763-778

The "Obesity Paradox"

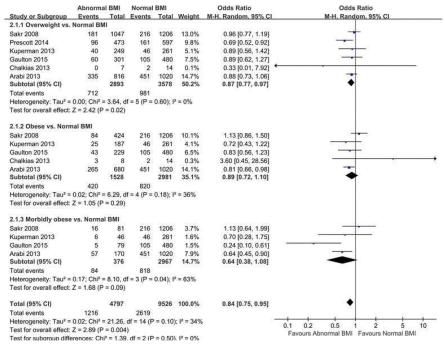


Romero-Corral A et al. (2006) Association of bodyweight with total mortality and with cardiovascular events in coronary artery disease: a systematic review of cohort studies. The Lancet 368:666-678.



Carnethon M et al. (2012) Association of Weight Status With Mortality in Adults With Incident Diabetes. JAMA 308:581-590.

BMI and Bloodstream Infection (BSI)/Sepsis Mortality



Wang S *et al.* (2017) The role of increased body mass index in outcomes of sepsis: a systematic review and meta-analysis. *BMC Anesthesiol* 17: 118.

Table 2. Associations of BMI and lifestyle factors with the risk of bloodstream infection among 64 027 participants in the HUNT Study, Norway 1995–2011

Lifestyle variable	Age- and sex-adjusted					Age-, sex-, education- and lifestyle-adjusted[b]				
	Person-years at risk	No. BSI	Incidence rate per 100 000 person-years	HR	95% CI	Person-years at risk	No. BSI	Incidence rate per 100 000 person-years	HR	95% CI
BMI, kg/m²										
< 18.5	5270	15	285	1.75	1.05–2.93	4402	8	182	1.41	0.70–2.85
18.5–24.9	321042	523	163	1.00	Reference	288423	388	135	1.00	Reference
25.0–29.9	340023	817	235	1.04	0.93–1.17	302915	568	188	1.02	0.90–1.17
30.0–34.9	103982	339	326	1.31	1.14–1.51	85402	234	274	1.18	1.17–1.63
35.0–39.9	21404	90	420	1.87	1.50–2.33	17137	58	338	1.77	1.34–2.33
≥ 40.0	3196	33	633	3.10	2.17–4.41	4181	23	550	3.14	2.05–4.81
Smoking										
Never	347417	703	191	1.00	Reference	317988	451	142	1.00	Reference
Previous	206261	626	304	1.30	1.16–1.46	180356	464	257	1.31	1.14–1.51
Current	231577	494	213	1.31	1.34–1.70	203405	364	179	1.53	1.32–1.78
Physical activity level[b]										
None	53199	207	389	1.71	1.42–2.07	48098	170	353	1.41	1.15–1.74
Slight	228970	497	217	1.18	1.01–1.38	216129	453	210	1.10	0.93–1.30
Moderate	230954	447	178	1.00	0.85–1.17	239738	417	174	0.97	0.83–1.15
High	204086	253	124	1.00	Reference	197784	237	120	1.00	Reference
Alcohol intake										
< 1 glass/2 weeks	272422	906	333	1.01	0.91–1.13	219150	572	261	0.97	0.85–1.10
1–7 glasses/2 weeks	401370	663	166	1.00	Reference	377542	569	151	1.00	Reference
8–14 glasses/2 weeks	85623	115	134	0.89	0.73–1.09	81741	100	122	0.86	0.69–1.06
≥ 15 glasses/2 weeks	24312	42	173	1.24	0.91–1.70	23316	38	163	1.21	0.87–1.69

[a]Adjusted for age, sex, education, smoking, BMI, physical activity and alcohol intake. The association of physical activity with BSI risk was additionally adjusted for chronic mobility impairment.
[b]None, no light or vigorous activity; slight, < 3 h light activity/week and no vigorous activity; moderate, ≥ 3 h light activity or < 1 h vigorous activity/week; high, ≥ 1 h vigorous activity/week.

Table 3. Associations of BMI and lifestyle factors with mortality from bloodstream infection[a] among 64 027 participants in the HUNT Study, Norway 1995–2011

Lifestyle variable	Age-and sex-adjusted					Age-, sex-, education-, and lifestyle-adjusted[b]				
	Person-years at risk	No. BSI deaths	Mortality rate per 100 000 person-years	HR	95% CI	Person-years at risk	No. BSI deaths	Mortality rate per 100 000 person-years	HR	95% CI
BMI, kg/m²										
< 18.5	5323	2	38	1.20	0.30–4.88	4445	1	23	0.56	0.13–6.94
18.5–24.9	323163	301	31	1.00	Reference	290075	71	24	1.00	Reference
25.0–29.9	339823	383	52	1.12	0.88–1.41	304978	137	38	1.09	0.81–1.47
30.0–34.9	105218	72	68	1.35	0.99–1.81	94596	48	36	1.49	1.02–2.17
35.0–39.9	21806	23	105	2.44	1.55–3.84	17420	15	86	2.56	1.45–4.53
≥ 40.0	3134	8	150	3.99	1.95–8.24	4272	7	164	5.47	2.56–12.58
Smoking										
Never	370124	159	41	1.00	Reference	320130	82	26	1.00	Reference
Previous	210418	137	66	1.50	1.01–1.68	182284	98	54	1.47	1.06–2.04
Current	233420	306	45	1.75	1.36–2.25	205073	79	39	2.21	1.59–3.14
Physical activity level[b]										
None	53918	51	95	2.08	1.37–3.16	48751	40	82	1.70	1.06–2.74
Slight	230884	307	46	1.39	0.97–2.00	218140	98	45	1.42	0.96–2.11
Moderate	232722	94	37	1.14	0.79–1.64	241663	85	35	1.19	0.80–1.77
High	205083	43	21	1.00	Reference	198931	36	18	1.00	Reference
Alcohol intake										
< 1 glass/2 weeks	275556	206	75	0.96	0.76–1.22	221482	122	55	0.89	0.67–1.19
1–7 glasses/2 weeks	404120	330	32	1.00	Reference	381017	108	28	1.00	Reference
8–14 glasses/2 weeks	86038	26	30	1.03	0.69–1.55	82233	22	27	0.98	0.62–1.56
≥ 15 glasses/2 weeks	24453	9	37	1.35	0.68–2.66	23453	7	30	1.13	0.52–2.45

[a]Defined as death within 30 days after detection of a BSI.
[b]Adjusted for age, sex, education, smoking, BMI, physical activity and alcohol intake. The association of physical activity with risk of BSI mortality was additionally adjusted for chronic mobility impairment.
[c]None, no light or vigorous activity; slight, < 3h light activity/week and no vigorous activity; moderate, ≥ 3 h light activity or < 1 h vigorous activity/week; high, ≥ 1 h vigorous activity/week.

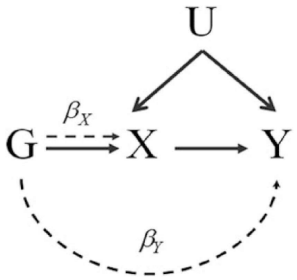
Paulsen J et al. (2017) Association of obesity and lifestyle with the risk and mortality of bloodstream infection in a general population: a 15-year follow-up of 64 027 individuals in the HUNT Study. Int J Epidemiol 46:1573-1581

Areas of Concern (BMI/BSI as an example)

- Selection Bias: If obesity is associated with BSI risk, non-obese patients may have other characteristics that cause their BSI that in turn are more strongly associated with mortality
- Reverse Causation: if measured BMI is affected by BSI
- Confounding: if factors such as chronic diseases and smoking habits that affect both BMI and BSI mortality are not adequately adjusted

Mendelian randomization

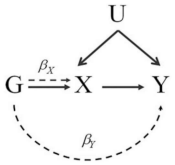
- Mimic randomized trial using genetic data as instruments for exposures
- Leverages information on genetic variants that segregate randomly at conception
- If an association between the instrument and outcome is detected, a causal relationship for this association is strengthened

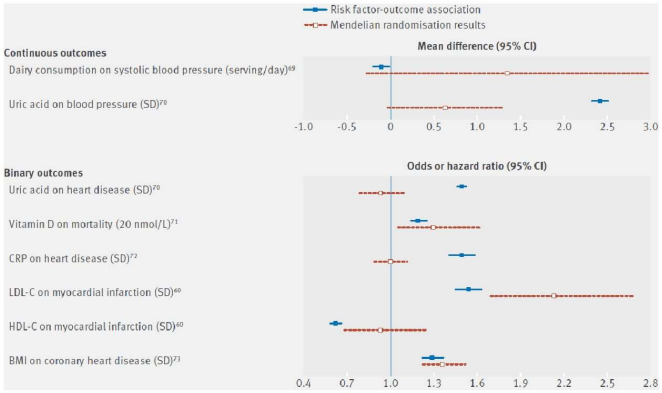


Dimou NL and Tsilidis KK. (2018) A primer in Mendelian Randomization Methodology with a Focus on Utilizing Published Summary Association Data. Methods Mol Biol. 2018;1793: 211-230

MR Assumptions

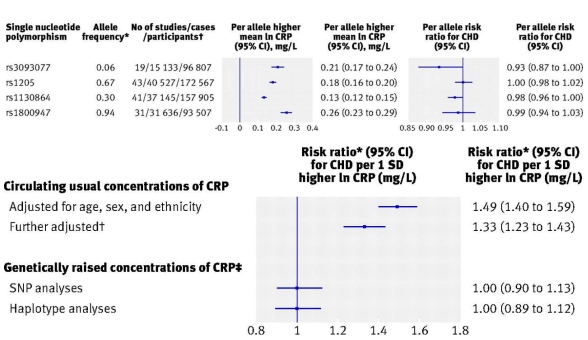
- The genetic instrument (G) is associated with the exposure (X)
- The genetic instrument is not associated with any confounder (U) of the exposure-outcome association
- The genetic instrument is conditionally independent of the outcome (Y) given the exposure and confounders





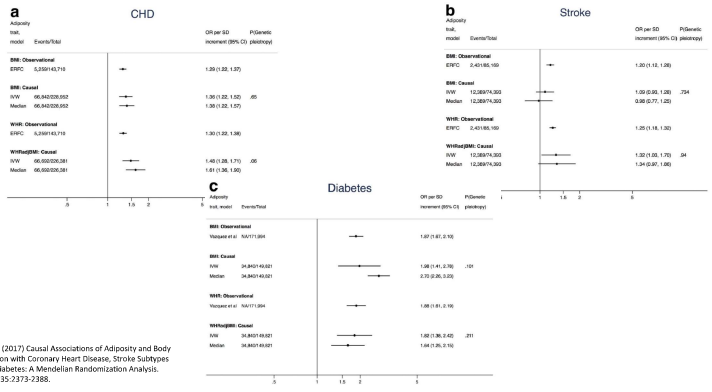
Davies et al. (2018) Reading Mendelian randomization studies: a guide, glossary, and checklist for clinicians. BMJ 362:k601

CRP and Heart Disease



C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC) BMJ 2011;342:bmj.d548

BMI and CHD|Stroke|Type 2 Diabetes



Dale CE et al. (2017) Causal Associations of Adiposity and Body Fat Distribution with Coronary Heart Disease, Stroke Subtypes and Type 2 Diabetes: A Mendelian Randomization Analysis. Circulation, 135:2373-2388.

One-sample vs. two-sample designs

One-sample

- Genotype(s), risk factor and outcome all measured in the same set of study subjects
- Individual level data must be available

Two-sample

- Genotype(s) and risk factor measured in one set of study subjects and genotype(s) and outcome measured in a separate set of study subjects
- Can use summary statistics or individual level data

One-sample vs. two-sample designs

Assumption/Issue	One-sample	Two-sample
Instrument variable related to risk factor	Weak instrument biases towards the confounded regression result	Weak instrument biases towards the null
Confounders	Can (and should) check this for measured confounders	Not often possible when using summary statistics
Pleiotropy	Multiple methods to explore this issue (including MR-Egger)	Multiple methods to explore this issue (including MR-Egger) and may be more powerful with large consortium datasets since methods tend to be statistically inefficient
Subgroup analyses	Possible if large sample sizes and data on relevant risk factors are available	Only possible if individual level data are available
Bias from adjustments made in GWAS	N/A as all adjustments made in the same set of subjects	Summary data may or may not have been adjusted

Adapted from: Lawlor DA (2016) Commentary: Two-sample Mendelian randomization: opportunities and challenges. Int J Epi 45: 908-915.

Selecting genetic variants for an instrument

- Single or multiple variants
- Current recommendation is to select variant(s) that are significantly associated with the exposure at the genome-wide level
- Want a strong genetic instrument to avoid weak instrument bias
  - A single variant or variants with modest effects in small samples are likely to have low power and can suffer from bias
- If selecting multiple variants these should not be in LD and assumes negligible gene-gene interaction among variants

Instrument strength

- Measured using the F statistic in the regression of the IV on the exposure

$$F = \frac{N-K-1}{K} * \frac{R^2}{1-R^2}$$

R^2: proportion of the variance of the exposure explained by IV

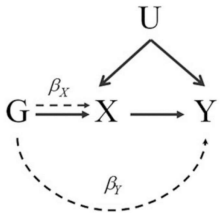
N: sample size

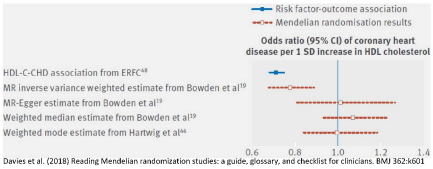
K: number of genetic variants

General Rule: F < 10 is an indication of a weak instrument

Pleiotropy

- Assumption that the IV is not associated with Y independently from X
- Presence of pleiotropy can bias the causal estimate
- Sensitivity analyses such as MR-Egger can be used to test whether or not the pleiotropy assumption has been violated





Davies et al. (2018) Reading Mendelian randomization studies: a guide, glossary, and checklist for clinicians. BMJ 362:k601

Testing MR: Wald Ratio

- Simple ratio of the effects of the instrument variable on the outcome over the instrument variable on the exposure
- Can be implemented in both one and two sample designs
  - One sample can use either a single variant or a GRS
  - Two sample design that uses multiple variants requires a method for combining Wald Ratios

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{ZY}}{\hat{\beta}_{ZX}}$$

Testing MR: 2 stage least squares (2SLS)

- Single continuous instrument (GRS)
- Only for one sample method
- Assumes a linear relationship between exposure and outcome

- Regress X on G
- Calculate genetically predicted values of X
- Regress Y on genetically predicted values of X
- Fix the standard errors (e.g. sandwich estimator)

Testing MR: Inverse variant weighted

- One or two sample designs
- Tends to give more reliable results in the presence of heterogeneity and when using large number of instruments

- Fixed (assumes no heterogeneity across SNP) or random effects meta-analysis

For each variant calculate the Wald ratio:

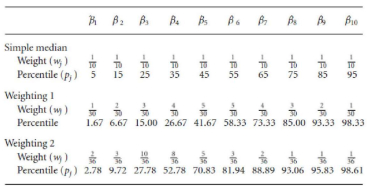
$$\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$$

Combine into an overall estimate using a formula from meta-analysis literature:

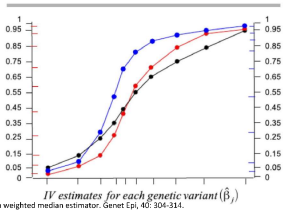
$$\hat{\beta}_{IVW} = \frac{\Sigma_j \hat{\gamma}_j^2 \sigma_{Yj}^{-2} \hat{\beta}_j}{\Sigma_j \hat{\gamma}_j^2 \sigma_{Yj}^{-2}}$$

Testing MR: Weighted Median

- Calculate the Wald ratio for each instrument
- Select the median value according to the weighted method



	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$
Simple median										
Weight (w_j)	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
Percentile (p_j)	5	15	25	35	45	55	65	75	85	95
Weighting 1										
Weight (w_j)	$\frac{1}{300}$	$\frac{2}{300}$	$\frac{3}{300}$	$\frac{4}{300}$	$\frac{5}{300}$	$\frac{5}{300}$	$\frac{4}{300}$	$\frac{3}{300}$	$\frac{2}{300}$	$\frac{1}{300}$
Percentile	1.67	6.67	15.00	26.67	41.67	58.33	73.33	85.00	93.33	98.33
Weighting 2										
Weight (w_j)	$\frac{1}{90}$	$\frac{1}{90}$	$\frac{10}{90}$	$\frac{8}{90}$	$\frac{5}{90}$	$\frac{9}{90}$	$\frac{5}{90}$	$\frac{1}{90}$	$\frac{1}{90}$	$\frac{1}{90}$
Percentile (p_j)	2.78	9.72	27.78	52.78	70.83	81.94	88.89	93.06	95.83	98.61

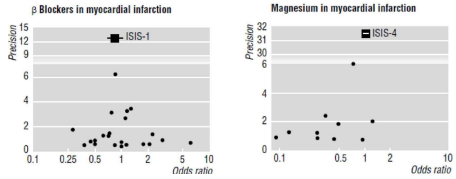


Bowden et al. (2016) Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epi, 40: 304-314.

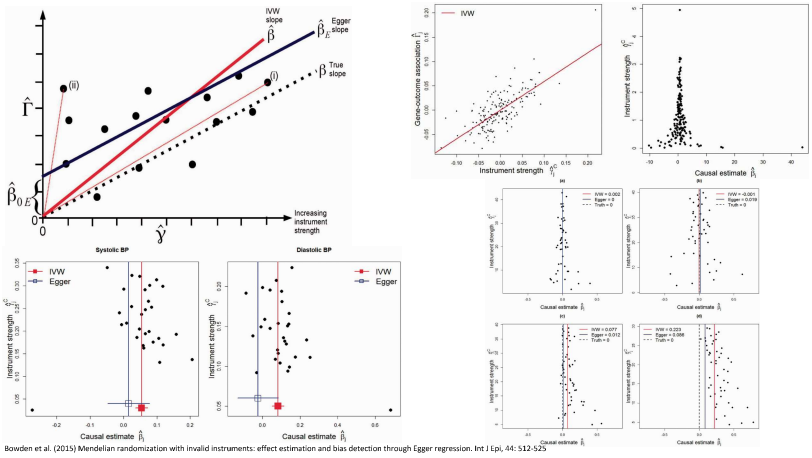
- Valid estimate when more than half of the genetic variants satisfy the IV assumptions
- No single IV contributes more than 50% of the weight

Testing MR: MR-Egger

- Provide a valid causal estimate in the presence of some violations of the MR assumptions (mainly pleiotropy)
- MR consisting of a single study with multiple IVs is analogous to a meta-analysis
- Bias resulting from pleiotropy is analogous to small study bias in meta-analysis
  - Small studies with less precise estimates tend to report larger estimates than big studies with more precise estimates
- Regress the standard normal deviate (odds ratio divided by its se) on the estimate's precision (inverse of the se)
  - Without bias, intercept = 0, and in the presence of bias the intercept is a measure of asymmetry



Egger et al. (1997) Bias in meta-analysis detected by a simple, graphical test. BMJ 315:629 - 634



Bowden et al. (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epi, 44: 512-525

Databases and software

Table 3 | Databases of genome-wide association study results

Data source	Description	Number of traits	Integrated with statistics package?
MR-Base	A curated database of genome-wide association study results with integrated R package for MR[23]	Over 1000	Yes
PhenoScanner	A curated database of genome-wide association study results with integrated R package for MR[37]	Over 500	Yes
GWAS catalog	Searchable database of genome-wide association study results[38]	Over 24 000	No

Davies et al. (2018) Reading Mendelian randomization studies: a guide, glossary, and checklist for clinicians. BMJ 362:k601

Body mass index and risk of dying from a bloodstream infection: A Mendelian randomization study

Tormod Rogne[1,2,3]*, Erik Solligård[1,3], Stephen Burgess[4,5], Ben M. Brumpton[6,7,8], Julie Paulsen[9], Hallie C. Prescott[10,11], Randi M. Mohus[1,3], Lise T. Gustad[1,12], Arne Mehl[12], Bjørn O. Åsvold[6,13], Andrew T. DeWan[1,2‡], Jan K. Damås[1,14,15‡]

PLOS Medicine | https://doi.org/10.1371/journal.pmed.1003413 November 16, 2020

Assess the causal association between BMI and risk of and mortality from BSI by overcoming the limitations of previous observational studies by conducting an MR study in a general population of approximately 56,000 participants in Norway with 23 years of follow-up

Study Population

- The Trondelag Health Study (HUNT) is a series of cross-sectional surveys carried out in Nord-Trondelag County, Norway
- 130,000 inhabitants who are representative of the general Norwegian population in terms of morbidity, mortality, sources of income and age distribution
- Based on HUNT2 survey conducted in 1995-1997 with 65,236 participants, 55,908 of whom had complete data for the analysis

Table 1. Background characteristics.

Characteristic	Total population (*n* = 55,908)	BSI incidence (*n* = 2,547)	BSI death (*n* = 451)
Age (years)[$]	48.3 (36.5–62.3)	63.6 (52.9–71.4)	67.3 (57.1–74.5)
Male sex[#]	26,324 (47.1)	1,345 (52.8)	263 (58.3)
BMI (kg/m^2)[^]	26.3 (4.1)	27.7 (4.5)	27.9 (4.8)
Median follow-up time (years)[$]	21.1 (17.1–21.8)	13.8 (8.4–18.3)	13.3 (7.7–17.9)
Self-reported cancer[#]	1,955 (3.7)	144 (6.2)	24 (5.9)
Smoking[#]			
Never	23,594 (43.0)	876 (35.2)	156 (35.6)
Previous	15,133 (27.6)	893 (35.8)	164 (37.4)
Current	16,117 (29.4)	723 (29.0)	118 (26.9)
Physical activity[#]			
None	3,821 (7.6)	243 (11.9)	54 (15.4)
Slight	15,662 (31.0)	714 (34.9)	117 (33.3)
Moderate	17,167 (34.0)	693 (33.9)	116 (33.1)
High	13,810 (27.4)	397 (19.4)	64 (18.2)
Education[#]			
≤9 years	19,033 (35.7)	1,305 (55.8)	240 (58.8)
10–12 years	23,468 (44.0)	762 (32.6)	125 (30.6)
≥13 years	10,832 (20.3)	274 (11.7)	43 (10.5)

BMI, body mass index; BSI, bloodstream infection. Data are presented as
^mean (standard deviation)
[$]median (25th–75th percentiles), or
[#]*n* (%). BSI incidence is based on first occurrence, otherwise, last occurrence is used. Education defined as follows: ≤9 years ("primary school 7–10 years, continuation school, folk high school"), 10–12 years ("high school, intermediate school, vocational school, 1–2 years high school" and "university qualifying examination, junior college, A levels"), and ≥13 years ("university or other post-secondary education, less than 4 years" and "university/college 4 years or more"). Activity defined as follows: none ("no light or vigorous activity"), slight ("<3 h light activity/week and no vigorous activity"), moderate ("≥3 h light activity/week or <1 h vigorous activity/week"), or high (">1 h vigorous activity/week").

Outcome

- Linked to all prospectively recorded blood cultures at the two community hospitals in the catchment area (Levanger and Namsos Hospitals) as well as St. Olav's Hospital in Trondheim (tertiary referral center)
- Data on blood cultures were available from January 1, 1995 through the end of 2017
- Date of death and emigration out of Nord-Trondelag County were obtained from the Norwegian population registry
- BSI was defined as a positive blood culture of pathogenic bacteria
- BSI mortality was defined as death within 30 days of BSI diagnosis

Genetic Instrument

- Based on a BMI meta-analysis of ~700,000 individuals [Yengo L et al. [2018] Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. Hum. Mol. Genet., 27, 3641–3649.)
- 939 of 941 SNPs identified as associated with BMI ($p<5x10^{-8}$, two SNPs did not pass imputation quality control)
- Genetic risk score (GRS) was calculated for BMI using the --score command in PLINK (version 1.9) and weighted based on the effect estimates from the meta-analysis
- GRS (939 variants) explained 4.2% of the variation in BMI in the population (F-statistic = 2,461)

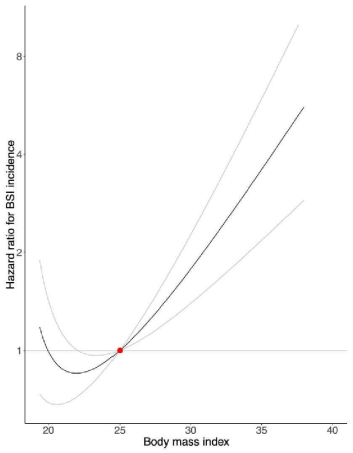
Analysis Methods

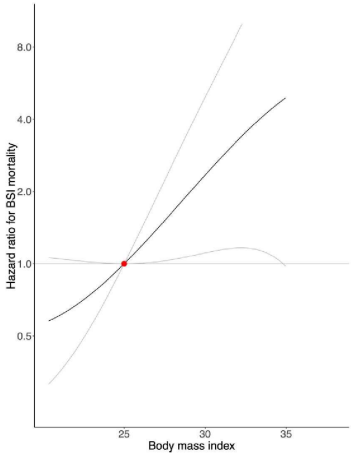
- Fractional polynomial model (suggestion of a nonlinear relationship between BMI and BSI)
- 2-stage least squares (with sandwich estimator) for analyses assuming a linear relationship between exposure and outcome
- Sensitivity analyses
  - MR Egger (random effects)
  - INW
  - Weighted median
  - 2-sample (using Yengo et al. for SNP-exposure associations)

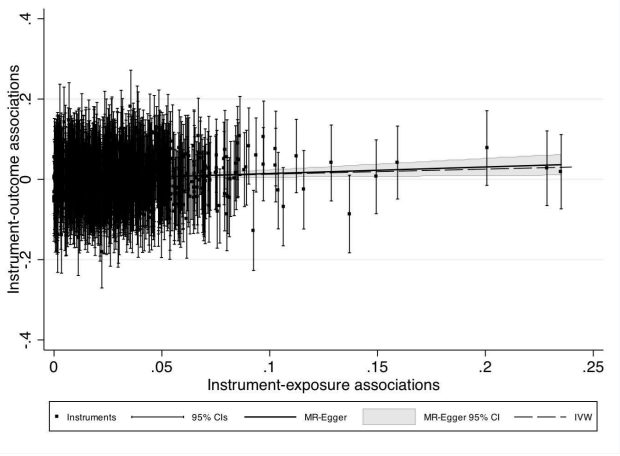
Table 1. Background characteristics.

Characteristic	Total population (*n* = 55,908)	BSI incidence (*n* = 2,547)	BSI death (*n* = 451)
Age (years)[b]	48.3 (36.5–62.3)	65.6 (52.9–71.4)	67.3 (57.1–74.5)
Male sex[c]	26,324 (47.1)	1,345 (52.8)	263 (58.3)
BMI (kg/m^2)[a]	26.3 (4.1)	27.7 (4.5)	27.9 (4.8)
Median follow-up time (years)[b]	21.1 (17.1–21.8)	13.8 (8.4–18.5)	13.3 (7.7–17.9)
Self-reported cancer[c]	1,955 (3.7)	144 (6.2)	24 (5.9)
Smoking[c]			
Never	23,594 (43.0)	876 (35.2)	156 (35.6)
Previous	15,133 (27.6)	893 (35.8)	164 (37.4)
Current	16,117 (29.4)	723 (29.0)	118 (26.9)
Physical activity[c]			
None	3,821 (7.6)	243 (11.9)	54 (15.4)
Slight	15,662 (31.0)	714 (34.9)	117 (33.3)
Moderate	17,167 (34.0)	693 (33.9)	116 (33.1)
High	13,810 (27.4)	397 (19.4)	64 (18.2)
Education[c]			
≤9 years	19,033 (35.7)	1,305 (55.8)	240 (58.8)
10–12 years	23,468 (44.0)	762 (32.6)	125 (30.6)
≥13 years	10,832 (20.3)	274 (11.7)	43 (10.5)

BMI, body mass index; BSI, bloodstream infection. Data are presented as
[a]mean (standard deviation)
[b]median (25th–75th percentiles), or
[c]*n* (%). BSI incidence is based on first occurrence otherwise, last occurrence is used. Education defined as follows: ≤9 years ("primary school 7–10 years, continuation school, folk high school"), 10–12 years ("high school, intermediate school, vocational school, 1–2 years high school" and "university qualifying examination, junior college, A levels"), and ≥13 years ("university or other post-secondary education, less than 4 years" and "university/college 4 years or more"). Activity defined as follows: none ("no light or vigorous activity"), slight ("<3 h light activity/week and no vigorous activity"), moderate ("≥3 h light activity/week or <1 h vigorous activity/week"), or high ("≥1 h vigorous activity/week").







S5 Table. Mendelian randomization sensitivity analyses of linear association between body mass index and bloodstream infection mortality in the general population

	HR/OR	Lower	Upper	P-value	Intercept	Lower	Upper	P-value
One-sample								
MR-Egger, random effects	1.18	1.04	1.33	0.011	1.00	0.99	1.00	0.476
IVW, random effects	1.13	1.05	1.23	0.002	-	-	-	-
Median estimator, weighted	1.13	0.99	1.30	0.081	-	-	-	-
Two-sample								
MR-Egger, random effects	1.98	0.95	4.18	0.070	1.00	0.99	1.01	0.877
IVW, random effects	1.89	1.33	2.67	<0.001				
Median estimator, weighted	2.09	1.10	3.97	0.025				

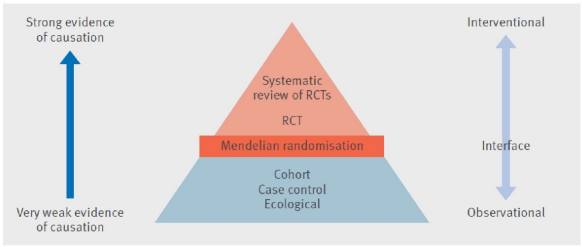
HR, hazard ratio; IVW, inverse-variance weighted; OR, odds ratio. Assuming a linear relationship between body mass index and bloodstream infection mortality in the general population using the same 939 single nucleotide polymorphisms (SNP) as used to create the genetic risk score. Two-sample analyses use SNP-exposure associations from Yengo et al [ref 2 in Supplementary text], and SNP-outcome associations from HUNT. The I^2 of the SNP-exposure associations were 54% in the one-sample MR-Egger regression, and 92% in the two-sample MR-Egger regression. Effect estimates reported as HR for one unit increase of body mass index in one-sample analyses and as OR for one standard deviation increase of body mass index in two-sample analyses.

STROBE-MR: Guidelines for strengthening the reporting of Mendelian randomization studies

Authors (in alphabetical order):

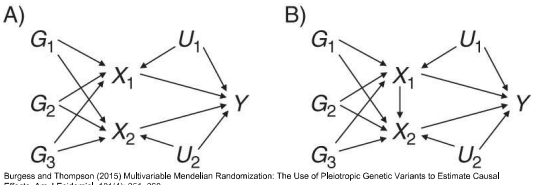
George Davey Smith, Neil M Davies, Niki Dimou, Matthias Egger, Valentina Gallo, Robert Golub, Julian PT Higgins, Claudia Langenberg, Elizabeth W Loder, J Brent Richards, Rebecca C Richmond, Veronika W Skrivankova, Sonja A Swanson, Nicholas J Timpson, Anne Tybjaerg-Hansen, Tyler J VanderWeele, Benjamin AR Woolf, James Yarmolinsky

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27857v1 | CC BY 4.0 Open Access | rec: 15 Jul 2019, publ: 15 Jul 2019

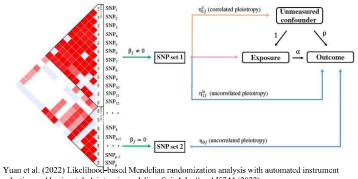


Davies et al. (2018) Reading Mendelian randomization studies: a guide, glossary, and checklist for clinicians. BMJ 362:k601

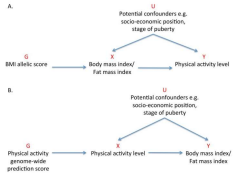
Some Advanced MR analysis approaches



Burgess and Thompson (2015) Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects. Am J Epidemiol. 181(4): 251–260.



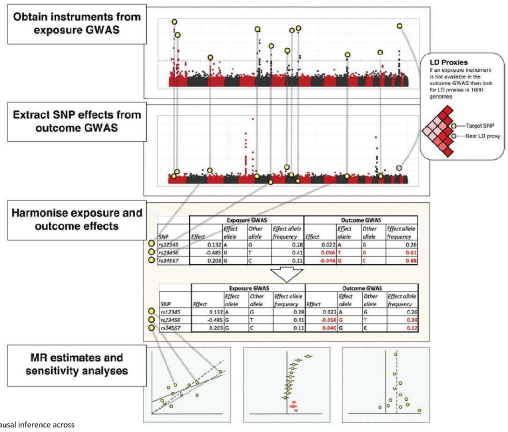
Yuan et al. (2022) Likelihood-based Mendelian randomization analysis with automated instrument selection and horizontal pleiotropic modeling. Sci. Adv. 8, eabl5744 (2022)



Richmond et al. (2014) Assessing Causality in the Association between Child Adiposity and Physical Activity Levels: A Mendelian Randomization Analysis. PLoS Med 11(3): e1001618.

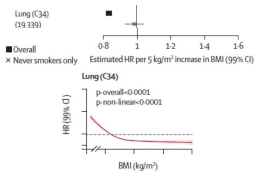
MRBASE

http://app.mrbase.org/



Hemani G et al. (2018) The MR-Base platform supports systematic causal inference across the human phenome. eLife 7:e34408.

BMI and Lung Cancer



Bhaskaran et al. (2014) Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults. Lancet 384:755-765

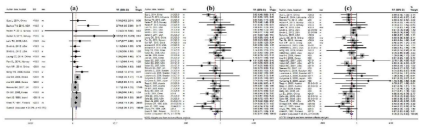


Figure 1. Forest plot of relative risks of underweight, overweight and obesity vs. normal weight for BMI with LC risk. Open blue diamonds denote the summary relative risks. Black diamonds indicate the RR in each study. The size of each box indicates the relative weight of each study in the meta-analysis. Horizontal lines represent the 95% confidence intervals (CIs). (**a**) Forest plots of risk of lung cancer associated with underweight (BMI ≤ 18.5 kg/m²). (**b**) Forest plots of risk of lung cancer associated with overweight (BMI = 25–29.9 kg/m²). (**c**) Forest plots of risk of lung cancer associated with obesity (BMI ≥ 30 kg/m²). RR, relative risk; normal weight, BMI = 18.5–24.9 kg/m²; BMI, body mass index; LC, lung cancer; m, men; w, women; mw, men and women.

Duan et al. (2015) Body mass index and risk of lung cancer: Systematic review and dose response meta-analysis. Sci Rep 5:16938

Prioritizing variants for a PRS

Advanced Gene Mapping Course, January 2025

Jurg Ott, Ph.D., Professor Emeritus
Rockefeller University, New York
https://lab.rockefeller.edu/ott/
https://jurgott.github.io/
ott@rockefeller.edu
PH +1 646 321 1013

Jurg Ott, PhD

Professor Emeritus and Director
Laboratory of Statistical Genetics
Rockefeller University
New York, NY 10065

https://lab.rockefeller.edu/ott/
EM: ott@rockefeller.edu
PH: +1 646 321 1013

Research Interests

Development of analysis methods for genetic data, genetic linkage and association analysis. Current topics: Digenic disease mapping; disease prediction based on genotype patterns.

Implementation in computer programs, dissemination on website

Collaboration with researchers world-wide on their data

Recent publications: [1-7] #1 now freely available from https://github.com/jurgott/handbook

1. Terwilliger, J.D. and Ott, J. (1994) *Handbook of human genetic linkage* Johns Hopkins University Press
2. Horpaopan, S. *et al.* (2020) Shared genomic segment analysis with equivalence testing. *Genet Epidemiol* 44, 741-747. DOI:10.1002/gepi.22335
3. Okazaki, A. *et al.* (2020) Population genetics: past, present, and future. *Human genetics*, 1-10. DOI:10.1007/s00439-020-02208-5
4. Okazaki, A. *et al.* (2021) Genotype pattern mining for pairs of interacting variants underlying digenic traits. *Genes* 12, 1160. DOI:10.3390/genes12081160
5. Okazaki, A. and Ott, J. (2022) Machine learning approaches to explore digenic inheritance. *Trends Genet.* DOI:10.1016/j.tig.2022.04.009
6. Ott, J. and Park, T. (2022) Overview of frequent pattern mining. *Genomics Inform* 20, e39. DOI:10.5808/gi.22074
7. Zhang, Q. *et al.* (2023) A multi-threaded approach to genotype pattern mining for detecting digenic disease genes. *Front Genet* 14, 1222517. DOI:10.3389/fgene.2023.1222517

Polygenic Risk Scores, PRS

https://choishingwan.github.io/PRS-Tutorial
Choi, *Nature Protocols* 15, 2759-2772 (2020) doi: s41596-020-0353-1

- What is it? "A PRS is an estimate of an individual's genetic liability to a trait or disease, calculated according to their genotype profile and relevant genome-wide association study (GWAS) data."
- Is it used a lot? Ubiquitously!
- Positive aspect: Combine single-SNP effects for disease association.
- Downsides: Main effects only, with few exceptions. Not very predictive of phenotype (case vs. control). Mosley et al (2020) *JAMA*. doi:10.1001/jama.2019.21782

Some prediction terminology

	test positive	test negative
Phenotype	predict case	predict control
case	*a*	*b*
control	*c*	*d*

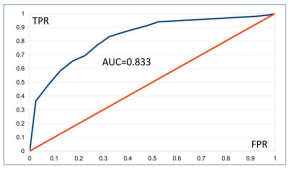
← sensitivity, power = a/(a + b) = TPR
← specificity, 1 - p = d/(c + d) = 1 - FPR

↑ Positive Predictive Value
PPV = a/(a + c)

↑ Negative Predictive Value
NPV = d/(b + d)

Prediction accuracy
ACC = (a + d)/(a + b + c + d)

ROC curve



Principles of Prediction

- Assume genotype 1/1 at any SNP perfectly predicts disease.
- SNP 1: 50% of individuals carry its 1/1 genotype.
- SNP 2: 30% of individuals carry its 1/1 genotype

	SNP 2		65%
SNP 1	1/1	other	sum
1/1	0.15	0.35	0.5
other	0.15	0.35	0.5
sum	0.3	0.7	1

	SNP 2		55%
SNP 1	1/1	other	sum
1/1	0.25	0.25	0.5
other	0.05	0.45	0.5
sum	0.3	0.7	1

- Many SNPs: Complicated. Solutions, mostly based on significant SNPs:
- Sums of pedigree maximum lod scores, MacLean et al, Am J Hum Genet. **1992**;50:1259–1266
- From *n* dimensions to 1 dimension: Sums of statistics over many SNPs (Hoh et al, **2001**, doi:10.1101/gr.204001)
- From *n* dimensions to 1 dimension: Multifactor dimensionality reduction, MDR (Moore et al, **2006**, doi:10.1016/j.jtbi.2005.11.036)
- Polygenic Risk Score, PRS → weighted sum of risk alleles over all individuals. Gidziela, 2023 (doi:10.36866/pn.132.21). PRS used in plant breeding; introduced to human genetics by Wray et al, **2007** (doi:10.1101/gr.6665407)

Prediction versus Significance

G. Shmueli, To Explain or to Predict? *Statistical Science* **25**, 289-310 (2010)
doi: 10.1214/10–STS330

- Current Approach
  - First step: Identify significant SNPs, based on some genetic association test
  - Add other, non-significant SNPs if necessary
  - Large number (10,000 – 100,000+) of SNPs used in PRS
  - Main aim of PRS: Prediction, identification of cases for polygenic traits like schizophrenia
- Lo, Chernoff, Zheng, Lo (2015) *PNAS* doi: 10.1073/pnas.1518285112
  - "Why significant variables aren't automatically good predictors". Abstract, short:

Thus far, genome-wide association studies (GWAS) have been disappointing in the inability of investigators to use the results of statistically significant variants in complex diseases to make predictions useful for **personalized medicine**. We demonstrate that highly predictive variables do not necessarily appear as highly significant, thus evading the researcher using significance-based methods. We point out that what makes variables good for prediction versus significance depends on different properties of the underlying distributions. If prediction is the goal, we must lay aside significance as the only selection standard.

Schizophrenia, males

807,119 SNPs; 660 cases and 1,504 controls

- Allelic association test in *plink*, `--assoc`
- Focus on 394,115 SNPs with OR > 1, that is, minor allele is disease-associated.
- Make PRS with *plink*, `--score`
- Classify as "case" if PRS > 95th %ile of PRS values in controls.
- Scrutinize results by cross-validation

- Order SNPs in 3 ways, by significance, *p*; odds ratio, OR = (a × d)/(b × c); or precision accuracy, ACC = (a + d)/*n*.
- PPV = a/(a + c) NPV = d/(b + d)
- `OR = PPV×NPV/[(1-PPV)(1-NPV)]`
- Rule: Order SNPs by OR. Use as many of the best SNPs so ACC > 0.95.

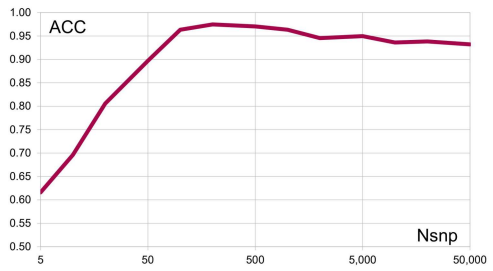
SNP	rank, p	rank, OR	rank, ACC
rs9275473	1	1,005	138,240
rs9260935	2	83	6
rs2647062	3	827	92,906
rs4387788	4	717	75,297
rs17212223	5	1,314	124,072
rs7209186	9	1	1
rs36054782	24	2	5
rs11064877	396	3	11
rs17141395	405	4	15
rs17071552	244	5	13
rs7209186	9	1	1
rs35147560	135,089	14,614	2
rs9501426	3,487	952	3
rs867831	60,052	4,362	4
rs36054782	24	2	5

Phenotype	predict case	predict control
case	*a*	*b*
control	*c*	*d*

Age-related Macular Degeneration, AMD

Klein et al (2005) *Science*, doi: 10.1126/science.1109557

Nsnp	ACC
5	0.616
10	0.696
20	0.806
50	0.897
100	0.963
200	0.975
500	0.970
1,000	0.963
2,000	0.945
5,000	0.950
10,000	0.936
20,000	0.938
50,000	0.932

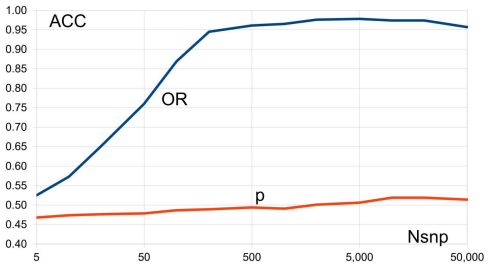


98,816 SNPs ⁘ 96 cases, 50 controls

Parkinson Disease

Downloaded 2006 from Coriell Institute

Nsnp	OR	p
5	0.525	0.468
10	0.573	0.474
20	0.652	0.477
50	0.76	0.479
100	0.869	0.487
200	0.945	0.489
500	0.961	0.494
1,000	0.965	0.491
2,000	0.976	0.501
5,000	0.978	0.506
10,000	0.974	0.519
20,000	0.974	0.519
50,000	0.957	0.514

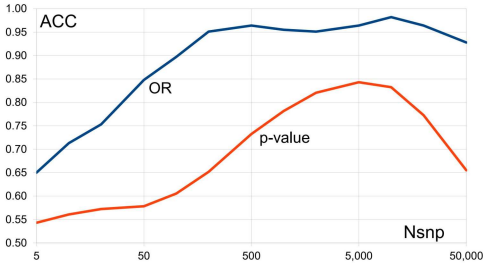


379,495 SNPs pruned to 106,141 SNPs ⁘ 270 cases, 271 controls

AMD collected in Hong Kong

Dewan et al (2006) *Science* doi: 10.1126/science.1133807

Nsnp	SNPs ordered by OR	SNPs ordered by p-value
5	0.650	0.543
10	0.713	0.561
20	0.753	0.572
50	0.848	0.578
100	0.897	0.605
200	0.951	0.652
500	0.964	0.732
1,000	0.955	0.782
2,000	0.951	0.821
5,000	0.964	0.843
10,000	0.982	0.832
20,000	0.964	0.773
50,000	0.928	0.655

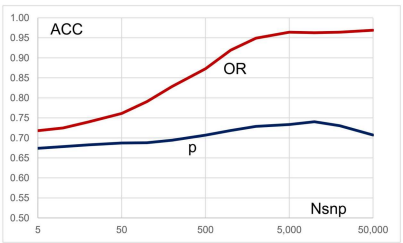


81,294 SNPs ⁘ 96 cases, 127 controls

Schizophrenia, males, from dbGaP

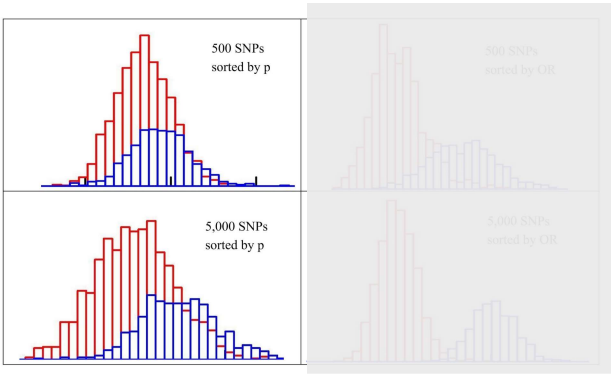
Ning et al (2024) *Med Res Arch*, doi: 10.18103/mra.v12i9.5723

Nsnp	OR	p
5	0.718	0.674
10	0.725	0.678
20	0.740	0.683
50	0.761	0.687
100	0.791	0.688
200	0.828	0.694
500	0.873	0.707
1,000	0.919	0.718
2,000	0.949	0.729
5,000	0.964	0.733
10,000	0.963	0.740
20,000	0.964	0.730
50,000	0.969	0.707

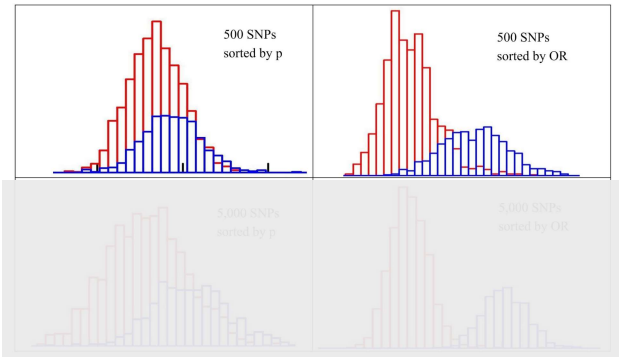


807,119 SNPs pruned to 179,104 SNPs ⁘ 660 cases, 1504 controls

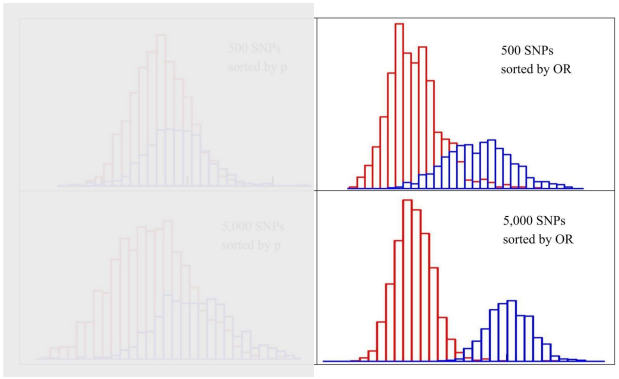
Histograms of 500 or 5,000 best SNPs ordered by *p* (male schizophrenia data)



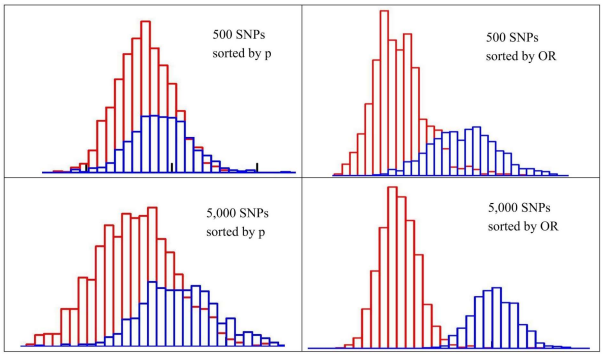
Histograms of 500 best SNPs ordered by *p* or *OR*



Histograms of 500 or 5,000 best SNPs ordered by *OR*



Histograms of 500 or 5,000 best SNPs ordered by *p* or *OR*



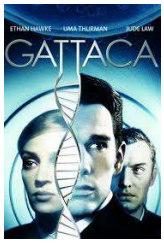
Number of misclassified individuals out of all 2,164 male schizophrenia cases and controls

	SNPs ordered by p	SNPs ordered by OR
Best 500 SNPs	732	271
Best 5,000 SNPs	720	80

	SNPs ordered by p	SNPs ordered by OR
Best 500 SNPs	33.8%	12.5%
Best 5,000 SNPs	33.3%	3.7%

Program L1outPRS, almost latest version:
https://github.com/jurgott/L1PRS

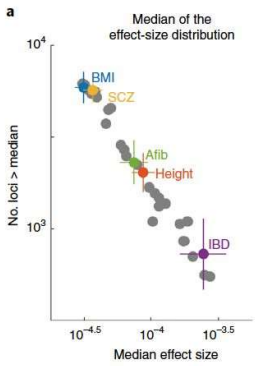
Genetic risk prediction



Genotype of an individual → Life-time risk of genetic disorders

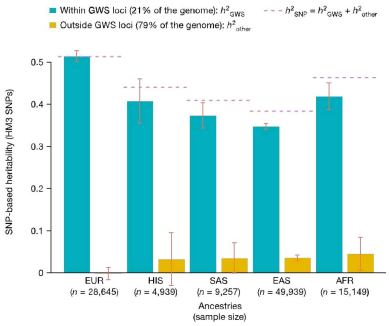
(Common SNPs) (Common complex genetic disorders)

Model-based estimates of effect sizes



O'Connor, *Nature Genetics* 2021

12,000 independent GWAS signals for height!

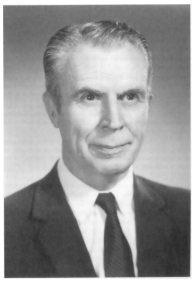


Yengo et al., *Nature* 2022

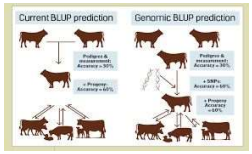
Effect sizes of individual variants are very small

- Genotype at a single locus carries very little information about phenotype.
- It does not mean that one cannot predict phenotype from genotype.
- Accuracy (r^2) of an ideal genetic predictor equals heritability.

BLUP – Best Linear Unbiased Predictor

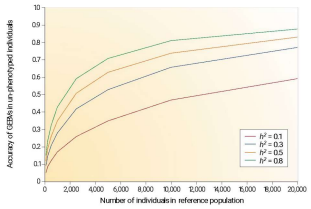


- Infinitesimal model
- Genetic effects are random
- Predict the expected genetic effect



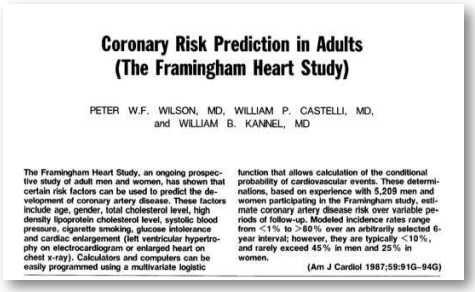
$$\hat{g}_i = \mathbf{G}_{i-}(\mathbf{G} + \mathbf{I}\lambda)^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu})$$

Accuracy of polygenic prediction in cattle



Poor transferability between breeds!

Measuring risk of myocardial infarction



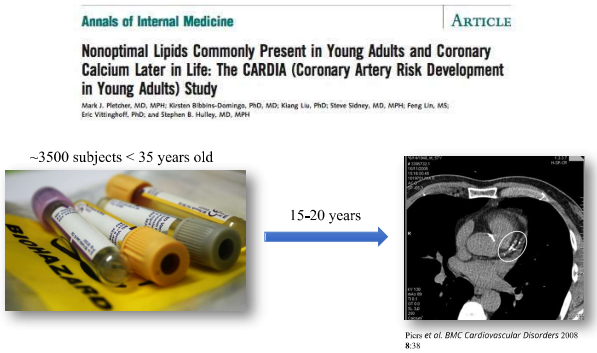
Coronary Risk Prediction in Adults (The Framingham Heart Study)

PETER W.F. WILSON, MD, WILLIAM P. CASTELLI, MD, and WILLIAM B. KANNEL, MD

The Framingham Heart Study, an ongoing prospective study of adult men and women, has shown that certain risk factors can be used to predict the development of coronary artery disease. These factors include age, gender, total cholesterol level, high density lipoprotein cholesterol level, systolic blood pressure, cigarette smoking, glucose intolerance and cardiac enlargement (left ventricular hypertrophy on electrocardiogram or enlarged heart on chest x-ray). Calculators and computers can be easily programmed using a multivariate logistic function that allows calculation of the conditional probability of cardiovascular events. These determinations, based on experience with 5,209 men and women participating in the Framingham study, estimate coronary artery disease risk over variable periods of follow-up. Modeled incidence rates range from <1% to >60% over an arbitrarily selected 6-year interval; however, they are typically <10%, and rarely exceed 45% in men and 25% in women.

(Am J Cardiol 1987;59:91G–94G)

LDL levels and risk of disease



Annals of Internal Medicine | ARTICLE

Nonoptimal Lipids Commonly Present in Young Adults and Coronary Calcium Later in Life: The CARDIA (Coronary Artery Risk Development in Young Adults) Study

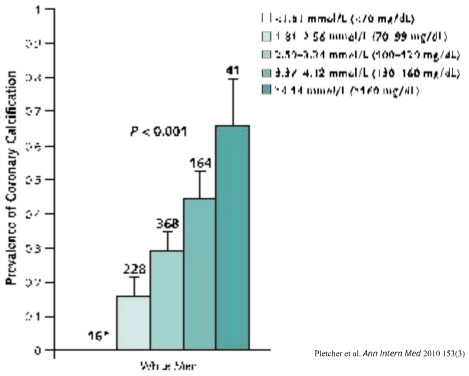
Mark J. Pletcher, MD, MPH; Kirsten Bibbins-Domingo, PhD, MD; Kiang Liu, PhD; Steve Sidney, MD, MPH; Feng Lin, MS; Eric Vittinghoff, PhD; and Stephen B. Hulley, MD, MPH

~3500 subjects < 35 years old

15-20 years

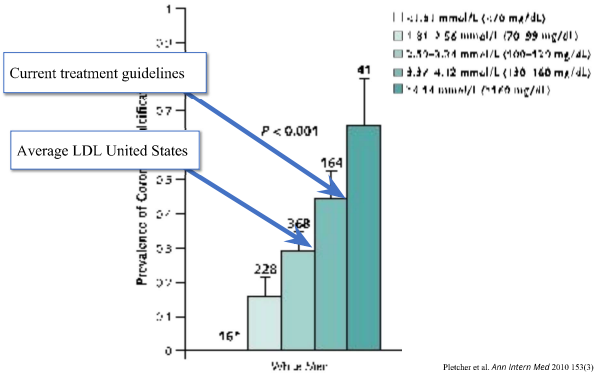
Pletcher et al. *Ann Intern Med* 2010 153(3)

LDL levels and risk of disease



Pletcher et al. *Ann Intern Med* 2010 153(3)

LDL levels and risk of disease



Pletcher et al. *Ann Intern Med* 2010 153(3)

Selecting populations for treatment

Why estimate genetic risk?

- An estimate of the long-term risk at birth
- Genetic risk can be combined with biomarkers and clinical features
- Genetics explains about 50% of risk. One cannot predict risk any better than that but 50% is a non-trivial proportion of risk

Applications in humans

Prediction of individual genetic risk to disease from genome-wide association studies

Naomi R. Wray, Michael E. Goddard and Peter M. Visscher

Genome Res. 2007 17: 1520-1528; originally published online Sep 4, 2007; Access the most recent version at doi:10.1101/gr.6665407

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

- LD-prune
- Exclude SNPs of very small effect

Extensions of BLUP – multiple variance scales and binary phenotypes

MultiBLUP:	Speed and Balding. *Genome Research* 2014
Bayesian analysis:	MacLeod et al. *Genetics* 2014
BSLMM:	Zhou et al. *PLOS Genetics* 2013
GeRSI:	Golan and Rossett. *AJHG* 2014

Methods that work with summary statistics

- Summary statistics are easily available
- Most methods require a separate small individual level dataset to tune parameters

16

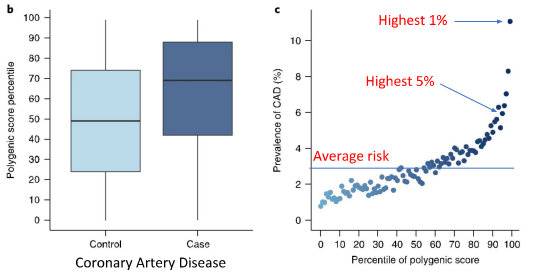
LDPred – a Bayesian method using summary statistics

$$\beta_i \sim_{iid} \begin{cases} N\left(0, \frac{h_g^2}{Mp}\right) \text{with probability } p \\ 0 \text{ with probability } (1-p), \end{cases}$$

Vilhjalmsson et al. 2015

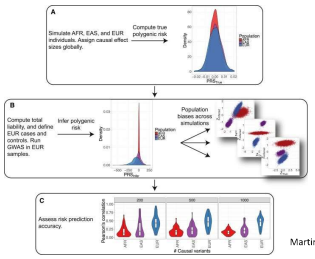
Also, check *BayesR*

Extreme tails in the distributions of genetic risk scores are highly predictive



Khera et al. 2018

With some caveats



Martin et al., *AJHG* 2017

Linear models for genetic risk prediction

$$y_i = \sum_j \beta_j \, x_{ij}$$

Genetic risk of individual i

Genotype of SNP j and individual i

Effect size of SNP j

"Polygenic scores" can leverage summary statistics from a large GWAS study

$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$

Predicted genetic risk

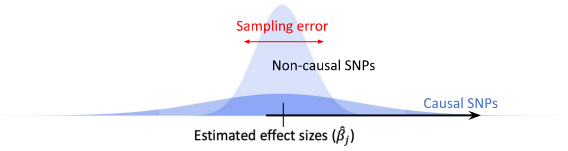
Estimated effect size

"Polygenic scores" can leverage summary statistics from a large GWAS study

$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$

Predicted genetic risk

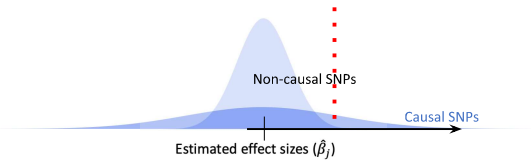
Estimated effect size



"Polygenic scores" can leverage summary statistics from a large GWAS study

P-value Thresholding

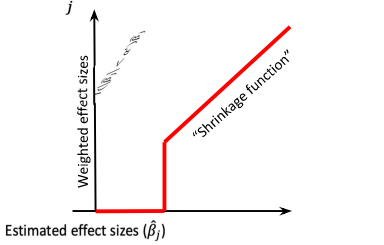
$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$



P-value thresholding can be reformulated as "shrinking" estimated effect sizes

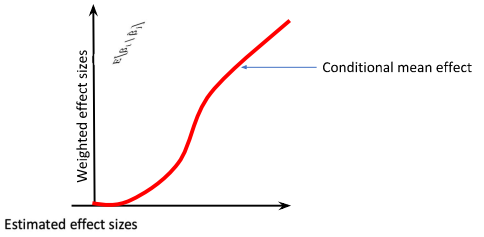
P-value Thresholding

$$\hat{y}_i = \sum_j I(|\hat{\beta}_j| < \alpha')\hat{\beta}_j x_{ij}$$



The optimal polygenic score can be constructed with "conditional mean effects"

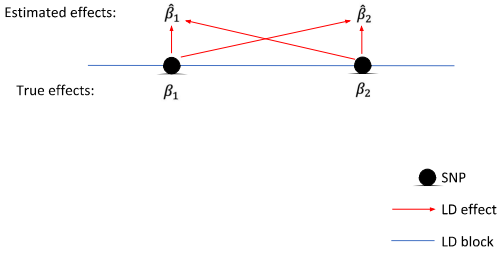
$$\hat{y}_i = \sum_j E[\beta_j \mid \hat{\beta}_j] x_{ij}$$



Goddard et al. 2009

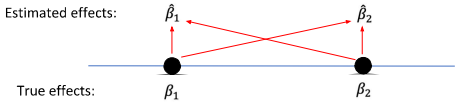
Accounting for LD in summary data is a major challenge

- Correlation between **apparent true genetic effects**



Accounting for LD in summary data is a major challenge

- Correlation between **apparent true genetic effects**



- Correlation between **sampling errors**

Our approach ("**N**on-**P**arametric **S**hrinkage" or NPS)

- No explicit specification of genetic architecture prior, thus "*non-parametric*"
- Learn conditional mean effects directly from training data
- Fully account for correlation in summary statistics

Our approach ("**N**on-**P**arametric **S**hrinkage" or NPS)

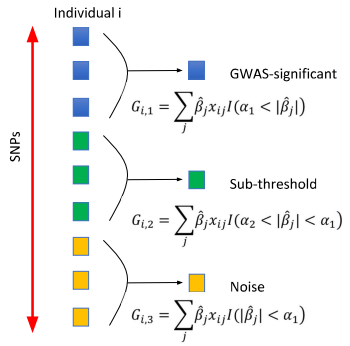
- No explicit specification of genetic architecture prior, thus "*non-parametric*"
- Learn conditional mean effects directly from training data

1. How to estimate $E[\beta_j \mid \hat{\beta}_j]$ without a Bayesian prior on $\boldsymbol{\beta}$

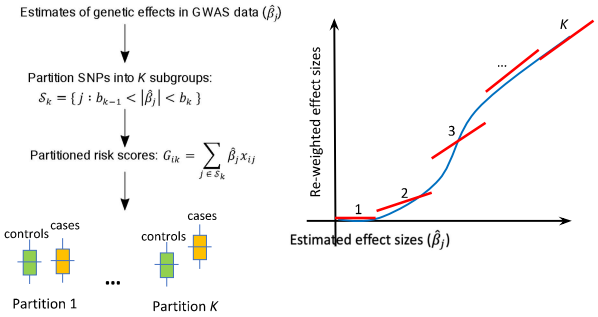
- Fully account for correlation in summary statistics

2. How to deal with LD

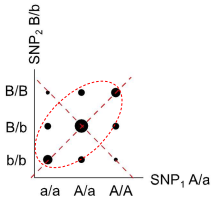
Partitioned risk scores



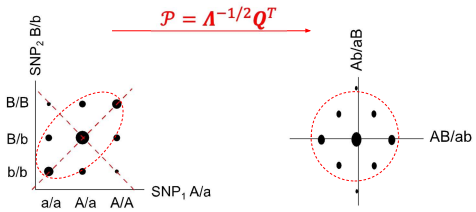
Piecewise linear interpolation on shrinkage curve



How to deal with LD?



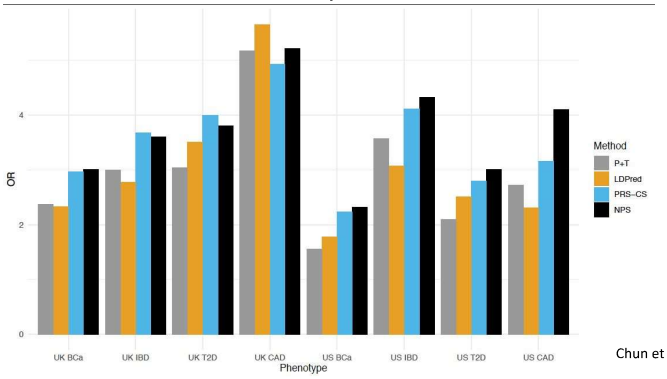
Decorrelating linear projection $\mathcal{P}$



$\boldsymbol{\Sigma}$ is a local LD matrix and $\boldsymbol{\Sigma} = \boldsymbol{Q}\,\boldsymbol{\Lambda}\,\boldsymbol{Q}^T$ by eigenvalue decomposition

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{Q}\,\boldsymbol{\Lambda}^{-1}\,\boldsymbol{Q}^T = (\boldsymbol{Q}\,\boldsymbol{\Lambda}^{-1/2})(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^T)$$

Accuracy of the 5% tail

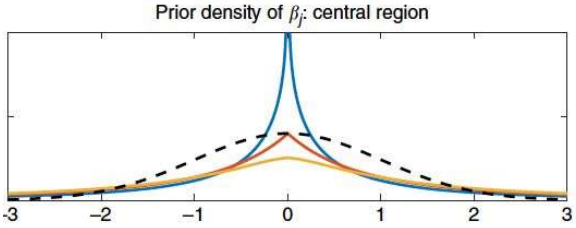


Chun et al. *AJHG* 2020

Other shrinkage methods: PRS-CS

$$\beta_j \sim \mathrm{N}\left(0, \frac{\sigma^2}{N}\phi\psi_j\right), \qquad \psi_j \sim g,$$

Prior density of β_j: central region

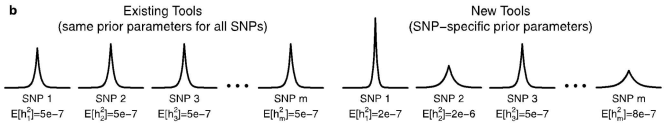


BayesR

$$\beta_j|\pi,\sigma_\beta^2 = \begin{cases} 0 & \text{with probability } \pi_1, \\ \sim N(0,\gamma_2\sigma_\beta^2) & \text{with probability } \pi_2, \\ \vdots & \\ \sim N(0,\gamma_C\sigma_\beta^2) & \text{with probability } 1-\sum_{c=1}^{C-1}\pi_c, \end{cases}$$

Lassosum – extension of *LASSO*

LDAK-Bolt-Predict

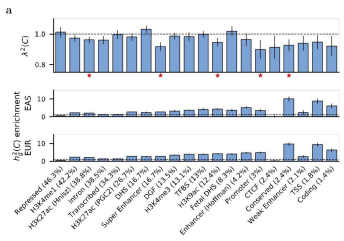


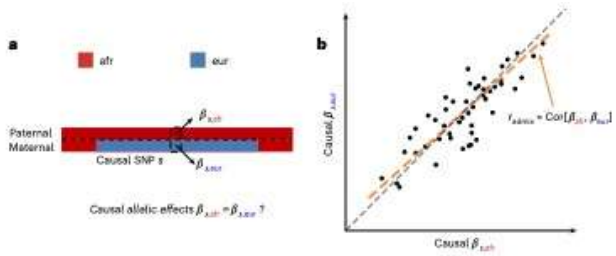
What makes PRS non-transferrable?

- Differences in allele frequencies between populations
- Differences in LD between populations
- Differences in effect sizes (although likely a minor contribution)

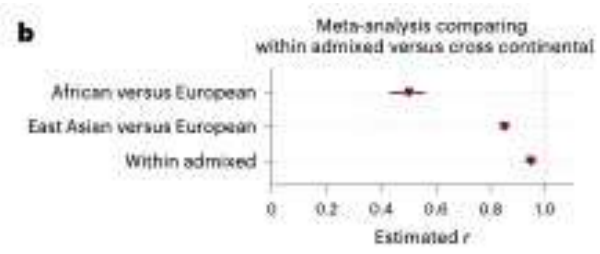
Slight differences in genetic effects between populations

Genetic correlations between populations are close but not equal to 1.
They are not uniformly distributed along the genome.



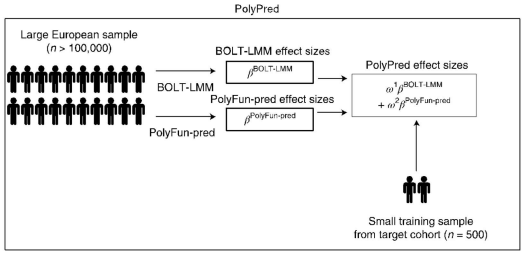


Hou et al., *Nature Genetics* 2023



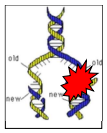
Hou et al., *Nature Genetics* 2023

PolyPred



Weissbrod *et al., Nature Genetics* 2022

Forces responsible for genetic change

Mutation  μ

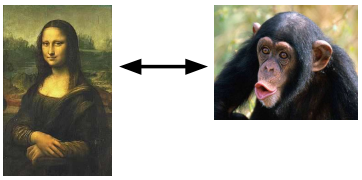
Selection  s

Drift  N_e

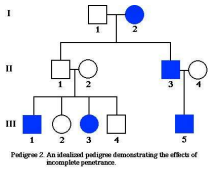
Population structure  F_{ST}

Mutations

Mutation rate in humans and flies



2.5x10^{-8} (Nachman & Crowell)



1.8x10^{-8} (Kondrashov)

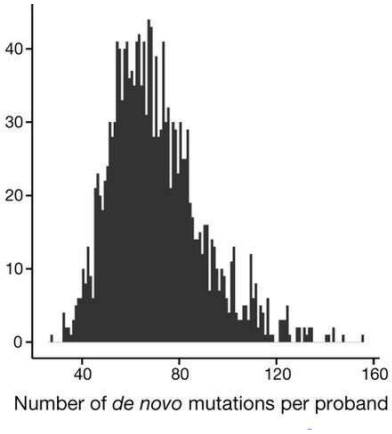
NGS estimates *~1.2X10*$^{-8}$ per nt changes genome

~70 per nt changes genome

Other events: indels (10^{-9}*)*

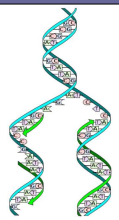
repeat extensions/contractions (10^{-5}*)*

Number of de novo mutations per individual

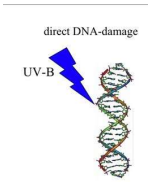


Jonsson et al., *Nature* 2017

Mutation rate is variable along the genome

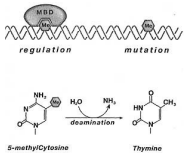


Replication fidelity



DNA damage

DNA repair



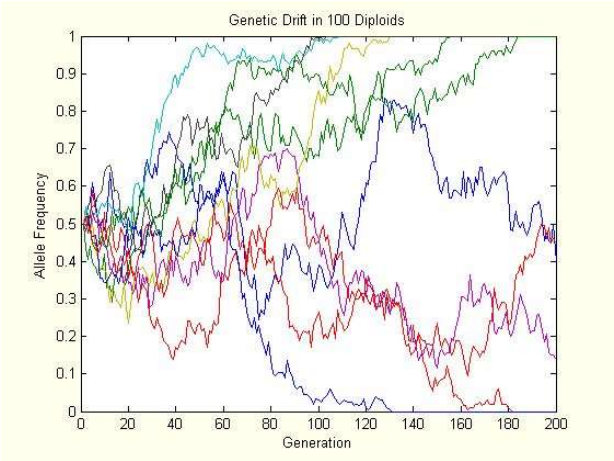
CpG deamination

Genetic drift

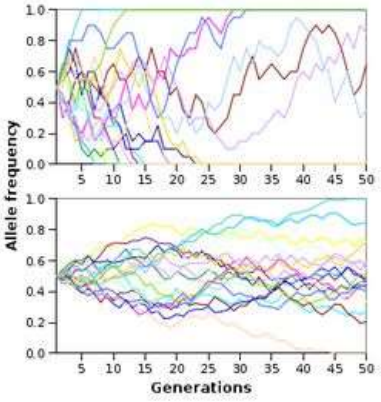
Regional variation of mutation rate

Context dependence of mutation rate

Drift is a random change of allele frequencies



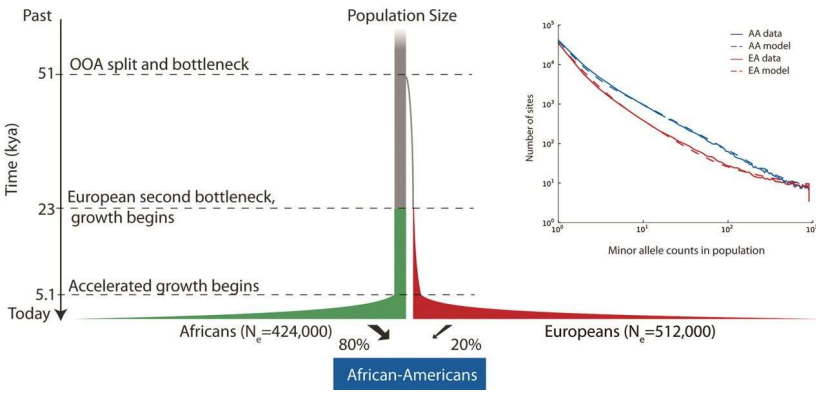
Drift depends on population size



Effective population size

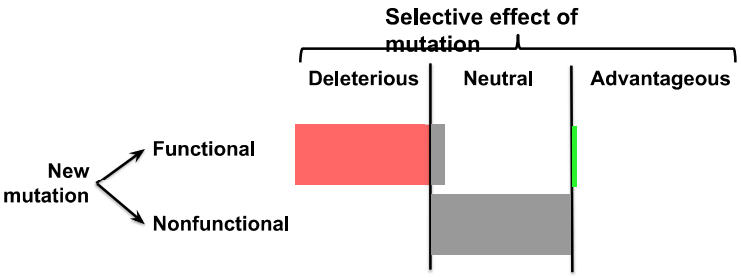
- In an idealized model, the intensity of genetic drift depends on population size (mean squared change in allele frequency is proportional to 1/Ne)
- In more realistic situations, effective population size (Ne) is a parameter characterizing intensity of drift

Demographic history



Selection

Most functional mutations are deleterious



Selection indicates functional mutations, whether or not the tested trait is under selection

Selection coefficient

- Selection coefficient (***s***) is the expected relative loss of fitness due to the sequence variant
- Variants with selection coefficients less than ~1/Ne are insensitive to selection. This is the drift barrier

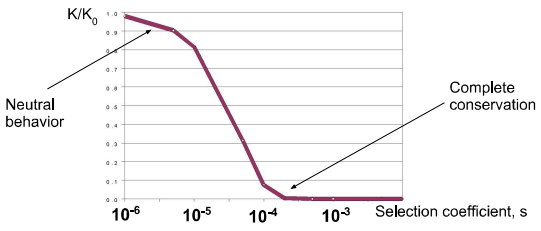
Conservation can be due to very weak selection!

Every new mutation eventually will be either fixed or lost

$$K = K_0 2 N_e \frac{(1-e^{-2s})}{(1-e^{-4N_e s})}$$

s – selection coefficient
N_e - effective population size

For humans estimated to be ~ 10 000



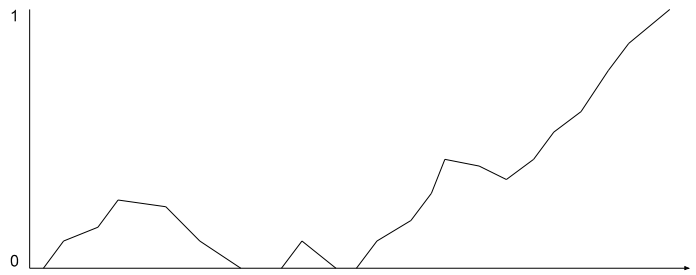
Basic facts about human genetic variation

- Nucleotide diversity (density of nucleotide differences between two randomly chosen chromosomes) is about 0.001
- Most common SNPs are very old (~300-400K years old)
- Protein coding regions are showing clear signs of selection (reduced diversity and excess of rare alleles)

Dynamic of allelic substitution

Mathematically, allele frequency change in a population follows a one-dimensional random walk

Methods of mathematical population genetics



Diffusion approximation

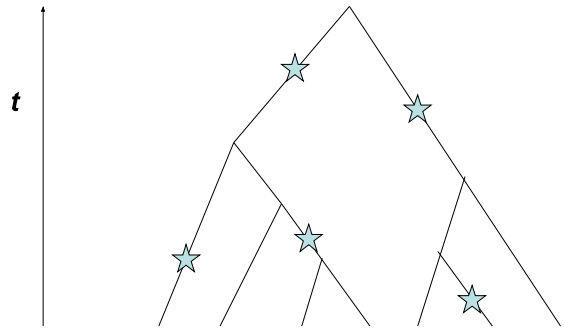
Random walk that does not jump long distances can be approximated by a diffusion process

$$\frac{\partial\phi(x,p,t)}{\partial t} = -\frac{\partial M\phi(x,p,t)}{\partial x} + \frac{1}{2}\frac{\partial^2 V\phi(x,p,t)}{\partial x^2}$$

Coalescent theory

Instead of modeling a population, we can model our sample

Time goes backwards !



Signatures of purifying selection

Reduced variation

Excess of rare alleles

Commonly used summary statistics to characterize variation

Number of segregating sites

```
. . . T C A A G T C A A G C G A T C A T G . . .
. . . T C A A G T C A A G C G A T C A G G . . .
. . . T C A G G T C A A G T G A T C A T G . . .
. . . T C A G G T C A A G T G A T C A T G . . .
. . . T C A A G T C A A G C G A T C A G G . . .
. . . T C A A G T C A A G C G A A C A G G . . .
```

k – number of sites variable in the sample

density of segregating sites is also frequently used

k is dominated by rare alleles

k strongly depend on sample size

Nucleotide diversity

$$\pi = \frac{2}{n(n-1)}\sum d_{ij}$$

d_{ij} - number of nucleotide differenced between sequences ***i*** and ***j***

$$\pi = \frac{n}{(n-1)}\sum 2p_k(1-p_k)$$

p_k – allele frequency at site ***k***

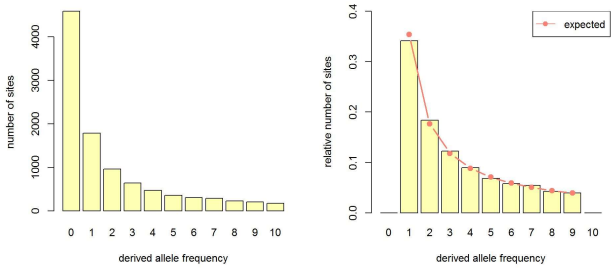
π – the average density of nucleotide differences between two sequences

π – per nucleotide heterozygosity

π is dominated by common alleles

π is independent of sample size

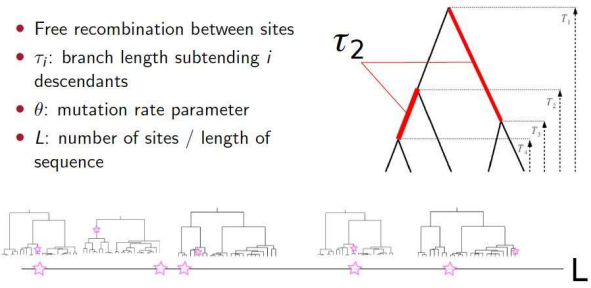
Site Frequency Spectrum (SFS)



SFS – expected number of variants at every frequency

A standard model of allele frequencies in a sample

- Free recombination between sites
- τ_i: branch length subtending i descendants
- θ: mutation rate parameter
- L: number of sites / length of sequence

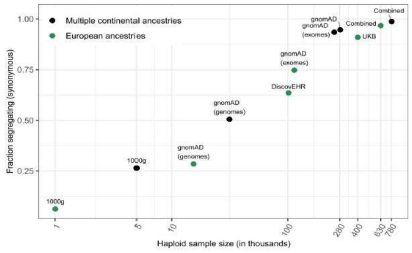


If every segregating site originated from just a single mutation, the distribution of allele frequencies (shape of SFS) does not depend on mutation rate!

Both π and k depend on mutation rate linearly!

Presence of recurrent mutations induces dependency of the shape of SFS on mutation rate!

- *Rapid recent growth of the human population*
  - *Rapid growth of available datasets*



Lek et al., Nature 2016
Harpak et al., PLOS Genetics 2016
Agarwal & Przeworski, eLife 2021

A mutation rate model at the basepair resolution identifies the mutagenic effect of Polymerase III transcription

Vladimir Seplyarskiy[1,2,*], Daniel J. Lee[1,2,*], Evan M. Koch[1,2,*], Joshua S. Lichtman[3], Harding H. Luan[3], Shamil R. Sunyaev[1,2]

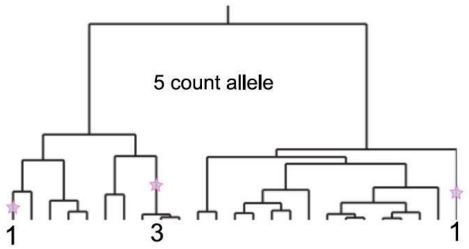
[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Brigham and Women's Hospital, Division of Genetics, Harvard Medical School, Boston, MA, USA
[3]NGM Biopharmaceuticals, South San Francisco, CA, USA
*Contributed equally

Recurrent mutation in the ancestry of a rare variant

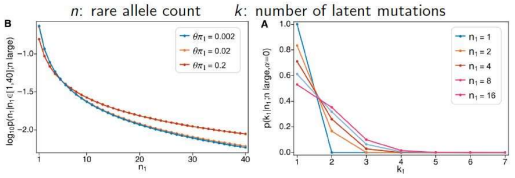
John Wakeley[1,*,†], Wai-Tong (Louis) Fan[2,3,†], Evan Koch[4,5], and Shamil Sunyaev[4,5]

[1]*Department of Organismic and Evolutionary Biology, Harvard University*
[2]*Department of Mathematics, Indiana University, Bloomington*
[3]*Center of Mathematical Sciences and Applications, Harvard University*
[4]*Department of Biomedical Informatics, Harvard Medical School*
[5]*Division of Genetics, Brigham and Women's Hospital, Harvard Medical School*
[†]*These authors contributed equally to this work.*
**Corresponding author: wakeley@fas.harvard.edu*

The effect of recurrent mutation



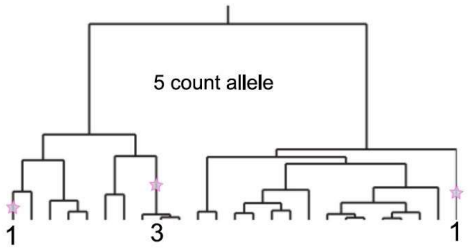
Constant Population Size



$$p(n) \propto \frac{\theta_{(n)}}{n!}$$

$$p(k|n) \propto \frac{S_n^{(k)}(\theta)^k}{\theta_{(n)}}$$

More generally, we can sum over latent mutations



Desai & Plotkin., Genetics 2008

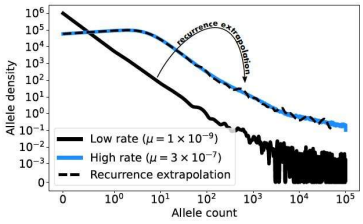
$$p(n) = \sum_{k=1}^{n} p(n|k)p(k)$$ • sum over recurrence

$$p(n|k) = \sum_{(i_1,\ldots,i_k)} \sum_{m=1}^{k} \frac{\mathrm{E}[\tau_i]}{\mathrm{E}[T_{total}]}$$ • sum over partitions, e.g. (n=5,k=3): 1+1+3, 2+2+1

$k \sim \mathrm{Poisson}(\theta \mathrm{E}[T_{total}])$ • latent mutations

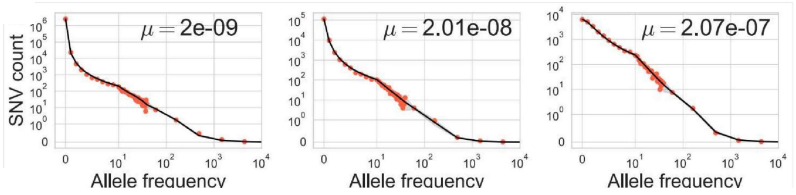
- $p(n)$: allele frequency distribution
- $p(k|n)$: recurrence distribution
- τ_i: total branch length with i descendants
- T_{total}: total size of the genealogy
- μ: mutation rate per generation
- θ: scaled mutation rate

Predict SFS for high mutation rate sites from low mutation rate sites



Estimate $\mathrm{E}[\tau_i]$ by assuming no recurrent mutations at low-rate sites.

This works very well on real data



In order to measure selection, we need a good handle on mutation rate!

RESEARCH

HUMAN GENETICS

Population sequencing data reveal a compendium of mutational processes in the human germ line

Vladimir B. Seplyarskiy[12]†, Ruslan A. Soldatov[2]†, Evan Koch[12], Ryan J. McGinty[12], Jakob M. Goldmann[3], Ryan D. Hernandez[45], Kathleen Barnes[6], Adolfo Correa[789], Esteban G. Burchard[910], Patrick T. Ellinor[11], Stephen T. McGarvey[121314], Braxton D. Mitchell[151617], Ramachandran S. Vasan[1819], Susan Redline[2021], Edwin Silverman[22], Scott T. Weiss[202122], Donna K. Arnett[23], John Blangero[2425], Eric Boerwinkle[2627], Jiang He[2829], Courtney Montgomery[30], D. C. Rao[31], Jerome I. Rotter[32], Kent D. Taylor[32], Jennifer A. Brody[33], Yii-Der Ida Chen[34], Lisa de las Fuentes[3135], Chii-Min Hwu[36], Stephen S. Rich[37], Ani W. Manichaikul[37], Josyf C. Mychaleckyj[37], Nicholette D. Palmer[38], Jennifer A. Smith[3940], Sharon L. R. Kardia[40], Patricia A. Peyser[40], Lawrence F. Bielak[40], Timothy D. O'Connor[414243], Leslie S. Emery[44], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium‡, TOPMed Population Genetics Working Group, Christian Gilissen[3], Peter V. Kharchenko[2], Shamil Sunyaev[12]*

The origin of human mutation in light of genomic data

Vladimir B. Seplyarskiy[1,2] and Shamil Sunyaev[1,2]✉

Features of mutation rate variation

Direction of transcription and replication (DNA repair recruitment)

Regional variation associated with replication timing

Methylation rate (CpG transitions)

Enzymatic demelythation rate (CpG transversions)

Regions mutagenic in arrested oocytes

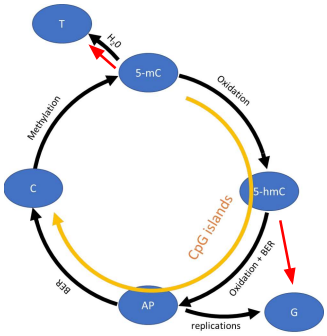
Mutagenic LINE elements

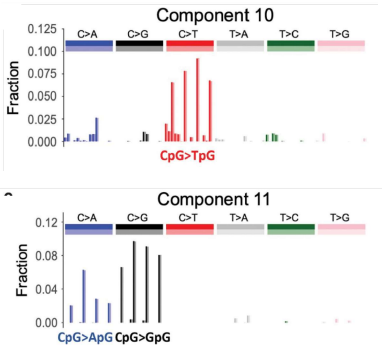
Sequence context

Transcription by RNA polymerase III

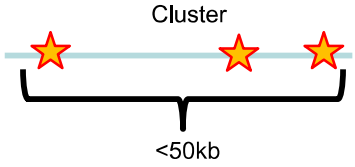
Transcription factor binding in testis

Deamination and demethylation

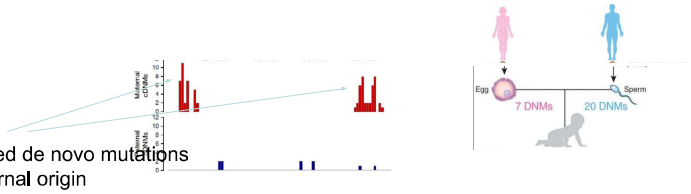




Oocyte-specific clusters

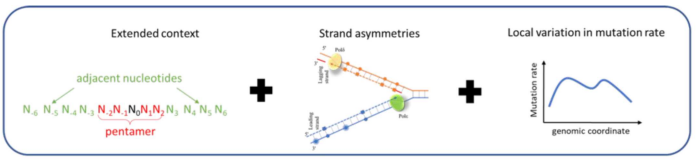


Oocyte-specific process



Clustered de novo mutations of maternal origin

Roulette: estimating mutation rate for each possible human mutation



At a given frequency deleterious and advantageous alleles are younger than neutral

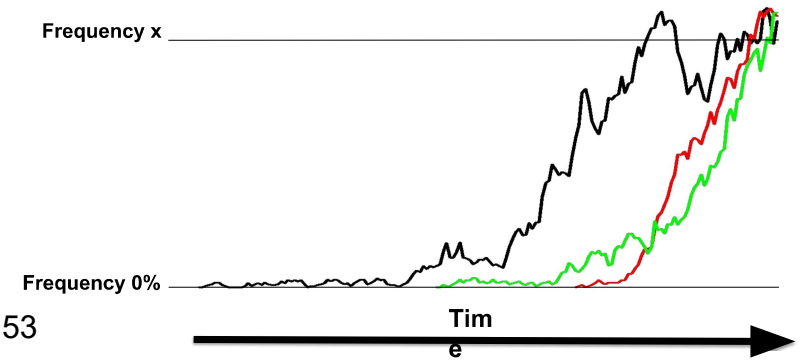
Am J Hum Genet 26:669–673, 1974

MAHLON V. R. FREEMAN, M.D.

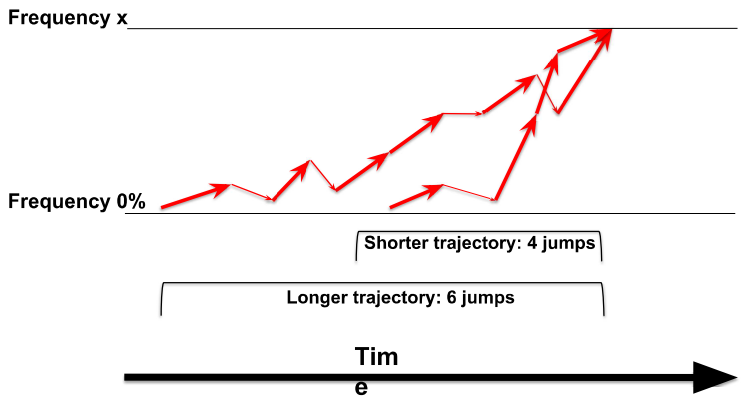
The Age of a Rare Mutant Gene in a Large Population

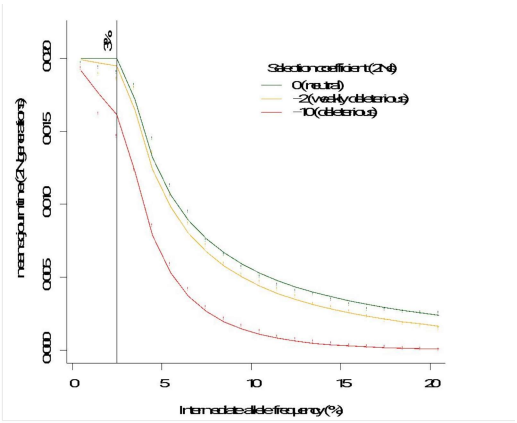
TAKEO MARUYAMA[1]

Maruyama effect (1974): at any frequency **advantageous** , or **deleterious** alleles are younger than **neutral** alleles



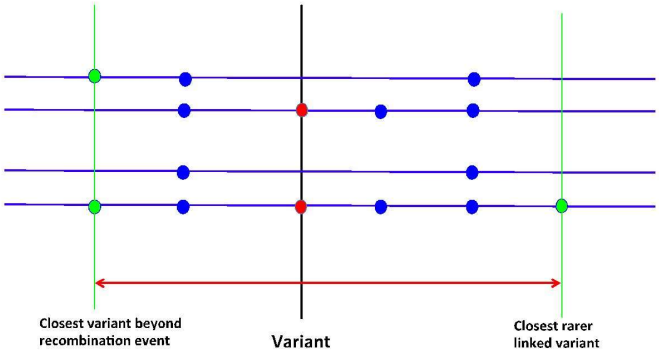
Intuition: shorter trajectories require fewer lucky jumps



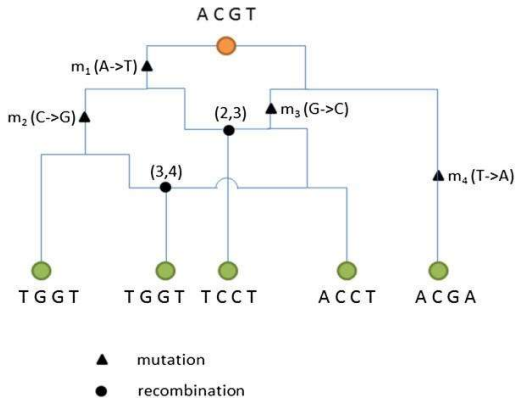


Kiezun et al. *PLOS Genetics* 2013

Neighborhood clock (fuzzy clock)



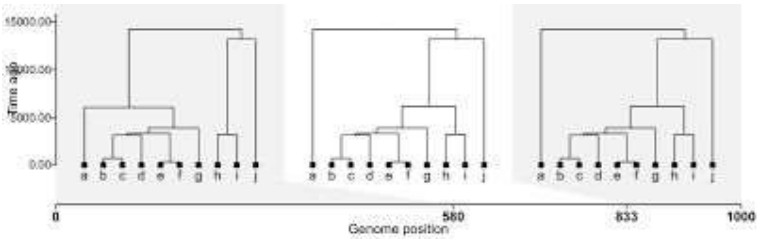
Ancestral Recombination Graph (ARG) is the full representation of the geneology



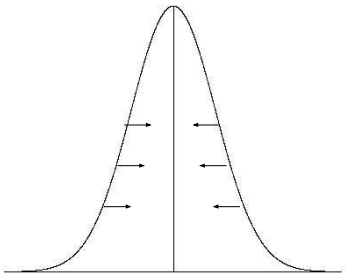
Tree sequences

Inferring whole-genome histories in large population datasets

Jerome Kelleher*, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers and Gil McVean



Stabilizing selection is the most common type of selection on a quantitative trait



Stabilizing selection

Selection may be related or unrelated to the trait

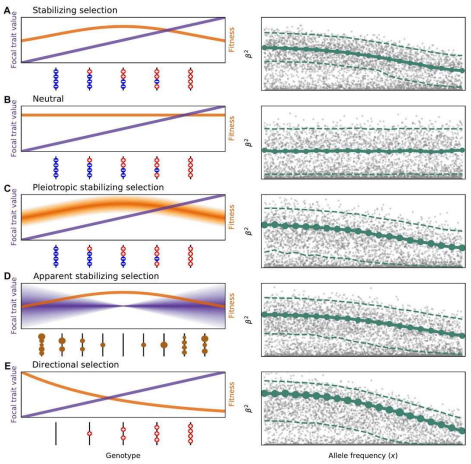
Technically, non-neutral genetic variation should not exist!

Forces to maintain variation:

Selection

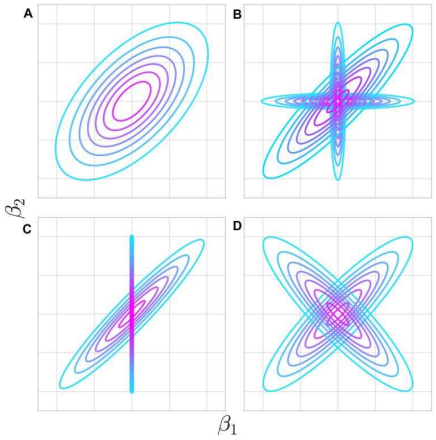
Mutation

Possible theoretical models



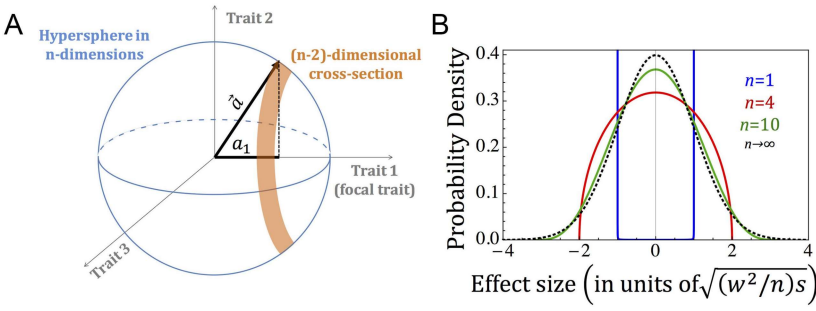
Koch & Sunyaev *Front. Genet.* 2021

Shades of pleiotropy



Koch & Sunyaev *Front. Genet.* 2021

A highly pleiotropic model



Simons et al., *PLOS Biology* 2018

Functional annotation of genes and variants

Map variants onto genomic annotation

Watch for multiple transcripts!

Watch for conflicting annotations!

Nonsense variants

One of most significant types of variants usually leading to the complete loss of function.

Nonsense variants are enriched in sequencing artifacts

Important considerations: i) location along the gene, ii) does the variant cause NMD? iii) is the variant in a commonly skipped exon?

Tool:
LOFTEE

Selection inference from frequency of individual SNVs

Change in allele frequency **=**

= ~~***Mutation***~~ **+** ***Selection*** **+** ***Drift***

Of the order of 10^{-8}

Demographic history

Population structure

Focusing on rare deleterious PTVs

PTV – protein truncating variant
(a.k.a. nonsense)

Combine all PTVs per gene – we assume that they have identical effects

Consider each gene as a bi-allelic locus – PTV / no PTV

Selection inference using combined frequency of PTVs

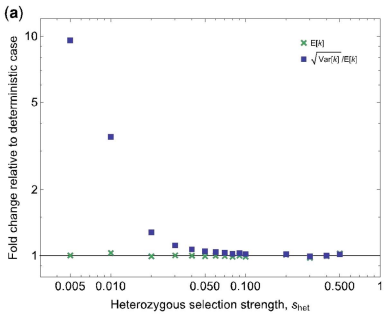
Change in allele frequency **=**

= ***Mutation*** **+** ***Selection*** **+** ~~***Drift***~~

Assuming string selection and a very large population, combined frequency of rare deleterious PTVs is expected to be Poisson distributed with $\lambda = U/hs$

Applicability of the Mutation–Selection Balance Model to Population Genetics of Heterozygous Protein-Truncating Variants in Humans

Donate Weghorn,[1,4,1,2] Daniel J. Balick,[1,1,2] Christopher Cassa,[1,2] Jack A. Kosmicki,[3,4] Mark J. Daly,[3,4] David R. Beier,[5,6] and Shamil R. Sunyaev*[,1,2]



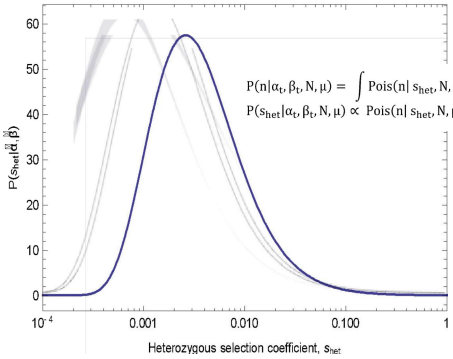
Loss-of-function observed/expected upper bound fraction (LOEUF)

- LOEUF is based on the number if segregating sites as the statistic
- LEOUF assumes Poisson distribution for the number of segregating sites. It computes the expectation. The constraint metric is based on the Poisson likelihood ratio upper bound.

Treating combined PTVs as a bi-allelic locus

- We can use the total frequency of PTVs in the locus
- Theoretically, we can simply treat all PTV variation as a single bi-allelic locus with high mutation rate

Distribution of selection coefficients



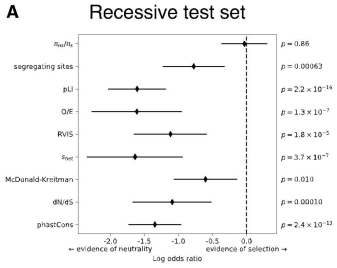
Cassa, Weghorn, Balick, Jordan et al. *Nature Genetics* 2017

Distribution of selection coefficients

1) The approach fails if selection is weak
2) The approach fails if mutational target is small
3) These considerations are important for regional constraint scores
4) Overall, the approach is non-informative in case of recessivity

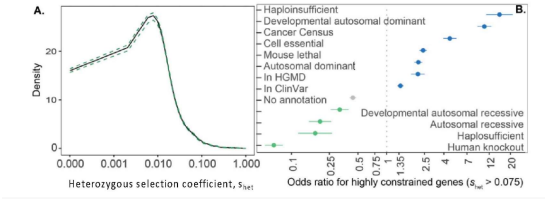
Overcoming constraints on the detection of recessive selection in human genes from population frequency data

Daniel J. Balick,[1,2,3,4,5] Daniel M. Jordan,[3,4,5] Shamil Sunyaev,[1,2,6,*] and Ron Do[3,4,6,*]



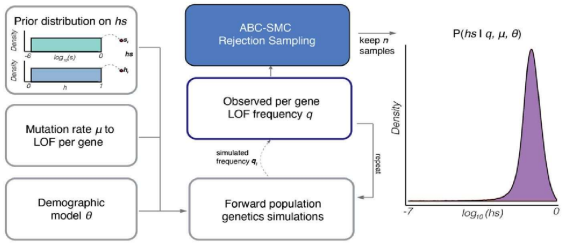
A deep catalog of protein-coding variation in 985,830 individuals

Kathie Y. Sun[1]*, Xiaodong Bai[1]*, Siying Chen[1], Suying Bao[1], Manav Kapoor[1], Joshua Backman[1], Tyler Joseph[1], Evan Maxwell[1], George Mitra[1], Alexander Gorovits[1], Adam Mansfield[1], Boris Boutkov[1], Sujit Gokhale[1], Lukas Habegger[1], Anthony Marcketta[1], Adam Locke[1], Michael D. Kessler[1], Deepika Sharma[1], Jeffrey Staples[1], Jonas Bovijn[1], Sahar Gelfman[1], Alessandro Di Gioia[1], Veera Rajagopal[1], Alexander Lopez[1], Jennifer Rico Varela[1], Jesus Alegre[2], Jaime Berumen[2], Roberto Tapia-Conyer[2], Pablo Kuri-Morales[2], Jason Torres[3], Jonathan Emberson[3,4], Rory Collins[3], Regeneron Genetics Center[15], RGC-ME Cohort Partners[5], Michael Cantor[1], Timothy Thornton[1], Hyun Min Kang[1], John Overton[1], Alan R. Shuldiner[1], M. Laura Cremona[1], Mona Nafde[1], Aris Baras[1], Goncalo Abecasis[1], Jonathan Marchini[1], Jeffrey G. Reid[1], William Salerno[1#], Suganthi Balasubramanian[1#]



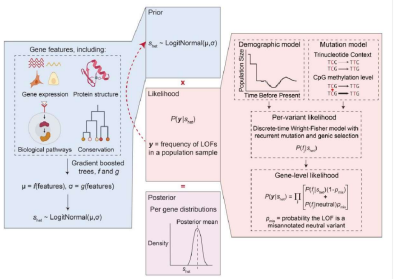
Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs

Ipsita Agarwal[1,2*†], Zachary L Fuller[1†], Simon R Myers[2,3], Molly Przeworski[1,4]

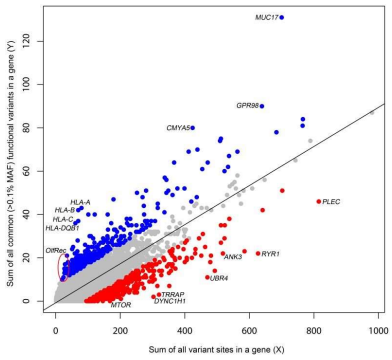


Bayesian estimation of gene constraint from an evolutionary model with gene features

Tony Zeng[1,*,†], Jeffrey P. Spence[1,*,†], Hakhamanesh Mostafavi[1], Jonathan K. Pritchard[1,2,†]

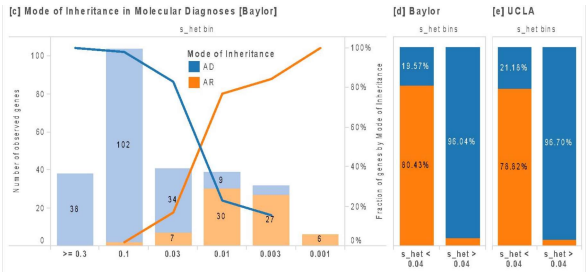


RVIS

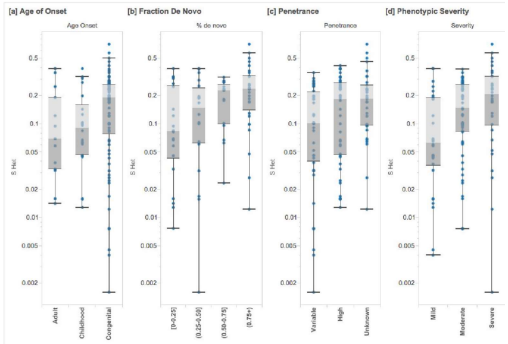


Petrovski et al. *PLOS Genetics* 2013

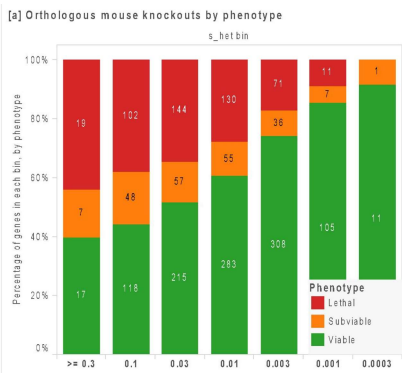
Dominant and recessive genes



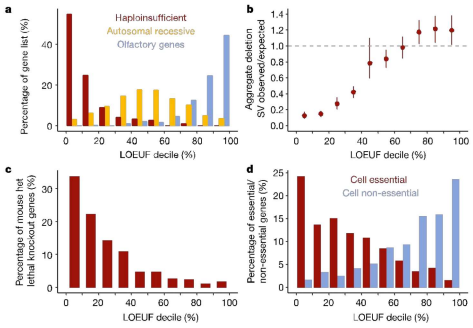
Age of onset, penetrance and severity



Concordance with the mouse knockout data



LOEUF (gnomAD)



Applications to Mendelian genetics –
large cohorts make Mendelian genetics a data science

Article

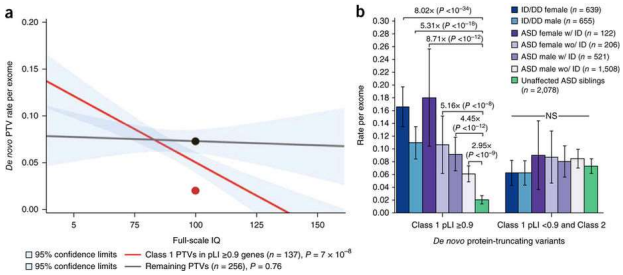
Evidence for 28 genetic disorders discovered by combining healthcare and research data

https://doi.org/10.1038/s41586-020-2832-5
Received: 8 October 2019
Accepted: 17 July 2020
Published online: 14 October 2020
Check for updates

Joanna Kaplanis[1,40], Kaitlin E. Samocha[1,40], Laurens Wiel[2,3,40], Zhancheng Zhang[4,40], Kevin J. Arvai[5], Ruth Y. Eberhardt[6], Giuseppe Gallone[6], Stefan H. Lelieveldt[7], Hillary C. Martin[1], Jeremy F. McRae[1], Patrick J. Short[1], Rebecca I. Torene[8], Elke de Boer[9], Petr Danecek[1], Eugene J. Gardner[1], Ni Huang[1], Jenny Lord[10], Iñigo Martincorena[1], Rolph Pfundt[11], Margot R. F. Reijnders[12], Alison Yeung[13], Helger G. Yntema[11], Deciphering Developmental Disorders Study*, Lisenka E. L. M. Vissers[11], Jane Juusola[8], Caroline F. Wright[14], Han G. Brunner[11,15,16], Helen V. Firth[1,17], David R. FitzPatrick[18], Jeffrey C. Barrett[1], Matthew E. Hurles[1,40] ✉, Christian Gilissen[2,40] & Kyle Retterer[8,40]

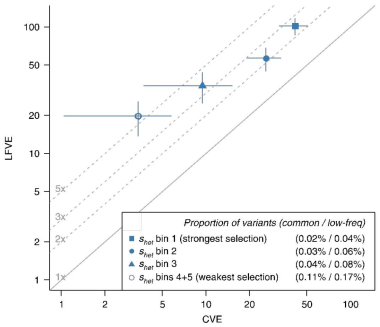
***DeNovo*WEST** – a method to identify significant recurrent *de novo* mutations controlling for mutation rate, weighting genes with s_{het} and weighting variants using variant effect predictors

De Novo mutations in ASD



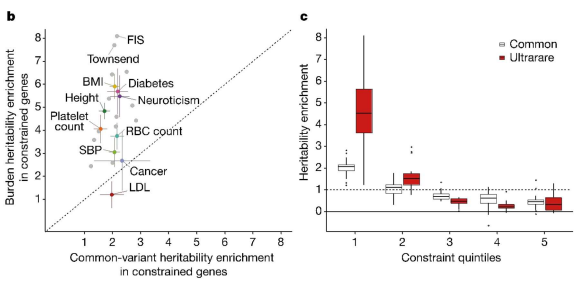
Kosmicki et al. *Nature Genetics* 2017

Heritability Enrichment



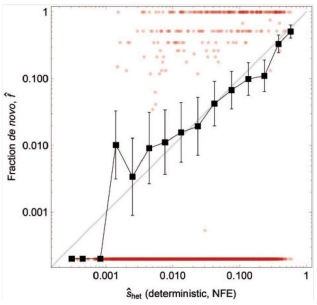
Gazal et al. *Nature Genetics* 2018

"Burden" heritability enrichment



Weiner, Nadig et al. *Nature* 2023

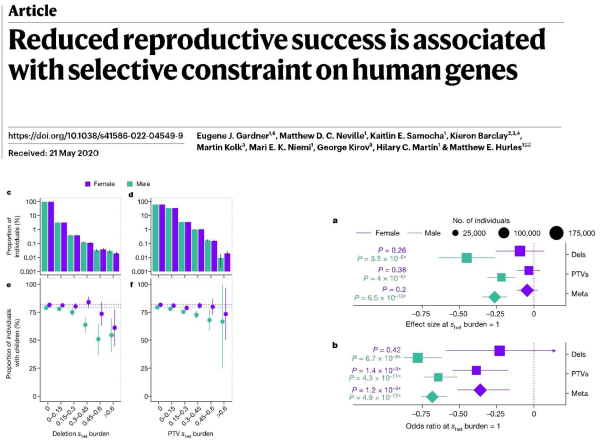
Selection in the present-day population



Fraction of *de novo* mutations (out of all variants) approximately equals selection coefficient.

This result does not depend on phenotypic ascertainment.

Weghorn et al., *MBE* 2019


Article

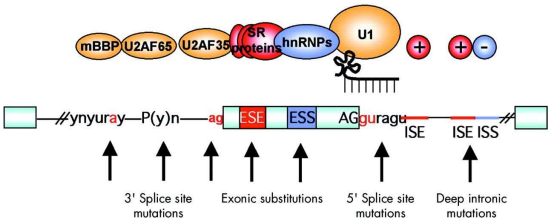
Reduced reproductive success is associated with selective constraint on human genes

https://doi.org/10.1038/s41586-022-04549-9

Received: 21 May 2020

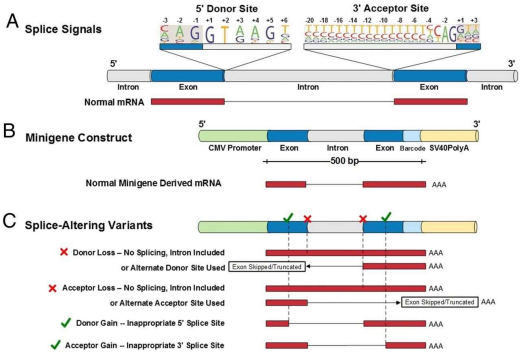
Eugene J. Gardner[1,4], Matthew D. C. Neville[1], Kaitlin E. Samocha[1], Kieron Barclay[2,3,4], Martin Kolk[3], Mari E. K. Niemi[3], George Kirov[3], Hilary C. Martin[1] & Matthew E. Hurles[1✉]

Variants involved in splicing

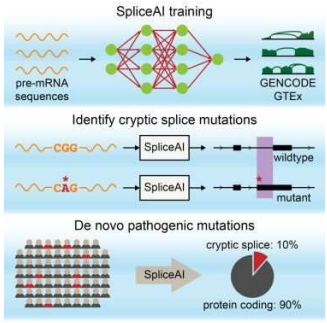


1) Variants in canonic splice sites
2) Variants in exonic or intronic splicing enhancers
3) Gain of splicing variants

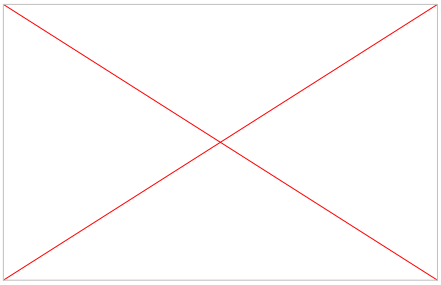
Experimental Methods: Minigene Assay and Massively Parallel Splicing Assay (MPSA)



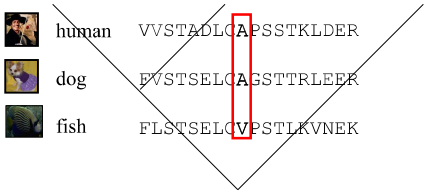
Computational Predictions: SpliceAI, Pangolin, MMSplice and other methods



Missense variants: computational predictions



Does the mutation fit the pattern of past evolution?

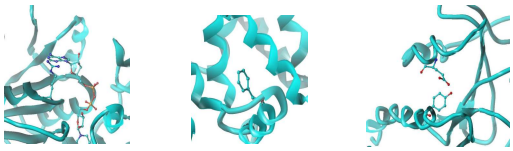


Statistical issues:
-sequences are related by phylogeny
-generally, we have too few sequences

Does the mutation fit the pattern of past evolution?

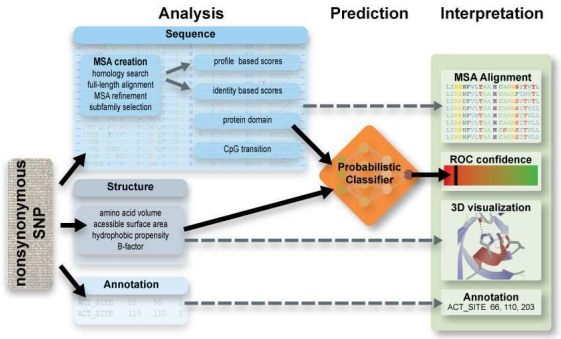
- We assume a constant fitness landscape: what is good for fish is good for human!
- We can estimate whether the mutation fits the pattern of amino acid changes.
- We can also estimate rate of evolution at the amino acid site

Protein structure view



- **Most of pathogenic mutations are important for stability (good news?).**
- ***ΔΔG* is difficult to estimate.**
- **Unfolded protein response pathway has to be taken into account.**
- **Heuristic structural parameters help but less than comparative genomics.**

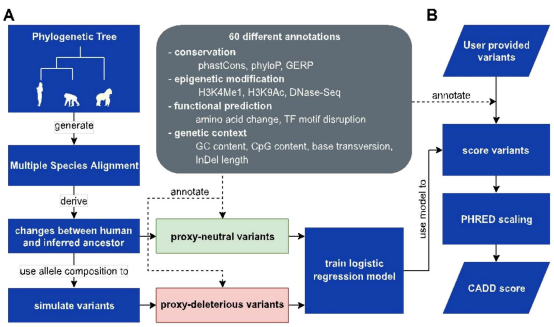
PolyPhen2



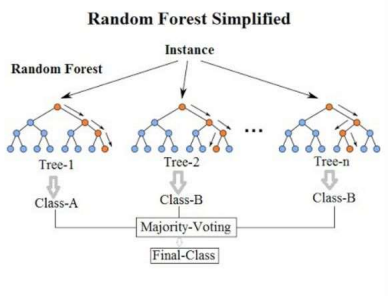
www.genetics.bwh.harvard.edu/pph2 Adzhubei, et al. Nature Methods 2010

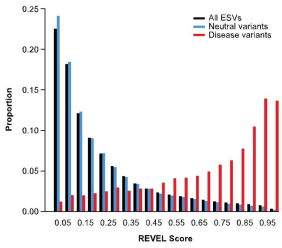
SIFT is based on multiple sequence alignment

Umbrella methods - **CADD**



Umbrella methods - **REVEL**





Umbrella methods

- **VEST4** – also an umbrella method using Random Forest
- **VARITY** – a new method using Gradient Boosting and focusing on de novo mutations and ultra rare variants

Weakly deleterious mutations

- Multiple independent lines of evidence suggest abundance of weakly deleterious alleles in humans
- Weakly deleterious variants may occur in highly conserved positions
- Weakly deleterious alleles probably contribute to complex phenotypes but not to simple Mendelian phenotypes

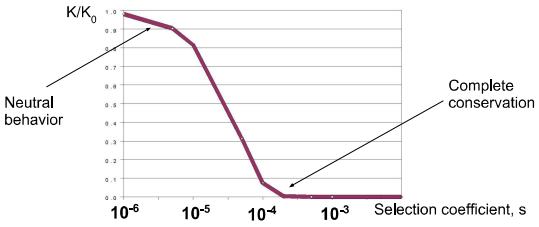
Conservation can be due to very weak selection!

Every new mutation eventually will be either fixed or lost

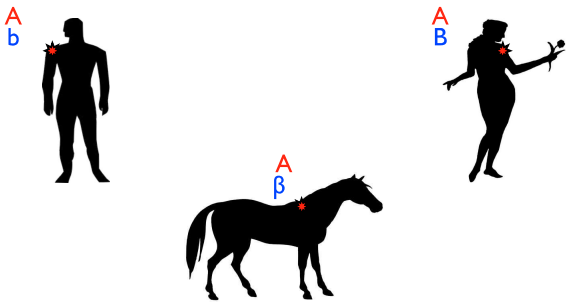
$$K = K_0 \, 2N_e \frac{(1-e^{-2s})}{(1-e^{-4N_e s})}$$

s – selection coefficient
N_e - effective population size

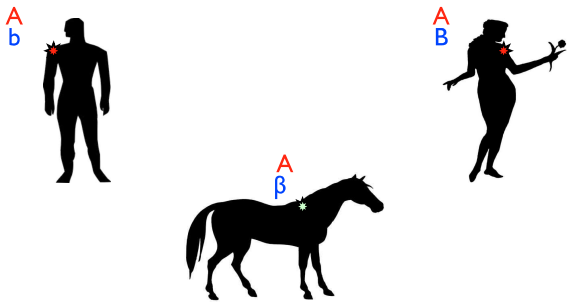
For humans estimated to be ~ 10 000



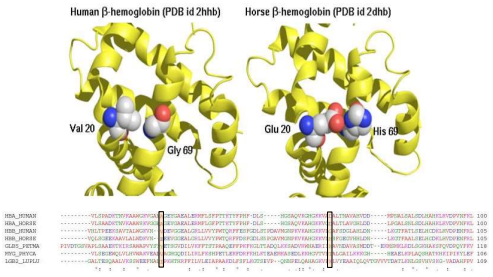
Constant fitness landscape



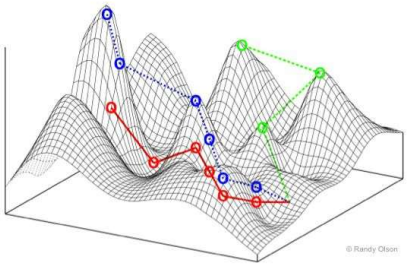
Epistatic interactions



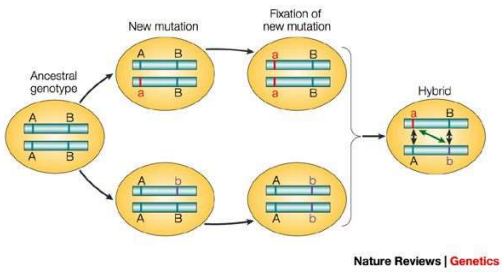
Compensatory mutations



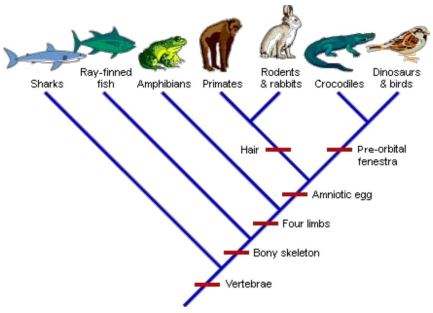
Ridges on the fitness landscape



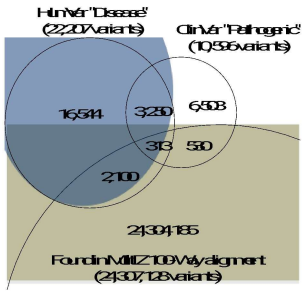
Dobzhansky-Muller incompatibility



Looking at vertebrate species



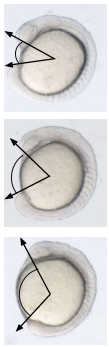
Many human pathogenic mutations are found in vertebrates



5.5-6.5% of presumably pathogenic human mutations are detected in mammals

Zebrafish model

- Model of Bardet-Biedl Syndrome (obesity, renal failure, vision loss)
- Caused by defects in primary cilium
- Embryonic convergence / extension phenotype in zebrafish
- Easily scorable phenotype



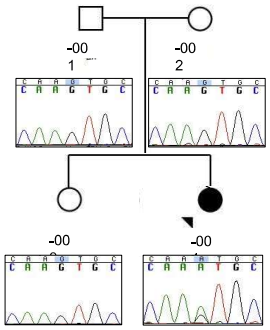
Images: Phoebe Liu

Testing double mutants

No injection		Human gene with disease mutant	
Knockdown		Double mutant (no suppression)	
Rescue with human gene		Double mutant (full suppression)	

Images: Phoebe Liu

A newly identified gene



Clinical features
Global developmental delay
microcephaly
feeding issues
failure to thrive
abnormal muscle tone
low immunoglobulins
frequent respiratory infections

Clinical testing
normal female microarray
metabolic testing – **negative**
extensive genetic testing – **negative**

BTG2 *De novo*	*TTN* Compound het
NOS2 *De novo*	*LAMA1* Compound het

Stephan Frangakis

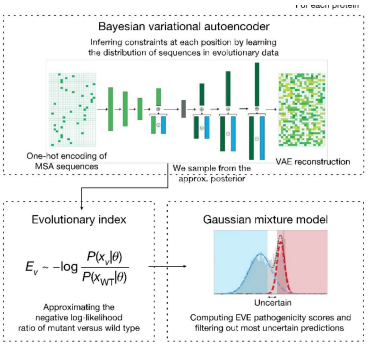
The mutation is a reversal to the mammalian ancestral state

BTG2	R80	L128	Q140	V141	L142
H. sapiens	R	L	Q	V	L
P. troglodytes	•	•	•	•	•
G. gorilla	•	•	•	•	•
M. musculus	K	V	•	M	M
R. norvegicus	K	V	•	M	M
H. glaber	•	V	•	M	M
S. domesticus	K	V	•	M	M
B. primigenius	K	V	•	M	M
E. ferus caballus	K	V	•	M	M
F. catus	K	V	•	M	M
C. lupus familiaris	K	V	•	M	M
D. novemcinctus	K	V	•	M	M
G. gallus	K	P	•	M	M

New methods directions

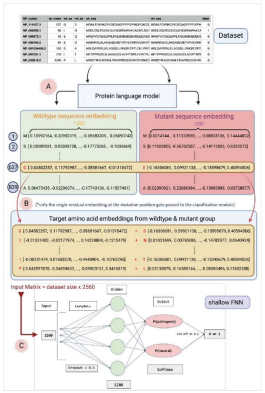
- Machine learning techniques have the potential to solve the epistasis problem
- Measures of population level constraint have the potential to solve the problem of distinguishing between strongly and weakly deleterious mutations.

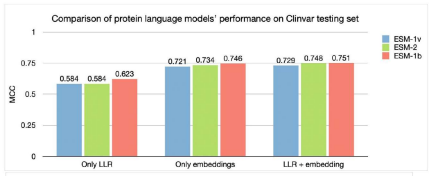
EVE – Variational Autoencoder



Frazer *et al.*, *Nature* 2021

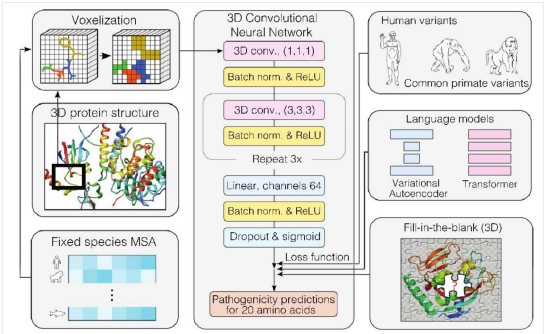
Large Language Models (**VariPred**)



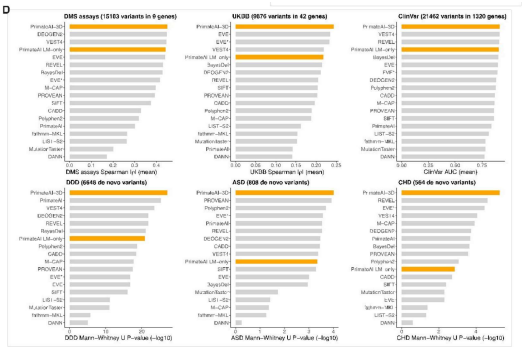


Lin *et al.*, *biorxiv* 2023

PrimateAI-3D

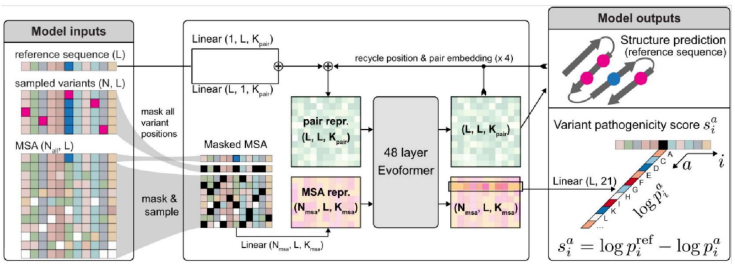


PrimateAI-3D



Gao *et al.*, *Science* 2023

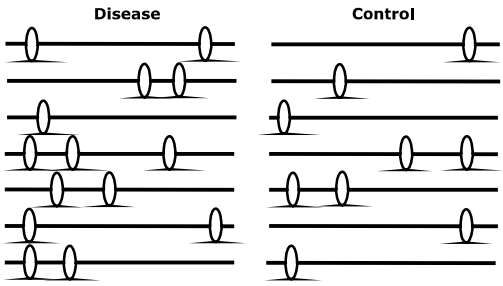
AlphaMissense



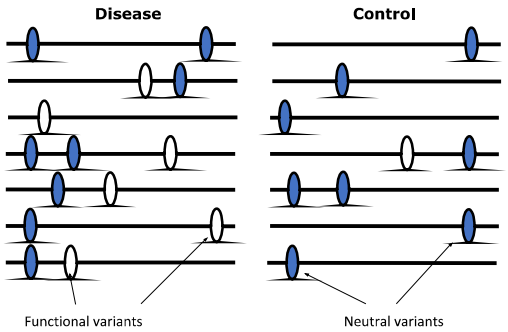
Applications

- Mendelian genetics
- Rare variant association studies

Rare variant collapsing study



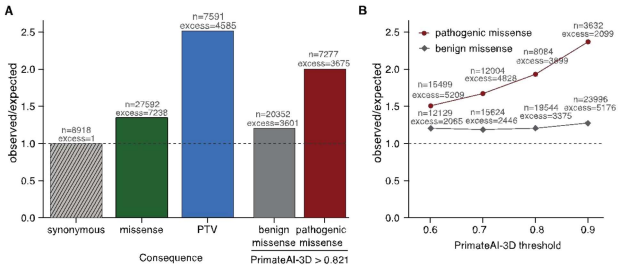
Rare variant collapsing study



Predicting functional consequences increases power

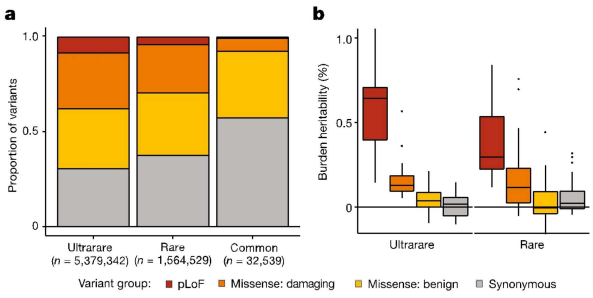
- Inclusion of neutral variants reduces power of the test
- Combining variants with vastly different effect sizes reduces power of the test
- Most groups limit the tests to nonsense, splicing and missense variants that are predicted functional
- Assigning quantitative weights is probably a better approach, but nobody uses it in practice

Damaging missense variants (as predicted by PrimateAI-3D) are enriched among de novo mutations in developmental disorders

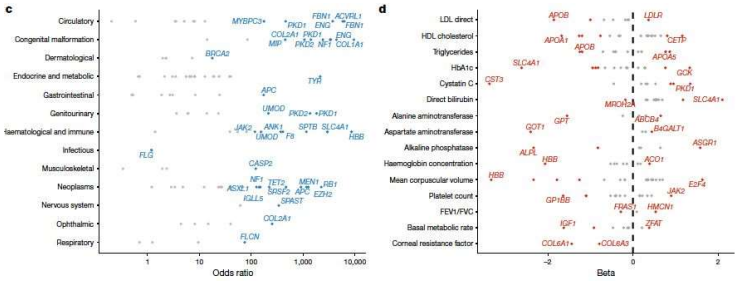


Gao et al., *Science* 2023

Burden heritability is significant for damaging missense variants (as predicted by *PolyPhen2*)

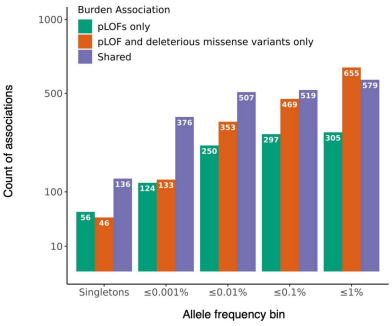


UK Biobank results (Wang et al.)



Variant grouping: nonsense, splicing, missense predicted by REVEL and MTR

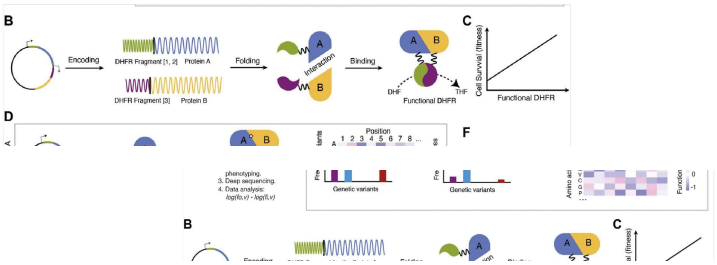
UK Biobank results (Backman et al.)



Deleterious missense variants:

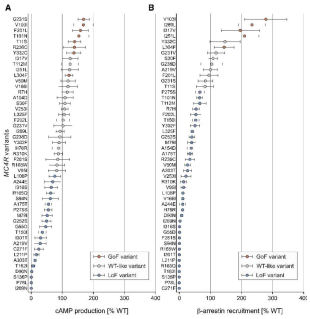
SIFT
PolyPhen2
LRT
Mutation Taster

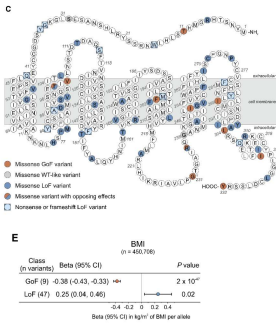
Experimental technologies – deep mutational scanning (DMS)



Wei & Li, *Frontiers in Genetics* 2023

MC4R example



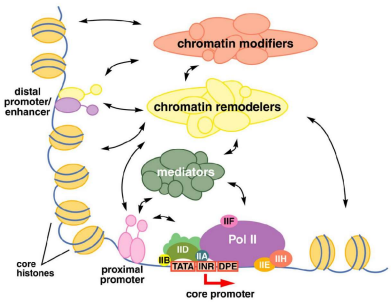


Lotta *et al.*, *Cell* 2023

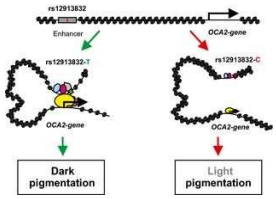
Non-coding variants

Regulatory variants

- Regulation: variants in promoters, enhancers, silencers, insulators

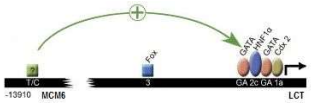


Non-disease alleles of large effect

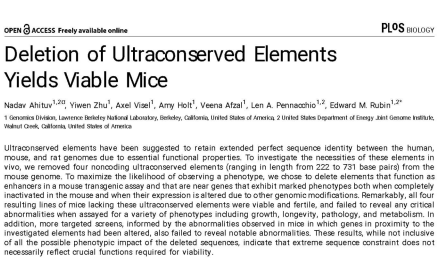


Eye color

Lactase tolerance



Ultraconserved elements



OPEN ACCESS Freely available online PLOS BIOLOGY

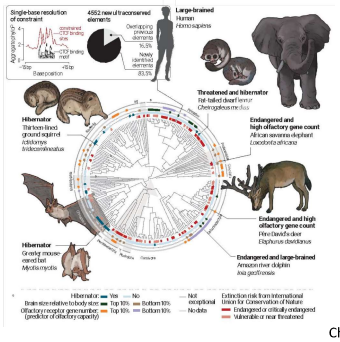
Deletion of Ultraconserved Elements Yields Viable Mice

Nadav Ahituv[1,2], Yiwen Zhu[1], Axel Visel[1], Amy Holt[1], Veena Afzal[1], Len A. Pennacchio[1,2], Edward M. Rubin[1,2]*

1 Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, 2 United States Department of Energy Joint Genome Institute, Walnut Creek, California, United States of America

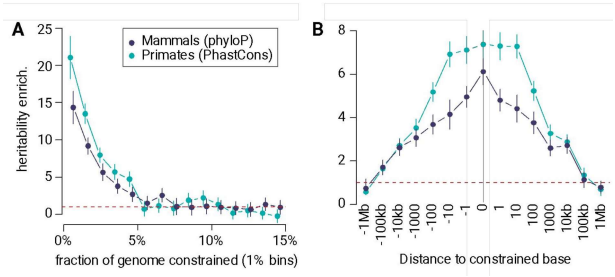
Ultraconserved elements have been suggested to retain extended perfect sequence identity between the human, mouse, and rat genomes due to essential functional properties. To investigate the necessities of these elements in vivo, we removed four noncoding ultraconserved elements (ranging in length from 222 to 731 base pairs) from the mouse genome. To maximize the likelihood of observing a phenotype, we chose to delete elements that function as enhancers in a mouse transgenic assay and that are near genes that exhibit marked phenotypes both when completely inactivated in the mouse and when their expression is altered due to other genomic modifications. Remarkably, all four resulting lines of mice lacking these ultraconserved elements were viable and fertile, and failed to reveal any critical abnormalities when assayed for a variety of phenotypes including growth, longevity, pathology, and metabolism. In addition, more targeted screens, informed by the abnormalities observed in mice in which genes in proximity to the investigated elements had been altered, also failed to reveal notable abnormalities. These results, while not inclusive of all the possible phenotypic impact of the deleted sequences, indicate that extreme sequence constraint does not necessarily reflect crucial functions required for viability.

Zoonomia conservation

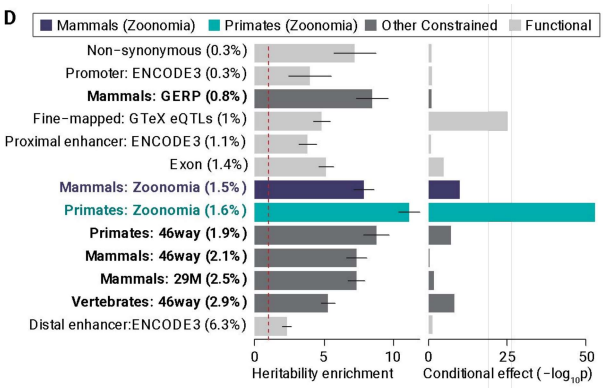


Christmas, Kaplow *et al.*, *Science* 2023

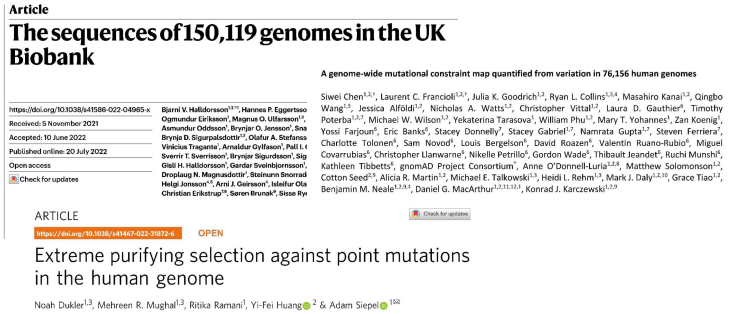
Heritability enrichment



Sullivan, Meadows *et al.*, *Science* 2023



Population constraint in non-coding regions



Article

The sequences of 150,119 genomes in the UK Biobank

A genome-wide mutational constraint map quantified from variation in 76,156 human genomes

https://doi.org/10.1038/s41586-022-04965-x
Received: 5 November 2021
Accepted: 10 June 2022
Published online: 20 July 2022
Open access
Check for updates

Bjarni V. Halldorsson[1,2,3], Hannes P. Eggertsson[4], Ogmundur Eiriksson[4], Magnus O. Ulfarsson[4], Asmundur Oddsson[4], Brynjar O. Jonsson[4], Sna Brynja D. Sigurpalsdottir[4], Olafur A. Stefansso Vinicius Tragante[1], Arnaldur Gylfason[4], Pall I. I Sverrir T. Sverrisson[4], Brynjar Sigurdsson[4], Sig Gisli H. Hatlidorsson[4], Gardar Sveinbjornsson[4], Droplaug N. Magnusdottir[4], Steinunn Snorrad Helgi Jonsson[1,4], Arni J. Geirsson[4], Isleifur Ola Christian Erikstrup[16], Soren Brunak[6], Sisse Ry

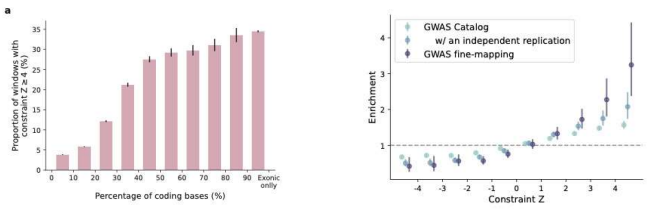
ARTICLE

https://doi.org/10.1038/s41467-022-31872-6 OPEN

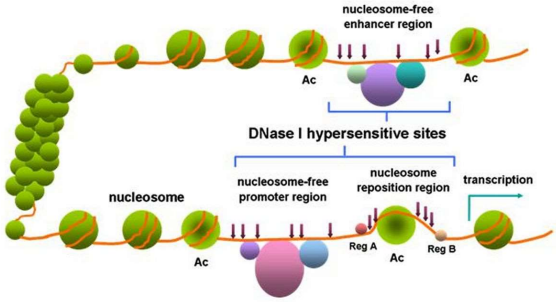
Extreme purifying selection against point mutations in the human genome

Noah Dukler[1,3], Mehreen R. Mughal[1,3], Ritika Ramani[1], Yi-Fei Huang[2] & Adam Siepel[1,2]

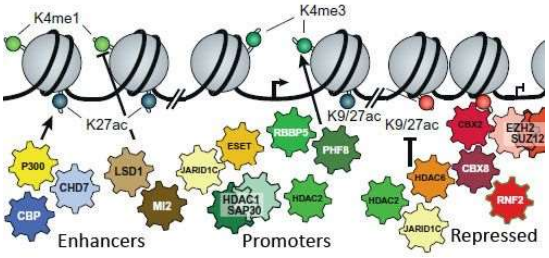
Population constraint in non-coding regions



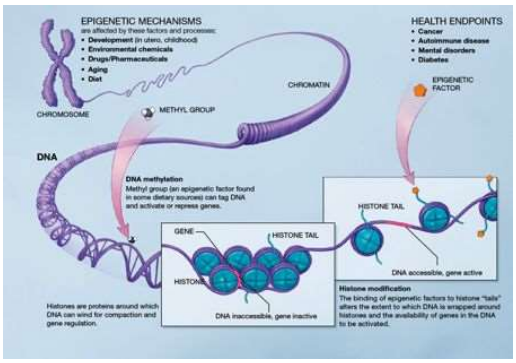
Chromatin accessibility



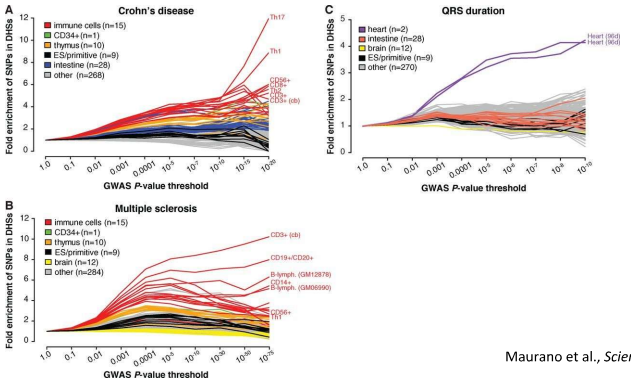
Chromatin modifications



Epigenomics

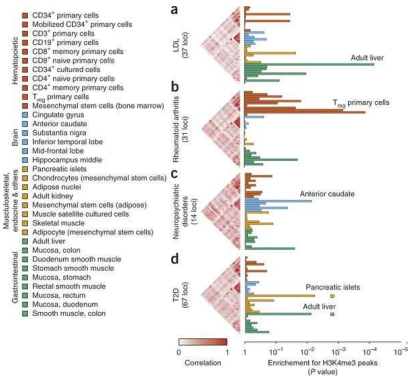


Enrichment of GWAS signals in regulatory elements



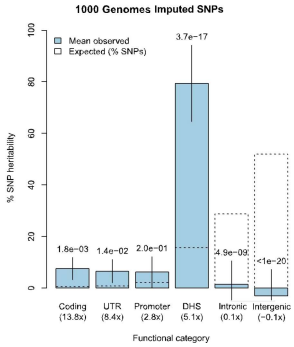
Maurano et al., *Science*, 2012

Enrichment of GWAS signals in regulatory elements



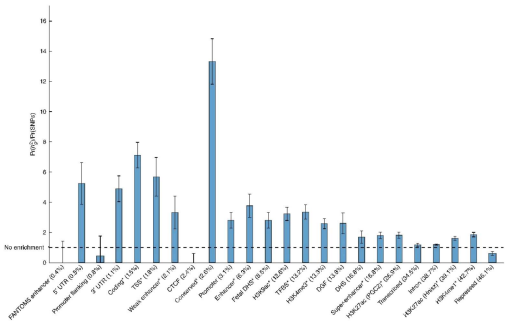
Trynka et al., *Nature Genetics*, 2014

Partitioning heritability



Gusev & Price, *AJHG*, 2014

Heritability partitioning across annotations



Finucane et al., *Nature Genetics*, 2015

Application – function informed fine-mapping

Functionally informed fine-mapping and polygenic localization of complex trait heritability

Omer Weissbrod[1,5], Farhad Hormozdiari[1], Christian Benner[2], Ran Cui[3], Jacob Ulirsch[3,4], Steven Gazal[1], Armin P. Schoech[1], Bryce van de Geijn[1], Yakir Reshef[1], Carla Márquez-Luna[2], Luke O'Connor[3], Matti Pirinen[2,6,7], Hilary K. Finucane[3,4] and Alkes L. Price[1,5]

- Estimate heritability enrichment and convert the estimates into prior probabilities
- Use these prior in fine-mapping (with SuSiE or FINEMAP)

Translating GWAS findings into mechanistic models

GWAS peak

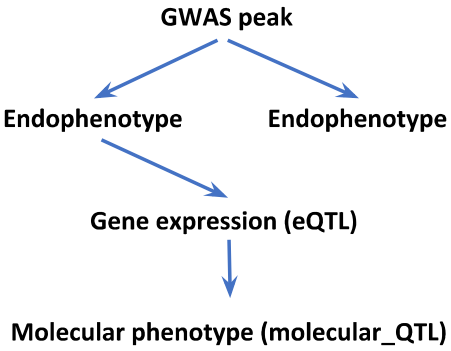
↓

Controlled model system

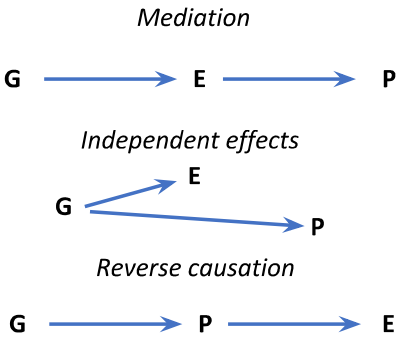
↓

Biochemistry

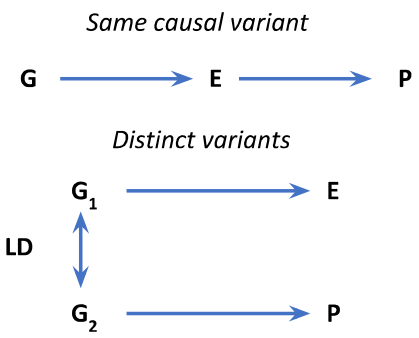
Human Genetics all the way



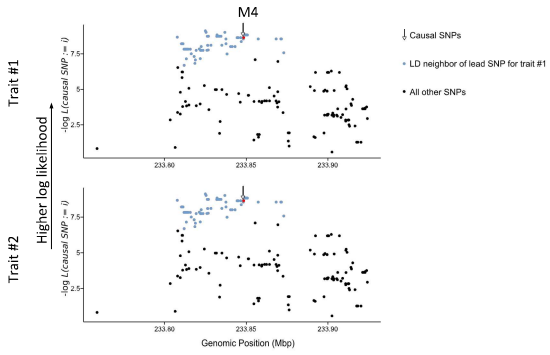
Causality



Co-localization



Co-localization problem



Methods

Coloc

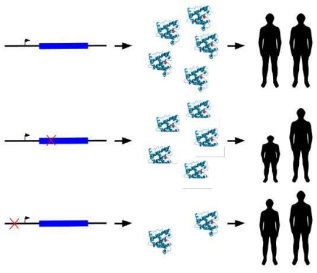
eCAVIAR

JLIM

Genetic variants differ between Mendelian and complex traits

- **Complex trait variants**
  - Small effect size
  - Extremely large number of loci
  - Mostly non-coding (regulatory)
- **Mendelian & somatic cancer variants**
  - Large effect sizes
  - Small number of loci
  - Mostly coding
  - Are in "putatively causative" genes

The basic model



https://commons.wikimedia.org/wiki/File:Myoglobin.png
https://commons.wikimedia.org/wiki/File:Human_outline_generic.svg

By now we know that most complex trait loci never harbor mutations of large effect

Hypothesis

- Most genes involved in Mendelian components of complex traits are also causative for cognate common forms.
- Variants involved in common forms alter regulatory sequence of these genes.
- This in turn induces changes in gene expression; regulatory variants are *eQTLs*.

Genes and phenotypes

(for complex traits, GWAS is restricted to non-coding variants)

Mend. trait	GWAS trait	Tissue
Breast cancer	Breast cancer	breast mammary tissue
Crohn disease	Crohn's disease	small intestine terminal ileum colon sigmoid colon transverse
Dyslipidemia Hyperlipidemia Tangier's disease	HDL	liver adipose whole blood
Dwarfism	Height	skeletal muscle
Blood pressure	Blood pressure	heart atrial appendage kidney heart left ventricle
Dyslipidemia Hyperlipidemia	LDL	liver adipose tissue whole blood
Monogenic diabetes	Type II diabetes	pancreas skeletal muscle adipose whole blood
Ulcerative colitis	Ulcerative colitis	small intestine terminal ileum colon sigmoid colon transverse

Overall, 139 genes

89 (64%) fall under a GWAS peak of a cognate complex trait

Examples include:

LDL Receptor under a GWAS peak for LDL Cholesterol

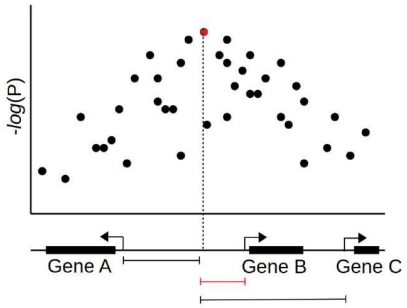
Estrogen receptor under a GWAS peak for breast cancer

These genes are highly likely to mediate the effects of regulatory variants

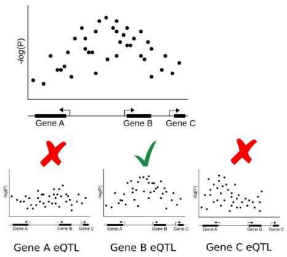
Statistical methods to locate the causative gene under GWAS peak

- Closest gene to peak
- Colocalization methods
  - JLIM
  - Coloc
  - eCAVIAR
- Transcriptome-wide association
  - FUSION
- Chromatin marks
  - Fine-mapping using SuSiE
  - Locate fine-mapped variants under chromatin modification peaks

Distance of fine-mapped SNPs (by SuSiE) to the closest gene

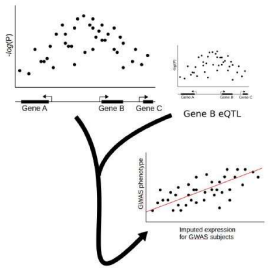


Colocalization of GWAS and eQTLs



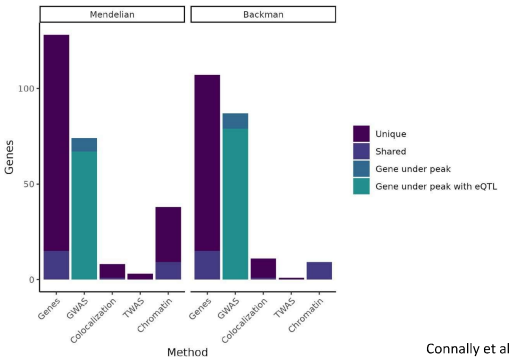
Methods effectively compare the shape of two peaks.
Colocalization often returns multiple hits per locus.

Transcriptome-wide association (TWAS)



TWAS often returns multiple hits per locus.

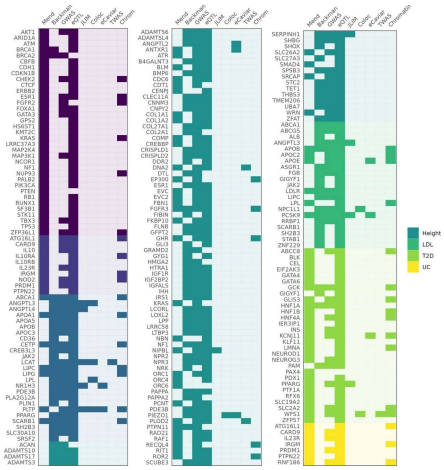
Results



Connally et al., *eLife*, 2022

Our curated genes rarely colocalize

- This is true across all tested traits
- We also tried a chromatin method
  - It worked better
  - In large part because it favors the closest gene



But *why*?

Are eQTLs specific to…

- certain cell types?
- certain developmental stages?
- certain environmental conditions?

Are there inconsistent relationships…

- between gene expression and protein levels?
- between rate of transcription and gene expression?

I find it highly surprising that

- A context independent large change in expression of LDLR due to a nonsense mutation leads to a large phenotypic change

- A smaller change in expression does not affect LDL levels, while non-coding effect on LDLR does

Quantifying genetic effects on disease mediated by assayed gene expression levels

Douglas W. Yao[1✉], Luke J. O'Connor[1,2,3], Alkes L. Price[2,3,4] and Alexander Gusev[3,5,6✉]

Feature Review

Where Are the Disease-Associated eQTLs?

Benjamin D. Umans,[1,*] Alexis Battle,[2,3,*] and Yoav Gilad[1,4,*]

Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery

Hakhamanesh Mostafavi[*,1], Jeffrey P. Spence[1], Sahin Naqvi[1,2], Jonathan K. Pritchard[*,1,3]

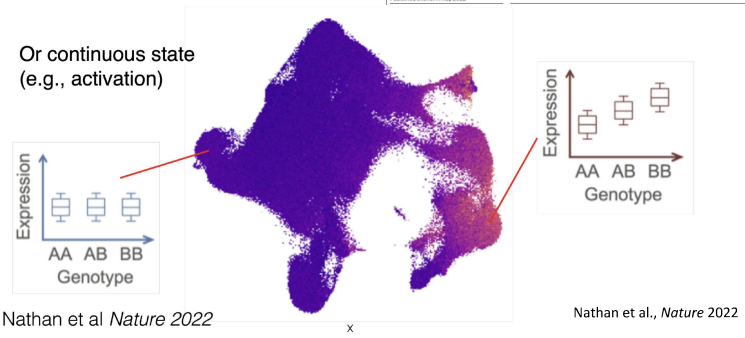
Modeling eQTL effects at single cell resolution

Article

Single-cell eQTL models reveal dynamic T cell state dependence of disease loci

https://doi.org/10.1038/s41586-022-04713-1
Received: 17 August 2021
Accepted: 31 March 2022
Published online: 11 May 2022

Aparna Nathan[1,2,3,4], Samira Asgari[1,2,3,4], Kazuyoshi Ishigaki[1,2,3,4], Cristian Valencia[1,2,3,4], Tiffany Amariuta[1,2], Yang Luo[1,2,3,4], Jessica I. Beynor[1,2,3,4], Yuriy Baglaenko[1,2,3,4], Sara Suliman[1], Alkes L. Price[2,3,4], Leonid Lecca[5,6], Megan B. Murray[5,6], D. Branch Moody[7] & Soumya Raychaudhuri[1,2,3,4,8,9]✉



Nathan et al *Nature 2022*

Nathan et al., *Nature* 2022