Genome Wide Association Studies

Postdoctoral Researcher

Bioinformatics Research Centre

Shannon D’Urso

What do you know about these topics already?

What is heritability?

How can we identify genetic variants that contribute to a phenotype?

Why do we care?

Overview of this lecture

- Genetic Linkage studies
- Transmission disequilibrium tests
- Candidate gene studies
- GWAS
- Imputation
- GWAS QC
- GWAS Results
- Multiple testing
- Population stratification

Types of Genetic Variation

- Chromosomal abnormalities
- Insertions/deletions
- Microsatellites
- Copy number variation
- Single nucleotide polymorphisms (SNPs)
- And others!

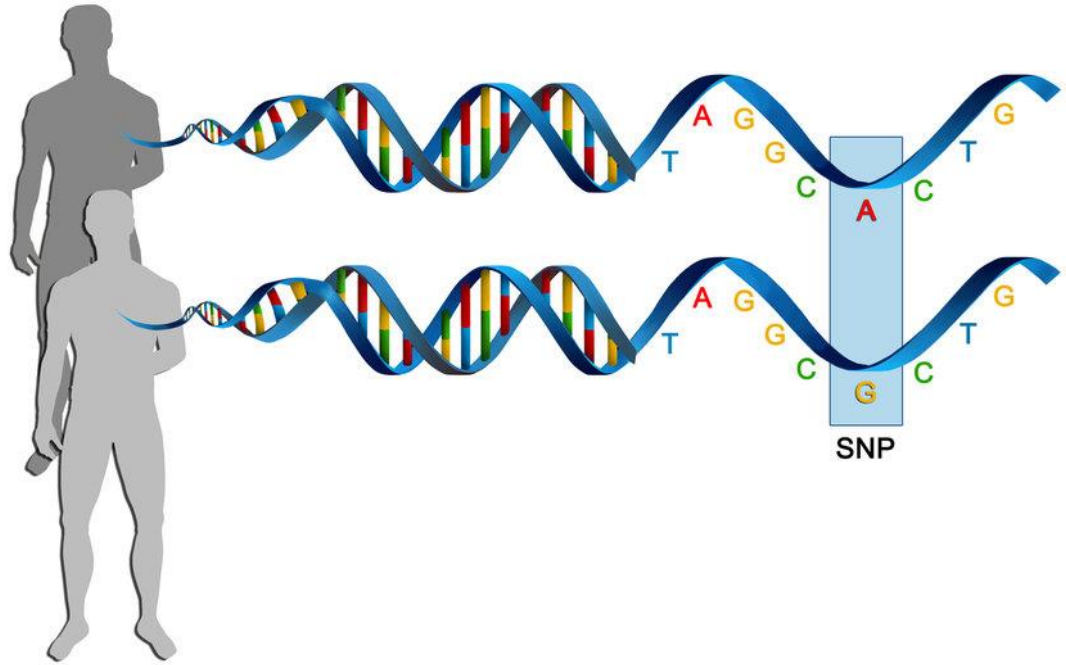


Image credit: Pereita et al., 2022 DOI 10.31219/osf.io/d96z3

- Alleles = variants at the locus (A and G)
- Genotypes = count of alleles at a locus (e.g. AA, AG or GG; 0, 1, 2)
- There are millions of known SNPs (e.g. rs12345)

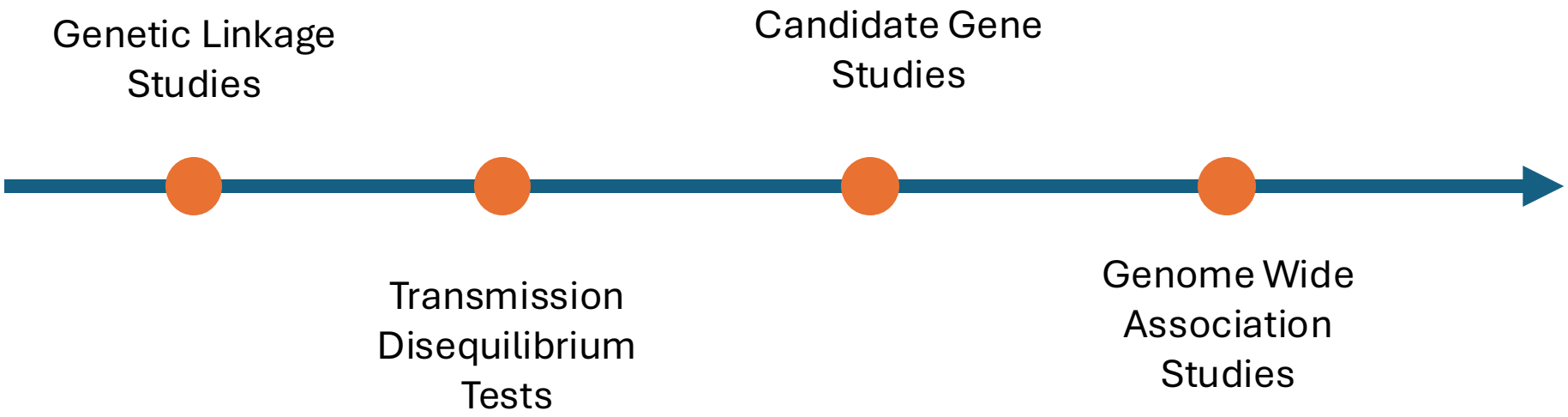
Monogenic vs Complex Traits

- Monogenic Traits (*mono = one, genic = gene*)
  - Typically explained by one or a few genes
  - Limited influence of environmental factors
  - *Dominant* or *recessive*
  - *e.g.* cystic fibrosis & sickle cell disease

	B	b
b	Bb	Bb
b	bb	bb

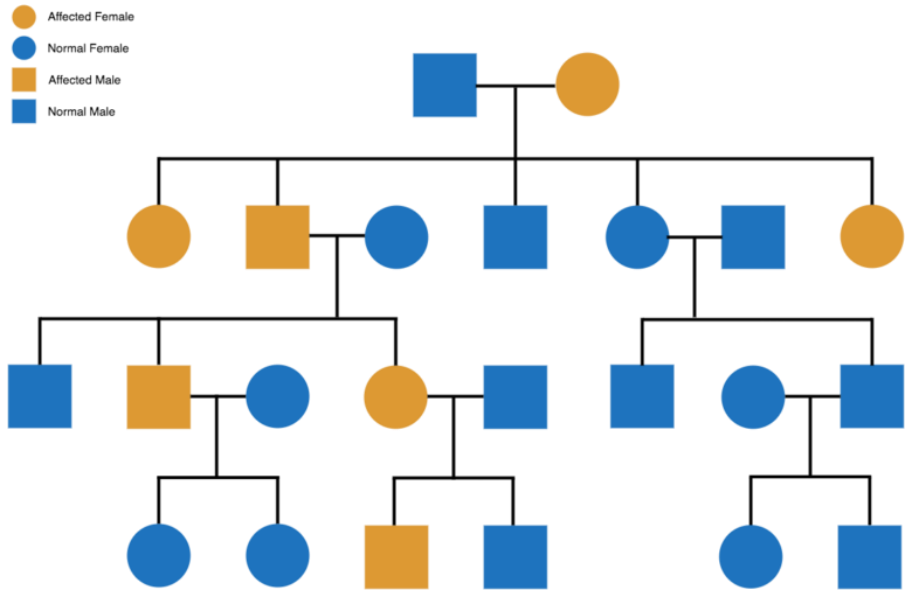
- Complex traits
  - Multifactorial: multiple genetic and environmental contributors
  - Cardiometabolic Diseases: *e.g.* Diabetes, heart disease, stroke
  - Anthropomorphic traits: *e.g.* height, BMI, weight, bone density, birth weight
  - Psychiatric traits: *e.g.* depression, autism, ADHD, schizophrenia

A brief history



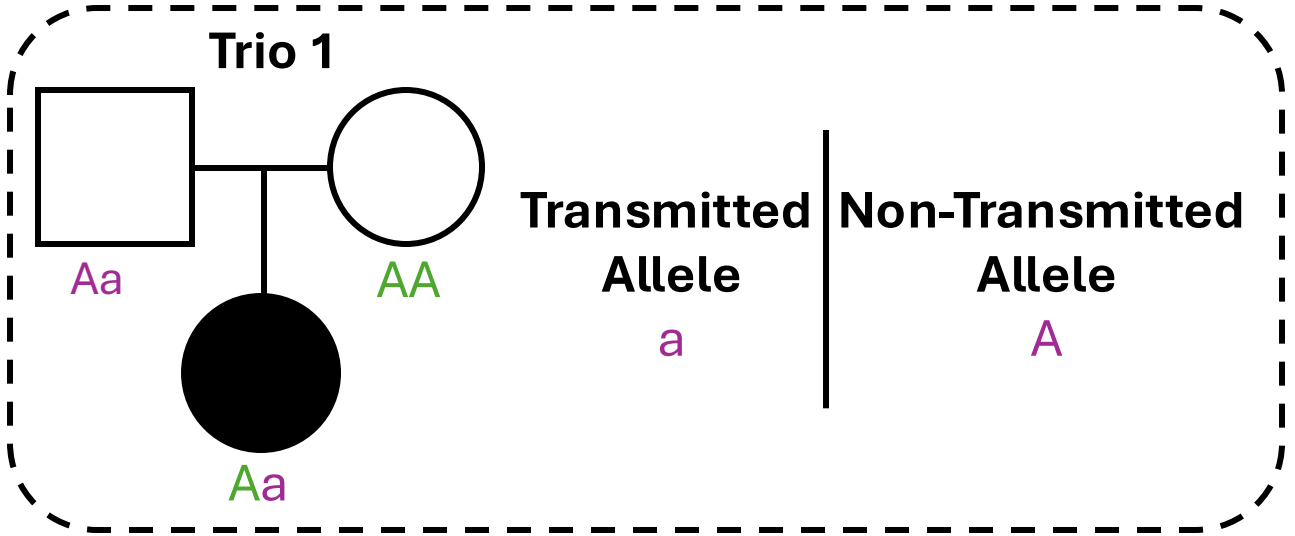
Genetic Linkage Studies

- Genetically related individuals to identify regions of the genome that may contain genes that contribute to trait variation
- Co-segregation of *genetic markers* with traits of interest in *pedigrees* containing affected relatives
- Usually microsatellite repeats – highly polymorphic regions
- Most successful for studies of rare diseases



Transmission Disequilibrium Tests

- Samples of parent-offspring trios where the offspring has a phenotype of interest
- Assess whether an allele is preferentially transmitted from heterozygous parents to affected offspring and therefore is associated with the phenotype



What is the null hypothesis?

Which SNP and allele is associated with the disease?

Repeat analysis across many trios + many SNPs

SNP	A1	A2	% Transmission of A1	% Non-transmission of A1
rs1	a	A	50	50
rs2	b	B	52	48
rs3	c	C	80	20

Per-SNP Statistical Test

	Non-Transmitted Allele		
Transmitted allele	**A**	**a**	**Total**
A	*a*	*b*	*a+b*
a	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*2n*

$$= \frac{(b-c)^2}{(b+c)}$$

Does this test look familiar?

Candidate Gene Studies

- Typically use a case—control study design
- Genotype cases and controls within the candidate gene
- Perform a chi-squared test or logistic regression
- Limited by current biological understanding of the trait of interest
- e.g. study a polymorphism within a genomic region previously implicated in a linkage study
- e.g. study a polymorphism previously associated with a closely related phenotype

Which candidate genes would you choose to study if you are interested in a psychiatry disorder?

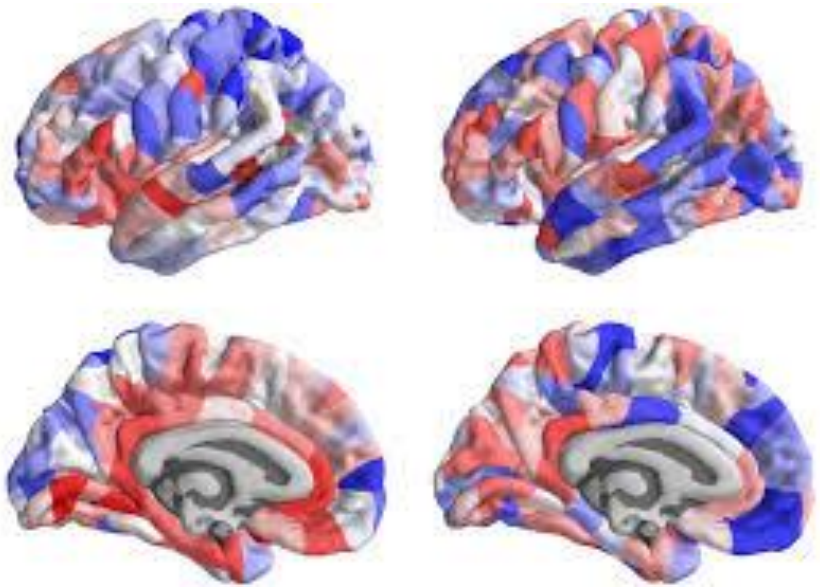


Image credit: https://www.thetransmitter.org/spectrum/gene-expression-maps-reveal-origins-brain-changes-from-autism-mutations/

Genome-Wide Association Studies

- Test for association between genetic variants and a phenotype
- Requires large samples of (usually unrelated) individuals

- Enabled by:
  - Human Genome Project (2003)
  - International HapMap Project (2005)
  - Cost-effective microarray (SNP-chip) technology

- Example of a microarray with red/green fluorescent dye
  - Green = AA
  - Red = TT
  - Yellow/Orange = AT
  - Brightness = amount of DNA that hybridizes to the chip

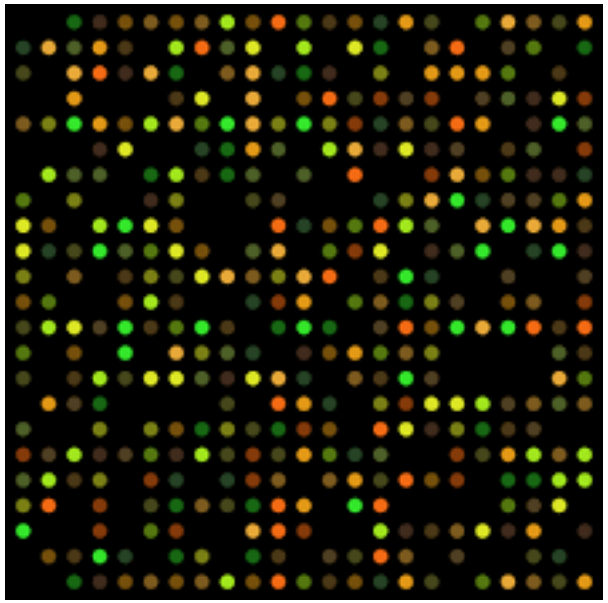
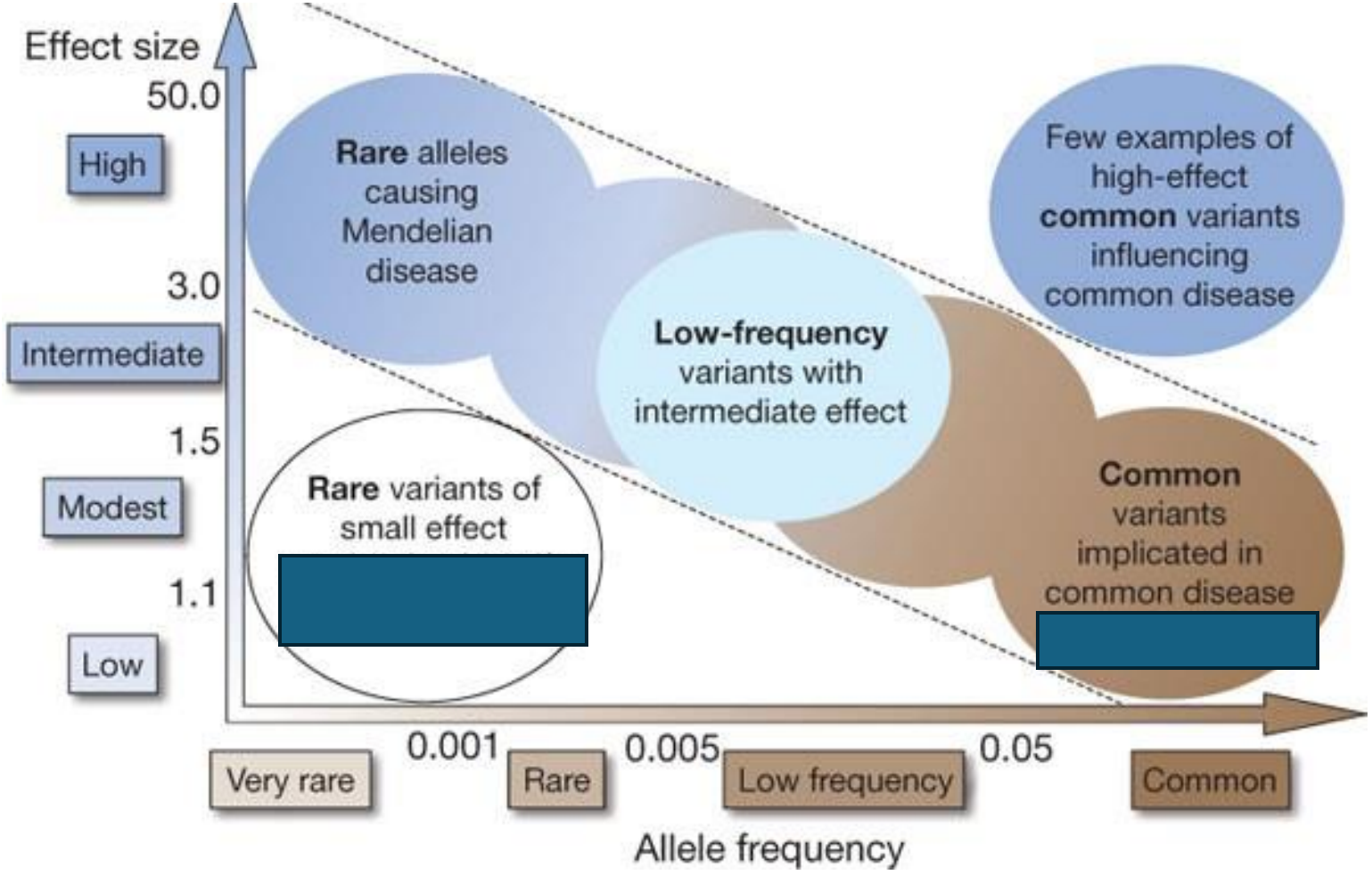


Image credit:
http://biotools.eu/en/52-microarrays

- *Which type of study would be best for detecting variants associated with rare monogenic diseases? What about for common diseases?*
- Common disease common variant hypothesis

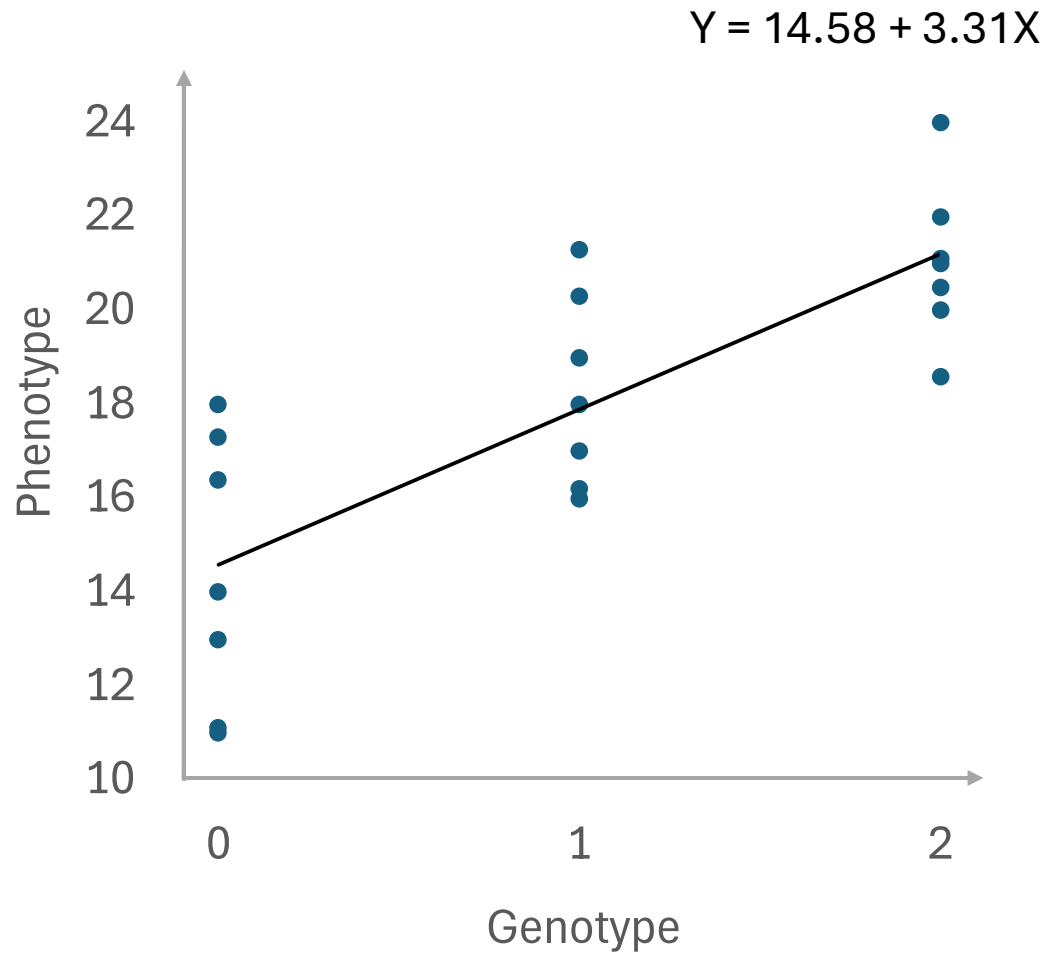


Linear regression (continuous traits)

- $Y = \alpha + \beta X + \varepsilon$
- Y = phenotype
- X = genotype (0, 1 or 2 copies of the effect allele)
- β = effect of the allele

What is the null hypothesis?

Each individual SNP is a predictor – how should multiple testing be handled?



Multiple Testing Correction

- *If we test 1000 markers that are not associated with the disease, how many of the tests do we expect to have a p-value less than 0.05?*
- Bonferroni correction:
  - Only consider p-values below 0.05/N significant
  - N = number of tests
- *If we test 10 million SNPs in a European population, should we then do Bonferroni correction for 10 million tests?*
- *Should the significance criterion be more or less stringent in an African population?*

Standard genome-wide significance threshold needed for publication is $5x10^{-8}$

Slide credit: Søren

Logistic regression (dichotomous traits)

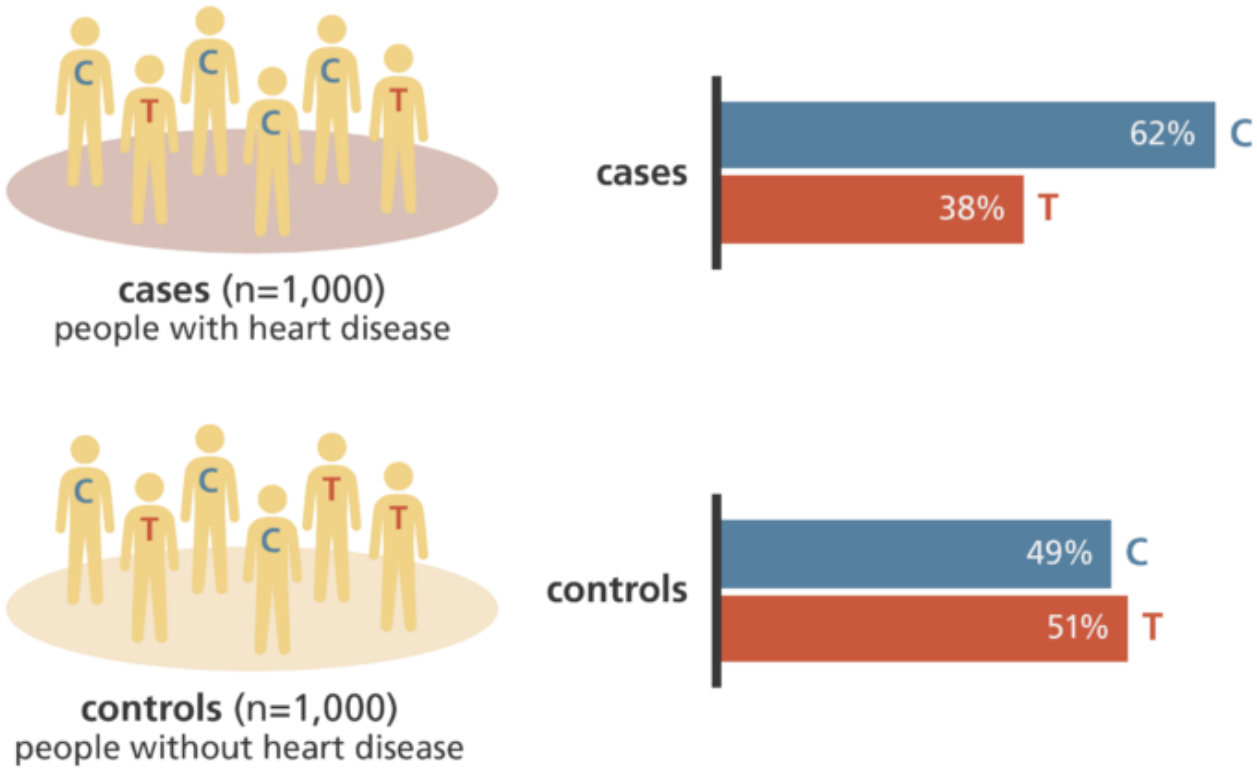


Image credit: Laura Olivares Boldú / Wellcome Connecting Science

Which allele is associated with heart disease?

- % correspond to the minor allele frequency in each group
- A variant is associated with the disease if the variant has different frequencies in a case group compared to a control group

- $\text{Ln}(P/1-P) = \alpha + \beta X + \varepsilon$
- OR > 1 = allele is associated with increased disease risk
- OR < 1 = allele is associated with decreased disease risk

Genotype Imputation

- Imputation can increase the number of tested SNPs from ~ 500K to several million
- Requires a densely genotyped reference panel
  - e.g. the International HapMap Project (International HapMap Consortium, 2003, 2005, 2007).
- Relies upon patterns of linkage disequilibrium (LD)
  - Genetic variants close together on a chromosome tend to be inherited together

- Step 1 Phasing: statistically estimate haplotypes from genotype data
  - A haplotype is the set of DNA variants inherited from a single parent (i.e. humans have 2 haplotypes)
- Step 2 Imputation: impute missing genotypes

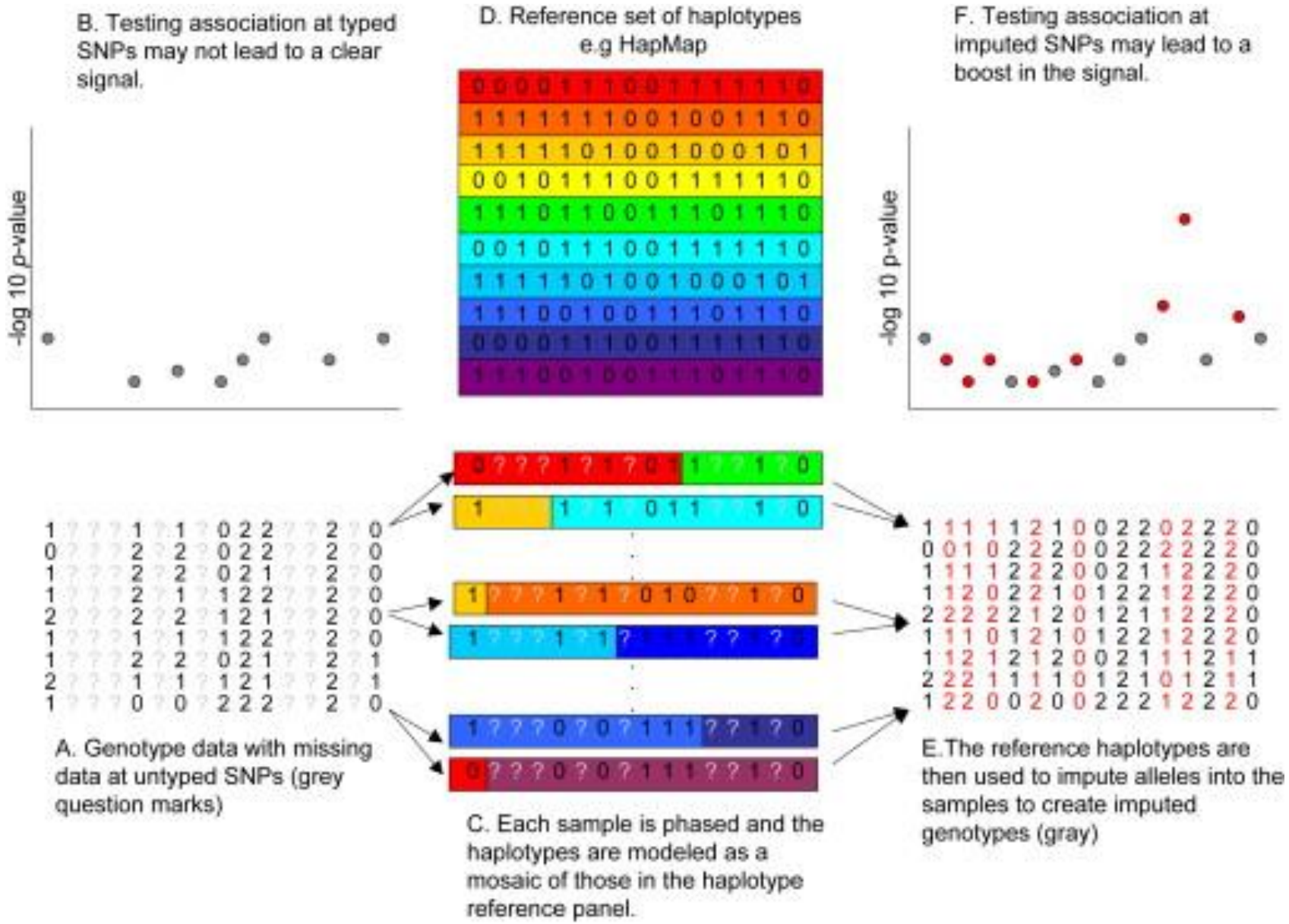
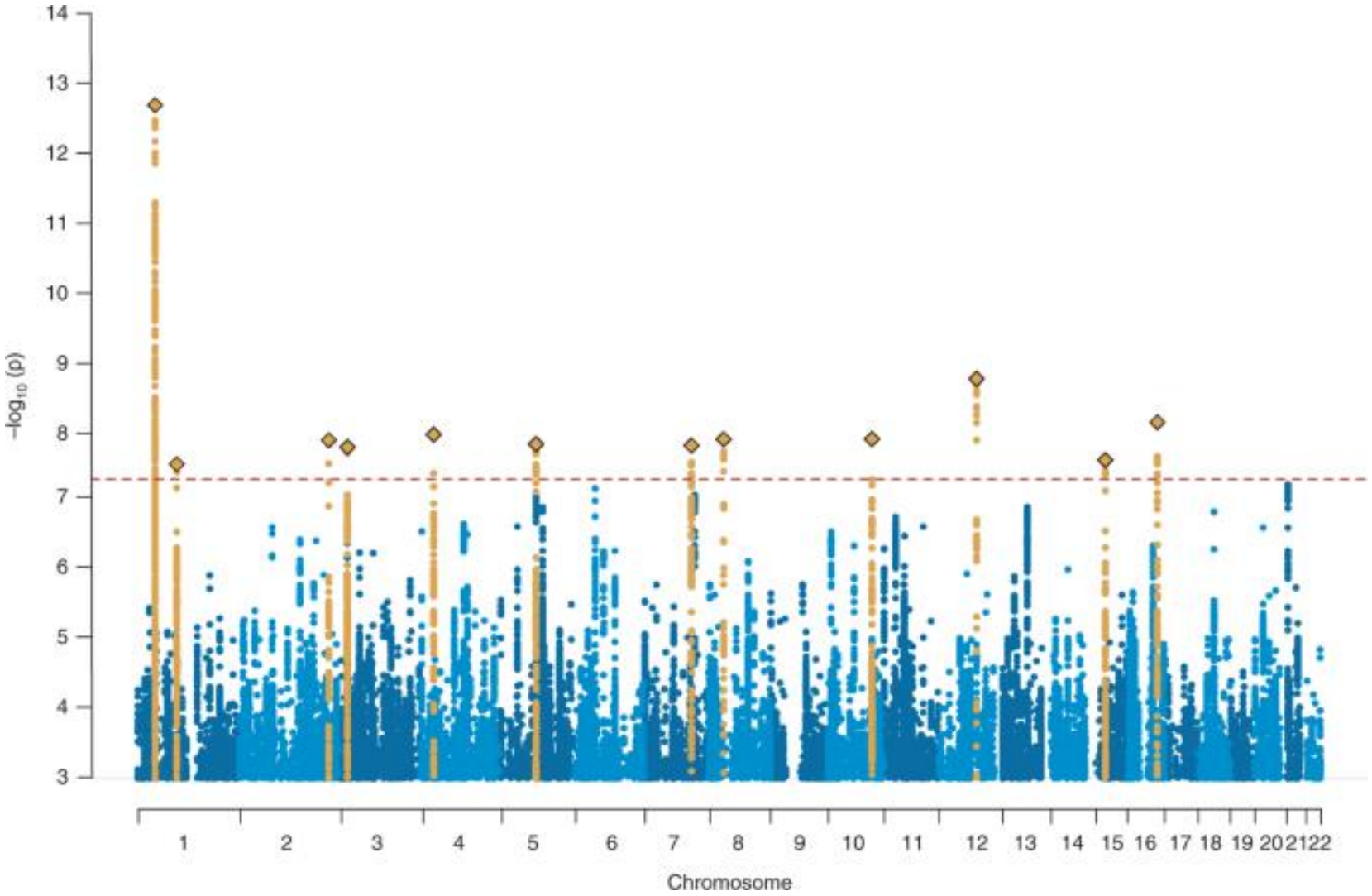


Image Credit: Marchini 2011https://doi.org/10.1016/B978-0-12-375142-3.10010-0

Manhattan plots

- First genome-wide significant risk loci for ADHD discovered in 2019
- 20,283 individuals with ADHD
- 35,1919 controls



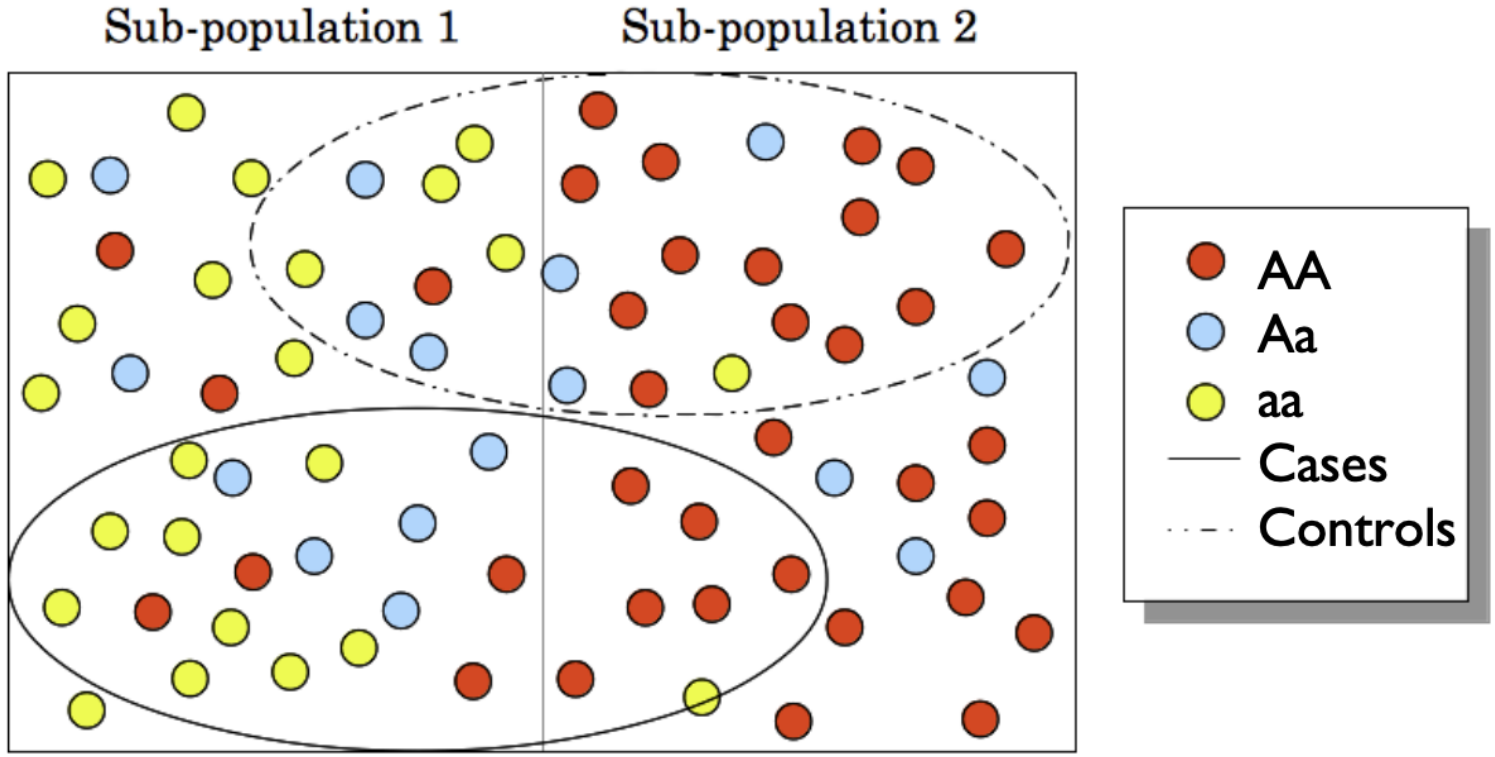
P = $5x10^{-8}$

Why do we see towers in the Manhattan plot?

Image credit: Demontis et al., 2019 https://doi.org/10.1038/s41588-018-0269-7

Population Stratification

- Presence of multiple subpopulations in the study that differ in:
  - allele frequencies
  - disease prevalence (or phenotypic trait mean)



Population Stratification

- Well-known Chopstick example (Uffelmann et al., 2021)
  - Conduct a GWAS for chopstick use
  - Cases = ‘using chopsticks regularly’
  - Controls = ‘never using chopsticks’
  - Cases are more likely to be of East Asian descent than controls
  - Any variant more common in east Asians than other populations will appear significant

Ancestry

- PCA of genotypes
- In a study with multiple ancestry groups, separation of PC1 and PC2 usually is due to Ancestry
- Include genetic PCs as covariates in the analysis

- Multi-ancestry GWAS methods have been developed (Mixed models)

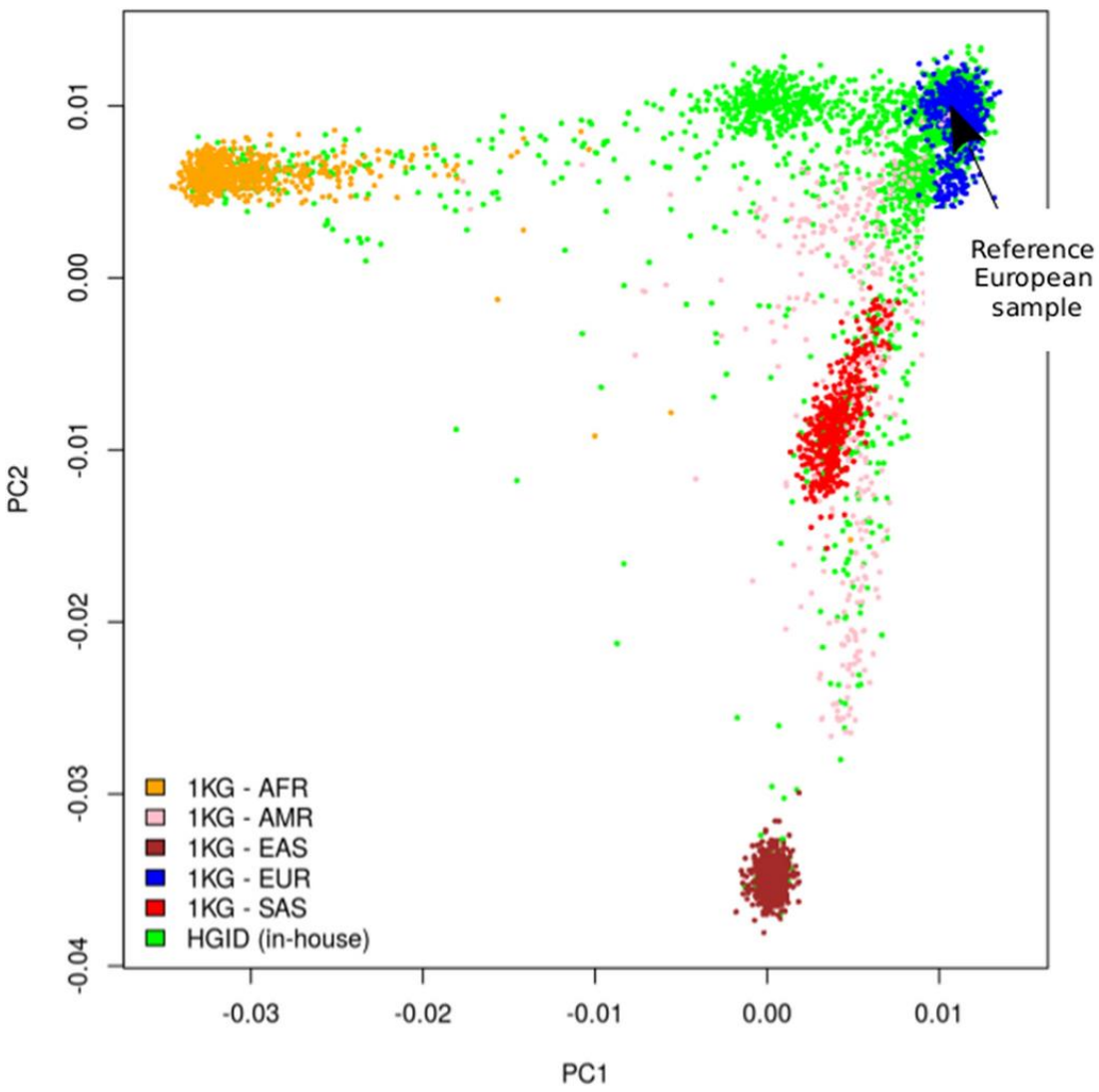
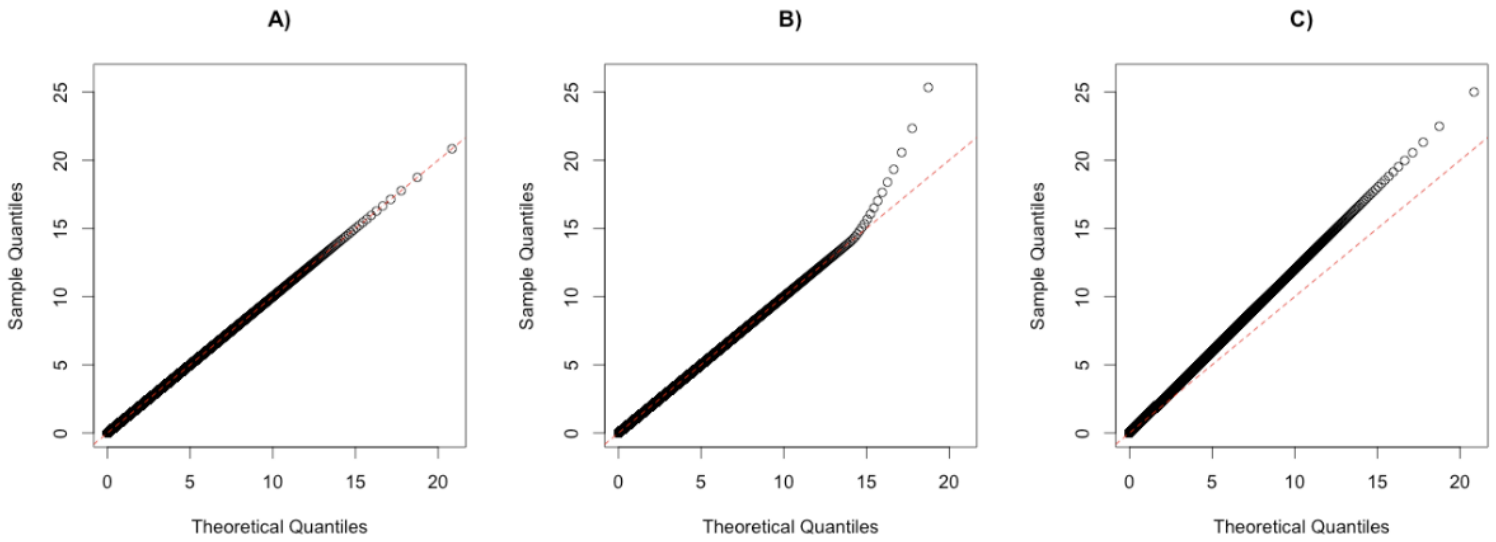


Image credit: Bouaziz et al., 2012; https://doi.org/10.1038/s41598-021-98370-5

Quantile-Quantile Plot (QQ-Plot)



A) The observed values follow the expected distribution

B) Some observed variants higher than expected.

C) All observed variants higher than expected: sign of a problem.

Image Credit; Soren

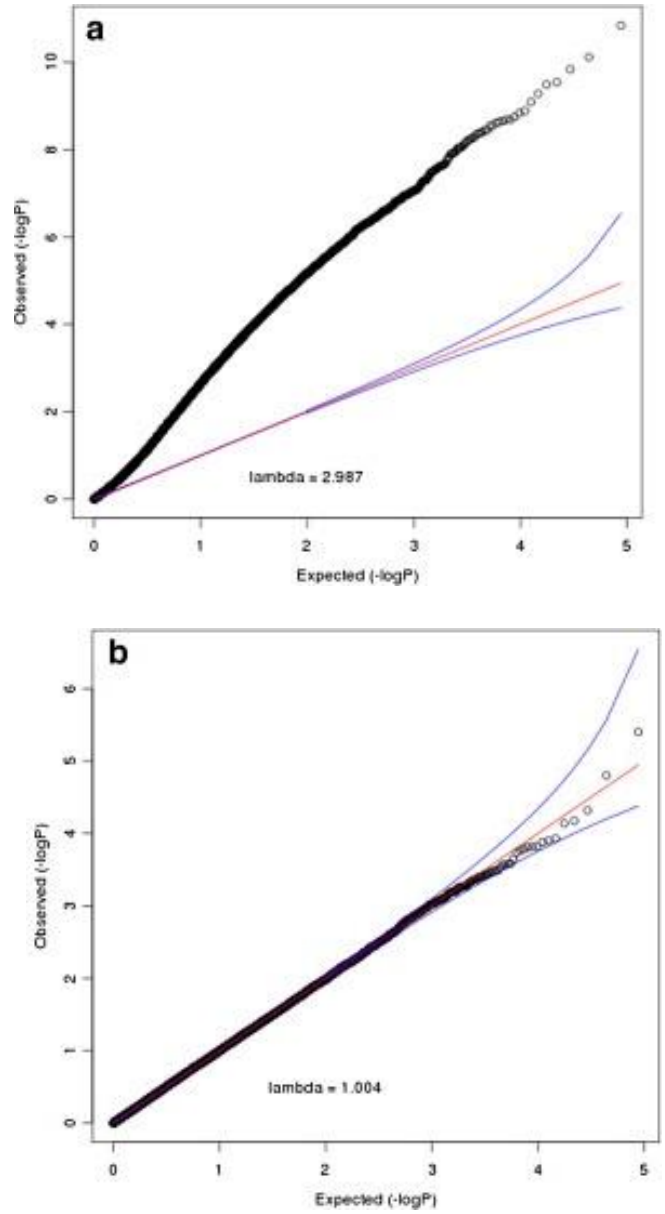
Genomic Inflation Factor (λ)

- λ is a measure of test statistic inflation
- λ = median(χ^2)/0.456
- λ = 1 (ideal)
- λ > 1 (population stratification or other confounding)
- λ < 1 (underpowered study)

- Genomic control: adjust every test statistic by dividing by the genomic inflation factor
- χ^2 adjusted = χ^2 / λ

Relatedness

- Linear and logistic regression assumes that each sample is independent
- Related individuals have correlated:
  - genotypes (due to common ancestry)
  - phenotypes (due to shared genetics and shared environment e.g. diet + lifestyle)
- Related individuals can be identified and 1 from each pair can be removed
- More advanced GWAS approaches can model relatedness



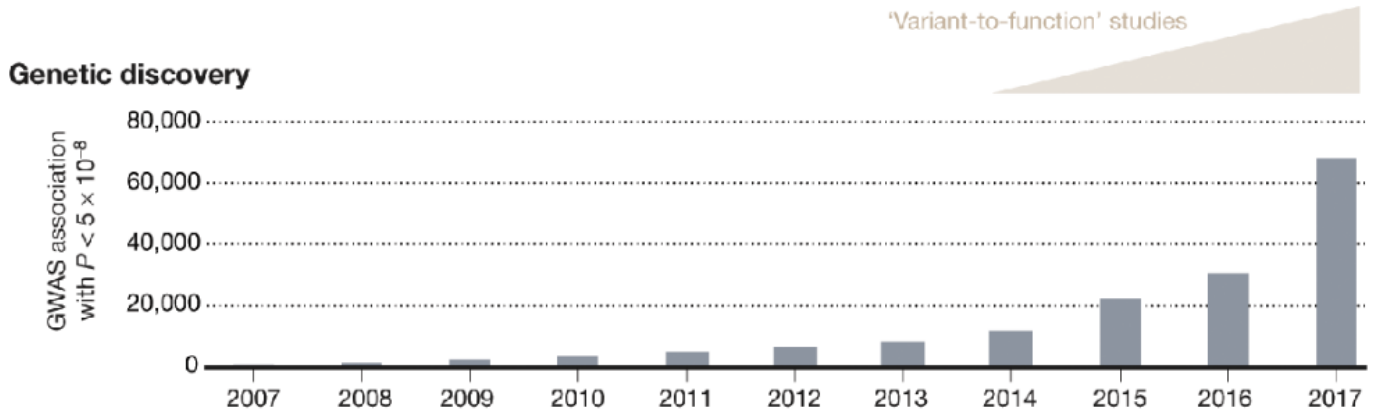
Chen et al., 2013;
DOI: 10.1186/1471-2105-14-74

Other Quality Control

- Genotyping errors are common:
  - Exclude markers with high missingness
  - Exclude markers with low minor allele frequency
  - Exclude markers that show large deviations from Hardy-Weinberg Equilibrium
- Some individuals have poor DNA quality
  - Exclude individuals with high missingness
- Sample mix-ups occur
  - Test that the sex information (based upon X chromosome genotypes) matches the reported sex of each sample

Summary of GWAS

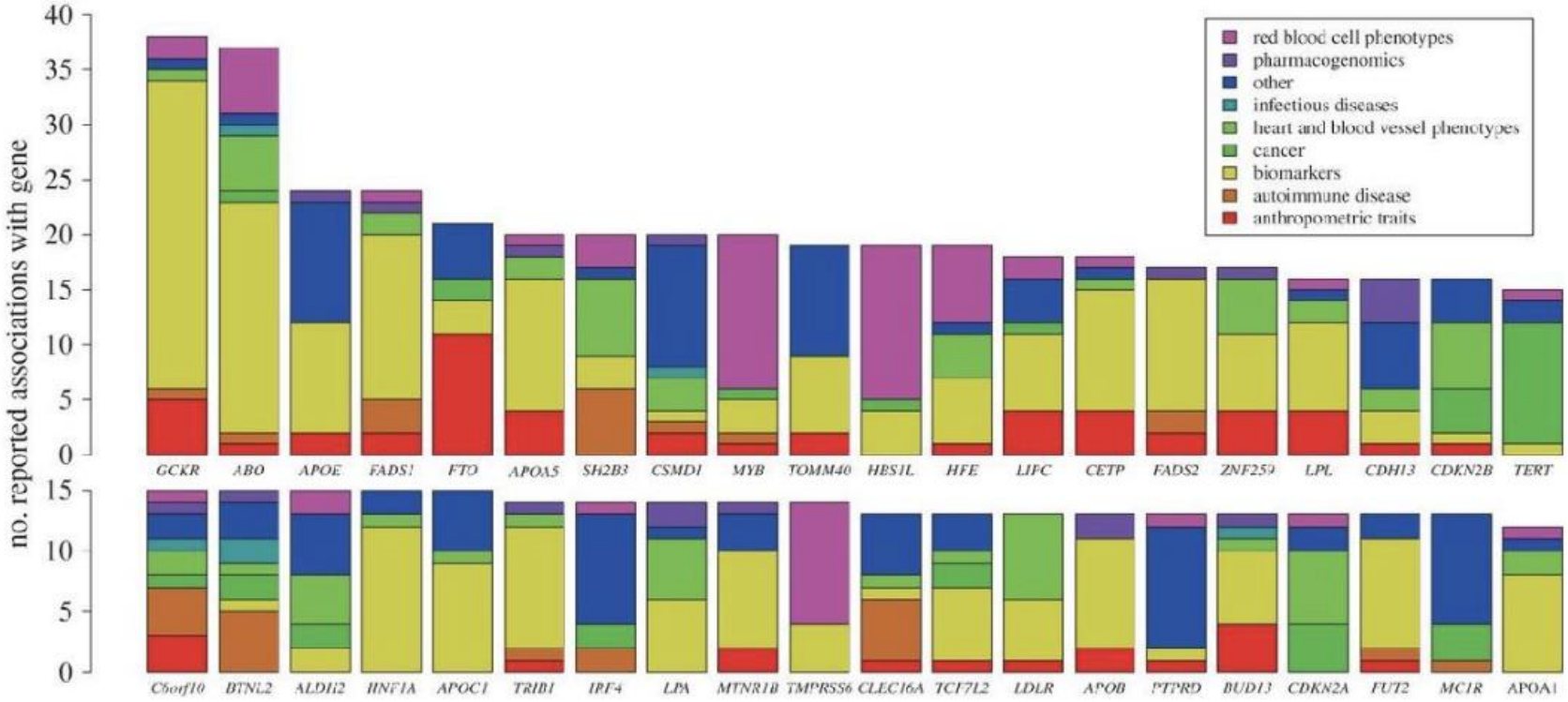
- Select study participants
- Quality Control
- Association testing
- Produce plots + check for abnormalities
- (Probably) redo QC
- Replicate findings using a different cohort



https://www.ebi.ac.uk/gwas/

Pleiotropy is widespread

- One variant affecting more than one trait



Large Genotyped cohorts

- 23 and Me (~ 10 Million individuals; US mostly)
- UK Biobank (~ 500K; UK)
- Million Veterans Program (~ 500K; various ancestries)
- ALL of Us (~ 250K individuals; US mostly)
- Norwegian Mother, Father and Child Cohort Study (~ 238K individuals; Norway)
- iPSYCH (~ 140K individuals; Danish)
- ALSPAC (~ 20K individuals, UK)
- FinnGen + Japan Biobank + Taiwan Biobank + many many others!

Discuss amongst yourselves some interesting GWAS applications
e.g. how might GWAS be used to identify pharmacological targets?

Have a look at some of the phenotypes available in the UK Biobank data catalogue:
https://biobank.ndph.ox.ac.uk/ukb/search.cgi?wot=0&srch=diet&yfirst=2000&ylast=2025

Questions?

- Next week: Heritability