GWAS Overview and Single SNP Tests using R

Adapted from Joanne Cole, PhD – GENC6120 Lecture

R Ladies Aurora

January 14, 2026

Presented by Kristen Sutton, PhD

Department of Biomedical Informatics
SCHOOL OF MEDICINE
UNIVERSITY OF COLORADO **ANSCHUTZ MEDICAL CAMPUS**

Learning Objectives

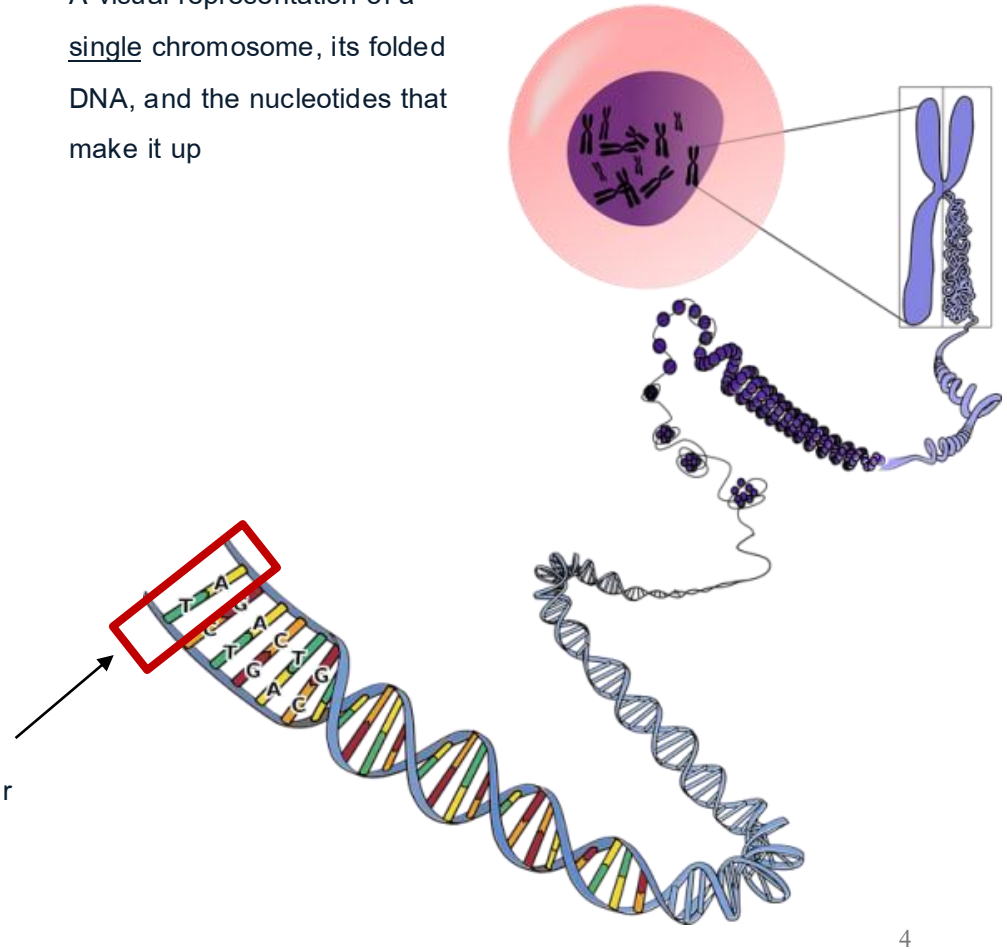
- Distinguish rare vs common variants in the genome
- Identify the major differences between rare and common disease
- Compare and contrast the different approaches to identify regions in the genome involved in disease
- Describe how a GWAS is conducted and interpret the results (P-values, Betas/ORs, Manhattan Plots)

Review: Basic Concepts

Genome Structure

- Humans have 22 pairs of autosomes + 1 pair of sex chromosomes
- Together, the chromosomes contain **>3 billion base-pairs**

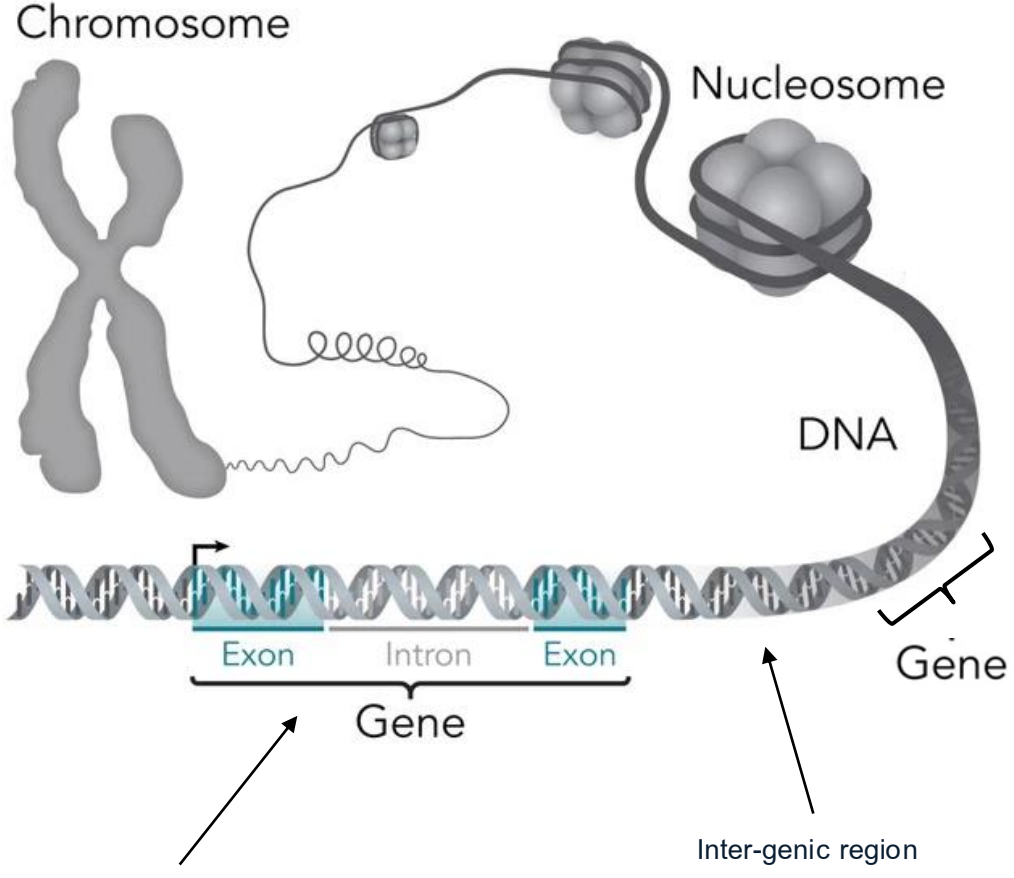
A visual representation of a single chromosome, its folded DNA, and the nucleotides that make it up



A single base-pair

Genome Structure

- Genes are segments of DNA that encode proteins
- Humans have ~20,000 protein-coding genes
- 40% of the genome is genic (introns + exons)
- **1% of the genome is protein-coding (exons)**



In reality, introns are typically much larger than exons

>99% of the genome is identical between any two people ~1% contributes to human diversity

- Mutation – a *rare* change to the DNA sequence that severely disrupts function and typically leads to disease
  - Occurs in <<1% of the population
  - *Technically mutation just means change to the DNA sequence, but this is more often how we use the word in human genetics*
- **Polymorphism – a *common* change to the DNA sequence in which there are two or more different alleles**
  - **The minor allele occurs in >1% of the population**
  - **Polymorphisms occur roughly every 300-1000bp (~10-15 million common polymorphisms in the genome)**
- Genetic variant/marker – a polymorphic DNA segment of known location

Types of genetic variation and mutations

- Single Nucleotide Variant (SNV)
  - Variation of a single DNA base pair (bp)
  - Single Nucleotide Polymorphism (SNP): an SNV which occurs >1% in the population
- Insertion/Deletion (InDel)
  - Less frequent than SNVs
  - Insertion or deletion of DNA bps (<50 bps)
  - Short tandem repeats (STR) are repeats of short sequences of DNA and are a type of InDel
- Structural Variation
  - Least frequent
  - Includes deletions, duplications, insertions, inversions, and translocations
  - Copy number variations (CNVs) are a subset of structural variations that lead to a change in copy number (loss or gain) of a DNA fragment >50 bps (typically >1kb)
  - Aneuploidies affect entire chromosomes (e.g., Trisomy 21)

SNPs and InDels are the two major types of common genetic polymorphisms

SNPs

- Occur every ~100-300 base-pairs
  - > 10 million SNPs in the human genome

rsID = a unique and stable identifier for every SNP. It links the alleles/genotypes to a chromosomal location (which is updated with each genome build)

An example SNP rs56372821

- Possible alleles
  - G is the major allele
  - A is the minor allele
- Possible genotypes
  - GG (homozygous major allele)
  - GA (heterozygous)
  - AA (homozygous minor allele)

SNP rs56372821
chr8:27578983 (build GRCh38)

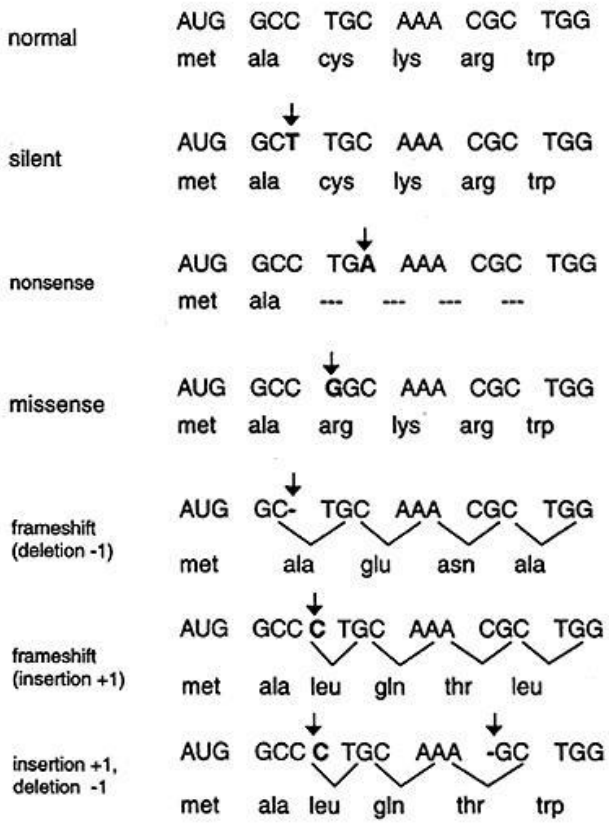
Allele	Allele Frequency*
G	84%
A	16%

* Individuals of European ancestry only
Frequencies may vary in other populations

Impact of genetic variation

- Genetic variation can impact a phenotype in many ways
- Variants that directly affect the coding sequence of a protein
  - Often have the highest impact
  - Easiest to predict their impact
- **Variants outside of the coding region often have a milder impact**
  - **Much more common than changes to the coding sequence**
    - **>80% of the genome is non-coding**
    - **Milder effects are less likely to be severely deleterious**
  - **Can affect the expression, stability, or function of a gene product (typically a protein)**

Variants that affect the coding sequence



+ splice variants, CNVs, and SV

https://step1.medbullets.com/biochemistry/102010/type-of-dna-mutations

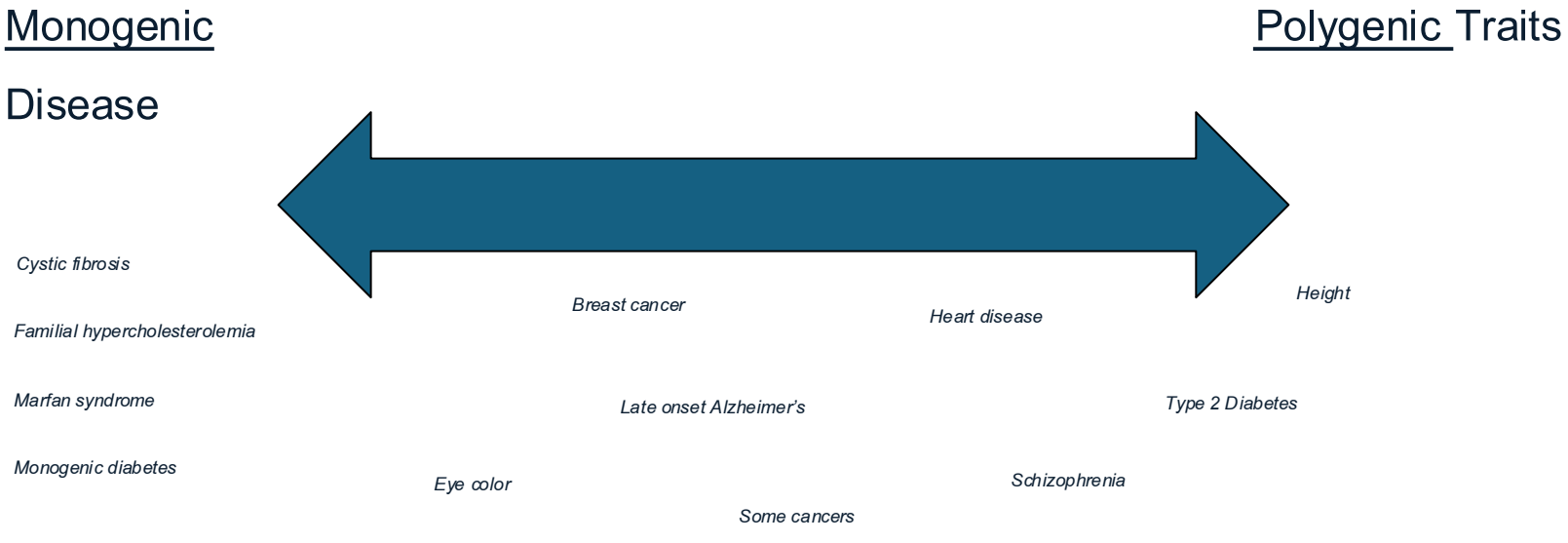
Monogenic disease vs. Common disease

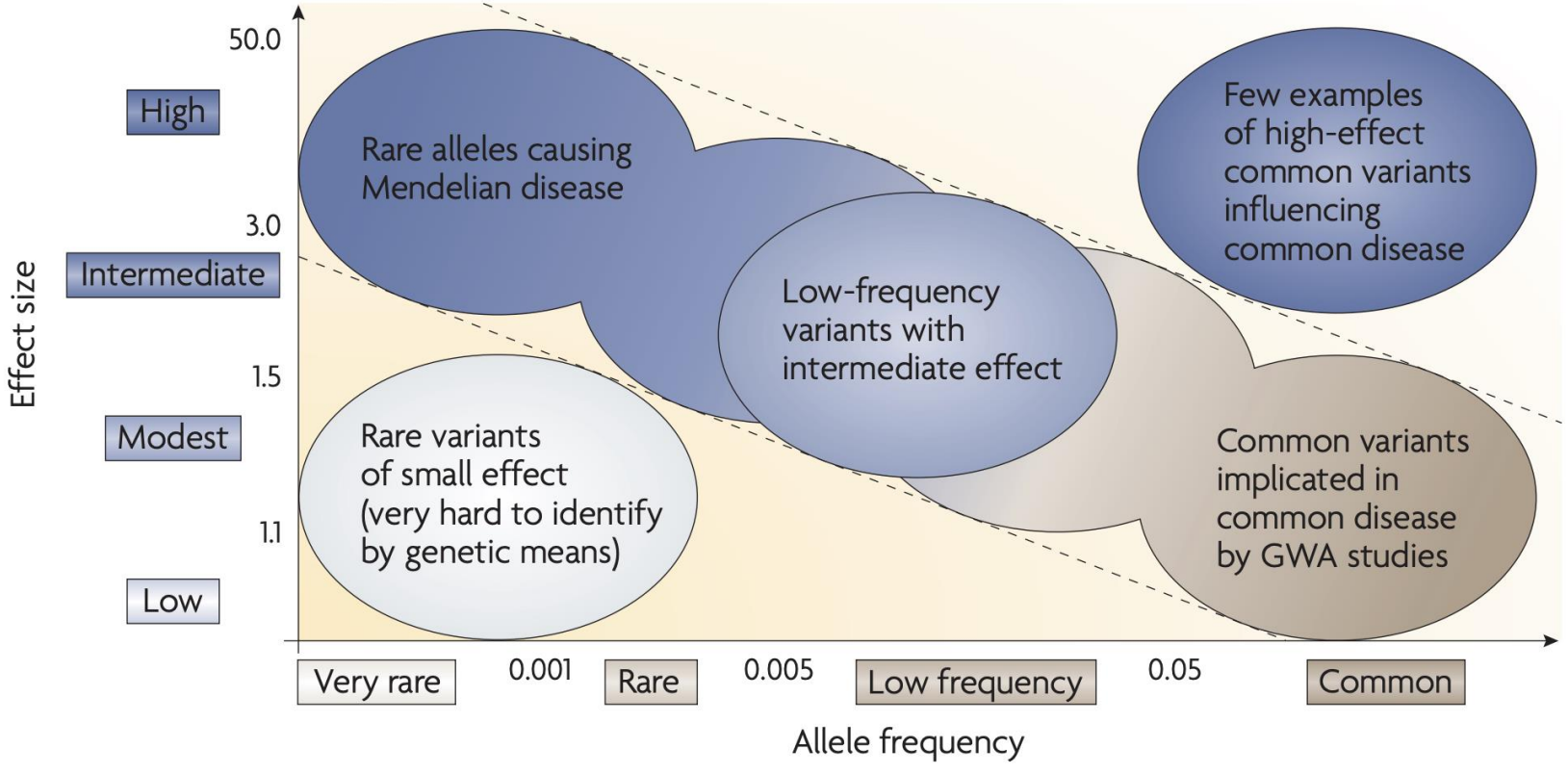
Monogenic disease = Mendelian disease

- Definition: A “single gene/genetic variant” causes disease
- Reality: Factors outside of the causative gene/genetic variant can contribute to variable disease presentation
  - Huntington’s disease: variable disease presentation (e.g., unaffected, mild, severe) based on the number of CAG repeats in the *HTT* gene
  - Duchenne and Becker muscular dystrophies: severity depends on which genetic variant in the dystrophin gene *(DMD)* you inherit

Common disease and Complex traits

- Diseases and traits which are caused by a combination of multiple genetic and environmental factors (and their interactions)





Atonarakis *et al.*, *Nature Reviews Genetics* 2010

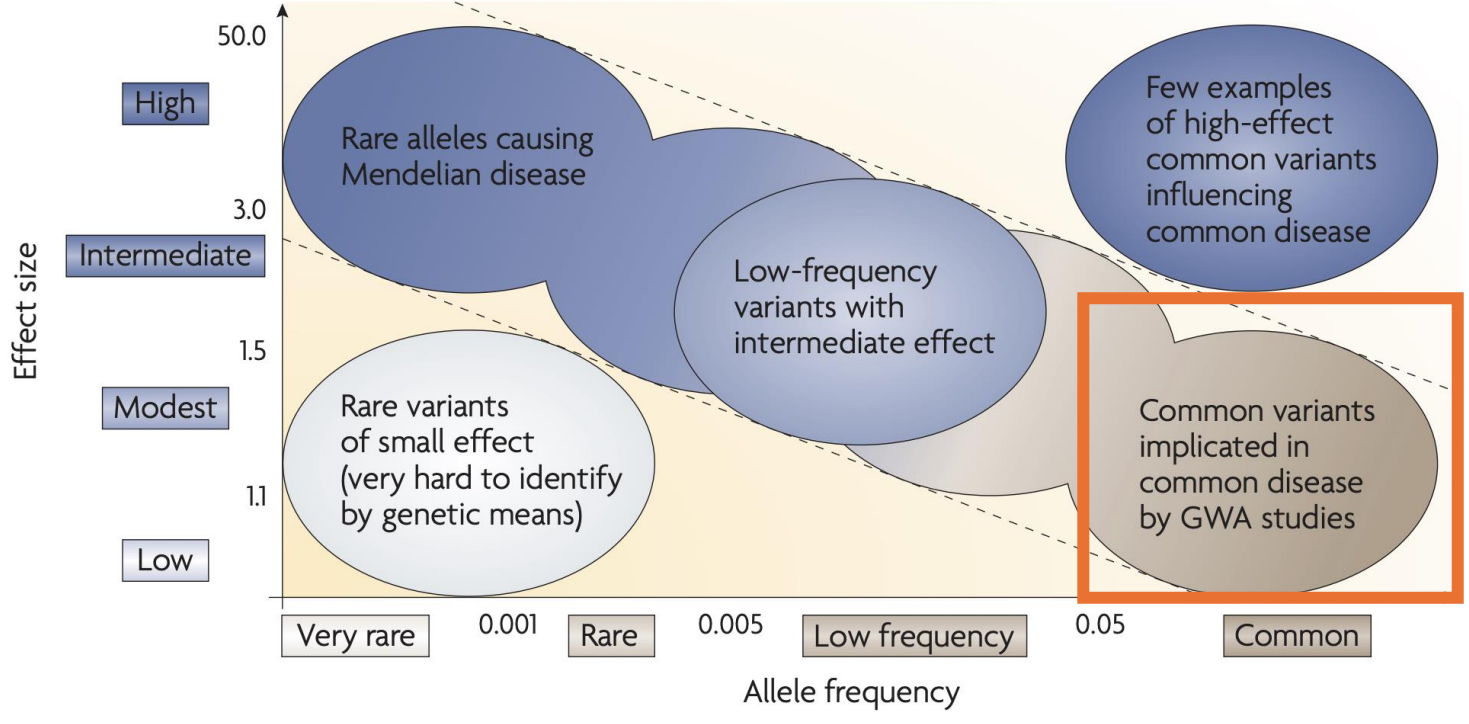
How to identify disease loci?

Locus (plural loci): a physical location in the genome. Typically used to refer to a region of the genome associated with your trait of interest.

Association Studies

- Test for the statistical relationship between genotype and phenotype
- Candidate gene association studies
- Genome-wide association studies (GWAS)
- Next generation sequencing association studies

Polygenic traits: Identify the genetic variants



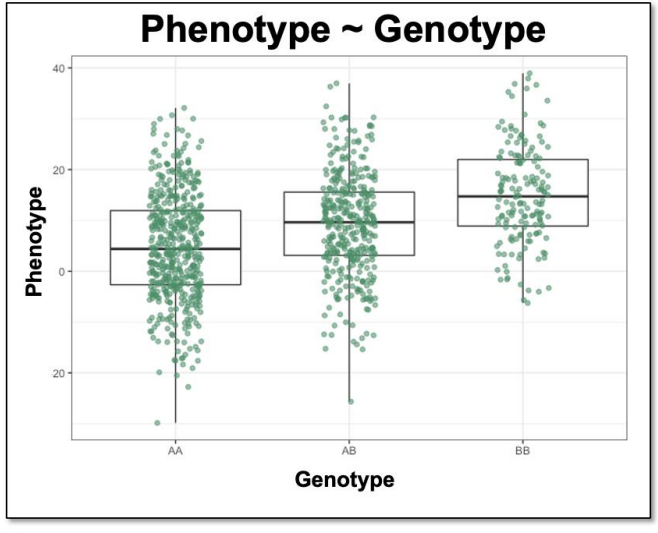
Atonarakis *et al.*, *Nature Reviews Genetics* 2010

The genetic underpinnings of common complex traits (i.e. polygenic traits)

- Multiple variants with small effect sizes
- Environment plays a role in etiology
- Many adult-onset phenotypes
- E.g., diabetes, asthma, height, blood pressure, Crohn's disease, schizophrenia, age-related macular degeneration, etc.

Association studies

- Test for the statistical association between genotype and phenotype
- Typically, under an additive genetic model, where each additional minor allele changes the phenotype by the same amount (See Plot)
- Though less frequently done, you can also test for an association under dominant, recessive, or genotype models



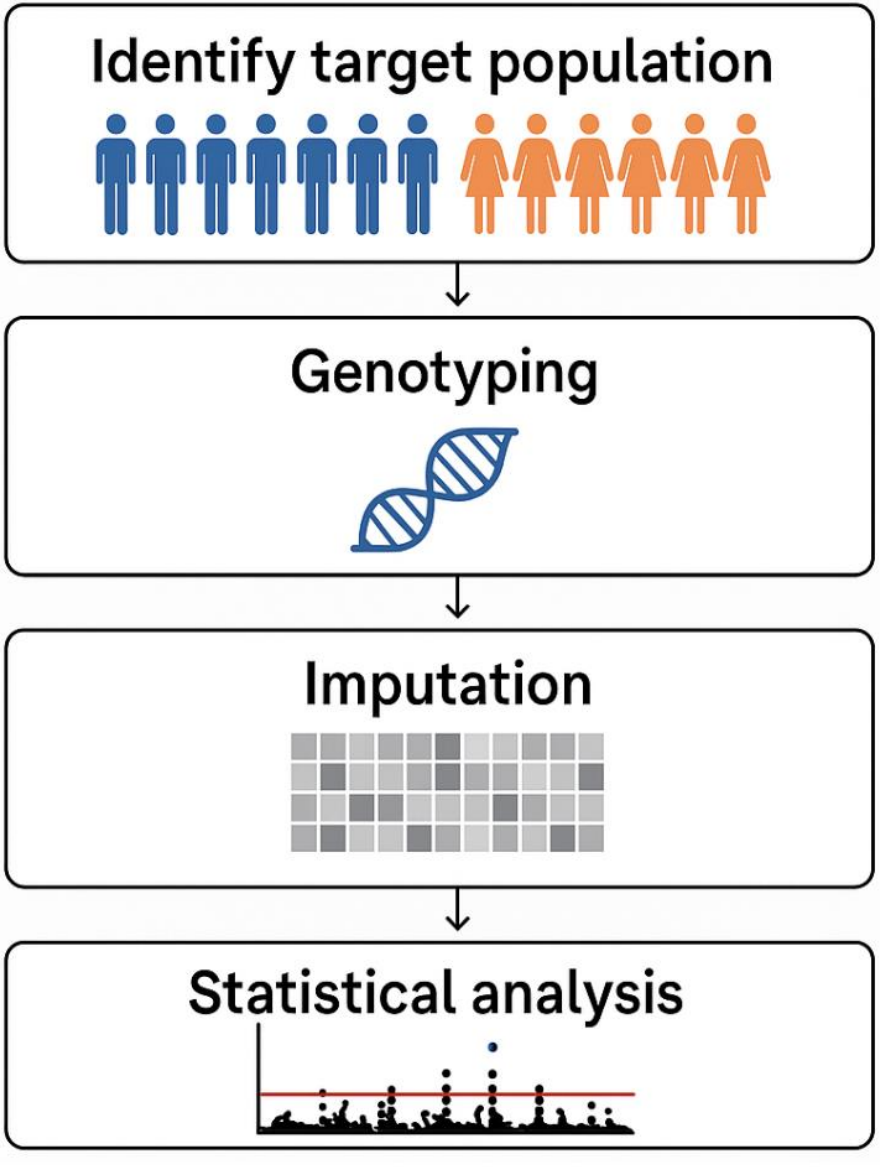
Quantitative outcomes are prettier to plot, but the same general approach (with slightly different statistical regressions) applies for case-control GWAS

Association Studies

- Candidate gene studies test a limited number of variants/genes based on a priori information (e.g., biological hypotheses or focus on a linkage peak)
  - Very high false positive rate

Association Studies

- Genome-wide association studies (GWAS) test millions of common genetic variants across the genome for a statistical relationship with your phenotype of interest
  - i.e. Is there a significant difference in genotypes between cases and controls? across quantitative outcomes?



Consider: homogeneity, trait prevalence, inclusion and equity, genetic ancestry, sample size

~1 million variants (e.g., SNPs and indels) are directly measured on a microarray with oligonucleotide probes

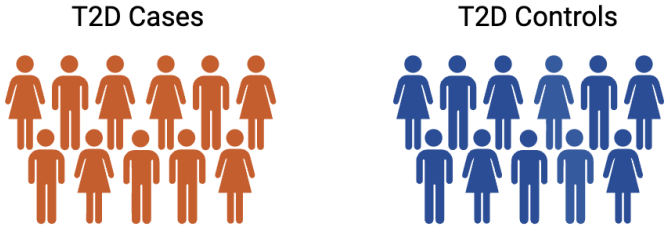
Then millions of additional variants are inferred through imputation using large whole-genome sequencing reference panels

Test for the statistical association between genotype and phenotype, one by one, at all variants

Target population

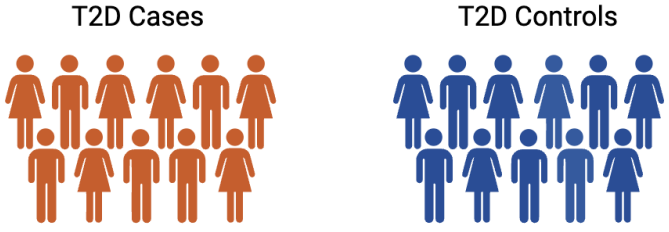
- Historically, conduct GWAS in a single homogeneous genetic ancestry
  - Most GWAS are in individuals of European ancestry
  - Avoids “population stratification” – ancestral differences in allele frequencies that lead to spurious associations with the phenotype

Population Stratification



GWAS: Find SNPs with different allele frequencies between groups

Population Stratification



GWAS: Find SNPs with different allele frequencies between groups

- South Asian
- European ancestry

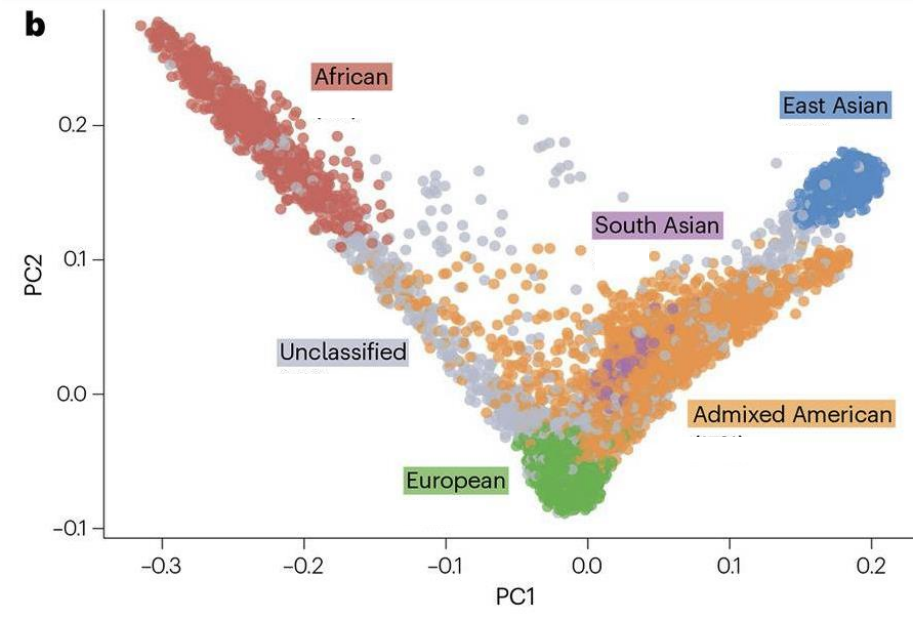
If cases and controls differ by other characteristics, your GWAS will find genetic associations with those other characteristics (false positive associations)

How to address population stratification

- Minimize population heterogeneity in your sample
- Adjust for genetic ancestry as covariates
- Fancier statistical models: mixed models with random effects for genetic ‘relatedness’

Genetic principal components analysis (PCA)

- Common approach to control for population stratification
- Genetic PCA identifies the largest axes of genetic variation across individuals
  - The biggest difference between all people is genetic ancestry

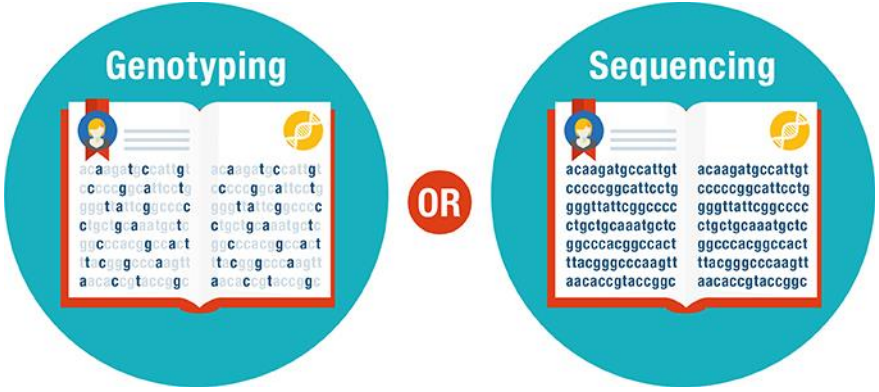


Kachuri et al 2024

Target population

- Historically, conduct GWAS in a single homogeneous genetic ancestry
  - Most GWAS are in individuals of European ancestry
  - Avoids “population stratification” – ancestral differences in allele frequencies that lead to spurious associations with the phenotype
- Now, many studies combine individuals of multiple ancestries
  - Use advanced statistical approaches to correct for relatedness and ancestry
  - More inclusive, larger sample sizes
  - Not always statistically appropriate if the phenotype greatly differs across populations

Genotyping arrays vs sequencing technologies



- Targets specific variants using pre-designed probes/primers on an array
- Cheaper and fairly comprehensive capture of <u>common</u> variants
- Misses any alleles or variants not included

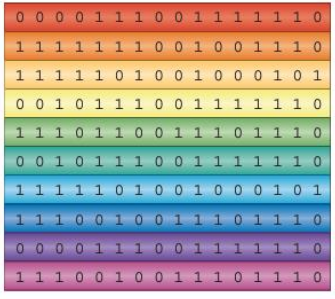
- Captures every base in a region, exome, or genome-wide using different technology
- Includes Sanger, Next Gen, and Third Gen
- Higher resolution, higher cost, more data, for better or for worse

Figure: www.aboutgeneticcounselors.com

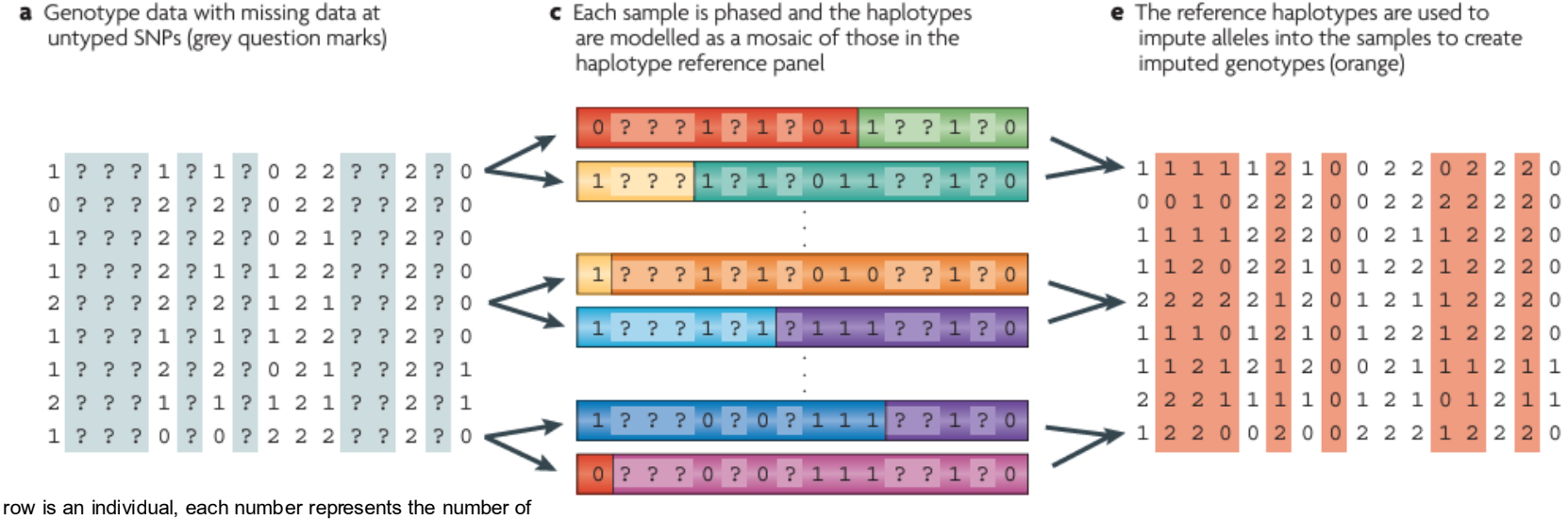
Imputation

1. We have genotyped data with missing information from genetic variants that were not on the genotyping array

Reference set of sequenced haplotypes



2. We match each row to a reference set of sequenced haplotypes to fill in the missing information



Each row is an individual, each number represents the number of alternate alleles present at that genetic marker (0/1/2 for the three possible genotypes)

Marchini and Howie 2010

nature reviews methods primers

I recommend this paper on GWAS

PRIMER

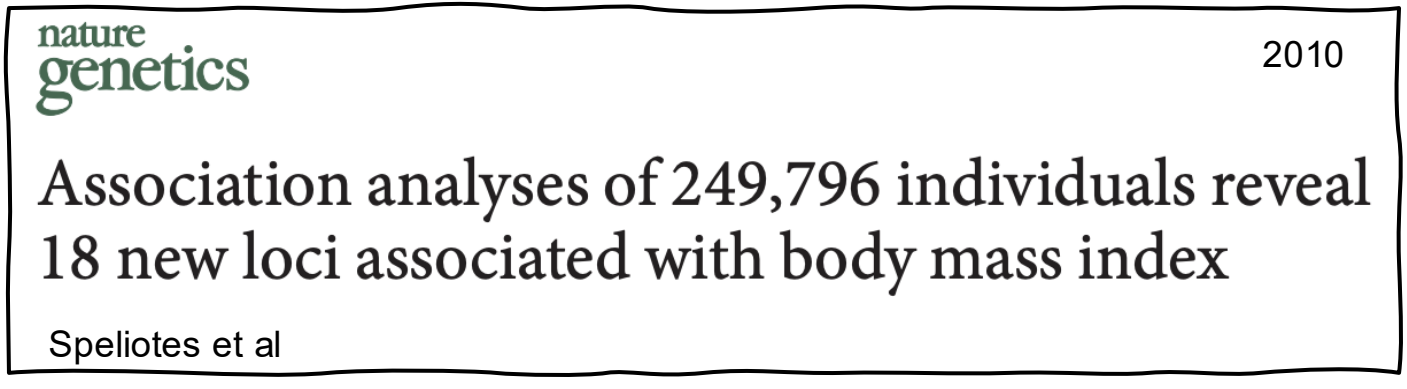
Check for updates

Genome-wide association studies

Emil Uffelmann[1], Qin Qin Huang[2], Nchangwi Syntia Munung[3], Jantina de Vries[3], Yukinori Okada[4,5], Alicia R. Martin[6,7,8], Hilary C. Martin[2], Tuuli Lappalainen[9,10,12] and Danielle Posthuma[1,11]

Abstract | Genome-wide association studies (GWAS) test hundreds of thousands of genetic variants across many genomes to find those statistically associated with a specific trait or disease. This methodology has generated a myriad of robust associations for a range of traits and diseases, and the number of associated variants is expected to grow steadily as GWAS sample sizes increase. GWAS results have a range of applications, such as gaining insight into a phenotype's underlying biology, estimating its heritability, calculating genetic correlations, making clinical risk predictions, informing drug development programmes and inferring potential causal relationships between risk factors and health outcomes. In this Primer, we provide the reader with an introduction to GWAS, explaining their statistical basis and how they are conducted, describe state-of-the art approaches and discuss limitations and challenges, concluding with an overview of the current and future applications for GWAS results.

Let’s do a GWAS!

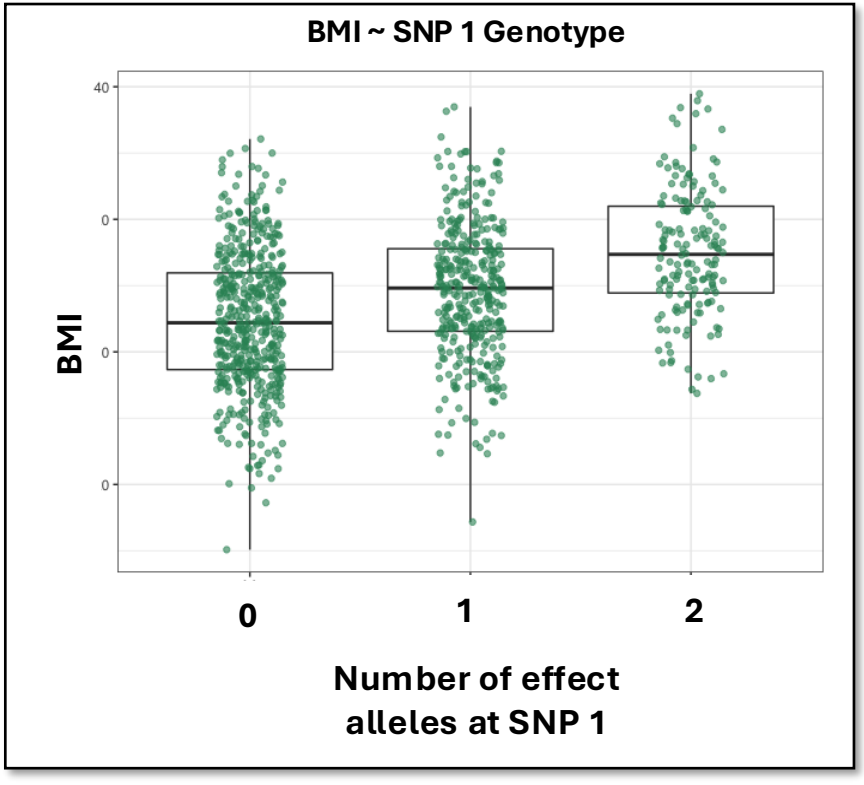


GWAS of Body Mass Index (BMI)

- Goal: identify genetic loci (regions in the genome) associated with BMI
- GWAS cannot identify causal genes
  - But it can point to loci that contain causal disease genes or regulatory elements that act on gene expression

GWAS linear regression model at a single SNP

1. For each individual, plot their:
   - Genotype in terms of number of effect alleles (x-axis)
   - BMI (y-axis)



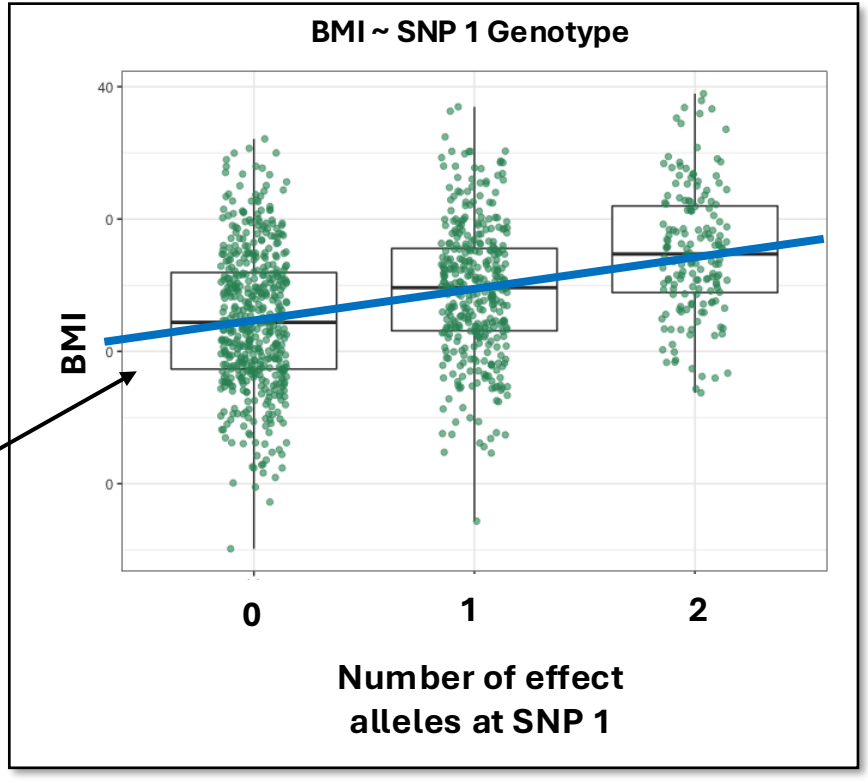
GWAS linear regression model at a single SNP

$$outcome = \beta * SNP_1 + intercept + covariates + error$$

BMI

Genotype at SNP1 in terms of number of effect alleles

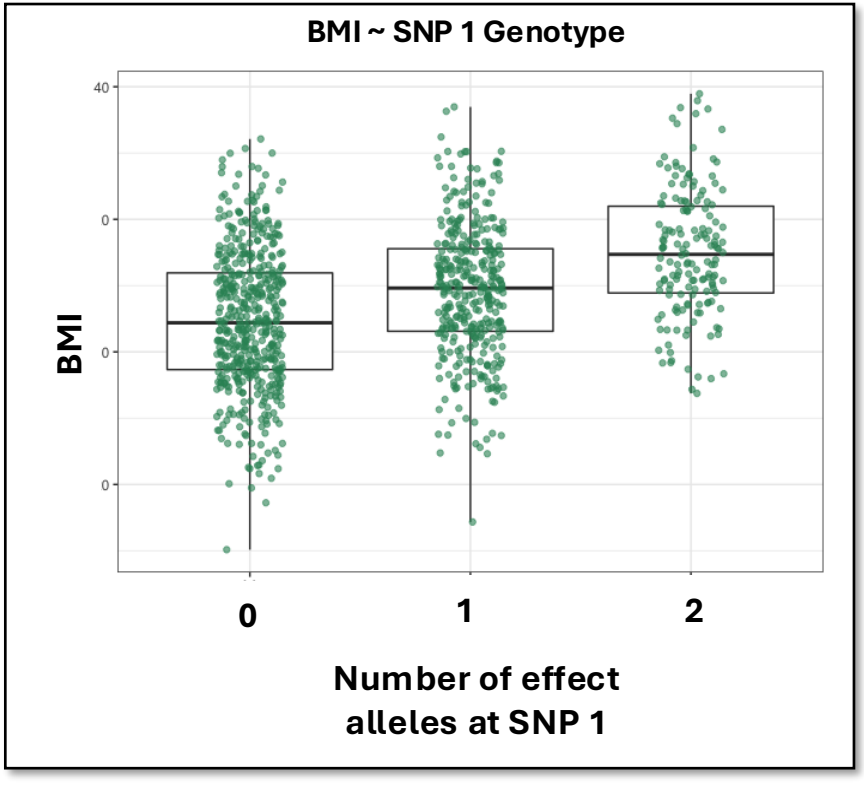
2. Using linear regression, estimate the relationship between genotype and BMI
   - β (Beta) = slope. The amount BMI changes for each additional effect allele at SNP 1



Note: logistic regression is a slightly different model used for binary traits (e.g., case-control status)

GWAS linear regression model at a single SNP

3. Repeat for each SNP in your data, genome-wide



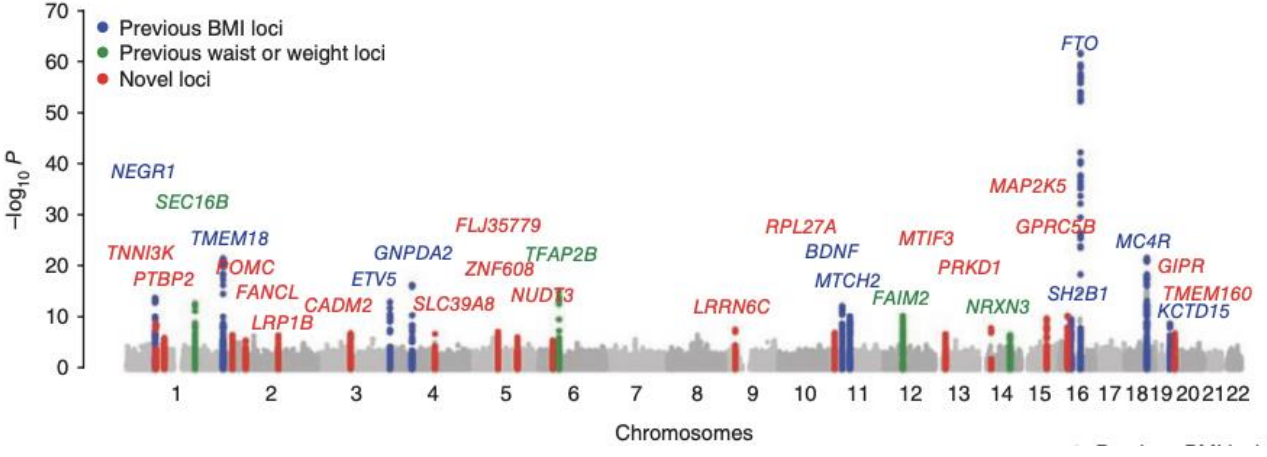
In reality, computers and specialized statistical genetics software do these calculations for us

GWAS Manhattan Plot

A tool to visualize our loci significantly associated with BMI

Each point is a SNP which we tested for association with BMI

Y-axis = Statistical strength of the association (transformed P-value)

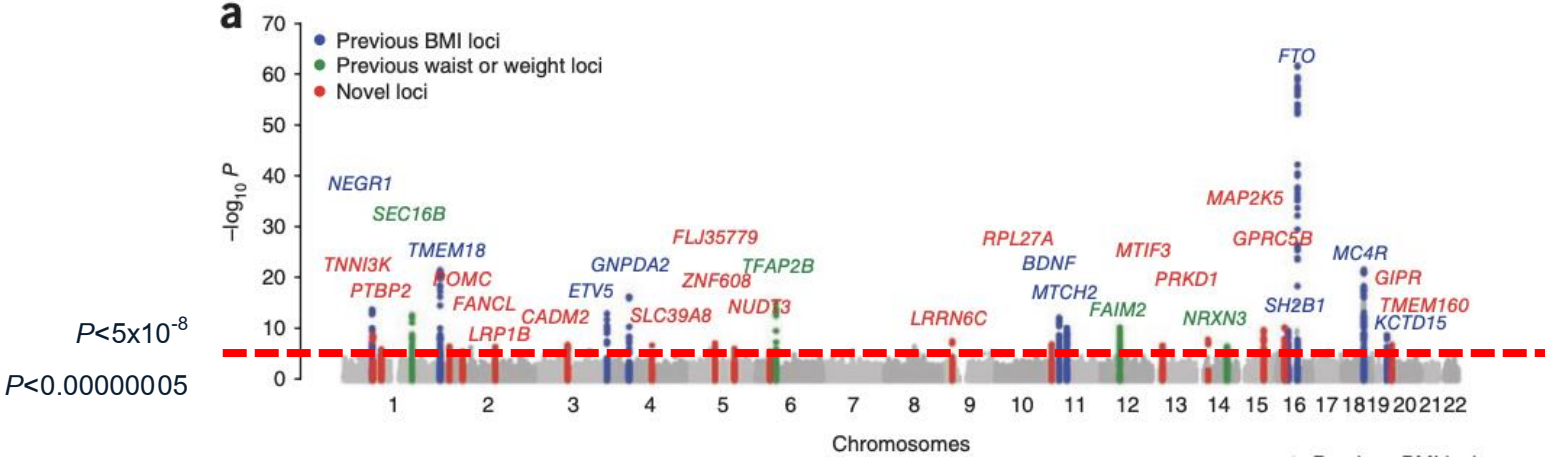


x-axis = chromosomal location of each tested SNP

GWAS Manhattan Plot

Additional concepts to take-away

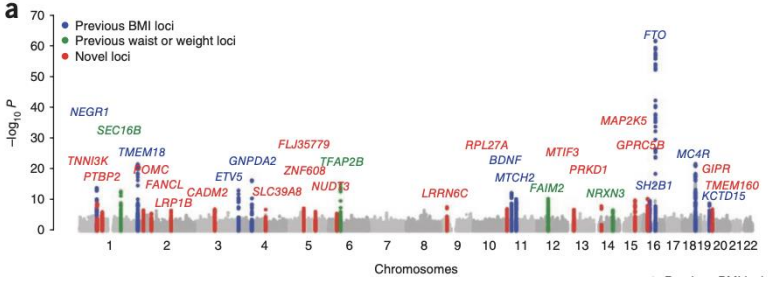
- Why is the genome-wide significance threshold so stringent ($P<5.0 \times 10^{-8}$)?
- Why do significant SNPs cluster together in peaks of association signal?



The $P<5.0x10^{-8}$ genome-wide significance threshold

- $P<0.05$ is the most widely used significance threshold for a single statistical test
  - The probability of getting your result due to chance alone (and not a real association) is less than 5%
- Multiple Testing Problem: What happens when you conduct 100 tests?
  - 5/100 tests will be false positives
  - The more tests we conduct, the more likely any given test will be a false positive
- Multiple Testing Correction: Adjust your significance threshold to still reflect the wanted 5% false positive rate

The $P<5.0x10^{-8}$ genome-wide significance threshold



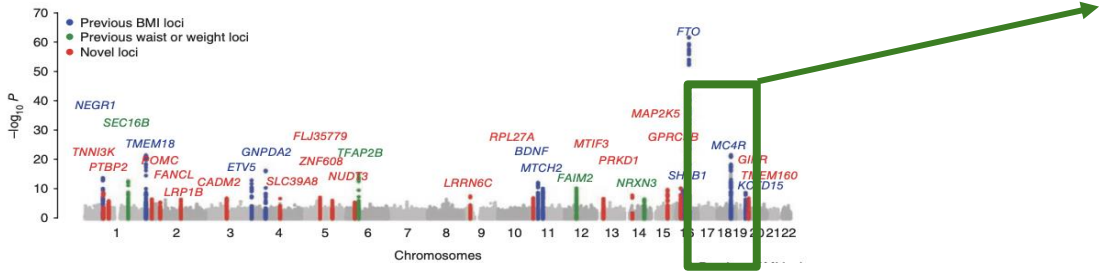
Should we adjust for the ~10 million common variants that we tested for association with BMI?

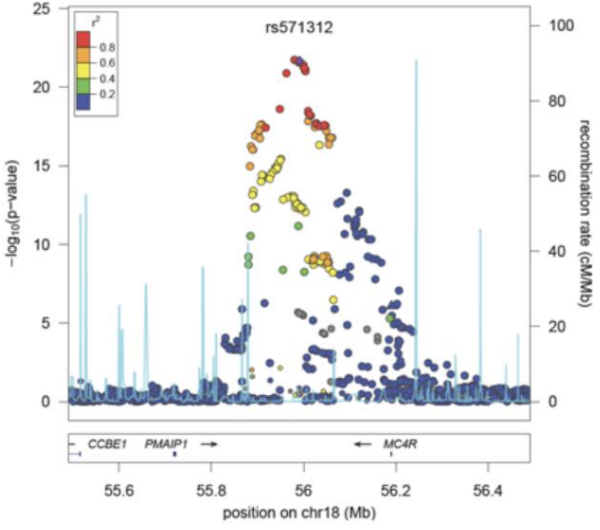
... Not quite

- The genotypes of genetic variants that are physically close together are correlated (they are in LD)
  - Linkage disequilibrium (LD) – the correlation between SNPs arising due to their physical proximity and the probability of recombination events
- Approximately only ~1 million independent common genetic variants in the genome (in individuals of European ancestry)
- $P < 0.05/1,000,000 = 0.00000005 = 5.0x10^{-8}$

Why do significant SNPs cluster together in peaks of association signal?

Regional Manhattan Plot = Zoomed in Manhattan plot





- The genotypes of SNPs close together (in linkage disequilibrium) are correlated → their association with BMI will be correlated too!

- Important. GWAS identify clusters of correlated SNPs in loci associated with your trait of interest. These loci may overlap genes, suggesting their importance. However, GWAS do not directly identify the causal genes in the locus.
- GWAS tend to name loci by their nearest gene, but this does not mean we are confident it is the causal gene

GWAS summary statistics table

SNP	Nearest gene	Other nearby genes[a]	Chr.	Position[b] (bp)	Alleles[b] Effect	Alleles[b] Other	Frequency effect allele	Per allele change in BMI β (s.e.m.)[c]	Explained variance (%)	Stage 1 *P*	Stage 2 *P*	Stage 1 + 2 *n*	Stage 1 + 2 *P*
Previously identified BMI loci													
rs1558902	*FTO*		16	52,361,075	A	T	0.42	0.39 (0.02)	0.34%	2.05×10^{-62}	1.01×10^{-60}	192,344	4.8×10^{-120}
rs2867125	*TMEM18*		2	612,827	C	T	0.83	0.31 (0.03)	0.15%	2.42×10^{-22}	4.42×10^{-30}	197,806	2.77×10^{-49}
rs571312	*MC4R* (B)		18	55,990,749	A	C	0.24	0.23 (0.03)	0.10%	1.82×10^{-22}	3.19×10^{-21}	203,600	6.43×10^{-42}
rs10938397	*GNPDA2*		4	44,877,284	G	A	0.43	0.18 (0.02)	0.08%	4.35×10^{-17}	1.45×10^{-15}	197,008	3.78×10^{-31}
rs10767664	*BDNF* (B,M)		11	27,682,562	A	T	0.78	0.19 (0.03)	0.07%	5.53×10^{-13}	1.17×10^{-14}	204,158	4.69×10^{-26}
rs2815752	*NEGR1* (C,Q)		1	72,585,028	A	G	0.61	0.13 (0.02)	0.04%	1.17×10^{-14}	2.29×10^{-9}	198,380	1.61×10^{-22}

- Binary traits/case-control outcomes (e.g., Obese vs. non-obese) are fit using a logistic regression and the beta is transformed into an odds ratio (OR)
  - Odds ratio (OR): the odds of having obesity for carriers of the effect allele / the odds of having obesity for non-carriers of the effect allele

Lead SNP for naming purposes. The locus makes up many correlated genetic variants in LD.

The utility and limitations of GWAS

- GWAS has been used to identify thousands of loci associated with polygenic traits in an unbiased manner
  - Works well for common variants and common traits
- Small effect sizes can have real impact
  - Nominate candidate genes in which bigger perturbations have a larger impact on the phenotype
  - Aggregation of many variants with small effect sizes into a polygenic score can together have a moderate effect on the phenotype
- Limitations
  - Association does not mean causality: GWAS does not identify the causal variant or gene, and getting to this next step is non-trivial
  - Requires very large sample sizes (in the thousands, at least)
  - Most GWAS have been conducted in people of European genetic ancestry and the GWAS associations and derived polygenic scores may not translate well to other more diverse populations

Sneak Peek at Polygenic Scores

- Polygenic risk scores (PRS) = Polygenic Scores (PGS)
- PGS is a measure of an individual's overall genetic risk or propensity for a given disease or trait
  - They are calculated as the sum of their genome-wide genetic effects, calculated from genome-wide GWAS data

How to calculate a simple PRS?

The simplest model:

1) For each individual, and for each SNP, multiply the number of effect alleles (0, 1, or 2) by the effect (beta) on the trait of interest
2) PRS = Sum of these values (the aggregate risk for each individual)

Example with 3 genome-wide significant loci
Using the lead SNP from each independent locus

Individual	SNP1* (beta = 2.2)	SNP2* (beta = 1.4)	SNP3* (beta = 1.6)	PRS
1	0	2	0	2.8
2	2	1	0	5.8
3	1	2	0	5.0
4	0	0	1	1.6

Individual 1 PRS = 0*2.2 + 2*1.4 + 0*1.6 = 2.8

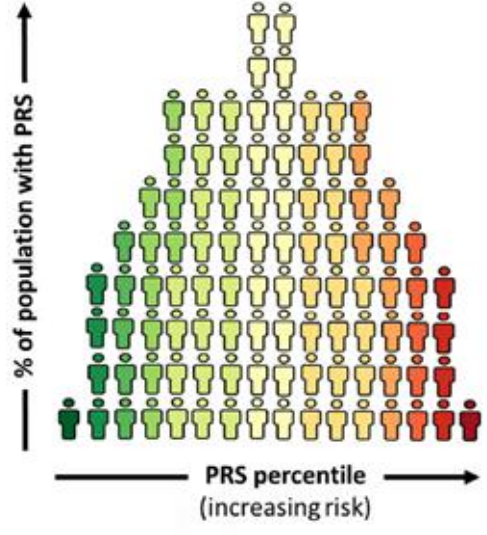
Note: More complex models take into account genome-wide genetic effects and correlation between SNPs

What does a PRS mean?

- The higher your PRS the higher your aggregated genetic risk for the disease of interest

Figure 2: PRS Distribution

PRS percentile	Risk of disease vs. reference group
0-1	Lowest
1-5	
5-10	
10-20	
20-40	
40-60 (reference)	1
60-80	
80-90	
90-95	
95-99	
99-100	Highest

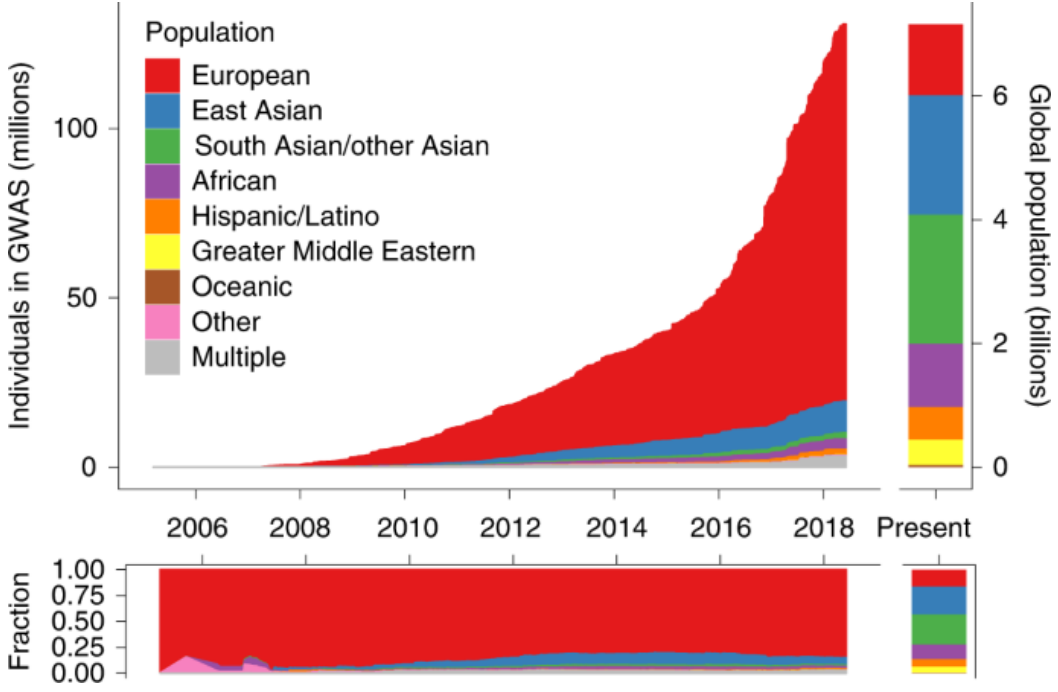


Source: RGA

A current important hot topic:

Increasing diversity in genetic studies

- Most genetic studies to date have been done in individuals of European ancestry
- How well the GWAS signals and PGS translate to non-White populations is a major area of research



Martin *et al., Nature Genetics* 2019