Announcement

- Final presentation schedules are posted in slack channel #general:
  - Thursday, April 6: **Lead, Copper, Gold**
  - Tuesday, April 11: **Aluminum, Platinum, Zinc**
  - Thursday, April 13: **Iron, Cobalt**

Statistical Methods for High-dimensional Biology

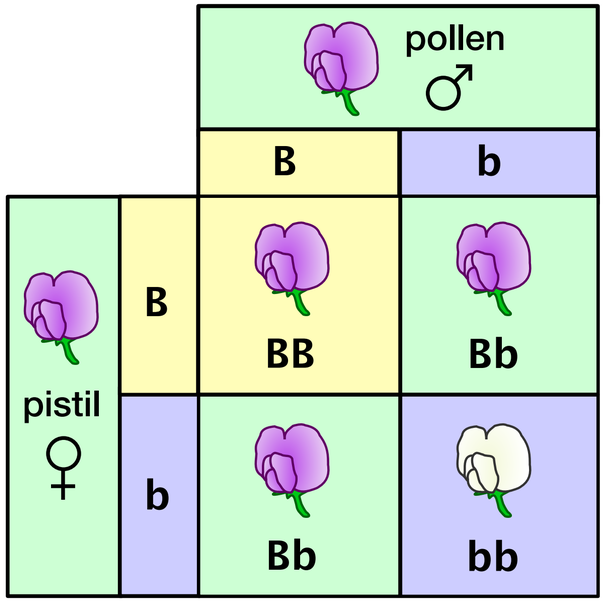
Genome-wide Association Studies

Yongjin Park, UBC Path&Lab, STAT, BC Cancer

Today’s lecture: GWAS and related topics

- **Human Genetics 101**
  - Variation in the human genome
  - How do we measure genetic associations?
- **Polygenic models**
  - Population structures
  - Linear mixed effect model
- **Systems Genetics**
  - Summary-based GWAS analysis
  - LD-score regression: “enrichment analysis” in GWAS

Genetics: It all started from pea plants

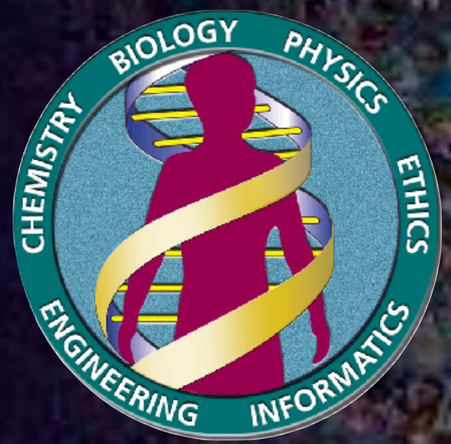


Gregor Mendel
(1822-1884)

Key concepts emerged:

- Gene = a unit of heredity that transfers from a parent to offspring
- Allele = a different form of a gene [from a Greek word, αλληλο, αλλος, "allos", other]

Human Genome Project



A reference DNA sequence for
3.2B basepairs x diploid for each individual

https://www.genome.gov/human-ger

Human genetics revolution

To Solve 3 Cold Cases, This Small County Got a DNA Crash Course

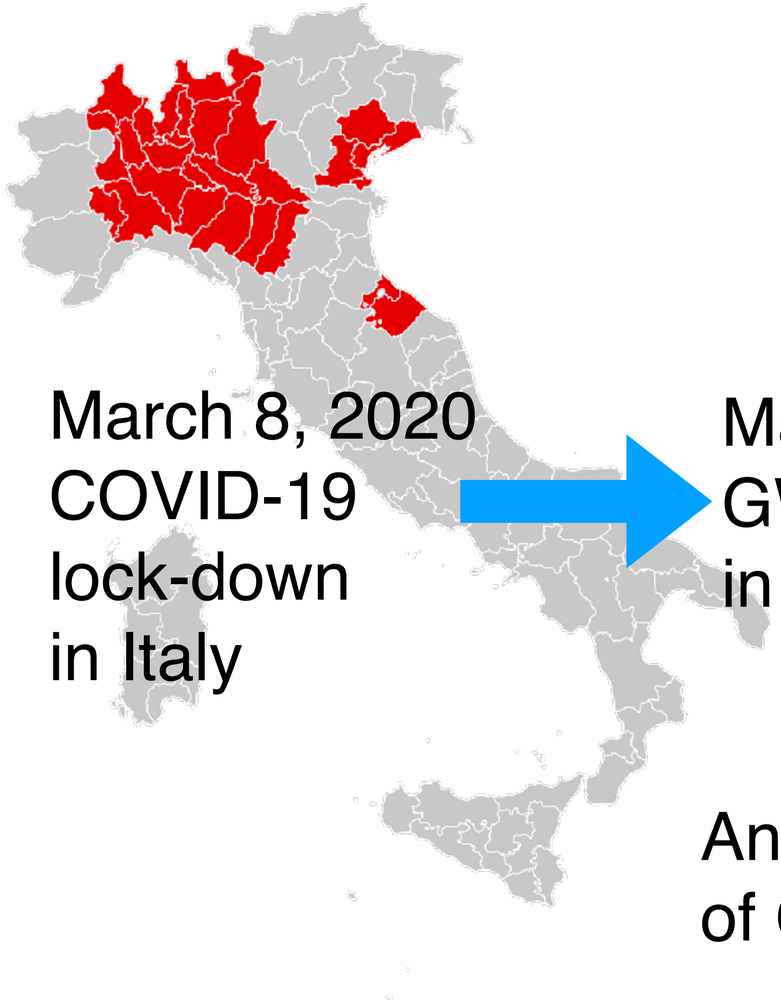
Forensic genealogy helped nab the Golden State Killer in 2018. Now investigators across the country are using it to revisit hundreds of unsolved crimes.

https://www.nytimes.com/2021/05/03/science/cold-cases-genetic-genealogy.html

Human Genomes help keep track of human evolution

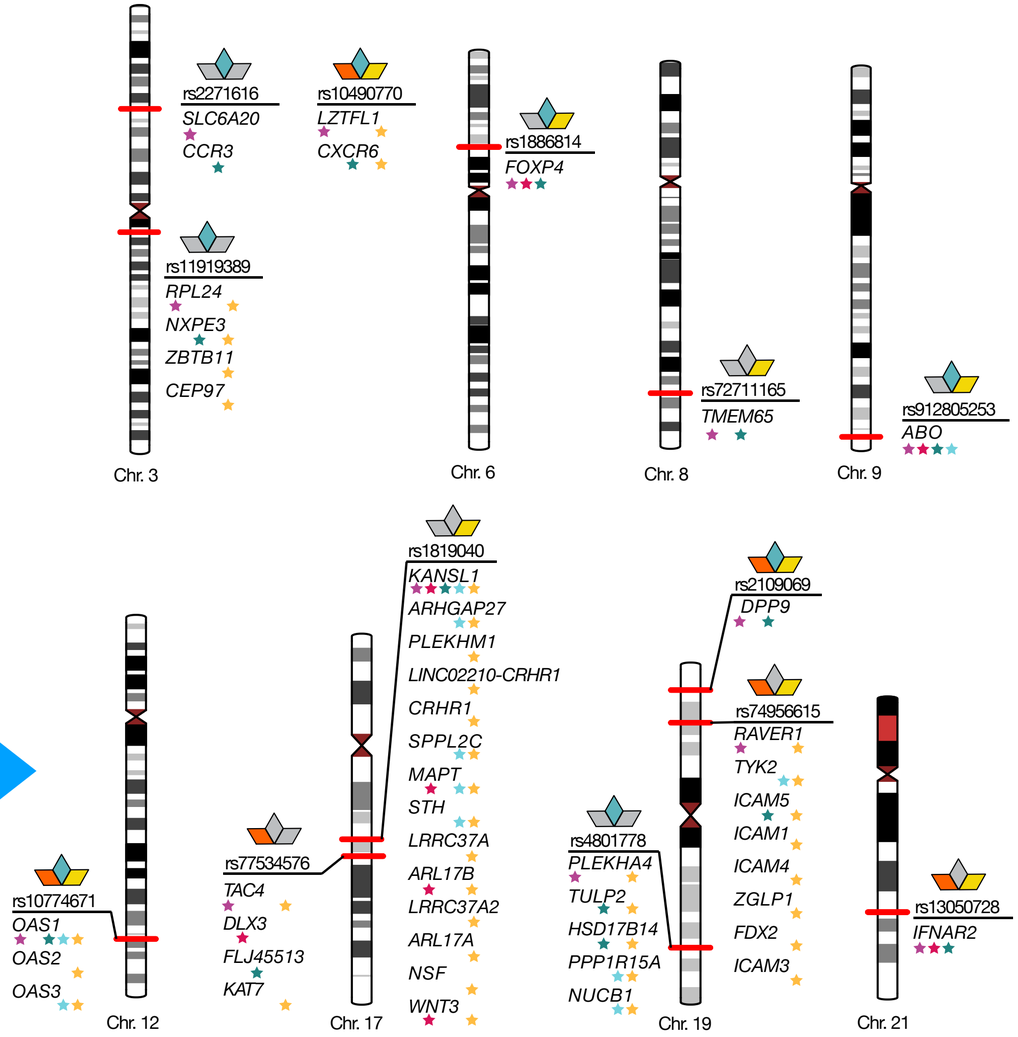
https://www.nytimes.com/2018/03/20/science/david-reich-human-migrations.html

COVID-19 Host Genetics



March 8, 2020
COVID-19
lock-down
in Italy

March 12, 2021,
GWAS paper
in medRxiv

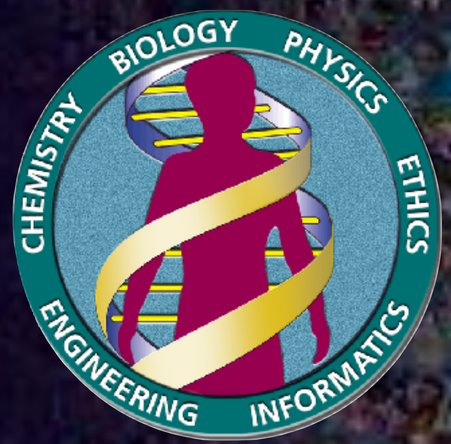


An unprecedented pace
of GWAS profiling

https://www.medrxiv.org/content/10.1101/2021.03.10.21252820v1.article-info

COVID-19 Host Genetics Initiative, *Nature* (2021)

Human Genome Project



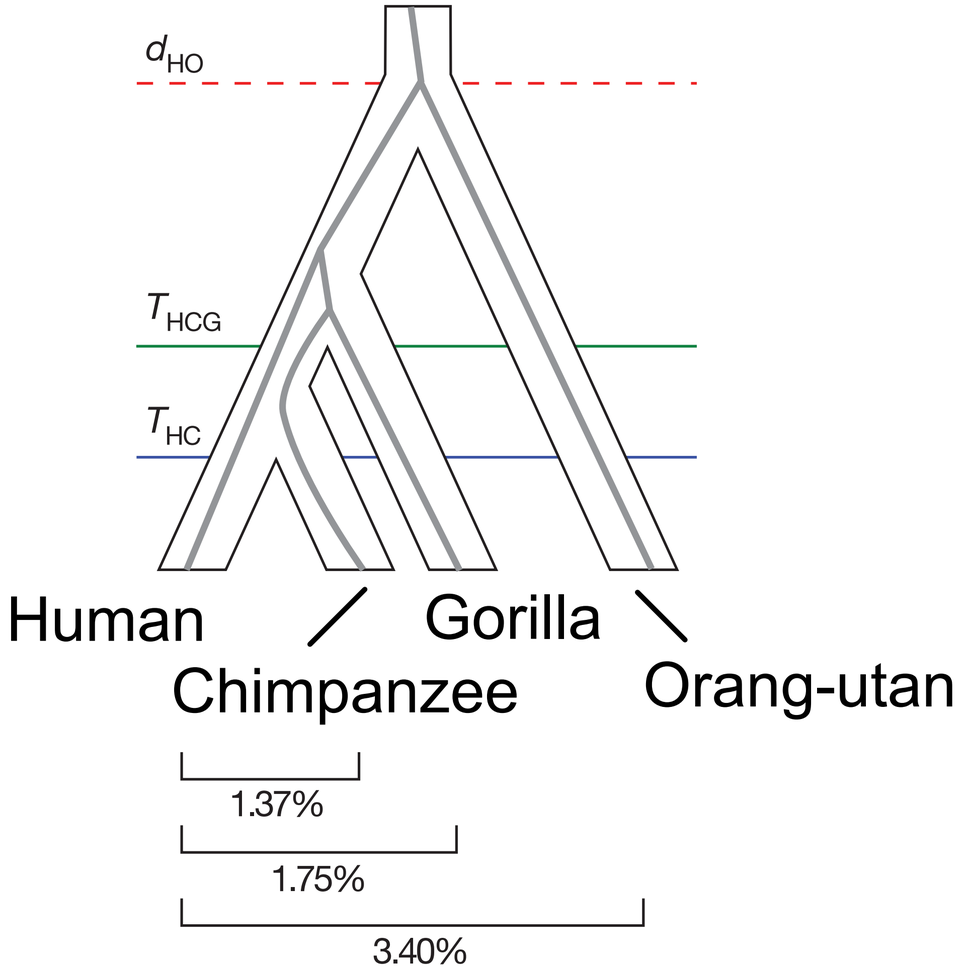
A reference DNA sequence for
3.2B basepairs x diploid for each individual

https://www.genome.gov/human-ger

How did Human Genome Project start the revolution?

We have a reference panel of genomic sequence information...

Why is it important to me?



99% genetic information shared across humans

Scally .. Durbin, *Natur*

If we are 99% identical, then
what is the 1% difference?

Which part of the human genome is variable?

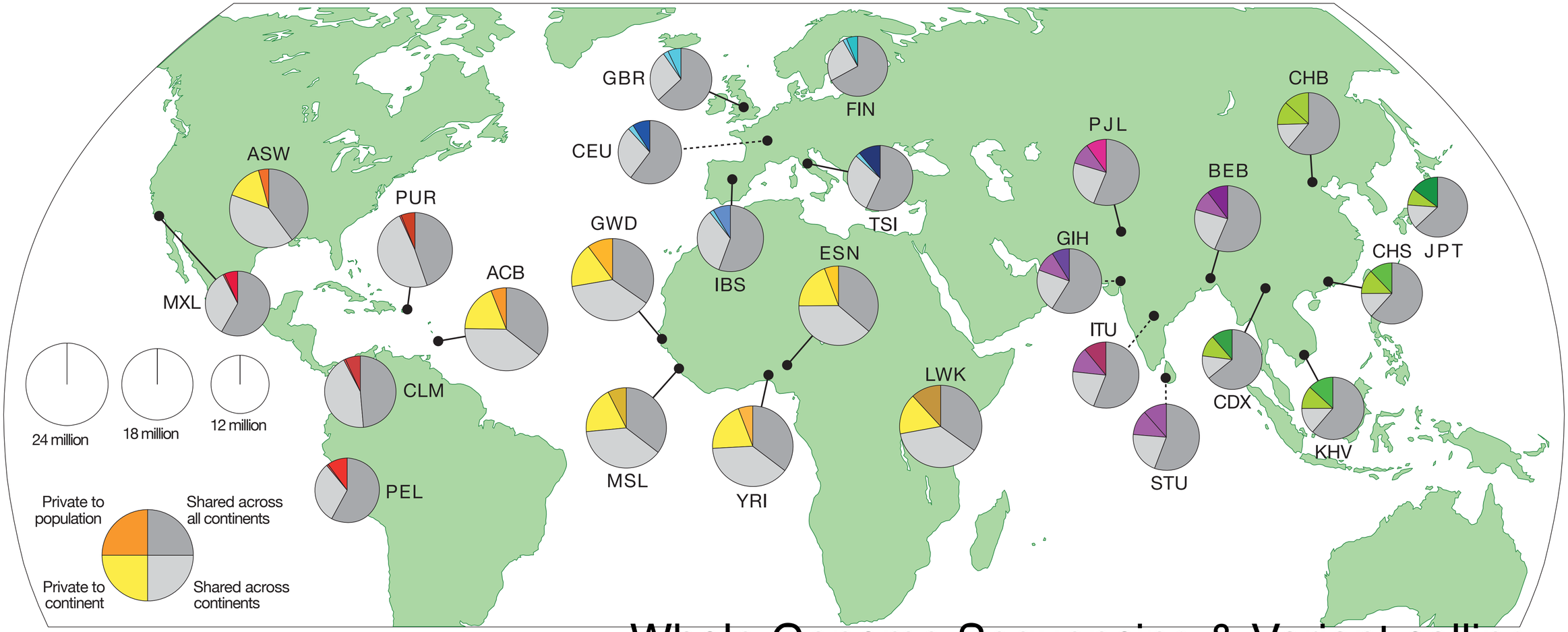
Published: 18 December 2003

The International HapMap Project

†The International HapMap Consortium

Nature **426**, 789–796 (2003) | Cite this article

The 1000 genomes project



Whole Genome Sequencing & Variant calling across 1000 individuals in many different groups

The 1000 Genomes Project Consortium, Nature, 2015

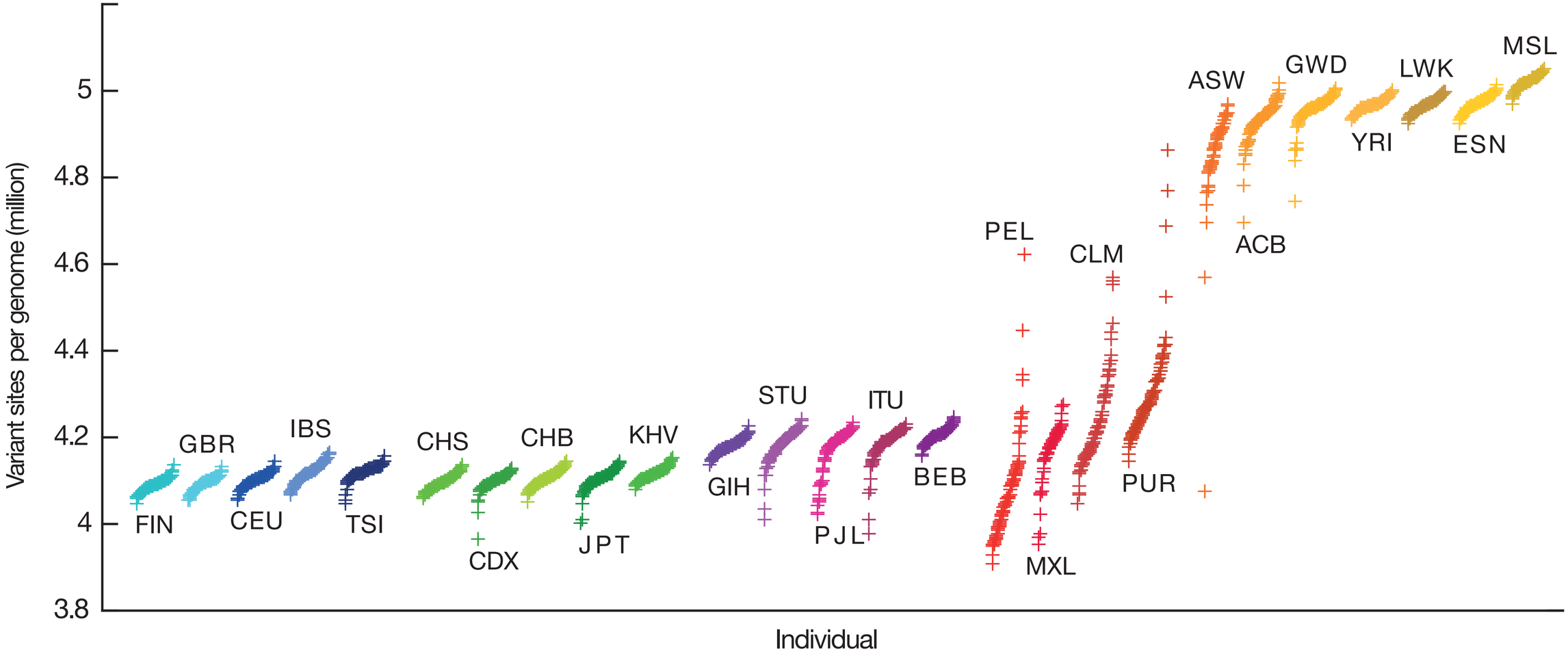
Genetic variation across human population

A typical genome

We find that a typical genome differs from the reference human genome at 4.1 million to 5.0 million sites (Fig. 1b and Table 1). Although >99.9% of variants consist of SNPs and short indels, structural variants affect more bases: the typical genome contains an estimated 2,100 to 2,500 structural variants (~1,000 large deletions, ~160 copy-number variants, ~915 Alu insertions, ~128 L1 insertions, ~51 SVA insertions, ~4 NUMTs, and ~10 inversions), affecting ~20 million bases of sequence.

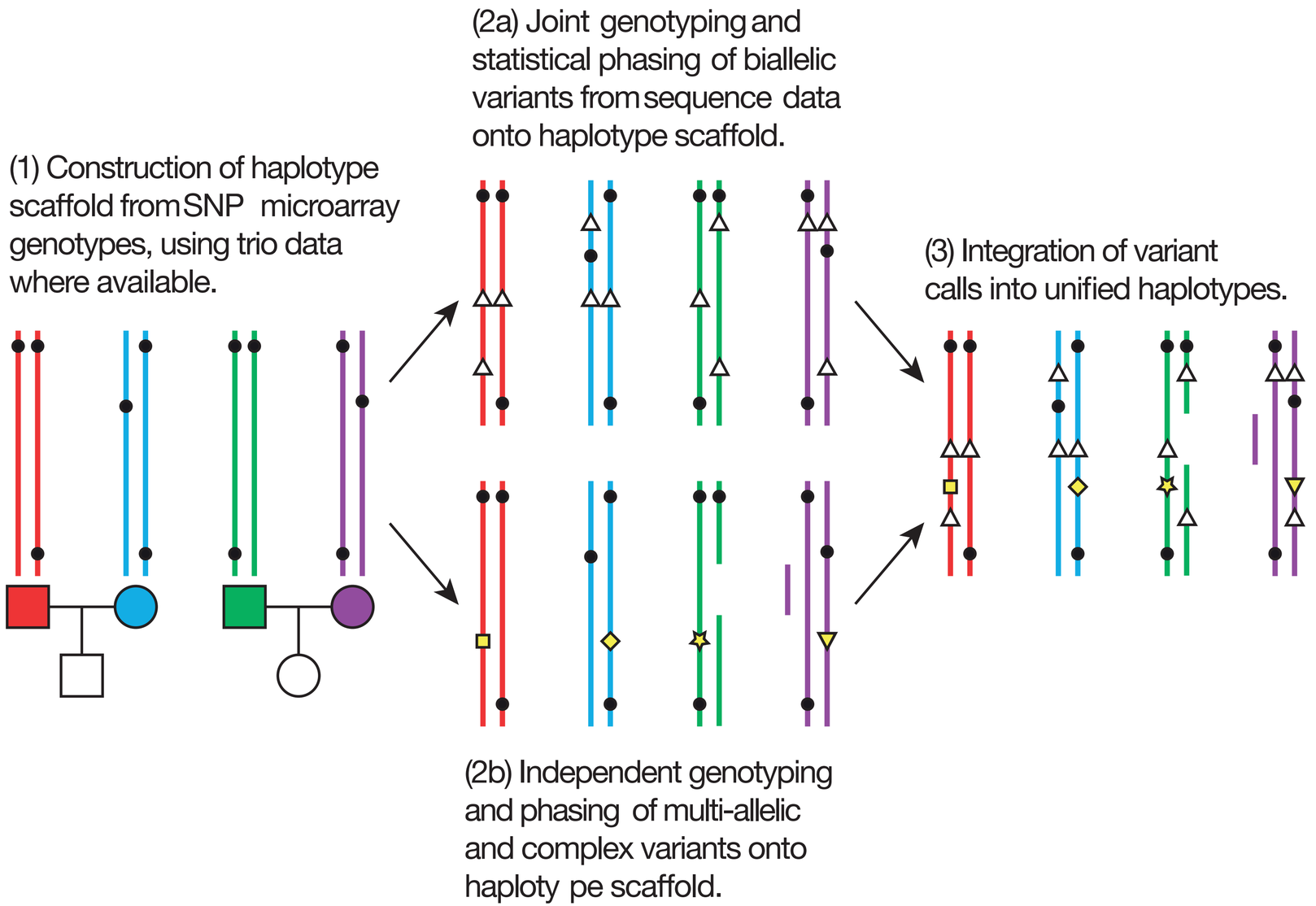
The 1000 Genomes Project Consortium, Nature, 2015

A typical genome differs from the reference at 4.1 to 5 million sites



The 1000 Genomes Project Consortium, Nature, 2015

Haplotype


(2a) Joint genotyping and statistical phasing of biallelic variants from sequence data onto haplotype scaffold.
(1) Construction of haplotype scaffold from SNP microarray genotypes, using trio data where available.
(3) Integration of variant calls into unified haplotypes.
(2b) Independent genotyping and phasing of multi-allelic and complex variants onto haploty pe scaffold.

Allele

A different form of a gene [from a Greek word, αλληλο, αλλος, "allos", other]

A different version of a gene

A different version of the same variant

Mostly used for a gene

Variant

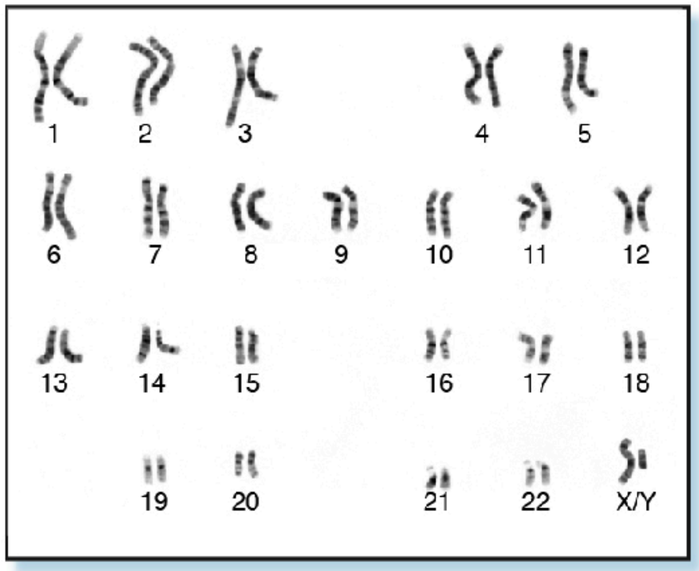
A specific region of the genome differs between two genomes.

Across two or more genomes, there could have genetic variants due to a mutational process

Ploidy (diploid)

One copy from maternal genome another copy from paternal genome

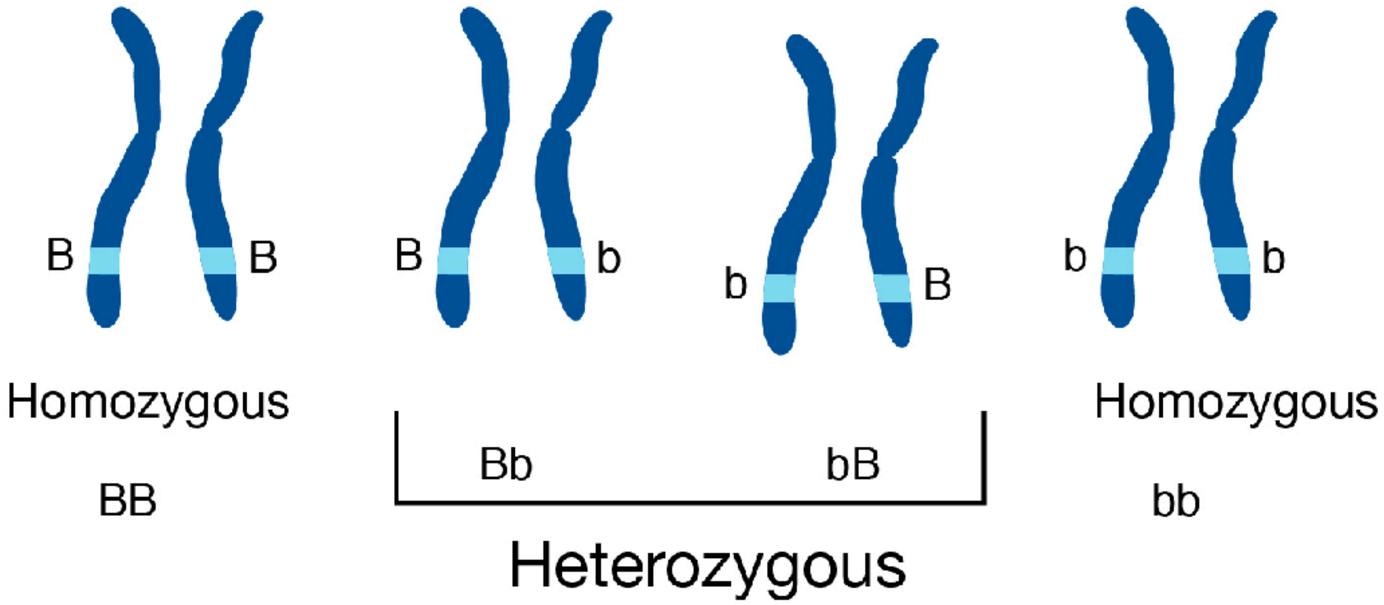
When do we have polyploidy or aneuploidy?



https://www.genome.gov/genetics-glossary/Diploid

Biallelic vs. triallelic vs. multiallelic

"Bi"-allelic: two different forms exist for this heritable unit (most of the human variants)
"Tri"-allelic: three different forms exist (much rare than biallelic variants)



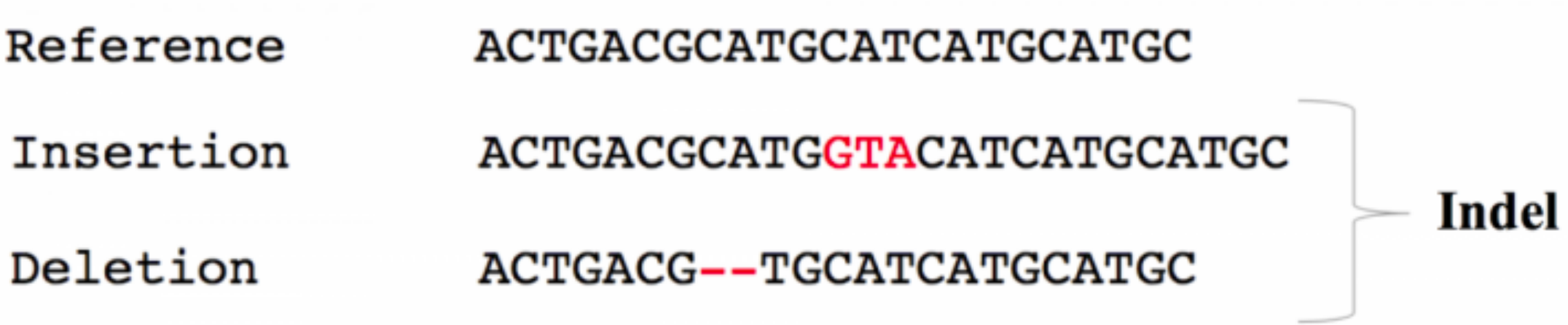
For biallelic variant:
- reference allele
- alternative allele

https://www.genome.gov/genetics-glossary/heterozygous

Many types of genetic variants

ACTCGTGACCGCATGCATCTTCATTCATGC **Reference**

ACTCGTGACCGCATGCATCGTCATTCATGC **Alternative**

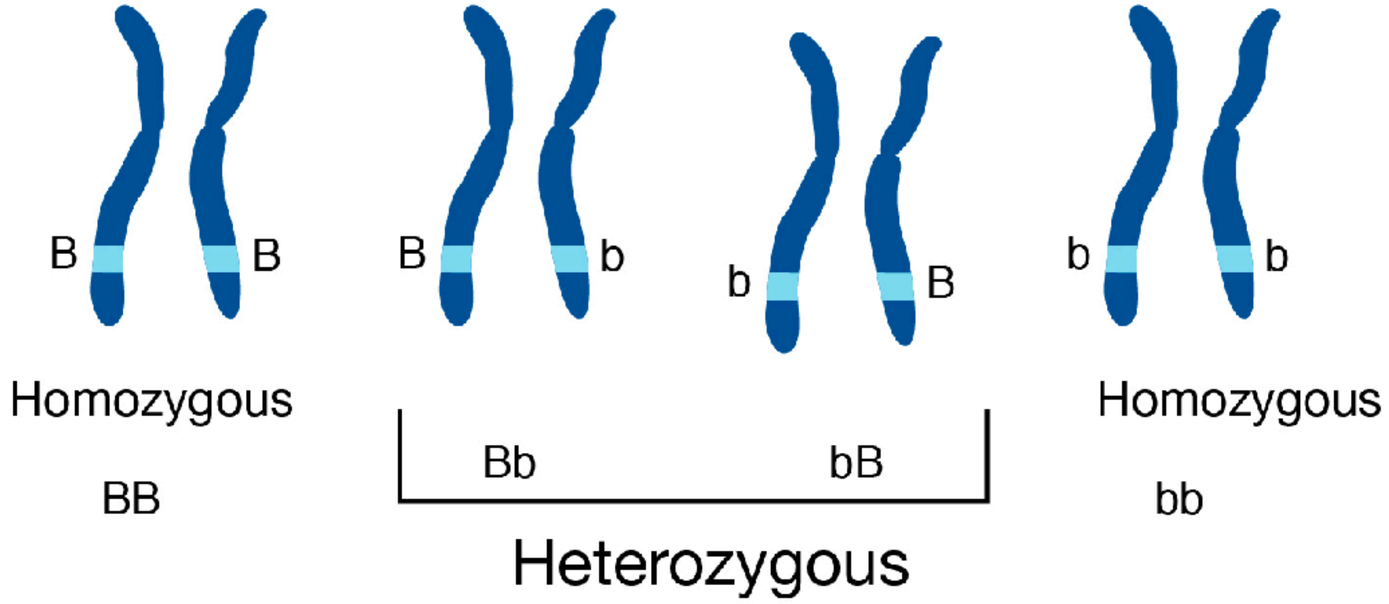


https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction

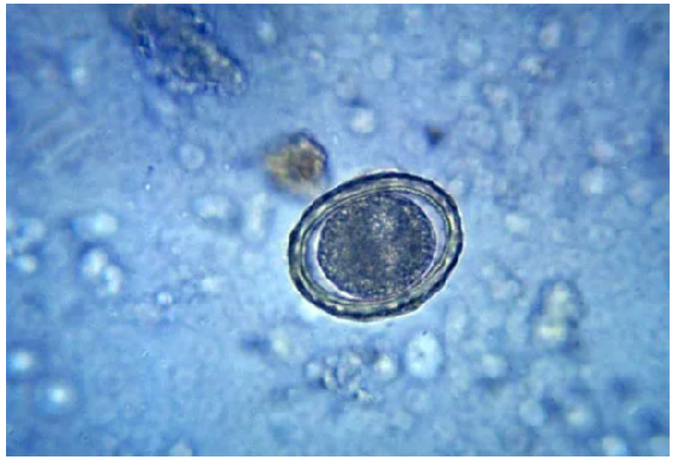
Homozygote vs. heterozygote

One copy from maternal genome
another copy from paternal genome

Homozygous: the maternal == paternal copy
Heterozygous: the maternal != paternal copy



- "Homo" vs. "hetero" ⟺ the same vs. different
- Zygote (fertilized egg cell)

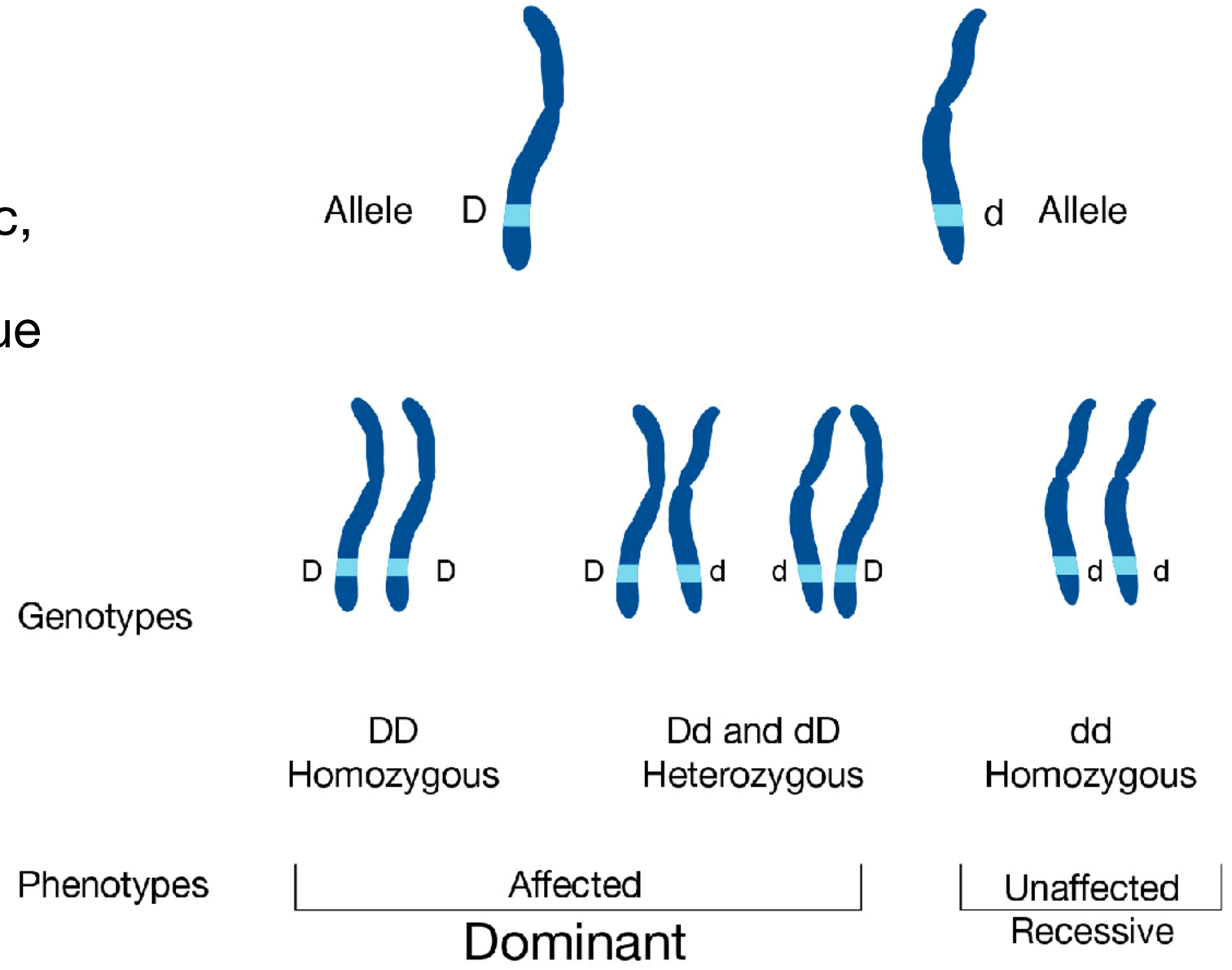


https://www.genome.gov/genetics-glossary/heterozygous https://www.britannica.com/scien

Huntington's Disease

Dominance

If one allele of a gene is toxic, the other allele of the same, wildtype gene may not rescue



https://www.genome.gov/genetics-glossary/Dominant

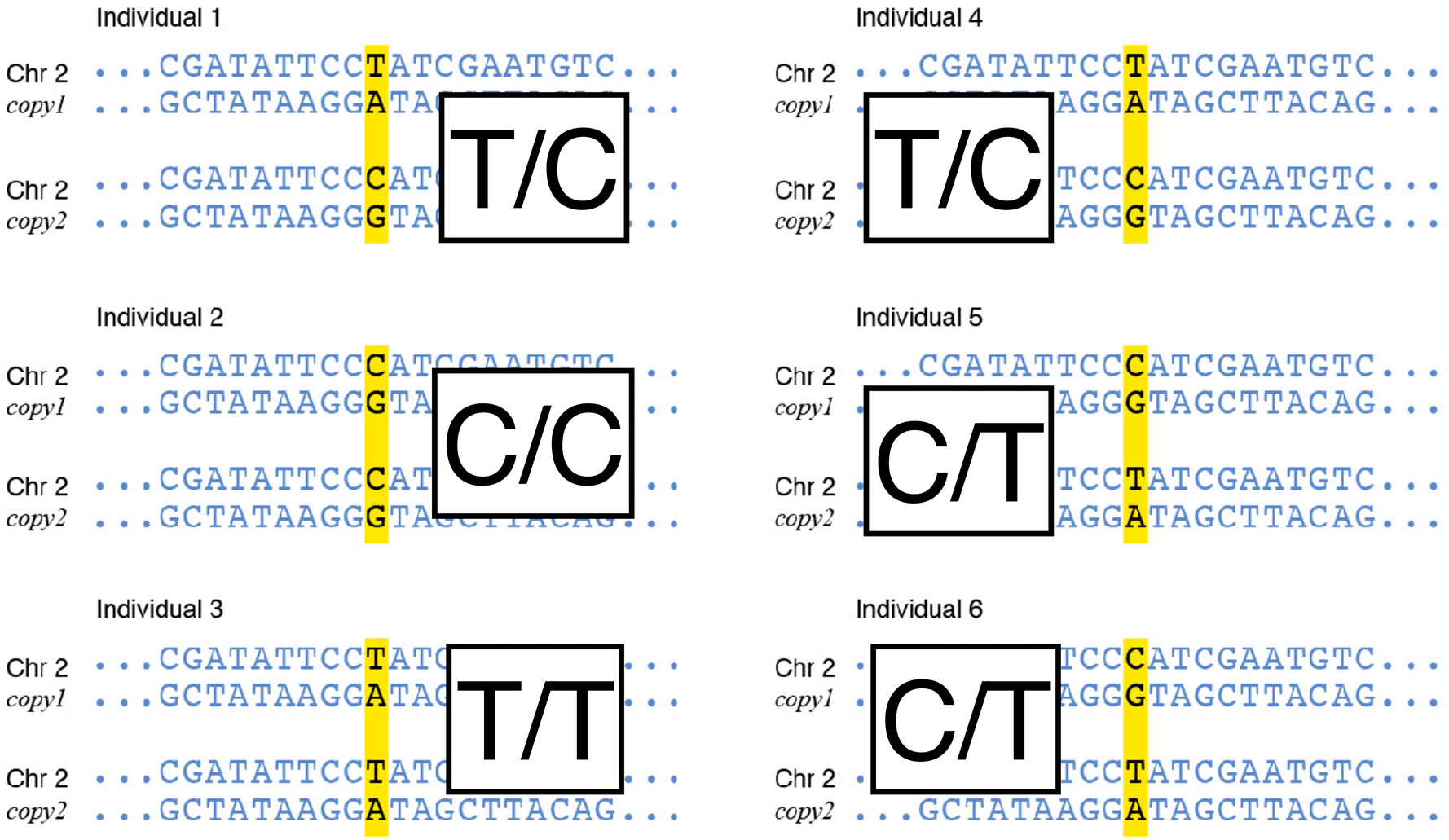
Single Nucleotide Polymorphism

~1 in every 1000bp

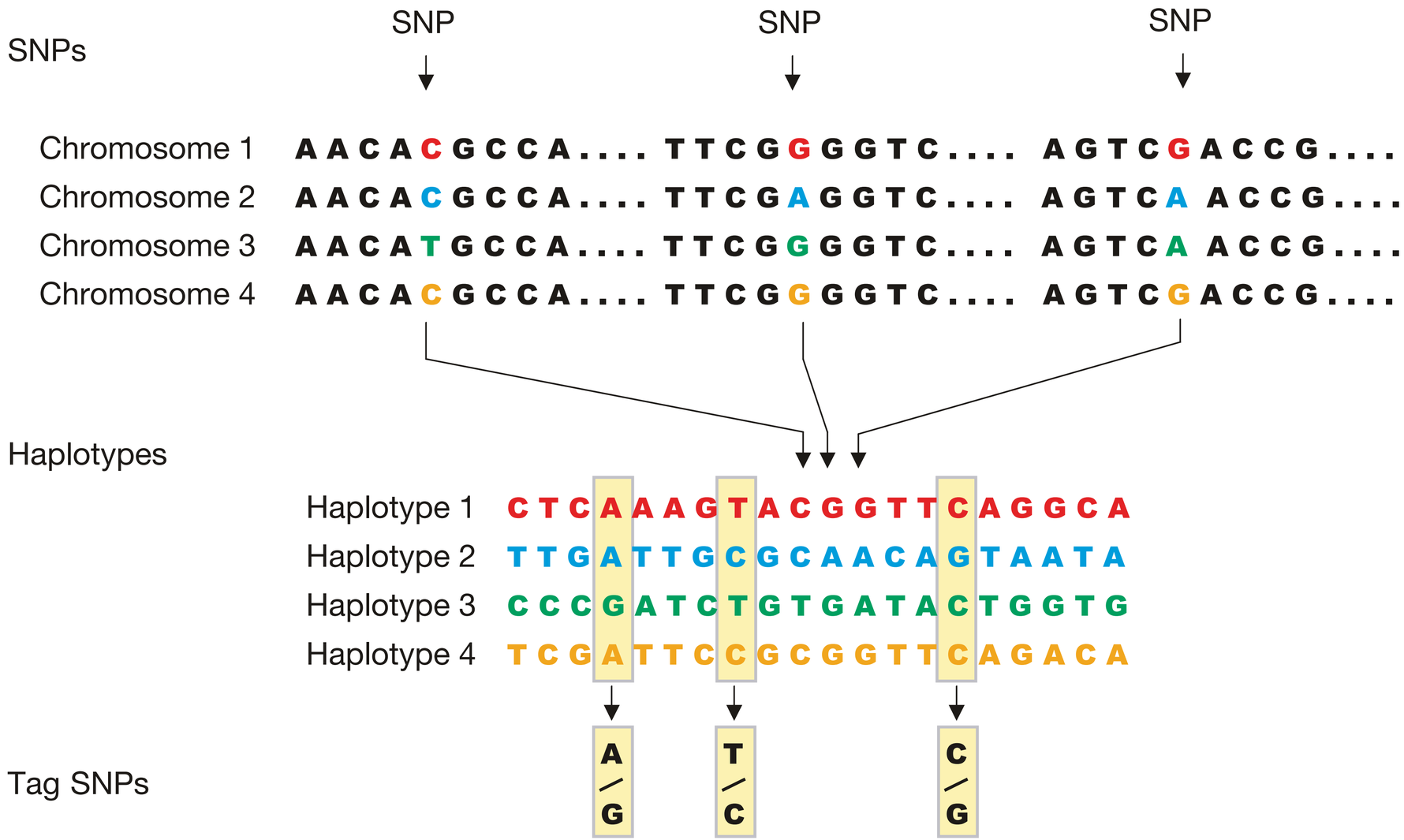
Single nucleotide polymorphism is way too many syllables, so you can understand why we just say "snip". And this is really a simple concept.

These are the places in the genome where people differ.

Francis Collins

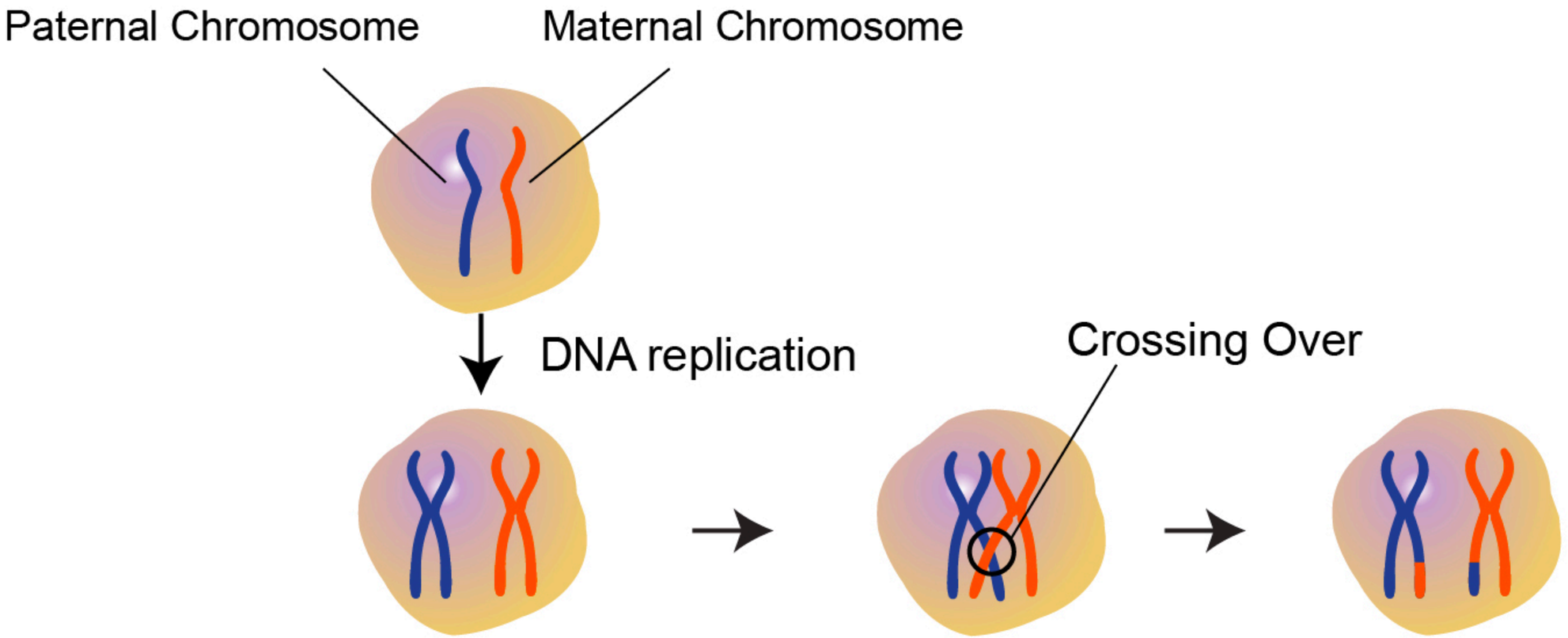


~ 1M common SNPs



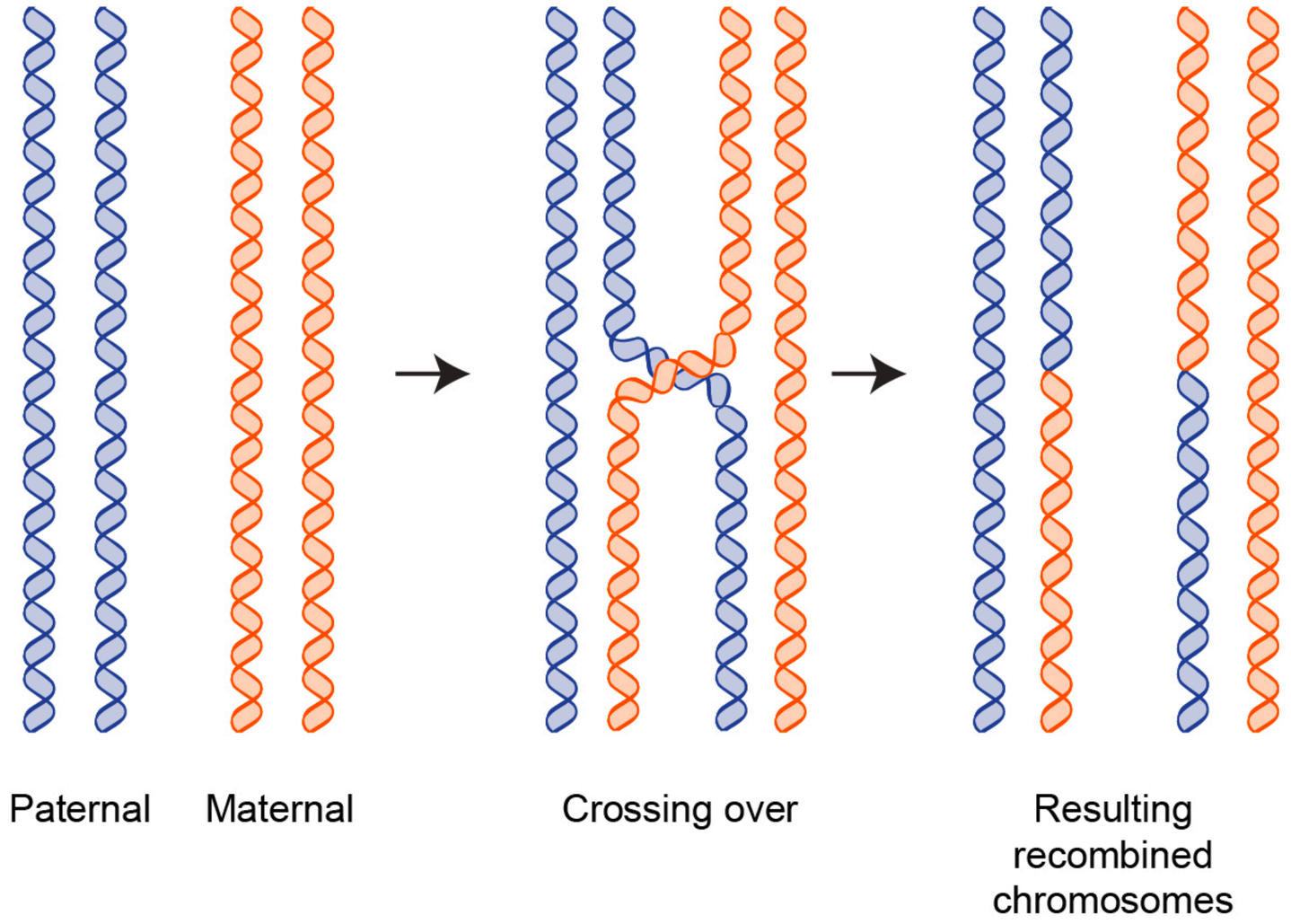
The International HapMap Project (2003)

Recombination: Mixing the maternal and paternal copies



https://www.genome.gov/genetics-glossary

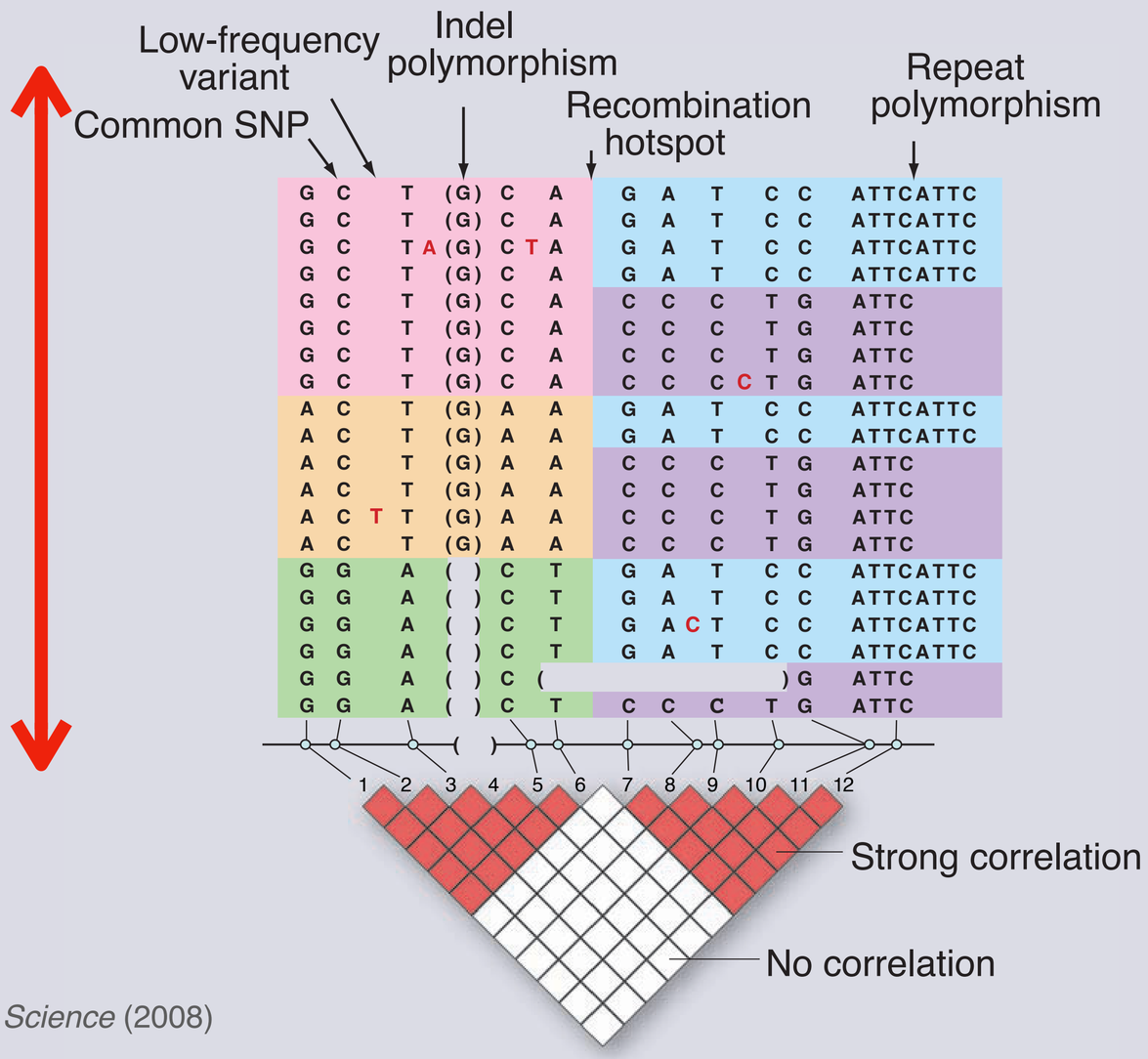
Recombination: Mixing the maternal and paternal copies



https://www.genome.gov/genetics-glossary

Genetic variation in one figure

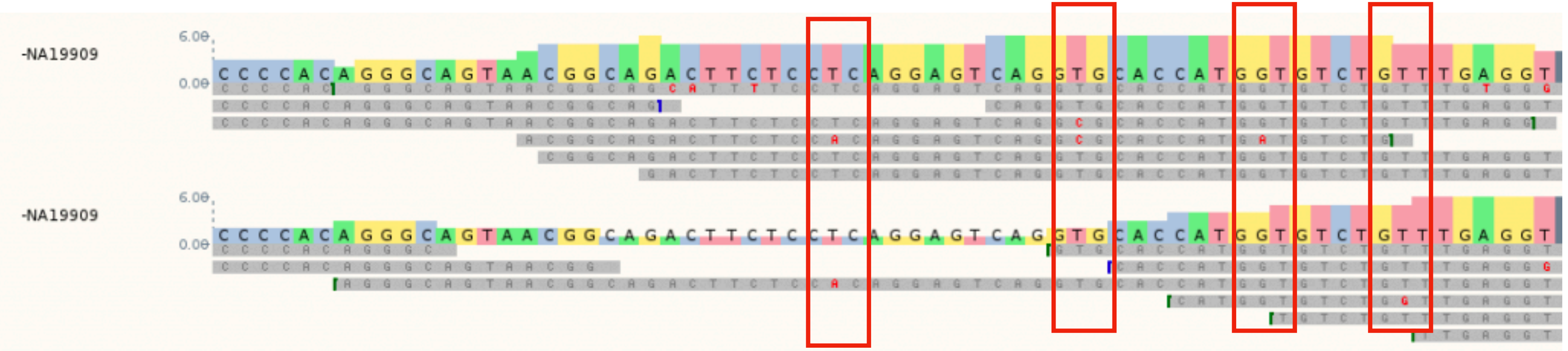
across many individuals (diploid genomes)



Common SNPs
Insertion/deletion
Other low-freq. variants
Other structural variants
Recombination hotspot

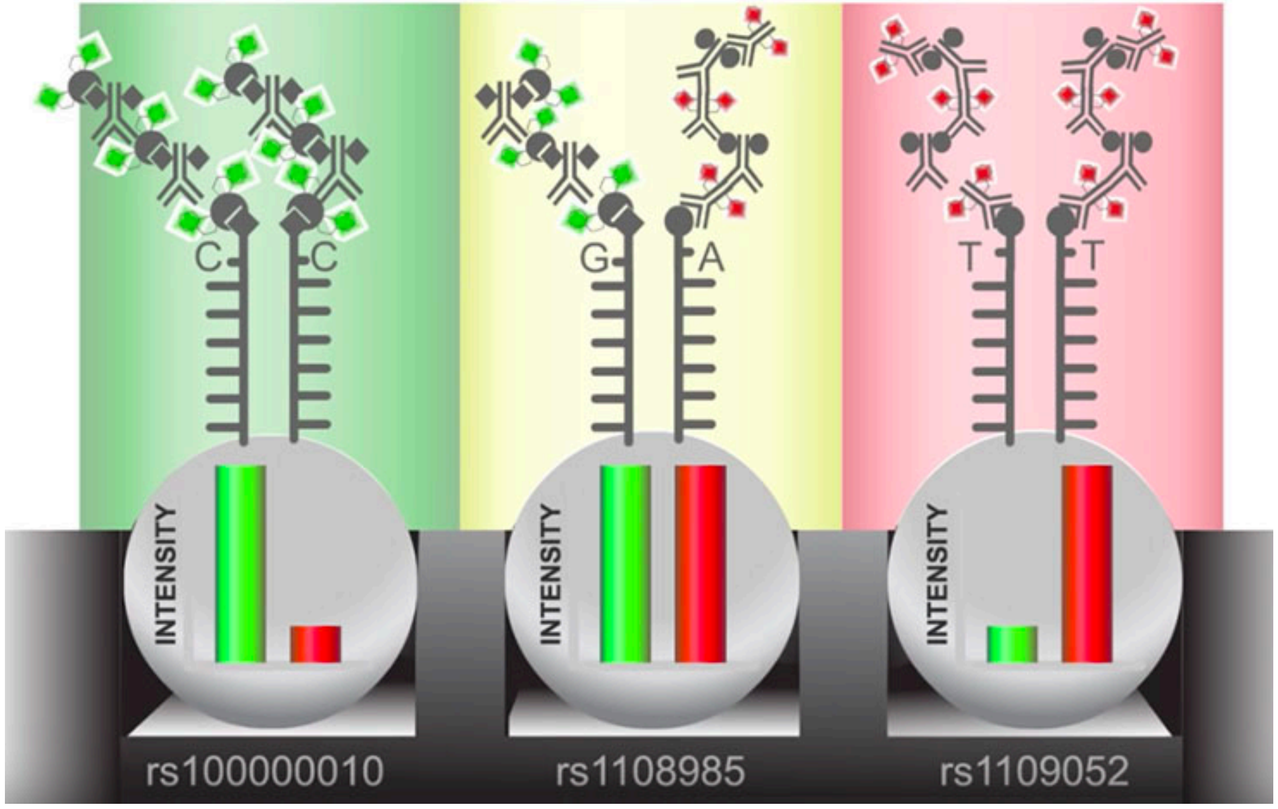
Altshuler, Daly, Lander, *Science* (2008)

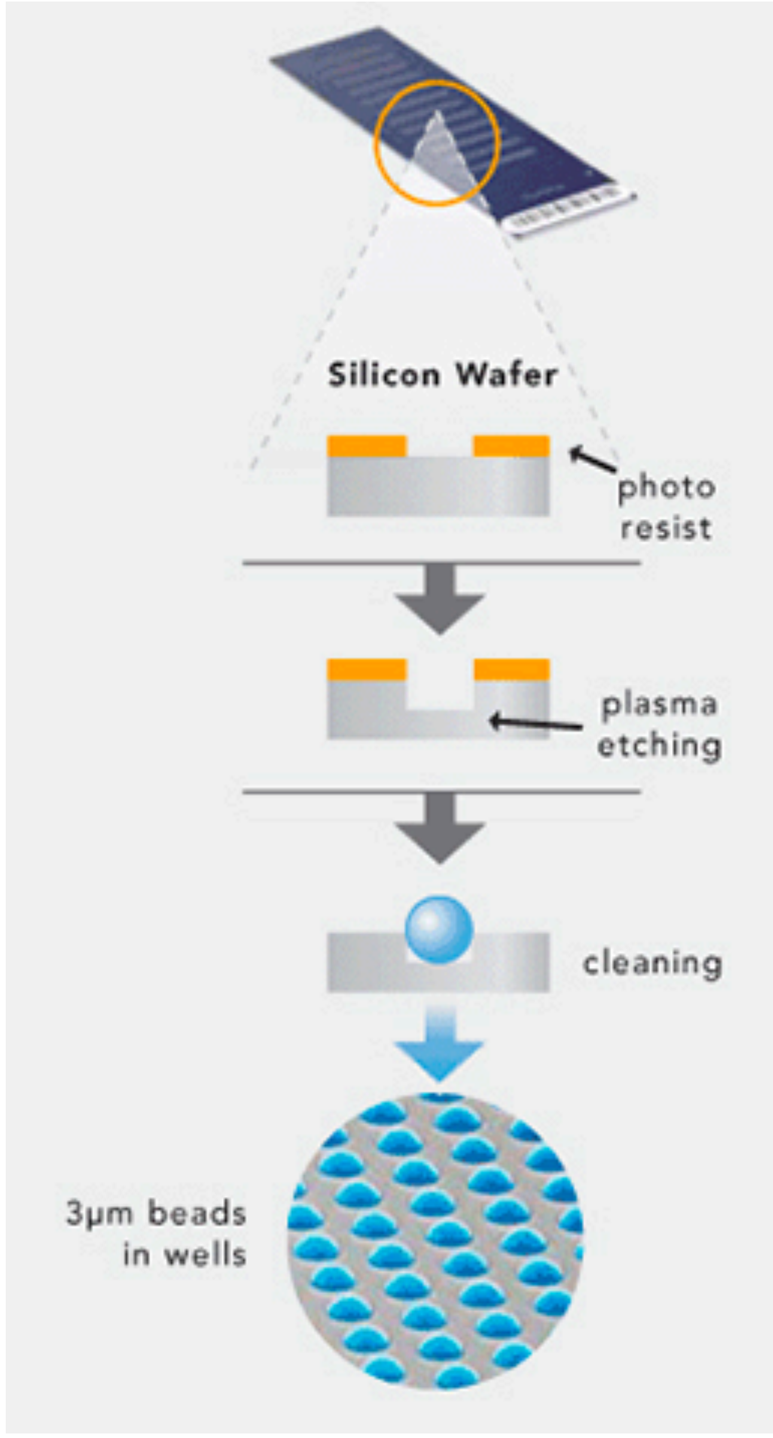
How do we call/quantify variants?



https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction

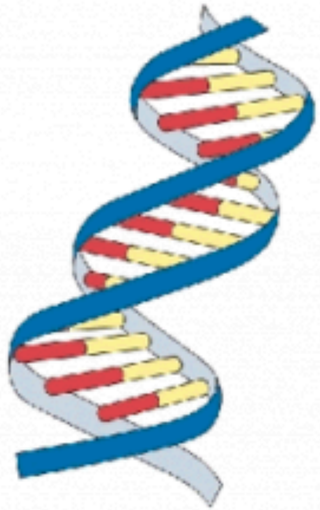
How do we measure variants?





https://www.illumina.com/science/technology/microarray.html

Genetic association studies: genotype vs. phenotype



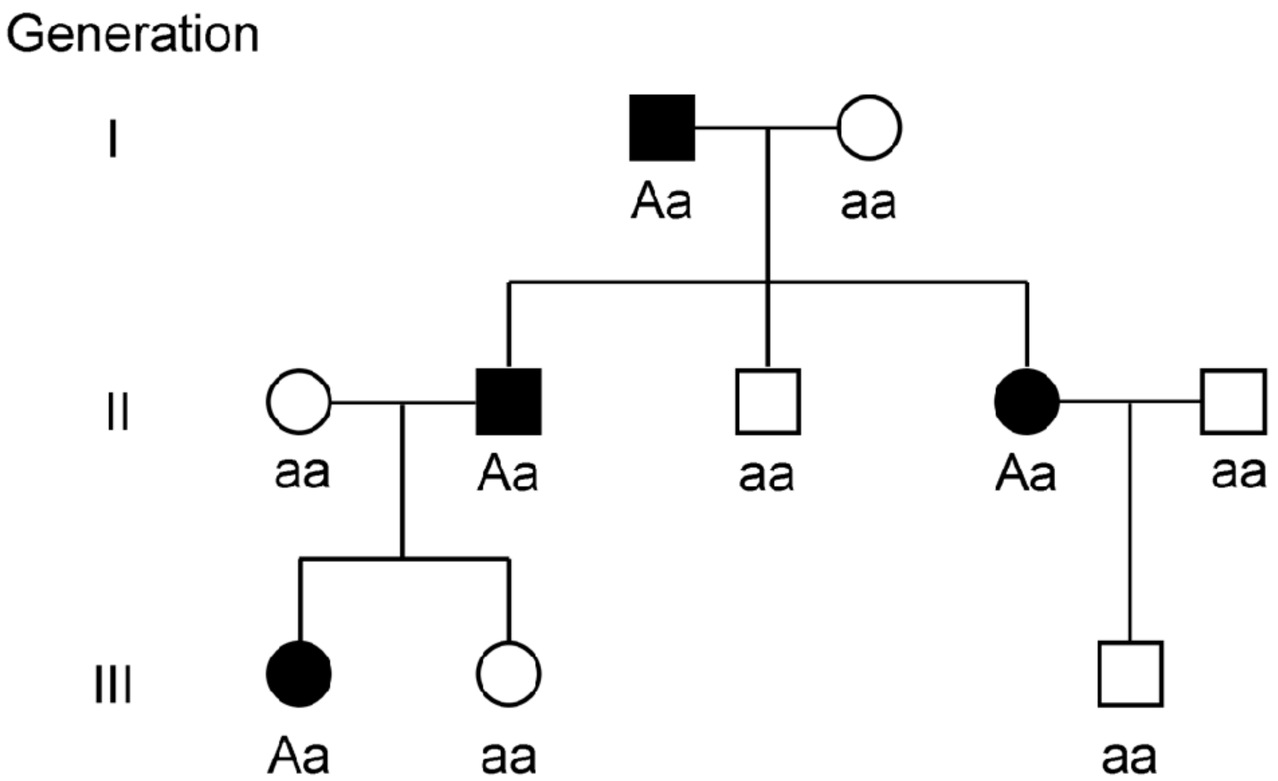
How do we test this?

Genotypes are the genetic make-up of an individual.

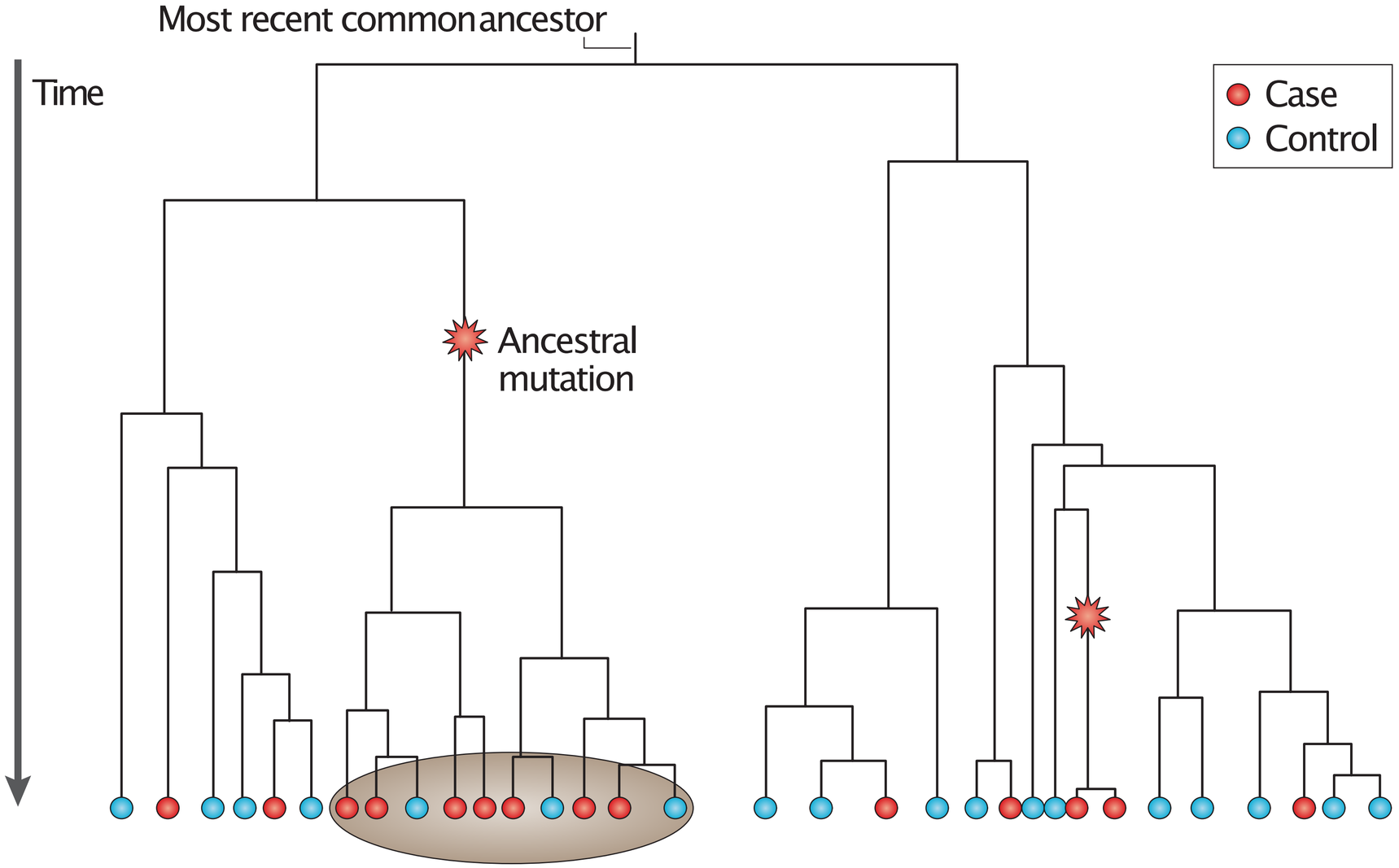
Phenotypes are the physical traits and characteristics of an individual and are influenced by their genotype and the environment.

https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction

How do we associate genetic variants to traits?

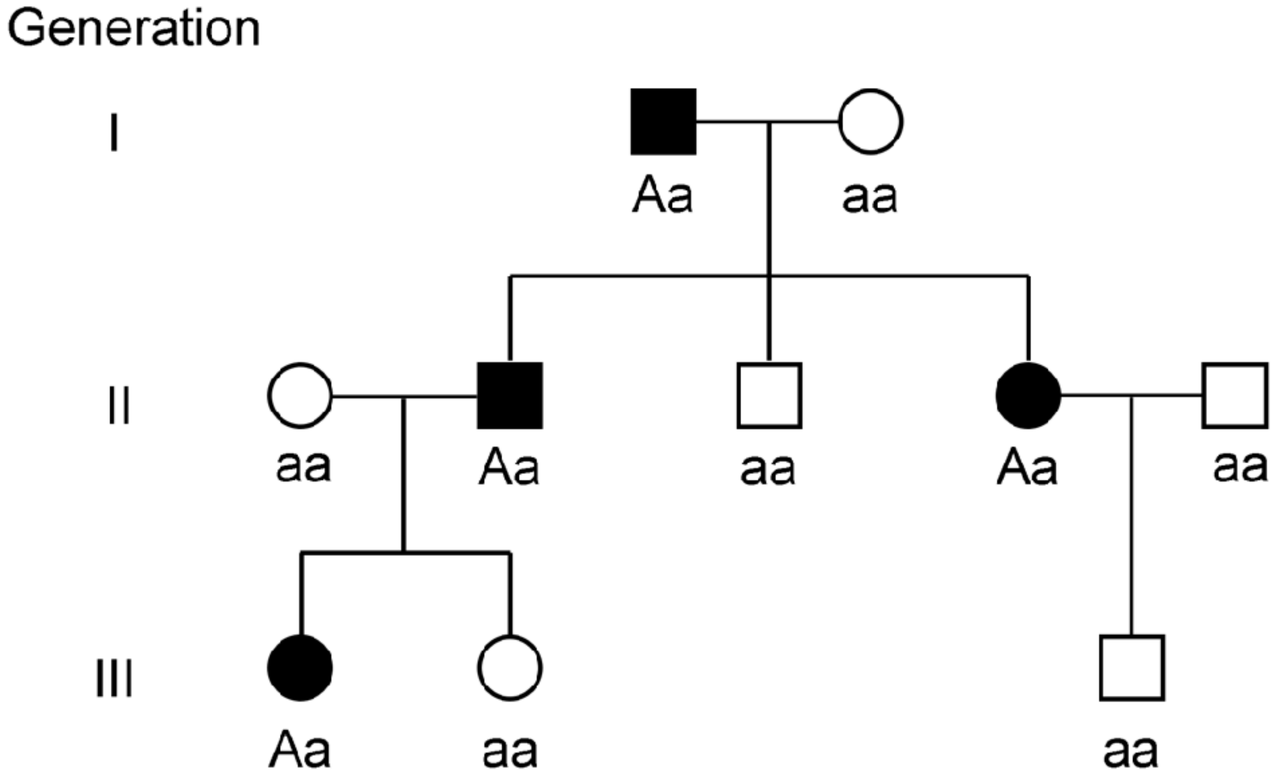


How can a genetic disease occur?



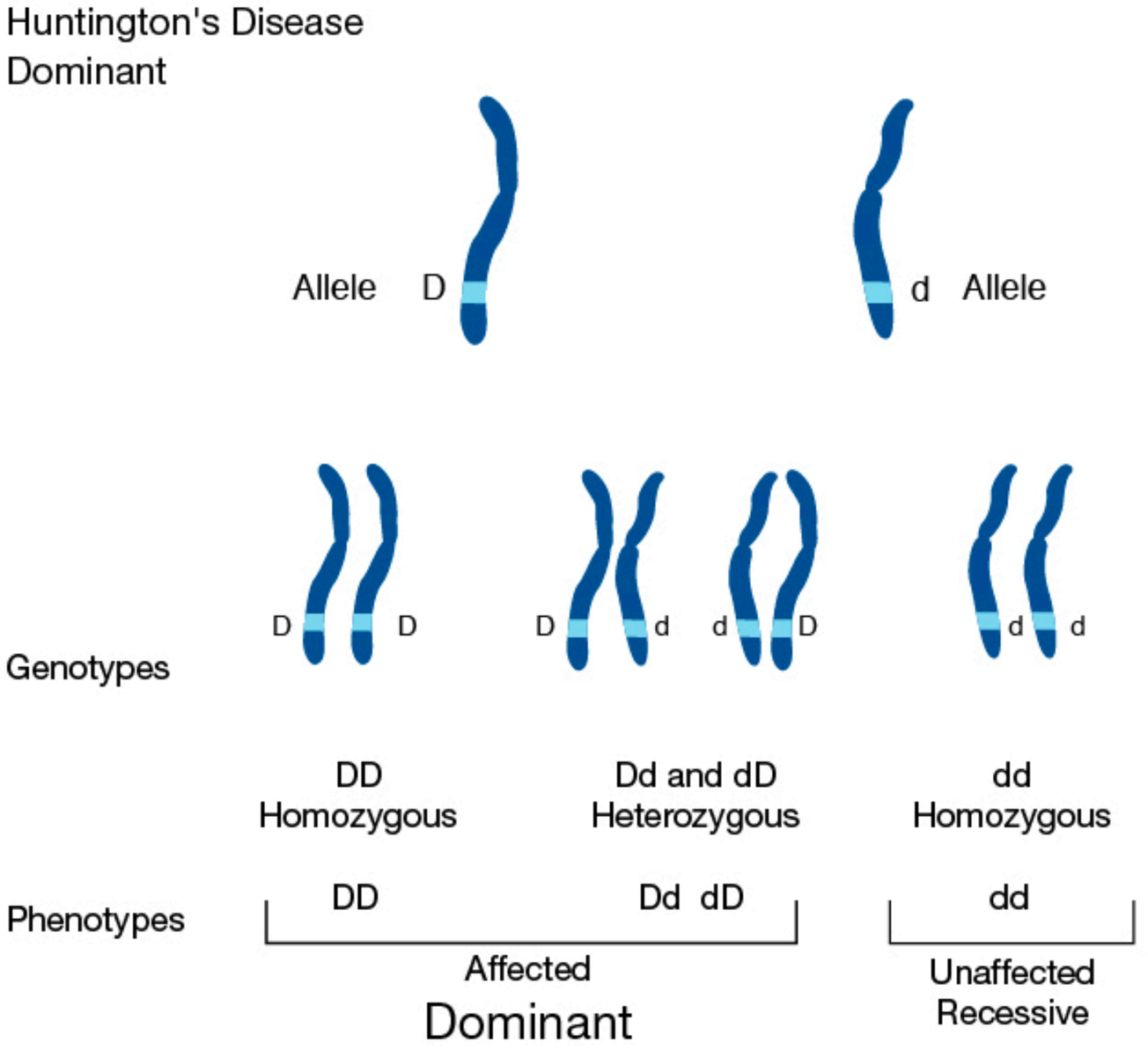
Balding, Nature Genetics Review, 2006

Mendelian disorders ≈ a single disease gene



- Cystic Fibrosis
- Sickle-cell anemia
- Phenylketonuria
- Huntington's disease
- ...

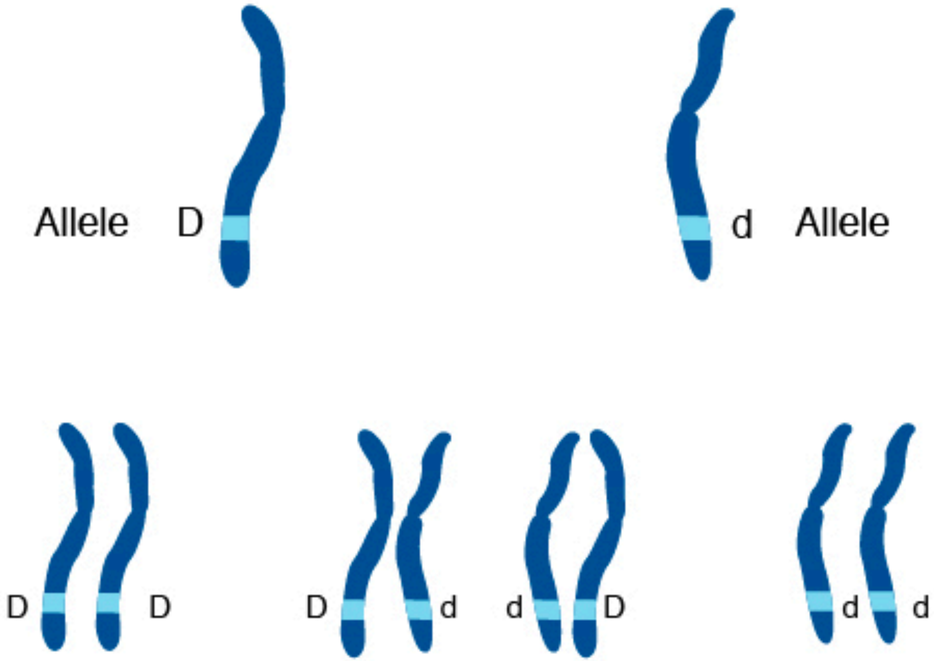
Very high penetrance, monogenic



https://www.genome.gov/genetics-glossary

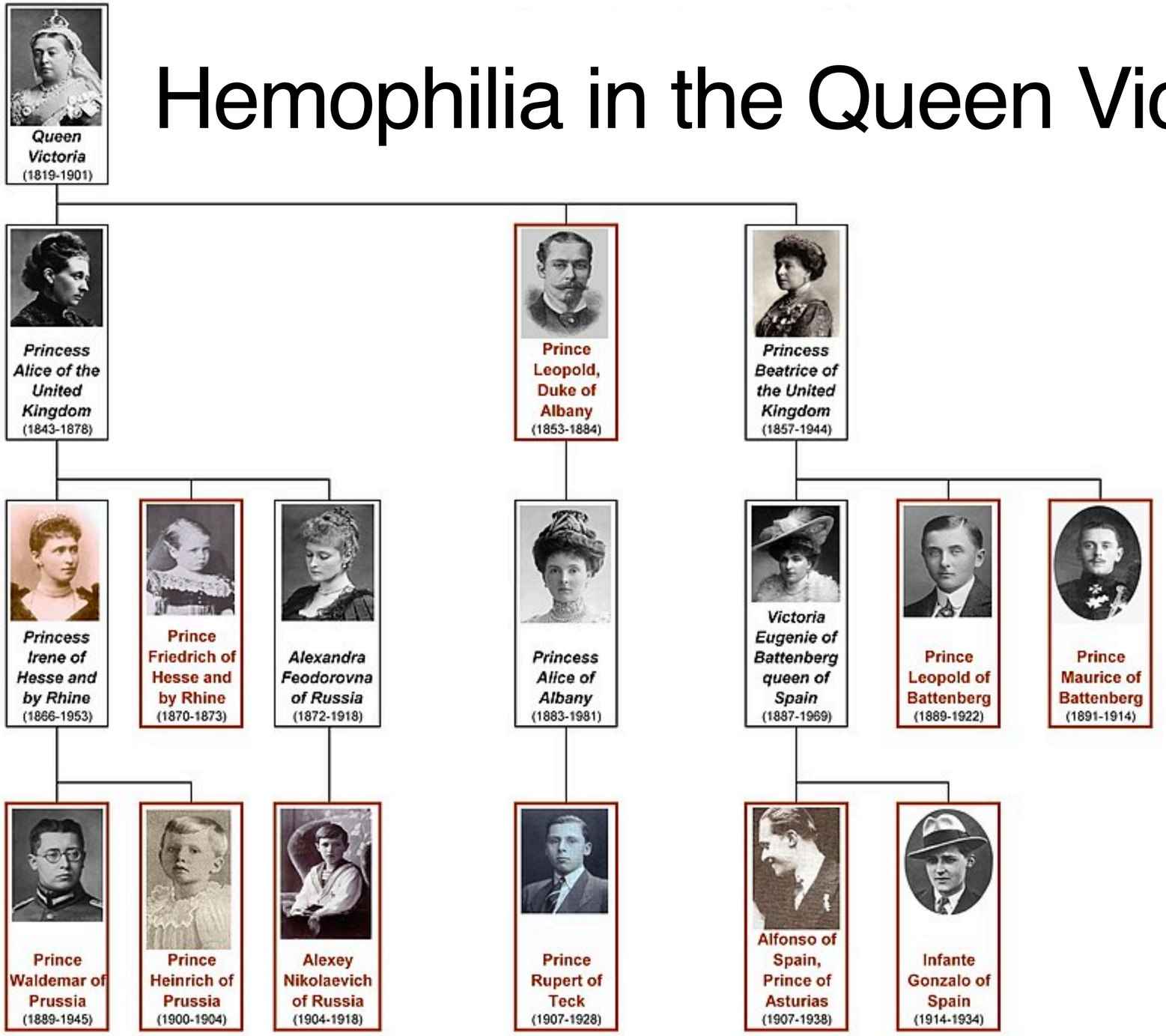
Sickle Cell Anemia or Cystic Fibrosis

Recessive



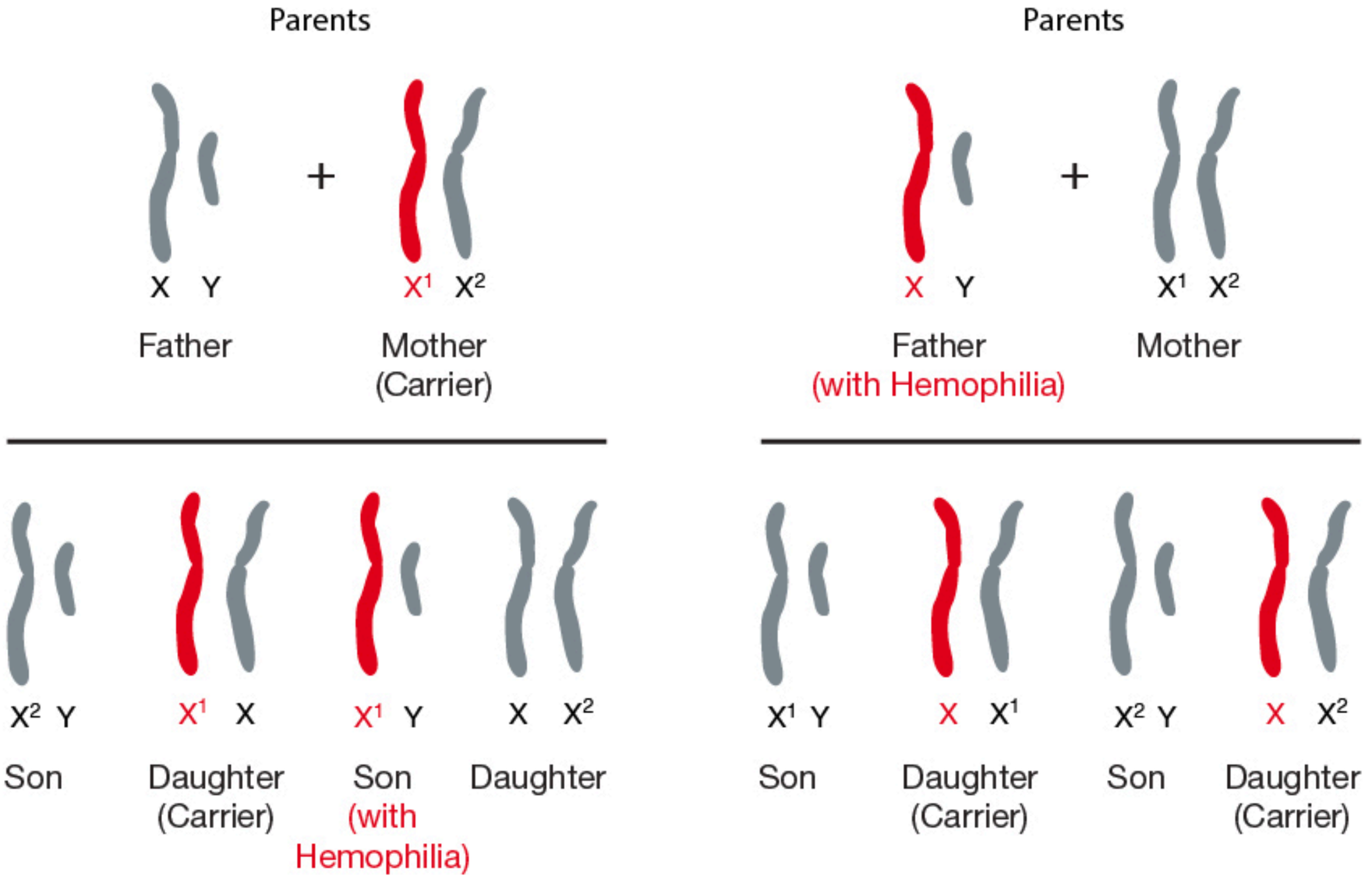
https://www.genome.gov/genetics-glossary

Hemophilia in the Queen Victoria's family



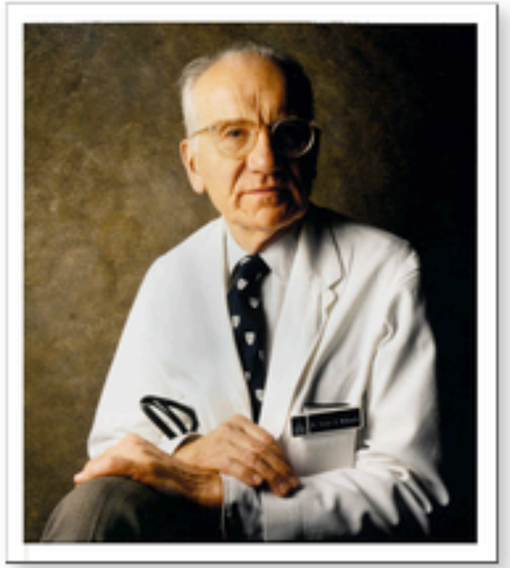
https://en.wikipedia.org/wiki/Haemophilia_in_European_royalty

Hemophilia



https://www.genome.gov/genetics-glossary

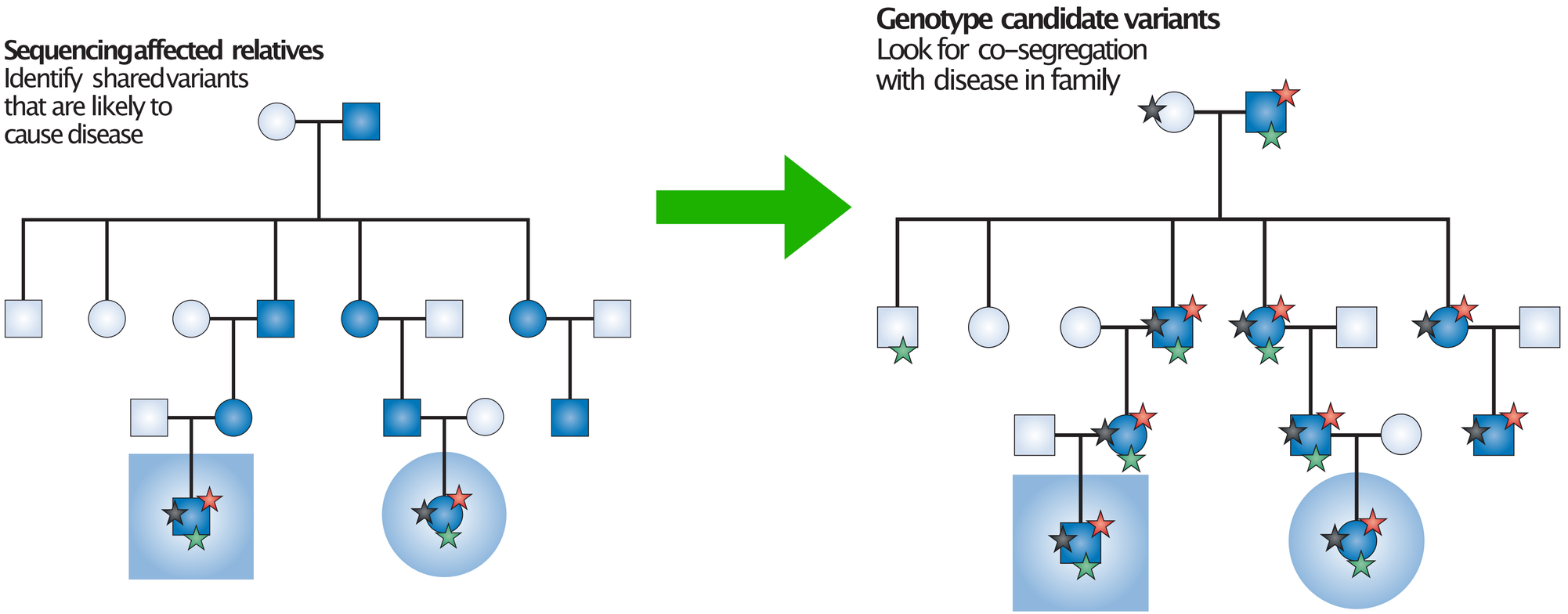
OMIM® – Online Mendelian Inheritance in Man®



Welcome to OMIM®, Online Mendelian Inheritance in Man®. OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 16,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources.

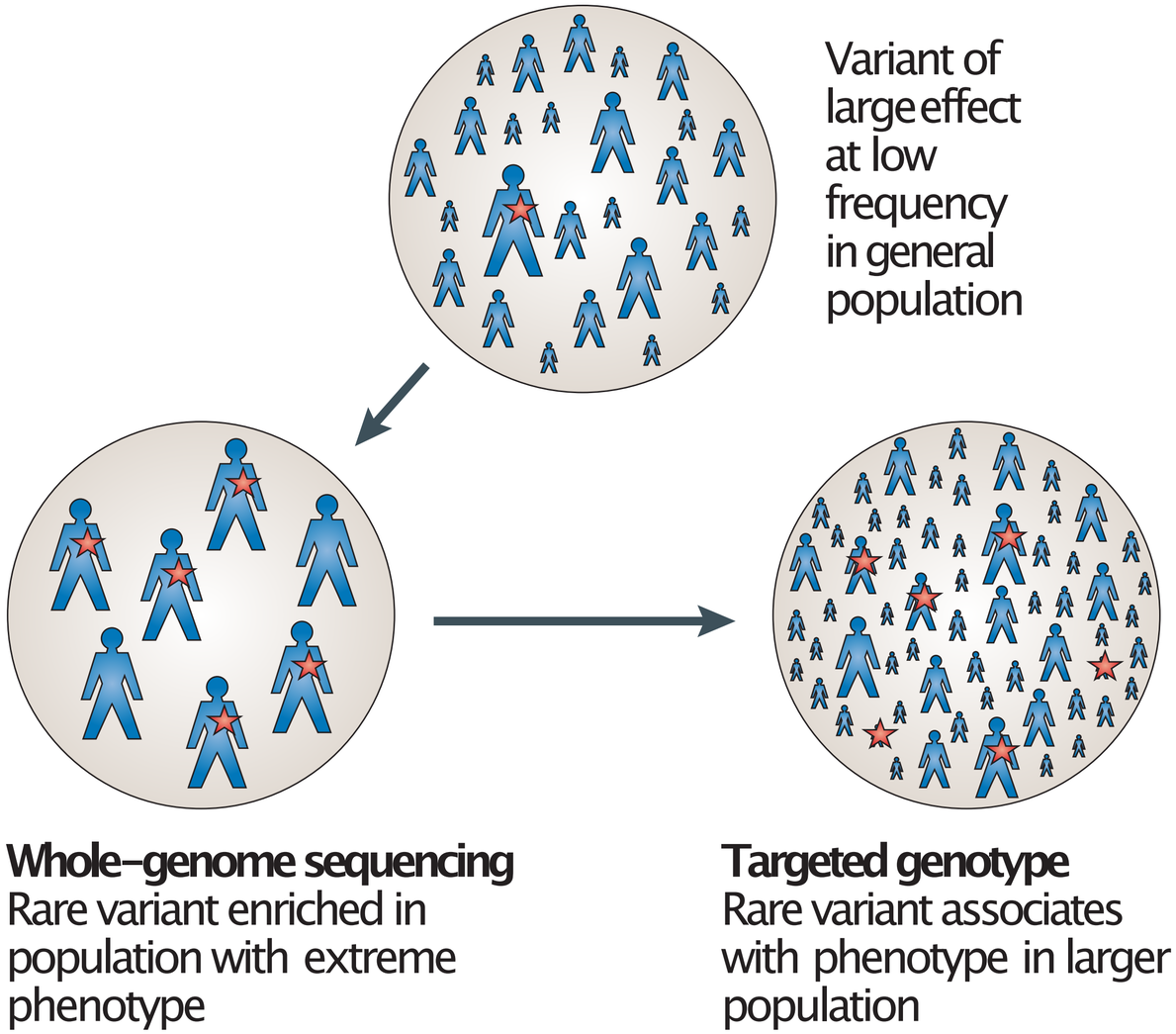
This database was initiated in the early 1960s by Dr. Victor A. McKusick as a catalog of mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM). Twelve book editions of MIM were published between 1966 and 1998. The online version, OMIM, was created in 1985 by a collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins. It was made generally available on the internet starting in 1987. In 1995, OMIM was developed for the World Wide Web by NCBI, the National Center for Biotechnology Information.

Targetted genotyping and sequencing



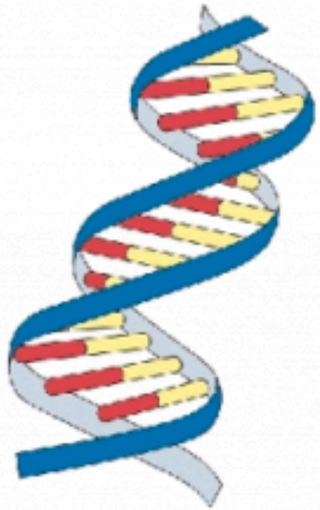
Cirulli and Goldstein, 2010

Population-level enrichment?



Cirulli and Goldstein, 2010

Can we identify variants associated with traits?

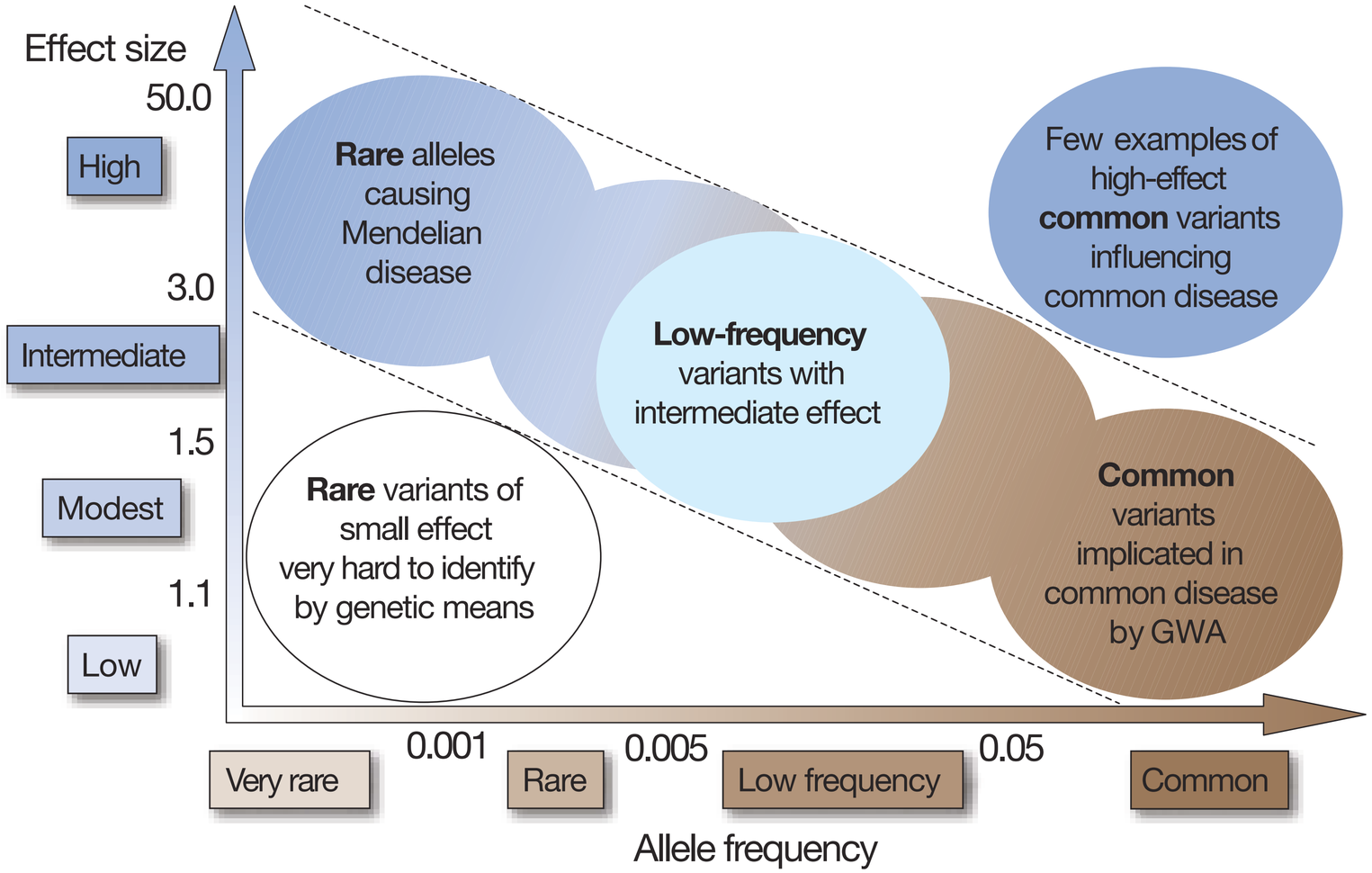


Genotypes are the genetic make-up of an individual.

Phenotypes are the physical traits and characteristics of an individual and are influenced by their genotype and the environment.

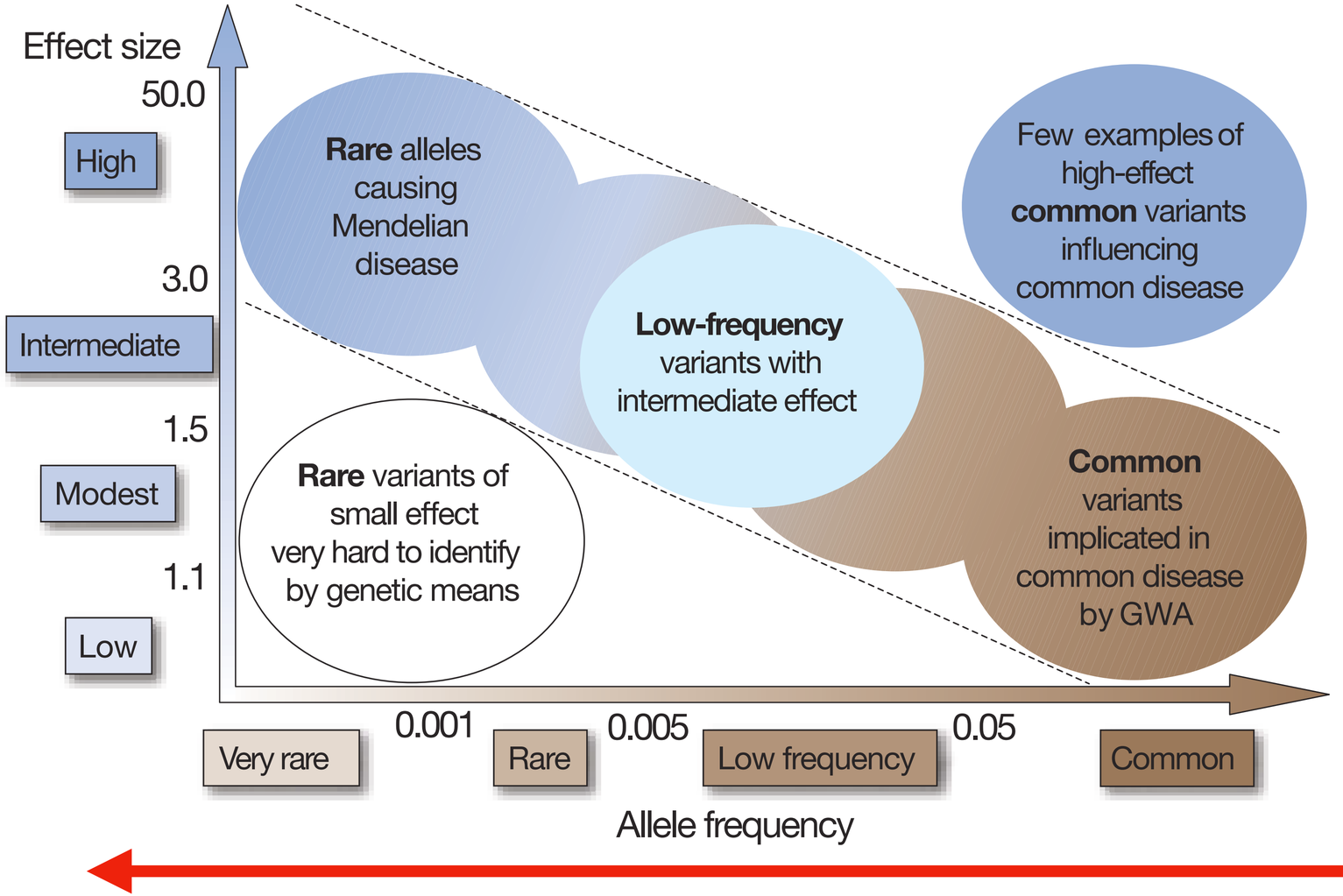
https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction

Will trait-associated variants emerge?



Manolio .. Vischer, *Nature* (2009)

Will population-level genetic studies be of worth?



Larger sample size

Larger sample size

Manolio .. Vischer, *Nature* (2009)

We need a large sample size!

Sample size matters (e.g., Schizophrenia GWAS)

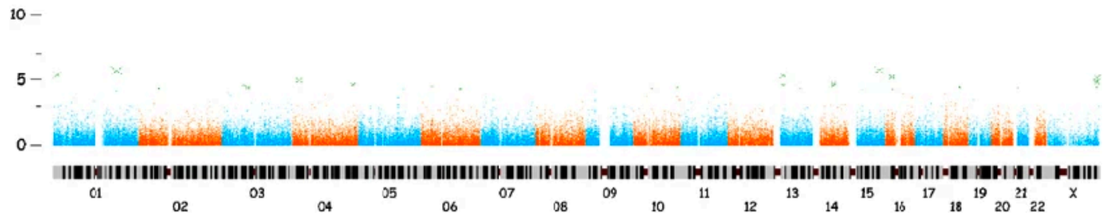
IMMEDIATE COMMUNICATION

Genomewide association for schizophrenia in the CATIE study: results of stage 1

PF Sullivan[1,2], D Lin[3], J-Y Tzeng[4], E van den Oord[5], D Perkins[6], TS Stroup[6], M Wagner[7], S Lee[3], FA Wright[3], F Zou[3], W Liu[8], AM Downing[9], J Lieberman[10] and SL Close[9]

N=733 cases vs. 733 controls

No variants found!

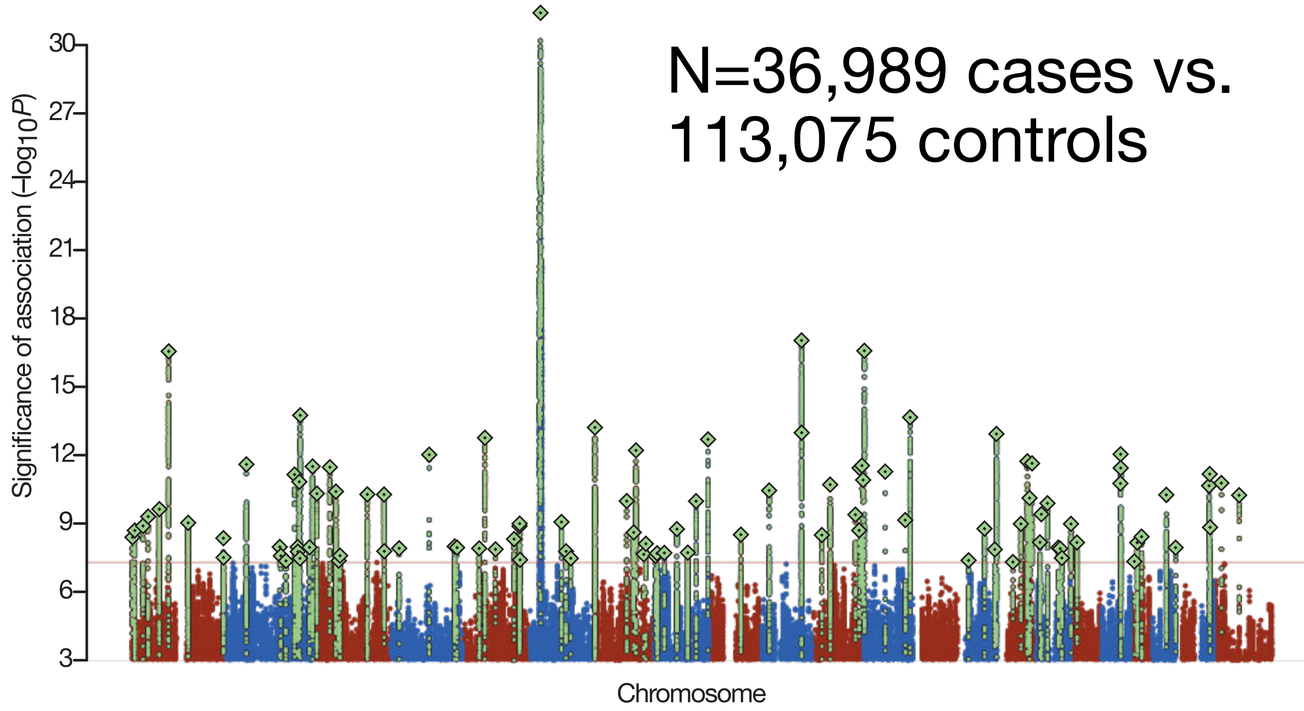


Sullivan *et al*. (2008)

Biological insights from 108 schizophrenia-associated genetic loci

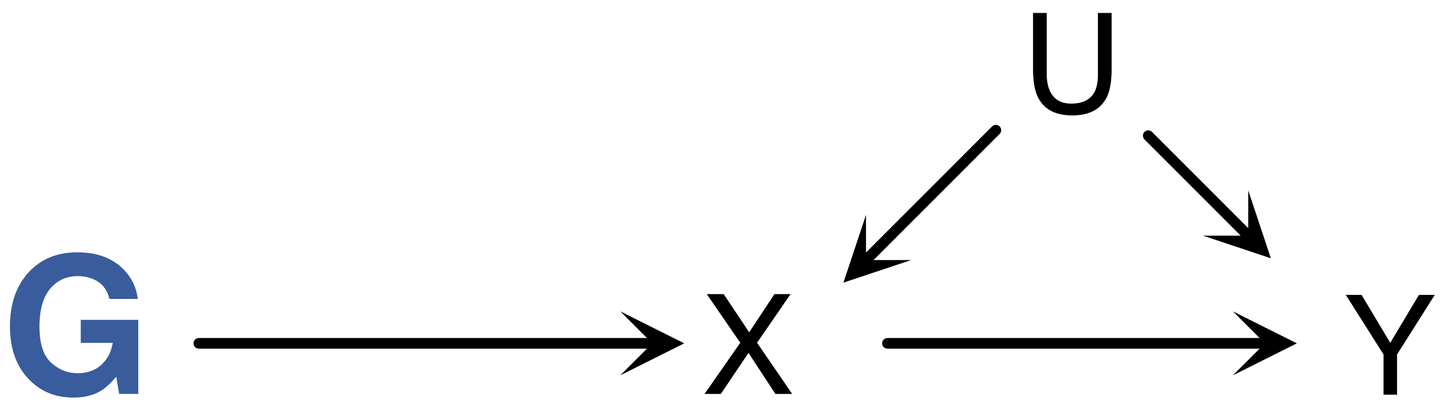
Schizophrenia Working Group of the Psychiatric Genomics Consortium*

N=36,989 cases vs. 113,075 controls



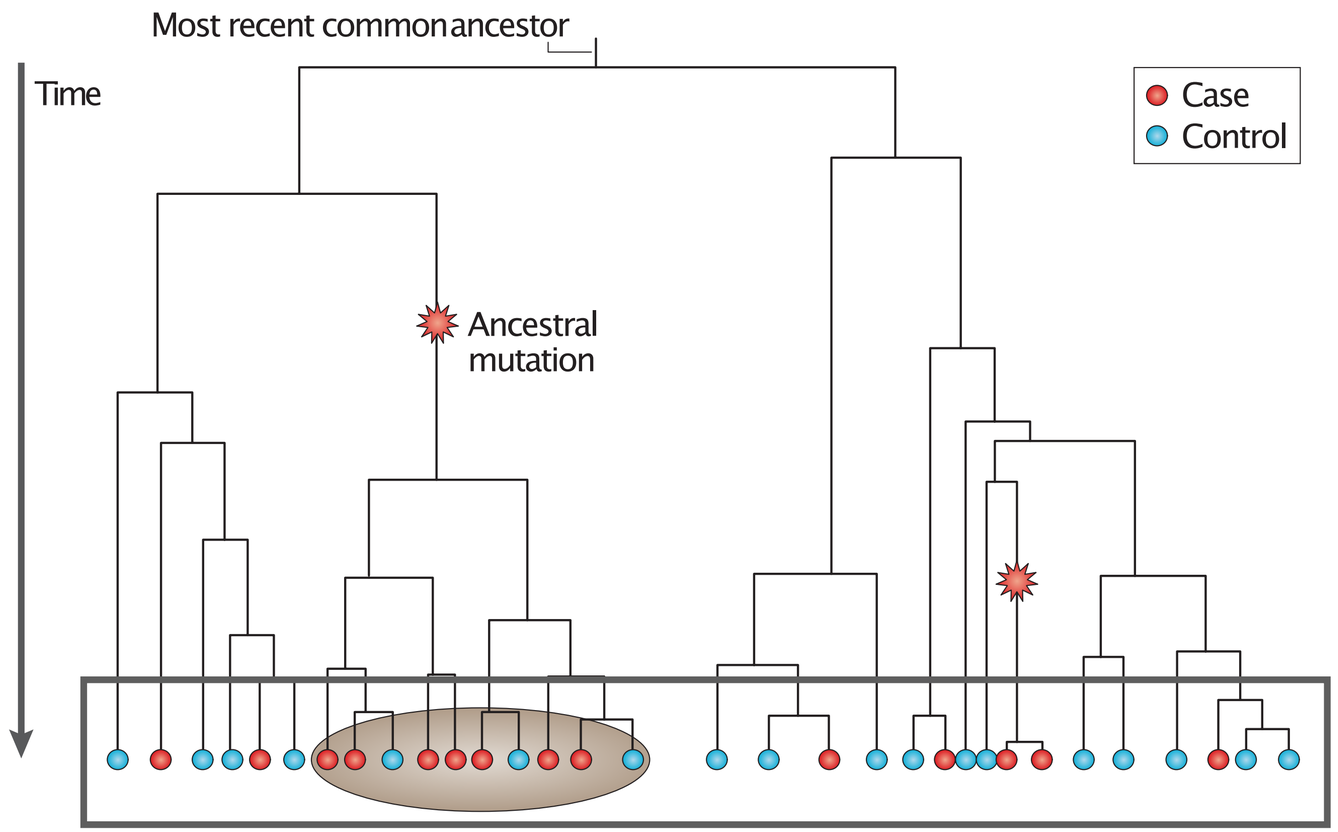
Schizophrenia Working Group of PGC, *Nature* (2014)

Genetic associations can lead to other discoveries



E.g., Mendelian Randomization (in the previous lecture)

Genetic association tests compare genotype vs. phenotype variation across individuals



Don't care about the history

Observed complex phenotype

vs.

Genotype at a certain locus

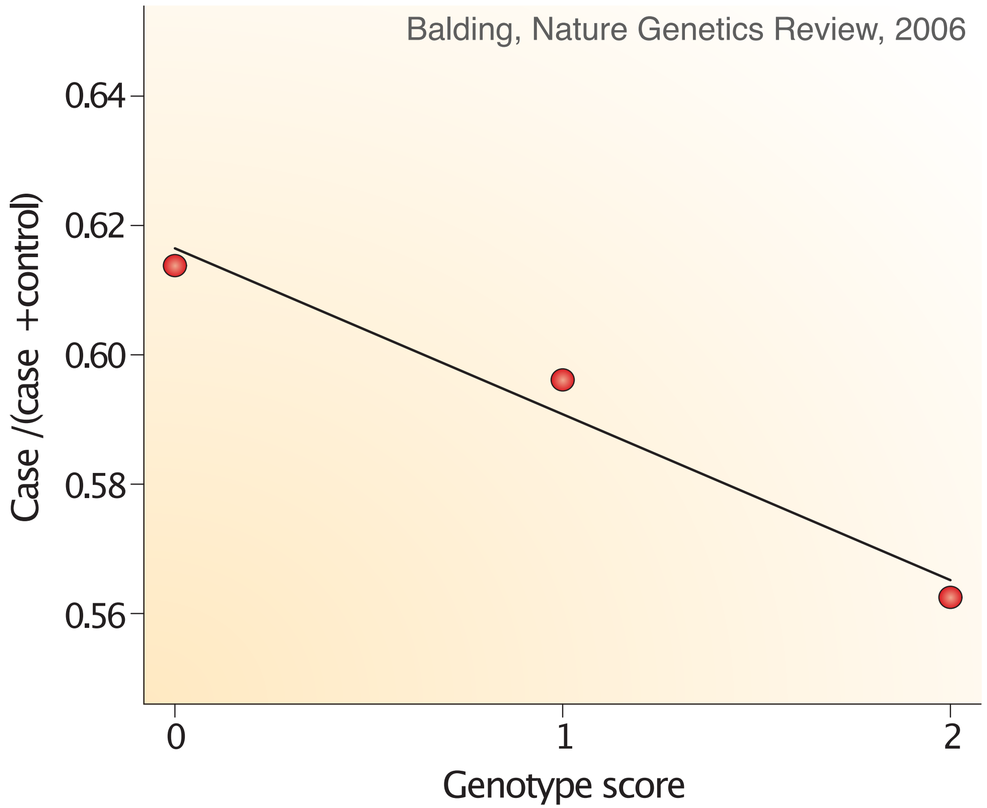
Balding, Nature Genetics Review, 2006

A typical genetics study design

Resistant
to a disease

More
susceptible
to a disease
(e.g., COVID)

T/T	0
C/C	**2**
T/T	0
T/**C**	**1**
T/T	0
T/T	0
T/T	0
T/T	0



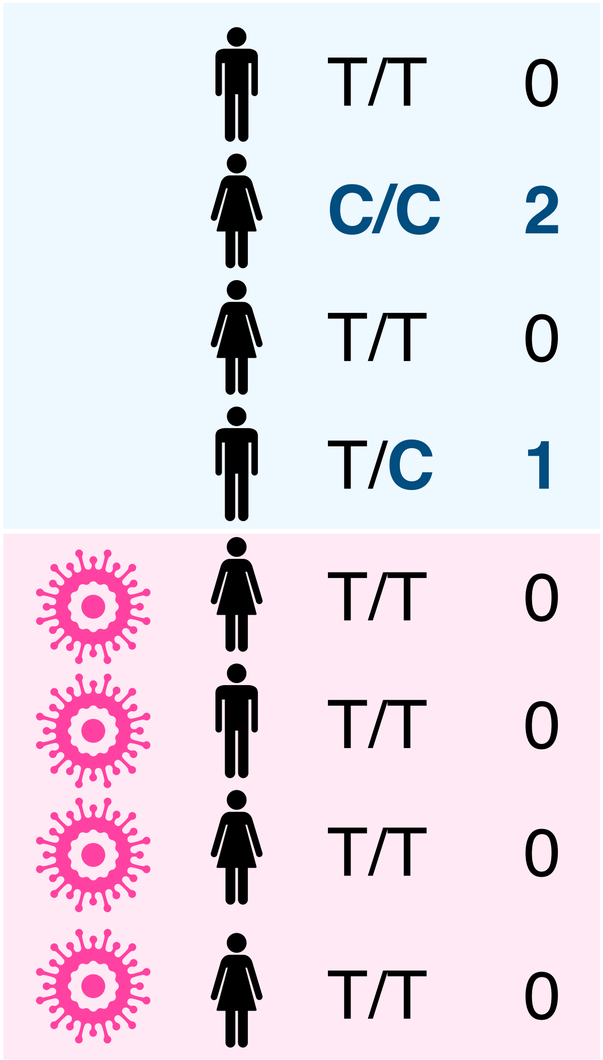
T = a major allele

C = a minor allele

A typical genetics study design

Resistant
to a disease

More
susceptible
to a disease
(e.g., COVID)



Genotype (dosage) of a variant *j*

$$X_{ij} \in \{0,1,2\}$$

$$Y_i \in \{0,1\}$$

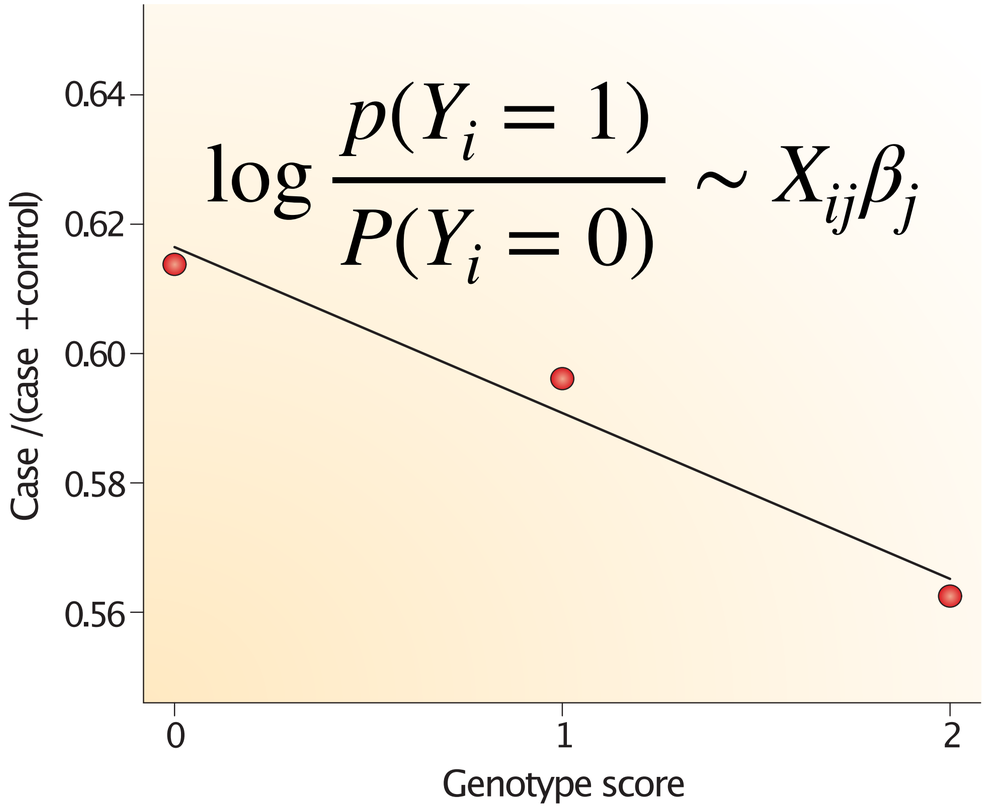
Genotype (dosage) of a variant *j*

$$\log \frac{p(Y_i = 1)}{P(Y_i = 0)} \sim X_{ij}\beta_j$$

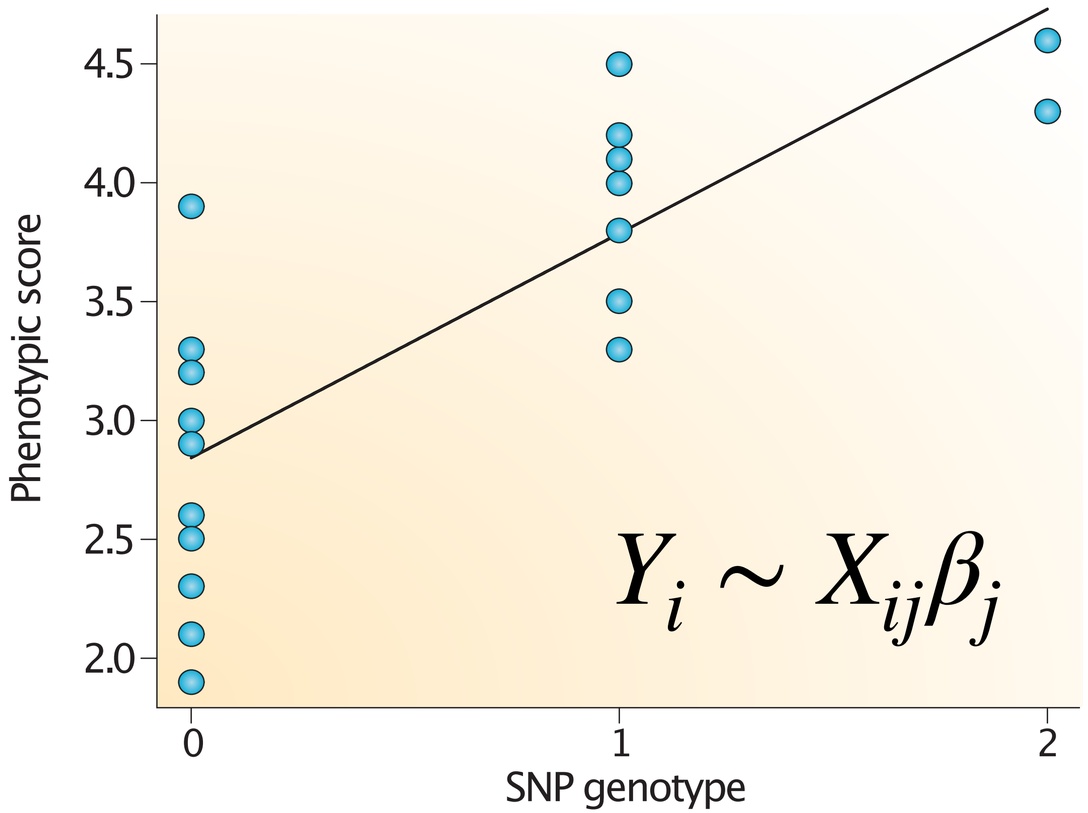
T = a major allele C = a minor allele

A traditional genetics study design

Case-control GWAS

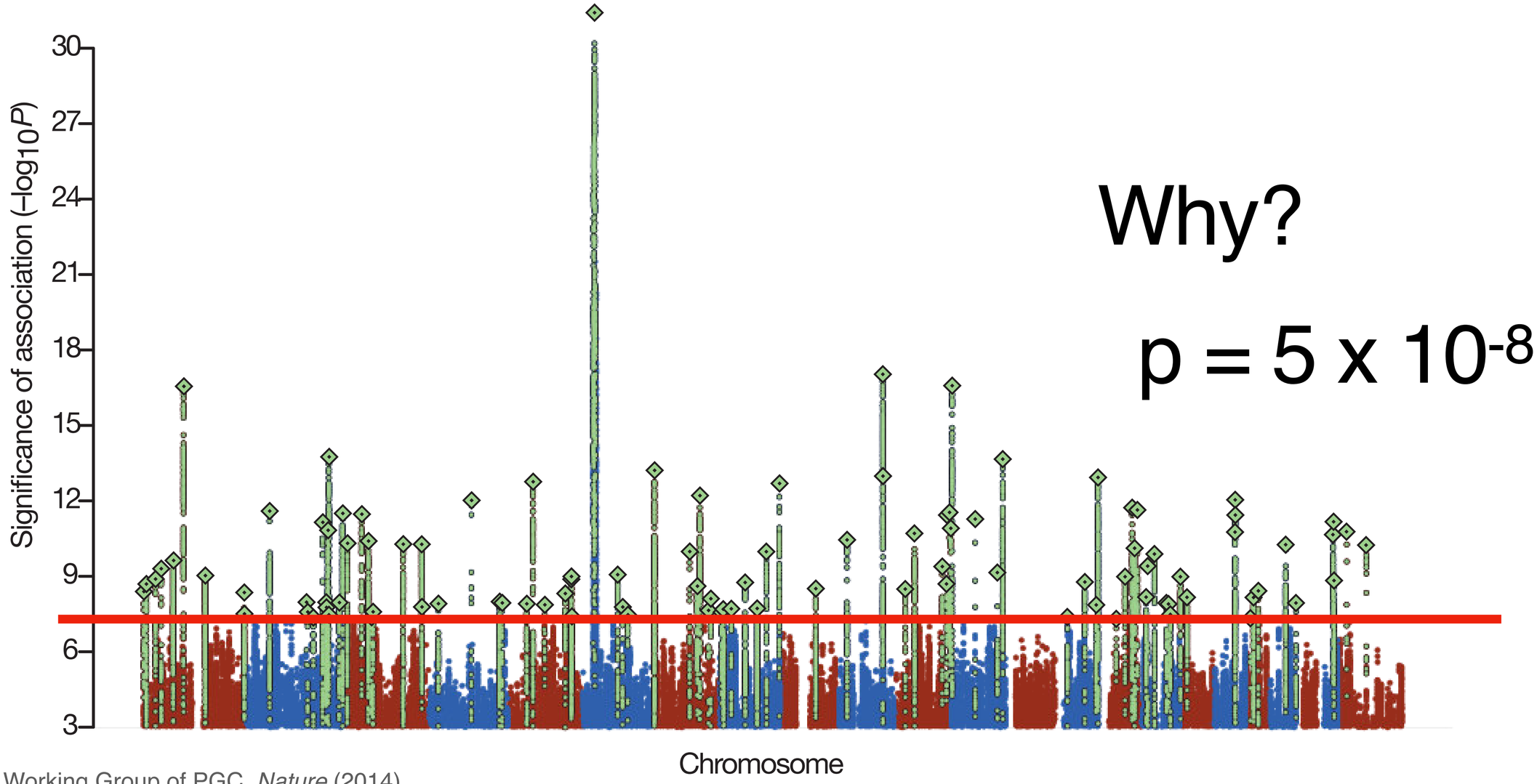


Quantitative trait GWAS



Balding, Nature Genetics Review, 2006

Genome-wide association study (many loci)



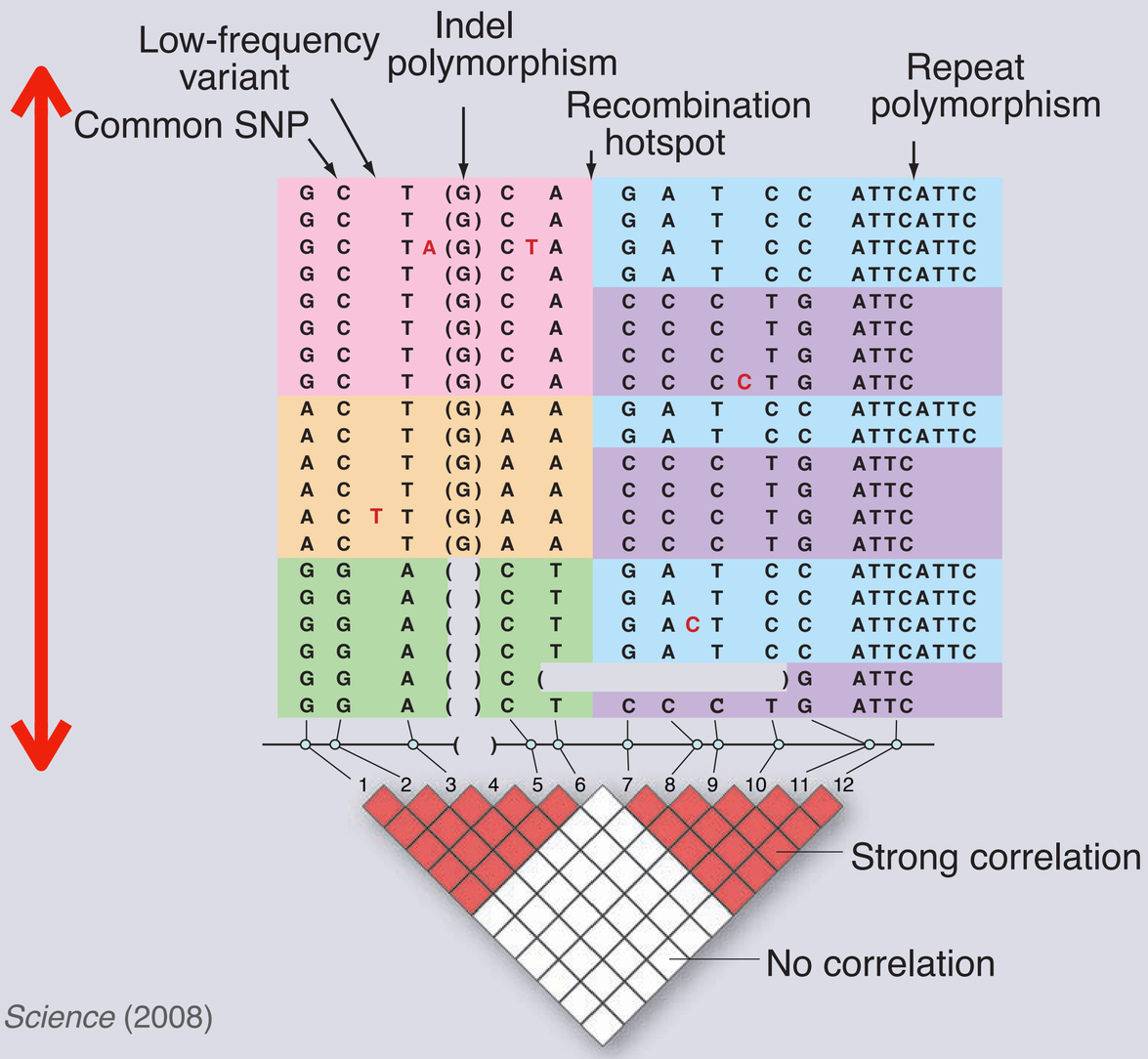
Why?

$p = 5 \times 10^{-8}$

Schizophrenia Working Group of PGC, *Nature* (2014)

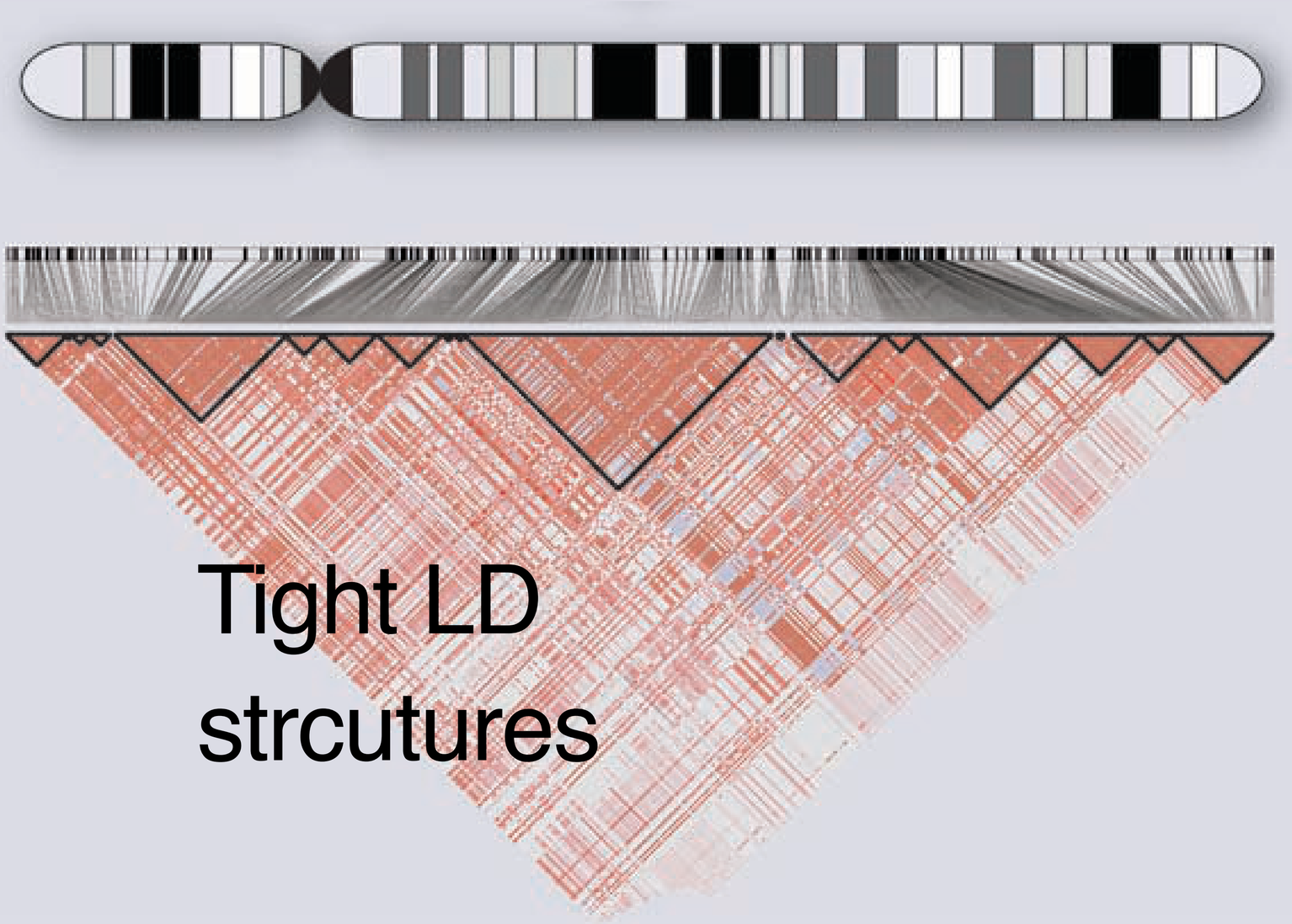
Genetic variation in one figure

across many individuals (diploid genomes)



Common SNPs
Insertion/deletion
Other low-freq. variants
Other structural variants
Recombination hotspot

Altshuler, Daly, Lander, *Science* (2008)



Tight LD
strcutures

1M "tagging" SNPs across the Human Genome

~ 1M independent loci (locations)

~ 5M variants

Genome-wide significance level

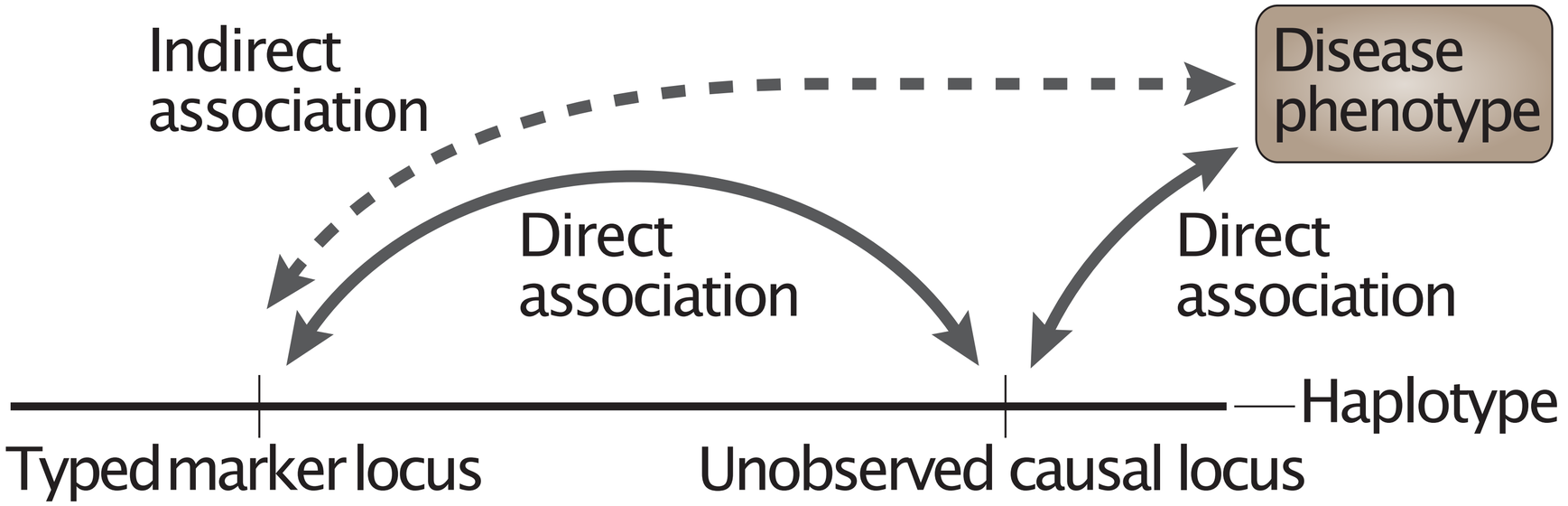
$$p_j < \frac{0.05}{\text{number of (independent) tests}} = \frac{0.05}{10^6} = 5 \times 10^{-8}$$

$$\text{FWER} = P\left(\cup_{j=1}^{10^6} \{p_j < 0.05/10^6\} \,|\, H_0 \right)$$

Union bound

$$< \sum_{j=1}^{10^6} P(p_j < 0.05/10^6 \,|\, H_0) = \sum_{j=1}^{10^6} \frac{0.05}{10^6} = 0.05$$

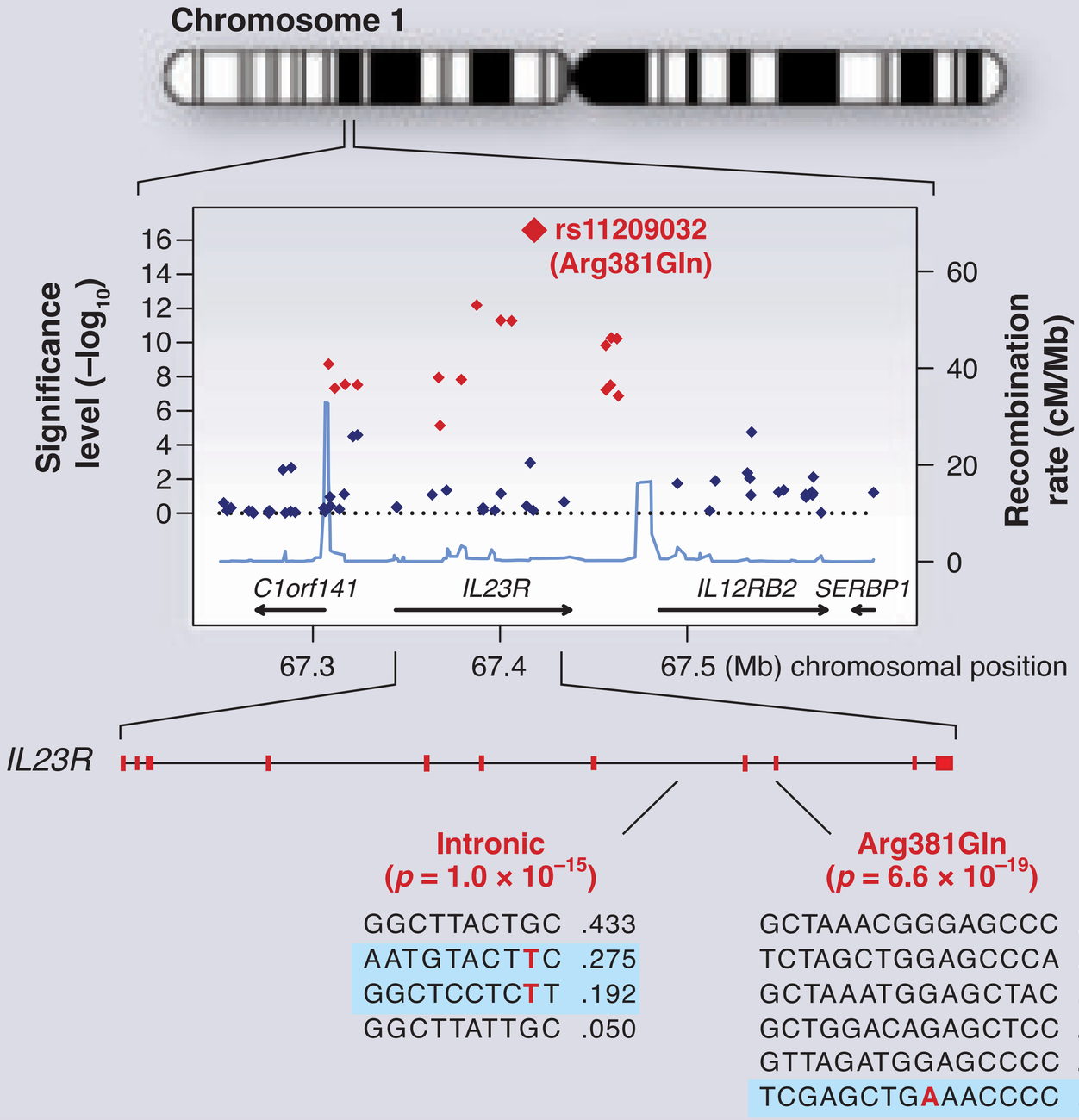
Types of association (typed ≈ tagging SNP)



GWAS only teach approximate locations of causal variants in the genome

Balding, Nature Genetics Review, 2006

A close-up view of Genome-wide-significant locuc

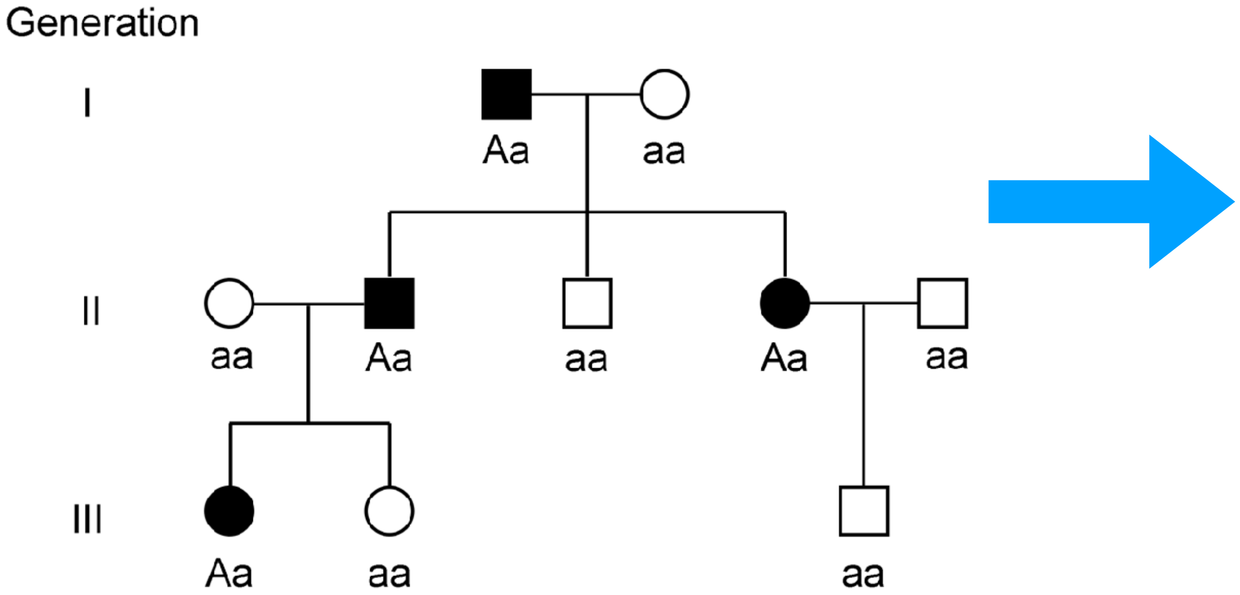


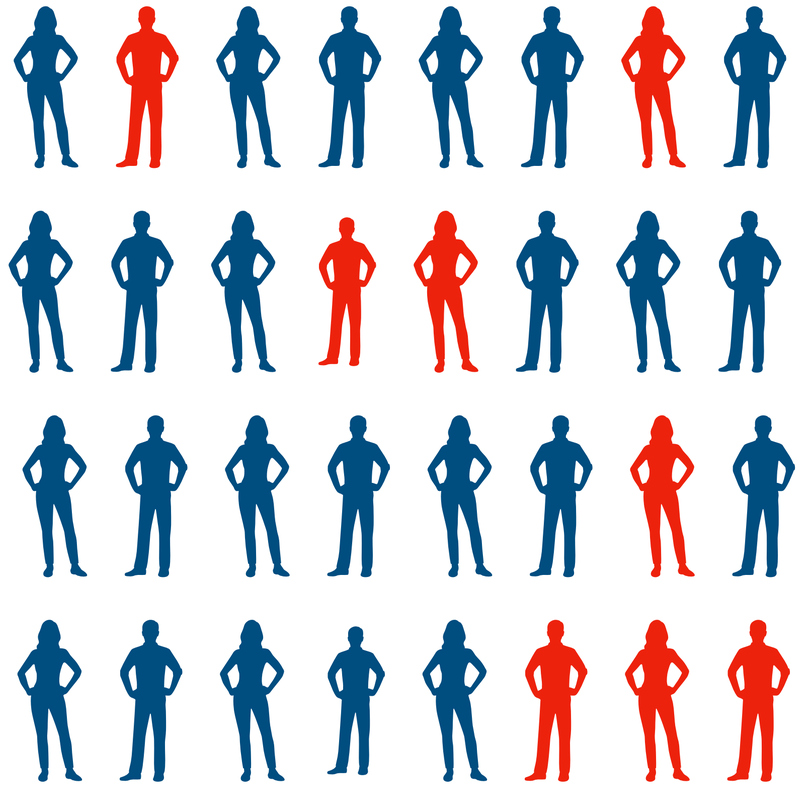
Altshuler, Daly, Lander, *Science* (2008)

Summary for GWAS

- Massive multiple hypothesis testing associations between genetic variants and phenotypic changes
- Most studies focus on biallelic variants, coding them as {0, 1, 2}
- Most association statistics concern an additive effect of genotypes (linearity)
- Genetic variants closely located in the genome are strongly correlated with each other due to recombination
- Implicitly assume ~ 1M genetic variants (of tagging, or representative SNPs)
- GWAS only teach an approximate genomic locations associated with phenotypes

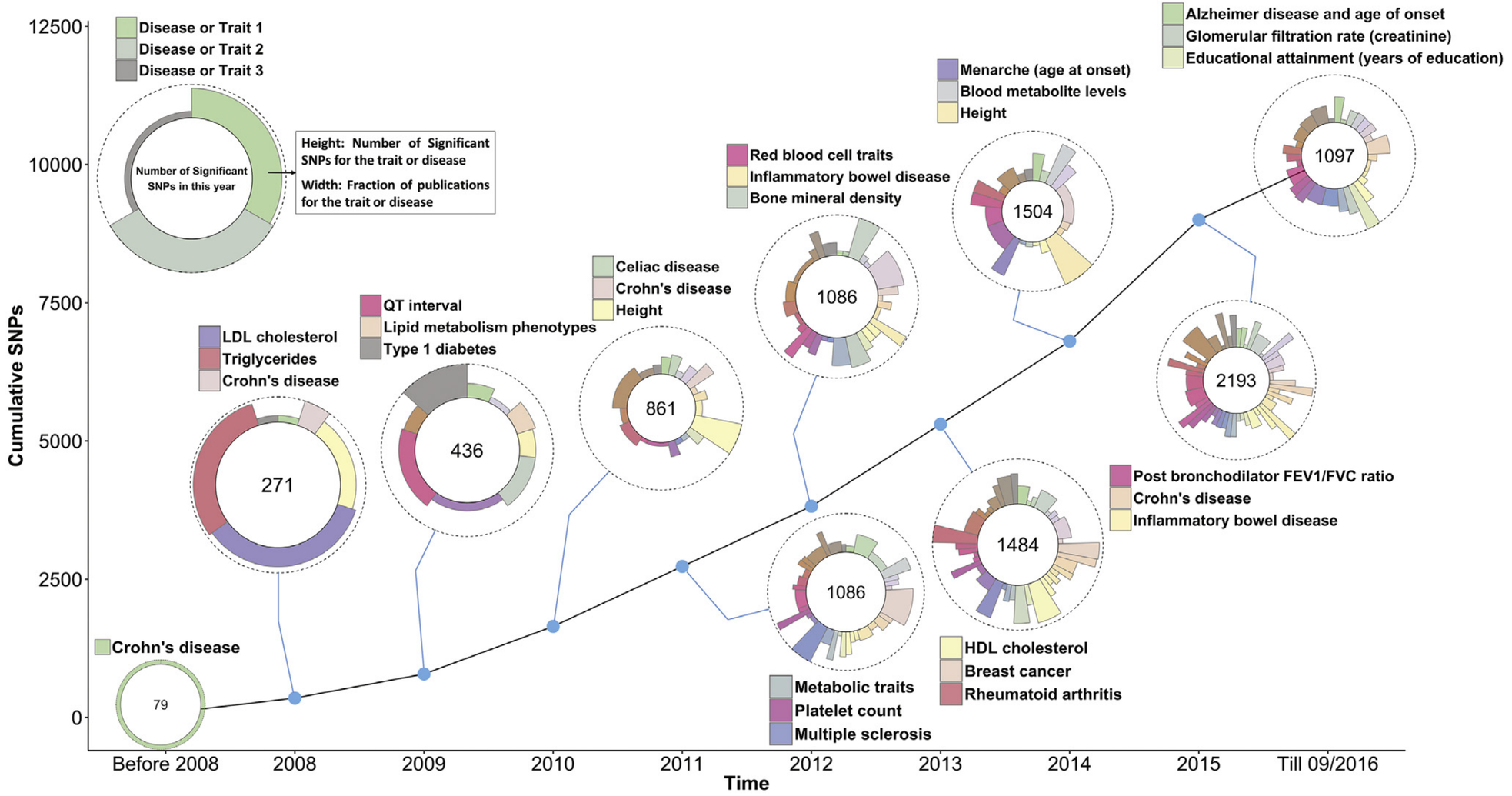
The success of GWAS changed the paradigm of human genetics studies





Are we assuming a homogeneous population?

GWAS data/results increased rapidly



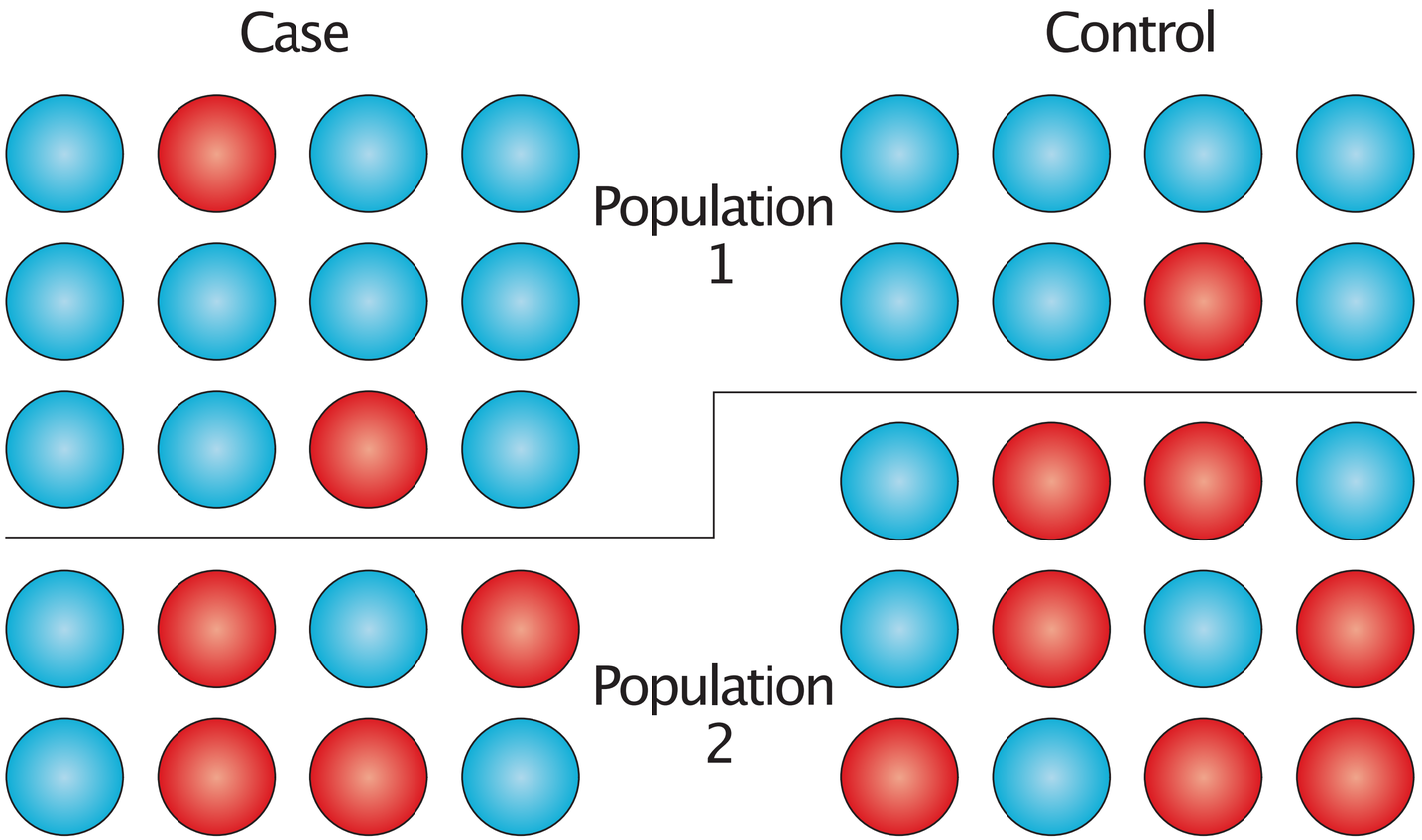
Vischer *et al.* (2017)

Big biobank-driven GWAS on prospective cohort

Today's lecture: GWAS and related topics

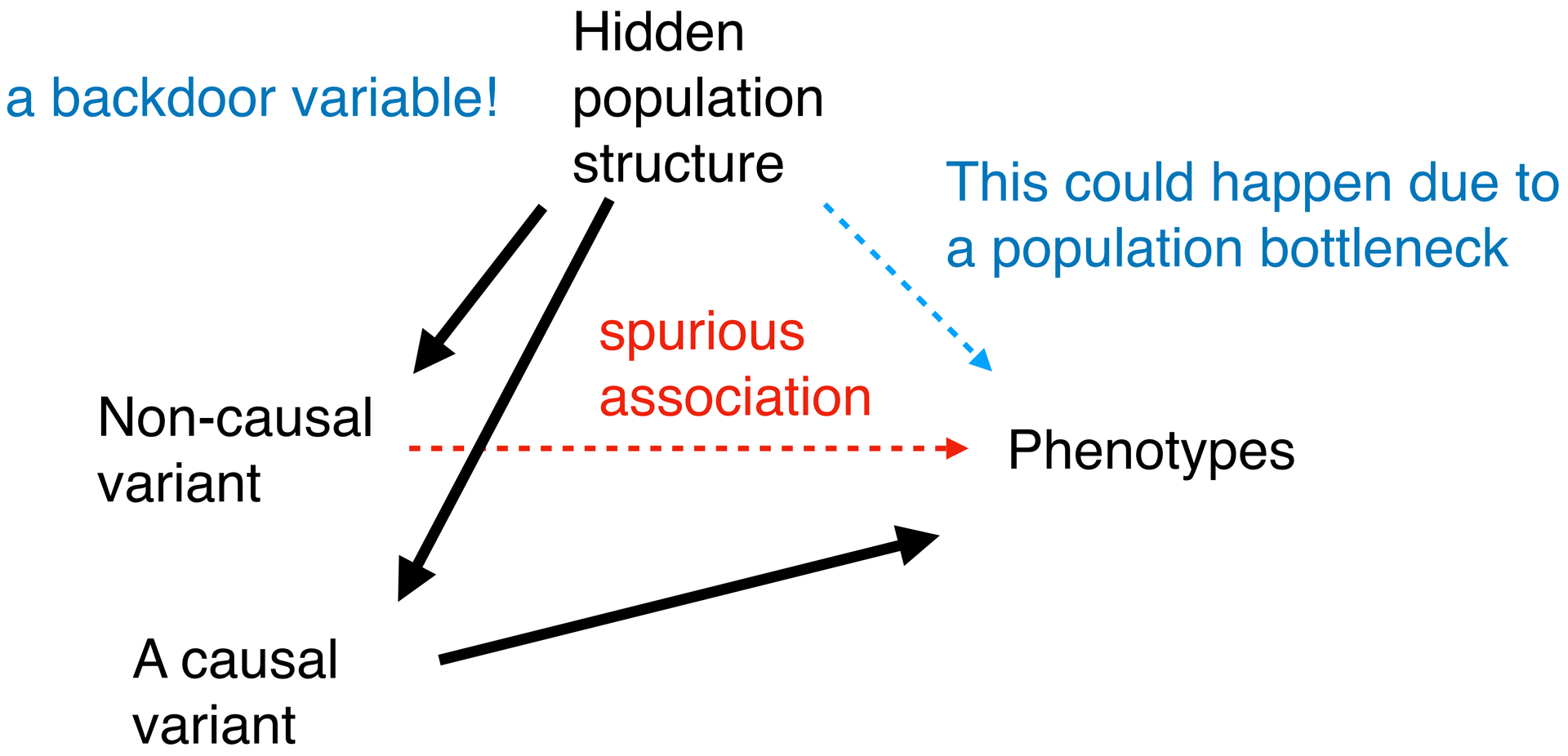
- **Human Genetics 101**
  - Variation in the human genome
  - How do we measure genetic associations?
- **Polygenic models**
  - Population structures
  - Linear mixed effect model
- **Systems Genetics**
  - Summary-based GWAS analysis
  - LD-score regression: "enrichment analysis" in GWAS

What if there were a hidden population structure?



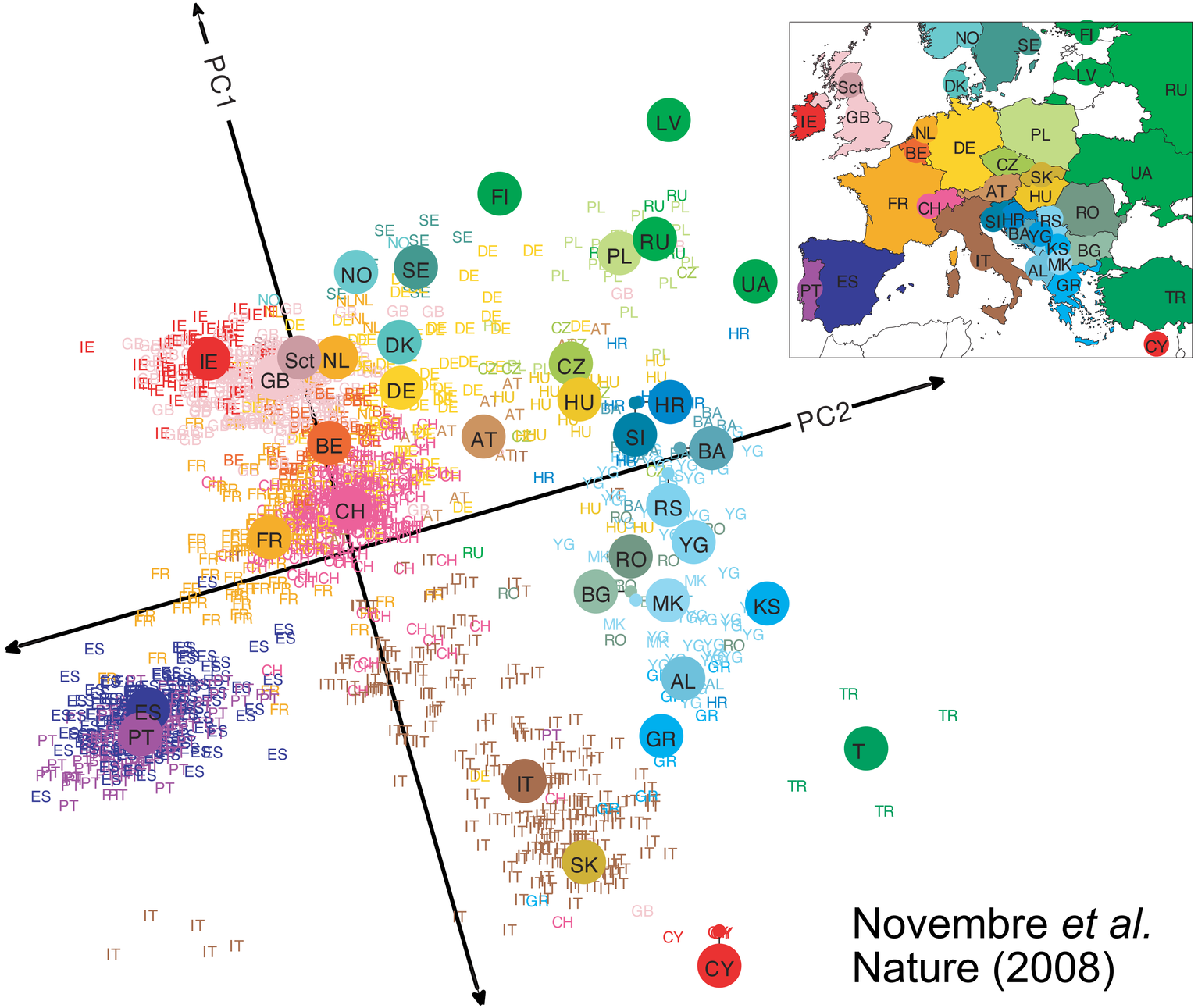
Balding, Nature Genetics Review, 2006

Hidden population structure in a causal graph



The human population as a result of the human migration history

https://www.nytimes.com/2018/03/20/science/david-reich-human-migrations.html



Novembre *et al.*
Nature (2008)

Much of human genetics problems centre on two covariance matrices

For a standardized $n \times p$ genotype matrix X (n: #individuals, p: #SNPs),

1. Genetic relatedness matrix (GRM)

individual by individual, $n \times n$ matrix

$$K \approx XX^\top/n$$

The matrix K captures population structure/correlation across different individuals.

- Kinship matrix; population admixture
- Human migration history

2. Linkage disequilibrium (LD)

SNP by SNP, $p \times p$ matrix

$$R \approx X^\top X/n$$

The matrix R captures localized correlation patterns along the genomic axis within a chromosome.

- LD matrix
- The results of many, many recombination events

1000 Genomes project data

1KG contains whole genome sequencing data of 2,490 individuals sampled from 26 groups based on the origins and geographical locations (as of 2013 phase3).

0	1	0	1	1	1	1	0	0	0
0	2	0	2	2	2	2	0	0	0
0	1	0	1	1	1	1	0	0	0
0	2	0	2	2	2	2	0	0	1
0	2	0	2	2	2	2	0	1	0
0	2	0	2	2	2	2	0	0	1
0	2	0	2	2	2	2	0	0	1
0	2	0	2	2	2	2	0	0	1
0	2	0	2	2	2	2	0	0	1
0	1	0	1	1	1	1	0	0	1

First 10 individuals and 10 variants

SVD captures principal components

$$X = UDV^\top$$

What is this?

$$\frac{1}{n}X^\top X = \frac{1}{n}VDU^\top UDV^\top = \frac{1}{n}VD^2V^\top$$

variant x variant

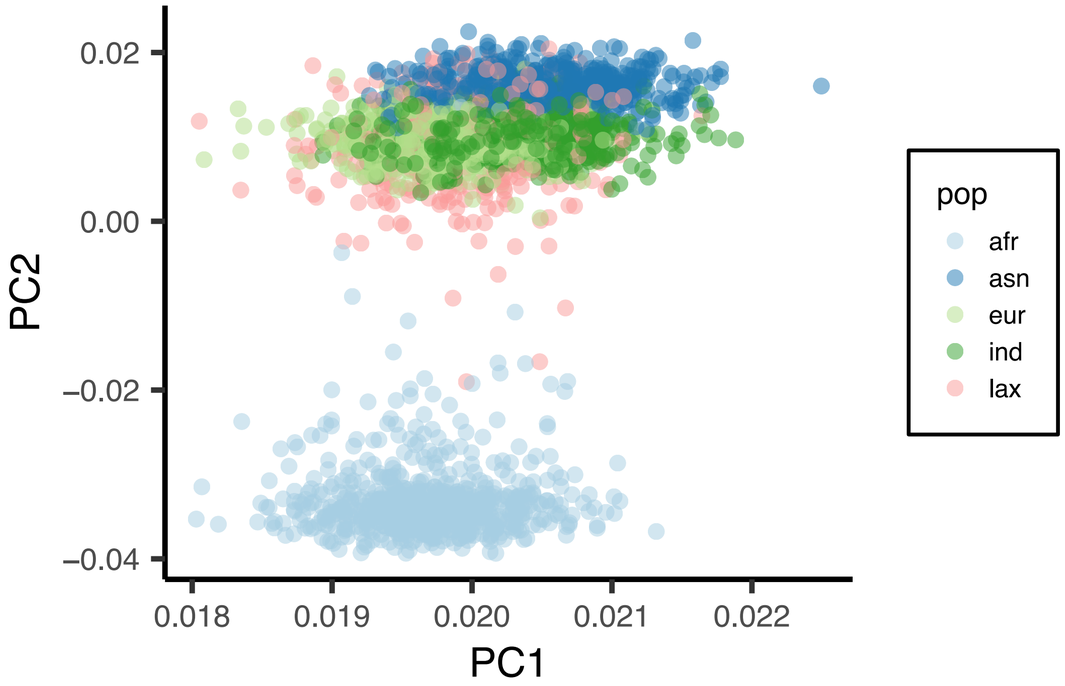
What is this?

$$\frac{1}{n}XX^\top = \frac{1}{n}UDV^\top VDU^\top = \frac{1}{n}UD^2U^\top$$

sample x sample

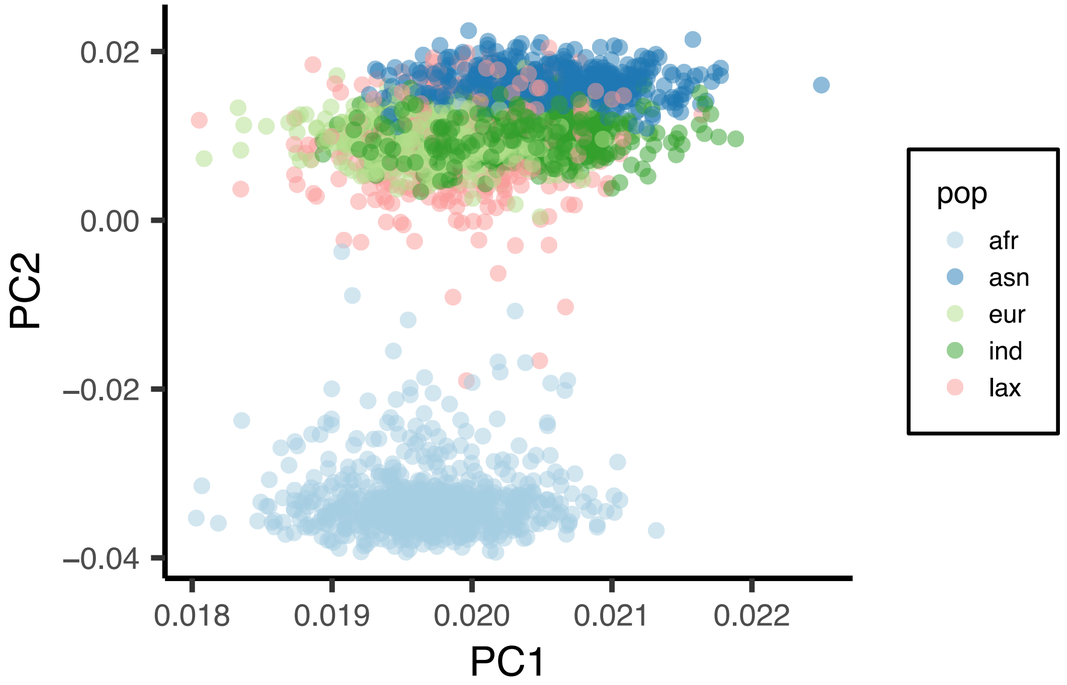
Let's take top 1000 most frequent variants

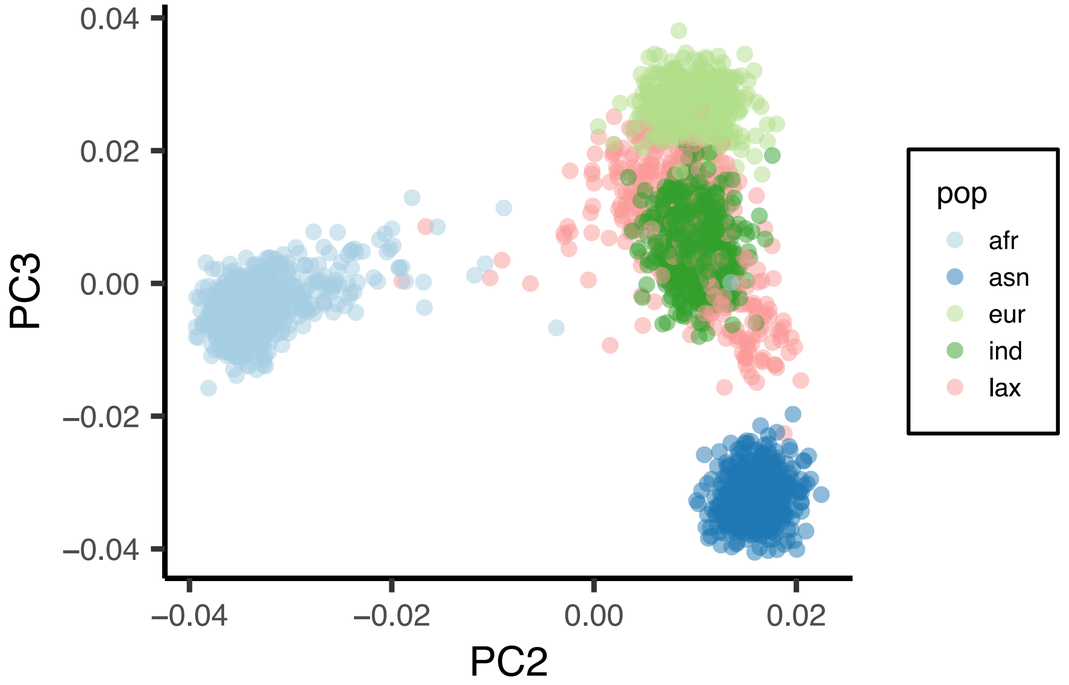
PCA already teaches us something interesting...



Let's take top 1000 most frequent variants

PCA already teaches us something interesting...





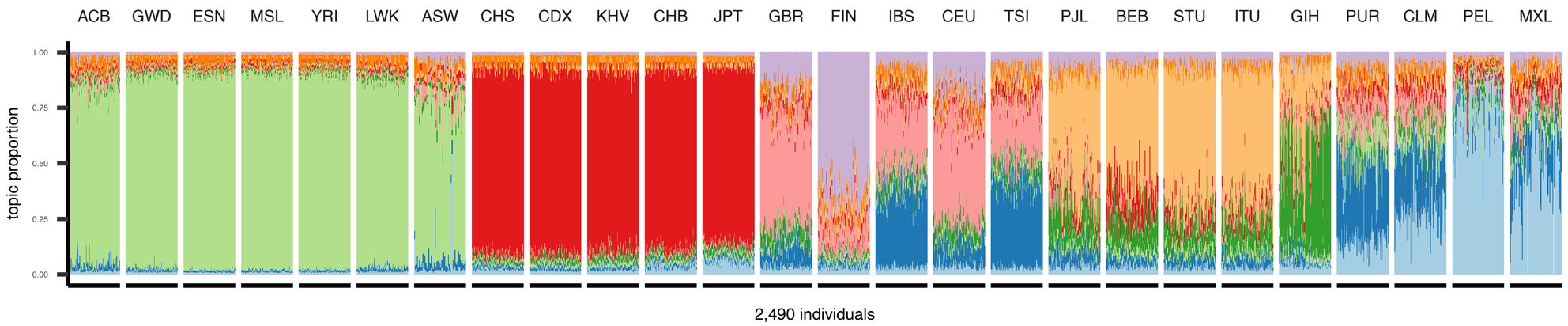
Population admixture learned from top 10k high MAF variants

A generative model:

- Sample each individual's topic proportion H_i
- Sample a topic membership for each variant j, say $Z_{ij} = k$ (could be implicitly handled)
- Genotype $X_{ij}|Z_{ij} = k \sim$ topic-specific β_{kj}

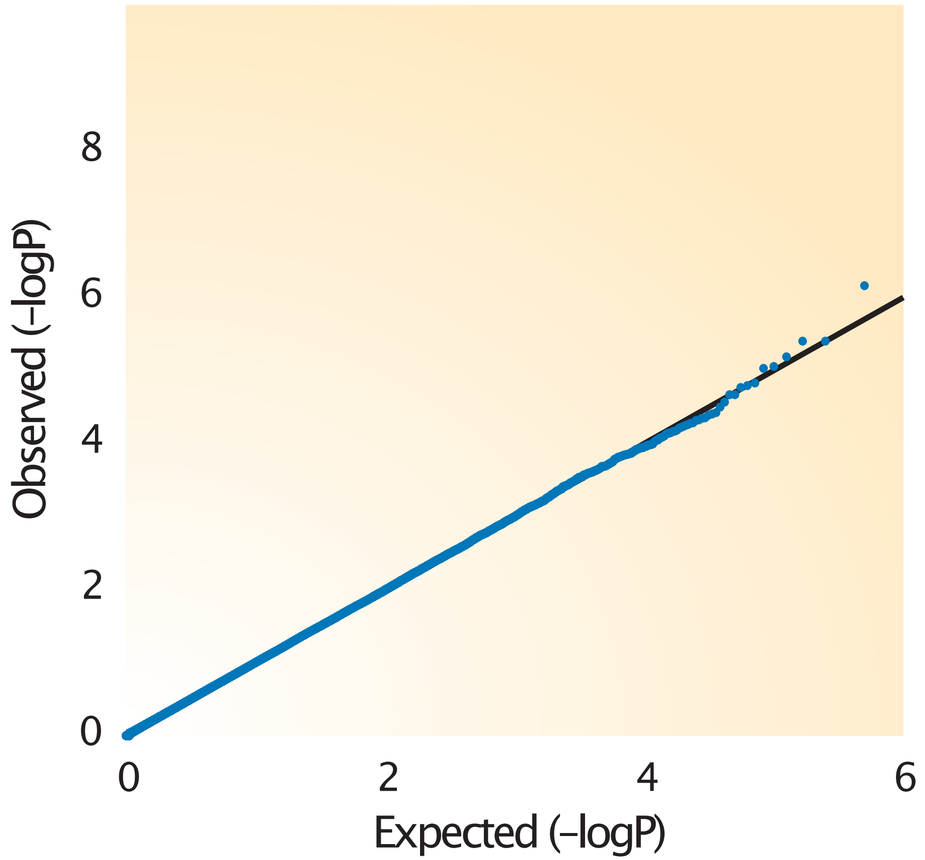
$$\mathcal{L} = \prod_{i=1}^{n} \prod_{j=1}^{10\text{k}} \left(\sum_{k=1}^{9} H_{ik} \beta_{kj} \right)^{X_{ij}}$$

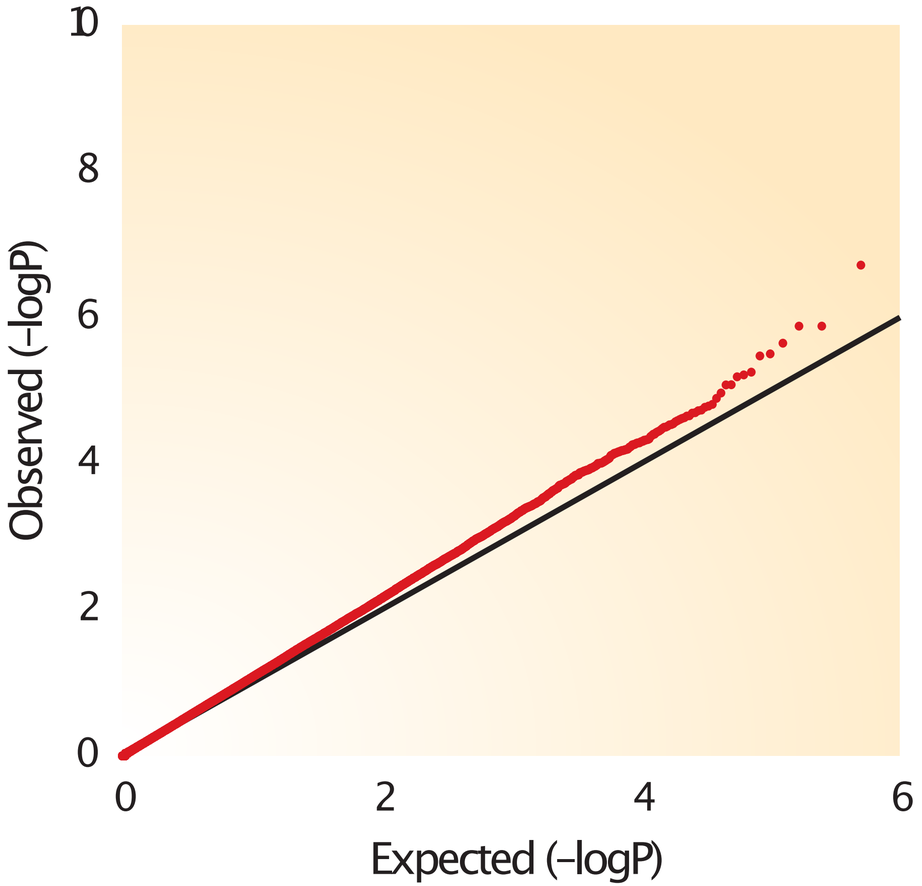
topic proportion H



Related work: Pritchard, Stephens, Donnelly, *Genetics* (2000)

Population structures may inflate GWAS stats





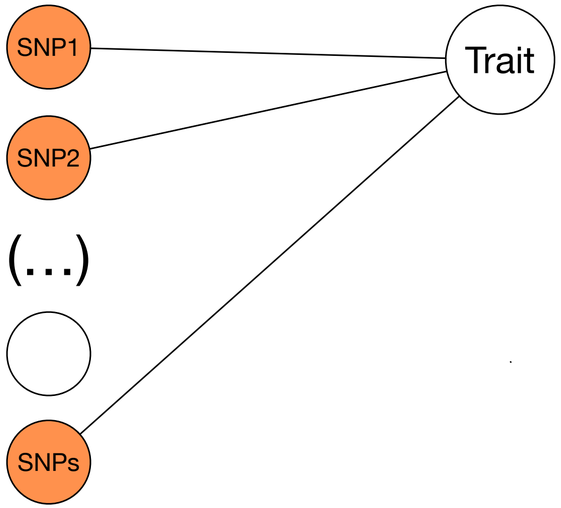
Price, Zaitlen, Reich, Patterson

Today’s lecture: GWAS and related topics

- **Human Genetics 101**
  - Variation in the human genome
  - How do we measure genetic associations?
- **Polygenic models**
  - Population structures
  - Linear Mixed Effect Model
- **Systems Genetics**
  - Summary-based GWAS analysis
  - LD-score regression: “enrichment analysis” in GWAS

A missing heritability problem

GWAS heritability ≪ Twin study heritability



A thousand, if not thousands of independent weak effect variants explain total heritability

The case of the missing heritability

Common SNPs explain a large part of heritability

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

using 300k SNPs

> 50% of height explained by genetic variation

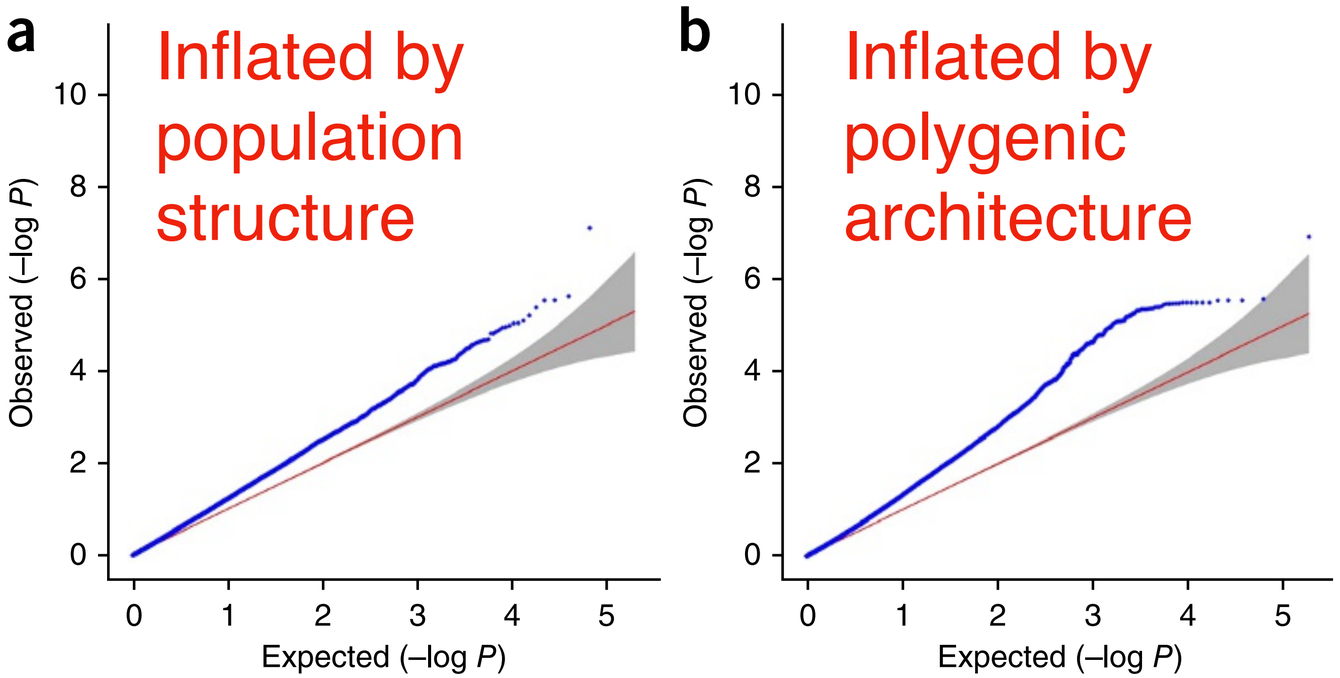
Table 1 Estimation of phenotypic variance explained from genetic relationships among unrelated individuals by restricted maximum likelihood

		No. SNPs	$L(H_0)$[a]	$L(H_1)$[b]	LRT[c]	σ_g^2 (s.e.)	σ_e^2 (s.e.)	σ_P^2 (s.e.)	h^2[d] (s.e.)
295K SNPs	Raw	294,831	−1950.89	−1936.12	29.53	0.445 (0.084)	0.546 (0.082)	0.991 (0.023)	0.449 (0.083)
	Adj.[e]	294,831	−1950.89	−1936.12	29.53	0.532 (0.101)	0.458 (0.098)	0.991 (0.023)	0.537 (0.100)
295K/516K SNPs[f]	Raw	294,831/516,345	−1950.89	−1935.94	29.89	0.449 (0.085)	0.536 (0.083)	0.986 (0.022)	0.456 (0.085)
	Adj.	294,831/516,345	−1950.89	−1935.87	30.04	0.536 (0.101)	0.449 (0.099)	0.985 (0.022)	0.544 (0.101)

[a]log-likelihood under the null hypothesis that $\sigma_g^2=0$. [b]log-likelihood under the alternative hypothesis that $\sigma_g^2 \neq 0$; [c]log-likelihood ratio test statistic, $LRT = 2[L(H_1) - L(H_0)]$. [d]Estimate of variance explained by all SNPs, with its s.e. given in the parentheses. [e]Raw estimate of genetic relationship adjusted for prediction error with equation (9) (assuming $c = 0$). [f]The genetic relationships are estimated from 1,318 individuals with 516,345 SNPs, and the other 2,607 individuals with 294,831 SNPs. See Online Methods for definitions of notations.

Using a linear mixed effect model

Population structures and polygenicity may inflate GWAS stats



Unmeasured confounding

Simply there are so many causal variants

Bulik-Sullivan *et al. Nature Genetics* (2015)

(digression) Useful facts on multivariate Gaussian distribution - 1

If we have $\mathbf{y}$

$$\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$$

then

$$\mathbb{E}[U^\top \mathbf{y}] = U^\top \mu, \quad \mathbb{V}[U^\top \mathbf{y}] = U^\top \Sigma U$$

and (affine transformation)

$$U^\top \mathbf{y} \sim \mathcal{N}(U^\top \mu, U^\top \Sigma U)$$

(digression) Useful facts on multivariate Gaussian distribution - 2

If we have two Gaussian random vectors, $\mathbf{y} \sim \mathcal{N}\left(\mu + \mathbf{u}, \Sigma_y\right)$ and $\mathbf{u} \sim \mathcal{N}(\mathbf{u}|\mathbf{0}, \Sigma_u)$

Bayesian integration:

$$\int \mathcal{N}\left(\mathbf{y}|\mu + \mathbf{u}, \Sigma_y\right) \mathcal{N}(\mathbf{u}|\mathbf{0}, \Sigma_u)\, d\mathbf{u} = \mathcal{N}\left(\mathbf{y}|\mu, \Sigma_y + \Sigma_u\right)$$

A key idea in the proof:

$$\left[\Sigma_y^{-1} - \Sigma_y^{-1}\left(\Sigma_y^{-1} + \Sigma_u^{-1}\right)^{-1}\Sigma_y^{-1}\right]^{-1} = \Sigma_y + \Sigma_u$$

by Woodbury identity.

A linear model with population-driven random effects

A linear regression model:

$$\mathbf{y} = \mathbf{x}_j \qquad \beta_j \qquad + \epsilon$$

a fixed genetic effect

What are we missing? Can we assume homo-scedasticity, i.e.,

$$\epsilon \overset{?}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

A linear model with population-driven random effects

A linear regression model:

$$\mathbf{y} = \mathbf{x}_j \quad \beta_j \quad + \epsilon$$

a fixed genetic effect

What are we missing? Can we assume homo-scedasticity, i.e.,

$$\epsilon \overset{?}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

A linear model with a random effect:

$$\mathbf{y} = \mathbf{x}_j \, \beta_j + \mathbf{u} + \epsilon$$

fixed; random effect

*N*ote: There is no specific parameterization for this $n \times 1$ random vector $\mathbf{u}$. Now, we assume:

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

A linear model with population-driven random effects - 2

We want to capture unwanted population, cohort-specific random effects by $n \times 1$ vector $\mathbf{u}$ and **remove** since our **goal** is to estimate the fixed genetic effect of a particular variant j.

$$\mathbf{y} = \underset{\text{goal}}{\mathbf{x}_j \beta_j} + \underset{\text{remove}}{\mathbf{u}} + \epsilon$$

A linear model with population-driven random effects - 2

We want to capture unwanted population, cohort-specific random effects by $n \times 1$ vector $\mathbf{u}$ and **remove** since our **goal** is to estimate the fixed genetic effect of a particular variant j.

$$\mathbf{y} = \underset{\text{goal}}{\mathbf{x}_j \beta_j} + \underset{\text{remove}}{\mathbf{u}} + \epsilon$$

1. Note that $\mathbf{u}$ shouldn't be tied to a particular variant (by definition)

A linear model with population-driven random effects - 2

We want to capture unwanted population, cohort-specific random effects by $n \times 1$ vector $\mathbf{u}$ and **remove** since our **goal** is to estimate the fixed genetic effect of a particular variant j.

$$\mathbf{y} = \underset{\text{goal}}{\mathbf{x}_j \beta_j} + \underset{\text{remove}}{\mathbf{u}} + \epsilon$$

1. Note that $\mathbf{u}$ shouldn't be tied to a particular variant (by definition)
2. Also, the covariation of $\mathbf{u}$ is primarily driven by relatedness among individuals, not the variants.

$$\mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, \tau^2 K\right), \quad K \approx \frac{1}{n} X X^\top$$

A linear mixed effect model (LMM) to test associations while adjusting population structure

We can define a hierarchical model:

$$\mathbf{y}|X, \beta, \mathbf{u}, \sigma \sim \mathcal{N}(X\beta + \mathbf{u}, \sigma^2 I) \quad (1)$$

$$\mathbf{u}|\tau, K \sim \mathcal{N}(\mathbf{0}, \tau^2 K) \quad (2)$$

If we integrate out $\mathbf{u}$,

$$\mathbf{y}|X, \beta \sim \mathcal{N}\left(\mathbf{y} \middle| X\beta, \underbrace{\tau^2 K}_{\text{genetic-relatedness matrix}} + \underbrace{\sigma^2 I}_{\text{irreducible}}\right)$$

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix
- We many not have a large matrix to learn about non-genetic confounders...

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix
- We many not have a large matrix to learn about non-genetic confounders...
- One LMM estimation can substitute multiple matrix factorization steps

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix
- We many not have a large matrix to learn about non-genetic confounders...
- One LMM estimation can substitute multiple matrix factorization steps
- We may have a good idea about relationships induced by random effects!

FaST Linear Mixed Model (Lippert *et al.* 2011)

We can resolve maximum likelihood estimate of the parameters, β, τ, σ,

$$\max \log \mathcal{N}(\mathbf{y} \,|\, X\beta, \sigma_2 \left(\delta K + I\right))$$

where $\tau^2 = \delta\sigma^2$.

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)

FaST Linear Mixed Model (Lippert *et al.* 2011)

We can resolve maximum likelihood estimate of the parameters, β, τ, σ,

$$\max \log \mathcal{N}(\mathbf{y} \,|\, X\beta, \sigma_2 \left(\delta K + I\right))$$

where $\tau^2 = \delta\sigma^2$.

We need to deal with this unfriendly form of likelihood:

$$-\frac{1}{2}\left(n \log(2\pi\sigma^2) + \log|I + \delta K| + \frac{1}{\sigma^2}\left[\mathbf{y} - X\beta\right]^\top \left(I + \delta K\right)^{-1}\left[\mathbf{y} - X\beta\right]\right)$$

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)

FaST Linear Mixed Model (Lippert *et al.* 2011)

Instead, we can transform the underlying distribution using spectral decomposition of the genetic-relatedness matrix (GRM),

$K = USU^\top$ where $U^\top U = I$, and S is a diagonal matrix.

$$\underbrace{U^\top \mathbf{y}}_{\text{projected output}} \sim \mathcal{N}\Big(\underbrace{U^\top X}_{\text{projected genotype}} \beta, \sigma^2 U^\top (I + \delta K) U \Big)$$

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)

FaST Linear Mixed Model (Lippert *et al.* 2011)

Instead, we can transform the underlying distribution using spectral decomposition of the genetic-relatedness matrix (GRM),

$K = USU^\top$ where $U^\top U = I$, and S is a diagonal matrix.

$$\underset{\text{projected output}}{U^\top \mathbf{y}} \sim \mathcal{N}\left(\underset{\text{projected genotype}}{U^\top X}\ \beta, \sigma^2 U^\top (I + \delta K) U\right)$$

$$\text{(by the affine transformation)} \sim \mathcal{N}\left(\underset{\text{projected genotype}}{U^\top X}\ \beta,\ \sigma^2 \underset{\text{diagonal matrix}}{(I + \delta S)}\right)$$

- We can find β by weighted least square
- We can find σ^2 and δ by fixing β

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)

A key research question in LMM: What covariance matrix?

If there were many types of random effects,

$$\mathbf{y} = \underbrace{X\beta}_{\text{fixed}} + \underbrace{\mathbf{u} + \mathbf{w} + \ldots}_{\text{random effects}} + \underbrace{\epsilon}_{\text{unknown}}$$

A key research question in LMM: What covariance matrix?

If there were many types of random effects,

$$\mathbf{y} = \underbrace{X\beta}_{\text{fixed}} + \underbrace{\mathbf{u} + \mathbf{w} + \ldots}_{\text{random effects}} + \underbrace{\epsilon}_{\text{unknown}}$$

We would need to many covariance matrices:

$$\mathbf{y}|\cdot \sim \mathcal{N}\left(X\beta,\, \sigma^2(I + \underbrace{\delta_u K_u + \delta_w K_w + \ldots}_{\text{random effects}})\right)$$

A key research question in LMM: What covariance matrix?

If there were many types of random effects,

$$\mathbf{y} = \underbrace{X\beta}_{\text{fixed}} + \underbrace{\mathbf{u} + \mathbf{w} + \ldots}_{\text{random effects}} + \underbrace{\epsilon}_{\text{unknown}}$$

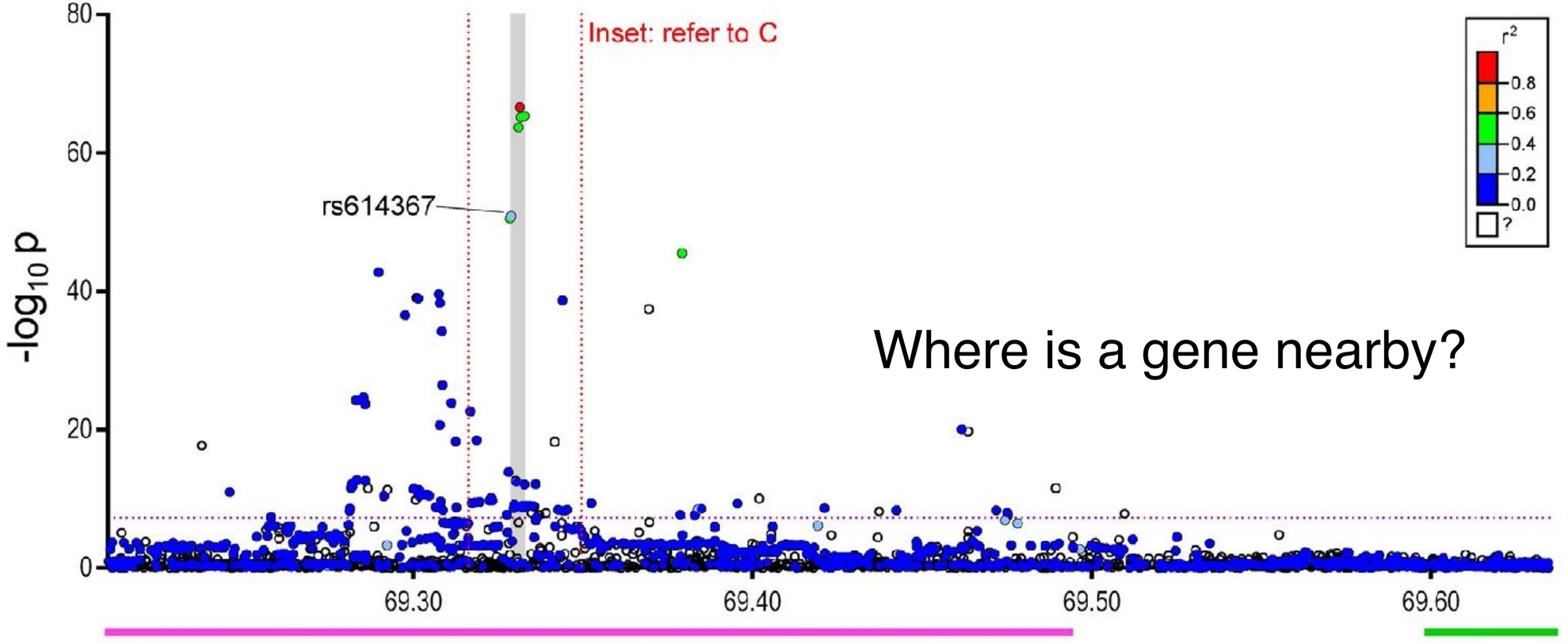
We would need to many covariance matrices:

$$\mathbf{y}|\cdot \sim \mathcal{N}\left(X\beta,\, \sigma^2(I + \underbrace{\delta_u K_u + \delta_w K_w + \ldots}_{\text{random effects}})\right)$$

If we only care about variance decomposition $\beta_j \sim \mathcal{N}(0, \tau)$:

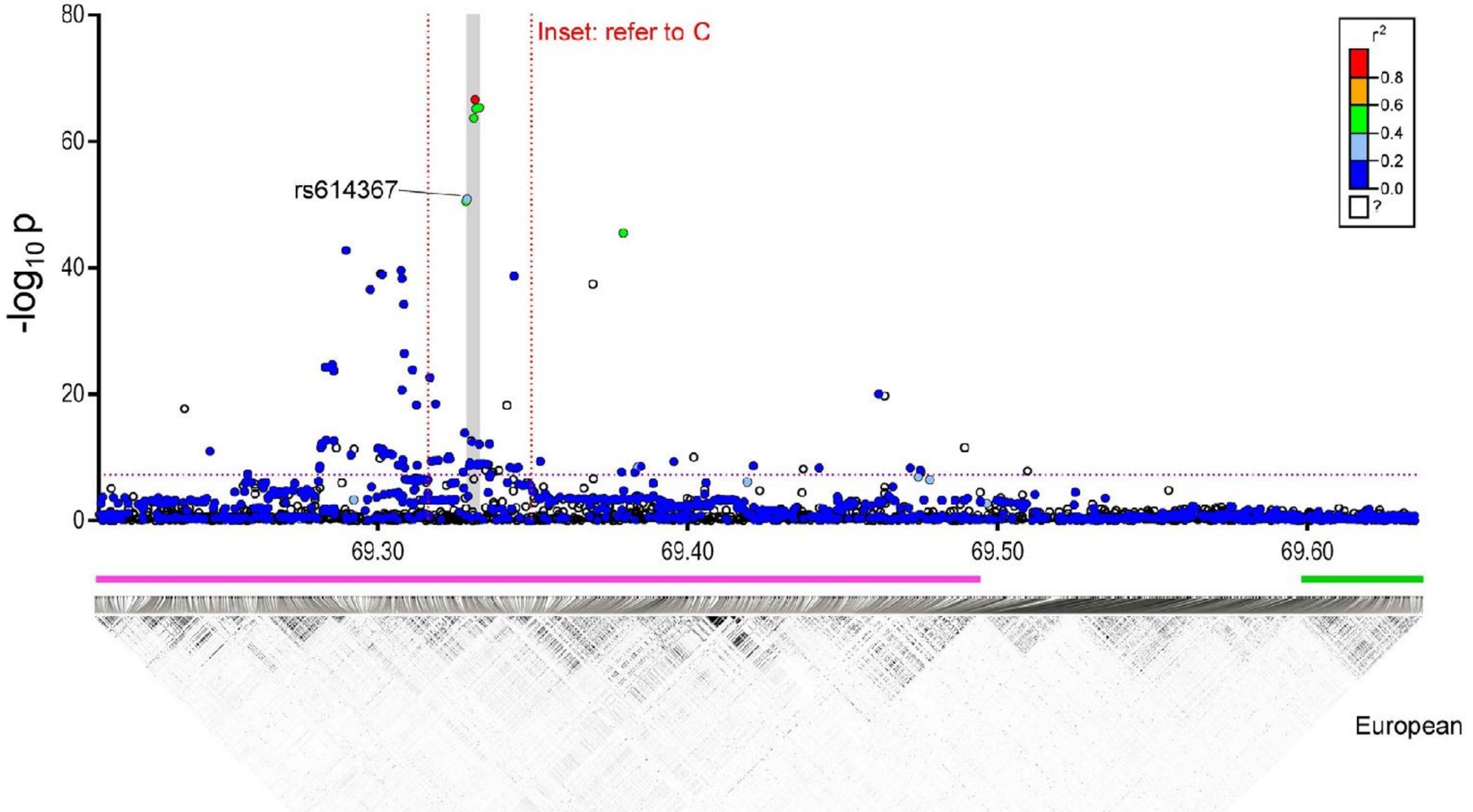
$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0},\, \sigma^2\left(\underbrace{\frac{\sigma^2_{\text{genetic}}}{n} XX^\top}_{\text{observed genetic}} + I + \underbrace{\delta_u K_u + \delta_w K_w + \ldots}_{\text{random effects}}\right)\right)$$

90% of GWAS hits on the non-coding regions



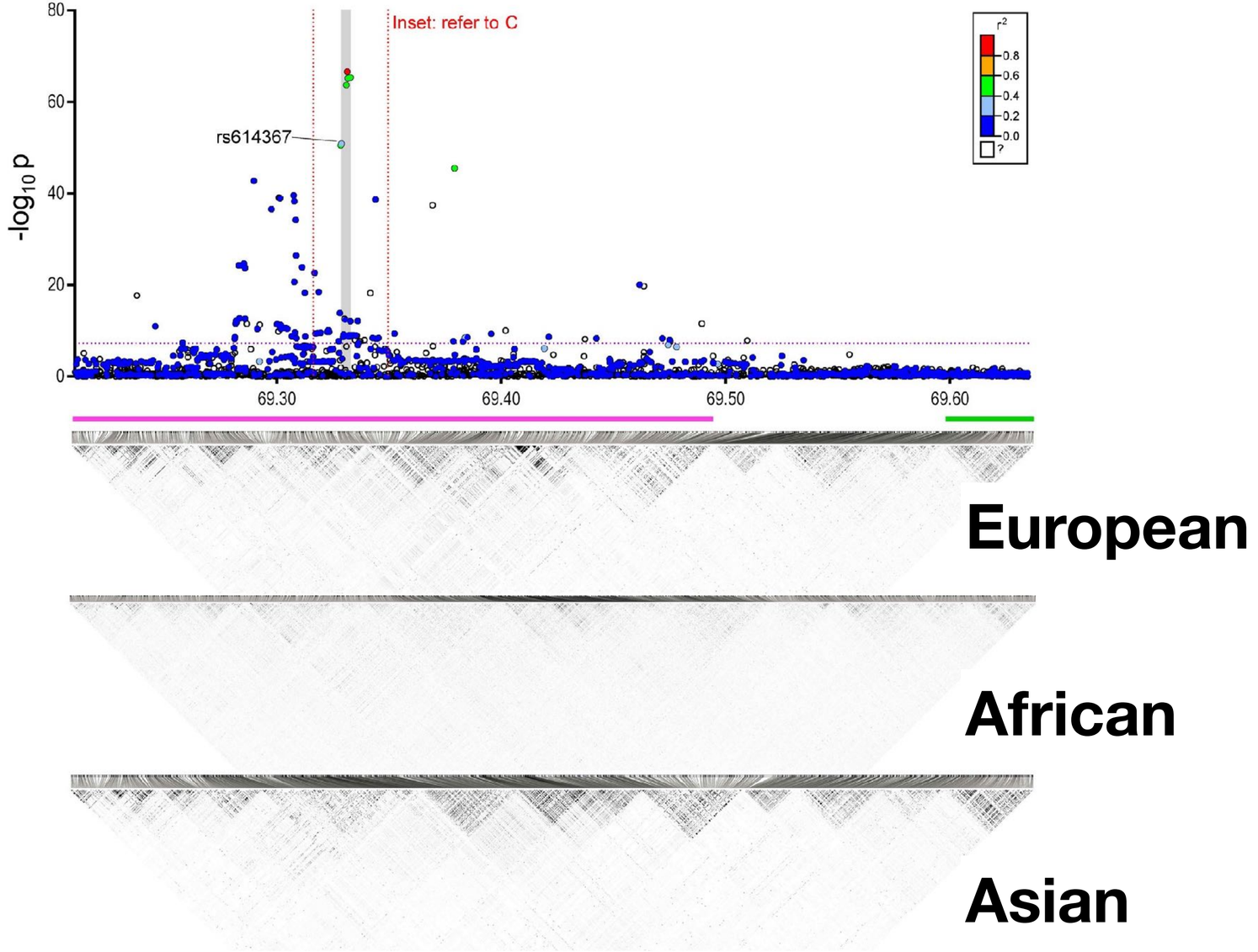
Edwards *et al. AJHG* (2013)

LD structure, many significantly associated variants



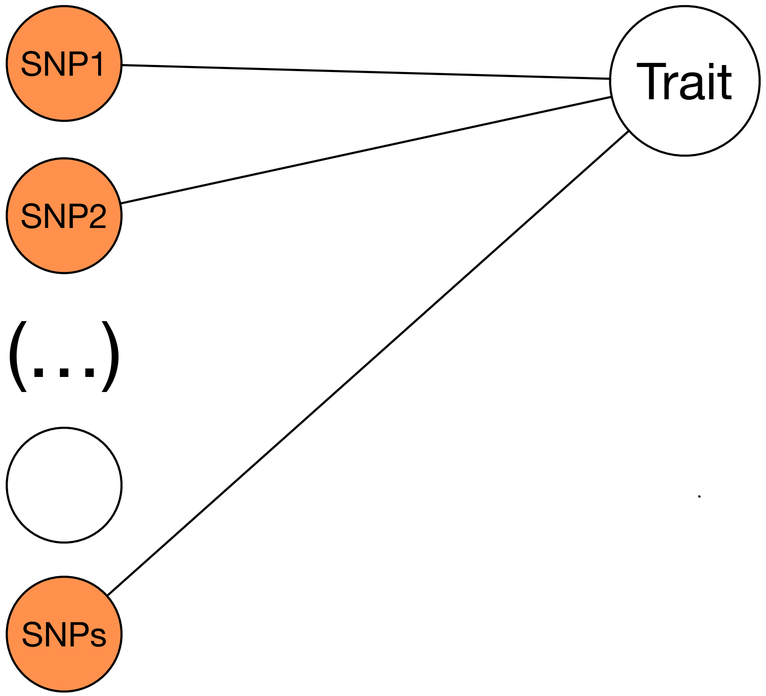
Edwards *et al. AJHG* (2013)

Different population structures, different LD patterns

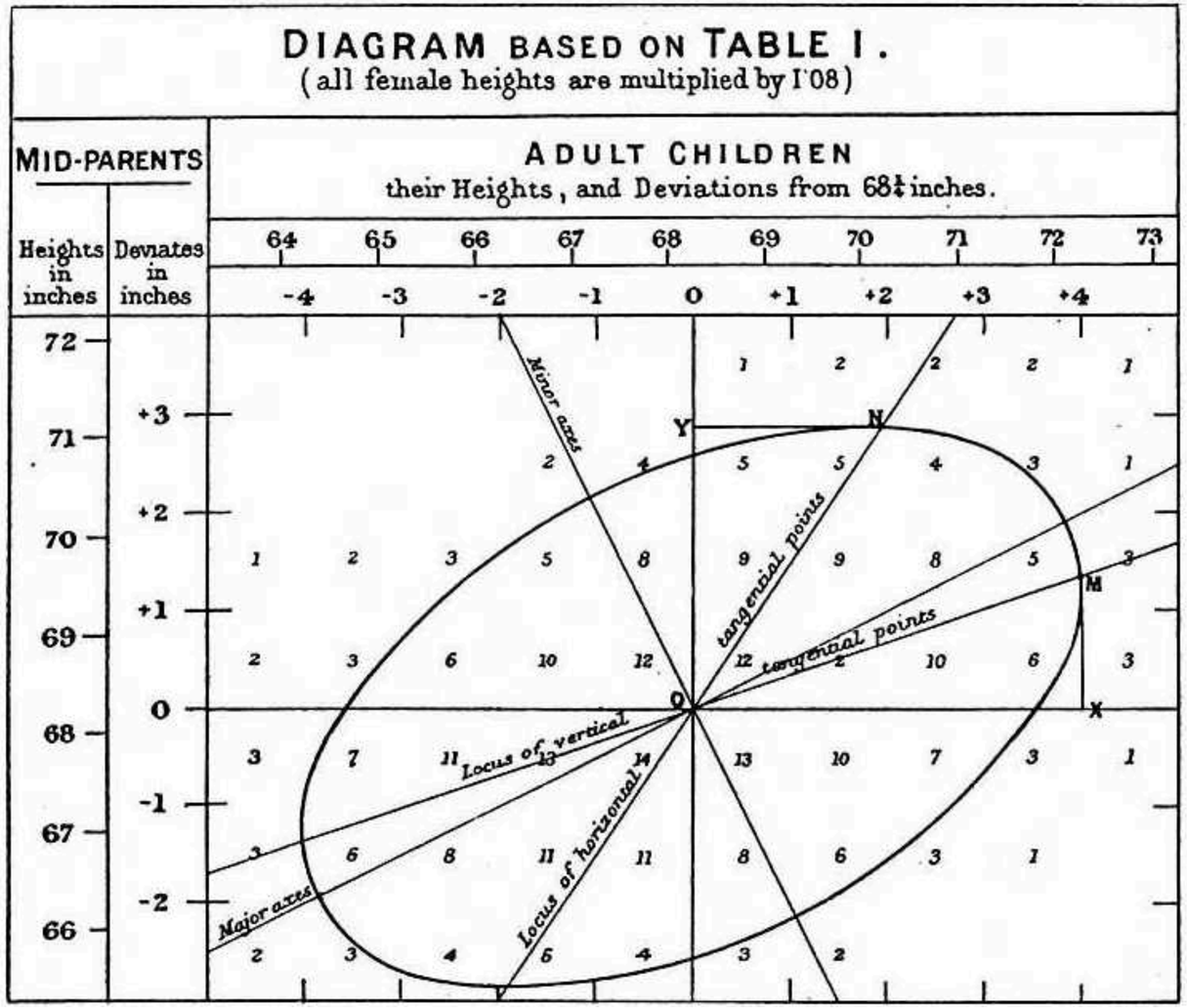


Edwards *et al. AJHG* (2013)

Many complex traits are polygenic



Galton's regression model finally confirmed



Parents' height

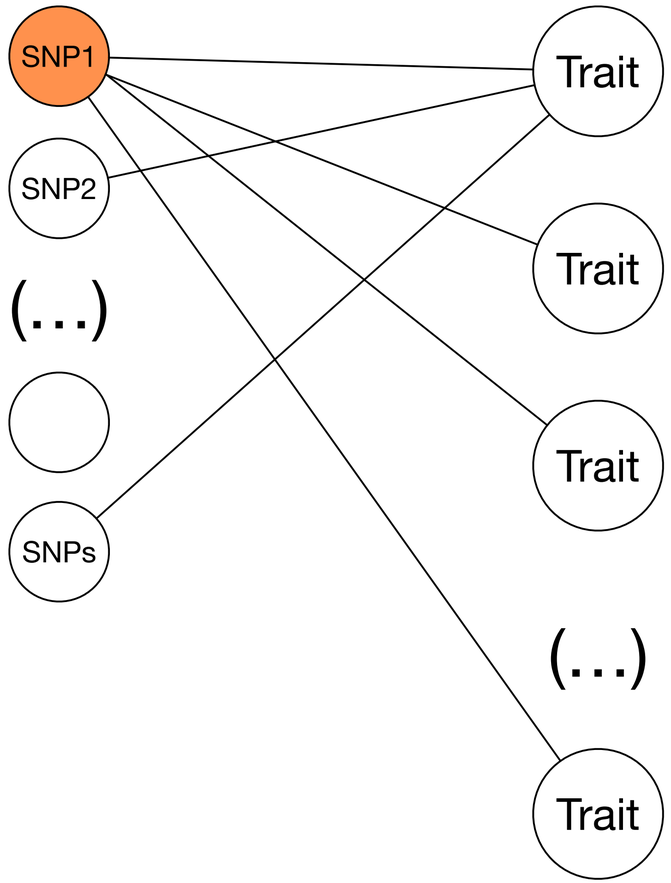
Children's height

Is there a "tall/small" gene?

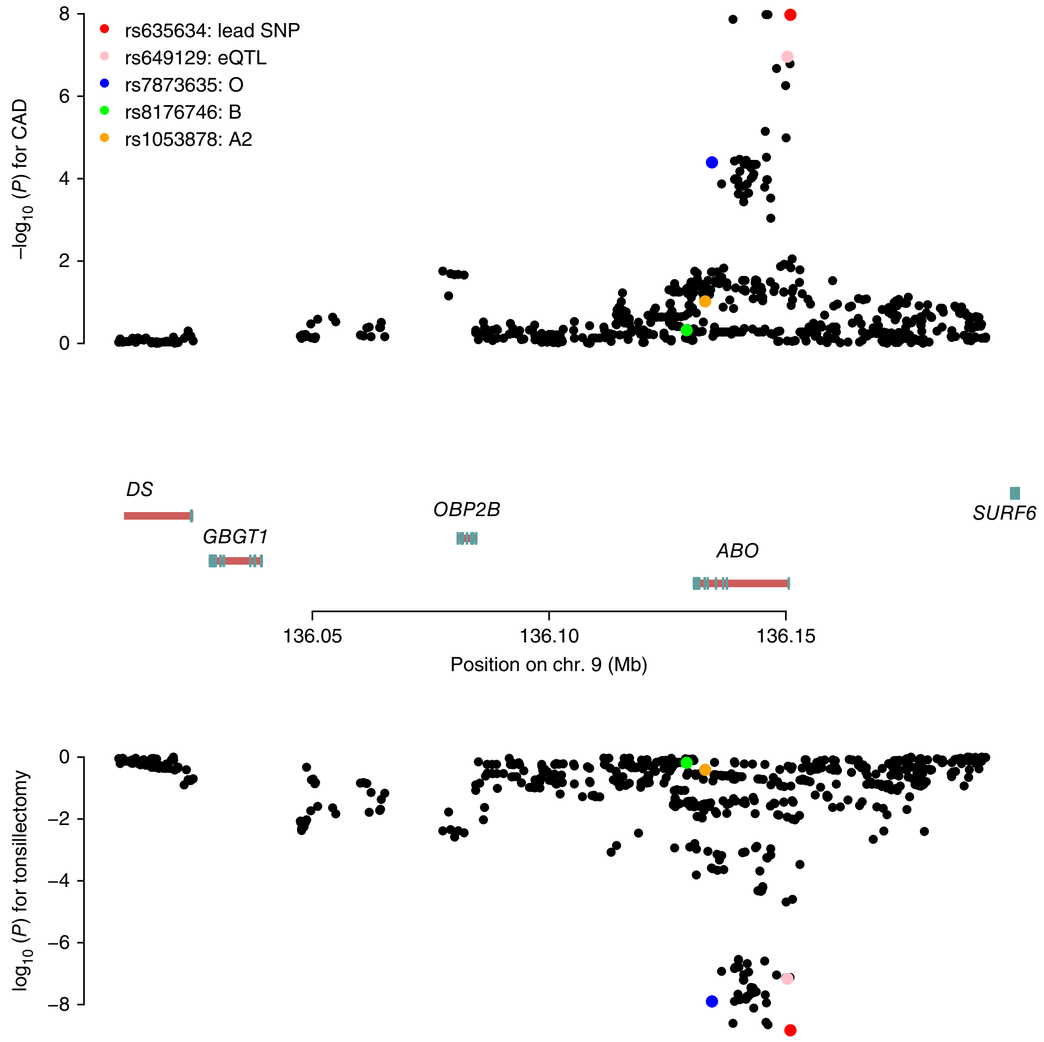
A single variant may act on many traits

Pleiotropy

A single variant is associated, over and over with many different human traits!

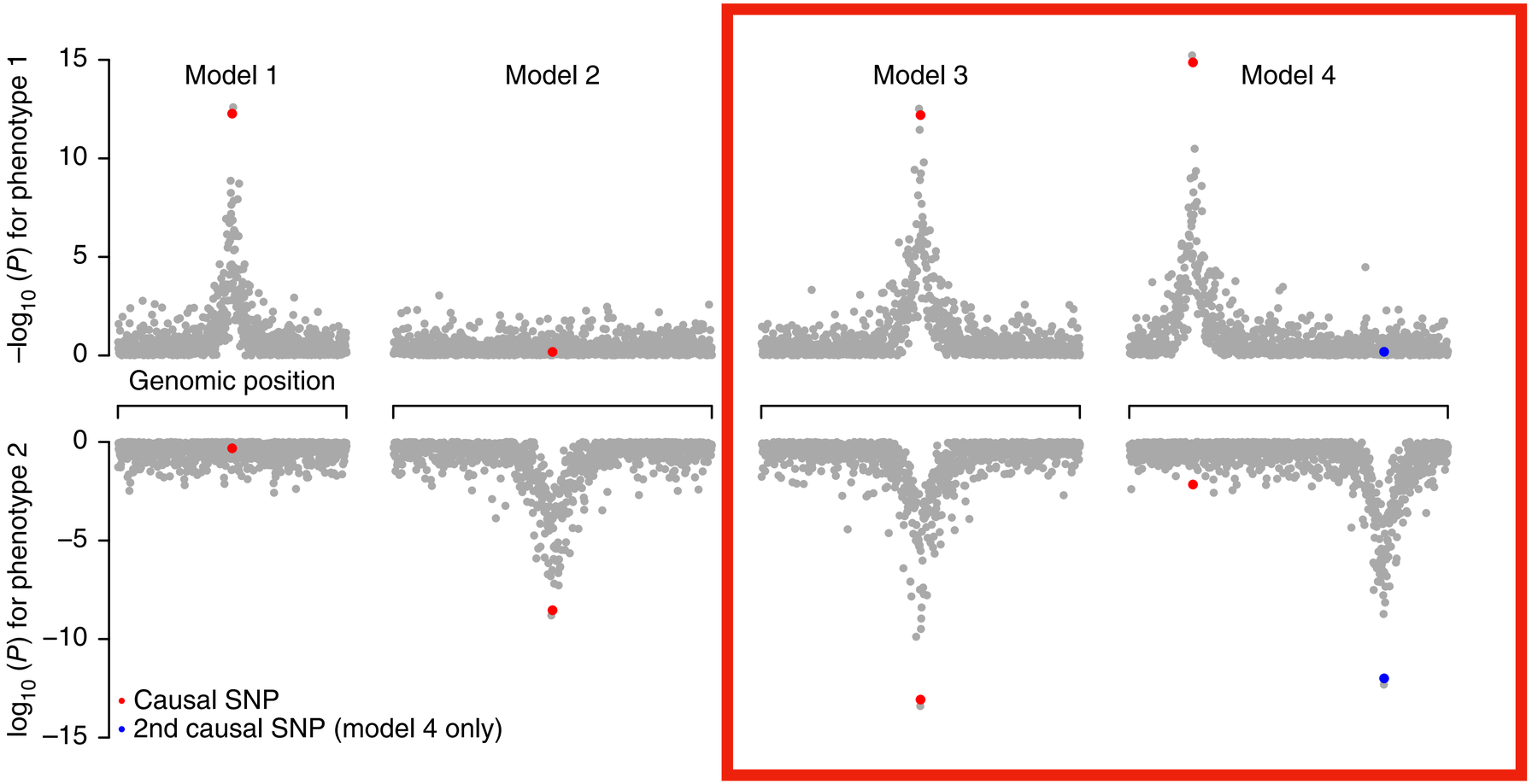


A single variant may act on many traits



Pickrell *et al. Nature Genetics* (2016)

A single variant may act on many traits

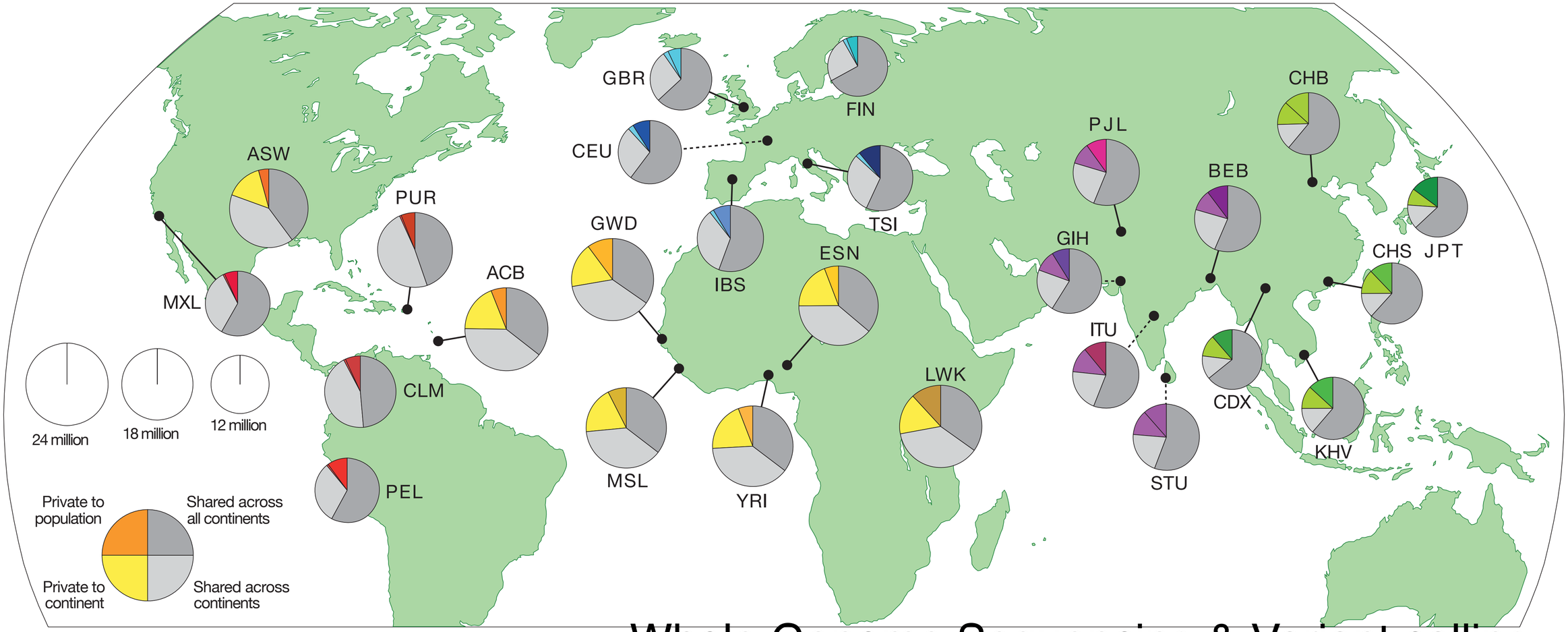


Pickrell *et al. Nature Genetics* (2016)

Today's lecture: GWAS and related topics

- **Human Genetics 101**
  - Variation in the human genome
  - How do we measure genetic associations?
- **Polygenic models**
  - Population structures
  - Linear Mixed Effect Model
- **Systems Genetics: post-GWAS analysis**
  - Summary-based GWAS analysis + polygenic risk prediction
  - LD-score regression: "enrichment analysis" in GWAS

The 1000 genomes project



Whole Genome Sequencing & Variant calling across 1000 individuals in many different groups

The 1000 Genomes Project Consortium, Nature, 2015

Much of human genetics problems centre on two covariance matrices

For a standardized $n \times p$ genotype matrix X (n: #individuals, p: #SNPs),

1. Genetic relatedness matrix (GRM)

individual by individual, $n \times n$ matrix

$$K \approx XX^\top/n$$

The matrix K captures population structure/correlation across different individuals.

- Kinship matrix; population admixture
- Human migration history

2. Linkage disequilibrium (LD)

SNP by SNP, $p \times p$ matrix

$$R \approx X^\top X/n$$

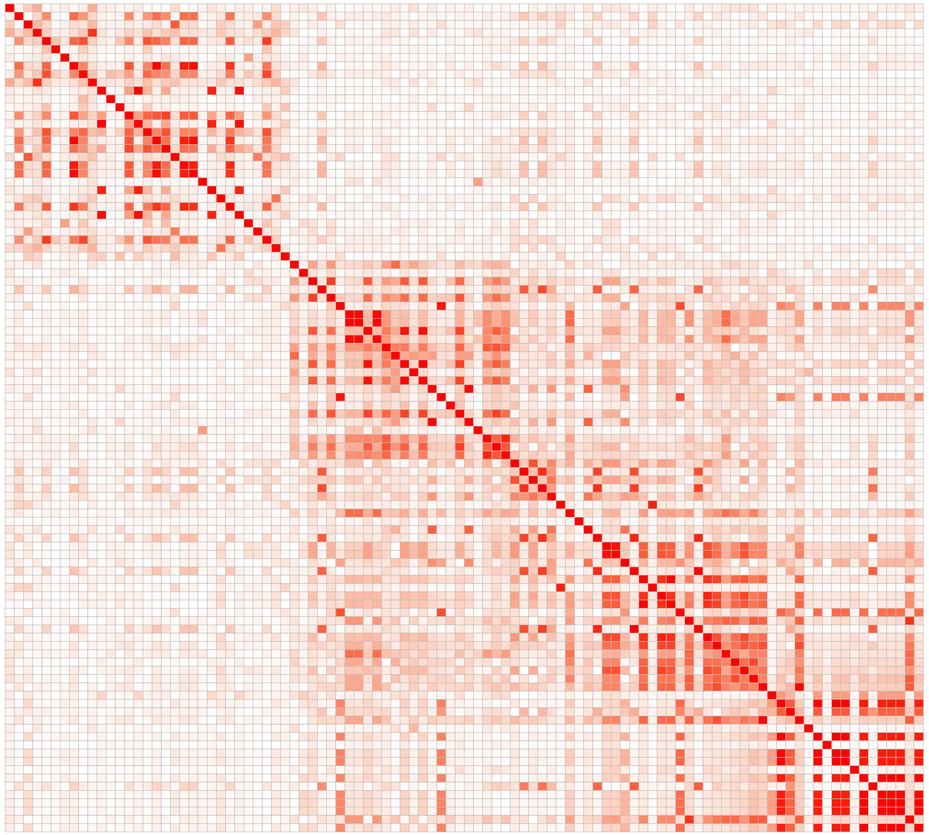
The matrix R captures localized correlation patterns along the genomic axis within a chromosome.

- LD matrix
- The results of many, many recombination events

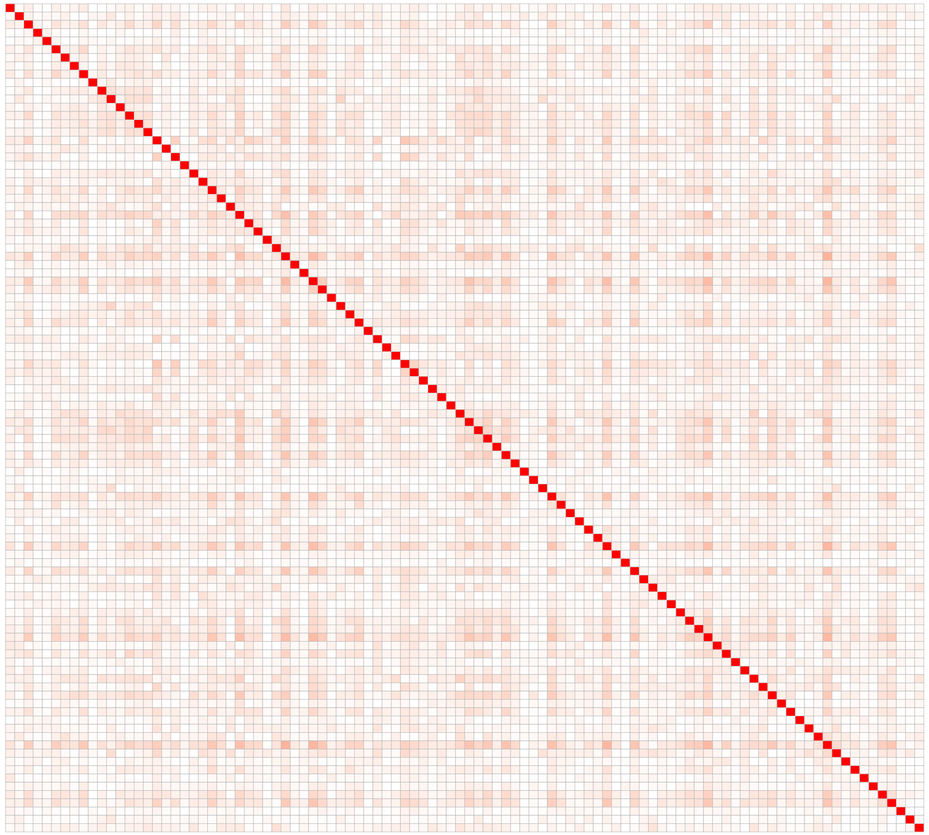
Let's discuss LD structures

Pairwise correlations between SNPs

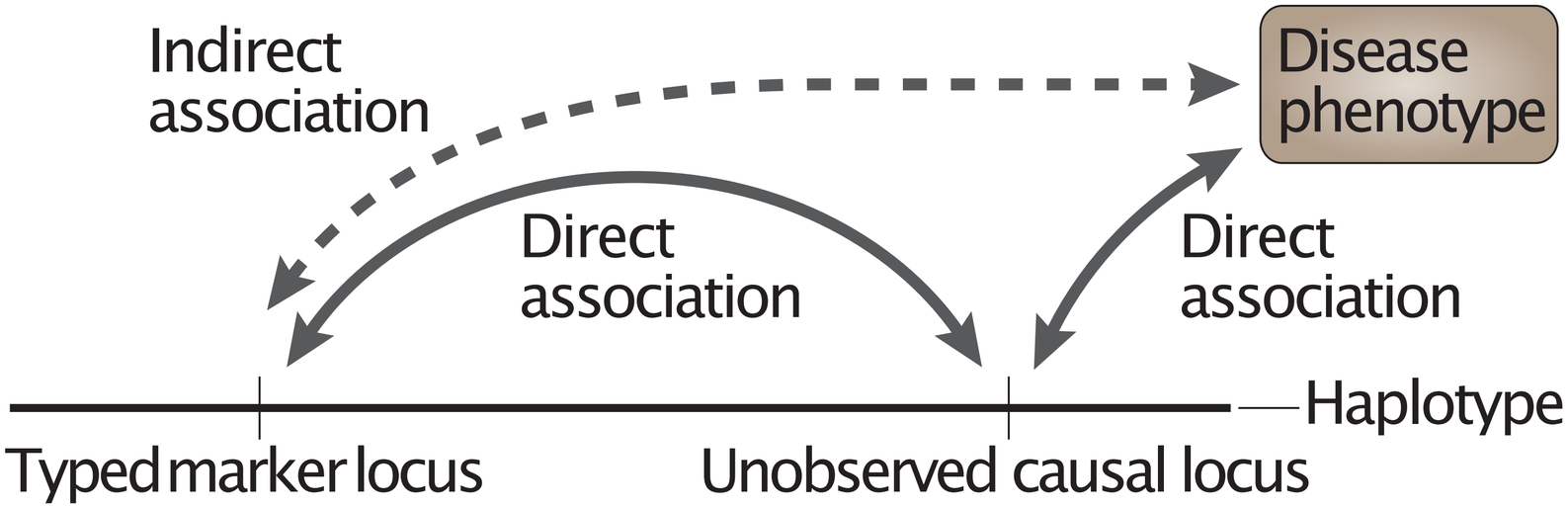
consecutive 100 SNPs



random 100 SNPs



GWAS fail to pinpoint exact locations associated with a disease



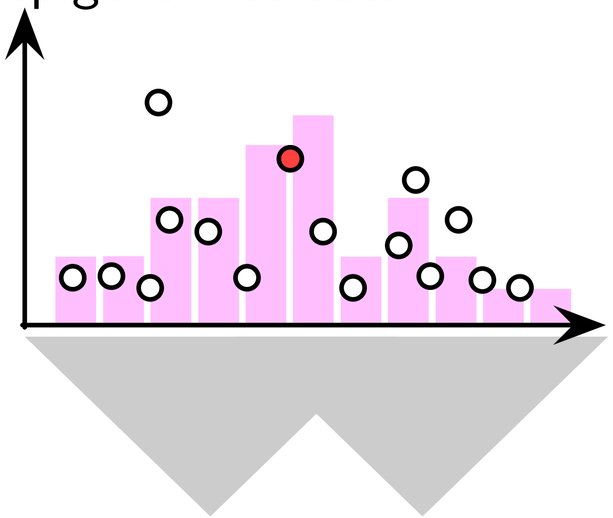
Balding, *Nature Genetics Review* (2006)

Common strategies to deal with LD structures

Strategy 1.

Fine-mapping to find a handful of causal ones

- Bayesian posterior estimation
- Overlap with epigenomics data



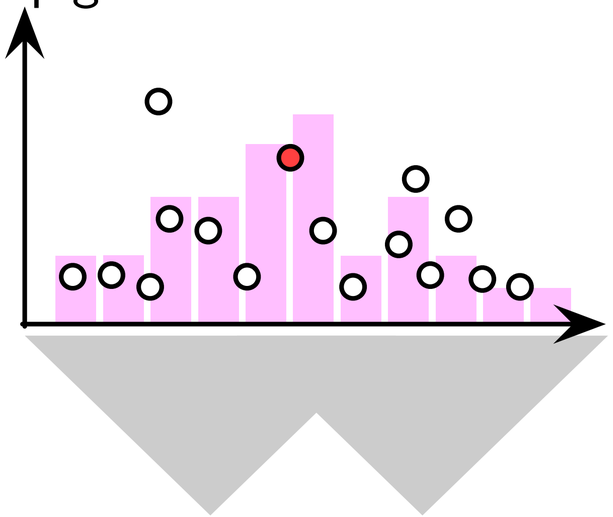
x-axis: genomic location; y-axis: -log10 p-value

Common strategies to deal with LD structures

Strategy 1.

Fine-mapping to find a handful of causal ones

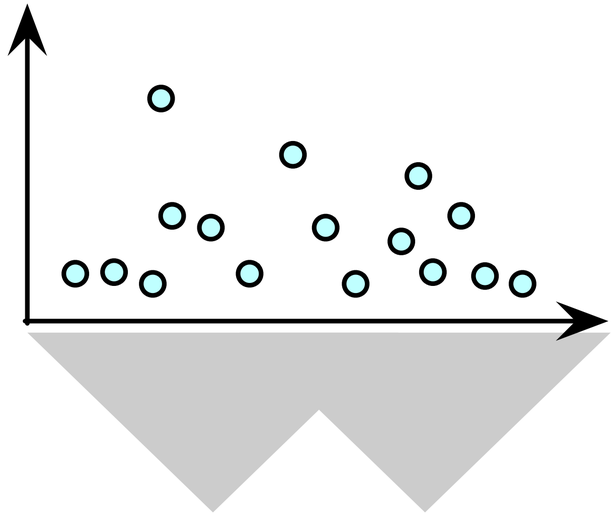
- Bayesian posterior estimation
- Overlap with epigenomics data



Strategy 2.

Aggregation to combine all the information:

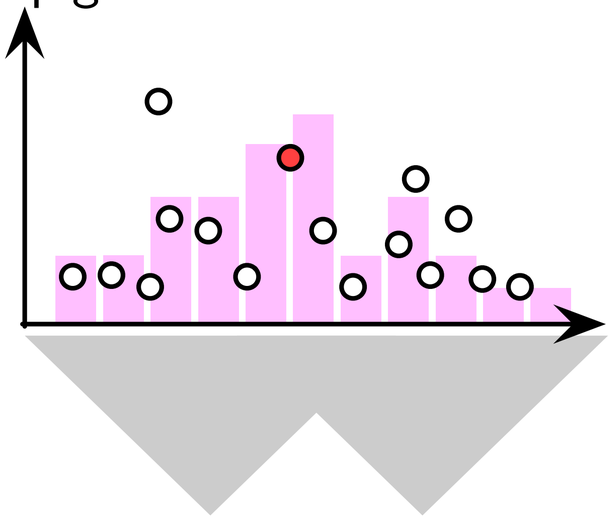
- Rare variant analysis
- Gene-level enrichment/association



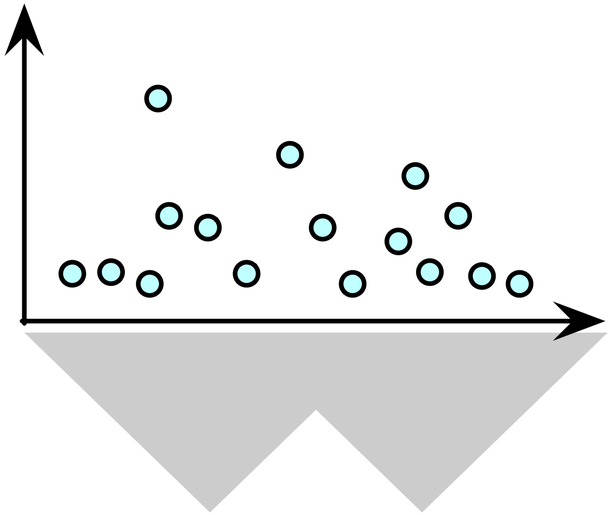
x-axis: genomic location; y-axis: -log10 p-value

Common strategies to deal with LD structures

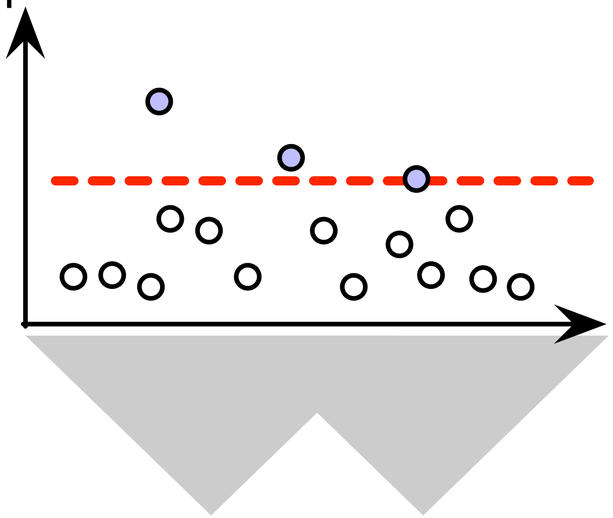
Strategy 1.
Fine-mapping to find a handful of causal ones
- Bayesian posterior estimation
- Overlap with epigenomics data



Strategy 2.
Aggregation to combine all the information:
- Rare variant analysis
- Gene-level enrichment/association



Strategy 3. **Pruning** to remove somewhat redundant information (heuristics)
- p-value thresholding
- Useful in polygenic risk prediction



x-axis: genomic location; y-axis: -log10 p-value

GWAS summary statistics

But we have summary statistics of meta-analysis:

A generative model of SNP-level statistics

For each β_j's, effect size, variance, z-score:

$$\hat{\beta}_j = \frac{\sum_i X_{ij} Y_i}{\sum_i X_{ij}^2}, \quad \hat{\mathbb{V}}[\beta_j] = \frac{\sigma_\epsilon^2}{\sum_{i=1} X_{ij}^2}, \quad Z_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\mathbb{V}}[\beta_j]}}$$

GWAS summary statistics

But we have summary statistics of meta-analysis:

A generative model of SNP-level statistics

For each β_j's, effect size, variance, z-score:

$$\hat{\beta}_j = \frac{\sum_i X_{ij} Y_i}{\sum_i X_{ij}^2}, \quad \hat{\mathbb{V}}[\beta_j] = \frac{\sigma_\epsilon^2}{\sum_{i=1} X_{ij}^2}, \quad Z_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\mathbb{V}}[\beta_j]}}$$

Although **the underlying multivariate regression model**:

$$\mathbf{y} = X\theta + \epsilon$$

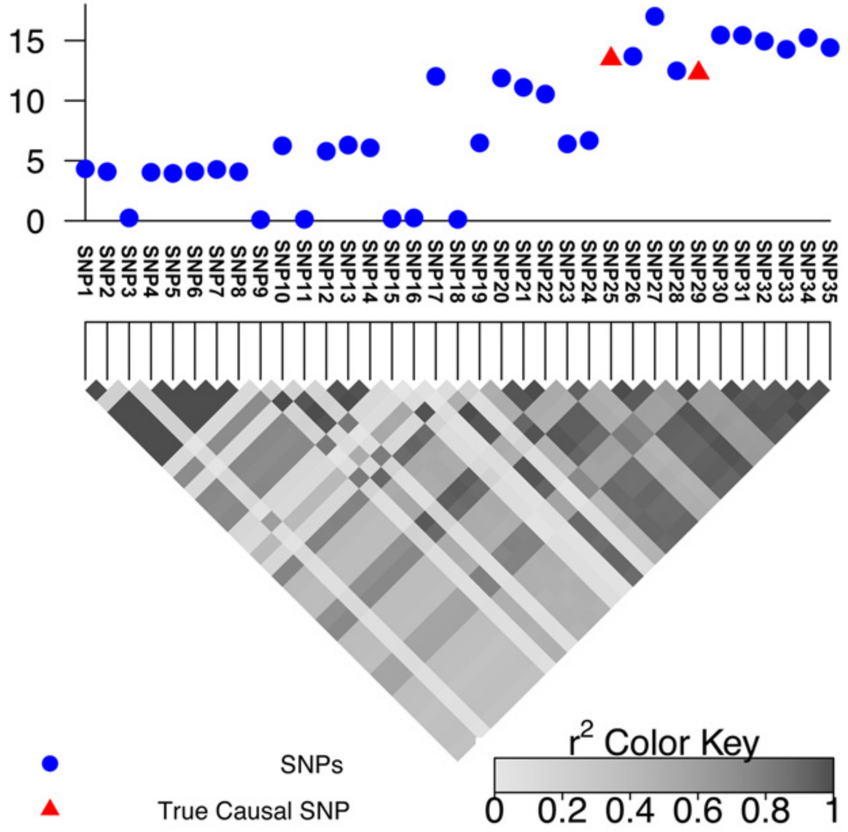
where $\theta_j \neq \beta_j$

What is the relationship between the summary (univariate) and multivariate effects?

For simplicity, let's assume standardized genotype matrix X, i.e., $\bar{X}_j = 0$ and $\hat{\sigma}^2_{X_j} = 1$. The we have z-score

$$\hat{Z}_j = \sum_{i=1}^{n} X_{ij} Y_i / \sigma_\epsilon \sqrt{n}$$

for all $j \in [p]$.

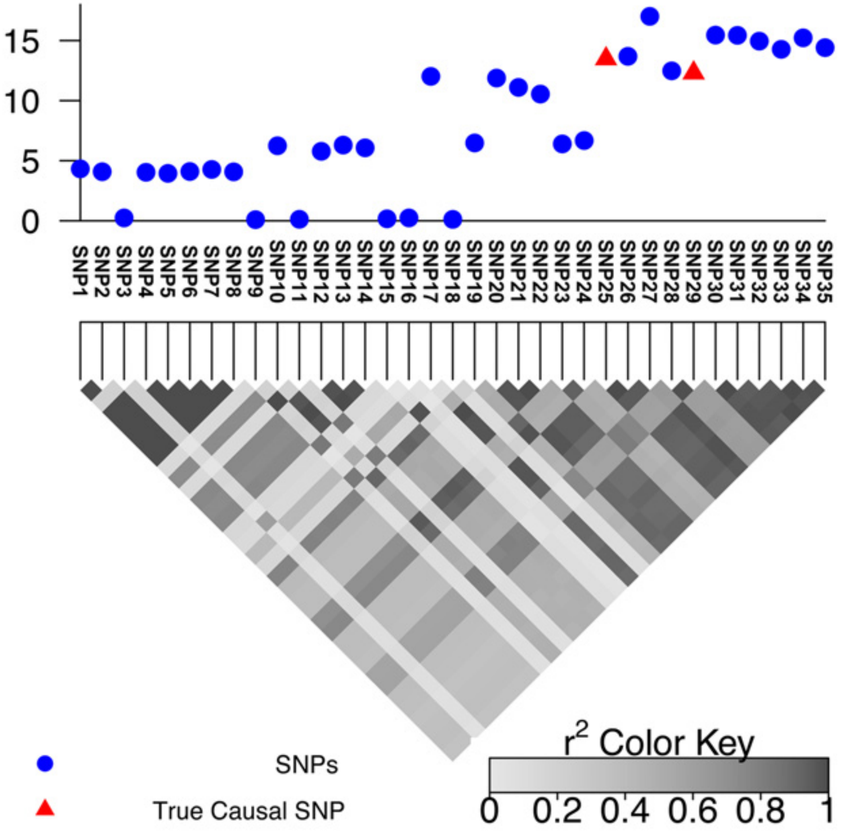


Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

What is the relationship between the summary (univariate) and multivariate effects?

For simplicity, let's assume standardized genotype matrix X, i.e., $\bar{X}_j = 0$ and $\hat{\sigma}^2_{X_j} = 1$. The we have z-score

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \mathbf{y}$$



Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

What is the relationship between the summary (univariate) and multivariate effects?

For simplicity, let's assume standardized genotype matrix X, i.e., $\bar{X}_j = 0$ and $\hat{\sigma}^2_{X_j} = 1$. The we have z-score

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \mathbf{y} = \frac{1}{\sigma\sqrt{n}} X^\top \underbrace{(X\theta + \epsilon)}_{\text{a multivariate model}}$$

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

What is the relationship between the summary (univariate) and multivariate effects?

For simplicity, let's assume standardized genotype matrix X, i.e., $\bar{X}_j = 0$ and $\hat{\sigma}^2_{X_j} = 1$. The we have z-score

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \mathbf{y} = \frac{1}{\sigma\sqrt{n}} X^\top \underbrace{(X\theta + \epsilon)}_{\text{a multivariate model}}$$

$$= \frac{\sqrt{n}}{\sigma} \underbrace{\left(\frac{1}{n} X^\top X\right)}_{\text{LD}} \theta + \frac{1}{\sigma\sqrt{n}} X^\top \epsilon$$

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

What is the relationship between the summary (univariate) and multivariate effects?

For simplicity, let's assume standardized genotype matrix X, i.e., $\bar{X}_j = 0$ and $\hat{\sigma}^2_{X_j} = 1$. The we have z-score

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \mathbf{y} = \frac{1}{\sigma\sqrt{n}} X^\top \underbrace{(X\theta + \epsilon)}_{\text{a multivariate model}}$$

$$= \mathbf{R}\frac{\sqrt{n}}{\sigma}\theta + \tilde{\epsilon}, \quad \tilde{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$$

where $\mathbf{R} = n^{-1} X^\top X$ is an empirical LD matrix.

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

How can GWAS p-value fool human genetics research?

Some thought experiment on causal variant i. We generated y as follows:

$$\mathbf{y} = \mathbf{x}_i\theta_i + ...$$

Because of LD and population structure, we have

$$\mathbf{x}_j = \mathbf{u} + \tilde{\mathbf{x}}_j \, ... \, , \, \mathbb{E}\left[\mathbf{x}_i^\top \tilde{\mathbf{x}}_j\right] = n\delta$$

This structural bias and correlation will simply remain in a z-score:

$$z_j = n^{-1/2}\mathbf{x}_j^\top \mathbf{y} = n^{-1/2}(\mathbf{u} + \tilde{\mathbf{x}}_j + ...)^\top(\mathbf{x}_i\theta_i + ...)$$

How can GWAS p-value fool human genetics research?

Some thought experiment on causal variant i. We generated y as follows:

$$\mathbf{y} = \mathbf{x}_i\theta_i + ...$$

Because of LD and population structure, we have

$$\mathbf{x}_j = \mathbf{u} + \tilde{\mathbf{x}}_j \ldots, \ \mathbb{E}\left[\mathbf{x}_i^\top \tilde{\mathbf{x}}_j\right] = n\delta$$

This structural bias and correlation will simply remain in a z-score:

$$z_j = \frac{\mathbf{u}^\top \mathbf{x}_i}{n^{1/2}}\theta_i + \frac{\overbrace{\mathbf{x}_i^\top \tilde{\mathbf{x}}_j}^{\to n\delta}}{n^{1/2}}\theta_i + ... = \left(\underbrace{\frac{\mathbf{u}^\top \mathbf{x}_i}{n^{1/2}}}_{\text{population structure}} + \underbrace{\delta n^{1/2}}_{\text{tight LD between } i,j} \right) \theta_i$$

How can GWAS p-value fool human genetics research?

Some thought experiment on causal variant i. We generated y as follows:

$$\mathbf{y} = \mathbf{x}_i \theta_i + ...$$

Because of LD and population structure, we have

$$\mathbf{x}_j = \mathbf{u} + \tilde{\mathbf{x}}_j \ldots, \mathbb{E}\left[\mathbf{x}_i^\top \tilde{\mathbf{x}}_j\right] = n\delta$$

This structural bias and correlation will simply remain in a z-score:

- Population structure-bias will diminish as $n \to \infty$
- LD-bias will aggravate as $n \to \infty$

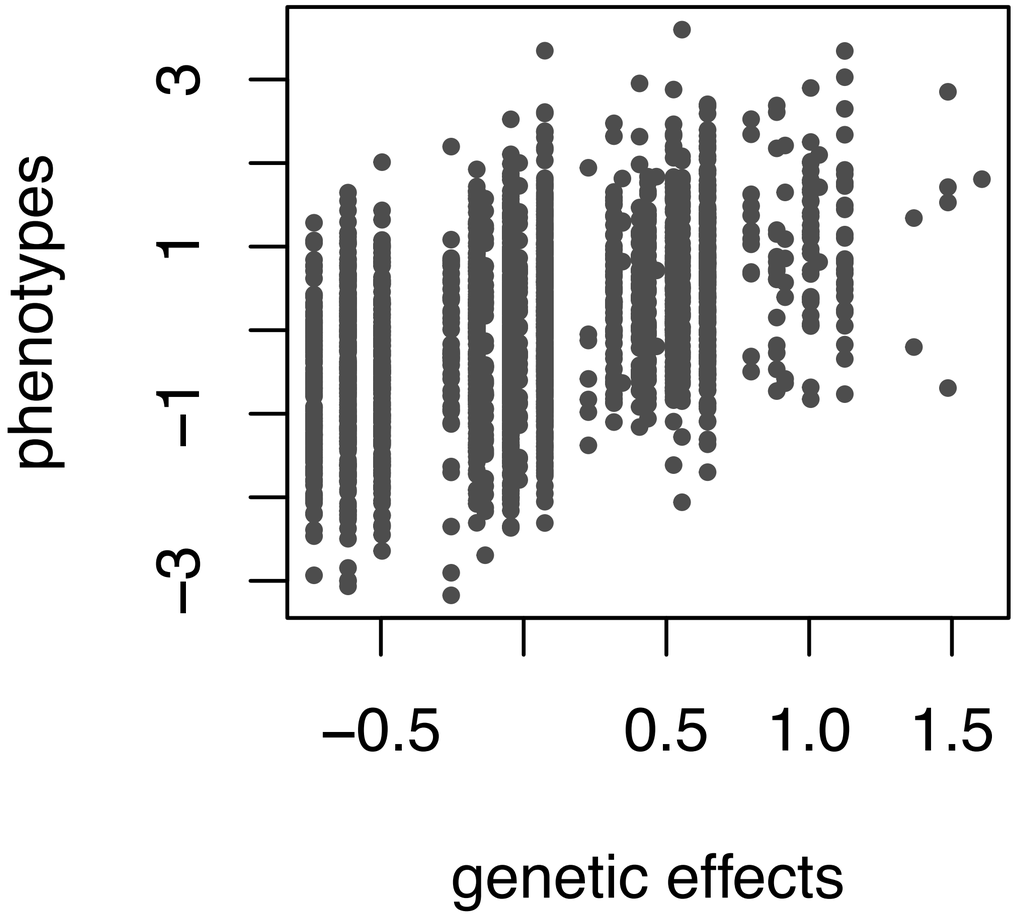
Polygenic regression model

- An “infinitesimal” (aka omnigenic) model: Every variant contributes to total heritability in complex traits.
- First coined by R.A. Fisher, trying to reconcile between Biometricians’ and Medelians’ different points of view.
- Pea (Mendelian) vs. height (biometrician)

Let's simulate GWAS data

```r
set.seed(47)
xtot <- apply(X, 2, scale)

## just remove missing values
xtot[is.na(xtot)] <- 0

## random subset 300
rr <- sample(nrow(xtot), 300)
xx <- xtot[-rr, , drop=F]

sim <- simulate.pgs(xx, 0.2, 3)
y <- sim$y
```

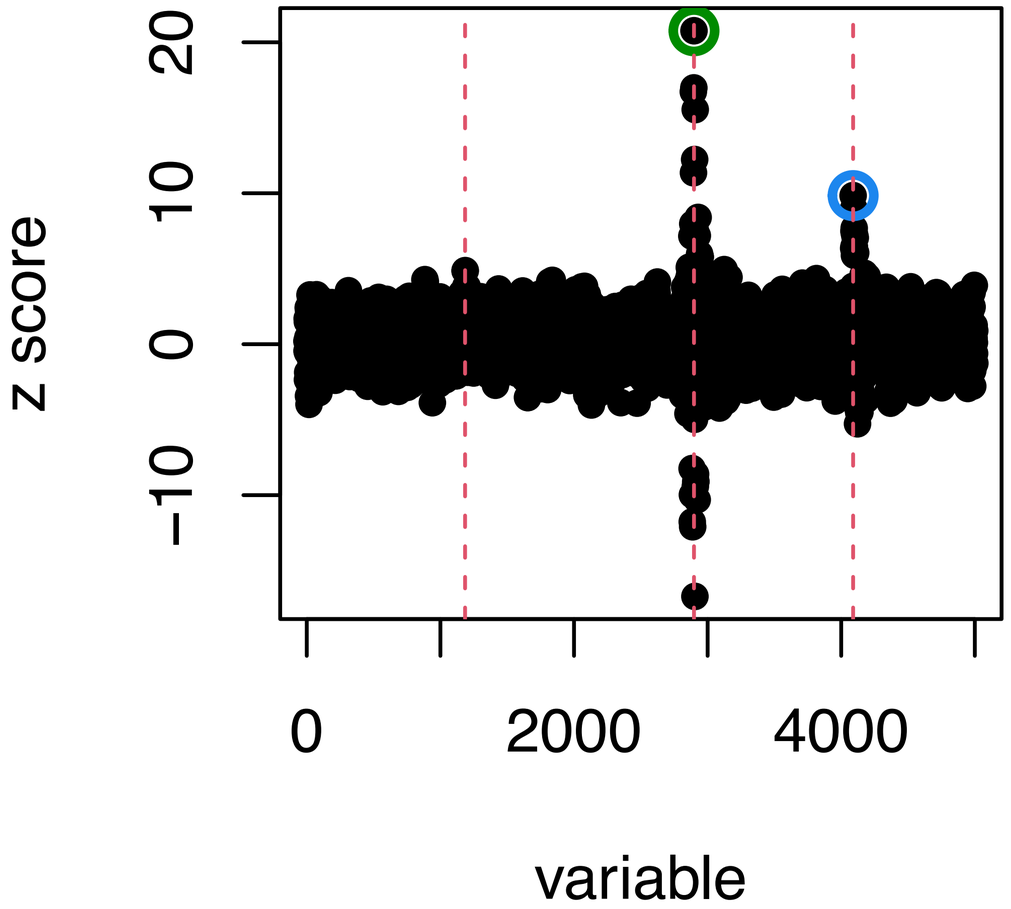
$$\mathbf{y} = \sum_{j \in \text{causal}} \mathbf{x}_j \beta_j + \epsilon$$



Run SuSiE to identify causal variants

```
library(susieR)

susie.full <-
    susie(xx, y, L = 30,
 compute_univariate_zscore=TRUE)
```

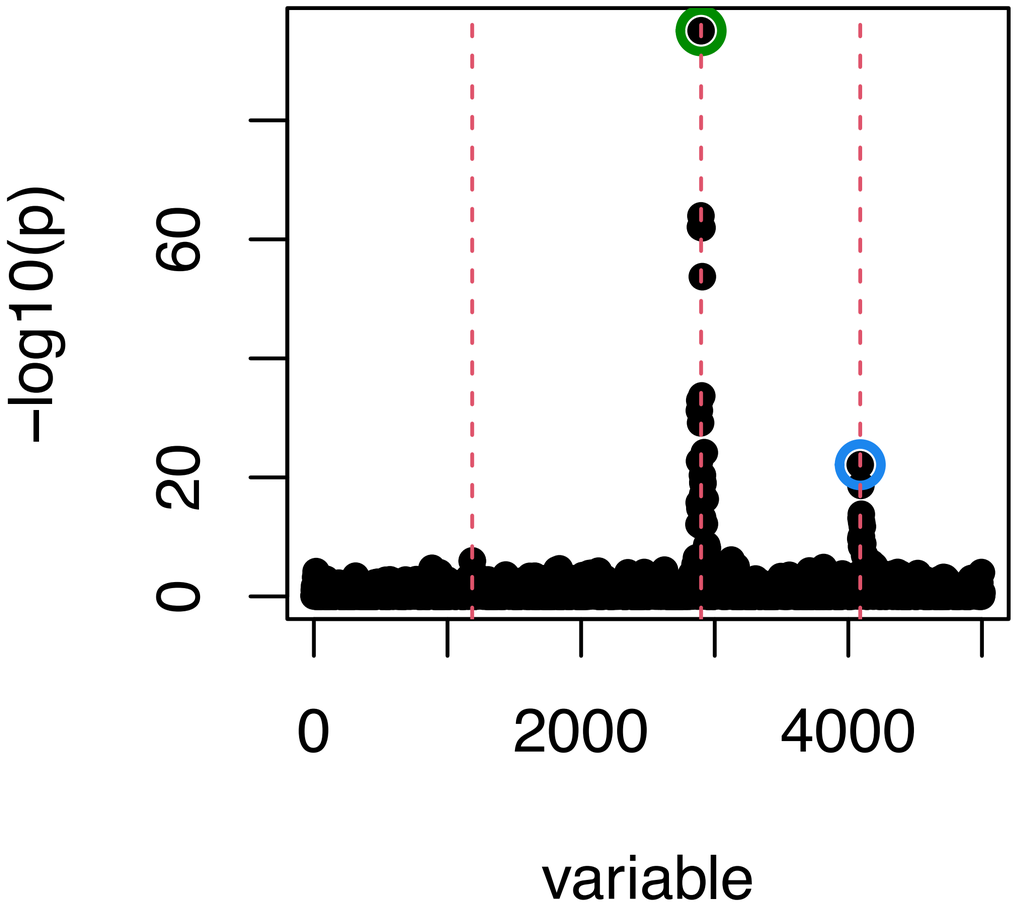
- 2190 $\times$ 5000 X genotype
- 2190 $\times$ 1 y response
- Dashed lines: causal variants



Run SuSiE to identify causal variants

```
library(susieR)

susie.full <-
    susie(xx, y, L = 30,
 compute_univariate_zscore=TRUE)
```

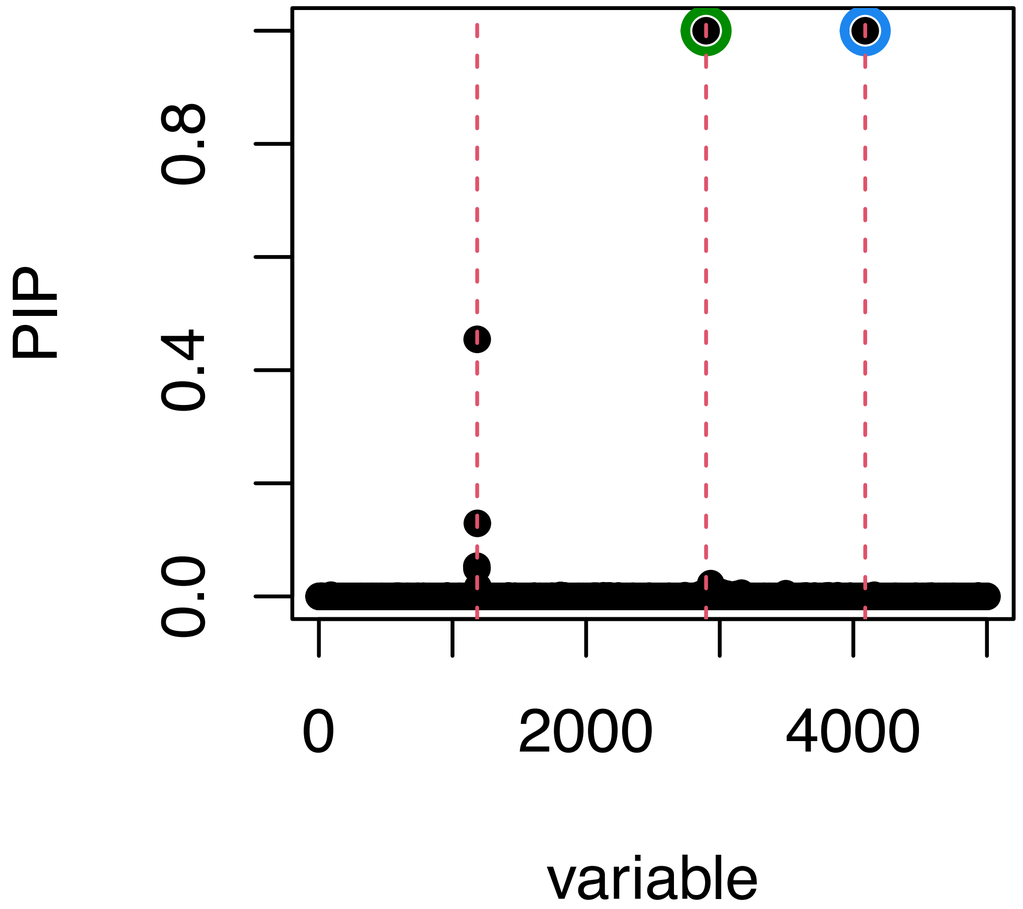
- 2190 $\times$ 5000 X genotype
- 2190 $\times$ 1 y response
- Dashed lines: causal variants



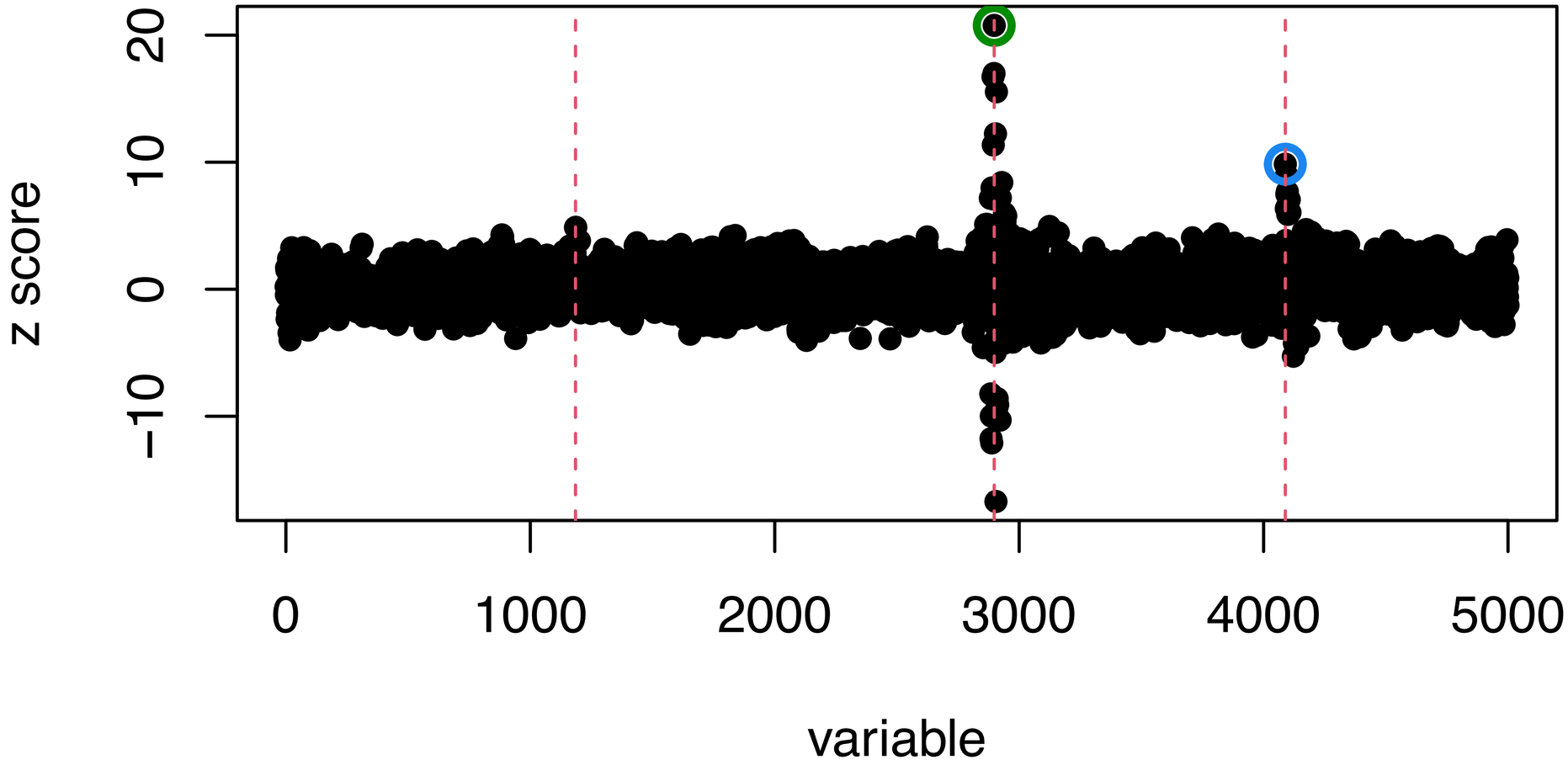
Run SuSiE to identify causal variants

```
library(susieR)

susie.full <-
    susie(xx, y, L = 30,
 compute_univariate_zscore=TRUE)
```

- 2190 $\times$ 5000 X genotype
- 2190 $\times$ 1 y response
- Dashed lines: causal variants



What if we only have a summary statistics vector?



z-scores...

How can we calculate summary z-scores?

```
## compute X'y
xty <- crossprod(xx, y)

## compute x'x
x2 <- colSums(xx^2)

## Maximum likelihood
beta.hat <- xty / x2
se.hat <- 1/sqrt(x2)
z <- beta.hat / se.hat
```

We can estimate z-scores...

$$\hat{Z}_j = \sum_{i=1}^{n} X_{ij} Y_i / \sigma_\epsilon \sqrt{n}$$

for all $j \in [p]$.
assuming $\sigma_\epsilon \approx 1$.

Note: The above assumption is not bad because most genetic effects explain a relatively small portion of phenotypic variability, and phenotype vectors are standardized.

Why do we model z-scores (or other summary statistics)?

- There is no privacy concern
- Directly translate to GWAS p-values
- A majority of GWAS variants are common
- LD structures are similar within each ancestry group

Moreover, we normally don't have a full genotype matrix

```
## a random subset of individuals
## rr <- sample(nrow(xx), 300)

## genotype information of this subset
xx.sub <- xtot[rr, , drop = FALSE]

## true phenotype vector y not observed
y.true <- xtot[rr, sim$causal, drop=F] %*% sim$theta
```

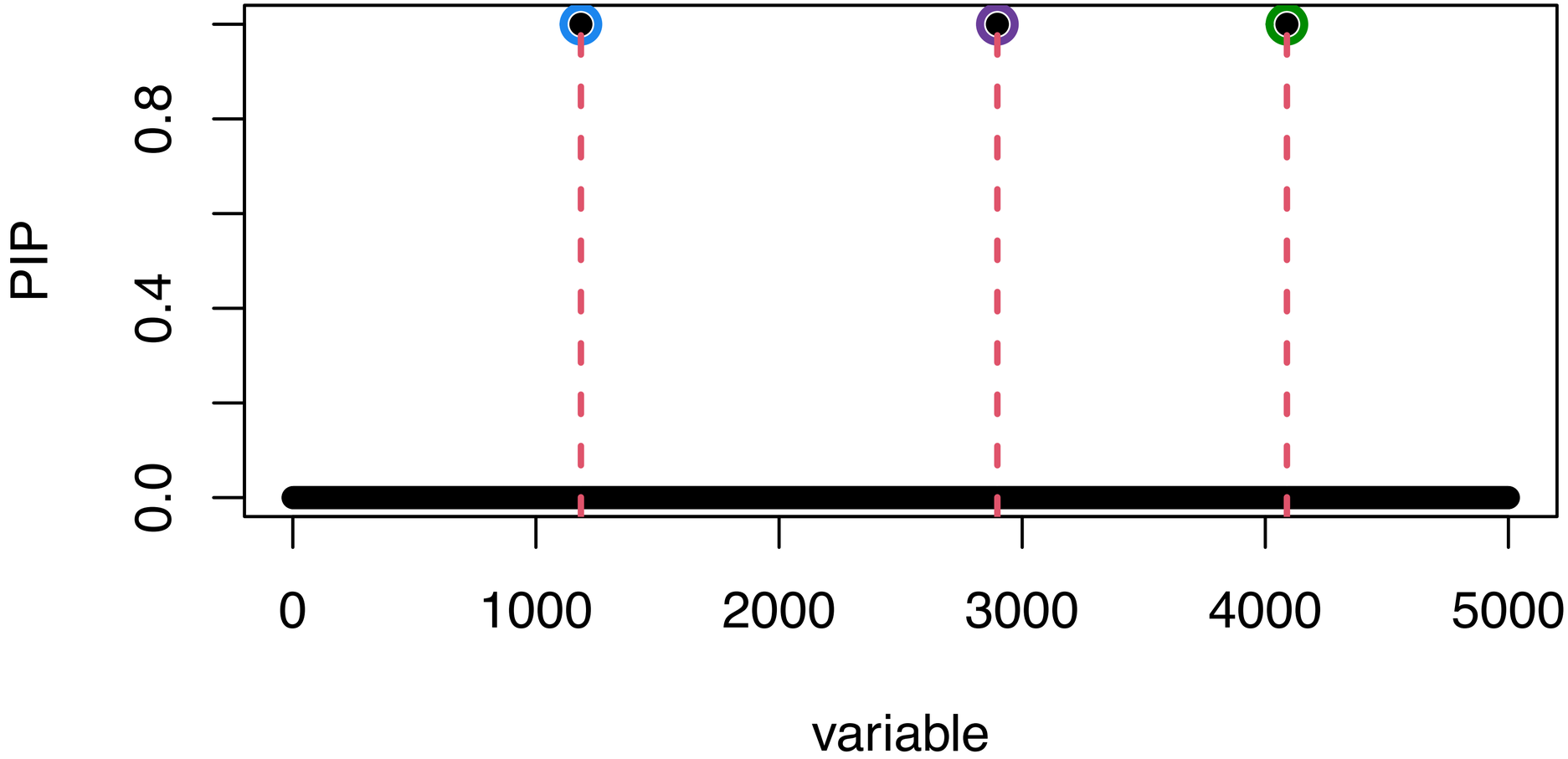
Let's say that we only have 300 individuals available independent of the full 2190 GWAS samples.

What do we have to carry out summary-based inference?

- $\mathbf{z}$: $p \times 1$ z-score vector
- $\hat{X}$: $m \times p$ genotype matrix $m \ll n$
- Typically we have $\hat{R}$: $X^\top X/m$ LD matrix
- If $\hat{X}$ is a submatrix of X, in-sample $\hat{R}$ LD
- Otherwise, we can use an LD matrix computed in some reference cohorts (e.g., 1000G, UK Biobank)

What if we had “true” phenotype values?

We could have run `susie.out <- susie(xx.sub, y.true)`



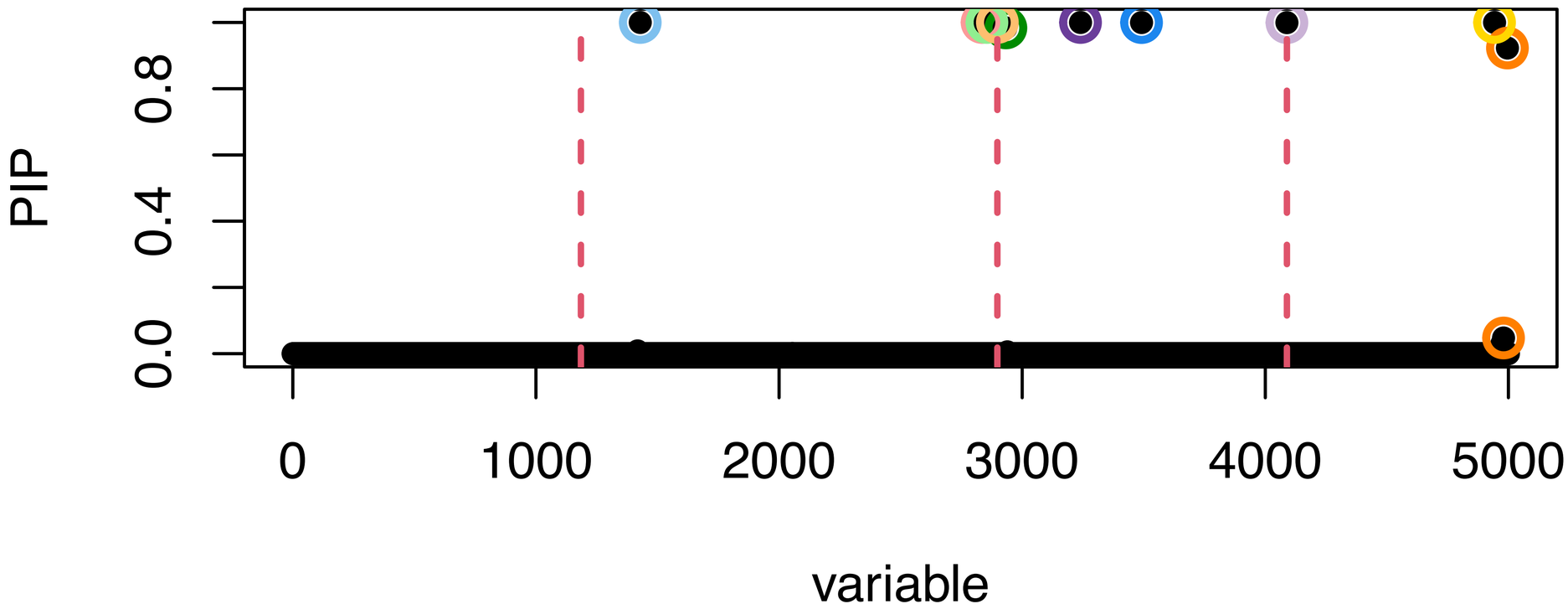
SuSiE-RSS (regression with summary statistics)

$$\mathbf{z} = \mathbf{R}\frac{\sqrt{n}}{\sigma}\theta + \tilde{\epsilon}, \quad \tilde{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$$

Zou, Carbonetto, Wang, Stephens, *PLoS Genetics* (2022)

SuSiE-RSS may produce so many false positives...

```
susie.rss <- susie_rss(z, R=cov(xx.sub), n=nrow(xx))
```



Can we simply predict unobserved Y?

```r
library(rsvd) # faster than svd

## [U, D, V'] = svd(X/sqrt(n))
nn <- nrow(xx.sub)
.svd <- rsvd(xx.sub / sqrt(nn))

## U * inv(D) * V'
U <- .svd$u
Vt <- t(.svd$v)
D <- .svd$d + .01
proj <-
    sweep(U,2,D,`/`) %*% Vt
y.hat <- proj %*% z
```

Polygenic risk prediction:

1. Roughly estimate parameter θ

$$\mathbf{z} \sim \mathcal{N}\left(\hat{R}\theta, \hat{R}\right)$$

$$\theta \sim \mathcal{N}\left(\hat{R}^{-1}\mathbf{z}, \hat{R}^{-1}\right)$$

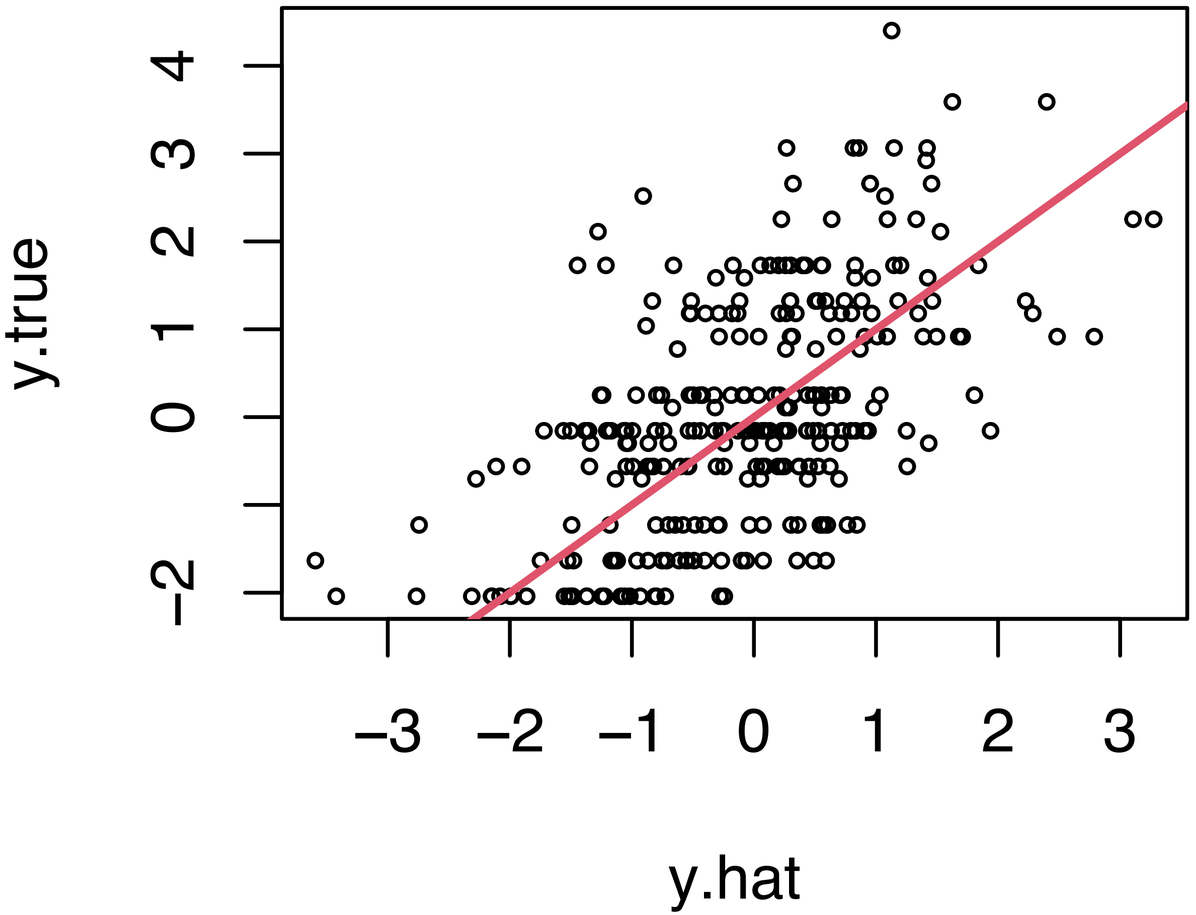
2. We can then predict/project the effects

$$\mathbf{y} \leftarrow \hat{X}\hat{\theta}$$

Here, we used the pseudo-inverse of the sample covariance matrix $\hat{R}^{-1} = VD^{-2}V^{\top}$ followed by SVD, $[U, D, V^{\top}] = \mathrm{svd}(n^{-1/2}X)$. For matrix algebra, refer to Matrix Cookbook.

Can we simply predict unobserved Y?

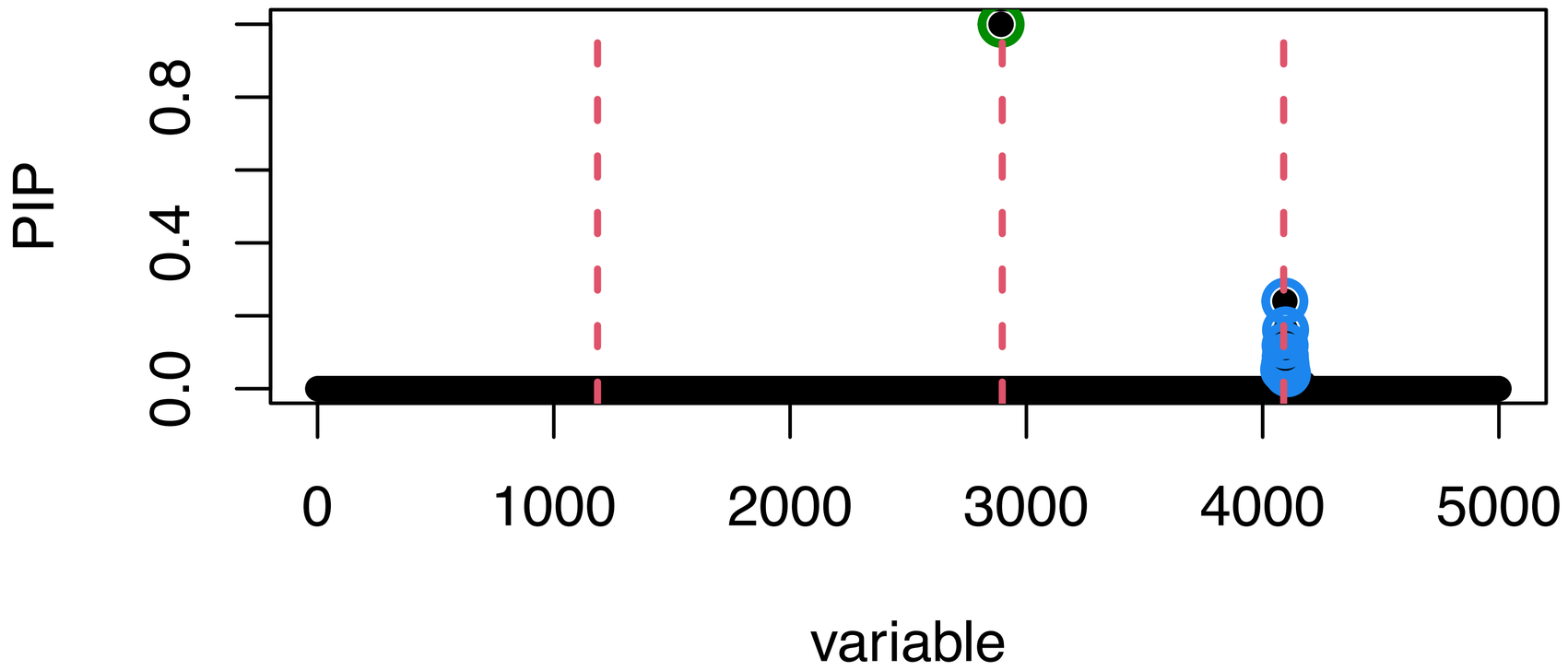
```
library(rsvd) # faster than svd

## [U, D, V'] = svd(X/sqrt(n))
nn <- nrow(xx.sub)
.svd <- rsvd(xx.sub / sqrt(nn))

## U * inv(D) * V'
U <- .svd$u
Vt <- t(.svd$v)
D <- .svd$d + .01
proj <-
    sweep(U,2,D,`/`) %*% Vt
y.hat <- proj %*% z
```



Here, we used the pseudo-inverse of the sample covariance matrix $\hat{R}^{-1} = VD^{-2}V^\top$ followed by SVD, $[U, D, V^\top] = \mathrm{svd}(n^{-1/2}X)$. For matrix algebra, refer to Matrix Cookbook.

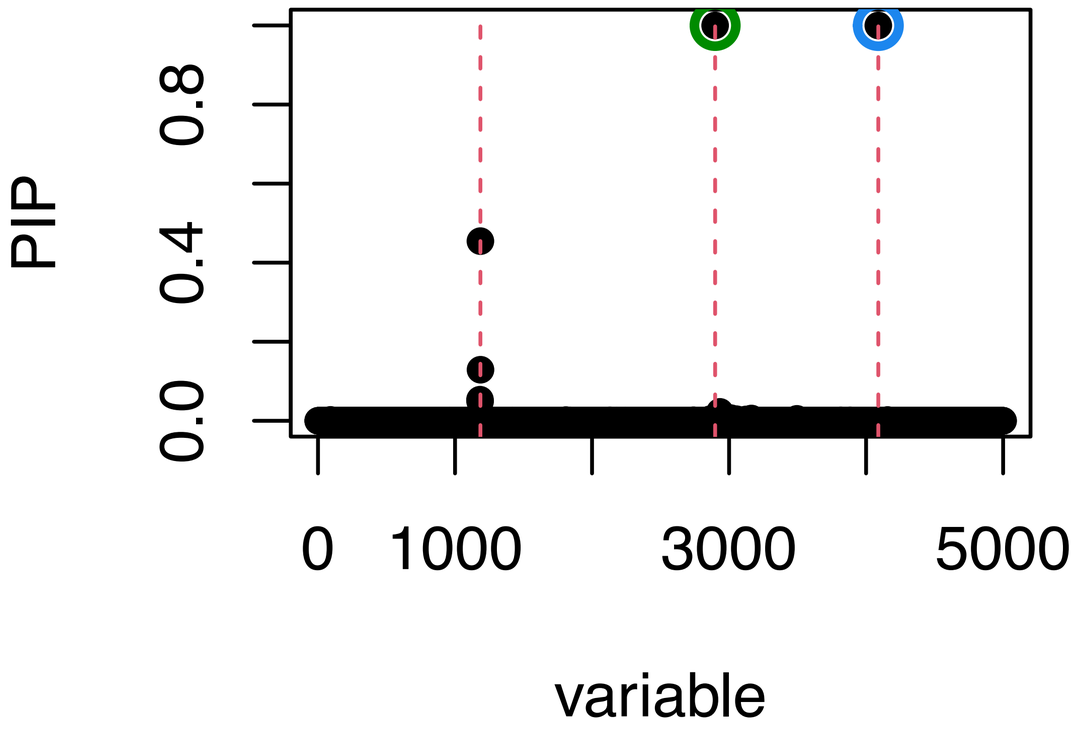
We can run the same susie method as if we observed phenotypes

```
susie.prs <- susie(xx.sub, y.hat)
```

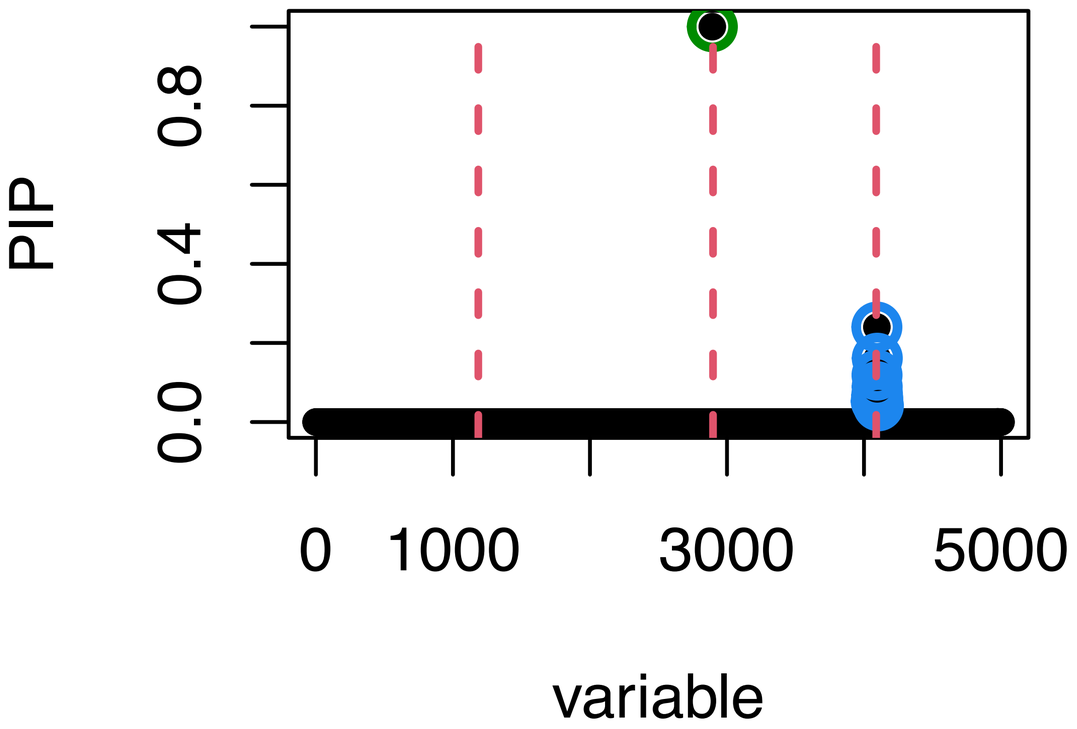


Much fewer false discoveries and competitive power

Full data analysis



Summary-based with out-of-sample LD



Today's lecture: GWAS and related topics

- **Human Genetics 101**
  - Variation in the human genome
  - How do we measure genetic associations?
- **Polygenic models**
  - Population structures
  - Linear Mixed Effect Model
- **Systems Genetics**
  - Summary-based GWAS analysis + polygenic risk prediction
  - LD-score regression: "enrichment analysis" in GWAS

LD score (LDSC) regression to model χ^2 statistics

What is a generative model for a χ^2_j $(= Z^2_j)$ statistics vector?

We have seen this relationship in the fine-mapping model:

$$\underset{\text{univariate, summary stat}}{Z_j} = \frac{\sqrt{n}}{\sigma} \sum_k \underset{\text{LD between j and k}}{R_{jk}} \underset{\text{multivariate, true effect}}{\theta_k} + \epsilon_j$$

where $\epsilon \sim \mathcal{N}(0, 1)$.

Bulik-Sullivan *et al.*, *Nature Genetics* (2014); Finucane *et al.*, *Nature Genetics* (2015)

LD score (LDSC) regression to model χ^2 statistics

What is a generative model for a χ_j^2 $(= Z_j^2)$ statistics vector?

Simply plugging Z_j in the equation,

$$\mathbb{E}\left[\chi_j^2\right] = \mathbb{E}\left[Z_j^2\right] = \mathbb{E}\left(\sqrt{n}\sum_k R_{jk}\theta_k + \epsilon_j\right)^2$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014); Finucane *et al.*, *Nature Genetics* (2015)

LD score (LDSC) regression to model χ^2 statistics

What is a generative model for a χ^2_j $(= Z^2_j)$ statistics vector?

Simply plugging Z_j in the equation,

$$\mathbb{E}\left[\chi^2_j\right] = \mathbb{E}\left[Z^2_j\right] = \mathbb{E}\left(\sqrt{n}\sum_k R_{jk}\theta_k + \epsilon_j\right)^2$$

If "true" multivariate effect for each variant is independent of other variants' effects, i.e., $\mathbb{E}\left[\theta_k\theta_j\right] = 0$ for all $k \neq j$,

$$\mathbb{E}\left[\chi^2_j\right] = n\underbrace{\sum_k R^2_{jk}}_{\text{LD-score}}\mathbb{E}\left[\theta^2_k\right] + 1$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014); Finucane *et al.*, *Nature Genetics* (2015)

Baseline LD-score regression to measure polygenic heritability

(1) Assuming that all the variants equally contribute,

$$\mathbb{E}[\theta_k^2] = \tau/p,$$

where p is the total number of SNPs,

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Baseline LD-score regression to measure polygenic heritability

(1) Assuming that all the variants equally contribute,

$$\mathbb{E}[\theta_k^2] = \tau/p,$$

(2) defining an LD score for a variant/SNP j as

$$l_j \overset{\text{def}}{=} \sum_k R_{jk}^2,$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Baseline LD-score regression to measure polygenic heritability

(1) Assuming that all the variants equally contribute,

$$\mathbb{E}[\theta_k^2] = \tau/p,$$

(2) defining an LD score for a variant/SNP j as

$$l_j \stackrel{\text{def}}{=} \sum_k R_{jk}^2,$$

We get

$$\mathbb{E}\left[\chi_j^2\right] = n \underbrace{\sum_k R_{jk}^2}_{\text{LD-score}} \mathbb{E}[\theta_k^2] + 1$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Baseline LD-score regression to measure polygenic heritability

(1) Assuming that all the variants equally contribute,

$$\mathbb{E}[\theta_k^2] = \tau/p,$$

(2) defining an LD score for a variant/SNP j as

$$l_j \overset{\text{def}}{=} \sum_k R_{jk}^2,$$

We get

$$\mathbb{E}\left[\chi_j^2\right] = n \underbrace{\sum_k R_{jk}^2}_{\text{LD-score}} \mathbb{E}[\theta_k^2] + 1 = \underset{\text{sample size}}{n} \underset{\text{LD score}}{l_j} \underset{\text{per SNP heritability}}{\frac{\tau}{p}} + 1$$

where p is the total number of SNPs.

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Baseline LD-score regression to measure polygenic heritability

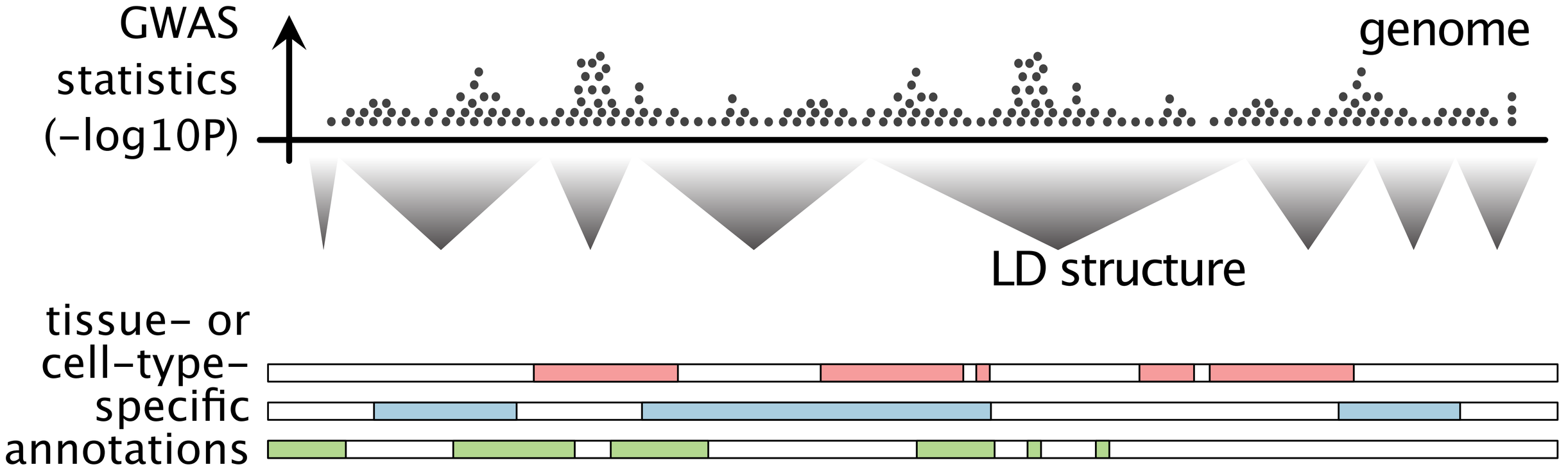
We can treat the relationships as a regression model and find the heritability parameters by regressing the observed χ^2 statistics on the reference LD scores l_j:

$$\begin{pmatrix} \chi_1^2 \\ \vdots \\ \chi_j^2 \\ \vdots \end{pmatrix} \sim \frac{n}{p} \begin{pmatrix} l_1 \\ \vdots \\ l_j \\ \vdots \end{pmatrix} \underbrace{\tau}_{\text{per SNP heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

If the intercept of $\{\chi_j^2\}$ deviate from 1, we can interpret that the GWAS statistics are inflated by some unadjusted population structures or other confounding factors.

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Stratified LD-score regression partitions total heritability into multiple genomic annotations



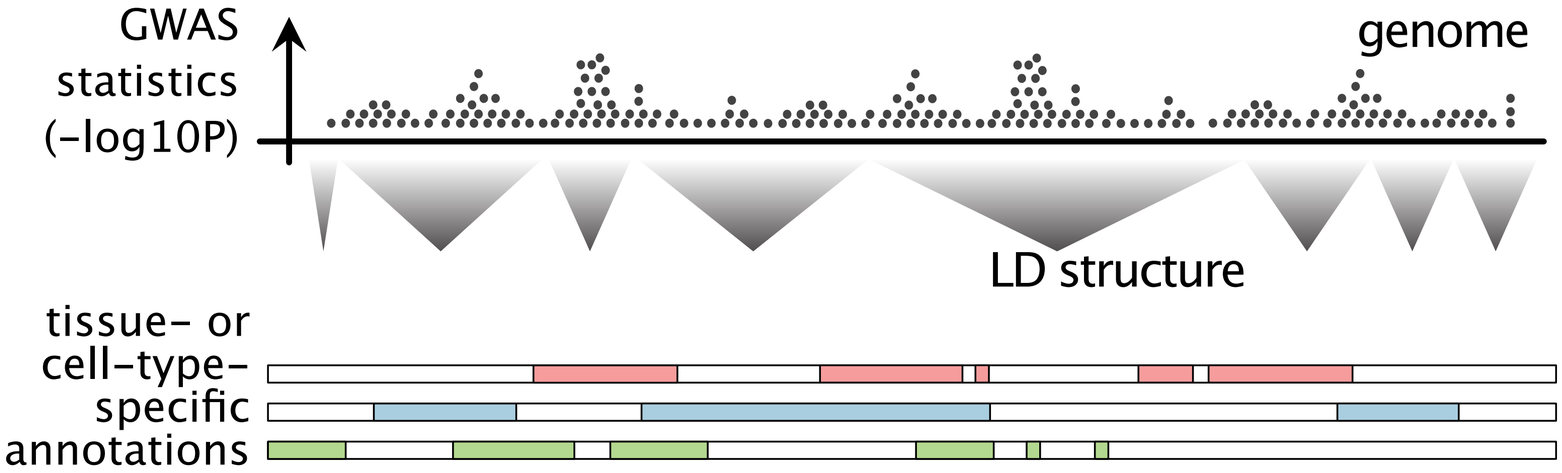
Stratified LD-score regression in math

When genome is partitioned by annotations (e.g., epigenetic tracks)

$$\mathbb{E}\left[\chi_j^2\right] = \frac{n}{p}\sum_t l_{jt}\underbrace{\tau_t}_{\text{stratified heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

where we use partitioned LD-scores for each annotation type t

$$l_{jt} = \sum_k R_{jk}^2 I\left\{k \in \mathcal{A}_t\right\}.$$



Finucane *et al.*, *Nature Genetics* (2015)

Stratified LD-score regression in math

When genome is partitioned by annotations (e.g., epigenetic tracks)

$$\mathbb{E}\left[\chi_j^2\right] = \frac{n}{p}\sum_t l_{jt}\underbrace{\tau_t}_{\text{stratified heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

where we use partitioned LD-scores for each annotation type t

$$l_{jt} = \sum_k R_{jk}^2 I\left\{k \in \mathcal{A}_t\right\}.$$

Instead of assuming a single parameter for the overall per-SNP heritability τ, we can "partition" this total heritability into annotation-type-specific ones, $\{\tau_t\}$.

Stratified LD-score regression in math

When genome is partitioned by annotations (e.g., epigenetic tracks)

$$\mathbb{E}\left[\chi_j^2\right] = \frac{n}{p}\sum_t l_{jt}\ \underbrace{\tau_t}_{\text{stratified heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

where we use partitioned LD-scores for each annotation type t

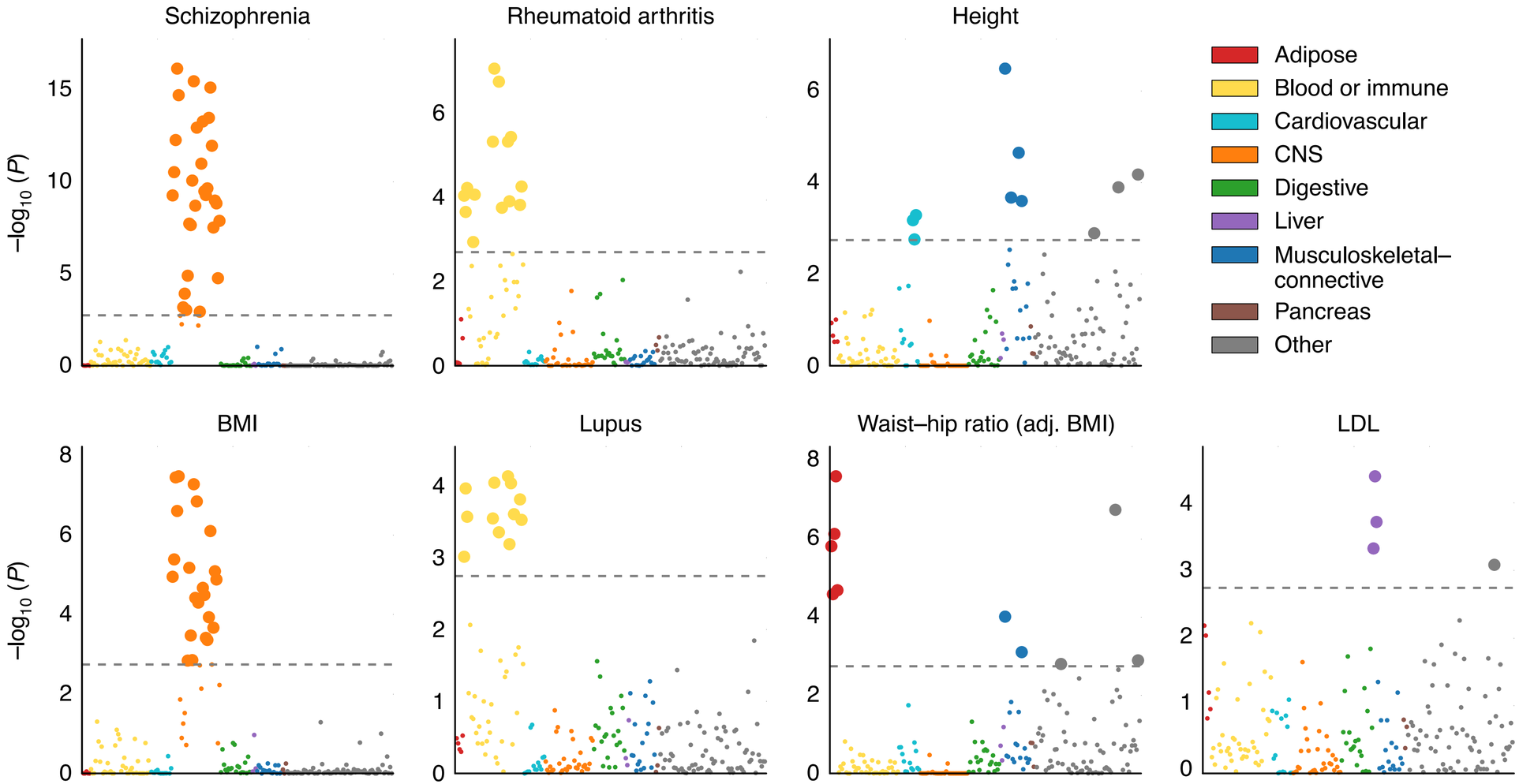
$$l_{jt} = \sum_k R_{jk}^2 I\left\{k \in \mathcal{A}_t\right\}.$$

More explicitly,

$$\begin{pmatrix} \chi_1^2 \\ \vdots \\ \chi_j^2 \\ \vdots \end{pmatrix} \sim \frac{n}{p} \underbrace{\begin{pmatrix} l_{11} & l_{12} & l_{1t} & \cdots \\ & \vdots & & \\ l_{j1} & l_{j2} & l_{jt} & \cdots \\ & \vdots & & \end{pmatrix}}_{\text{stratified LD scores}} \underbrace{\begin{pmatrix} \tau_1 \\ \vdots \\ \tau_t \\ \vdots \end{pmatrix}}_{\text{stratified heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

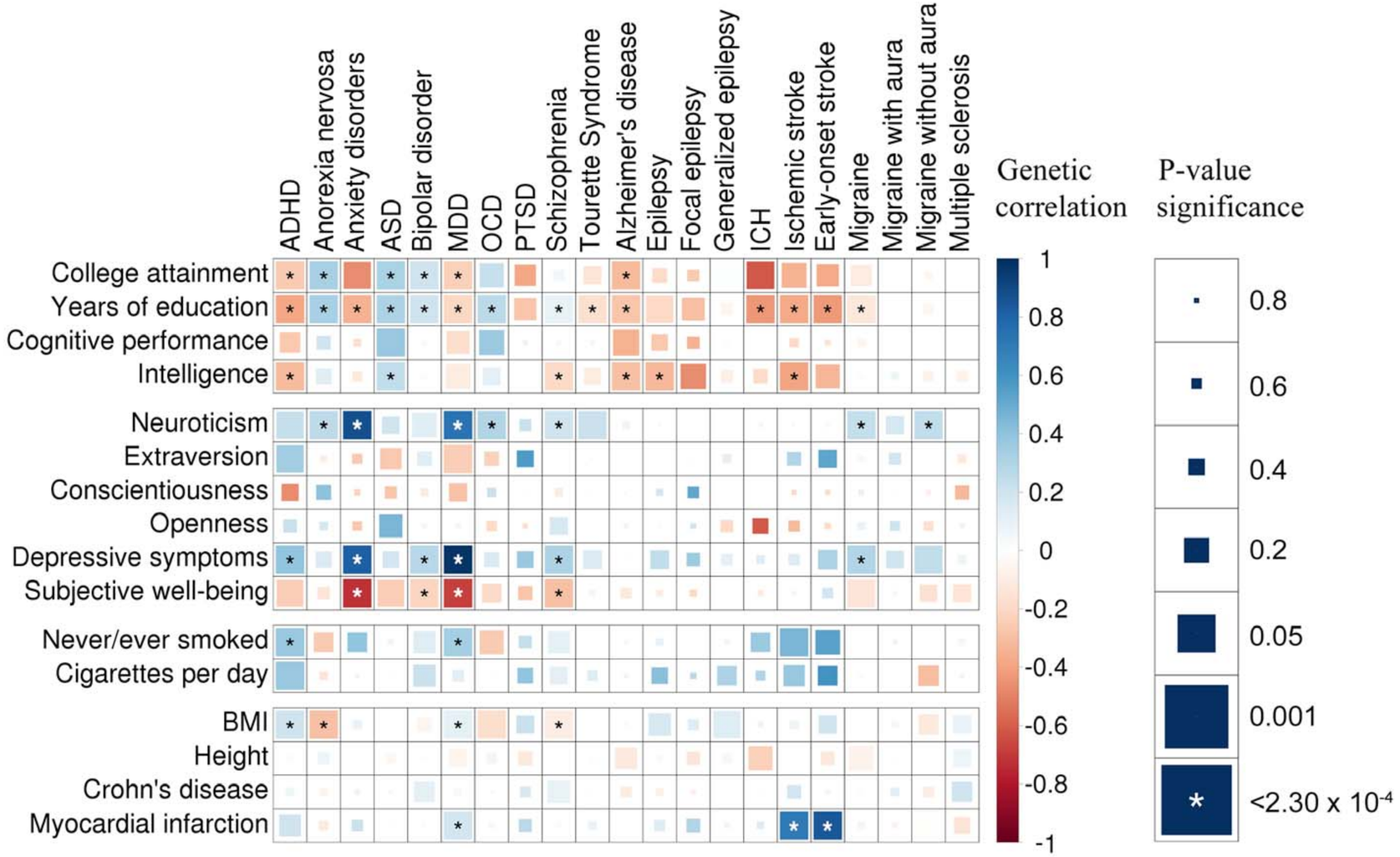
Finucane *et al.*, *Nature Genetics* (2015)

Stratified LDSC can identify tissue-specific enrichment of GWAS signals



Finucane … Prices, *Nature Genetics* (2018)

When multiple GWAS were done, post-GWAS analysis begins



Bivariate LD-score regression

Instead of one χ^2 vector, we need to deal with the element-wise product of two vectors of z-scores (between a trait 1 and 2):

$$\begin{pmatrix} z_1^{(1)} z_1^{(2)} l_1 \\ \vdots \\ z_j^{(1)} z_j^{(2)} l_j \\ \vdots \end{pmatrix} \sim \frac{\sqrt{N_1 N_2}}{p} \begin{pmatrix} l_1 \\ \vdots \\ l_j \\ \vdots \end{pmatrix} \underbrace{\rho}_{\text{genetic correlation}} + \underbrace{\frac{\rho_0 N_s}{\sqrt{N_1 N_2}}}_{\text{sample sharing}}$$

where N_1 and N_2 count sample size of the GWAS 1 and 2; N_s is the number of control individuals shared between the two traits.

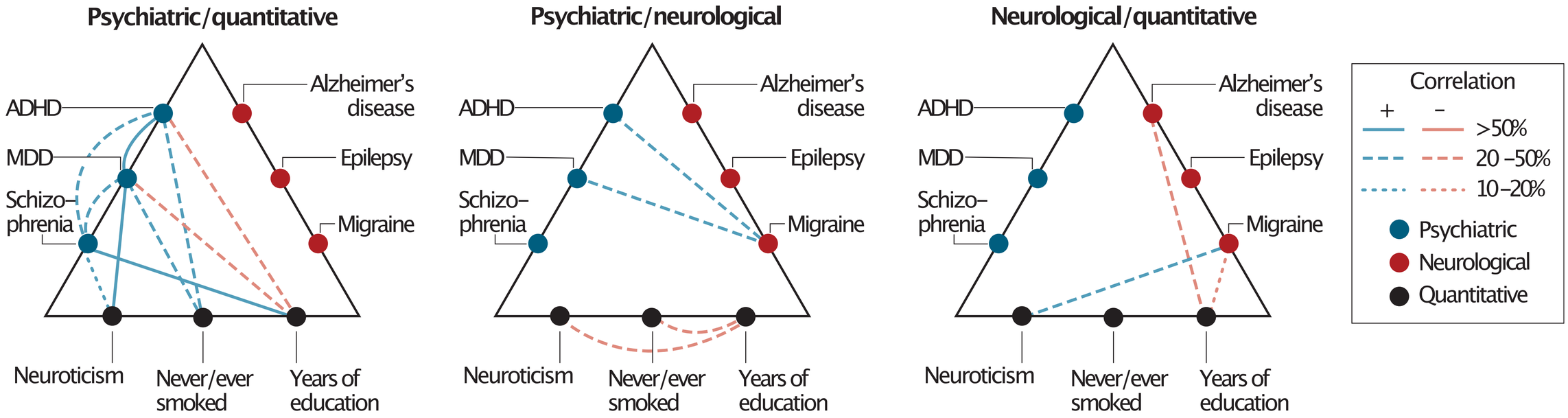
Bulik-Sullivan *et al.*, *Nature Genetics* (2015)

Post-GWAS analysis example: genetic correlations across many traits

Psychiatric disorders				Neurological disorders			
Disorder	**Source**	**Cases**	**Controls**	**Disorder**	**Source**	**Cases**	**Controls**
Attention deficit hyperactivity disorder	PGC-ADD2	12,645	84,435	Alzheimer's disease	IGAP	17,008	37,154
Anorexia nervosa	PGC-ED	3495	10,982	Epilepsy	ILAE	7779	20,439
Anxiety disorders	ANGST	5761	11,765	Focal epilepsy	"	4601*	17,985*
Autism spectrum disorder	PGC-AUT	6197	7377	Generalized epilepsy	"	2525*	16,244*
Bipolar disorder	PGC-BIP2	20,352	31,358	Intracerebral hemorrhage	ISGC	1545	1481
Major depressive disorder	PGC-MDD2	66,358	153,234	Ischemic stroke	METASTROKE	10,307	19,326
Obsessive-compulsive disorder	PGC-OCDTS	2936	7279	Cardioembolic stroke	"	1859*	17,708*
Posttraumatic stress disorder	PGC-PTSD	2424	7113	Early onset stroke	"	3274*	11,012*
Schizophrenia	PGC-SCZ2	33,640	43,456	Large-vessel disease	"	1817*	17,708*
Tourette syndrome	PGC-OCDTS	4220	8994	Small-vessel disease	"	1349*	17,708*
				Migraine	IHGC	59,673	316,078
				Migraine with aura	"	6332*	142,817*
				Migraine without aura	"	8348*	136,758*
				Multiple sclerosis	IMSGC	5545	12,153
				Parkinson's disease	IPDGC	5333	12,019
Total psychiatric		158,028	365,993	Total neurologic		107,190	418,650

The Brainstorm Consortium, *Science* (2018)

Bivariate LDSC reveals disease comorbidity at the common genetic variants' level



The Brainstorm Consortium, *Science* (2018)

Today’s lecture: GWAS and related topics

- **Human Genetics 101**
  - Variation in the human genome
  - How do we measure genetic associations?
- **Polygenic models**
  - Population structures
  - Linear Mixed Effect Model
- **Systems Genetics**
  - Summary-based GWAS analysis + polygenic risk prediction
  - LD-score regression: “enrichment analysis” in GWAS