STAT 540

Genome-wide Association Studies

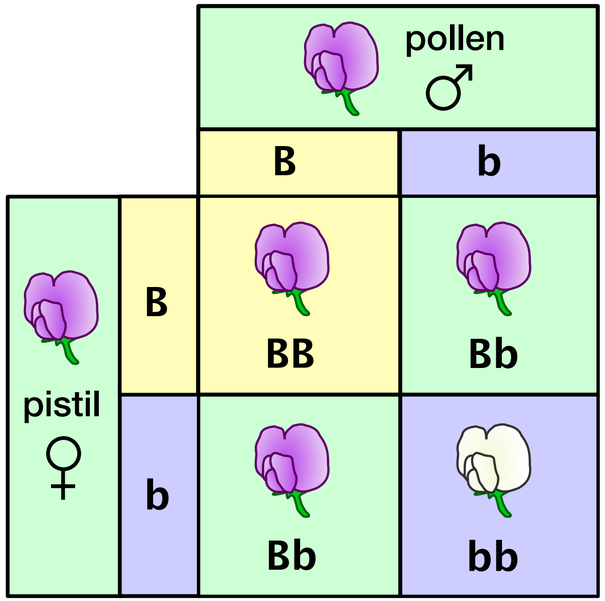
Yongjin Park

a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

Today's lecture

- Fundamentals in Population Genetics
- Genome-wide association studies
- What are the limitations of 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-gen

Human genetics revolution

23andMe

deCODE genetics

CanPath
Canadian Partnership for Tomorrow's Health

Counsyl

biobank uk

ancestry

FIMM

MyHeritage

MVP

BIOBANK JAPAN

All of Us
RESEARCH PROGRAM

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

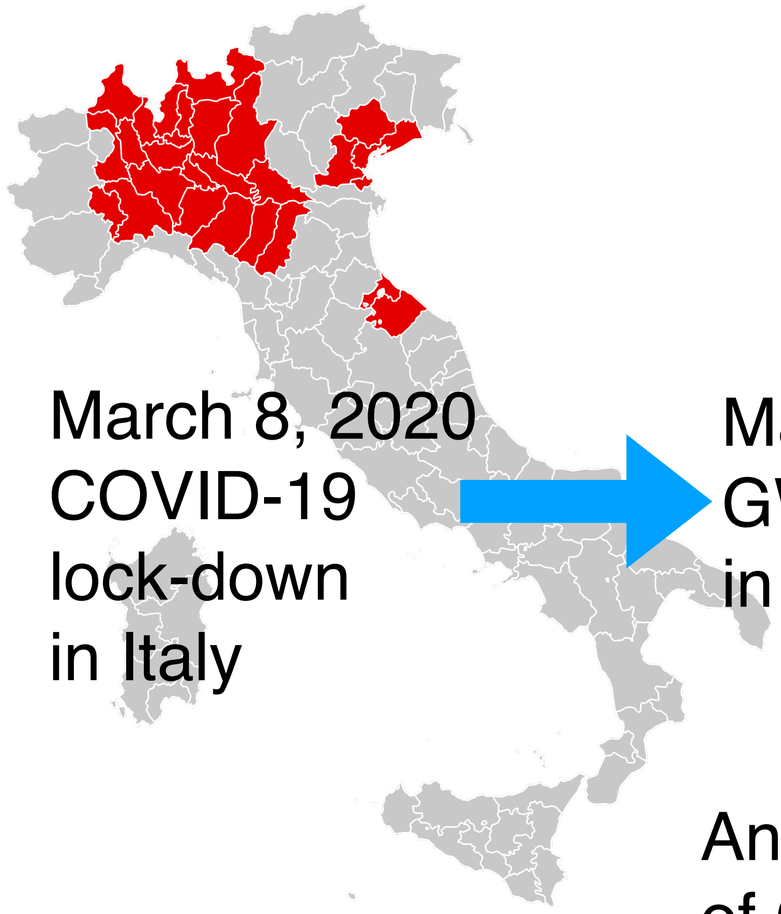
Forensic genealogy helped nab the Golden State Killer in 2018. Now investigators across the country are using it to revisit hundreds of unsolved crimes.

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

Human Genomes help keep track of human evolution

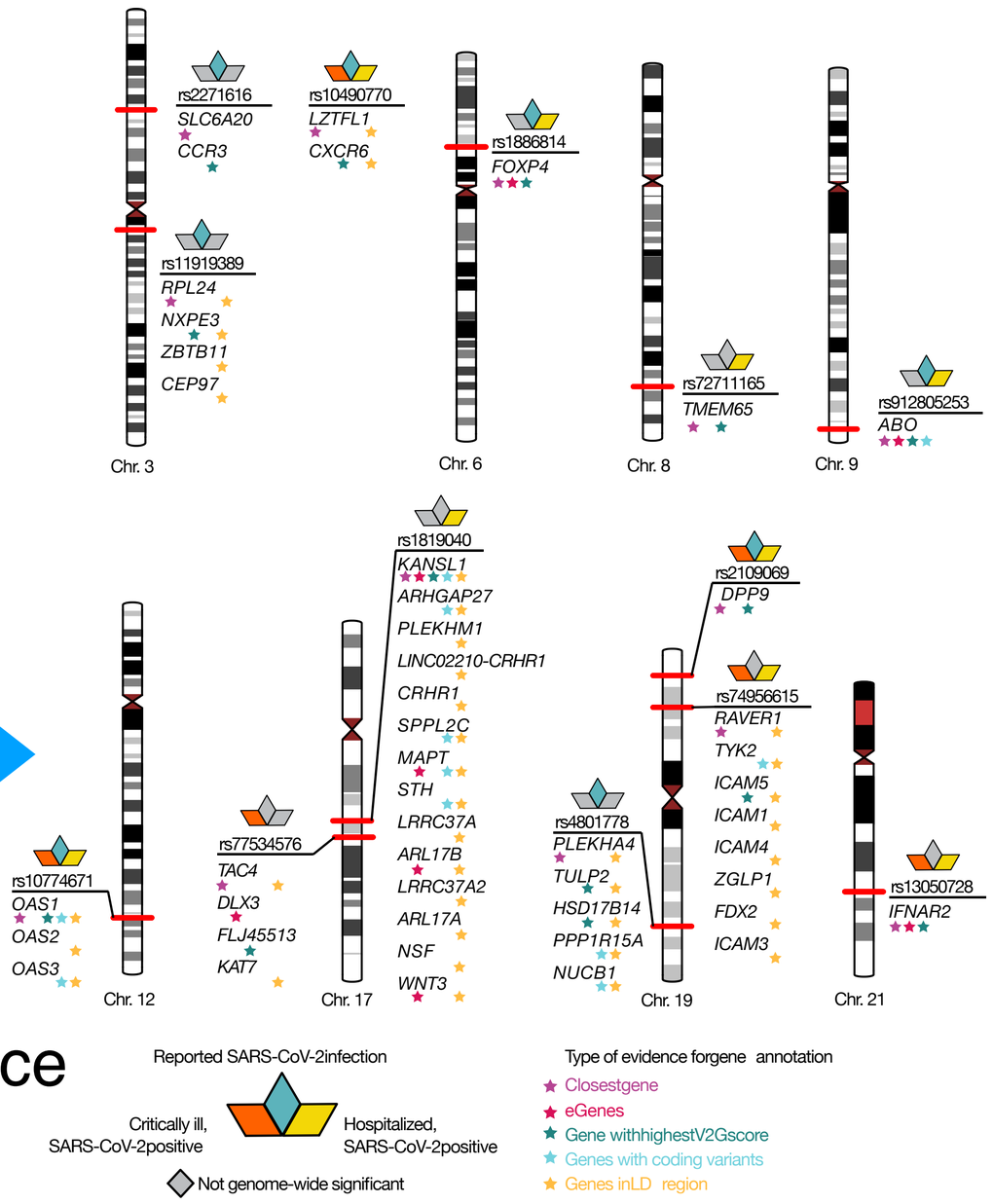
https://www.nytimes.com/2018/03/20/science/david-reich-human-migrations.html

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

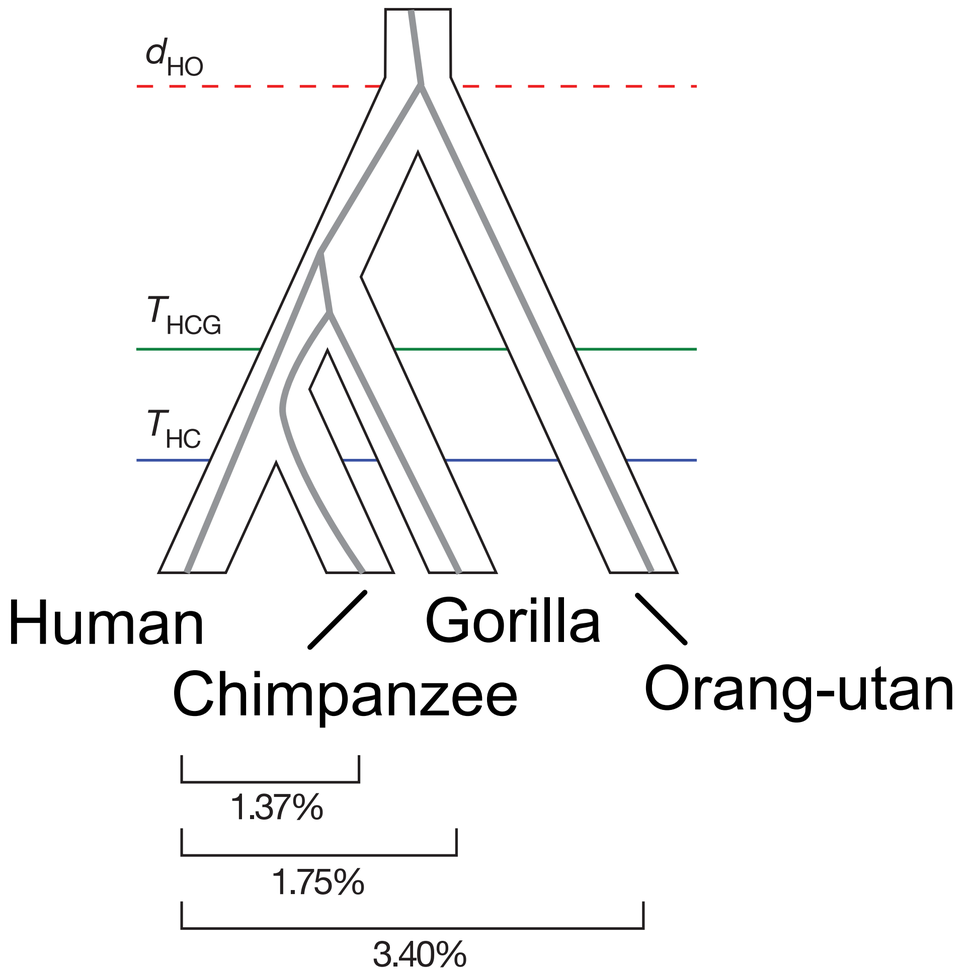
A reference DNA sequence for
3.2B basepairs x diploid for each individual

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

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

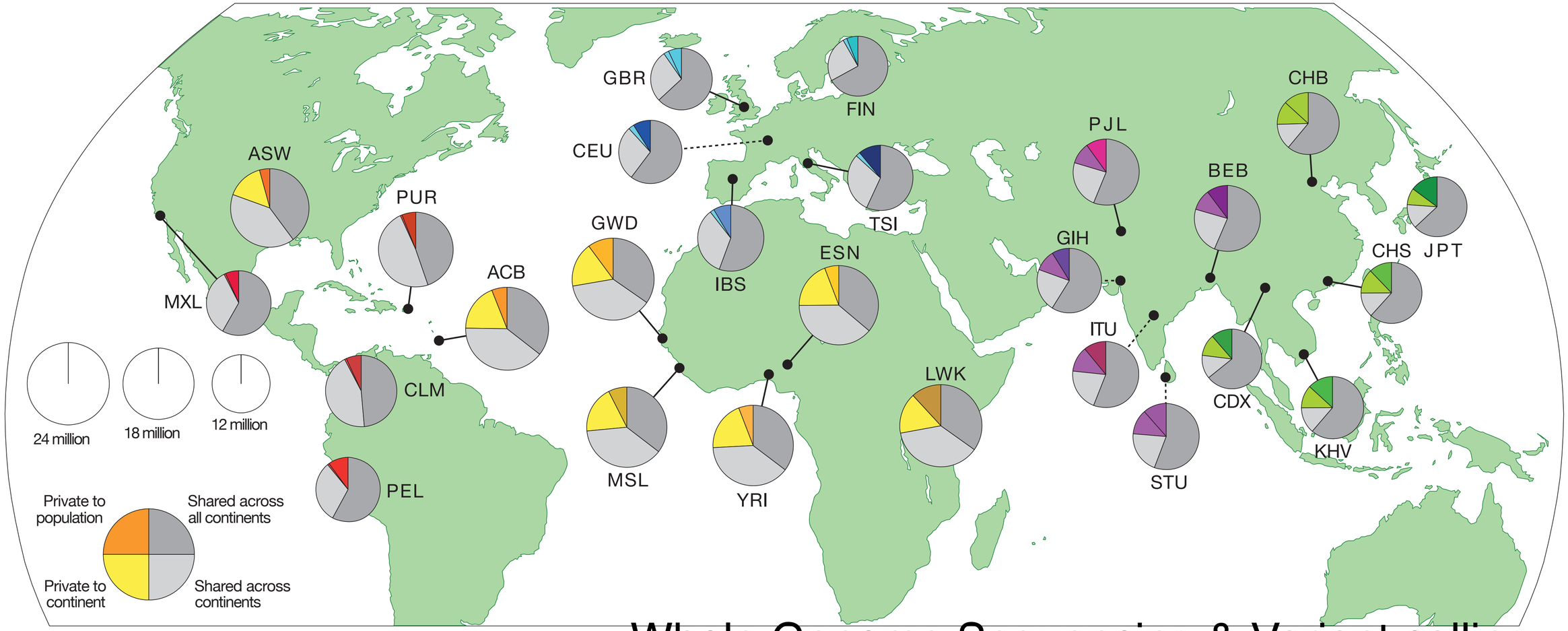
The International HapMap Project

†The International HapMap Consortium

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

80k Accesses | **4231** Citations | **59** Altmetric | Metrics

The 1000 genomes project



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

The 1000 Genomes Project Consortium, Nature, 2015

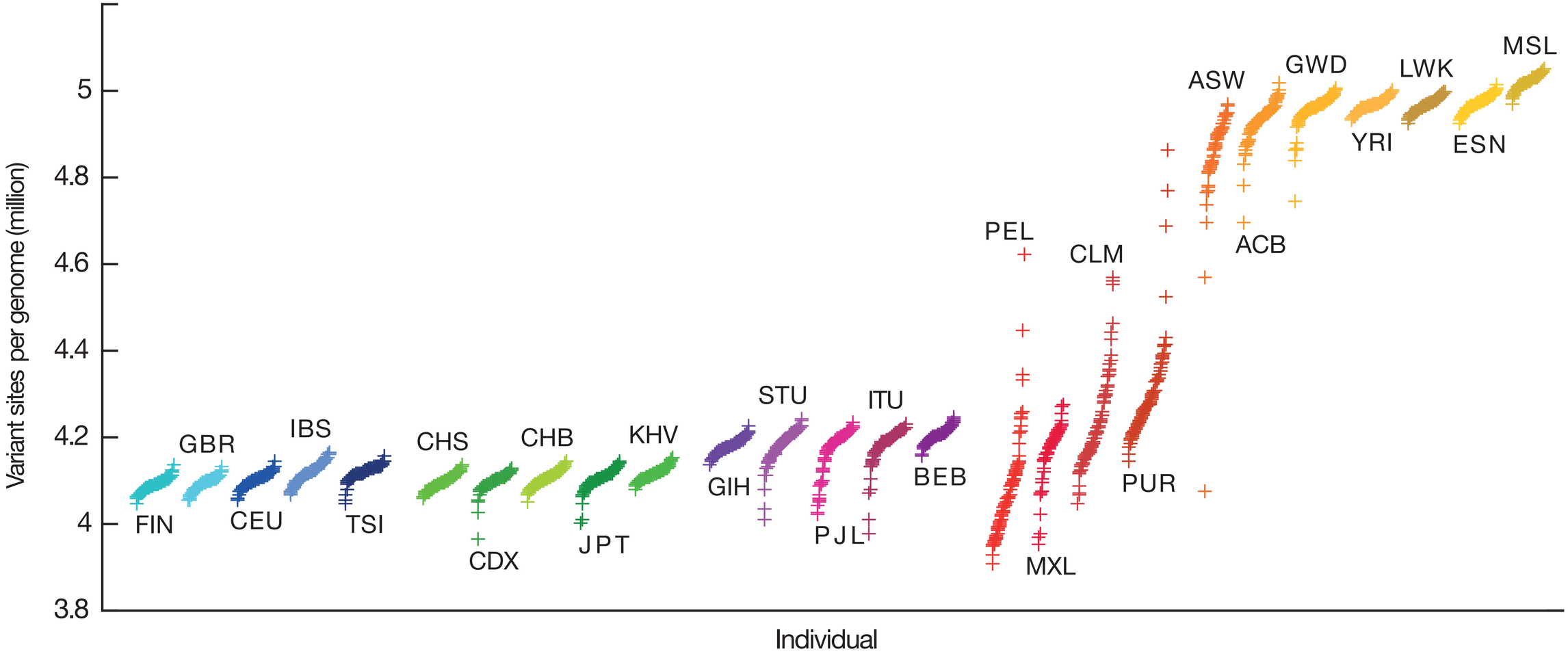
Genetic variation across human population

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

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

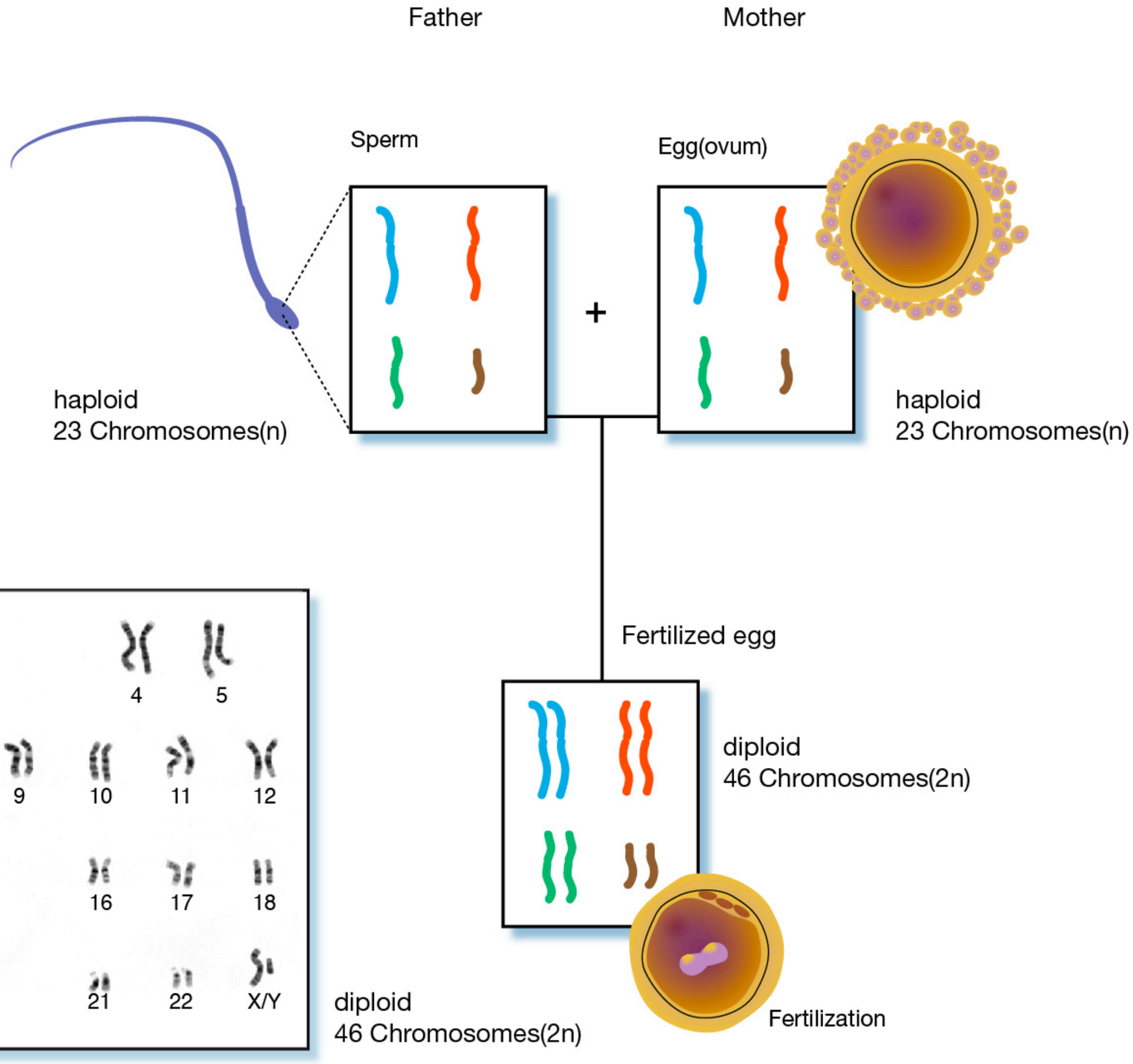
A specific region of the genome differs between two genomes.

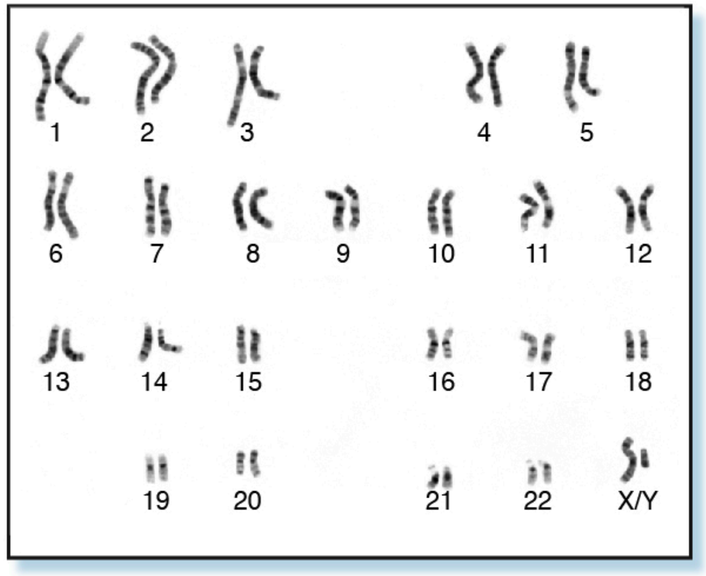
Across two or more genomes, there could have genetic variants due to a mutational process

Ploidy (diploid)

One copy from maternal genome another copy from paternal genome

When do we have polyploidy or aneuploidy?

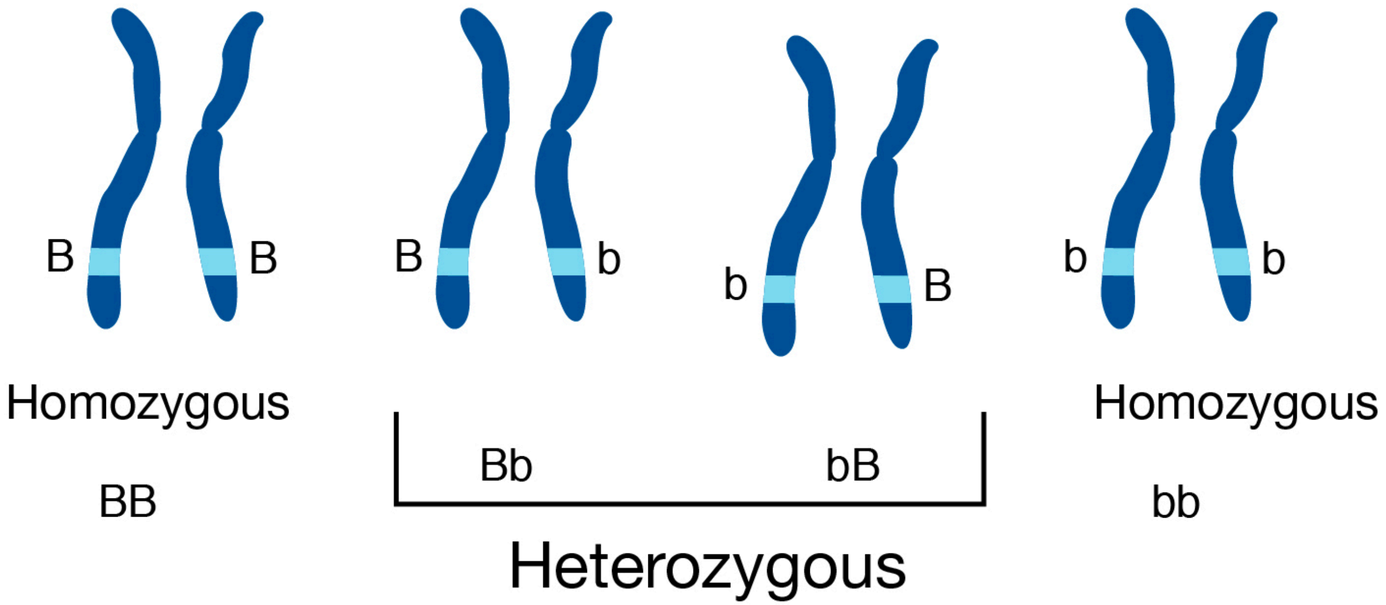




https://www.genome.gov/genetics-glossary/Diploid

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**

Reference	ACTGACGCATGCATCATGCATGC
Insertion	ACTGACGCATGGTACATCATGCATGC
Deletion	ACTGACG--TGCATCATGCATGC

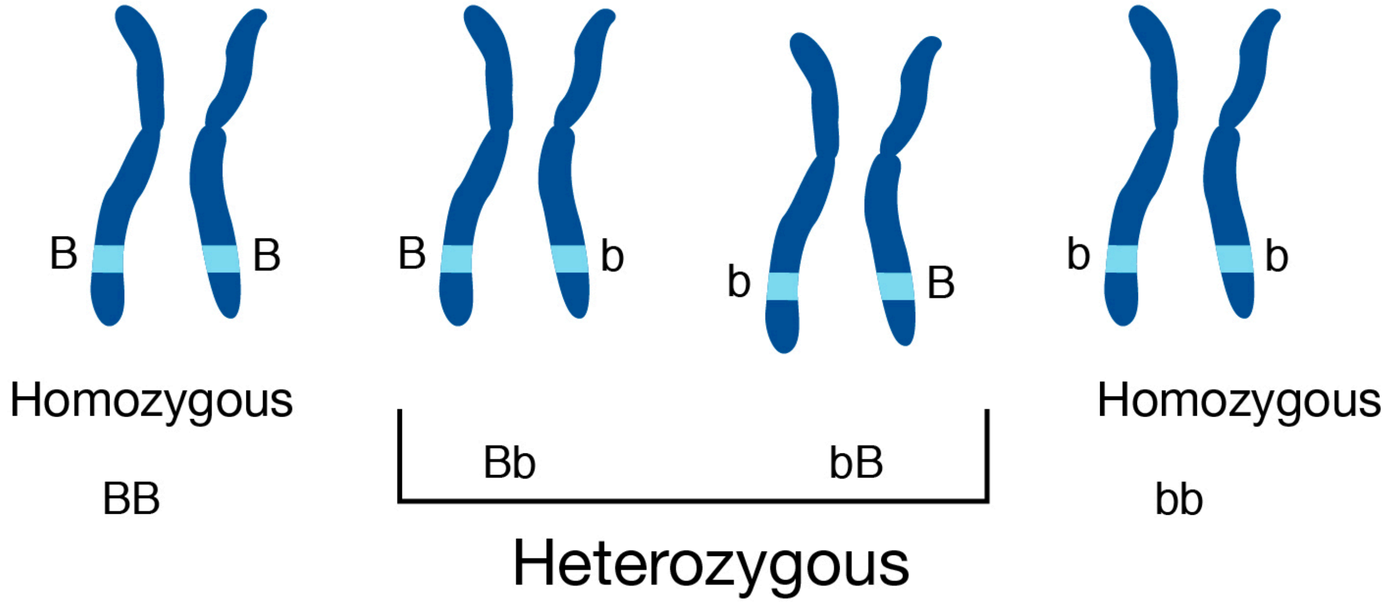
Indel

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

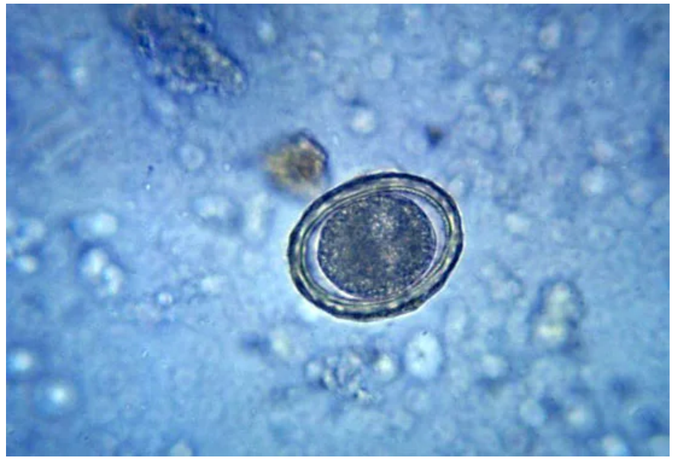
Homozygote vs. heterozygote

One copy from maternal genome
another copy from paternal genome

<u>Homozygous</u>: the maternal == paternal copy
<u>Heterozygous</u>: the maternal != paternal copy



- "Homo" vs. "hetero" $\Longleftrightarrow$ the same vs. different
- Zygote (fertilized egg cell)



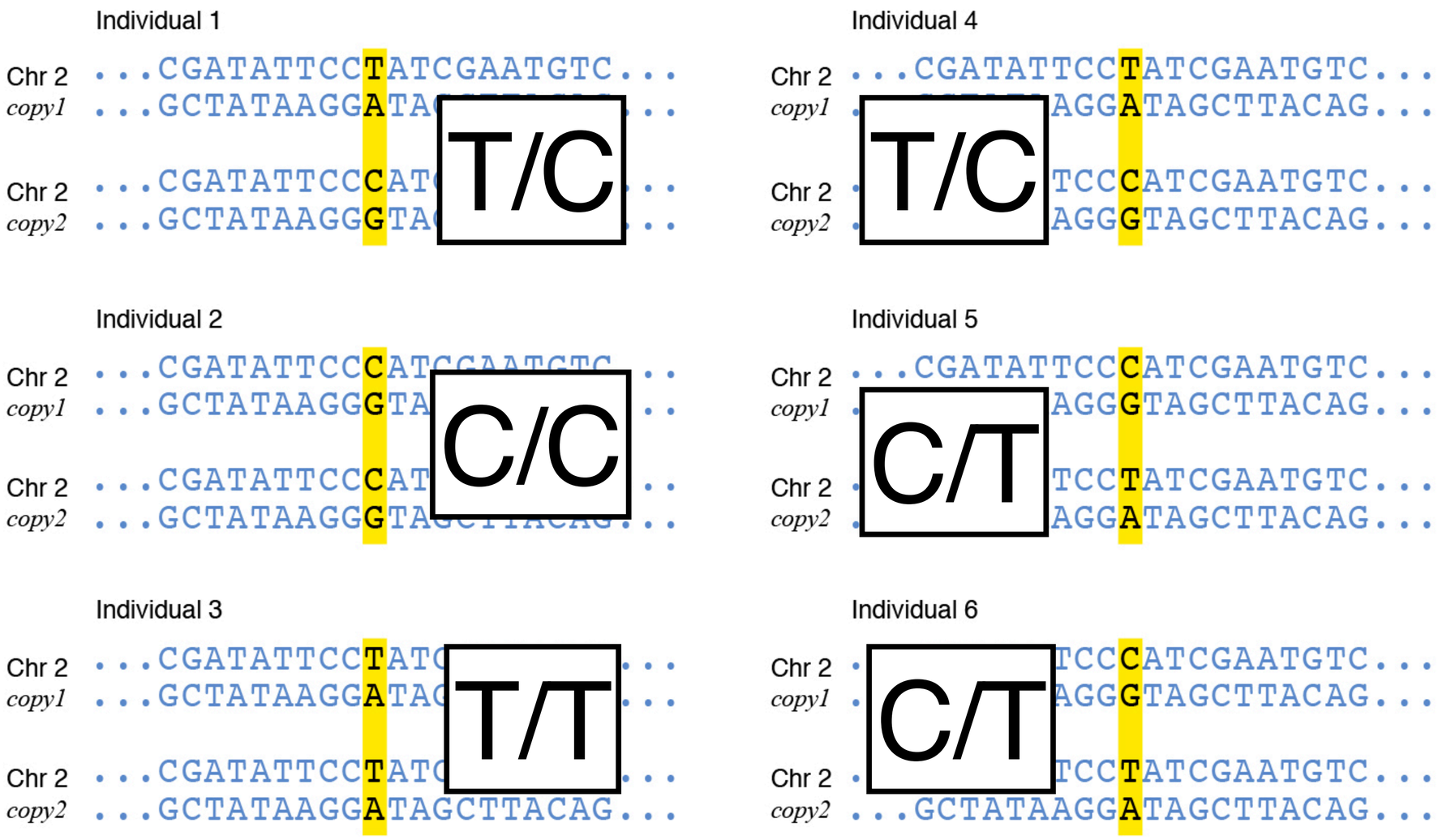
https://www.genome.gov/genetics-glossary/heterozygous https://www.britannica.com/scien

Single Nucleotide Polymorphism

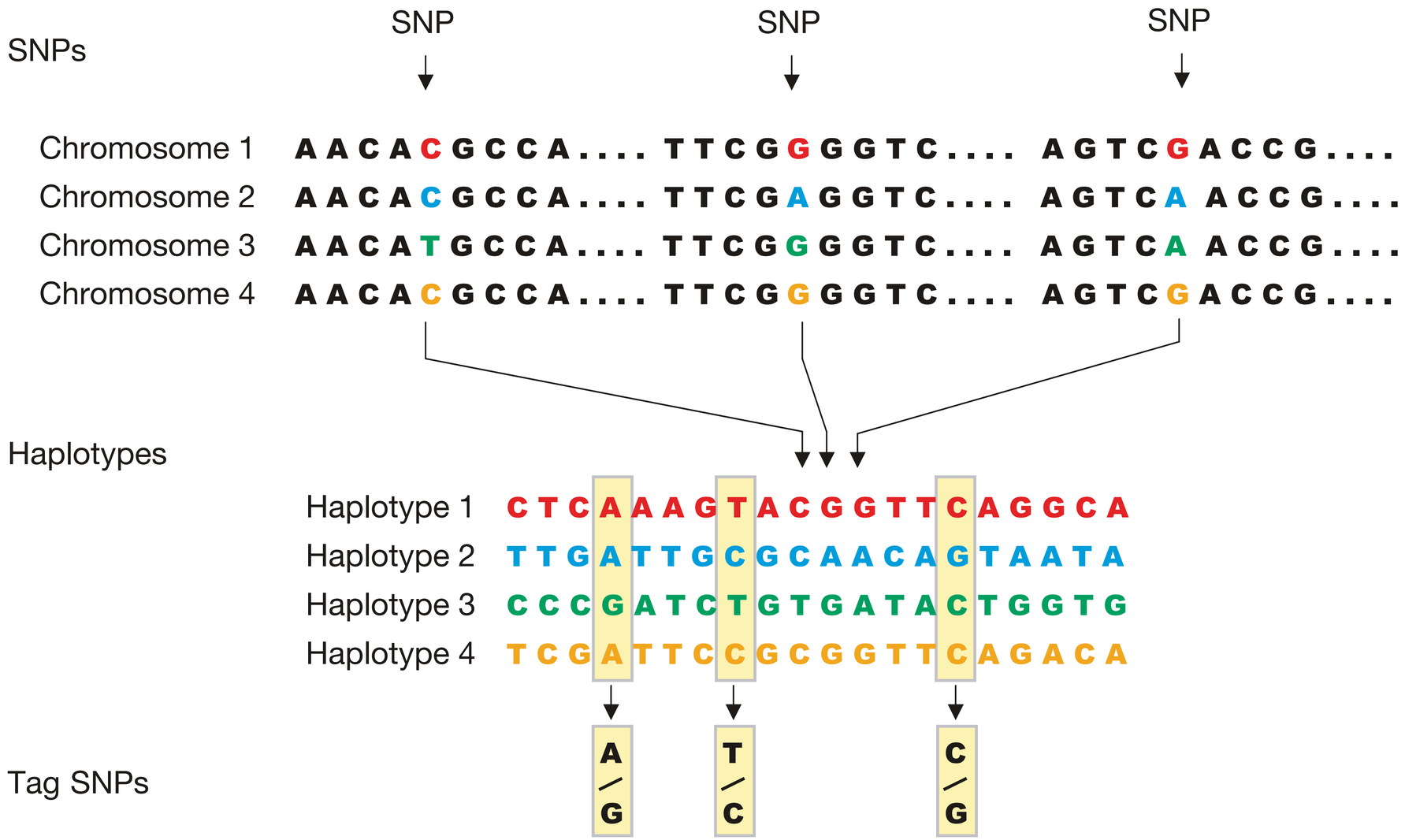
~1 in every 1000bp

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

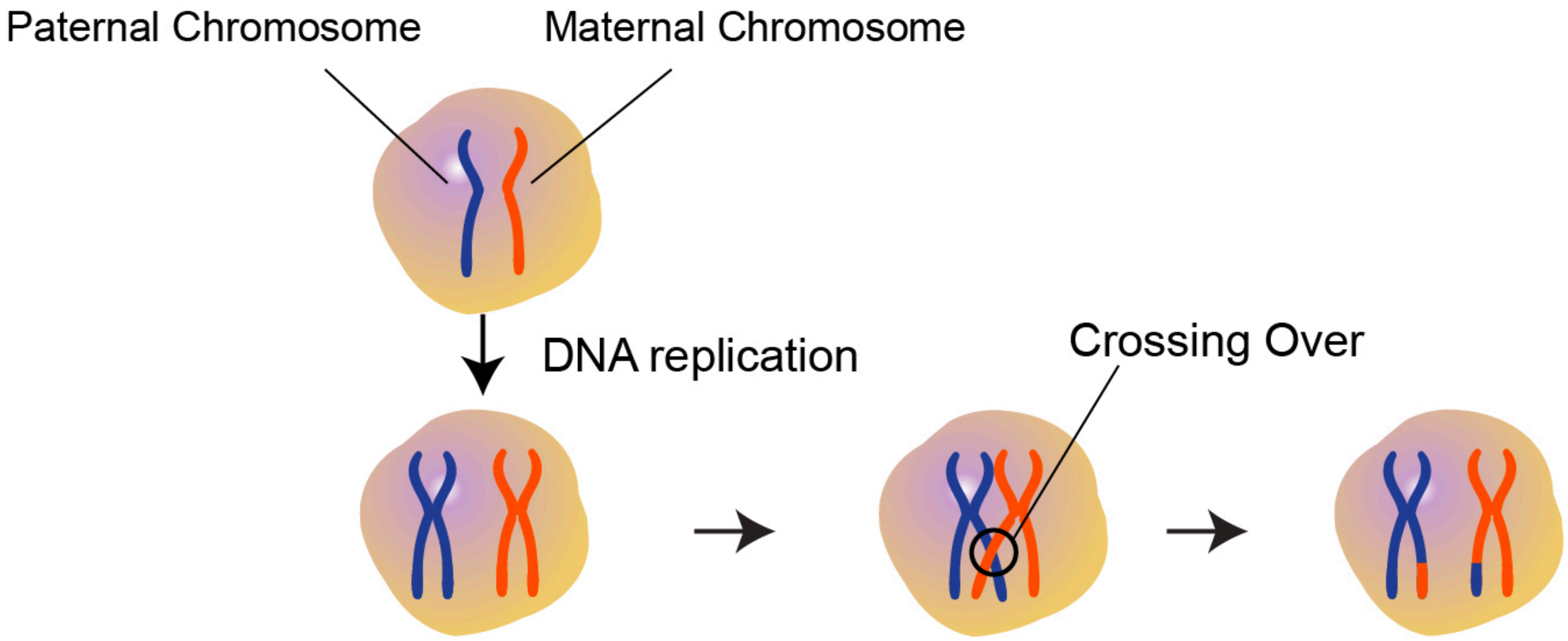


~ 1M common SNPs



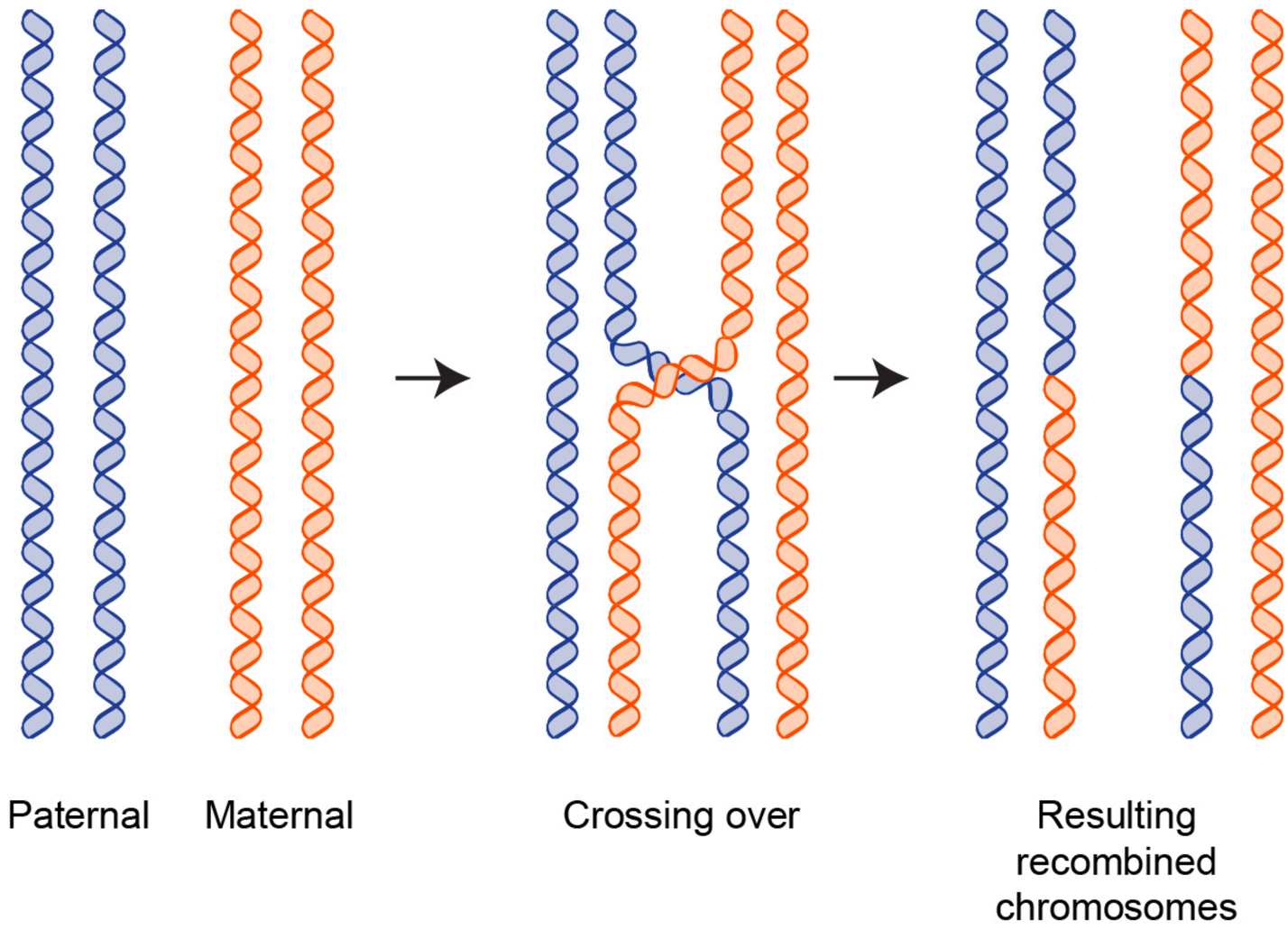
The International HapMap Project (2003)

Recombination: Mixing the maternal and paternal copies



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

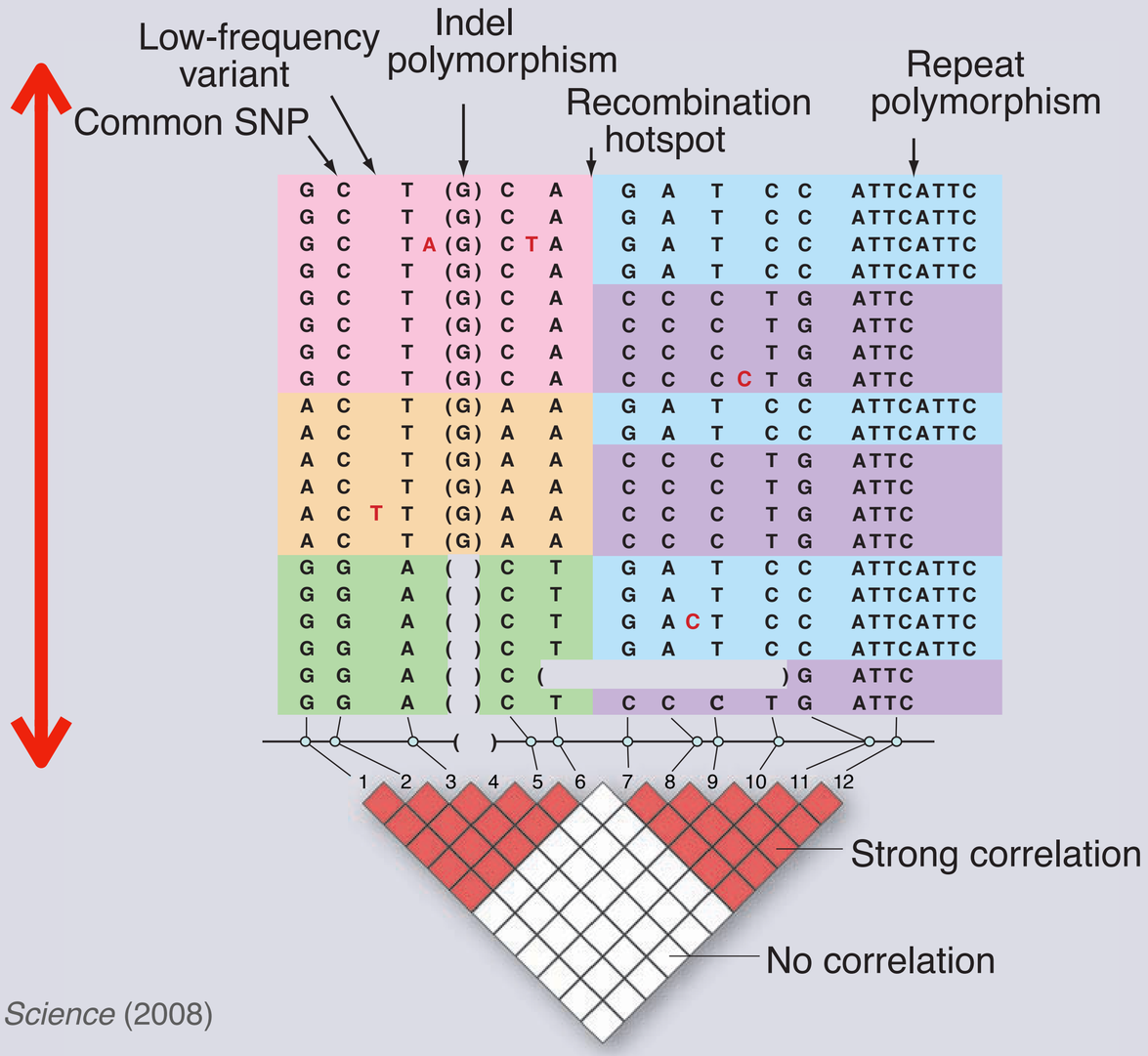
Recombination: Mixing the maternal and paternal copies



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

Genetic variation in one figure

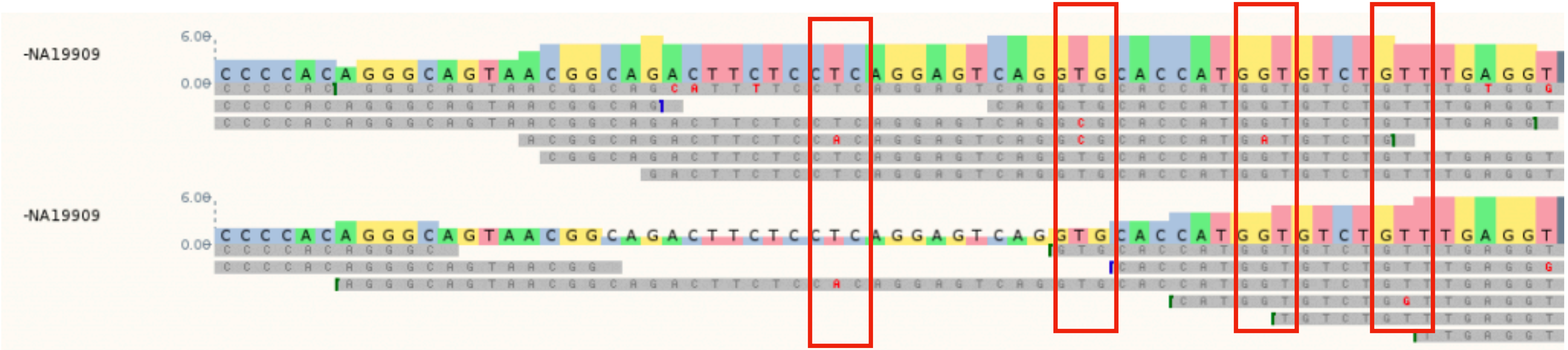
across many individuals (diploid genomes)



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

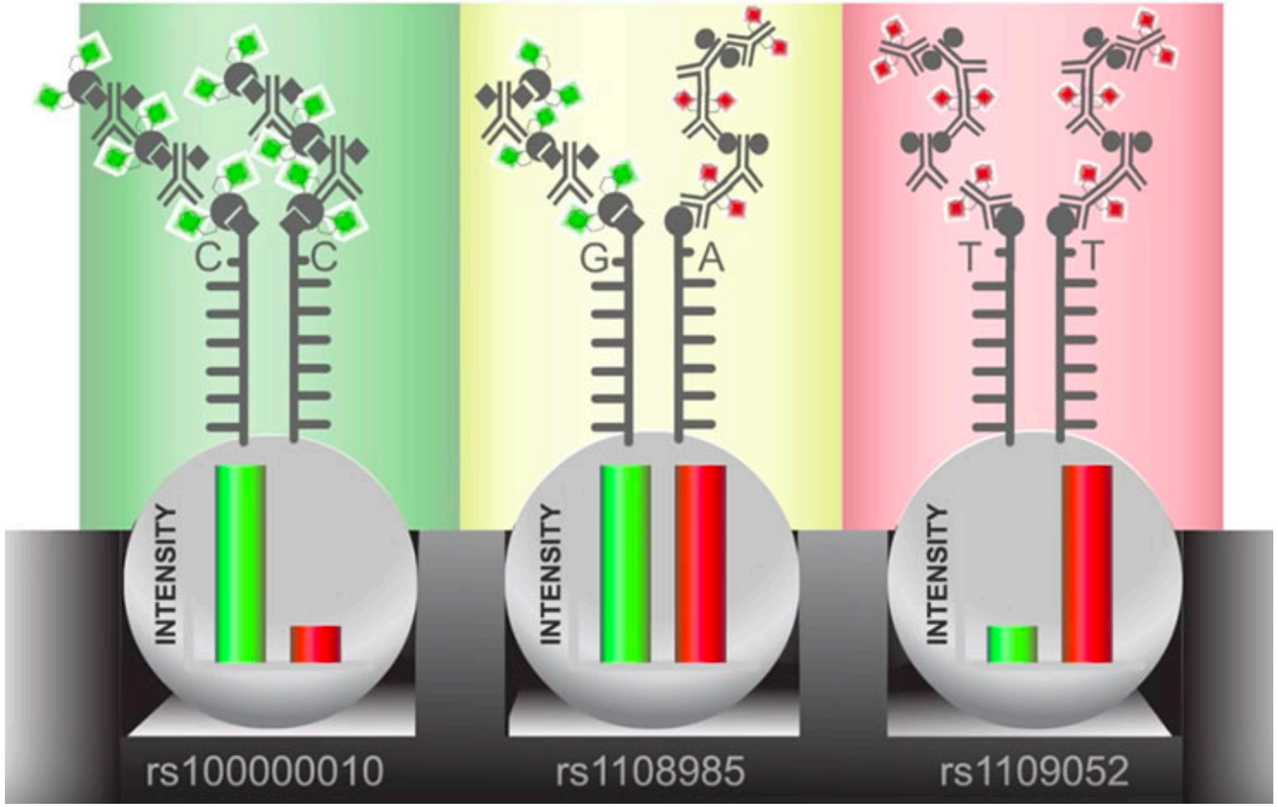
Altshuler, Daly, Lander, *Science* (2008)

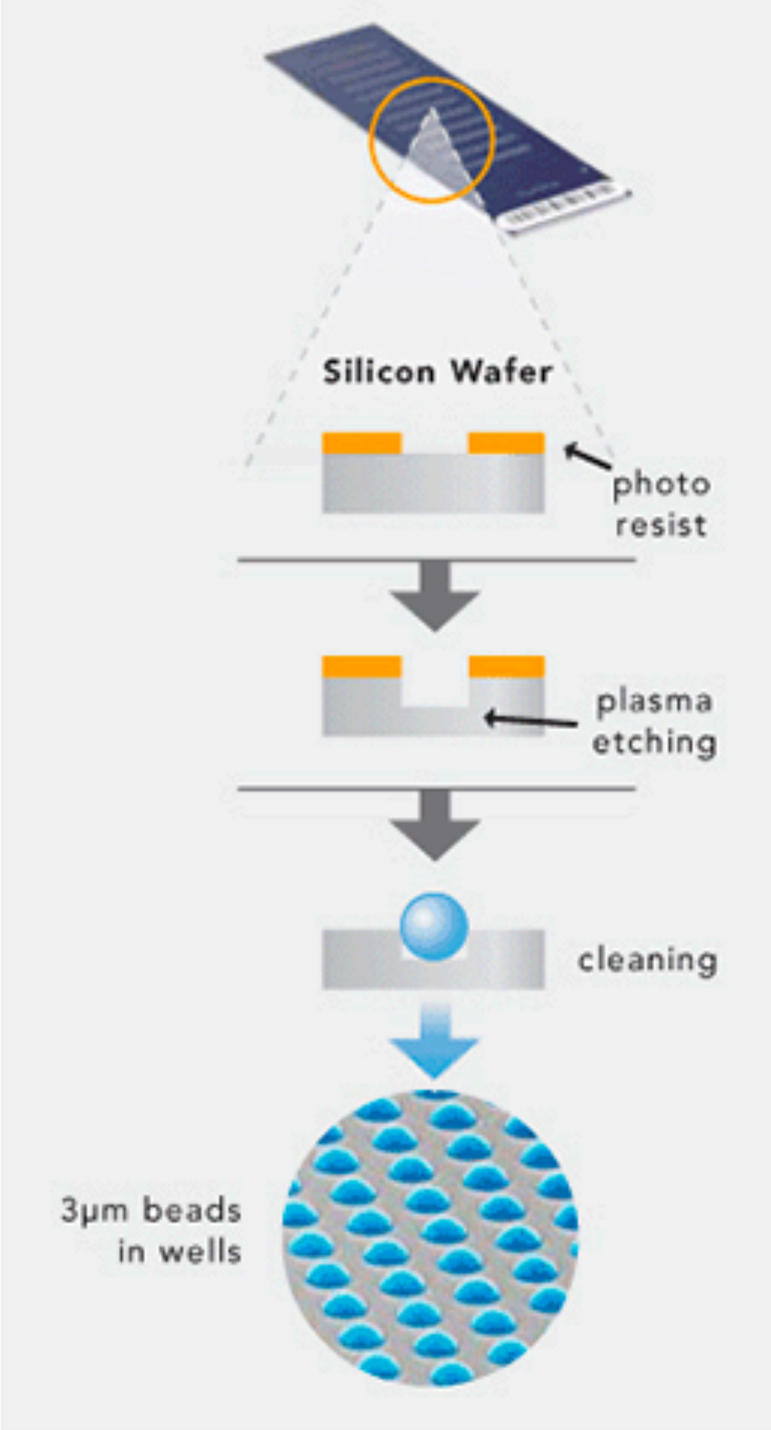
How do we call/quantify variants?



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

How do we measure variants?



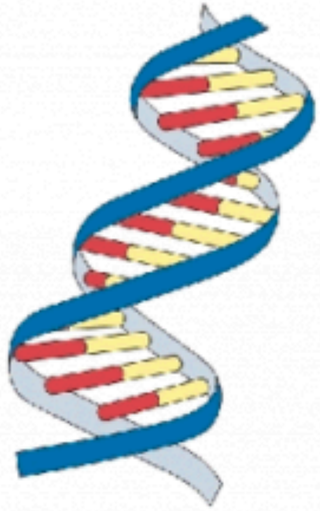


https://www.illumina.com/science/technology/microarray.html

Today's lecture

- Fundamentals in Population Genetics
- Genome-wide association studies
- What are the limitations of GWAS?

Genotype vs. phenotype



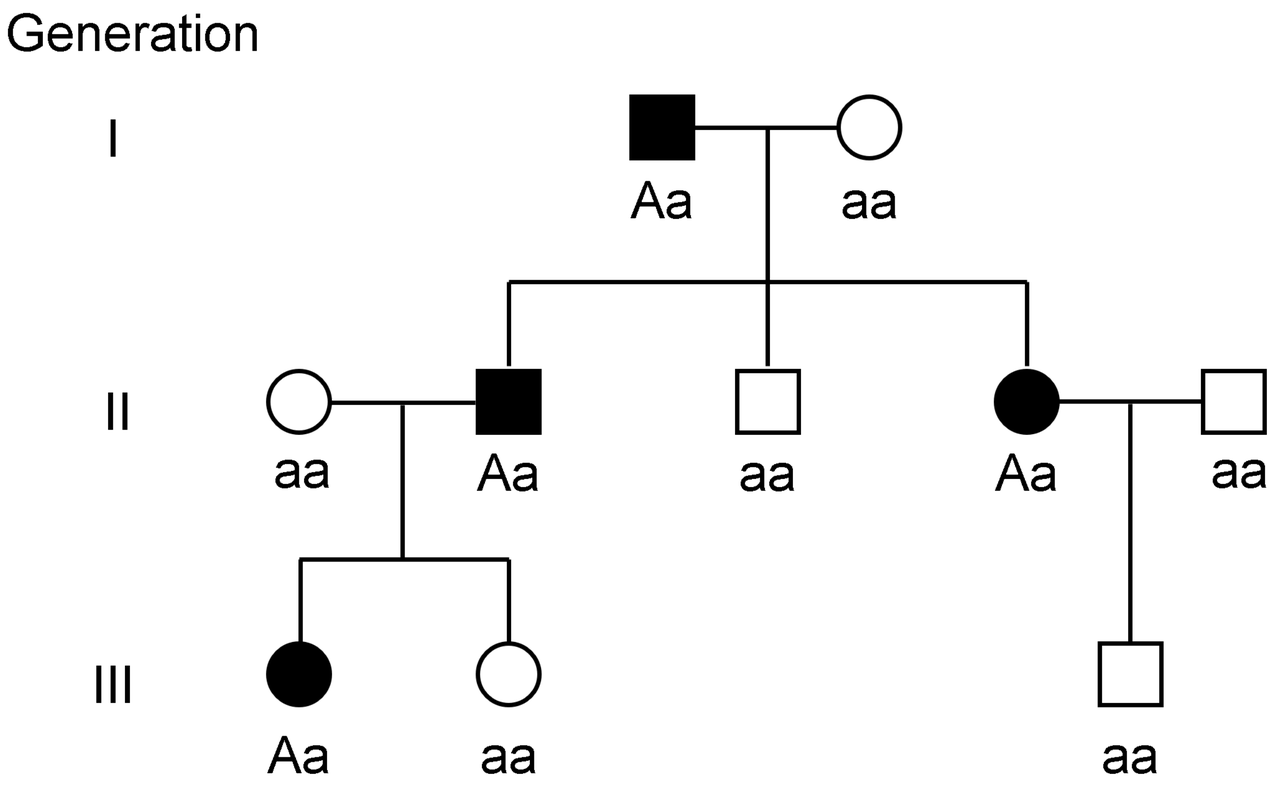
How do we test this?

Genotypes are the genetic make-up of an individual.

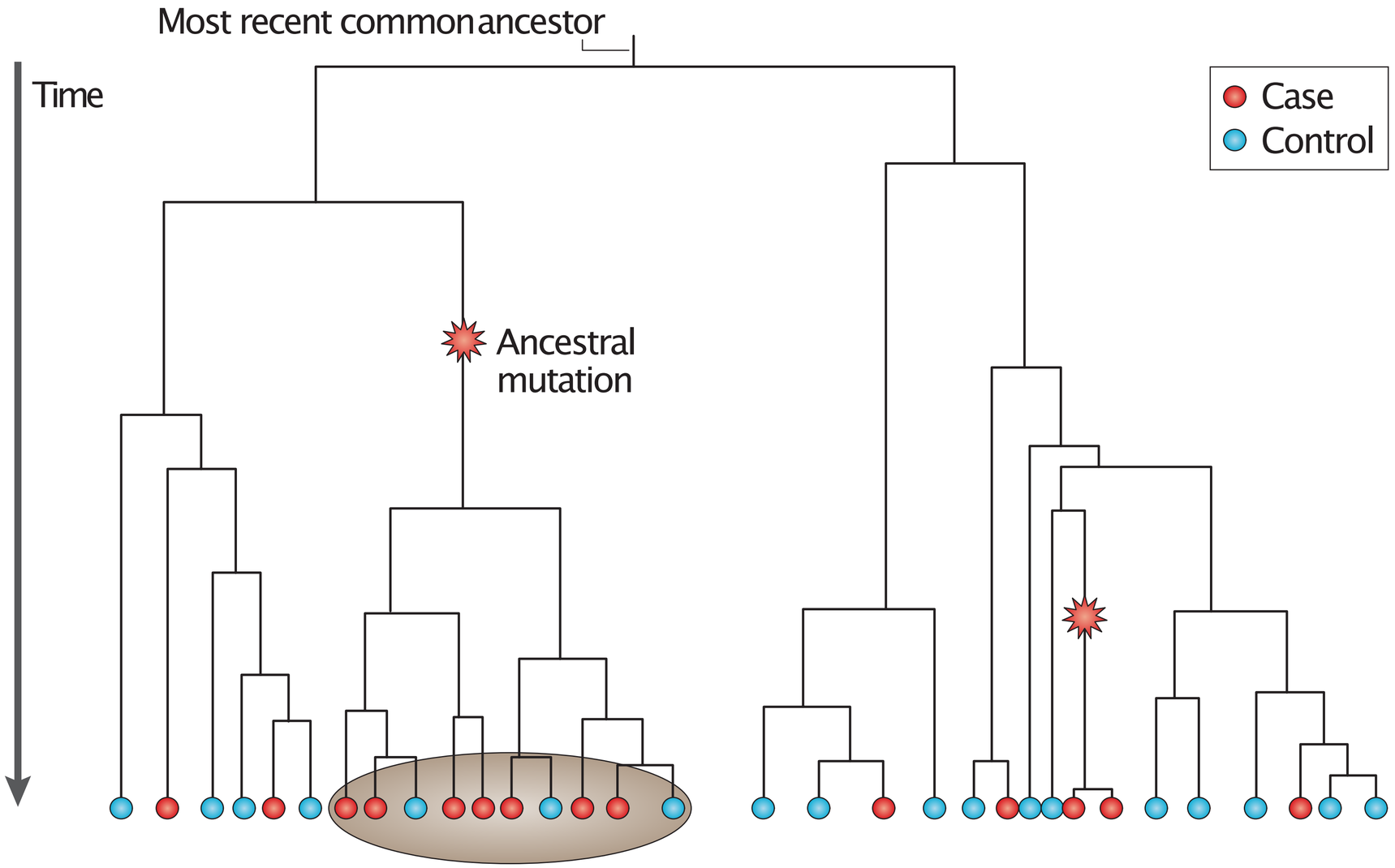
Phenotypes are the physical traits and characteristics of an individual and are influenced by their genotype and the environment.

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

How do we associate genetic variants to traits?

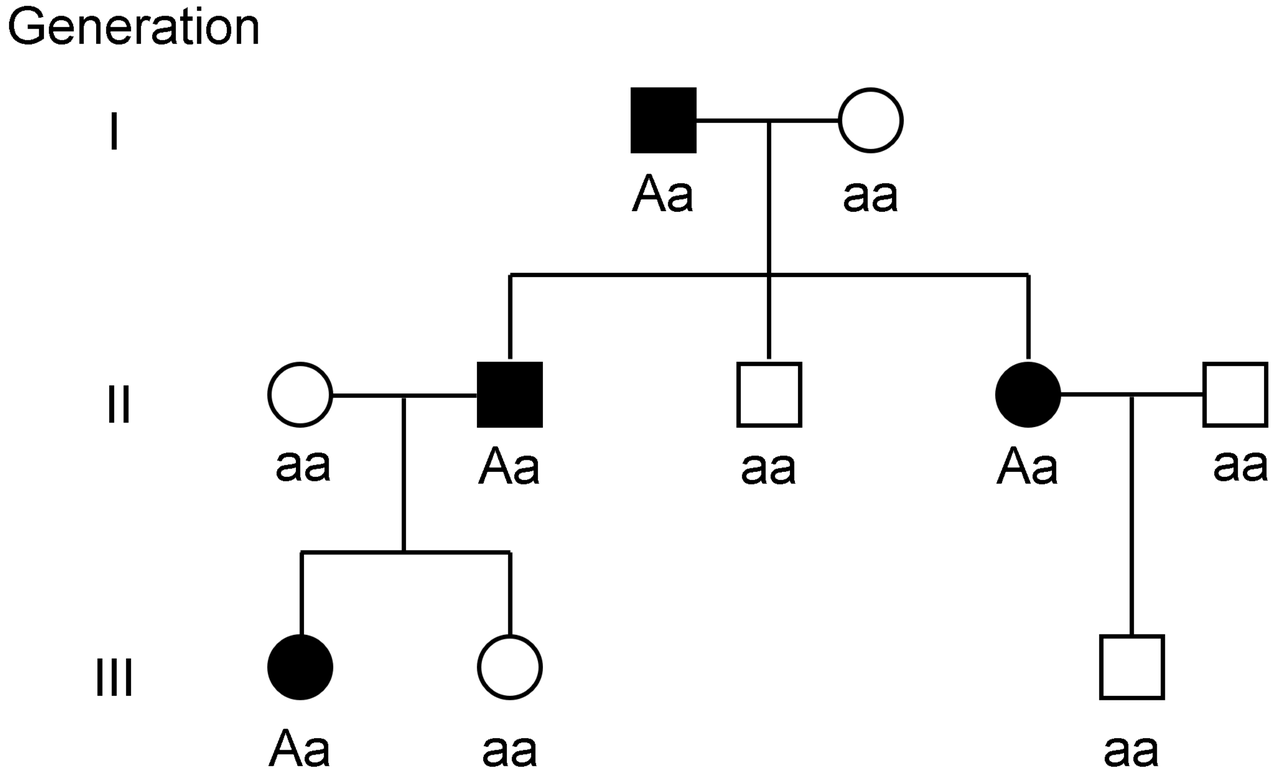


How can a genetic disease occur?



Balding, Nature Genetics Review, 2006

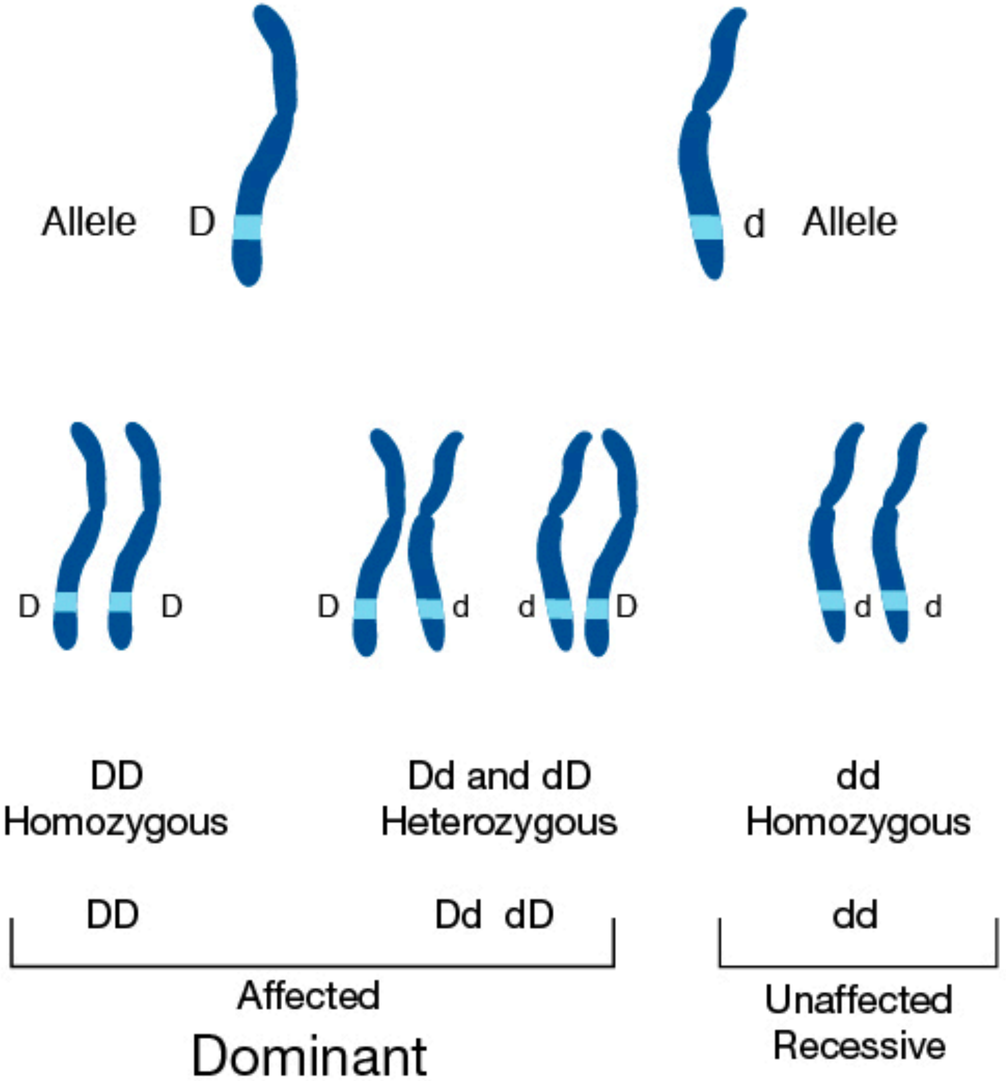
Mendelian disorders ≈ a single disease gene



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

Very high penetrance, monogenic

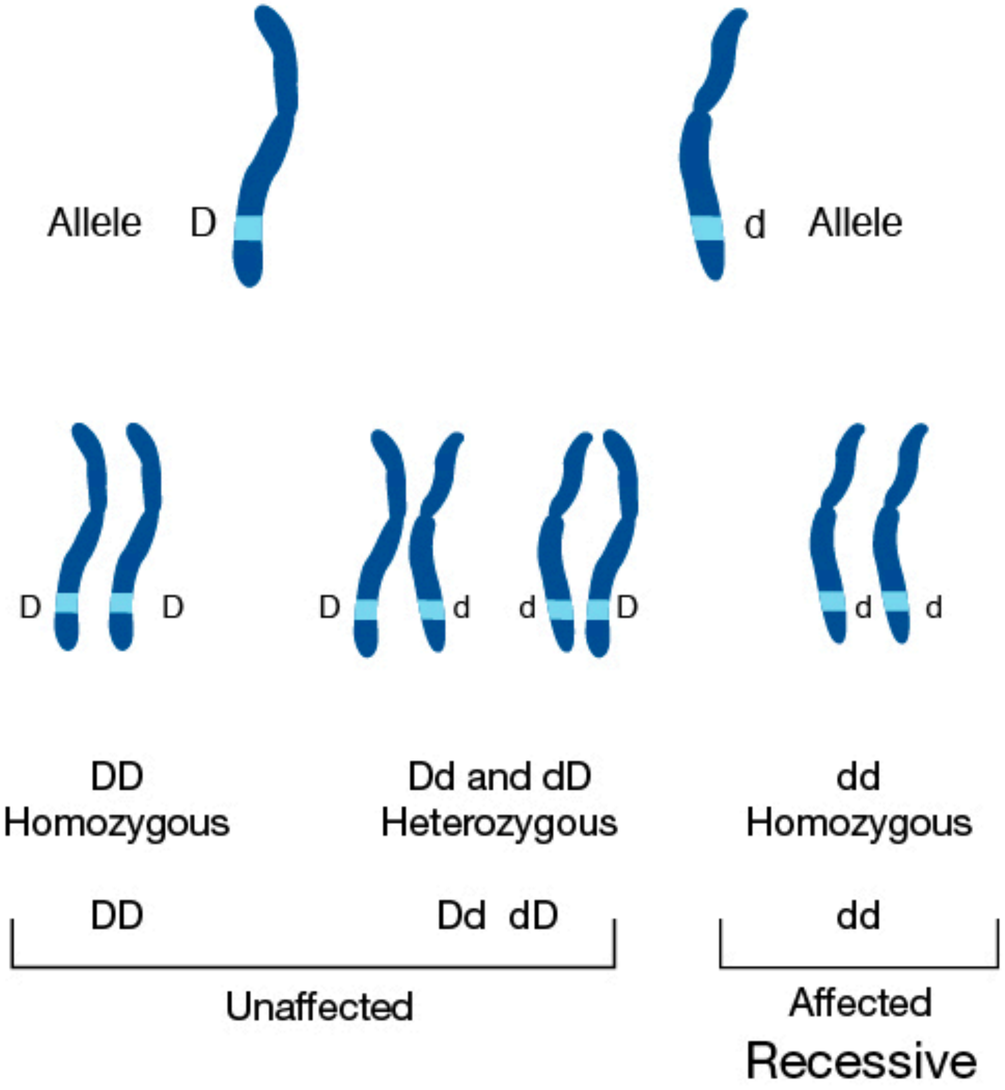
Huntington's Disease
Dominant



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

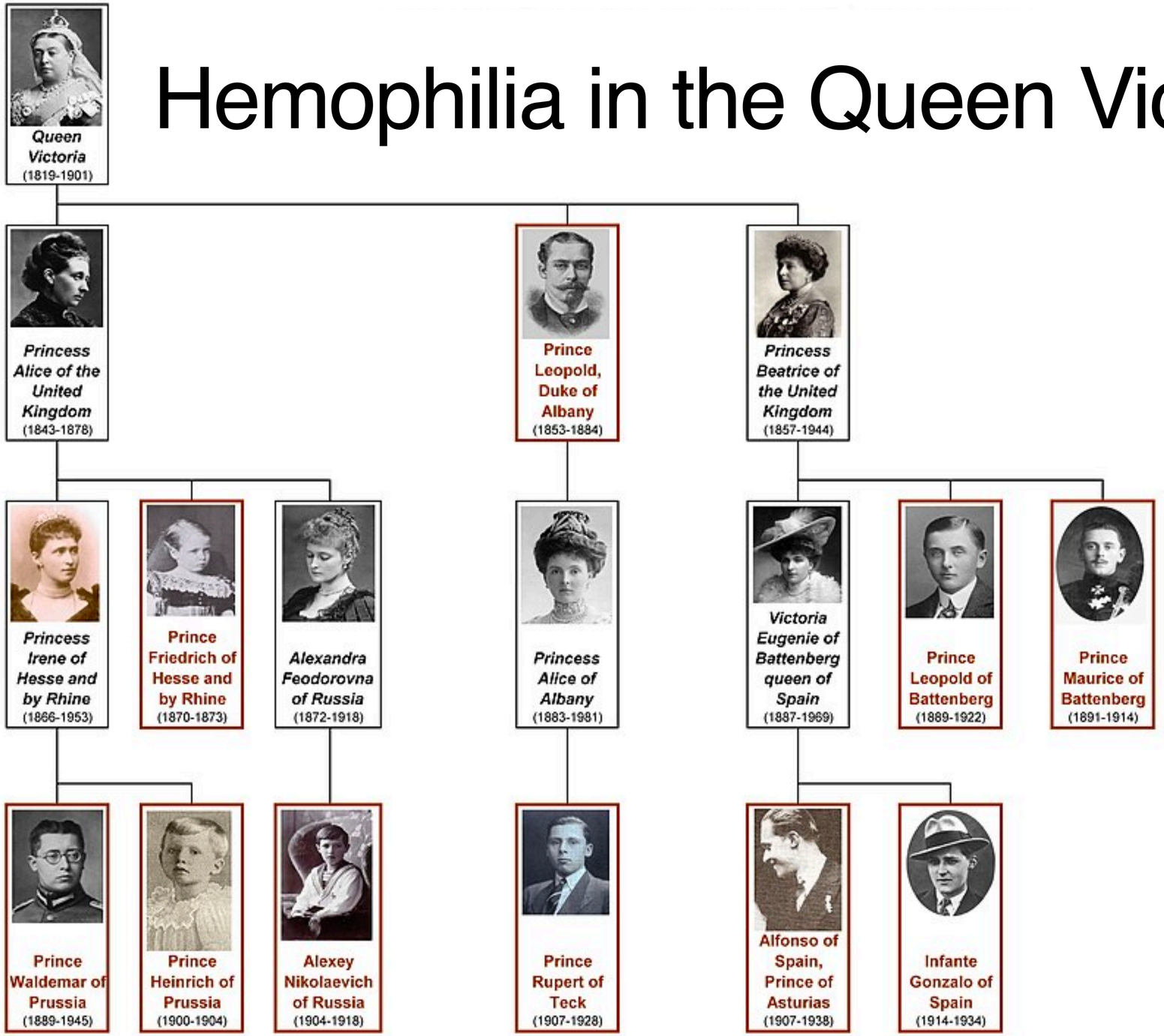
Sickle Cell Anemia or Cystic Fibrosis

Recessive



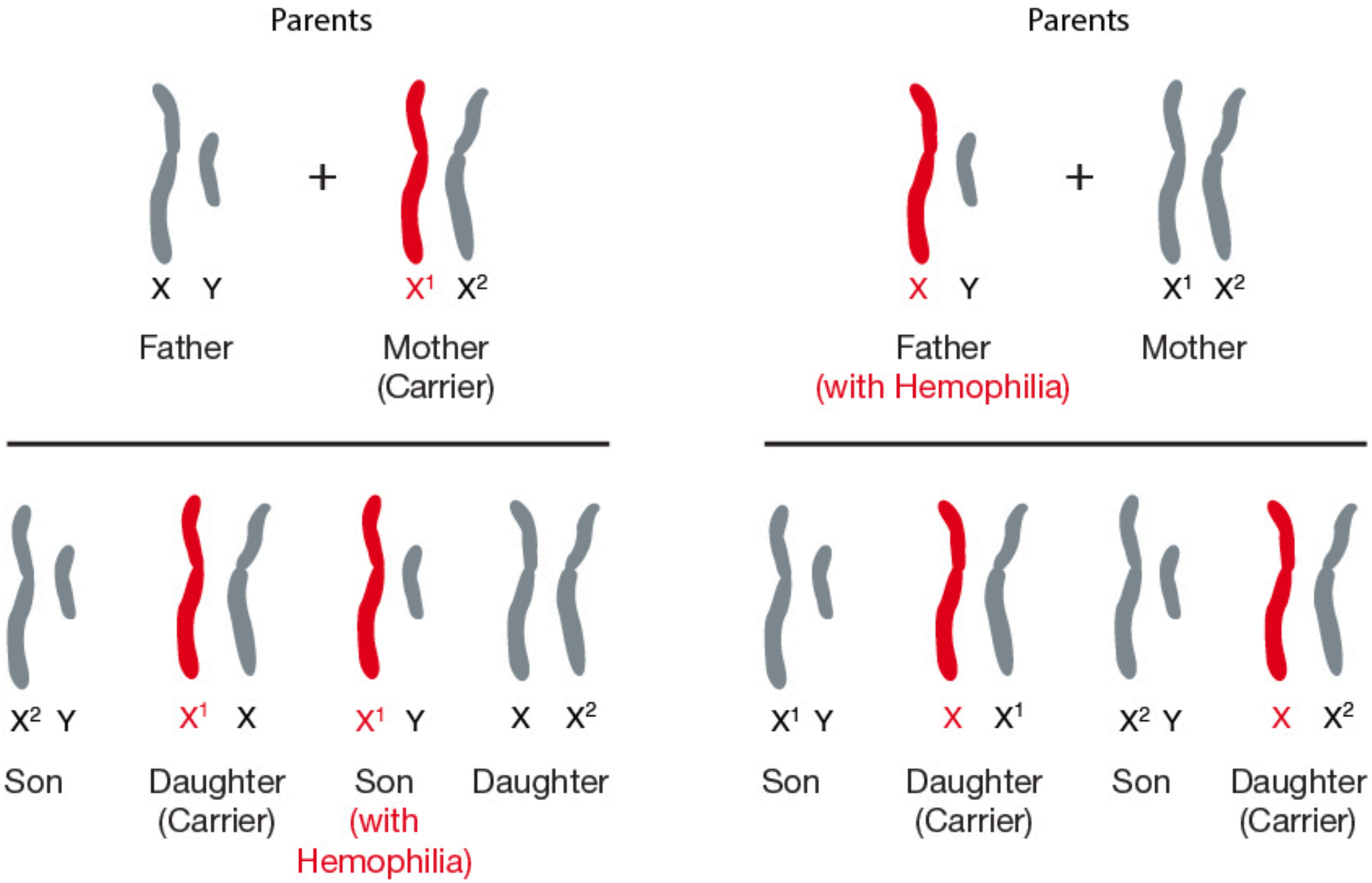
https://www.genome.gov/genetics-glossary

Hemophilia in the Queen Victoria's family



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

Hemophilia



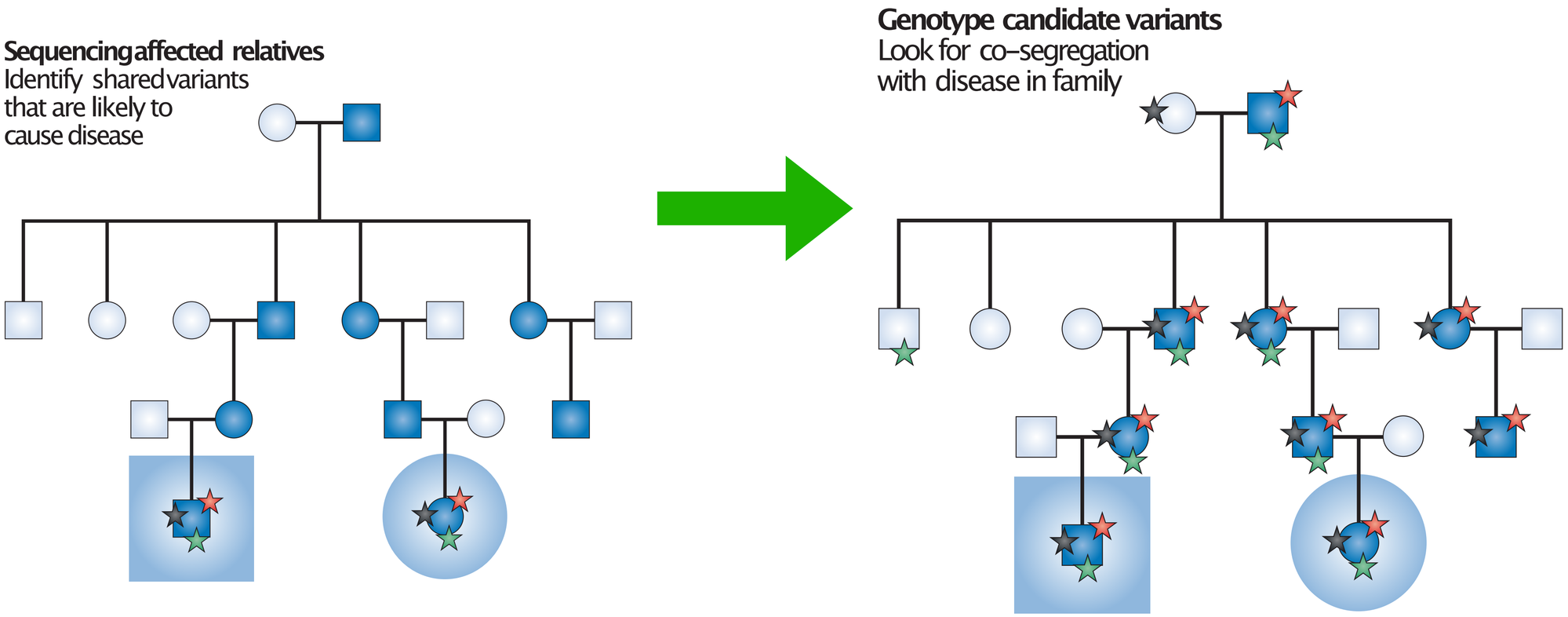
https://www.genome.gov/genetics-glossary

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.

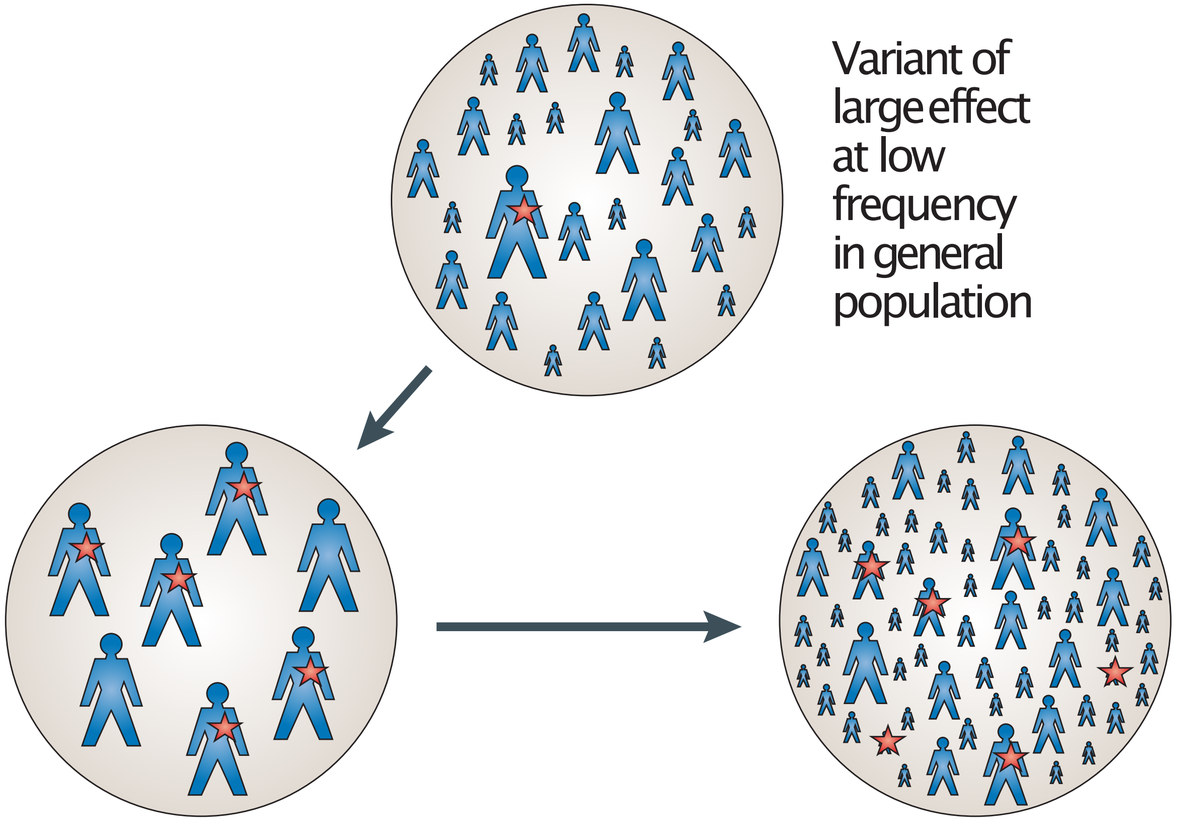
This database was initiated in the early 1960s by Dr. Victor A. McKusick as a catalog of mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM). Twelve book editions of MIM were published between 1966 and 1998. The online version, OMIM, was created in 1985 by a collaboration between 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?

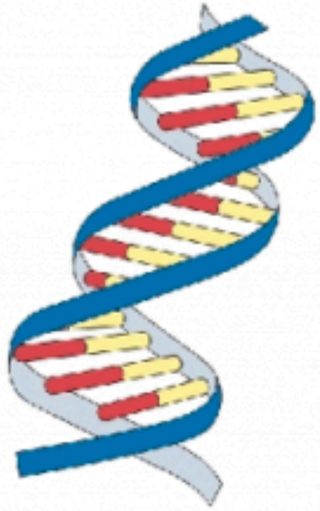


Whole–genome sequencing
Rare variant enriched in population with extreme phenotype

Targeted genotype
Rare variant associates with phenotype in larger population

Cirulli and Goldstein, 2010

Can we identify variants associated with traits?

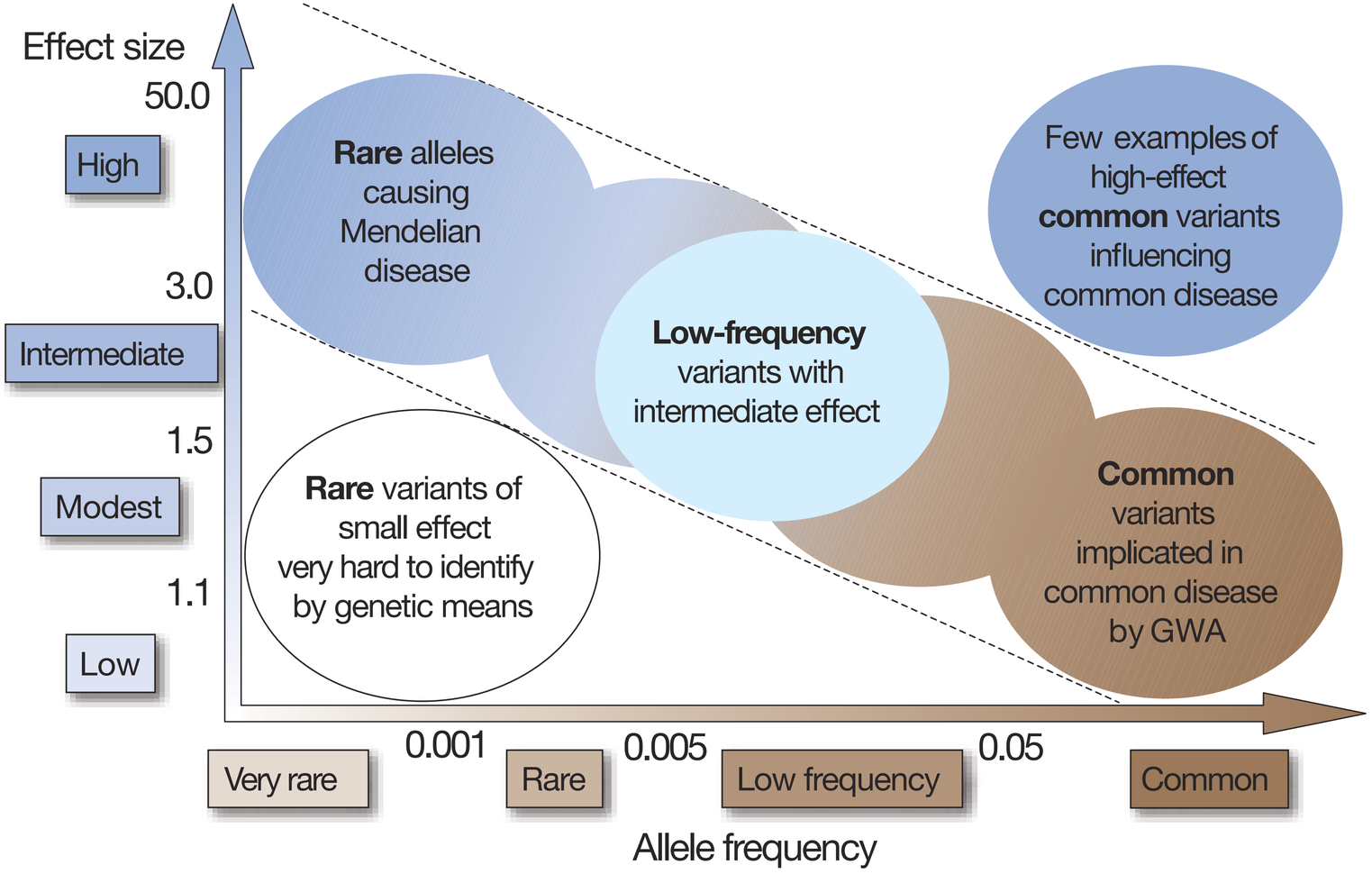


Genotypes are the genetic make-up of an individual.

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

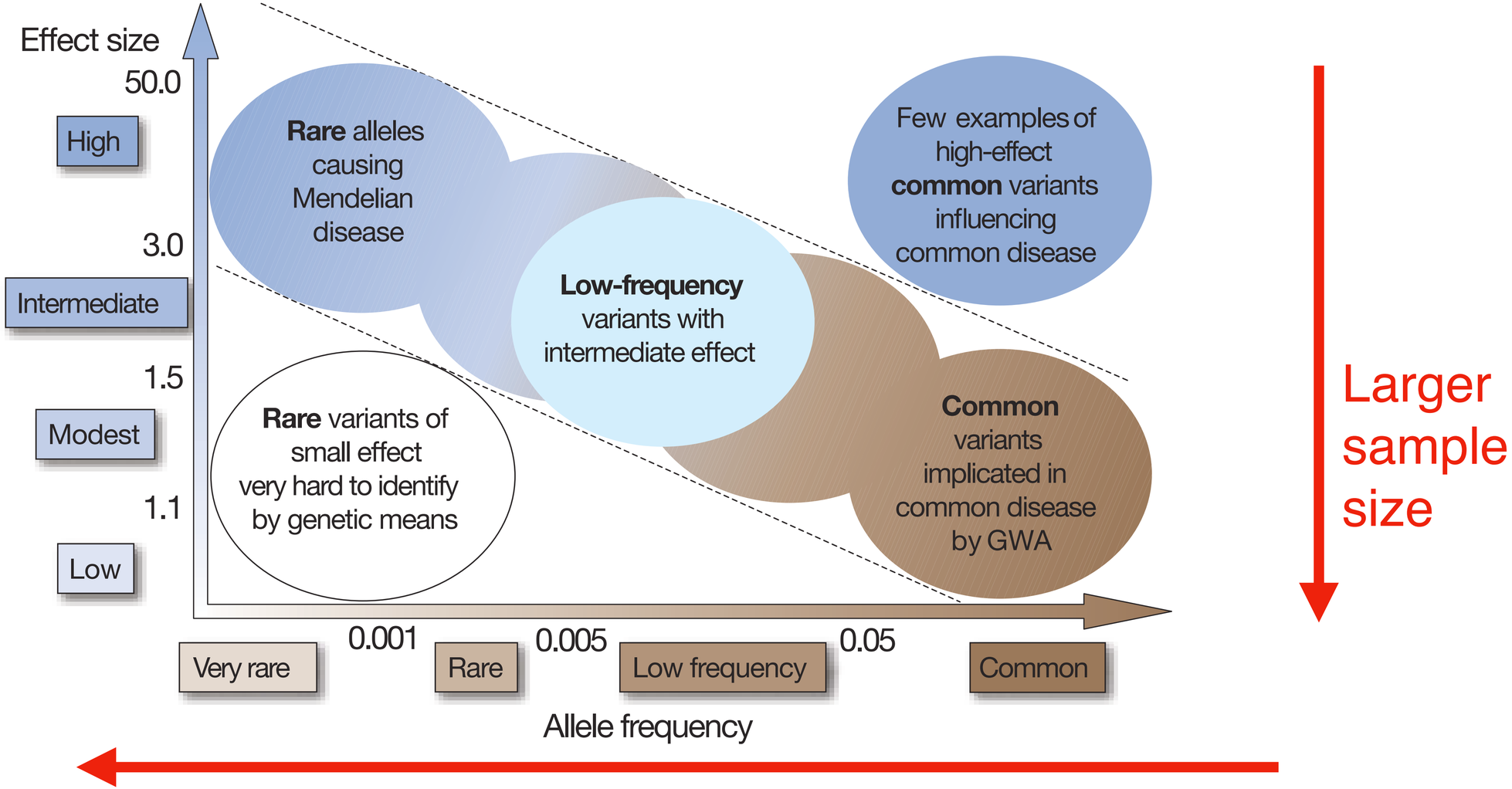
https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction

Will trait-associated variants emerge?



Manolio .. Vischer, *Nature* (2009)

Will population-level genetic studies be of worth?



Manolio .. Vischer, *Nature* (2009)

We need a large sample size!

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

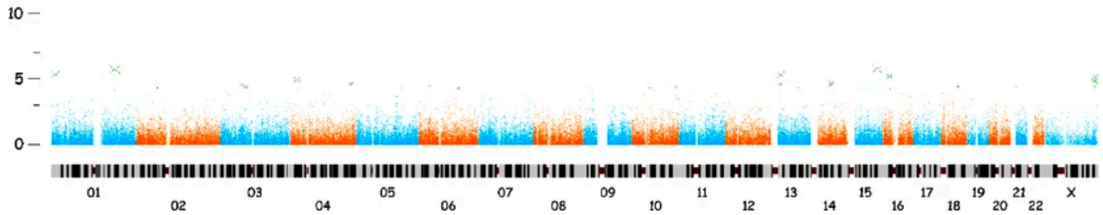
IMMEDIATE COMMUNICATION

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!

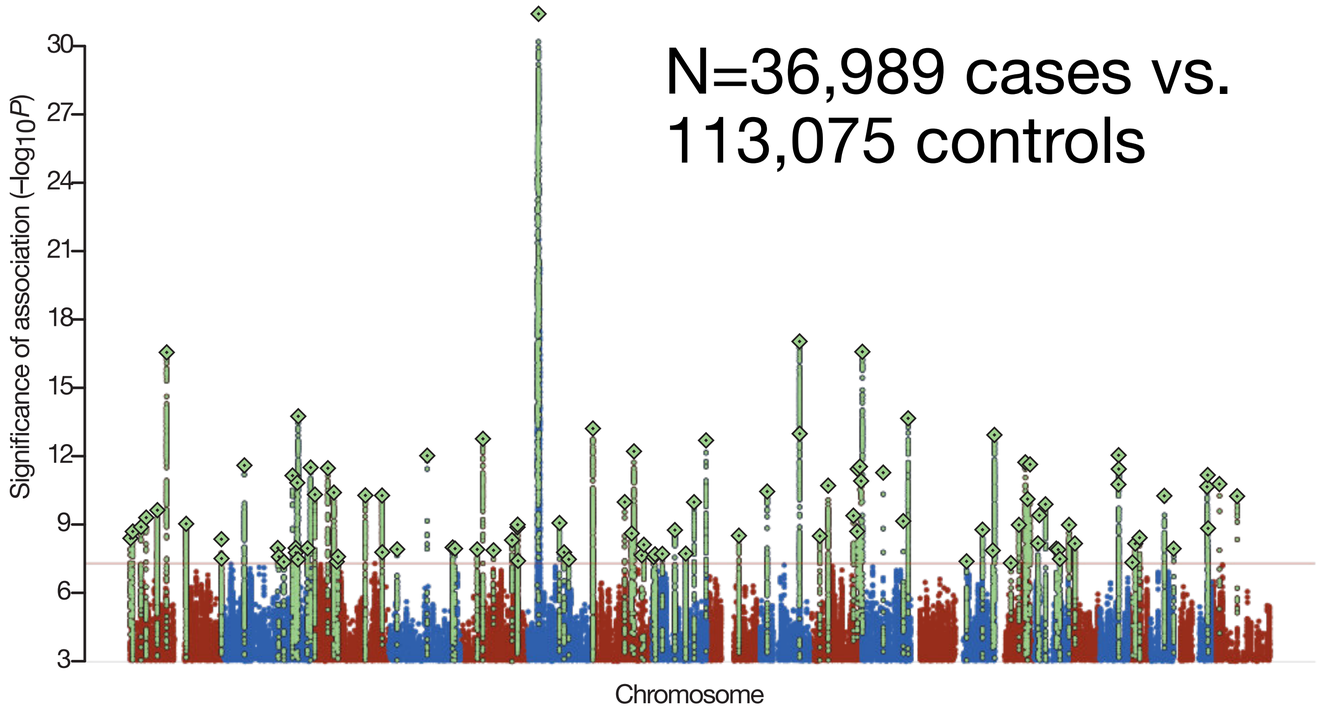


Sullivan *et al.* (2008)

Biological insights from 108 schizophrenia-associated genetic loci

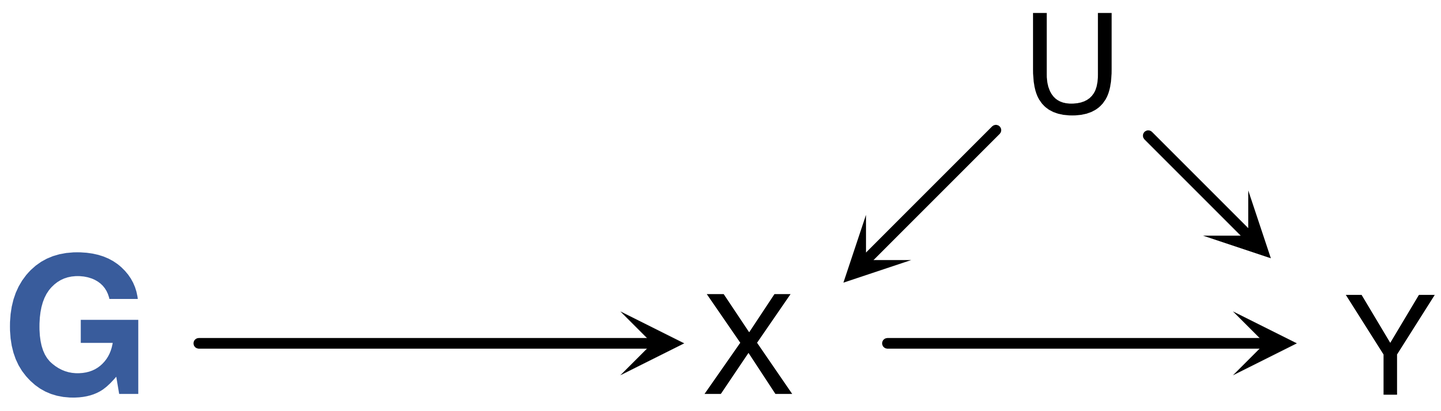
Schizophrenia Working Group of the Psychiatric Genomics Consortium*

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



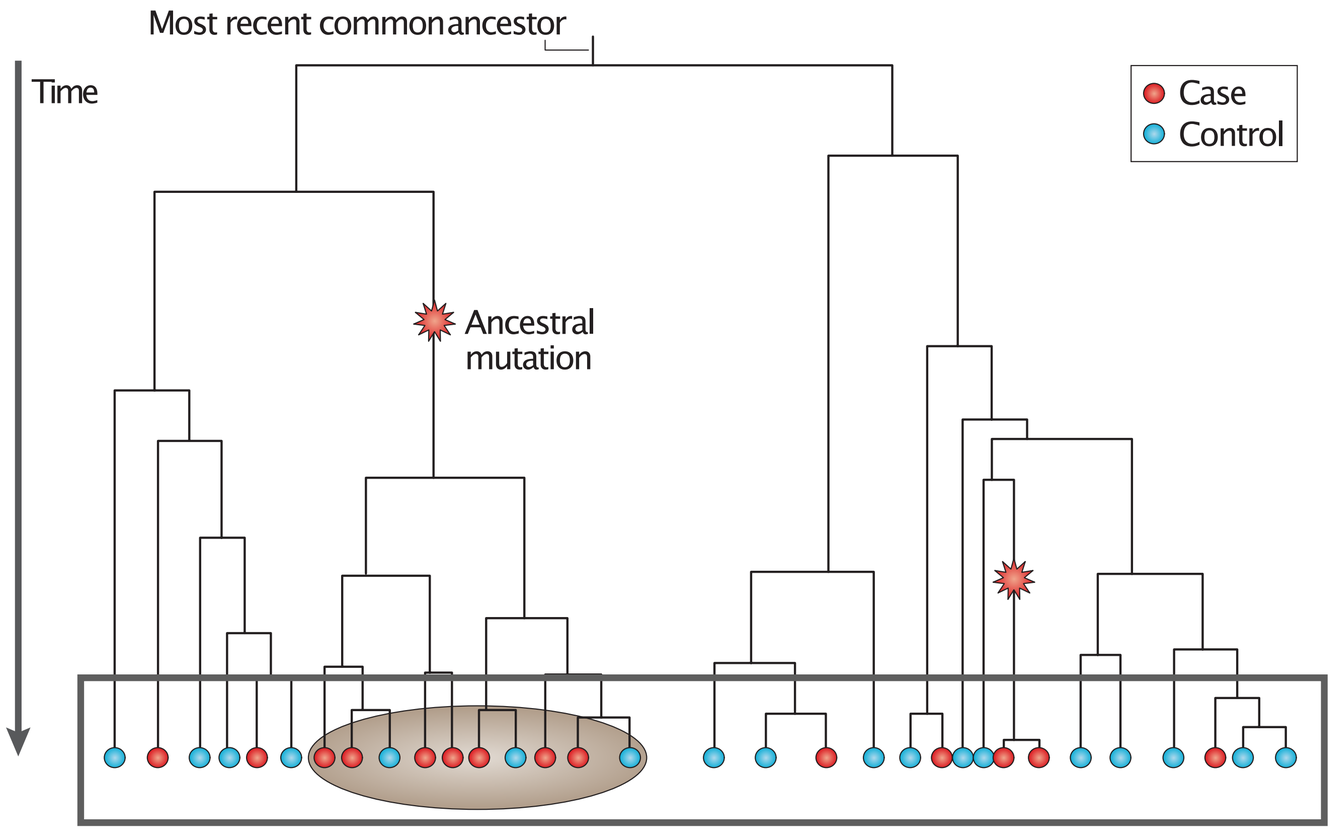
Schizophrenia Working Group of PGC, *Nature* (2014)

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



T/T C/C T/T T/C T/T

Don't care about the history

Observed complex phenotype

vs.

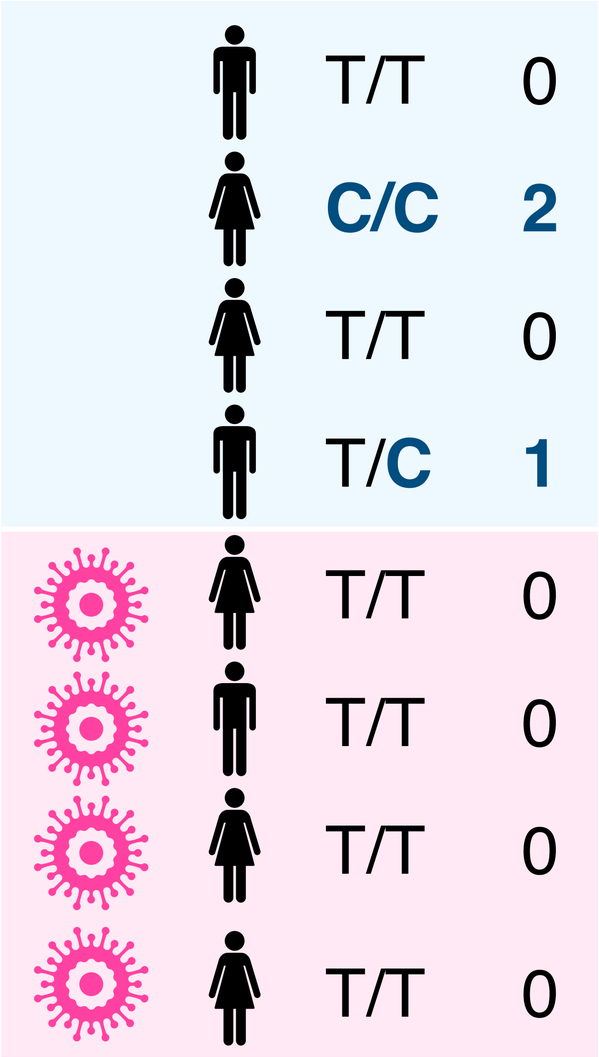
Genotype at a certain locus

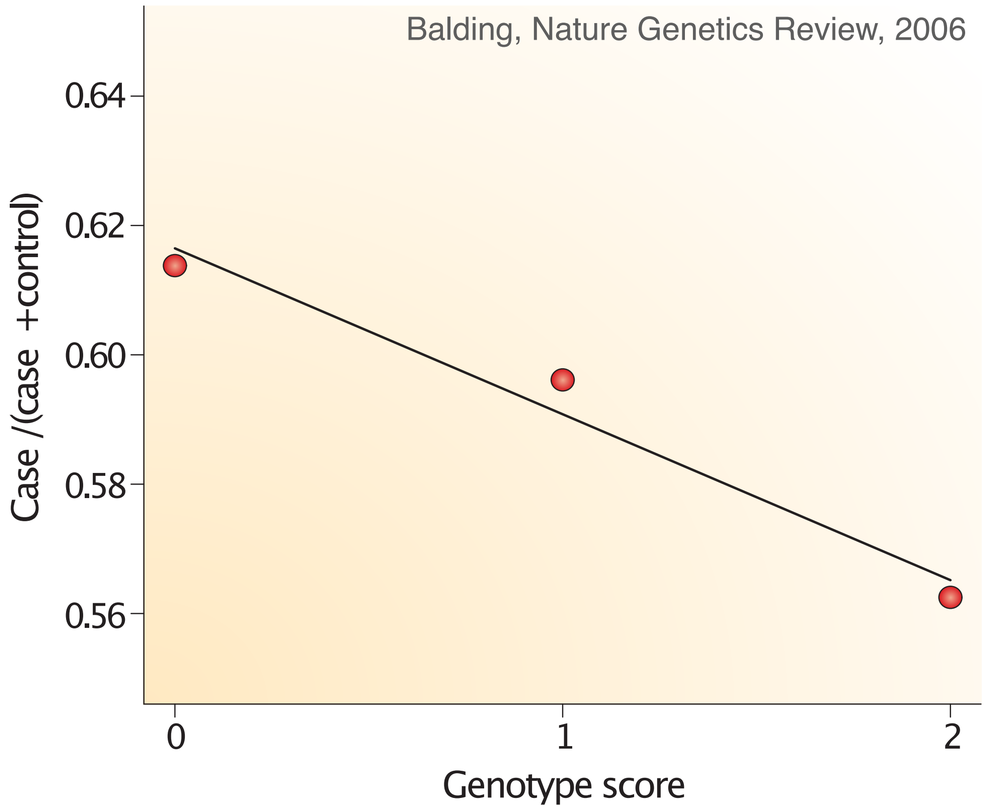
Balding, Nature Genetics Review, 2006

A typical genetics study design

Resistant
to a disease

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





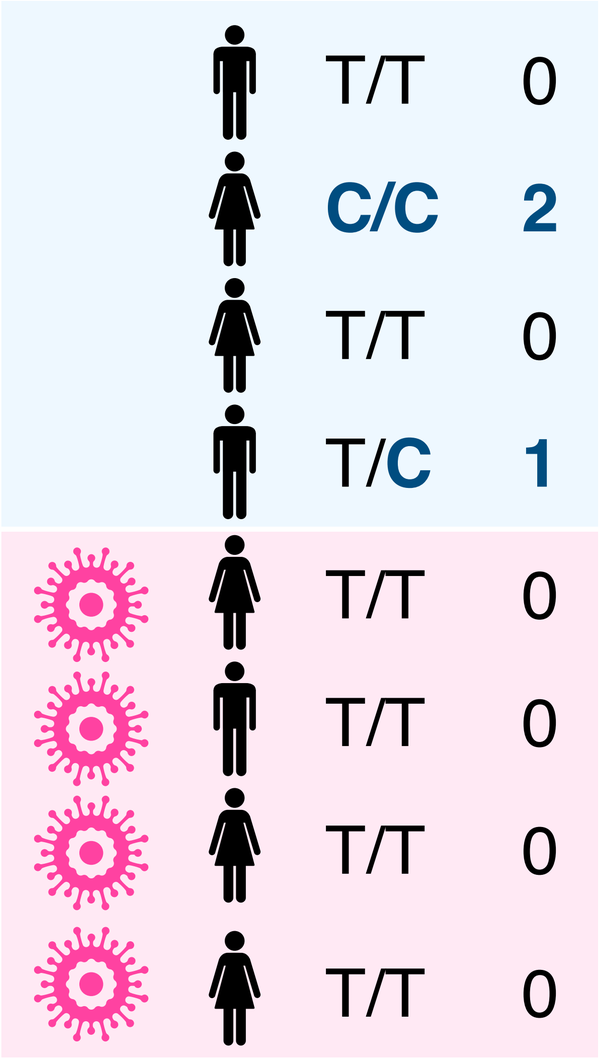
T = a major allele

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*

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

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

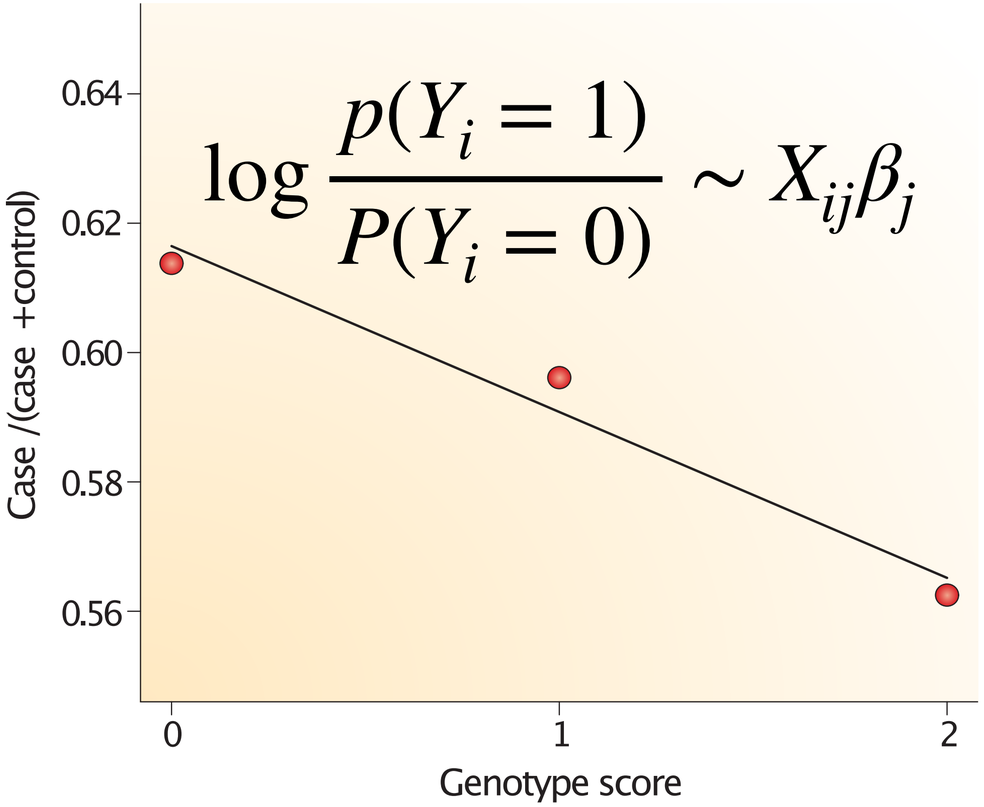
Genotype (dosage) of a variant *j*

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

T = a major allele C = a minor allele

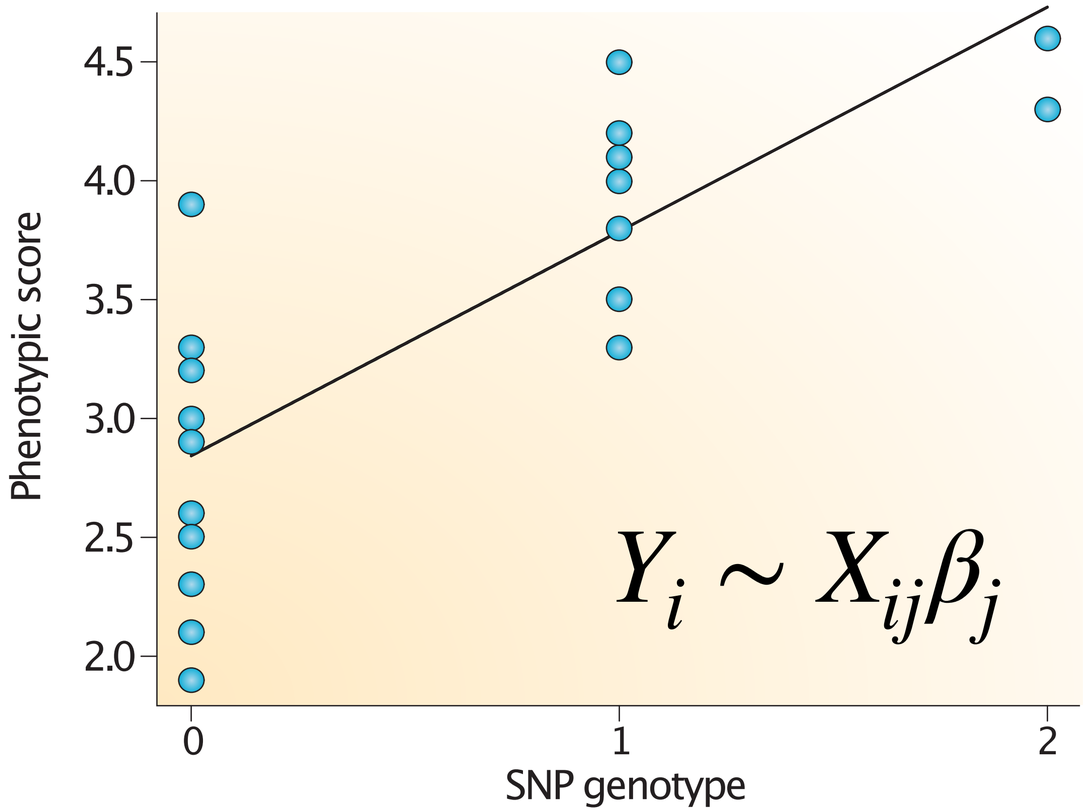
A traditional genetics study design

Case-control GWAS



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

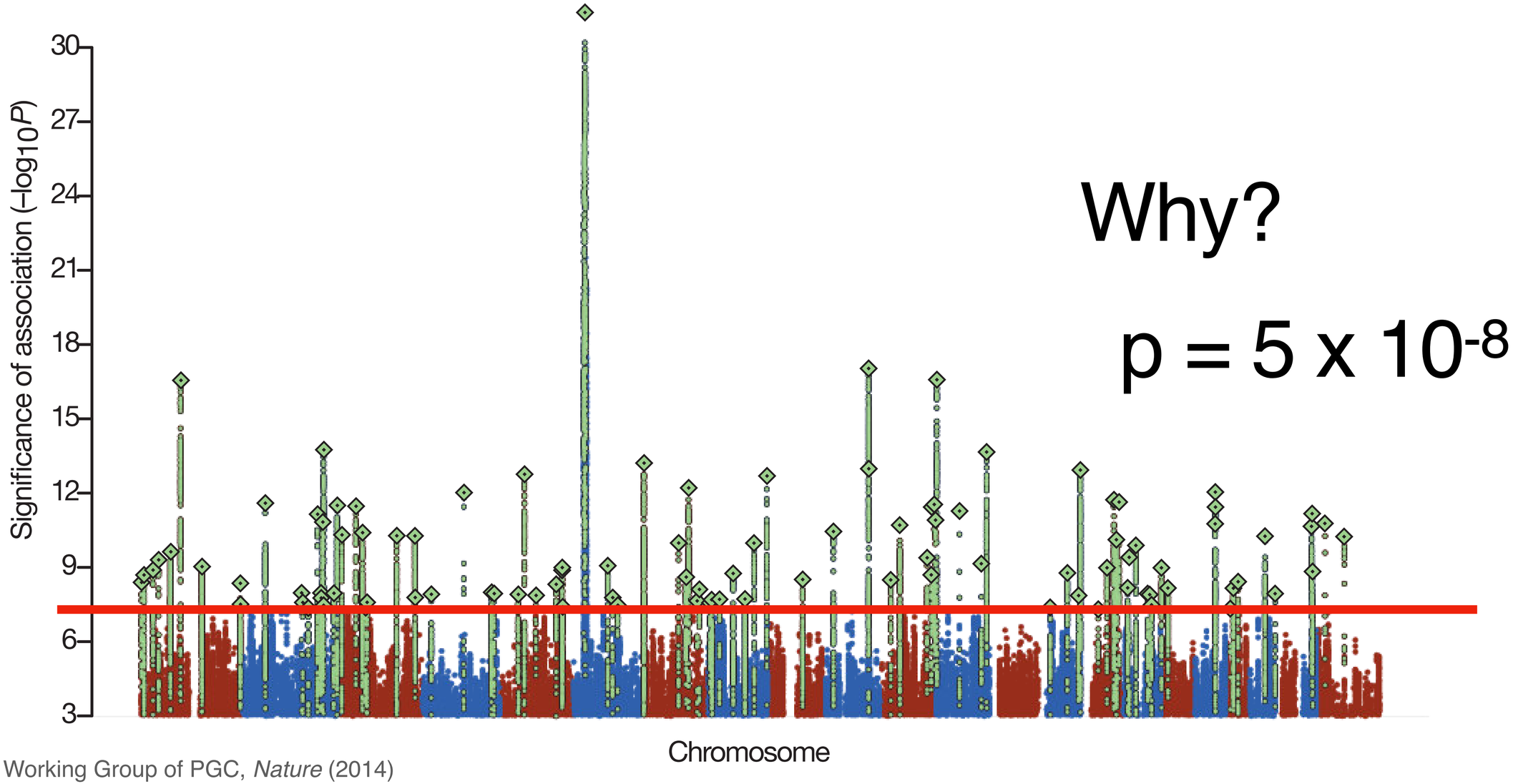
Quantitative trait GWAS



$$Y_i \sim X_{ij}\beta_j$$

Balding, Nature Genetics Review, 2006

Genome-wide association study (many loci)



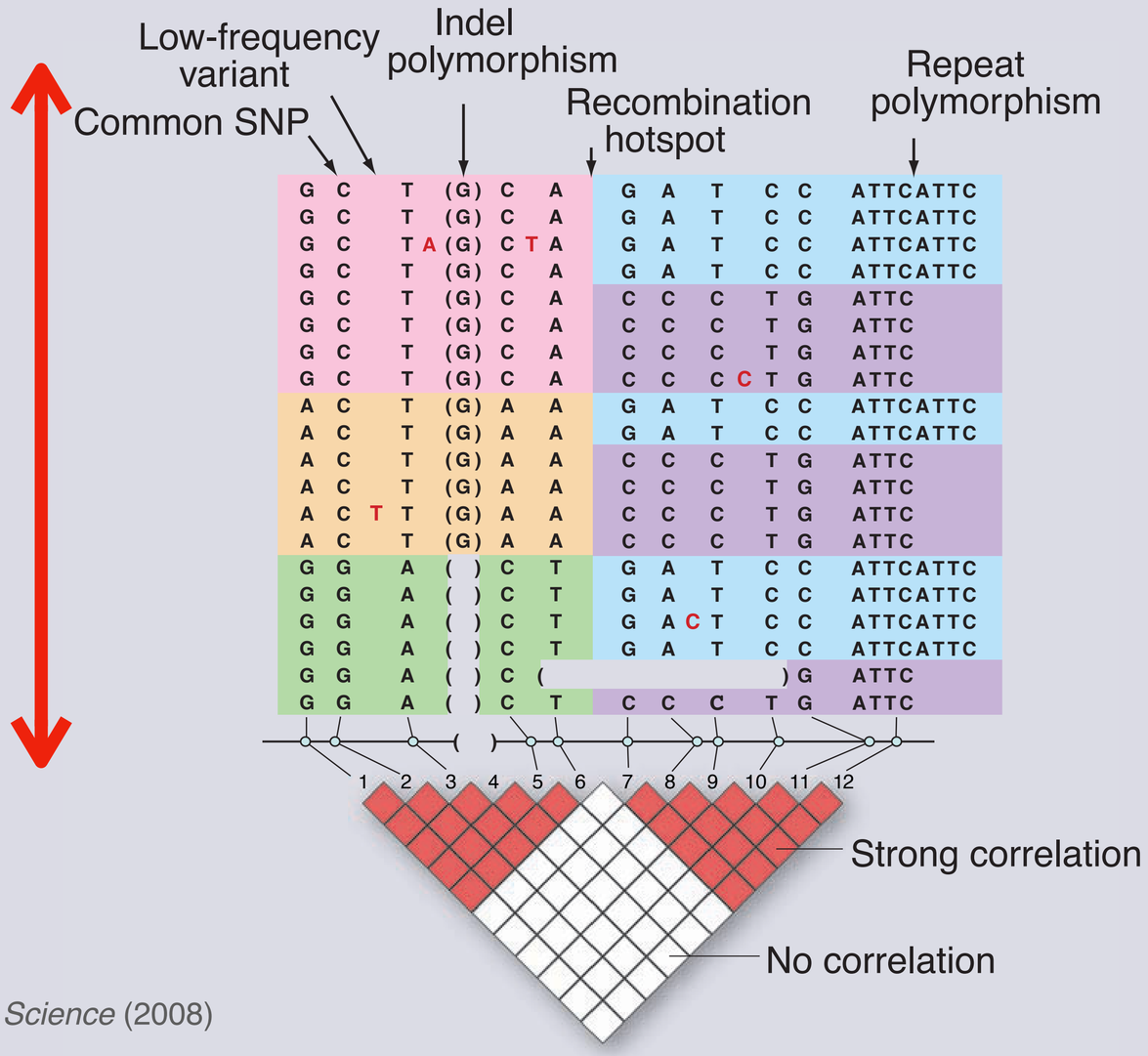
Why?

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

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

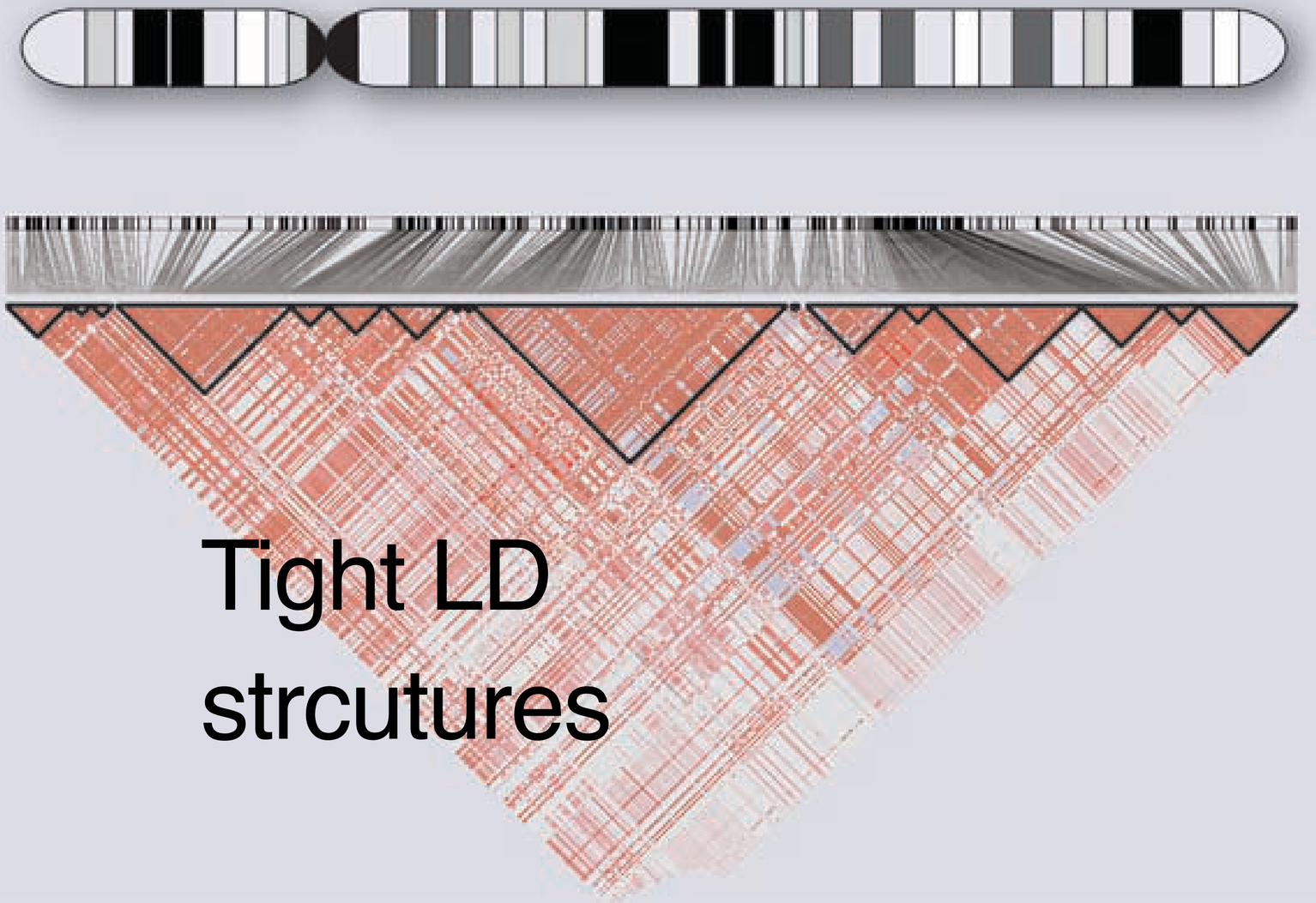
Genetic variation in one figure

across many individuals (diploid genomes)



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

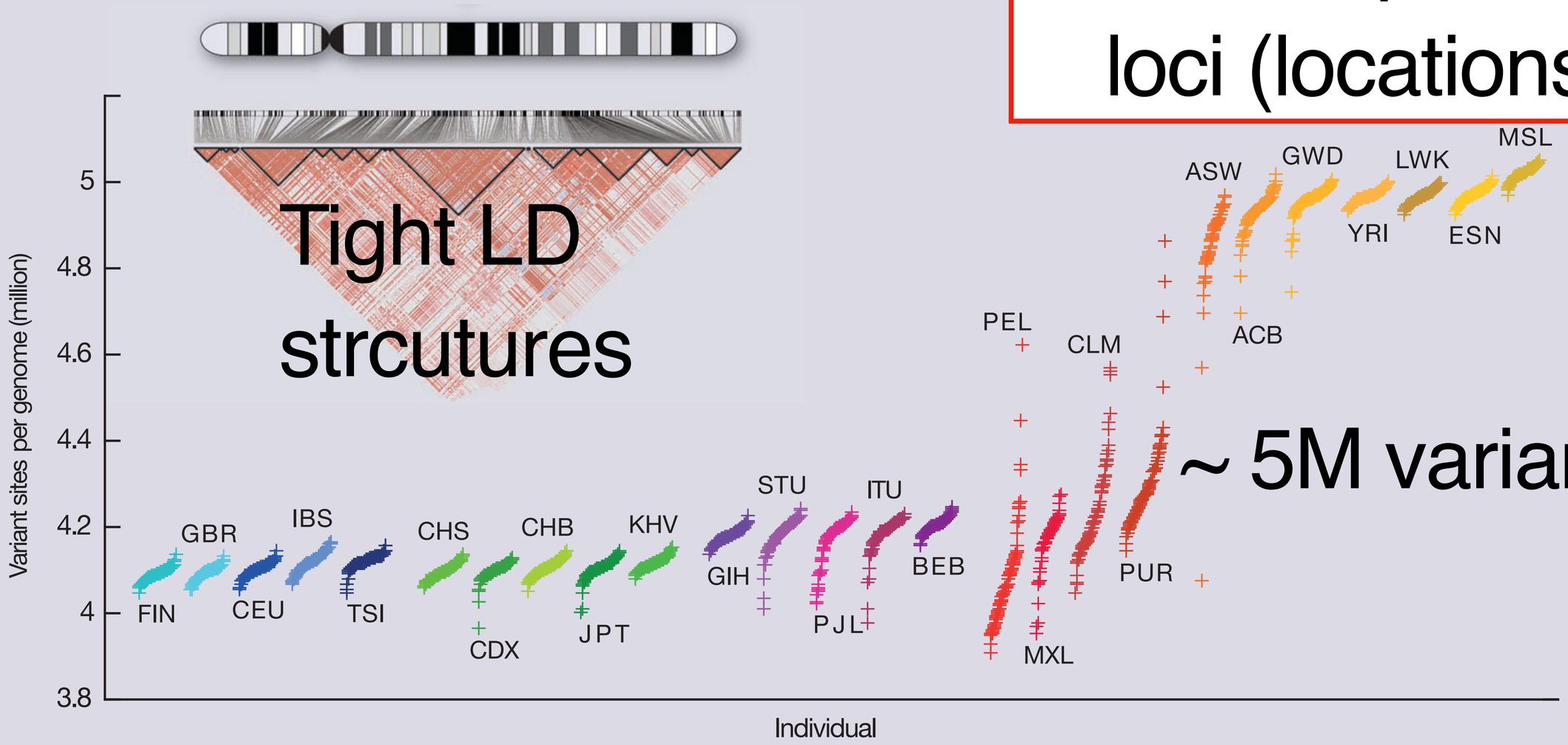
Altshuler, Daly, Lander, *Science* (2008)



Tight LD strcutures

1M "tagging" SNPs across the Human Genome

~ 1M independent loci (locations)



~ 5M variants

Genome-wide significance level

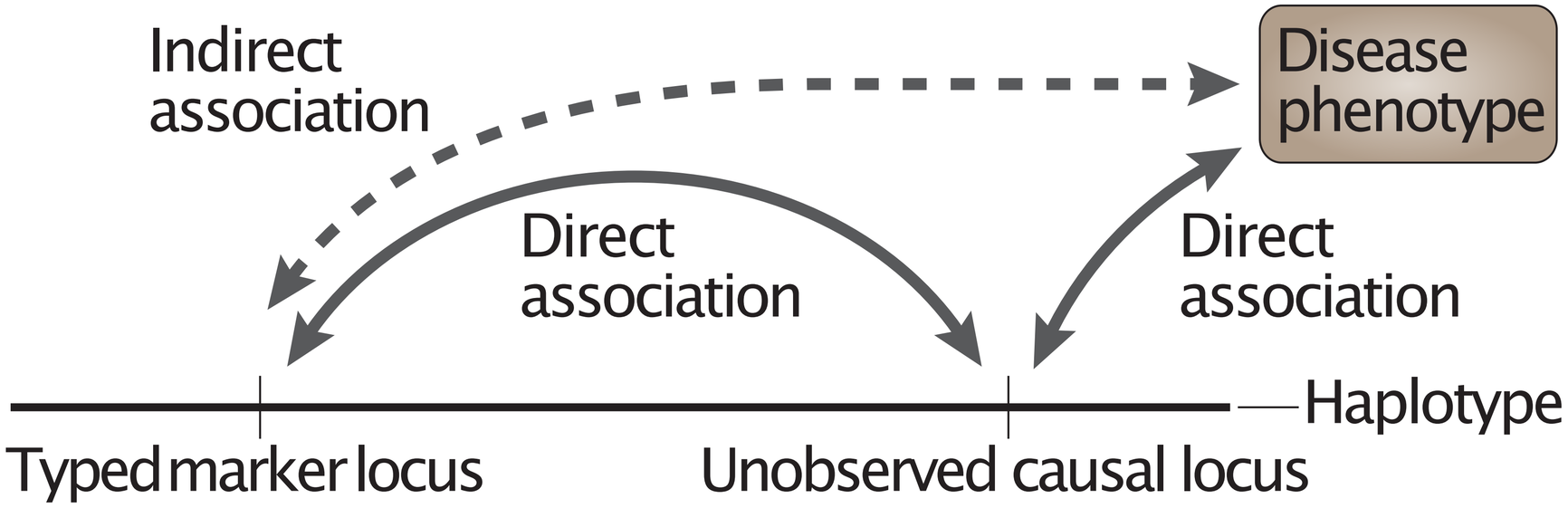
$$p_j < \frac{0.05}{\text{number of (independent) tests}} = \frac{0.05}{10^6} = 5 \times 10^{-8}$$

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

Union bound

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

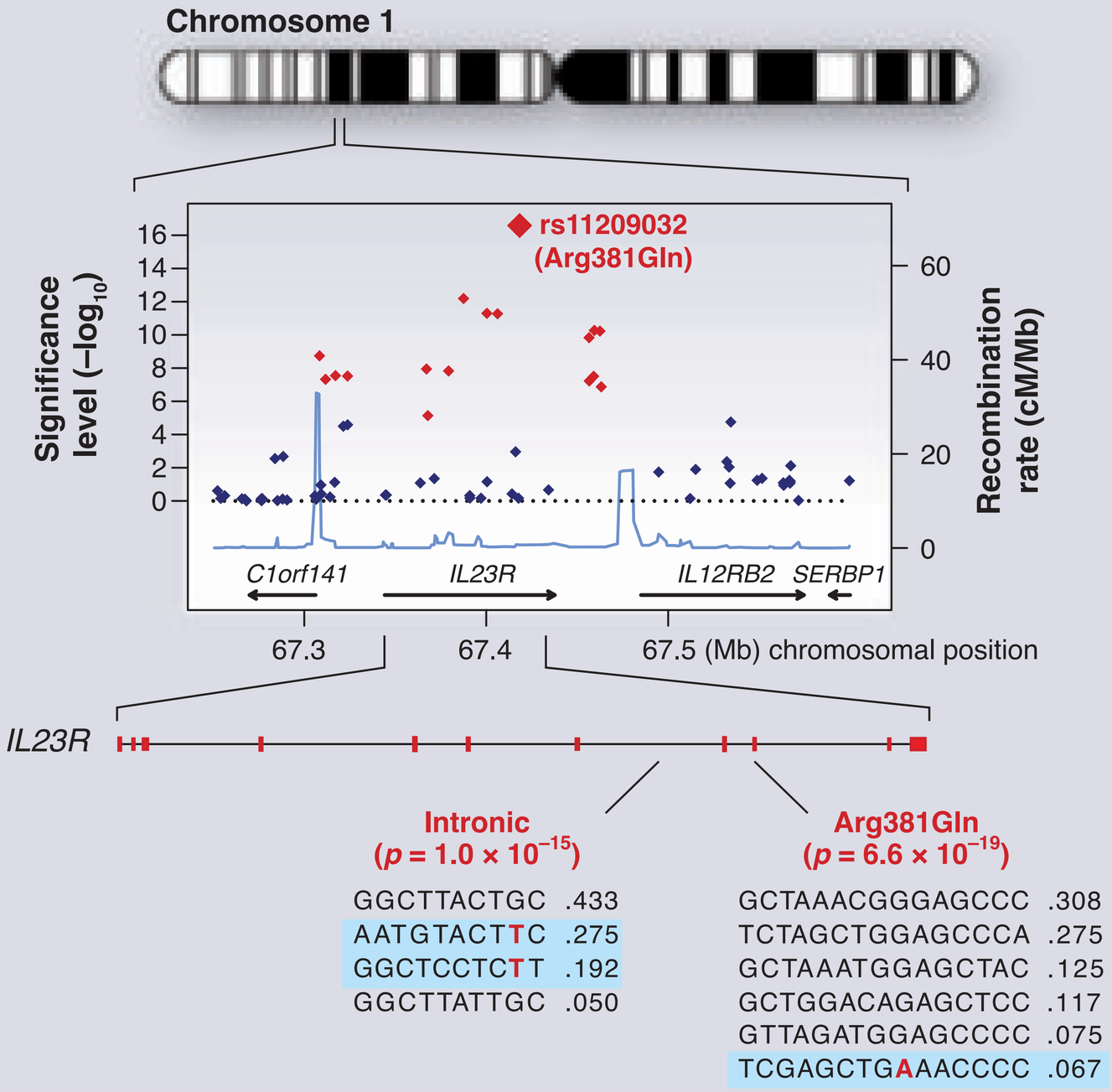
Types of association (typed ≈ tagging SNP)



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)

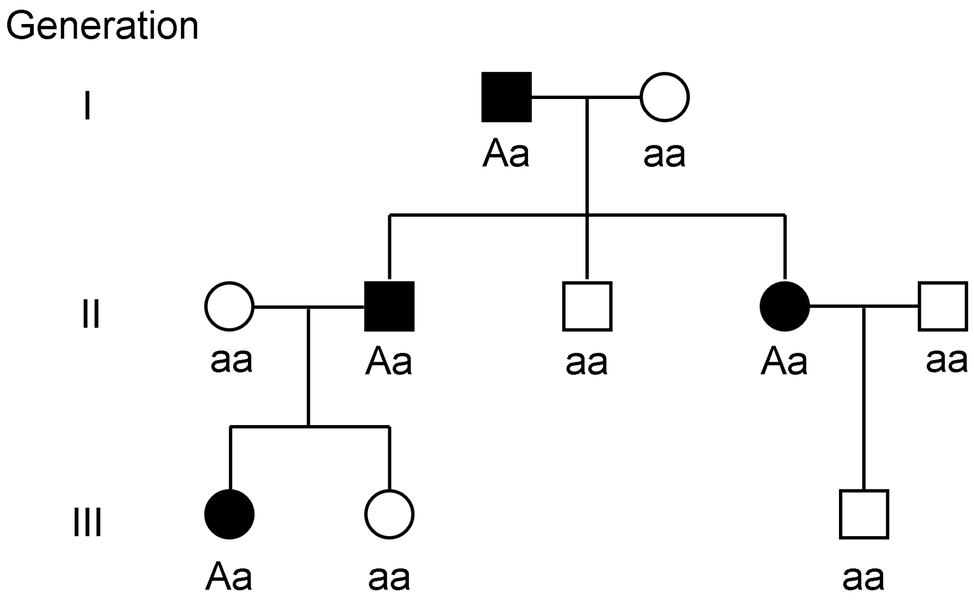
Summary for GWAS

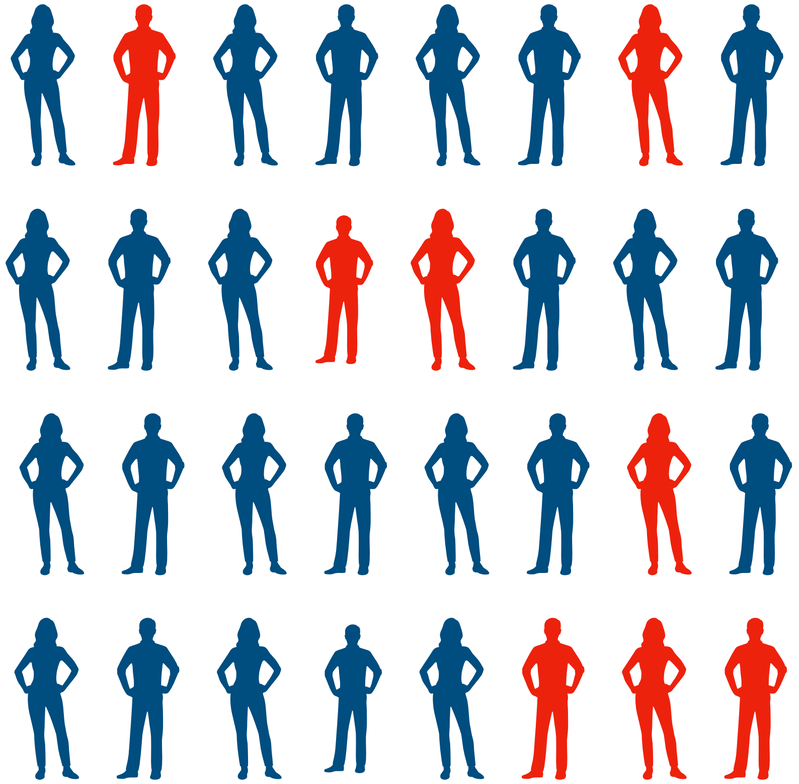
- Massive multiple hypothesis testing associations between genetic variants and phenotypic changes
- Most studies focus on biallelic variants, coding them as {0, 1, 2}
- Most association statistics concern an additive effect of genotypes (linearity)
- Genetic variants closely located in the genome are 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

Today's lecture

- Fundamentals in Population Genetics
- Genome-wide association studies
- What are the limitations & challenges in GWAS?

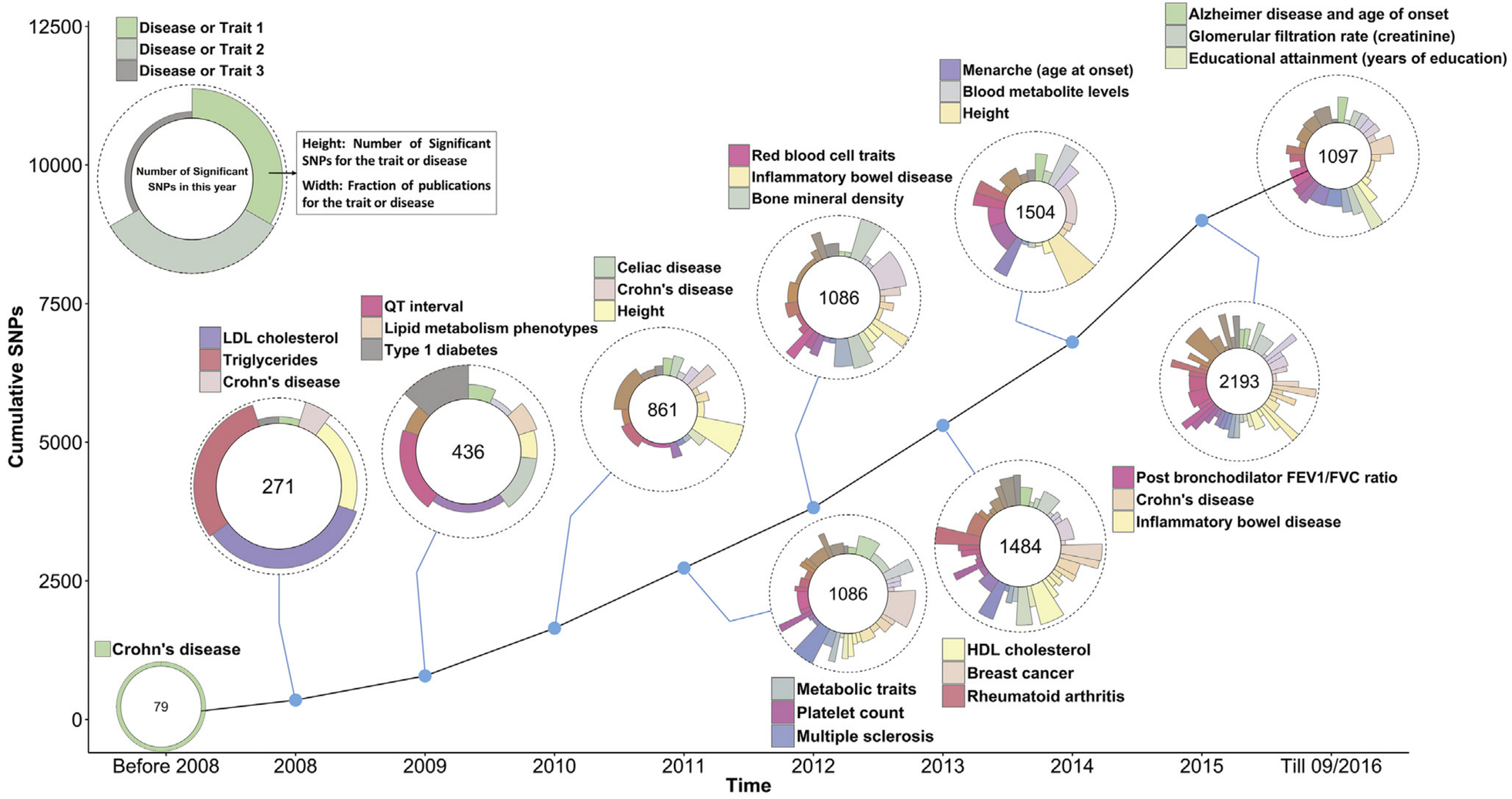
The success of GWAS changed the paradigm of human genetics studies





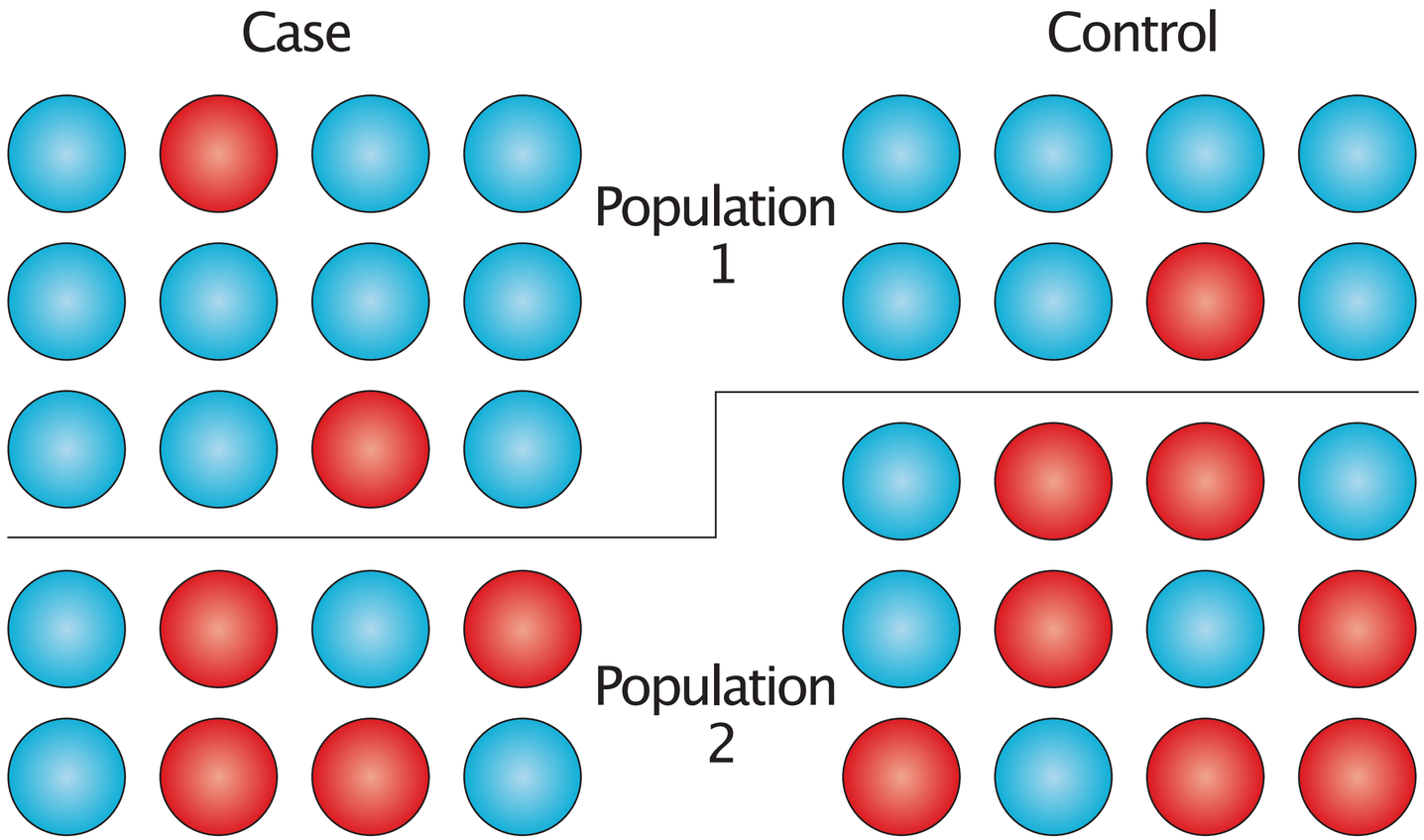
Are we assuming a homogeneous population?

GWAS data/results increased rapidly



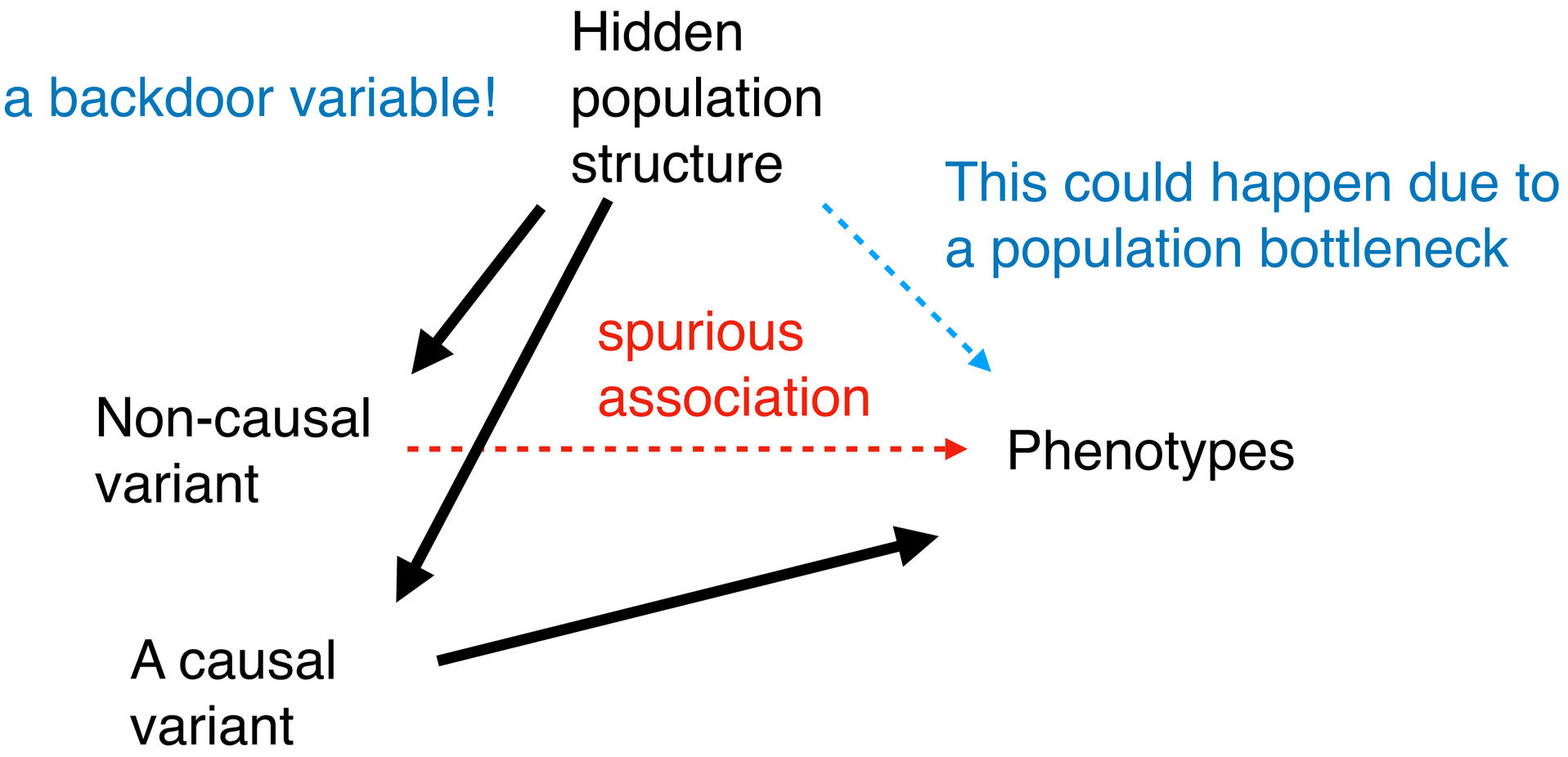
Vischer *et al.* (2017)

What if there were a hidden population structure?



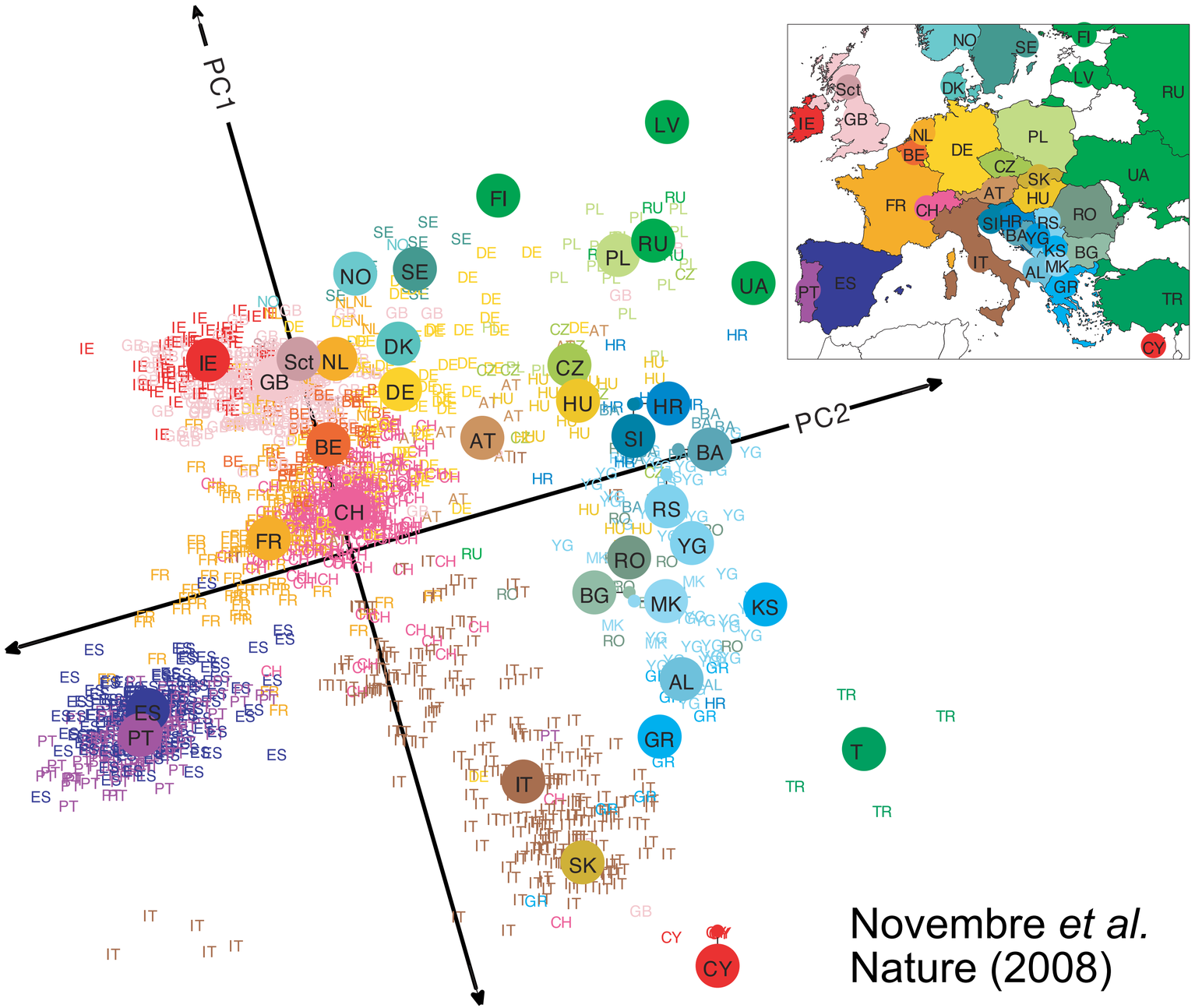
Balding, Nature Genetics Review, 2006

Hidden population structure in a causal graph



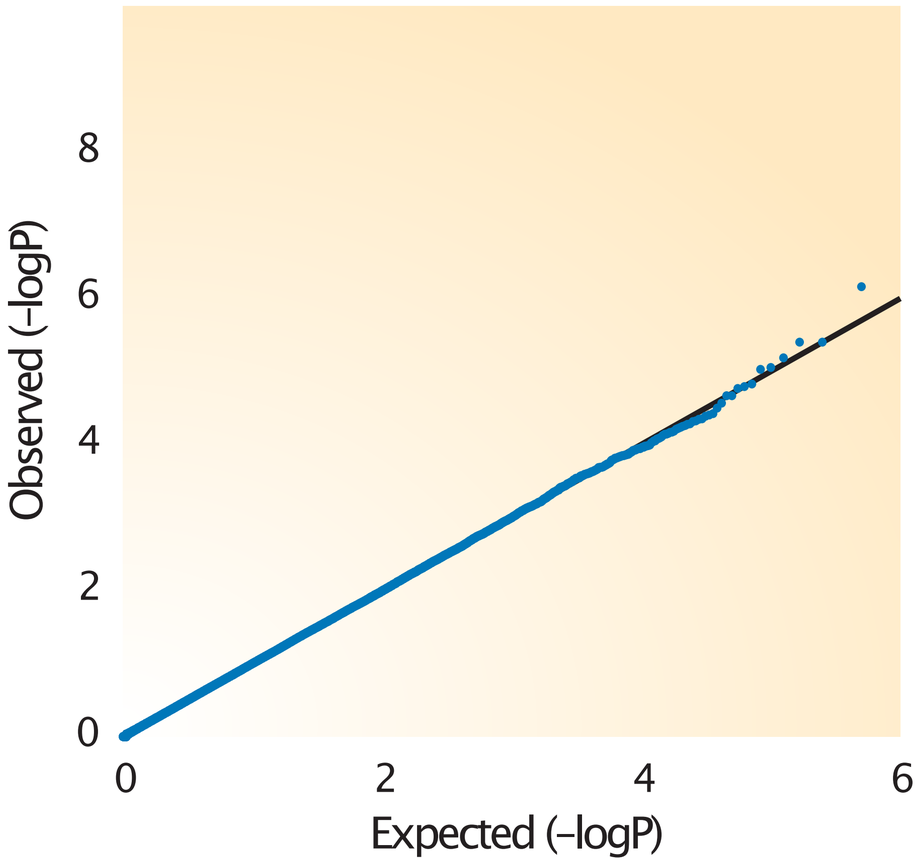
The human population as a result of the human migration history

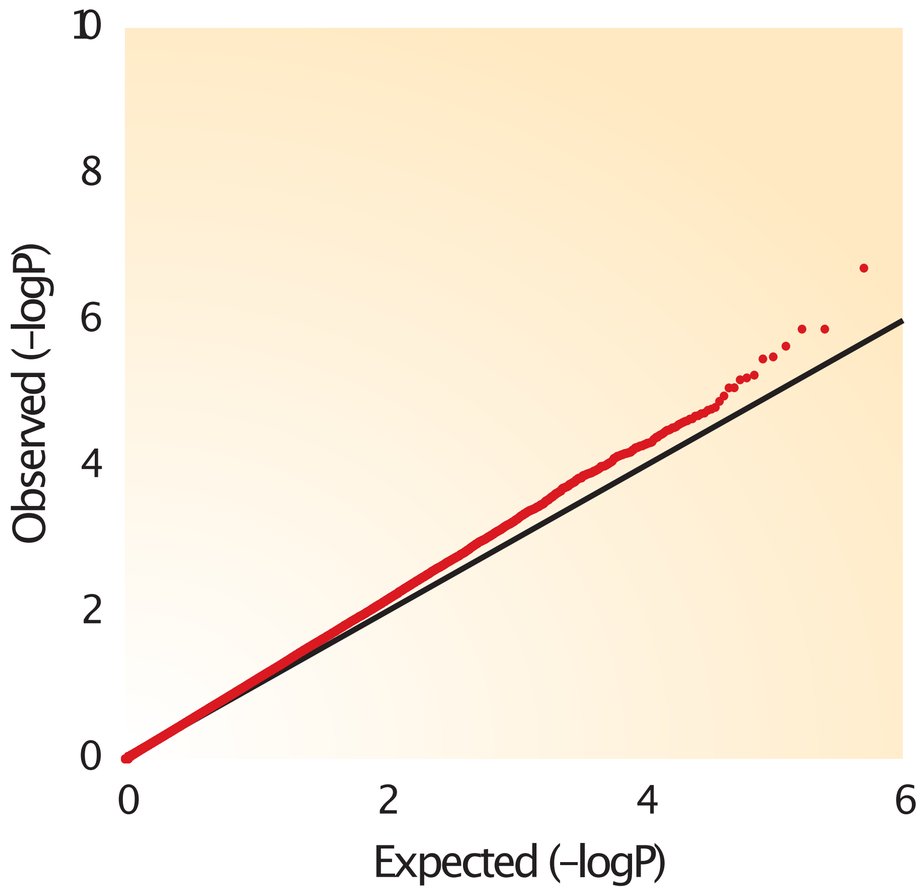
https://www.nytimes.com/2018/03/20/science/david-reich-human-migrations.html



Novembre *et al.*
Nature (2008)

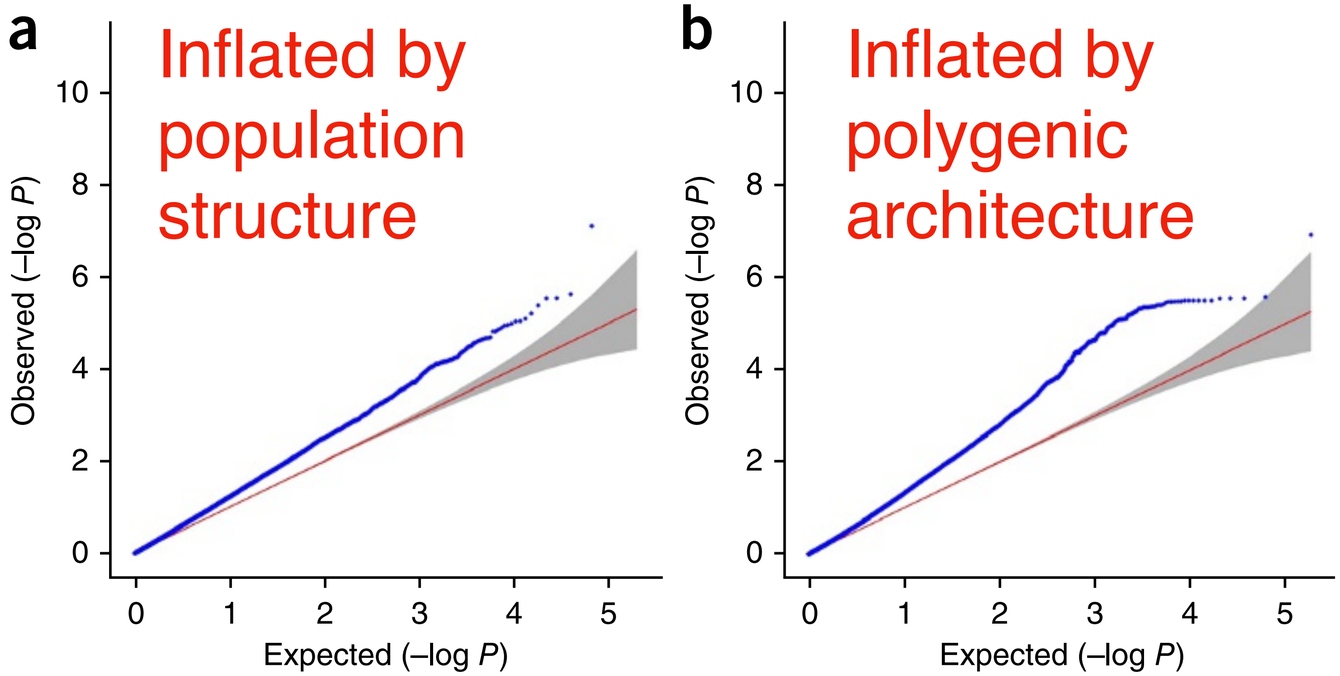
Population structures may inflate GWAS stats





Price, Zaitlen, Reich, Patterson

Population structures and polygenicity may inflate GWAS stats



Unmeasured confounding

Simply there are so many causal variants

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

What could be a generative model?

genotype

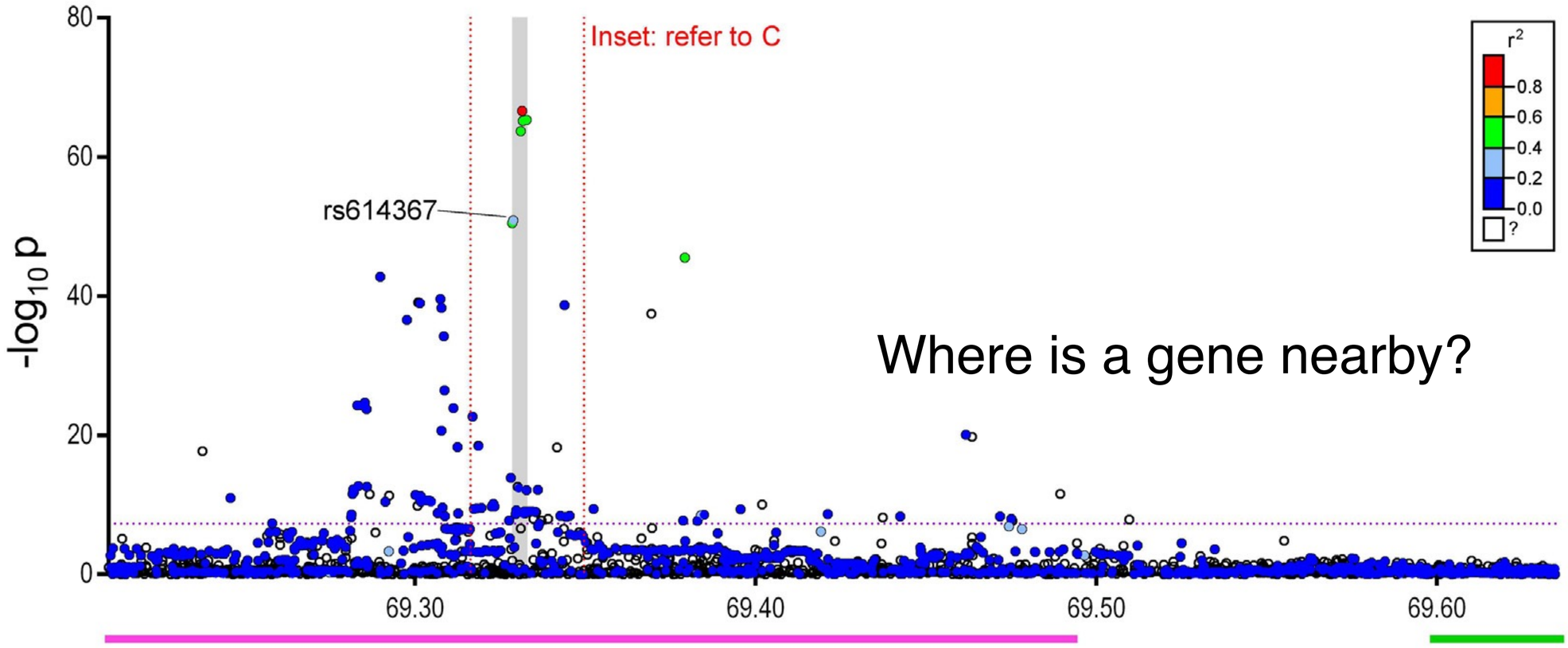
$$Y_i = \sum_j X_{ij}\beta_j + \sum_k U_{ik}\gamma_k + \epsilon_i$$

phenotype

popuation

$$X_{ij} = \sum_k U_{ik}\alpha_k + \delta_{ij}$$

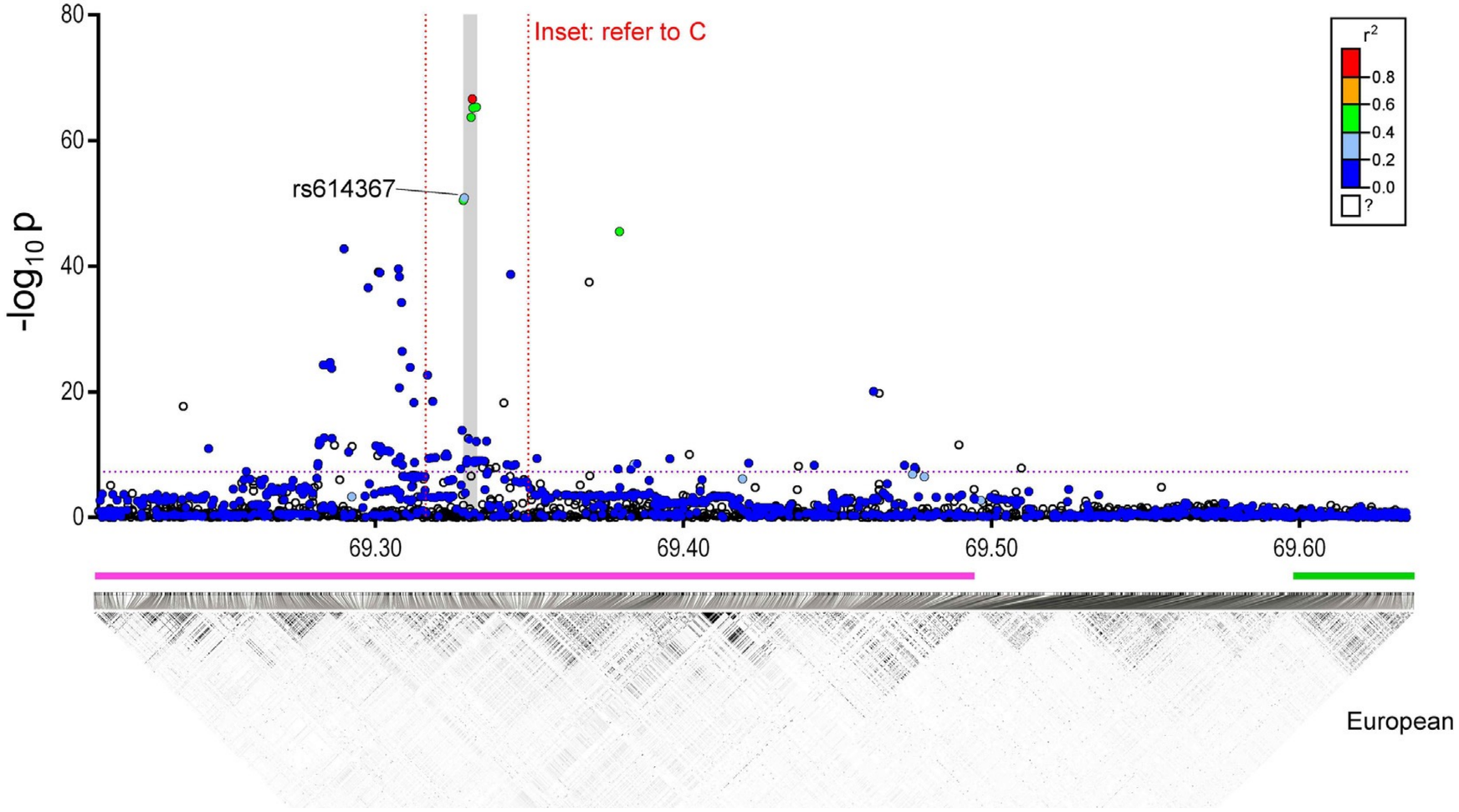
90% of GWAS hits on the non-coding regions



Where is a gene nearby?

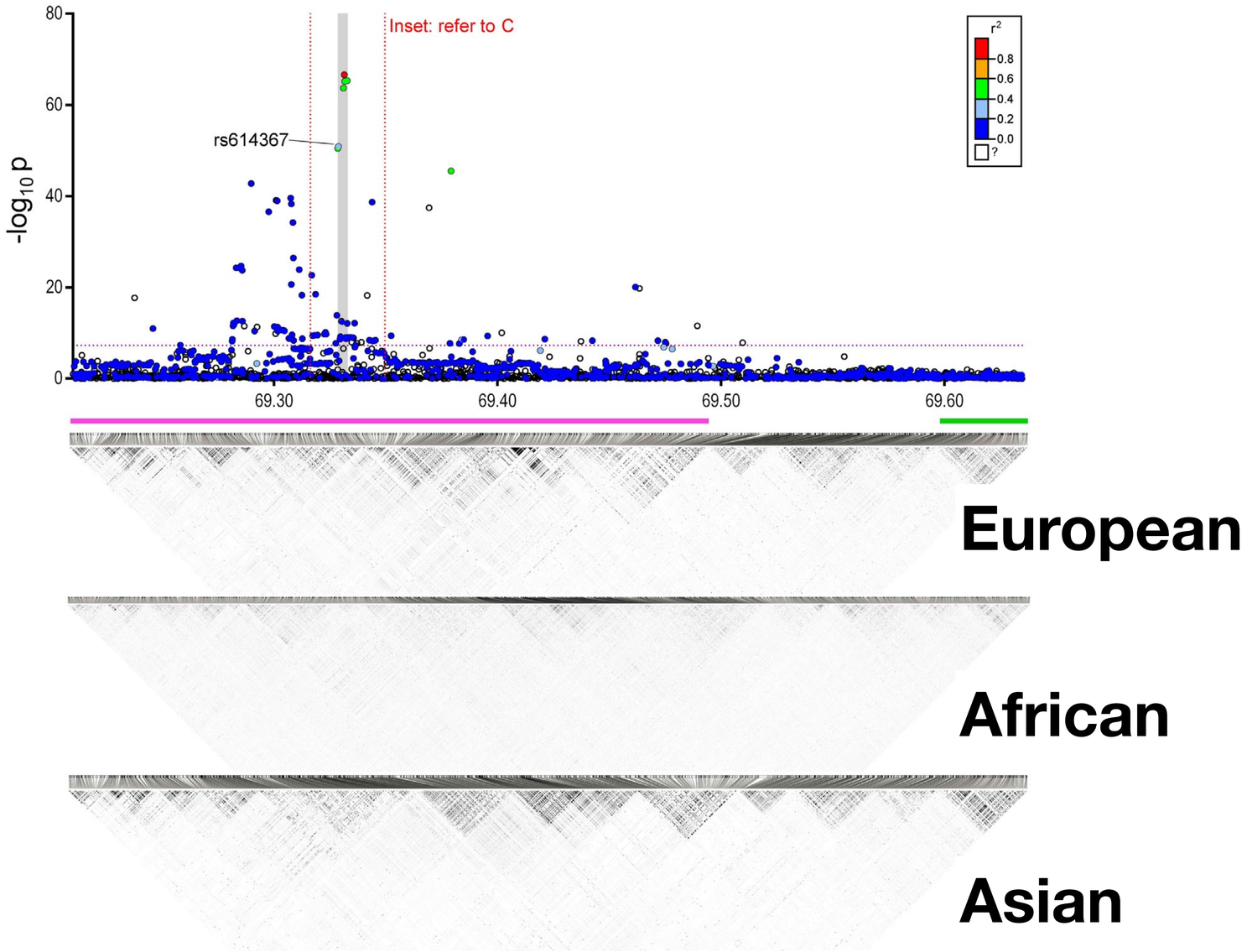
Edwards *et al. AJHG* (2013)

LD structure, many significantly associated variants



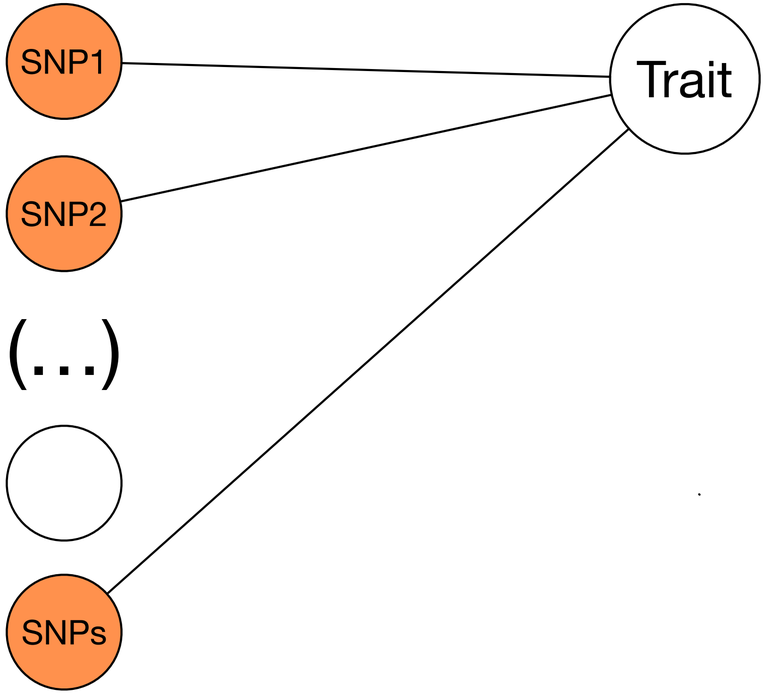
Edwards *et al. AJHG* (2013)

Different population structures, different LD patterns



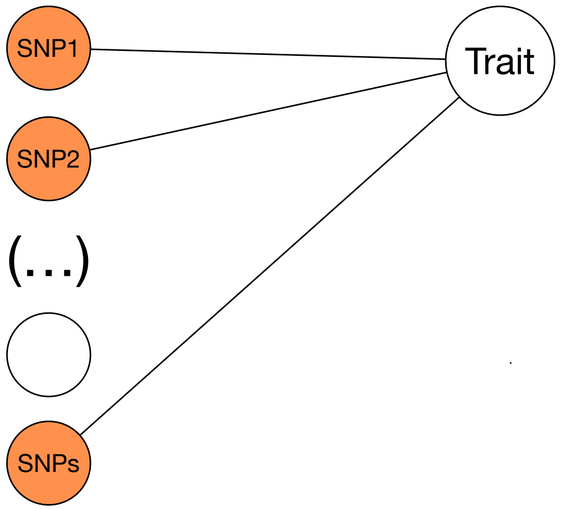
Edwards *et al. AJHG* (2013)

Many complex traits are polygenic



A missing heritability problem

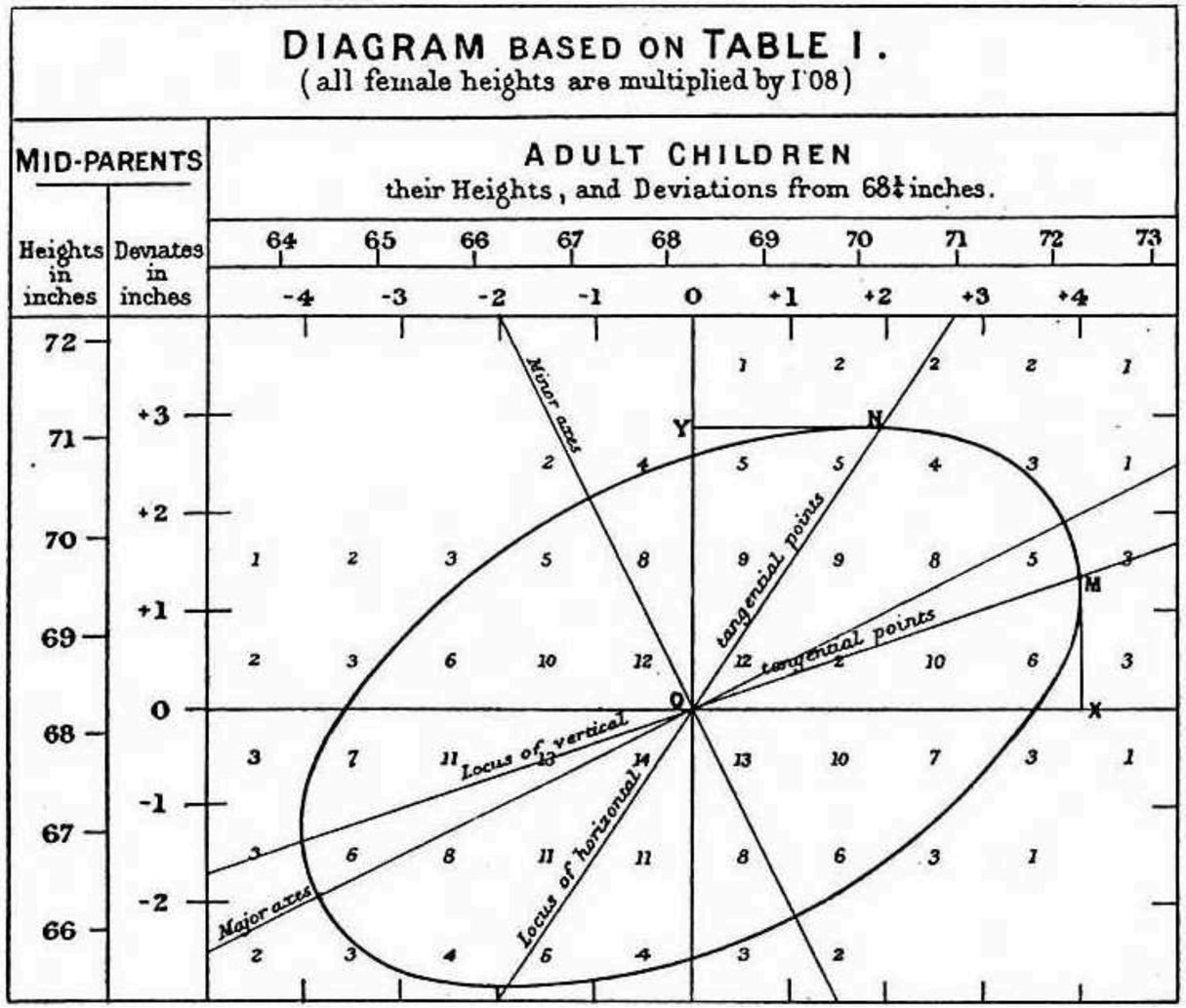
GWAS heritability $\ll$ Twin study heritability



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

The case of the missing heritability

Galton's regression model finally confirmed



Parents' height

Children's height

Is there a "tall/small" gene?

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)

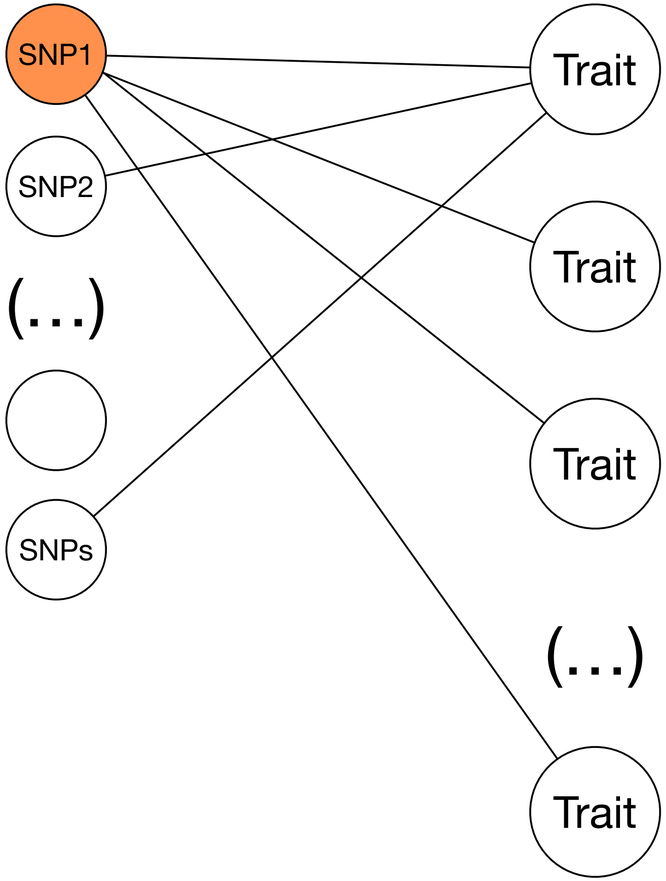
[a]log-likelihood under the null hypothesis that $\sigma_g^2=0$. [b]log-likelihood under the alternative hypothesis that $\sigma_g^2 \neq 0$; [c]log-likelihood ratio test statistic, $LRT = 2[L(H_1) - L(H_0)]$. [d]Estimate of variance explained by all SNPs, with its s.e. given in the parentheses. [e]Raw 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 (next week)

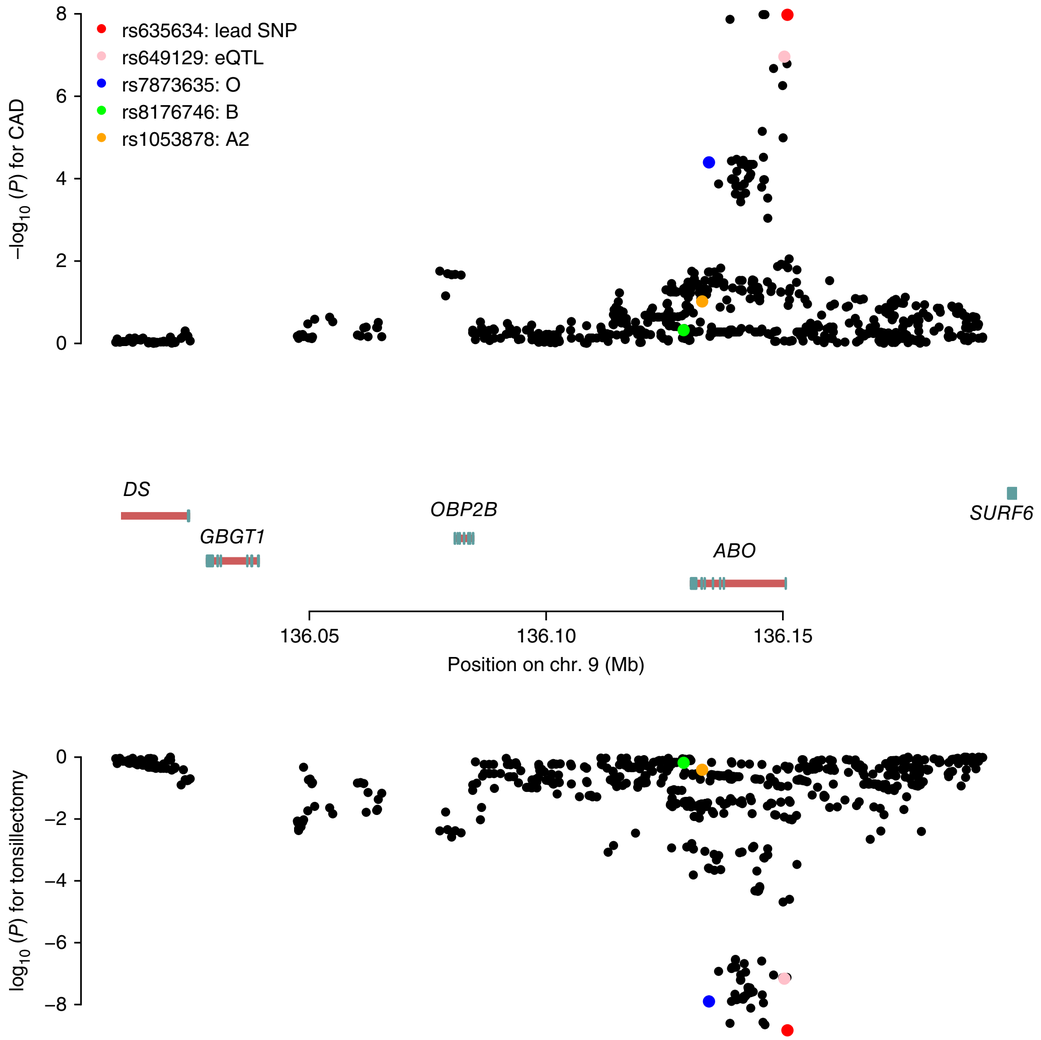
A single variant may act on many traits

Pleiotropy

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

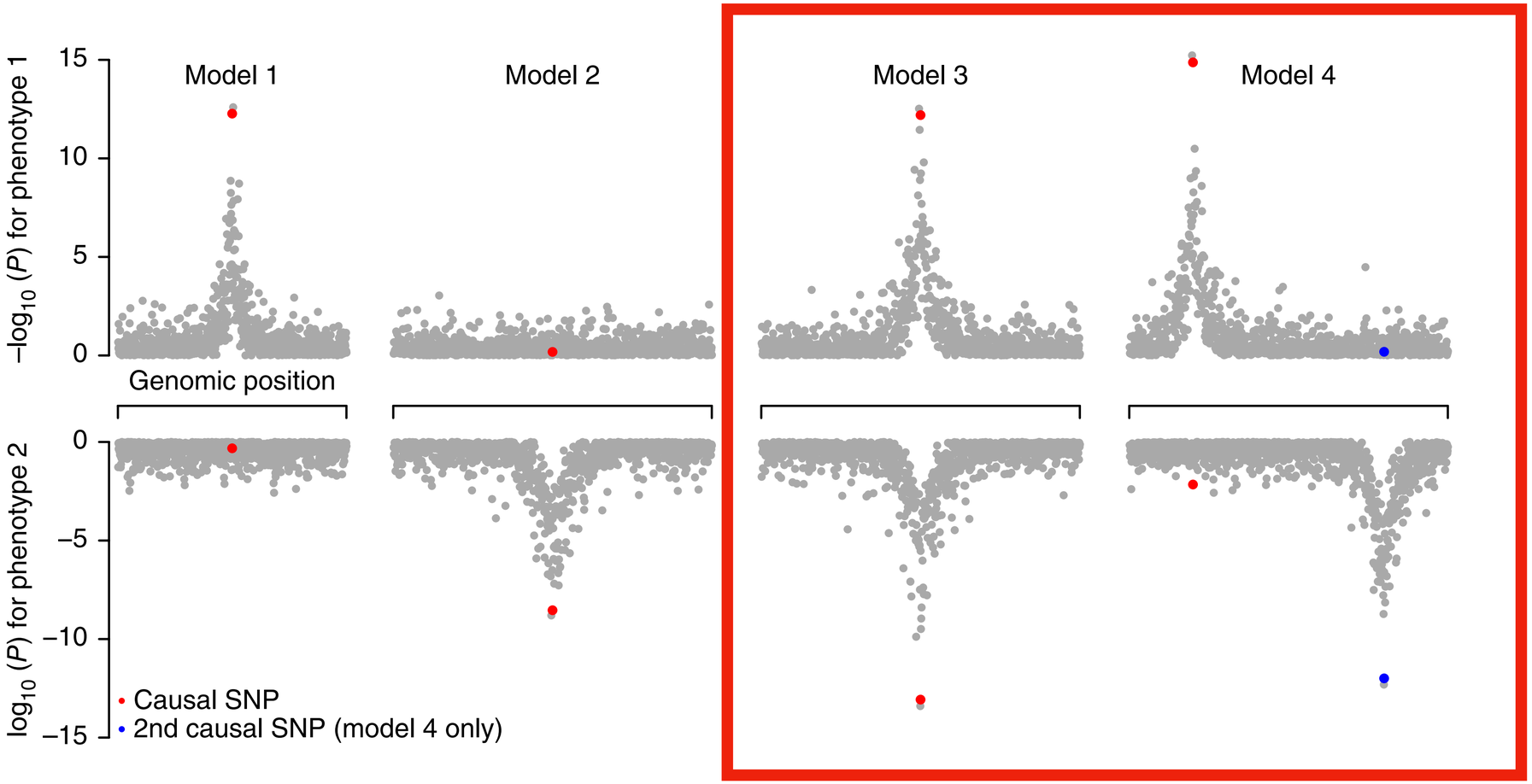


A single variant may act on many traits



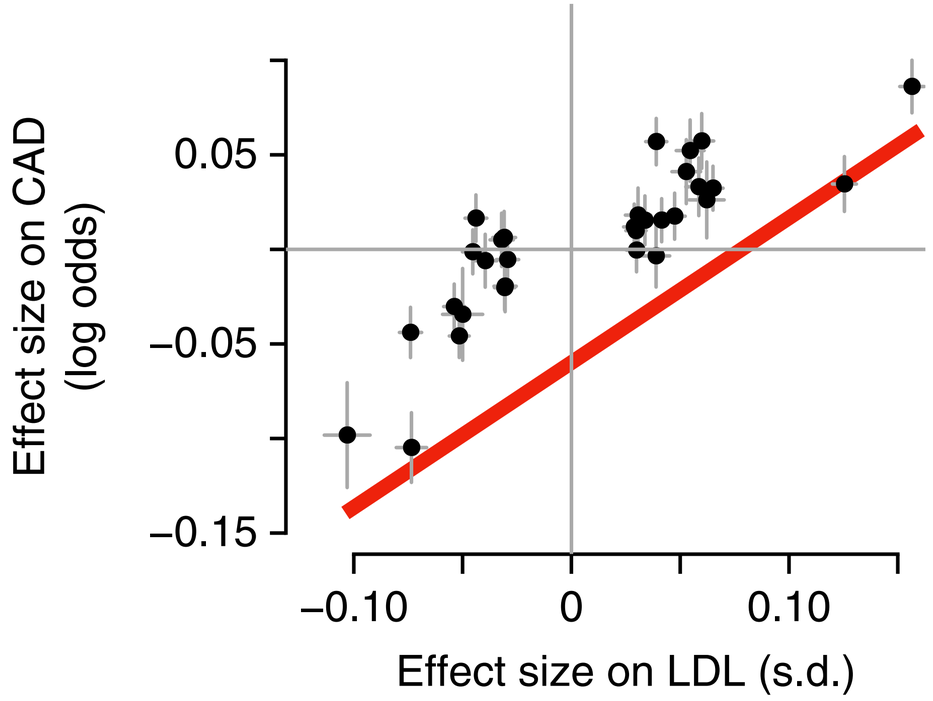
Pickrell *et al. Nature Genetics* (2016)

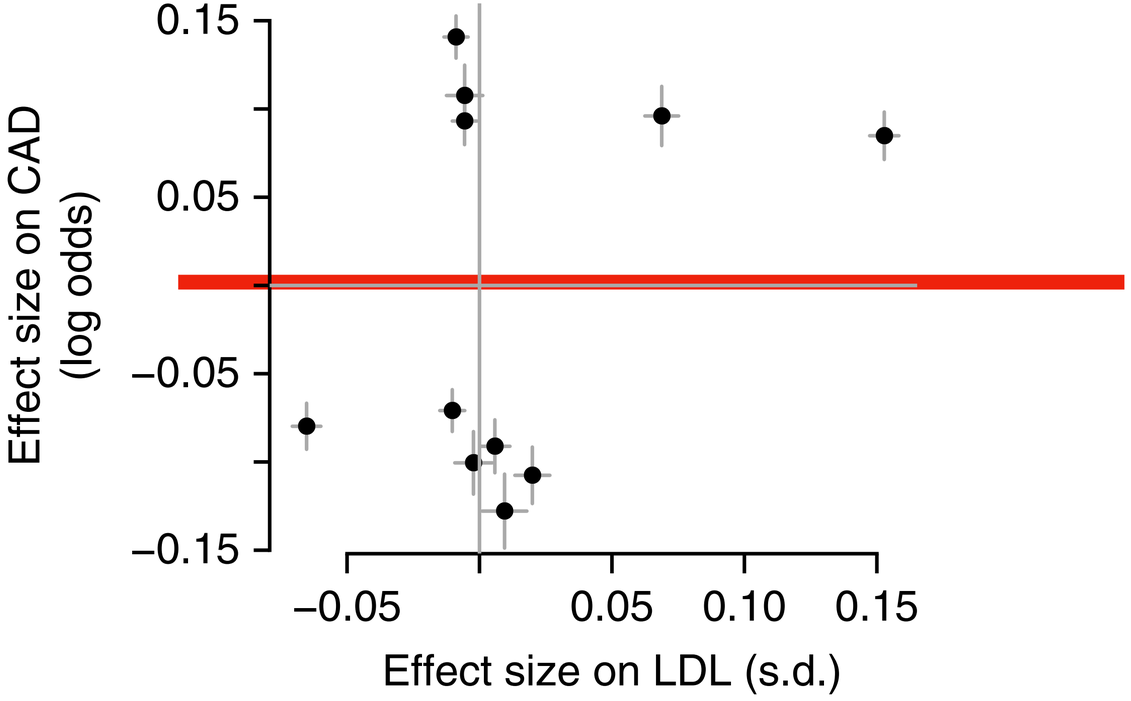
A single variant may act on many traits



Pickrell *et al. Nature Genetics* (2016)

Comparing GWAS effect sizes between traits may shed light on resolving causal directions





Pickrell *et al. Nature Genetics* (2016)

LDL cholesterol → Coronary Artery Disease?

So many variants significantly associated with traits

Table 1 | Count and proportion of pleiotropic trait-associated loci, genes, SNPs and gene sets

	Loci		Genes		SNPs		Gene set	
	Length (Mb)	%	Count	%	Count	%	Count	%
Total in genome	2,796.10	100.00	17,518	100.00	1,740,179	100.00	10,086	100.00
Associated	1,706.97	61.05	11,544	65.90	236,638	13.60	1,030	10.21
Pleiotropic[a]	1,593.32	93.34	9,374	81.20	142,439	60.19	587	56.99
Multidomain	1,537.06	90.04	7,754	67.17	76,703	32.41	353	34.27
Domain specific	56.26	3.30	1,620	14.03	65,736	27.78	234	22.72
Trait specific	113.64	6.66	2,170	18.80	94,199	39.81	443	43.01
Nonassociated	1,089.13	38.95	5,974	34.10	1,503,541	86.40	9,056	89.79

[a]The count of pleiotropic loci, genes, SNPs and gene sets is the sum of the multidomain and domain-specific categories. Proportion of pleiotropic, multidomain, domain-specific and trait-specific categories are relative to the associated loci, SNPs, genes or gene sets, respectively.

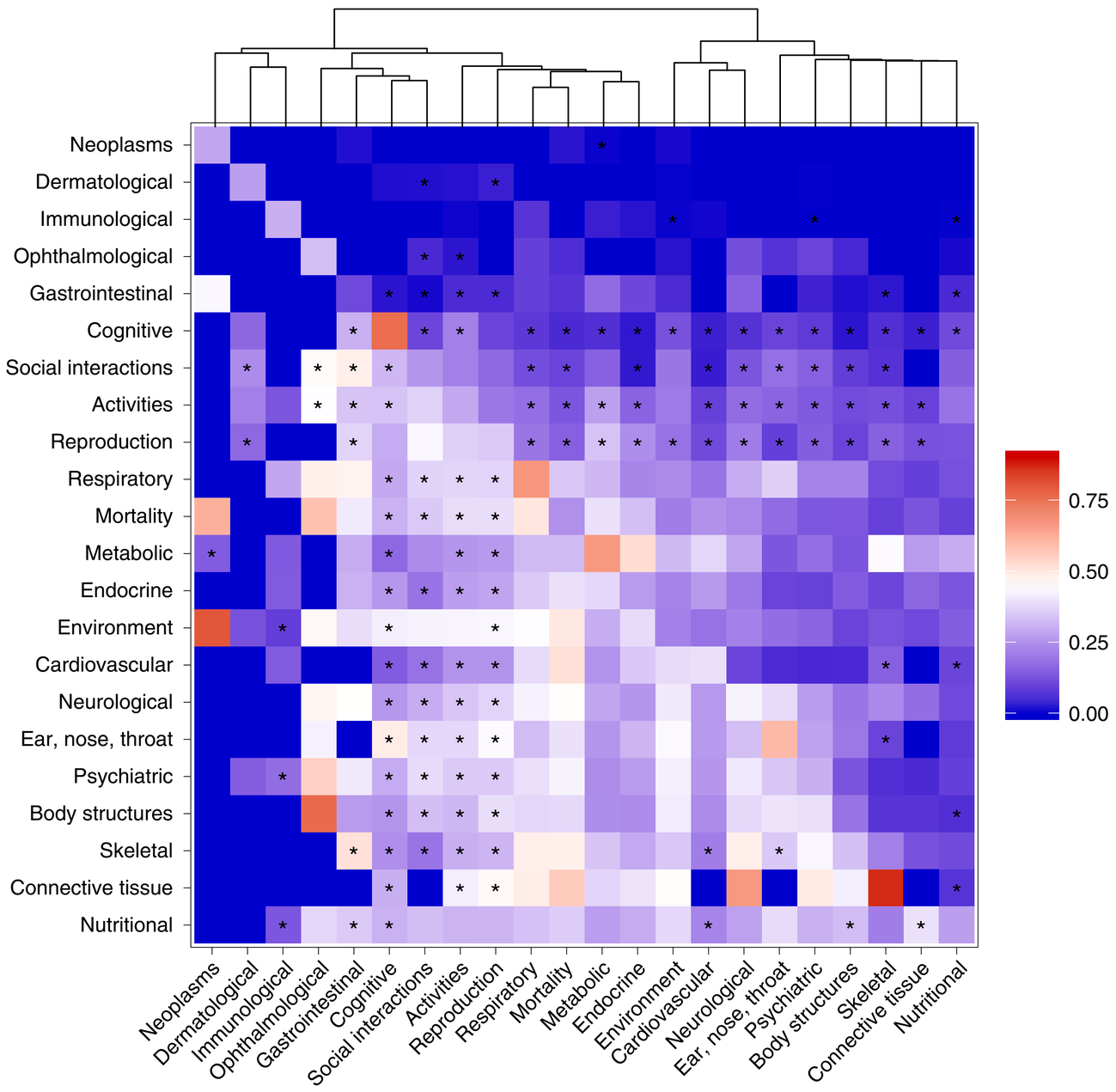
A global overview of pleiotropy and genetic architecture in complex traits

Kyoko Watanabe[1], Sven Stringer[1], Oleksandr Frei[2], Maša Umićević Mirkov[1], Christiaan de Leeuw[1], Tinca J. C. Polderman[1], Sophie van der Sluis[1,3], Ole A. Andreassen[2,4], Benjamin M. Neale[5,6,7] and Danielle Posthuma[1,3*]

> 90% of GWAS variants are pleiotropic!

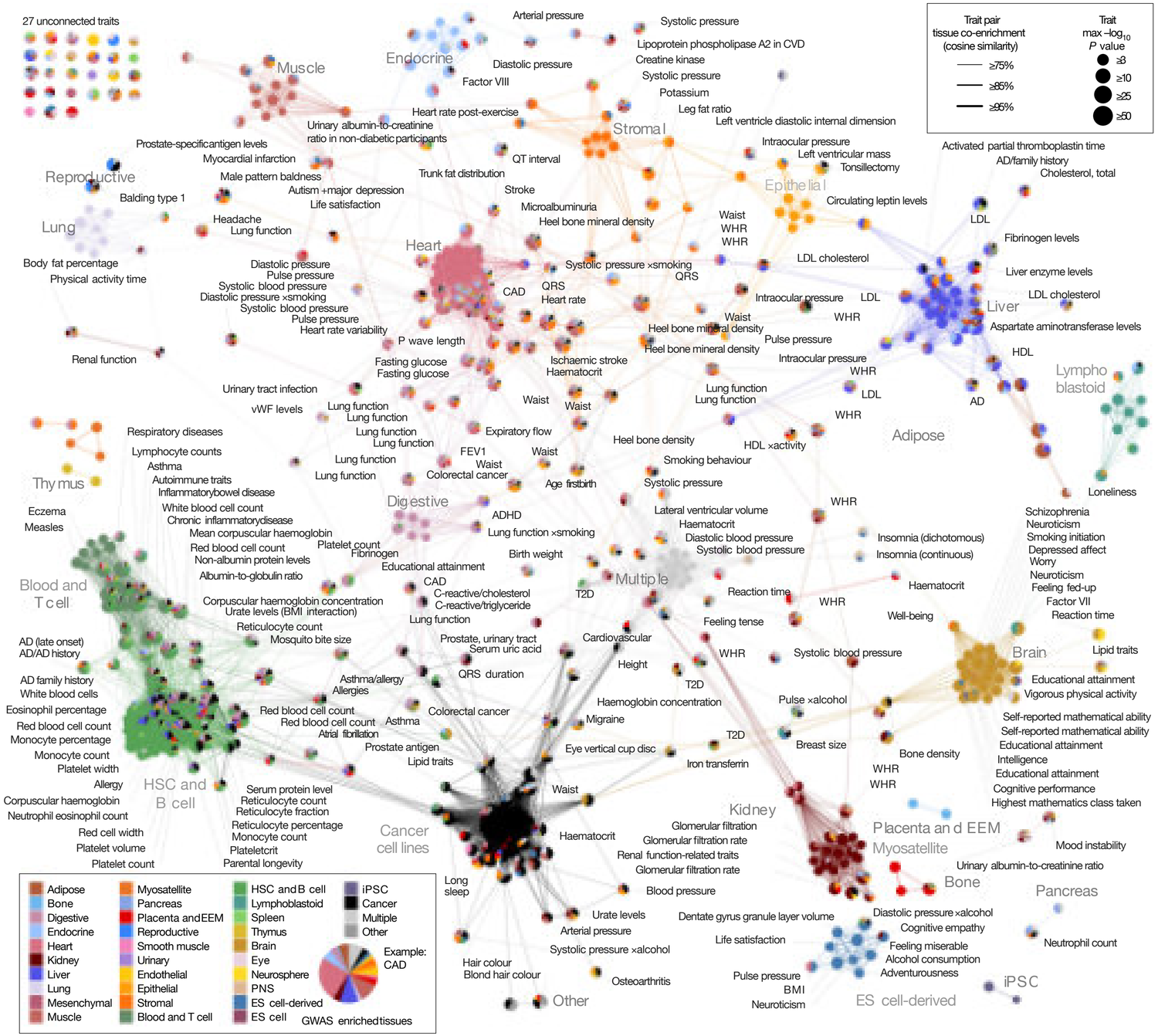
Watanabe *et al. Nature Genetics* (2019)

Many common variants induce genetic correlation across different traits



Watanabe *et al. Nature Genetics* (2019)

Most traits are also connected with other traits because of pleiotropy



Boix *et al. Nature* (2021)

Big biobank-driven GWAS on prospective cohort

23andMe

deCODE genetics

biobank uk

CanPath
Canadian Partnership for Tomorrow's Health

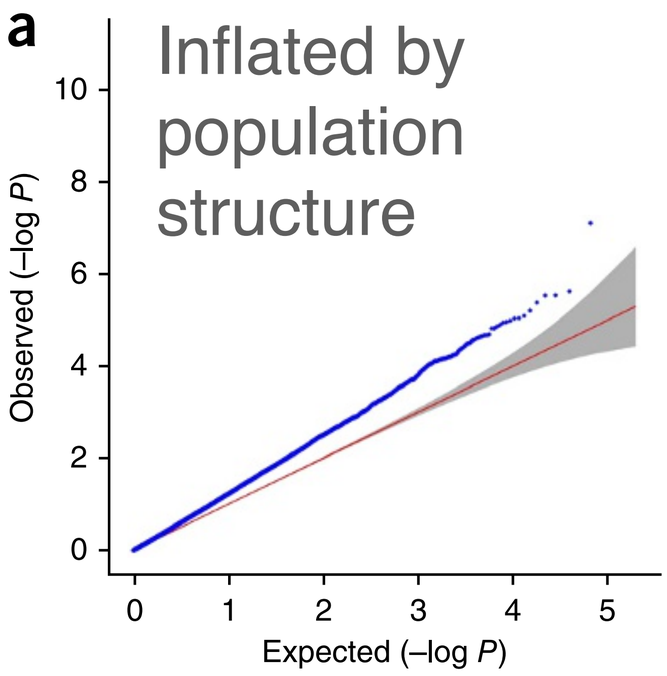
MVP

BIOBANK JAPAN

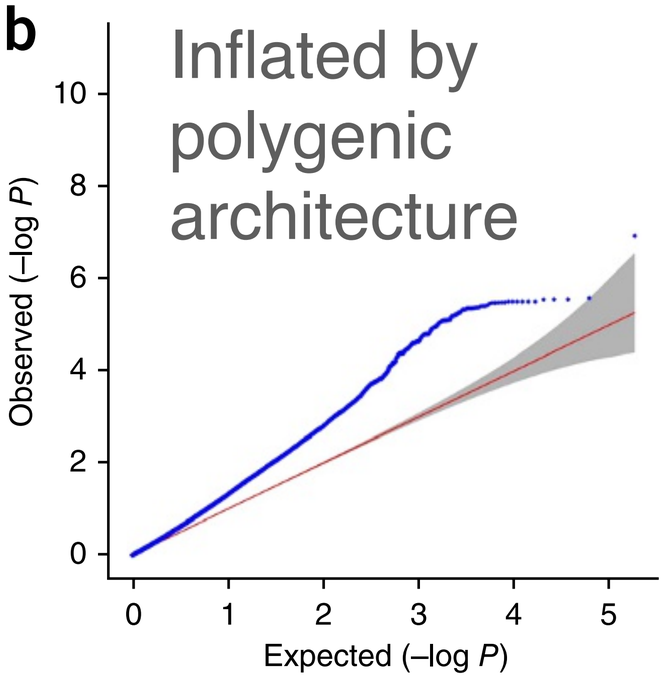
FIMM

All of Us
RESEARCH PROGRAM

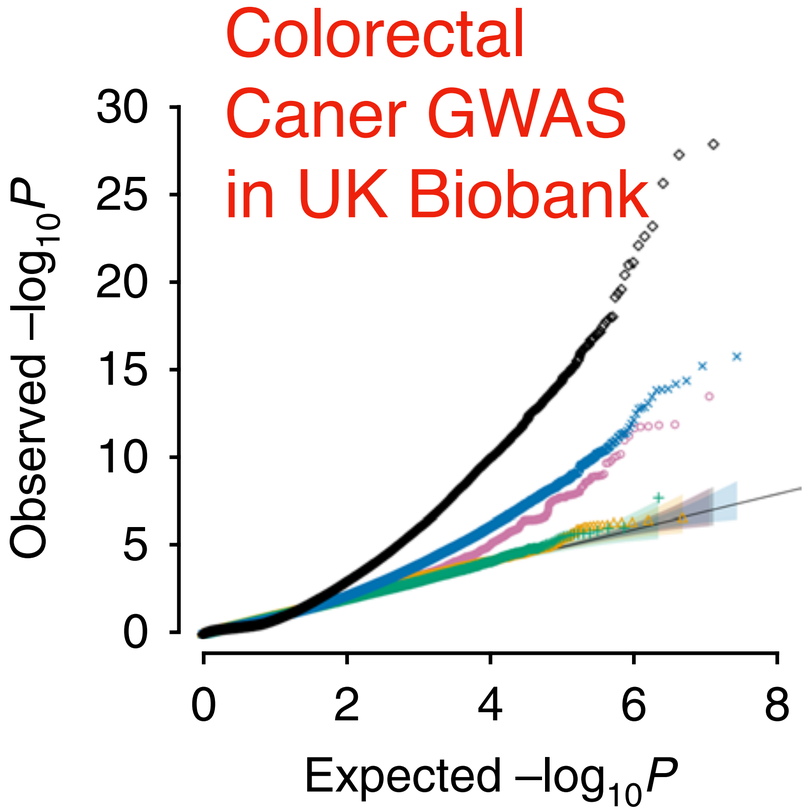
Case-control imbalance can inflate GWAS stat



Unmeasured confounding



Simply there are so many causal variants

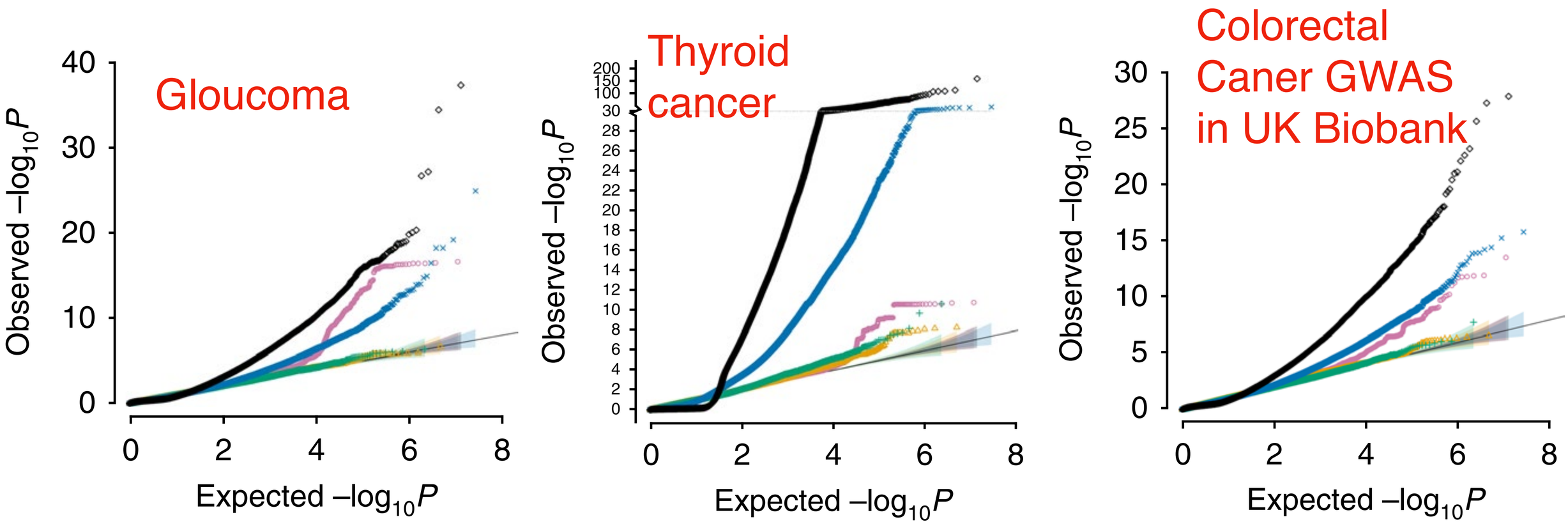


Too few cases vs. too many controls

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

Zhou *et al. Nature Genetics* (2018)

Case-control imbalance can inflate GWAS stat



Too few cases vs. too many controls

Zhou *et al. Nature Genetics* (2018)

Today's lecture

- Fundamentals in Population Genetics
- Genome-wide association studies
- What are the limitations & challenges in GWAS?