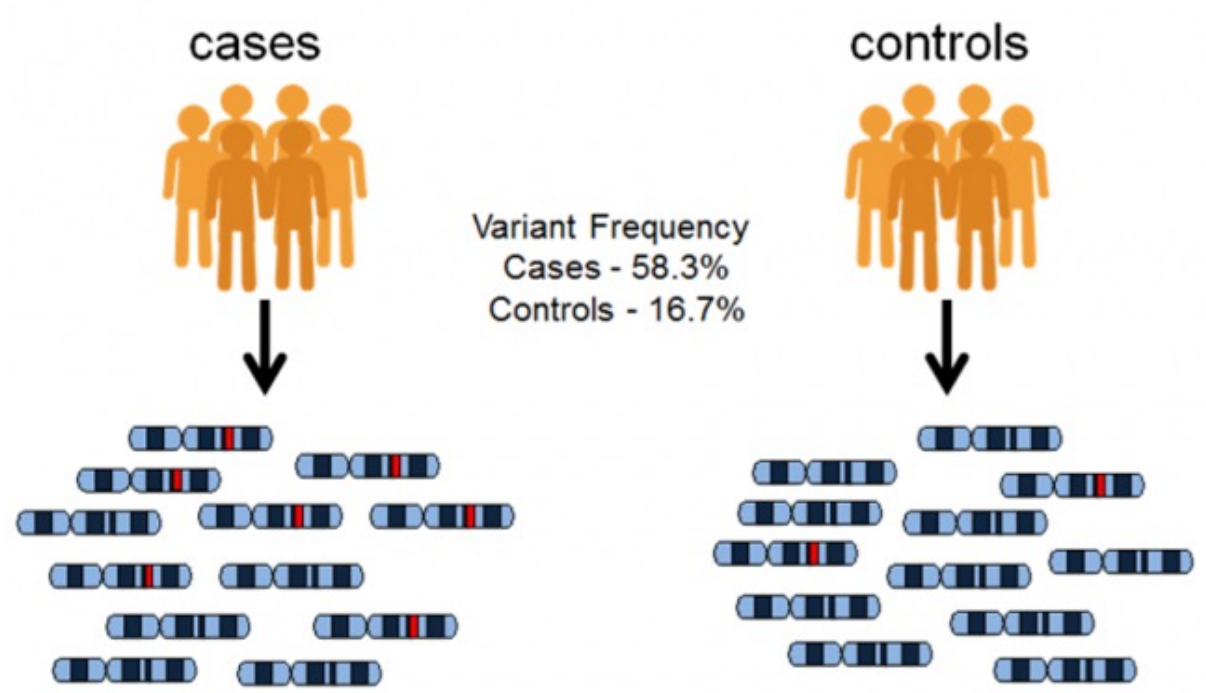
Genome-wide association studies (GWAS)

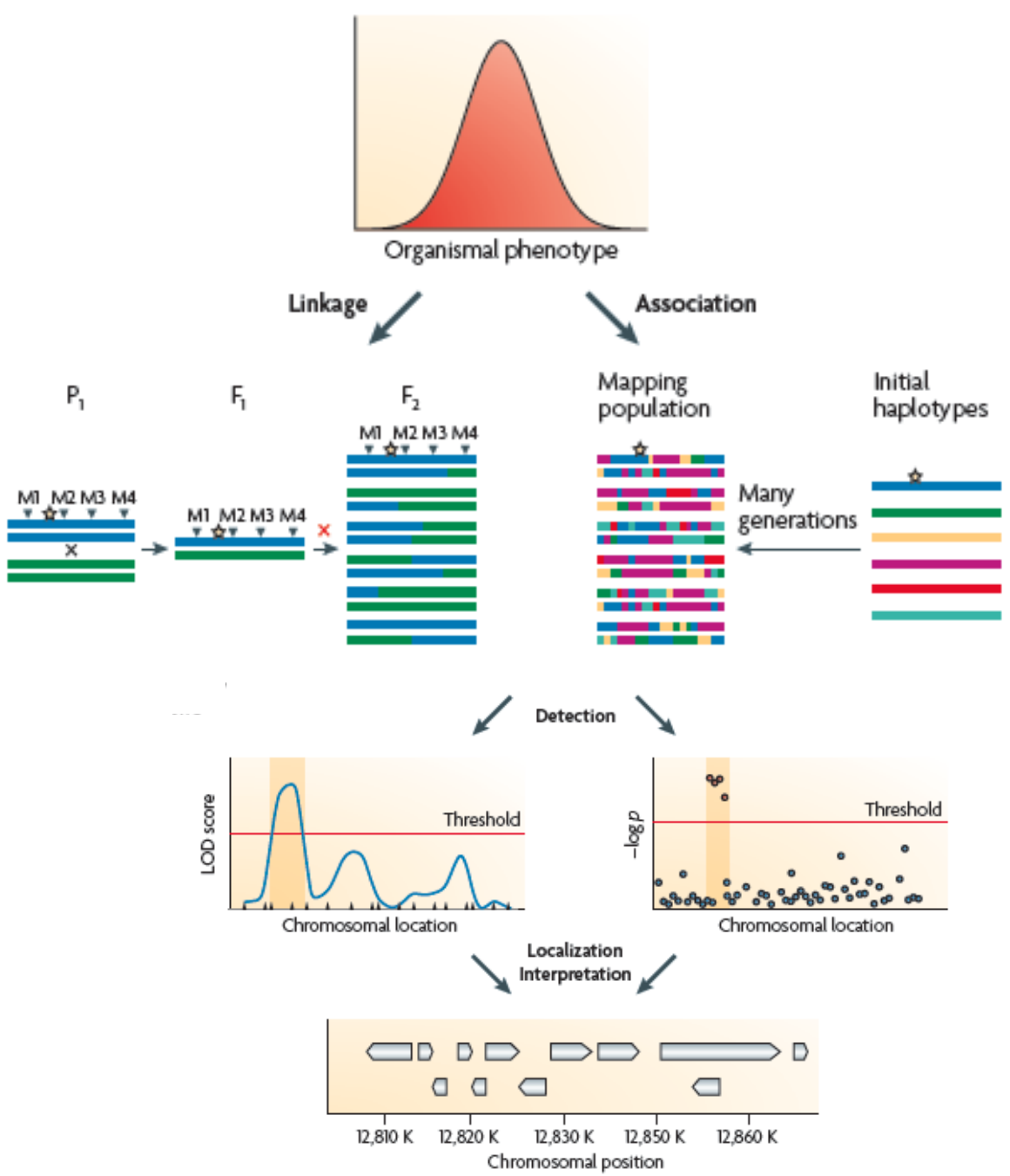


Hancock

18. April 2024

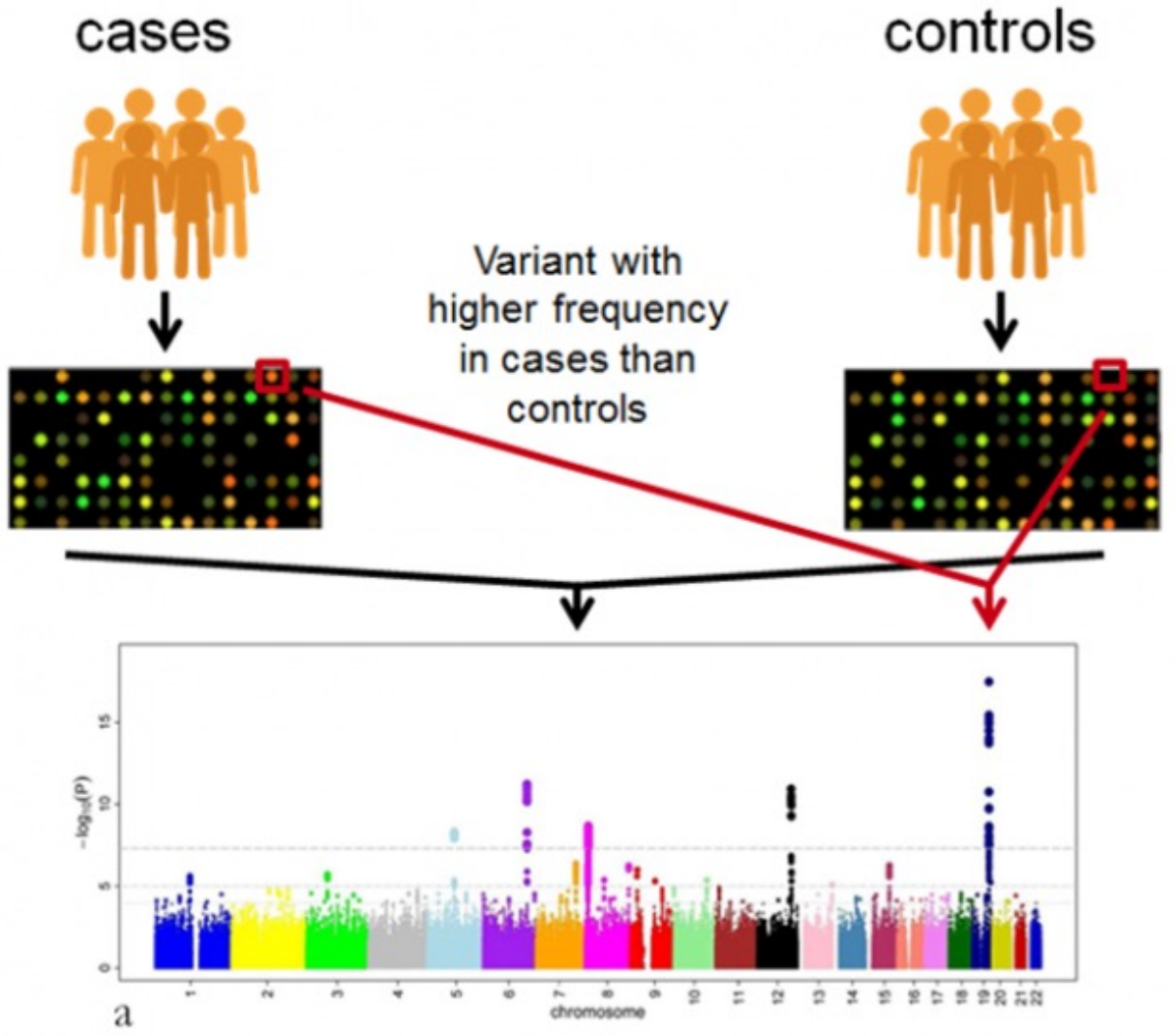
Linkage versus association mapping

- Linkage mapping uses information from a pedigree while association mapping uses information from ‘unrelated’ individuals in a population
- In linkage mapping, the genetic architecture may be simpler and population structure is not a problem, but the resolution is low due to the limited number of recombination events on the necessarily short time scale

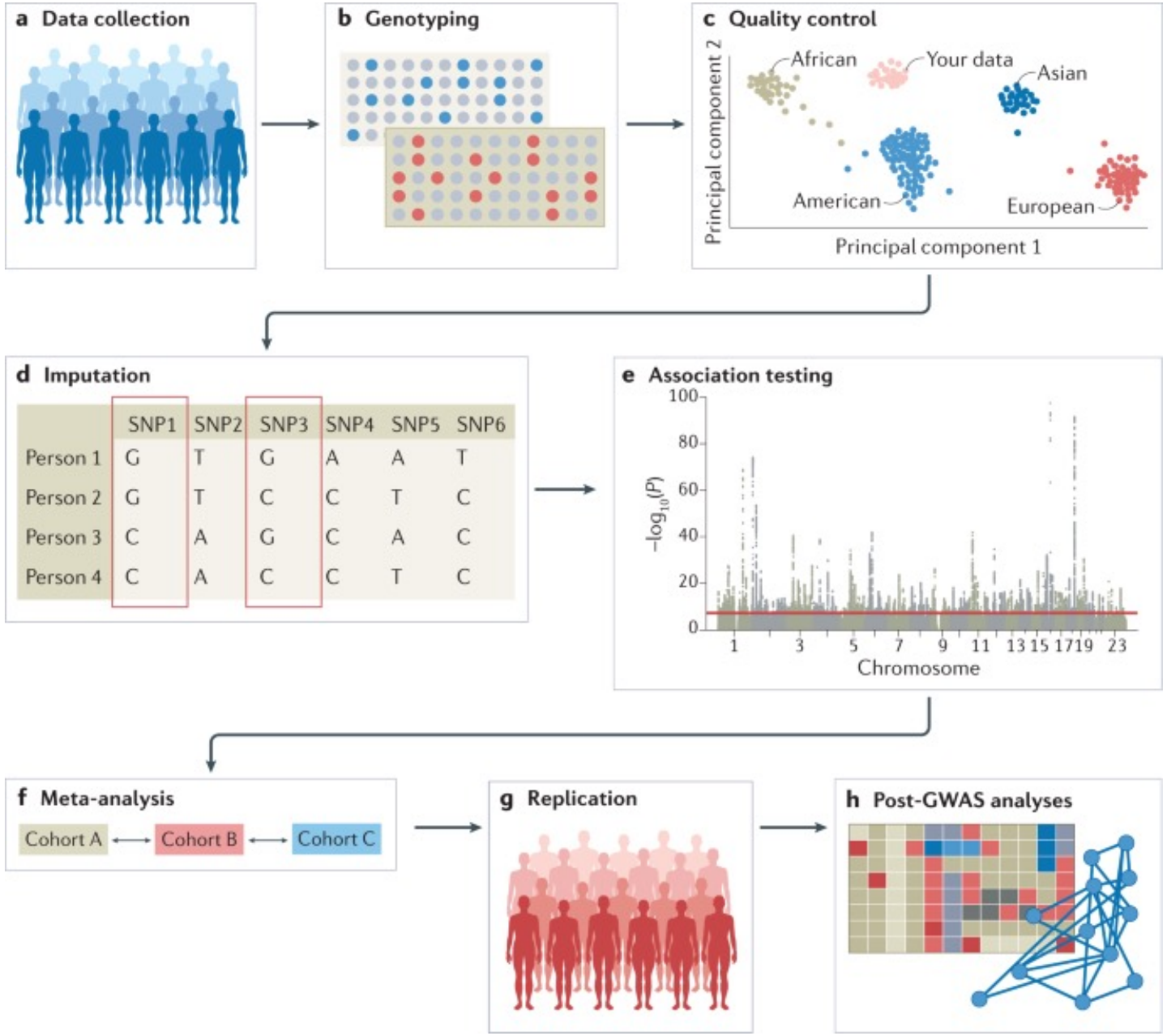


McKay et al., NRG 2009

Genome-wide association studies (GWAS)



Schematic of a GWAS pipeline



https://www.nature.com/articles/s43586-021-00056-9

Argument for genome-wide association studies

- Risch and Merikangas argued that we can use the information from the Human Genome Project
- They argued that as a next step, a project to assay common polymorphisms in human populations was needed

The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas

Geneticists have made substantial progress in identifying the genetic basis of many human diseases, at least those with conspicuous deter-

age analysis we have chosen for this argument is a popular current paradigm in which pairs of siblings, both with the disease, are

linkage analysis for loci conferring GRR of about 2 or less will never allow identification because the number of families required (more than ~2500) is not practically achievable.

Although tests of linkage for genes of modest effect are of low power, as shown by the above example, direct tests of association with a disease locus itself can still be quite strong. To illustrate this point, we use the transmission/disequilibrium test of Spielman *et al.* (3). In this test, transmission of a particular allele at a locus from heterozygous parents to their

We argue below that the method that has been used successfully (linkage analysis) to find major genes has limited power to detect genes of modest effect, but that a different approach (association studies) that utilizes candidate genes has far greater power, even if one needs to test every gene in the genome. Thus, the future of the genetics of complex diseases is likely to require large-scale testing by association analysis.

Science 1996

The HapMap Project

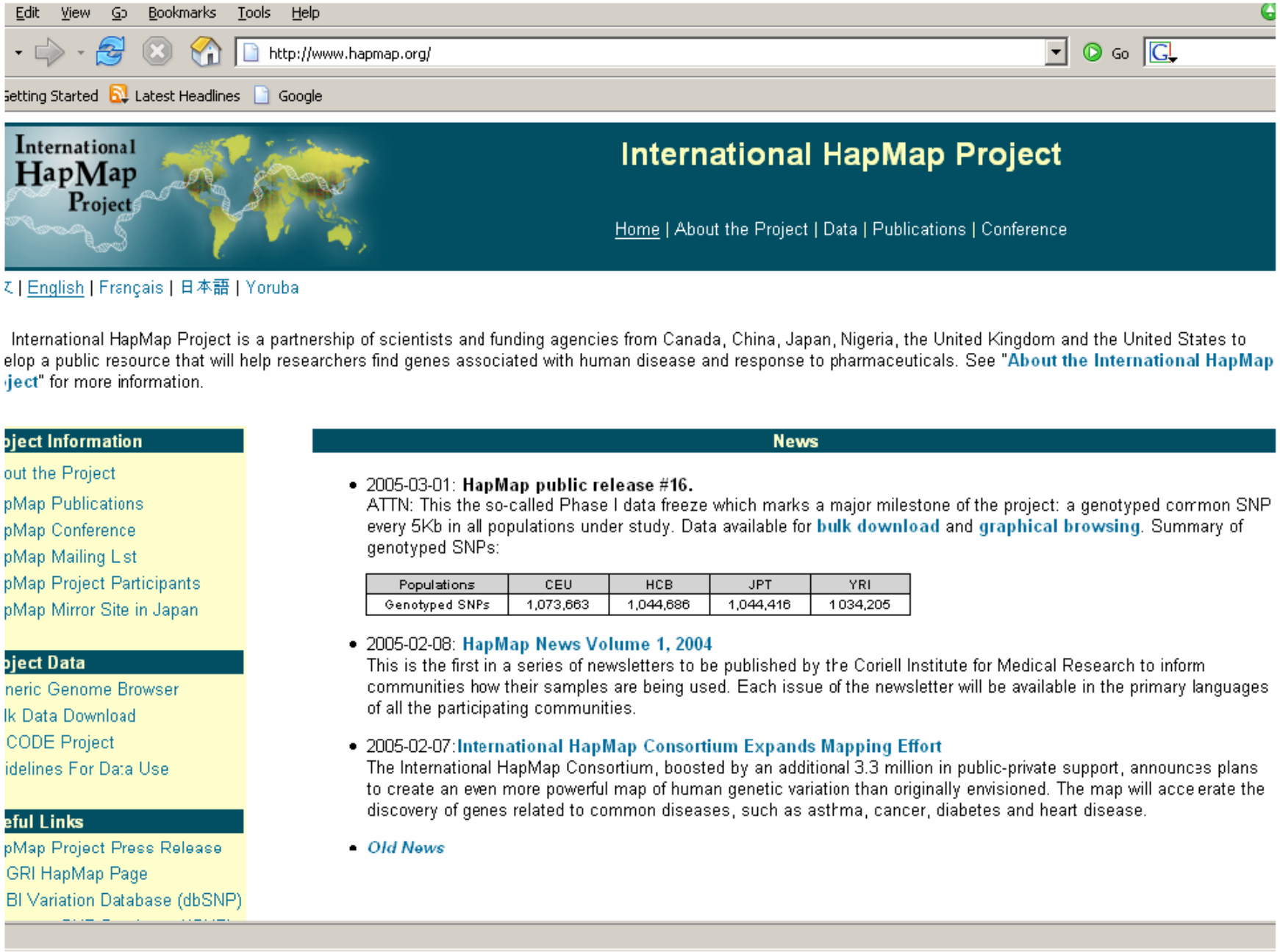
The human genome project can have more than one reward. In addition to sequencing the entire human genome, it can lead to identification of polymorphisms for all the genes in the human genome and the diseases to which they contribute. It is a charge to the molecular technologists to develop the tools to meet this challenge and provide the information necessary to identify the genetic basis of complex human diseases.

Risch and Merikangas, 1996

The **Human HapMap Project** aimed to characterize polymorphism and linkage disequilibrium across the genome with the goal of identifying common representative DNA polymorphisms that could be used for genome-wide association studies

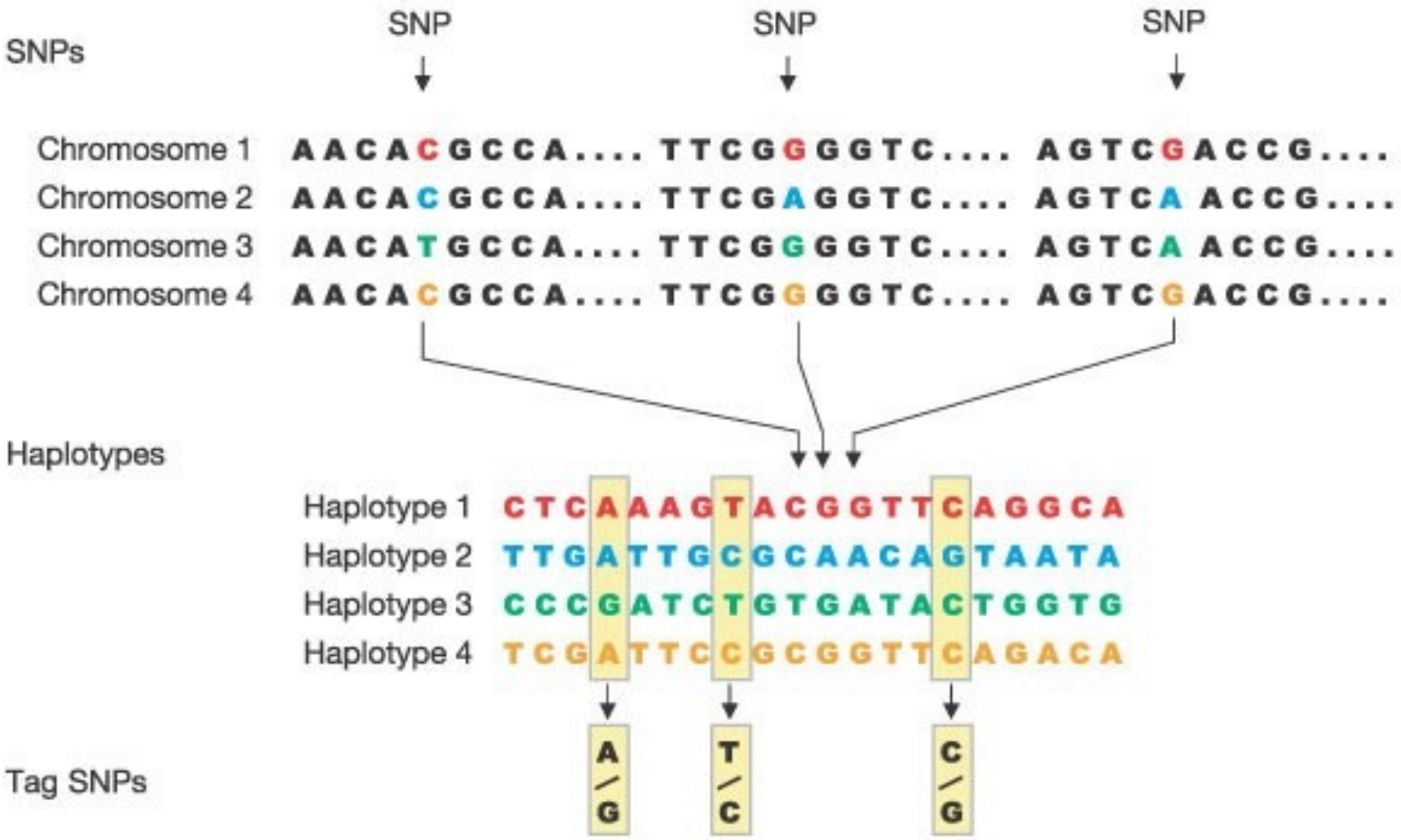
Q: Why identify ‘representative’ polymorphisms?

A: Because at the time it was infeasible to sequence entire genomes


Edit View Go Bookmarks Tools Help
http://www.hapmap.org/
Go
Getting Started
Latest Headlines
Google
International HapMap Project
International HapMap Project
Home | About the Project | Data | Publications | Conference
ζ | English | Français | 日本語 | Yoruba
International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to elop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "About the International HapMap ject" for more information.
oject Information
out the Project
pMap Publications
pMap Conference
pMap Mailing List
pMap Project Participants
pMap Mirror Site in Japan
oject Data
neric Genome Browser
lk Data Download
CODE Project
idelines For Data Use
eful Links
pMap Project Press Release
GRI HapMap Page
BI Variation Database (dbSNP)
News
2005-03-01: HapMap public release #16.
ATTN: This the so-called Phase I data freeze which marks a major milestone of the project: a genotyped common SNP every 5Kb in all populations under study. Data available for bulk download and graphical browsing. Summary of genotyped SNPs:
Populations | CEU | HCB | JPT | YRI
Genotyped SNPs | 1,073,663 | 1,044,686 | 1,044,416 | 1034,205
2005-02-08: HapMap News Volume 1, 2004
This is the first in a series of newsletters to be published by the Coriell Institute for Medical Research to inform communities how their samples are being used. Each issue of the newsletter will be available in the primary languages of all the participating communities.
2005-02-07: International HapMap Consortium Expands Mapping Effort
The International HapMap Consortium, boosted by an additional 3.3 million in public-private support, announces plans to create an even more powerful map of human genetic variation than originally envisioned. The map will accelerate the discovery of genes related to common diseases, such as asthma, cancer, diabetes and heart disease.
Old News

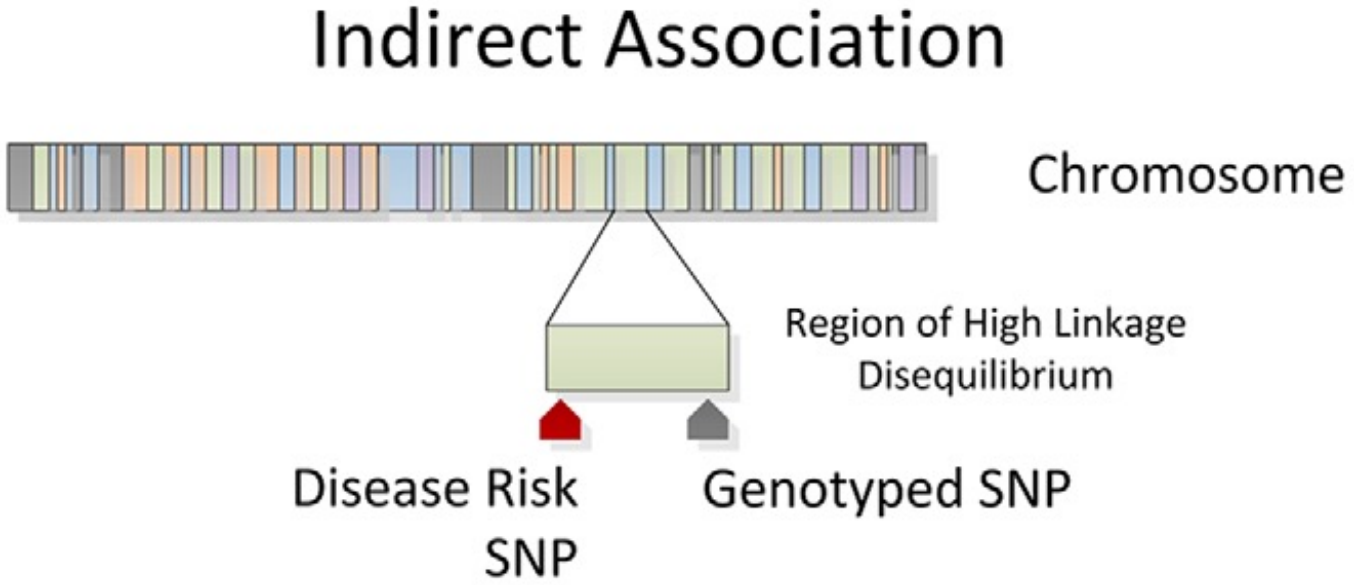
The HapMap Project Goal:

Identify SNPs that could tag haplotypes and be used to represent all the common polymorphisms in worldwide populations



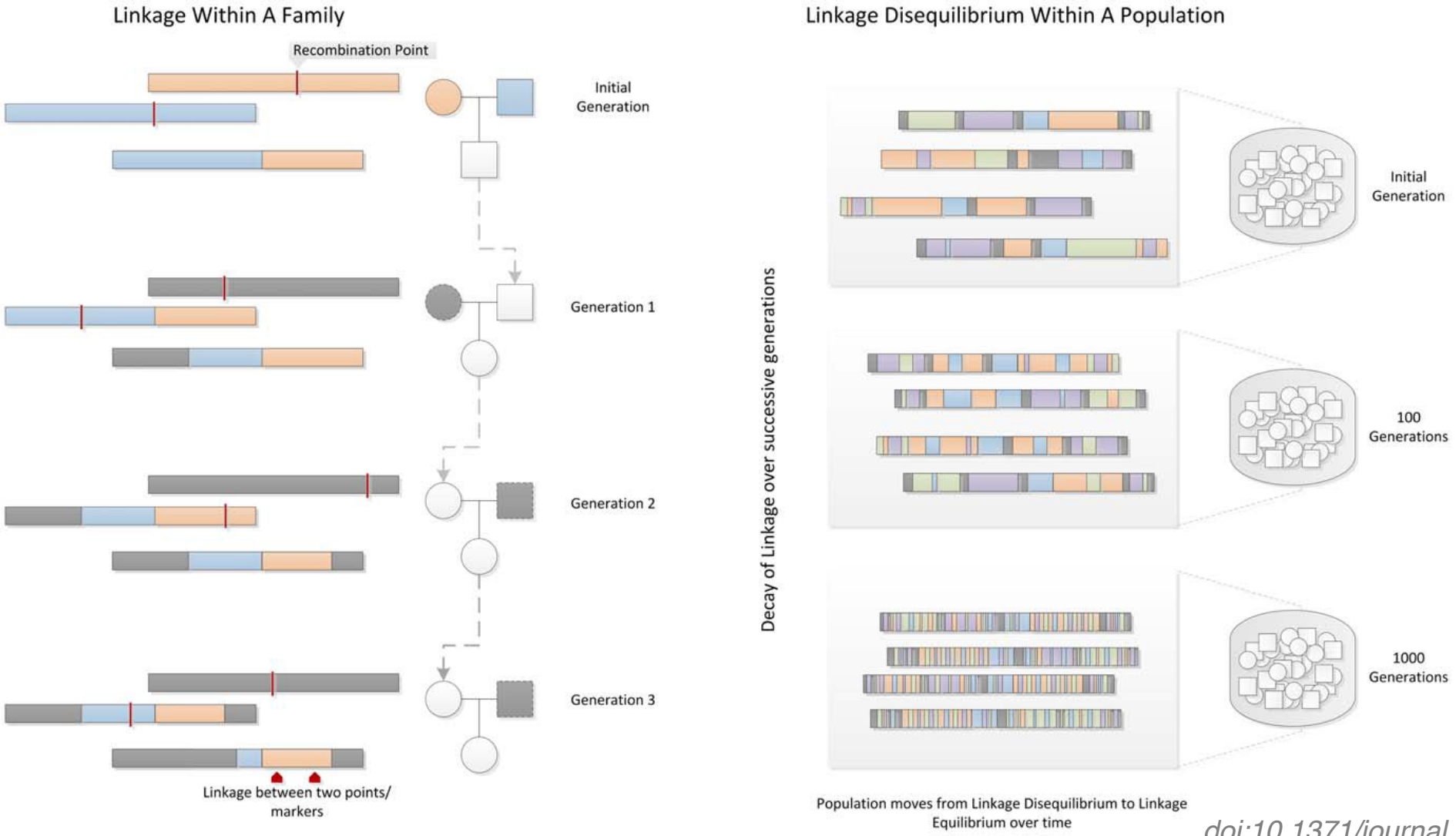
HapMap relied on the assumption that ‘tag SNPs’ could be chosen the represent linked disease SNPs

- Sequencing large enough numbers of individuals was not possible at the time, so the plan was to identify ‘tag SNPs’ that could be genotyped in large populations and represent disease SNPs
- Because LD was so central, association mapping was also called ‘**LD-mapping**’



doi:10.1371/journal.pcbi.1002822.g003

Familial linkage versus population-level LD



doi:10.1371/journal.pcbi.1002822.g002

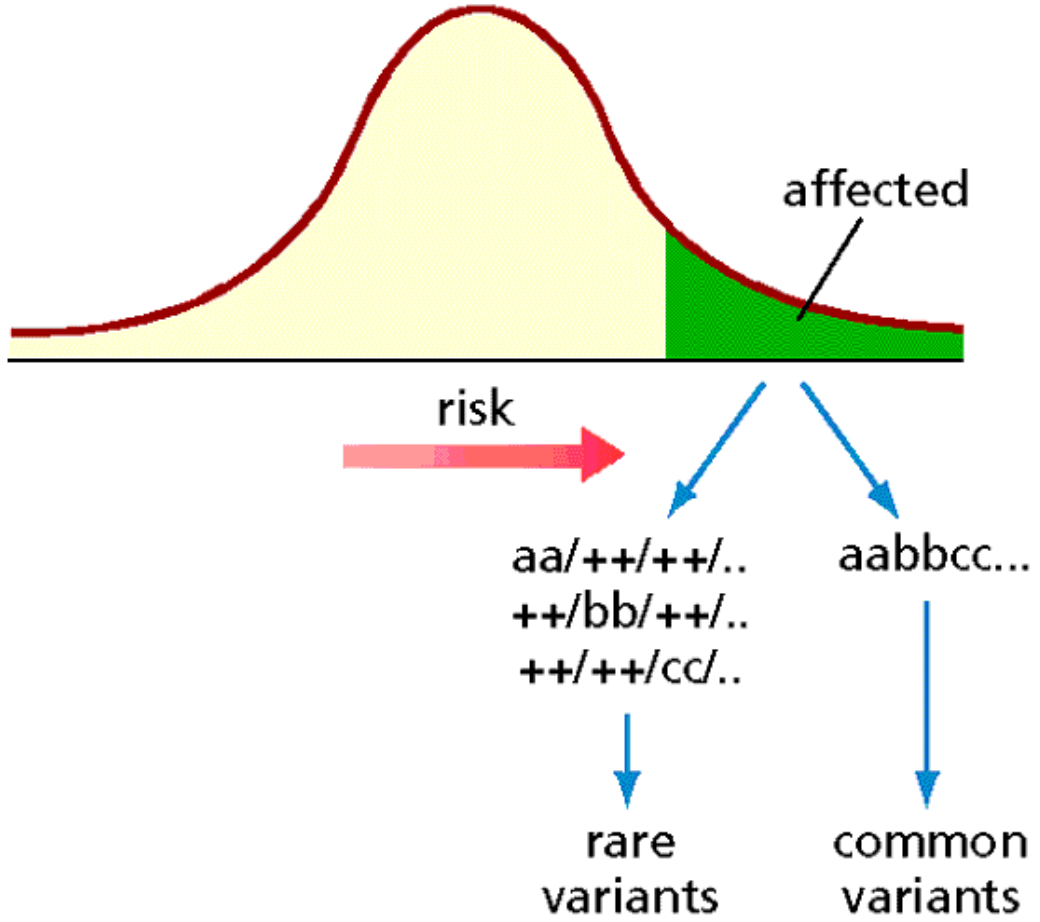
What assumptions were being made?

- The idea that genome wide association studies would work hinged on the idea that risk variants would be common in populations
- And that the risk variants would be represented by haplotype-tagging variants that were mostly randomly selected

Competing hypotheses about complex disease inheritance

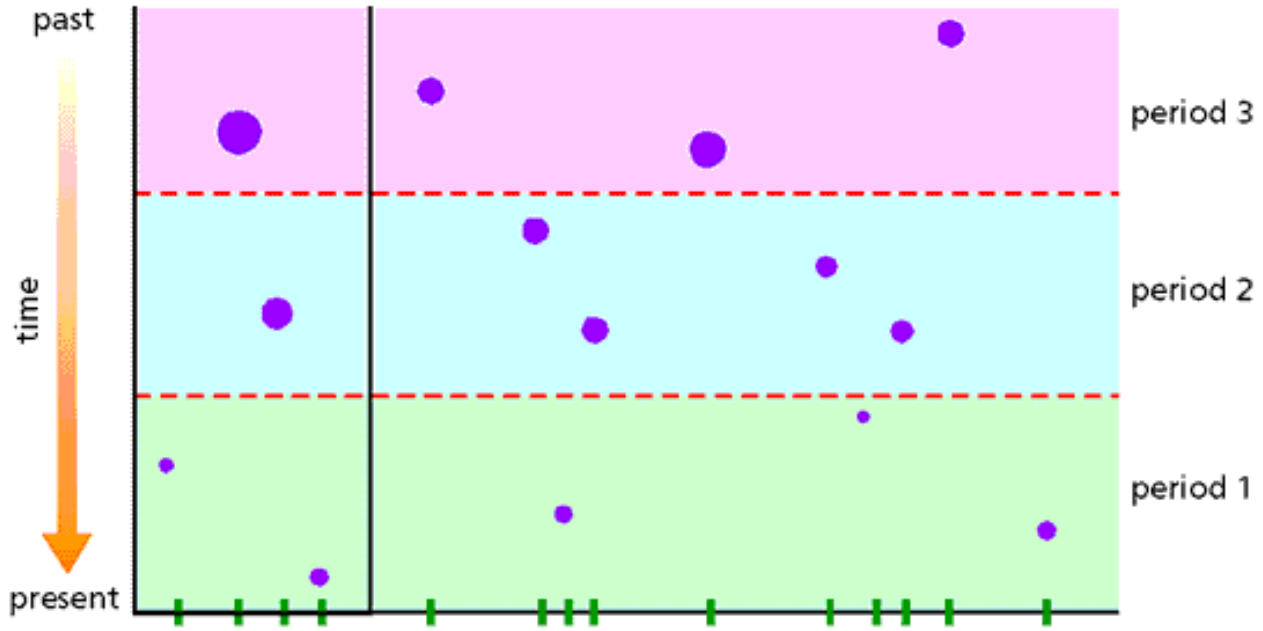
For a fixed disease incidence, individuals who are clinically affected can either have mutations at only one of many possible disease loci (**in which case the mutant alleles are rare** in the population) or harbour mutations at multiple loci simultaneously (**in which case the mutant alleles are common** in the population).

These hypotheses are the extremes of many other possible intermediate scenarios.



Chakravarti, 1999

Within a region, variants arose at different times and therefore have different expected frequencies (under neutrality) and patterns of LD



"We need to answer some central questions regarding the nature of genetic variation of complex diseases. Are they at single or multiple genes? Is the mutational diversity high or low? Are the relevant alleles rare or common? Are they young or old? What is the nature of selection for or against them? Are individuals affected because they harbour too many susceptibility alleles or because they have too few protective alleles? Almost all of the contemplated studies of nucleotide sequence will assist in ferreting out the answers."

Chakravarti, 1999

Competing hypotheses about complex disease inheritance

On the allelic spectrum of human disease

David E. Reich and Eric S. Lander

Human disease genes show enormous variation in their allelic spectra; that is in the number and population frequency of the disease-predisposing alleles at the loci. For some genes, there are a few predominant disease alleles. For others, there is a wide range of disease alleles, each relatively rare. The allelic spectrum is important: disease genes with only a few deleterious alleles can be more readily identified and are more amenable to clinical testing. Here, we weave together strands from the human mutation and population genetics literature to provide a framework for understanding and predicting the allelic spectra of disease genes. The theory does a reasonable job for diseases where the genetic etiology is well understood. It also has bearing on the Common Disease/Common Variants (CD/CV) hypothesis, predicting that at loci where the total frequency of disease alleles is not too small, disease loci will have relatively simple spectra.

Common vs. rare allele hypotheses for complex diseases

Nicholas J Schork, Sarah S Murray, Kelly A Frazer and Eric J Topol

There has been growing debate over the nature of the genetic contribution to individual susceptibility to common complex diseases such as diabetes, osteoporosis, and cancer. The 'Common Disease, Common Variant (CDCV)' hypothesis argues that genetic variations with appreciable frequency in the population at large, but relatively low 'penetrance' (or the probability that a carrier of the relevant variants will express the disease), are the major contributors to genetic susceptibility to common diseases. The 'Common Disease, Rare Variant (CDRV)' hypothesis, on the contrary, argues that multiple rare DNA sequence variations, each with relatively high penetrance, are the major contributors to genetic susceptibility to common diseases. Both hypotheses have their place in current research efforts.

known today as Mendelian genetics as espoused by the 'Mendelian' camp at the time owing to the fact that discrete units of heredity, such as Mendelian-segregating genes, could not, it seemed to them, explain the continuous range of phenotypic variation seen in real populations.

The debate between the Mendelians and Biometricians was resolved, to a high degree, by RA Fisher among others.

Am. J. Hum. Genet. 69:124–137, 2001

Are Rare Variants Responsible for Susceptibility to Complex Diseases?

Jonathan K. Pritchard

Department of Statistics, University of Oxford, Oxford

Little is known about the nature of genetic variation underlying complex diseases in humans. One popular view proposes that mapping efforts should focus on identification of susceptibility mutations that are relatively old and at high frequency. It is generally assumed—at least for modeling purposes—that selection against complex disease mutations is so weak that it can be ignored. In this article, I propose an explicit model for the evolution of complex disease loci, incorporating mutation, random genetic drift, and the possibility of purifying selection against susceptibility mutations. I show that, for the most plausible range of mutation rates, neutral susceptibility alleles are unlikely to be at intermediate frequencies and contribute little to the overall genetic variance for the disease. Instead, it seems likely that the bulk of genetic variance underlying diseases is due to loci where susceptibility mutations are mildly deleterious and where there is a high overall mutation rate to the susceptible class. At such loci, the total frequency of susceptibility mutations may be quite high, but there is likely to be extensive allelic heterogeneity at many of these loci. I discuss some practical implications of these results for gene mapping efforts.

Competing hypotheses about complex disease inheritance

- **Common disease common variant hypothesis**
  - This hypothesis states that the variants that cause common disease are likely to be old and thus at intermediate frequencies and accessible by association mapping
  - It fits with the idea that genetic variants that cause diseases late in life may not be under strong purifying selection
  - GWAS is likely to work if this is true
- **Common disease, rare variant hypothesis**
  - Stabilizing selection should keep most disease-causing alleles at low frequency in the population
  - Thus, common disease variants are likely at low frequency and there may be high allelic heterogeneity
  - If this is true, GWAS may **not** work very well

First genome-wide association mapping study (WTCCC)

- The first set of genome-wide association studies were conducted by the Wellcome Trust Case-Control Consortium
- Traits examined included bipolar disorder, coronary artery disease, Crohn's disease, hypertension, type 1 diabetes, type 2 diabetes and rheumatoid arthritis
- Across all traits, there were 2,000 cases for each disease and 3,000 shared controls (no quantitative trait measures)
- Trend tests were conducted across 500K SNPs for each disease
- The authors attempted to focus on samples from a broadly defined UK/European ancestry populations and removed outliers from genetic clustering (MDS), but there was no population structure control incorporated into the model

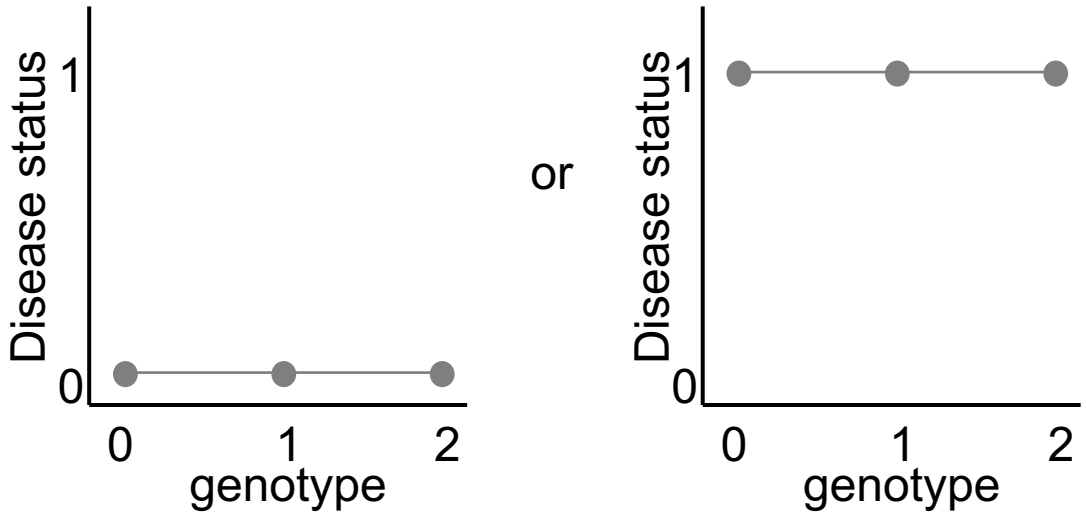
McKay et al., NRG 2009

Testing for signficant associations

The authors used logistic regression to test for associations between each genotyped SNP and trait

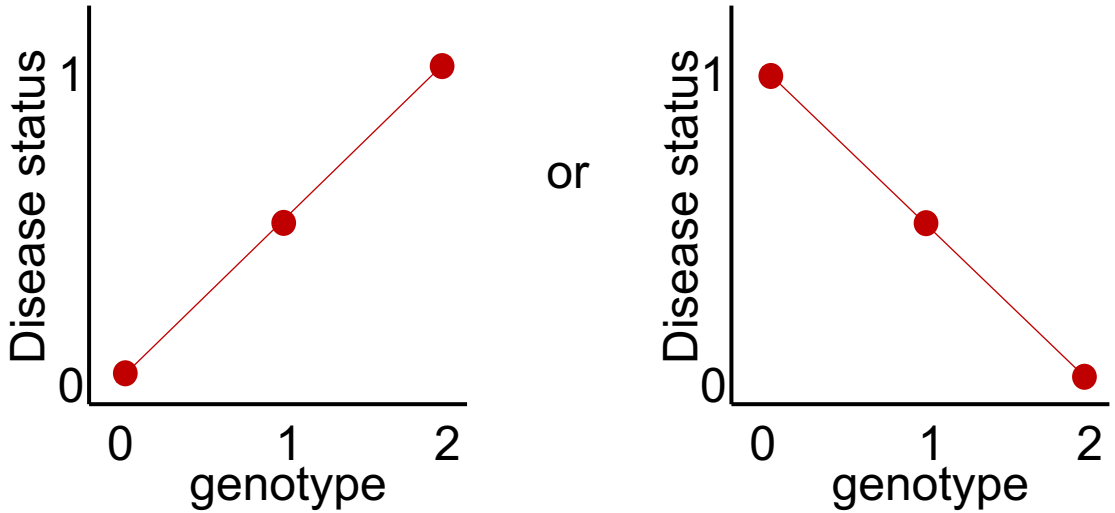
Expected under the null

Slope = 0



Expected under the alternative

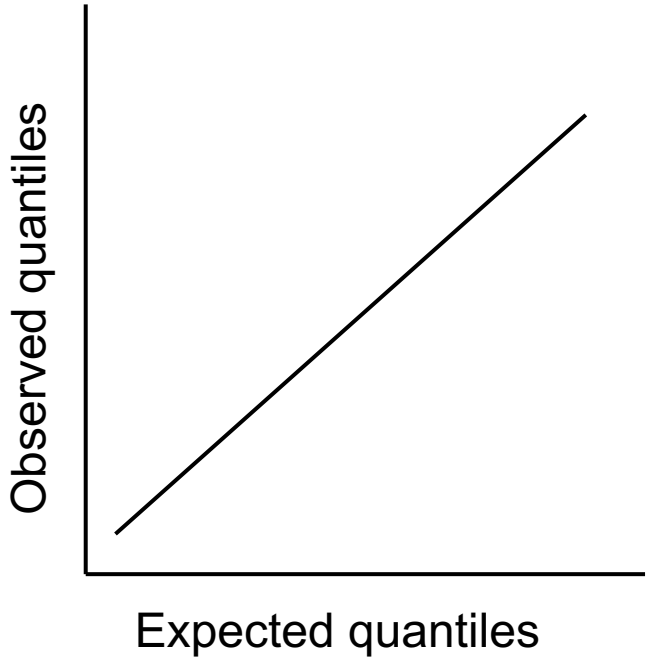
Slope ≠ 0



The test statistic is expected to be χ^2 distributed under the null

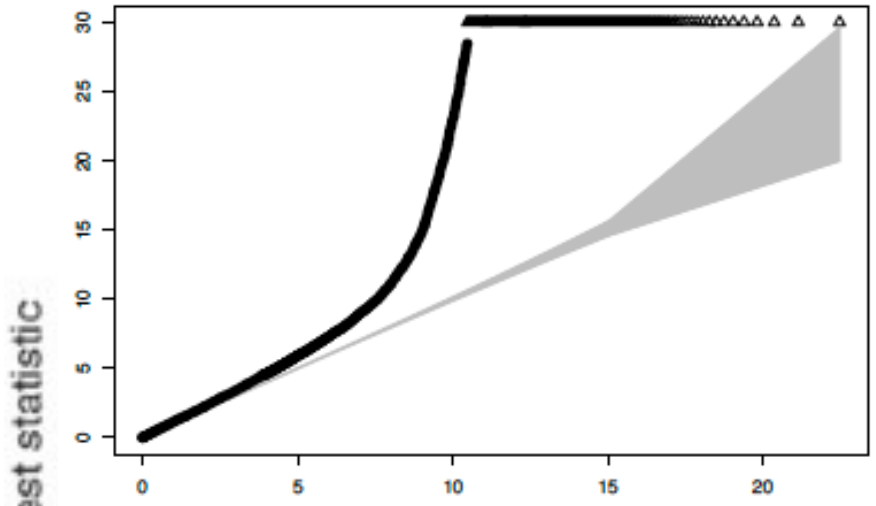
Comparing the genome-wide distribution of the test statistic against the expected distribution

- A quantile-quantile plot is used to compare two distributions
- This approach can be used to assess whether there is an excess of signficant associations relative to expected genome-wide

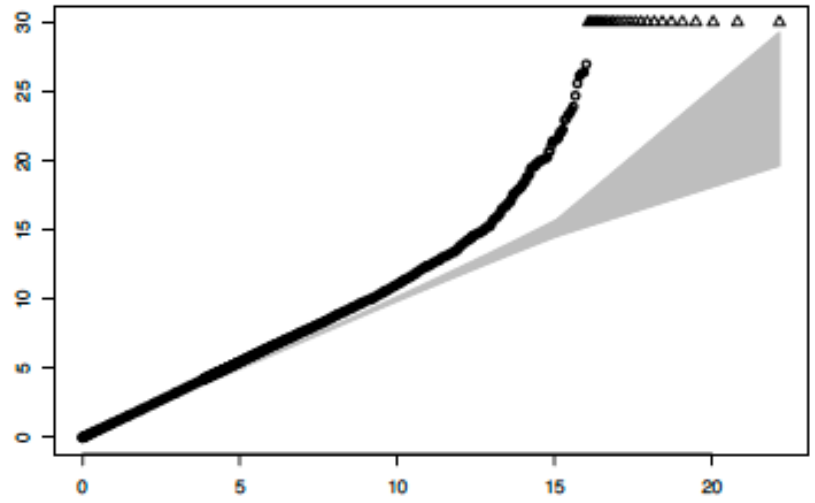


Q-Q plots for T2D at different levels of filtering

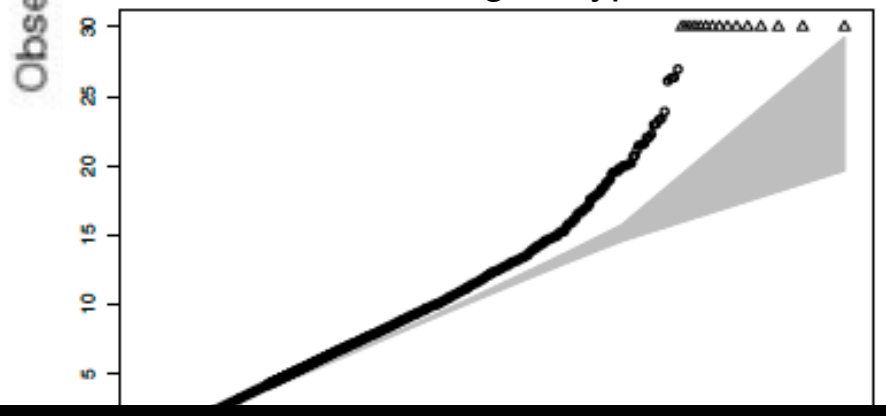
All SNPs



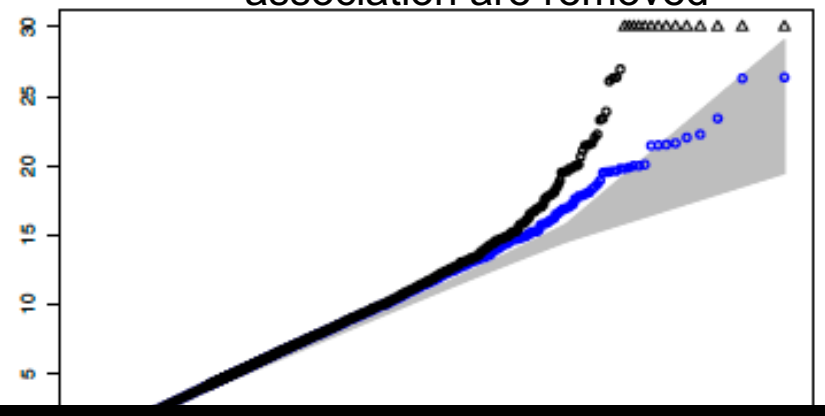
SNPs passing standard filter & MAF>1%



SNPs removed after manual examination of genotype calls



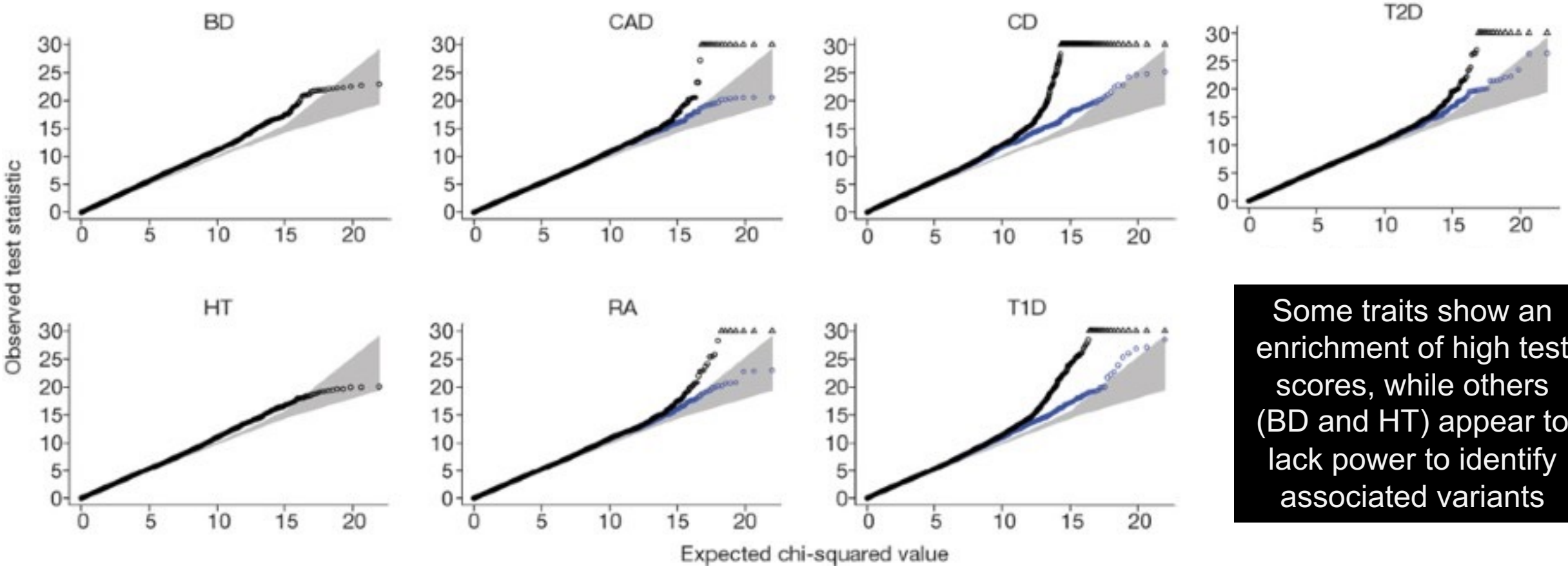
Same as left, showing (blue) the results when regions with strong evidence of association are removed



Greater inflation of high test statistic values in the unfiltered data

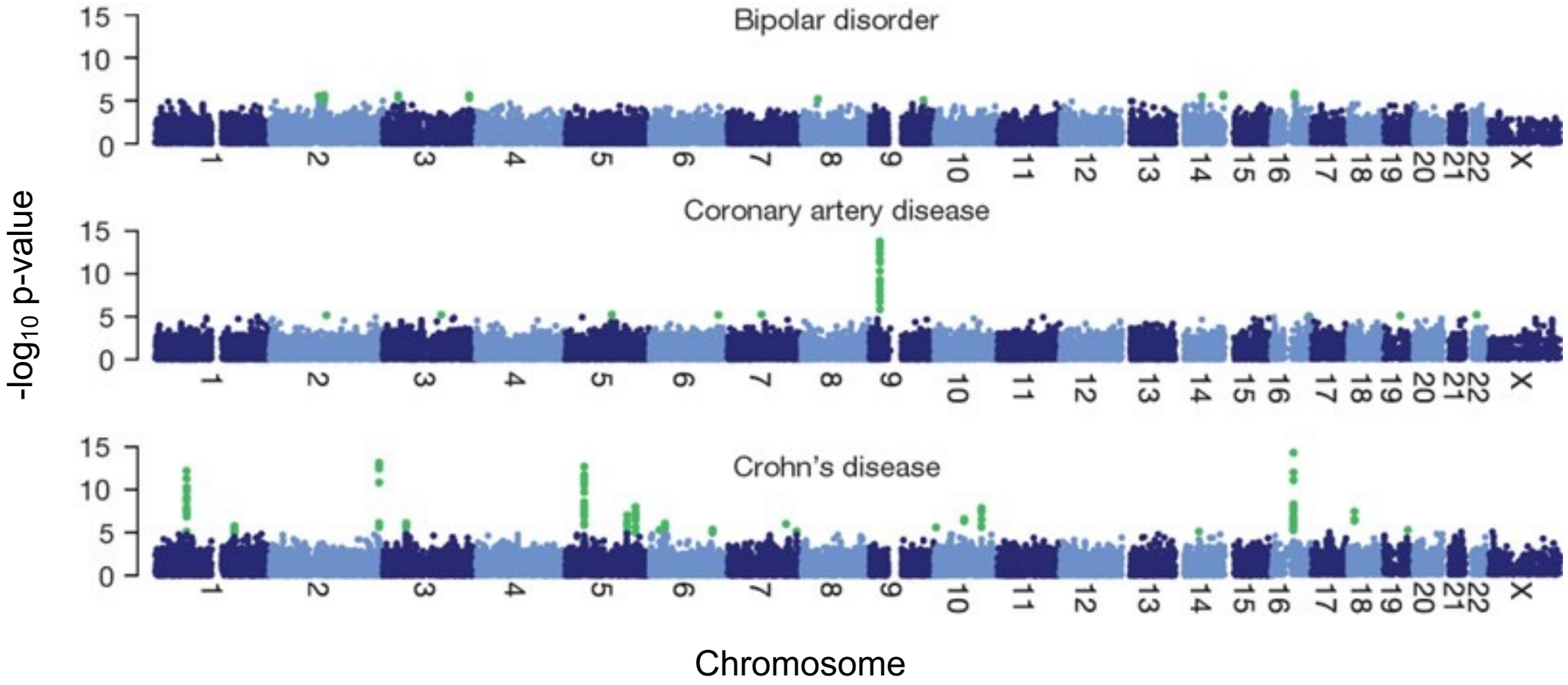
Q-Q plots for genome-wide scans (after filtering)

Since the test statistic is chi-square distributed under the null, they compared the distribution of observed test statistics to the expected under a chi-square distribution

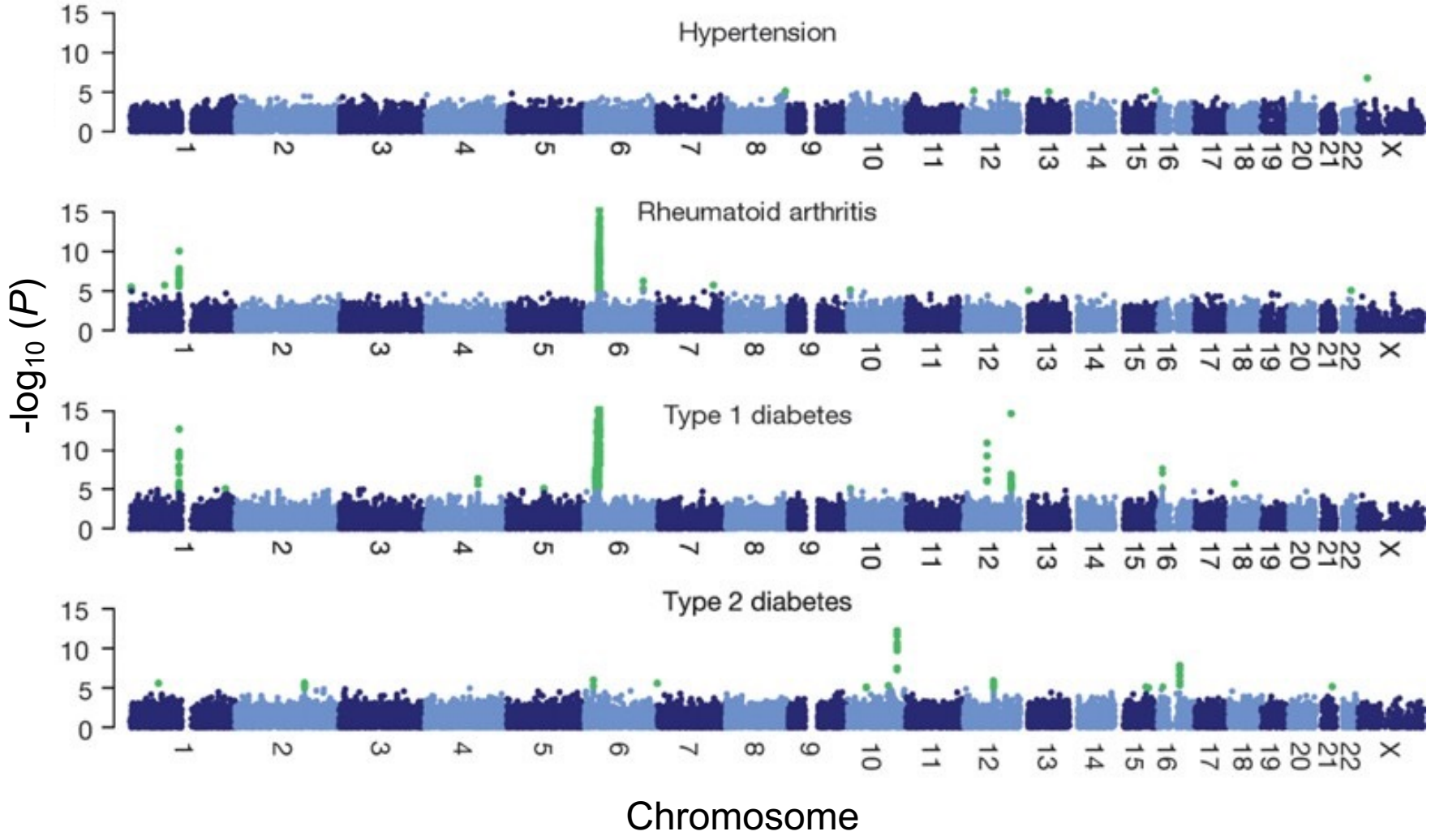


Some traits show an enrichment of high test scores, while others (BD and HT) appear to lack power to identify associated variants

GWAS results for BD, CAD, CD



GWAS for hypertension, RA, T1D, T2D



Positive controls: known loci are detected in the GWAS

Table 2 | Evidence for signal of association at previously robustly replicated loci

Collection	Gene	Chromosome	Reported SNP	WTCCC SNP	HapMap r^2	Trend *P* value	Genotypic *P* value
CAD	*APOE*	19q13	*	rs4420638	-	1.7×10^{-01}	1.7×10^{-01}
CD	*NOD2*	16q12	rs2066844	rs17221417	0.23	9.4×10^{-12}	4.0×10^{-11}
CD	*IL23R*	1p31	rs11209026	rs11805303	0.01	6.5×10^{-13}	5.9×10^{-12}
RA	*HLA-DRB1*	6p21	*	rs615672	-	2.6×10^{-27}	7.5×10^{-27}
RA	*PTPN22*	1p13	rs2476601	rs6679677	0.75	4.9×10^{-26}	5.6×10^{-25}
T1D	*HLA-DRB1*	6p21	*	rs9270986	-	4.0×10^{-116}	2.3×10^{-122}
T1D	*INS*	11p15	rs689	†	-	-	-
T1D	*CTLA4*	2q33	rs3087243	rs3087243	1	2.5×10^{-05}	1.8×10^{-05}
T1D	*PTPN22*	1p13	rs2476601	rs6679677	0.75	1.2×10^{-26}	5.4×10^{-26}
T1D	*IL2RA*	10p15	rs706778	rs2104286	0.25	8.0×10^{-06}	4.3×10^{-05}
T1D	*IFIH1*	2q24	rs1990760	rs3788964	0.26	1.9×10^{-03}	7.6×10^{-03}
T2D	*PPARG*	3p25	rs1801282	rs1801282	1	1.3×10^{-03}	5.4×10^{-03}
T2D	*KCNJ11*	11p15	rs5219	rs5215	0.9	1.3×10^{-03}	5.6×10^{-03}
T2D	*TCF7L2*	10q25	rs7903146	rs4506565	0.92	5.7×10^{-13}	5.1×10^{-12}

Where information on the strength of association at a particular SNP had been previously published and replicated we tabulated the *P* value of both the trend and genotype test at the same SNP (if in our study), or the best tag SNP (defined to be the SNP with highest r^2 with the reported SNP, calculated in the CEU sample of the HapMap project). Positions are in NCBI build-35 coordinates. *Previous reports relate to haplotypes rather than single SNPs. †Not well tagged by SNPs that pass the quality control, see main text.

Q: How do you determine significance with so many tests?

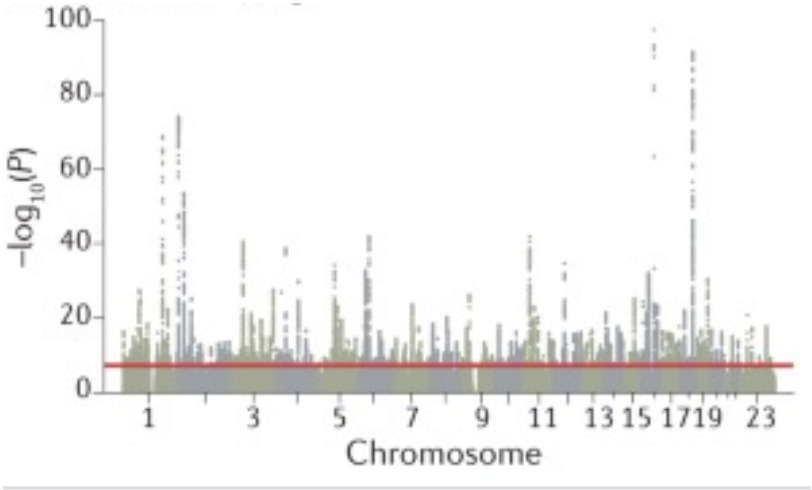
A: Bonferroni correction based on the inferred number of independent tests

Bonferroni correction

- In genome-wide association studies, a very large number of tests are conducted, which leads to a multiple testing problem
- The Bonferroni correction can be used to adjust a significance test to correct for multiple tests
- Using a 5% significance threshold (α = 0.05), we would expect 5% of the markers whose true marker effect is 0 to be significant just by random chance
- This error is called the “type I error rate” , i.e., the probability of rejecting the null when the null is true
- When testing multiple hypotheses, the Bonferroni correction is used to control the type I error rate across hypotheses

Bonferroni correction

- If *m* is the total number of hypotheses tested, the Bonferroni correction rejects the null hypothesis for each $p_i \leq \frac{\alpha}{m}$
- So, if you are testing 10^6 unlinked (independent) markers, the p-value cut-off would be $\frac{0.05}{10^6} = 5 \text{ x } 10^{-8}$ or – $\log_{10}7.3$
- For simplicity, GWAS p-values are plotted on a $-\log_{10}$ scale, as in this example:



Not enough data?

Make some up!

Genotype imputation

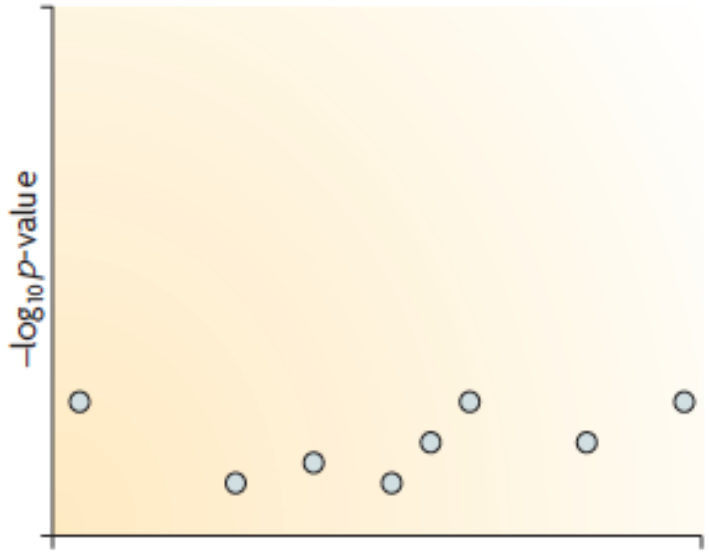
Genotype imputation is can be used to infer missing data and boost power in GWAS

In a sample of unrelated individuals, some genotype data may be missing due to technical issues

1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

(and missing data is an even bigger problem in whole genome sequencing data!)

Power to detect associations may be low due to missing data



…in other cases, a researcher may want to combine data sets to conduct a meta-analysis, but different SNP sets might be genotyped in different data sets

https://www.nature.com/articles/s43586-021-00056-9

A reference set of samples can be used to impute based on haplotype similarity

Genomes sequenced in a reference panel, e.g., HapMap

Haplotypes are matched to the reference set based on the non-missing data in the sample

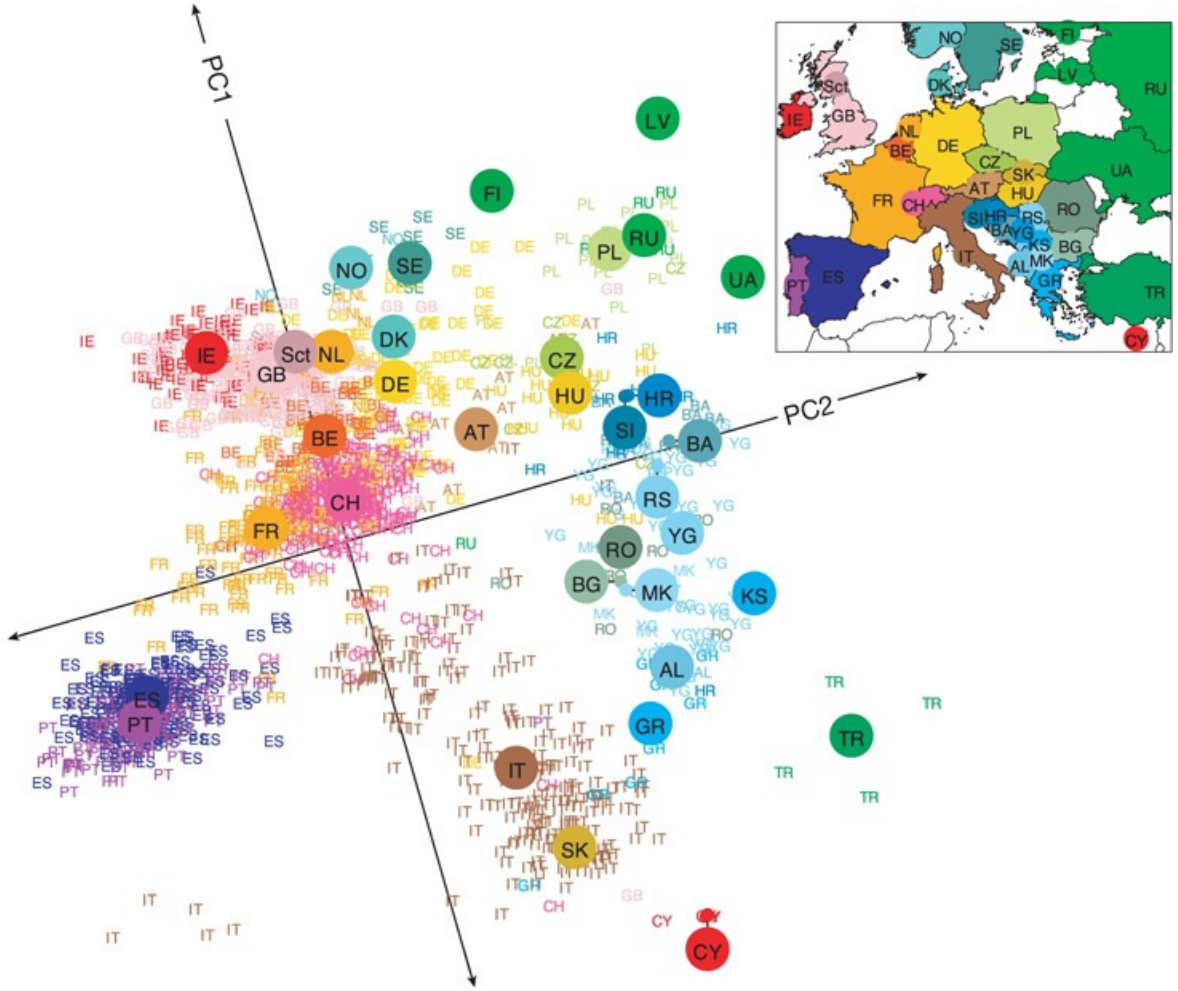
And missing data are imputed

https://www.nature.com/articles/s43586-021-00056-9

Considerations

- Haplotype imputation works best when samples of the population which imputation is needed are drawn from the same population as the reference sequenced reference
- Diverged samples cannot be imputed with high accuracy

What about population structure?

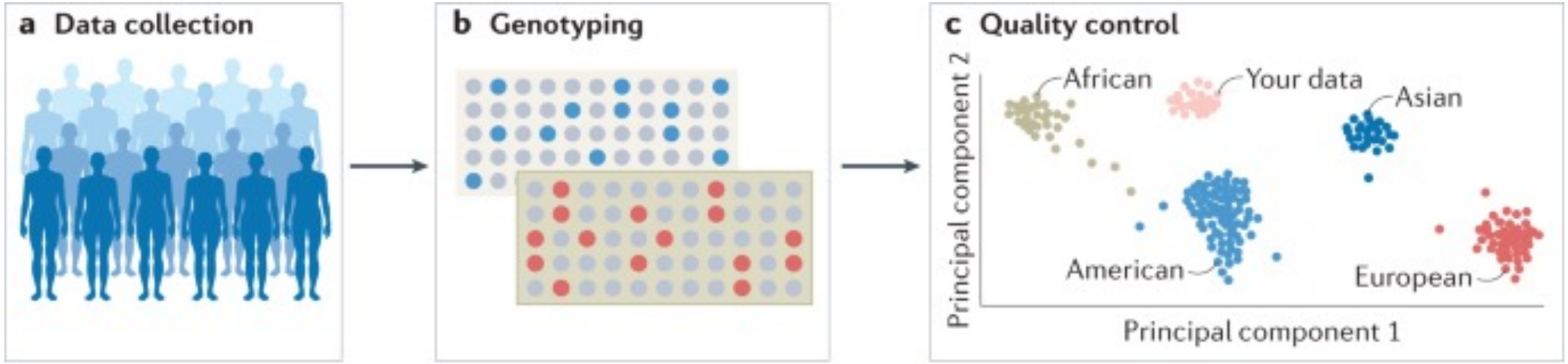


*Novembre et al. Nature **000**, 1-4 (2008) doi:10.1038/nature07331*

Population structure can confound association analysis

- Relatedness among individuals that is not accounted for can result in false positives or loss of power
  - False positives may result from correlation between structure and the trait
  - False negatives can result if the effects of structure are strong relative to the effects of true variants
- Including population structure in the model to detect genetic effects on phenotype can help solve these problems

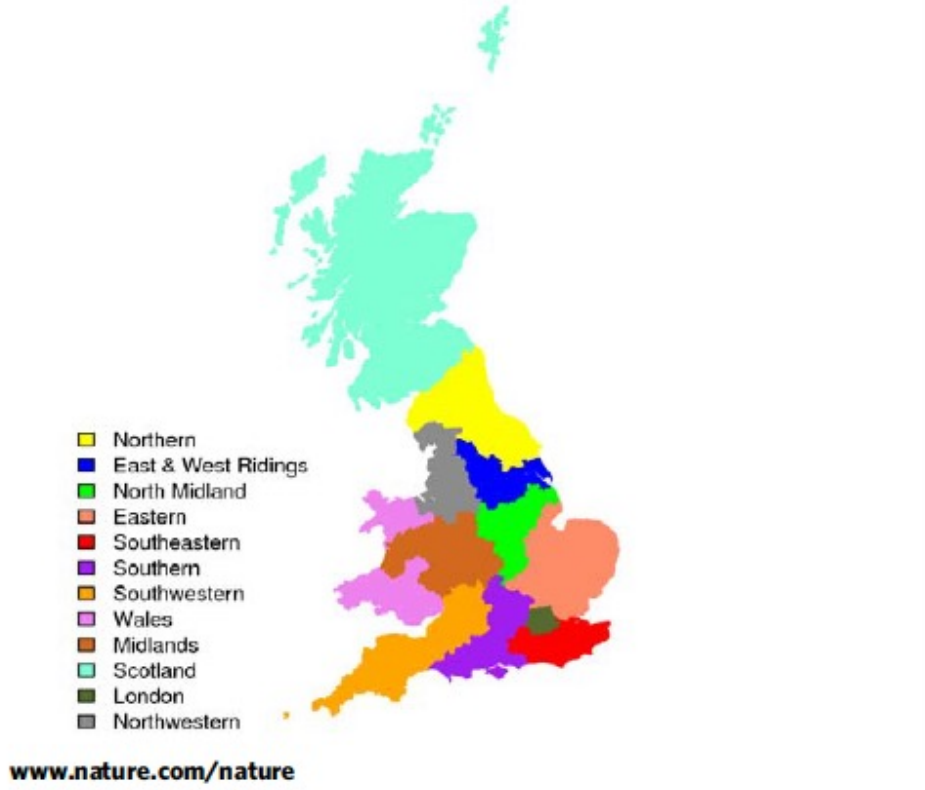
PCA can be helpful for finding mistakes or individuals who do not cluster as expected



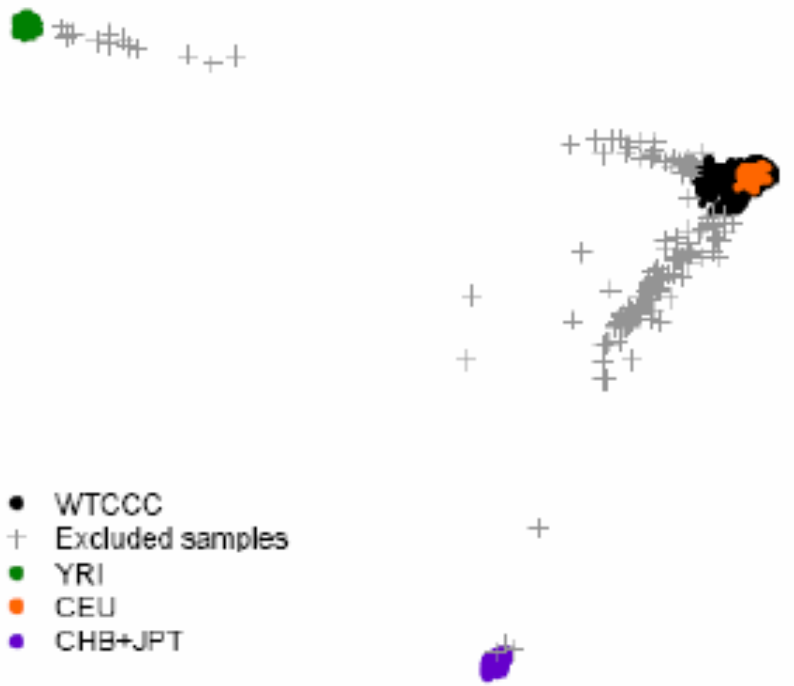
https://www.nature.com/articles/s43586-021-00056-9

Population structure in the WTCCC

Regions of origin in the WTCCC set



Population structure based on MDS scaling



Approaches to deal with population structure

- Incorporate it into the model to estimate the effect of the genetic variant in the presence of population structure
- A covariance matrix of individuals, derived from the matrix of individuals x variants is often used to control for population structure in the model
- This is either done by including the covariance matrix directly in the model or by including eigen vectors derived from the covariance matrix (i.e., principal components) in the model

Linear model for genotype association analysis

Assuming the phenotype is quantitative and genetic basis additive, we can model it as

$$y_i = \sum_{j=1}^{m} \beta_0 + x_i \beta_j + \varepsilon_i$$

where y_i is the phenotype of the i^{th} individual, β_0 is the mean, $x_{i,j}$ is the genotype of the i^{th} individual at the j^{th} variant, m is the number of variants, β_j is the effect size of the j^{th} variant, and ε_i is the error or noise term for the i^{th} individual.

The noise terms are assumed to be independent with a Gaussian (i.e., normal) distribution.

The genotypes are assumed to be fixed (not random) variables

Linear Model for genotype association analysis

This model is run consecutively on individual SNPs, so in practice, for each SNP we assess the evidence for its marginal effect:

$$y_i = \beta_0 + x_i\beta + \varepsilon_i$$

where y_i is the phenotype of the i^{th} individual, β_0 is the mean, x_i is the genotype of the i^{th} individual, β is the effect size variant, and ε_i is the error or noise term for the i^{th} individual.

The noise terms are assumed to be independent with a Gaussian (i.e., normal) distribution.

The genotypes are assumed to be fixed (not random) variables

Linear Mixed Model for single SNP analysis

Estimate the effect of each allele on the phenotype, while controlling for population structure:

$$\mathbf{y} = \beta_0 + x_i\beta + \mathbf{Zu} + \varepsilon_i$$

where *Zu* is the random term that accounts for the covariance structure among individuals. Z is an *n* x *m* matrix individuals x variants, u is an *m* x 1 vector of random effects, and ε is an n x 1 vector of errors.

Calculating *u* and ε are computationally expensive steps due to the need to invert the matrix of residual error variance

Some work-arounds have been developed to improve computational speed, e.g., GEMMA

Using principal components to account for population structure

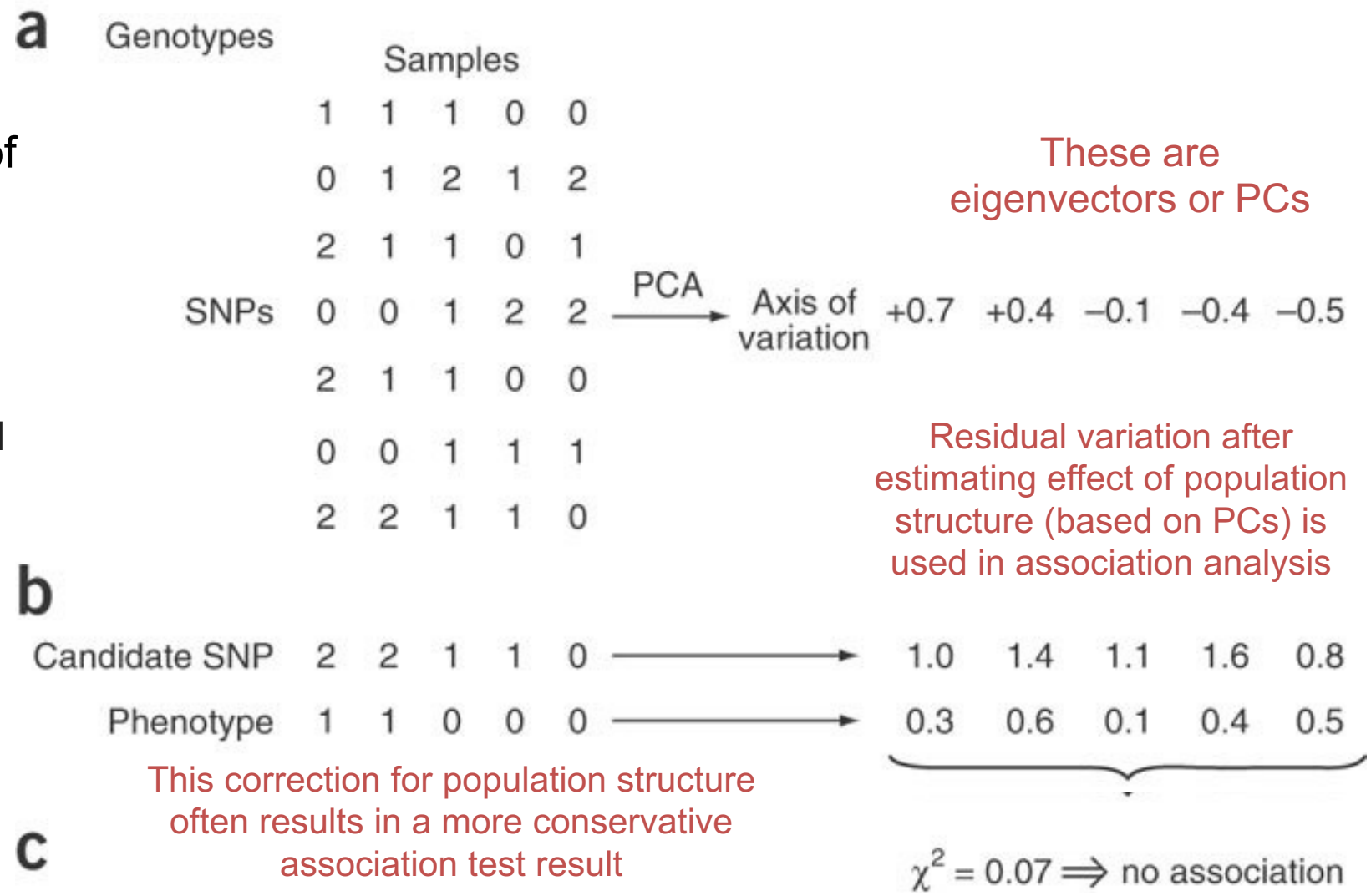
- Start with the matrix of individuals by variants
- Identify vectors (eigenvectors) that maximize the variance explained from the total matrix
- Some number of principal components (eigenvectors) can then be included in the linear model to represent population structure
- Choosing the number of eigenvectors (principal components) to include in the model is not always straight-forward, but can be determined based on the amount of the total variance explained

PCA to control for population structure (Eigenstrat algorithm)

Step 1:
PCA is applied to genotype data to infer continuous axes of genetic variation (a single axis is shown here)

Step 2:
Genotype at a candidate SNP and phenotype are adjusted by amounts attributable to ancestry, removing correlation to ancestry

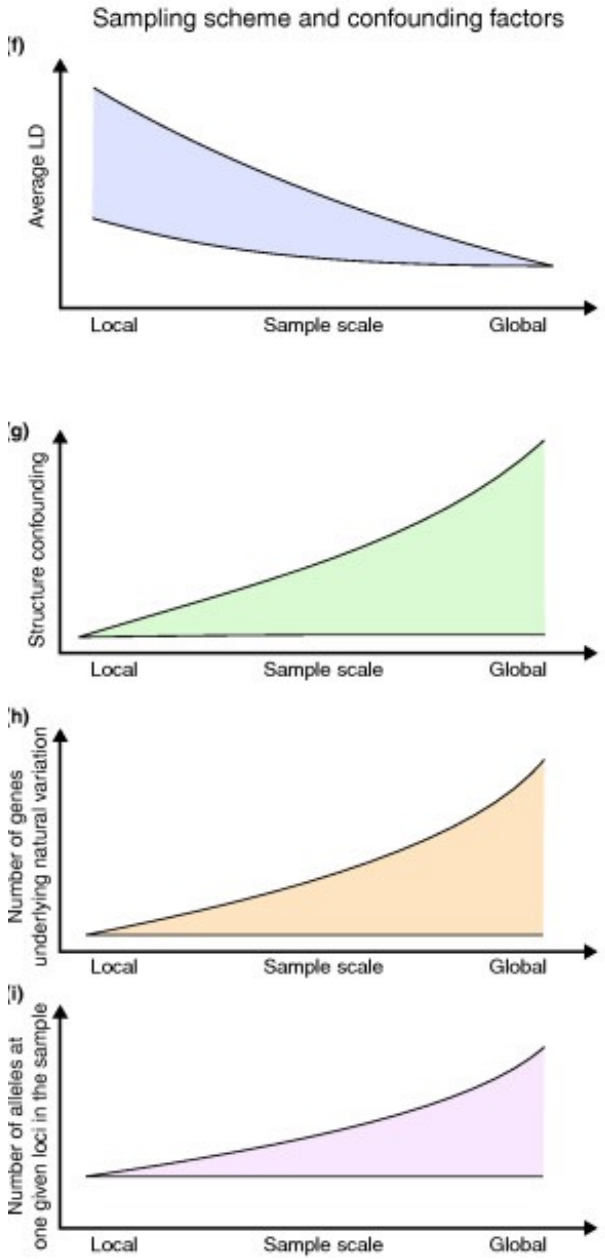
Step 3:
A corrected association test statistic results



CHALLENGE: What is the appropriate scale for defining a population

- Tradeoff between inclusive and specific definition
  - Benefits of inclusive design:
    - larger Ne: more genetic variation, potentially more phenotypic variation, less LD
  - Benefits of specific definition:
    - less genetic heterogeneity
    - less allelic heterogeneity

Local versus global samples for GWAS



With global population generally have more time (in the past) for recombination, so LD is lower

Population structure may be less complex in a local population

Fewer genes underlying trait: less genetic heterogeneity in a local population

Fewer alleles underlying the trait: less allelic heterogeneity in a local population

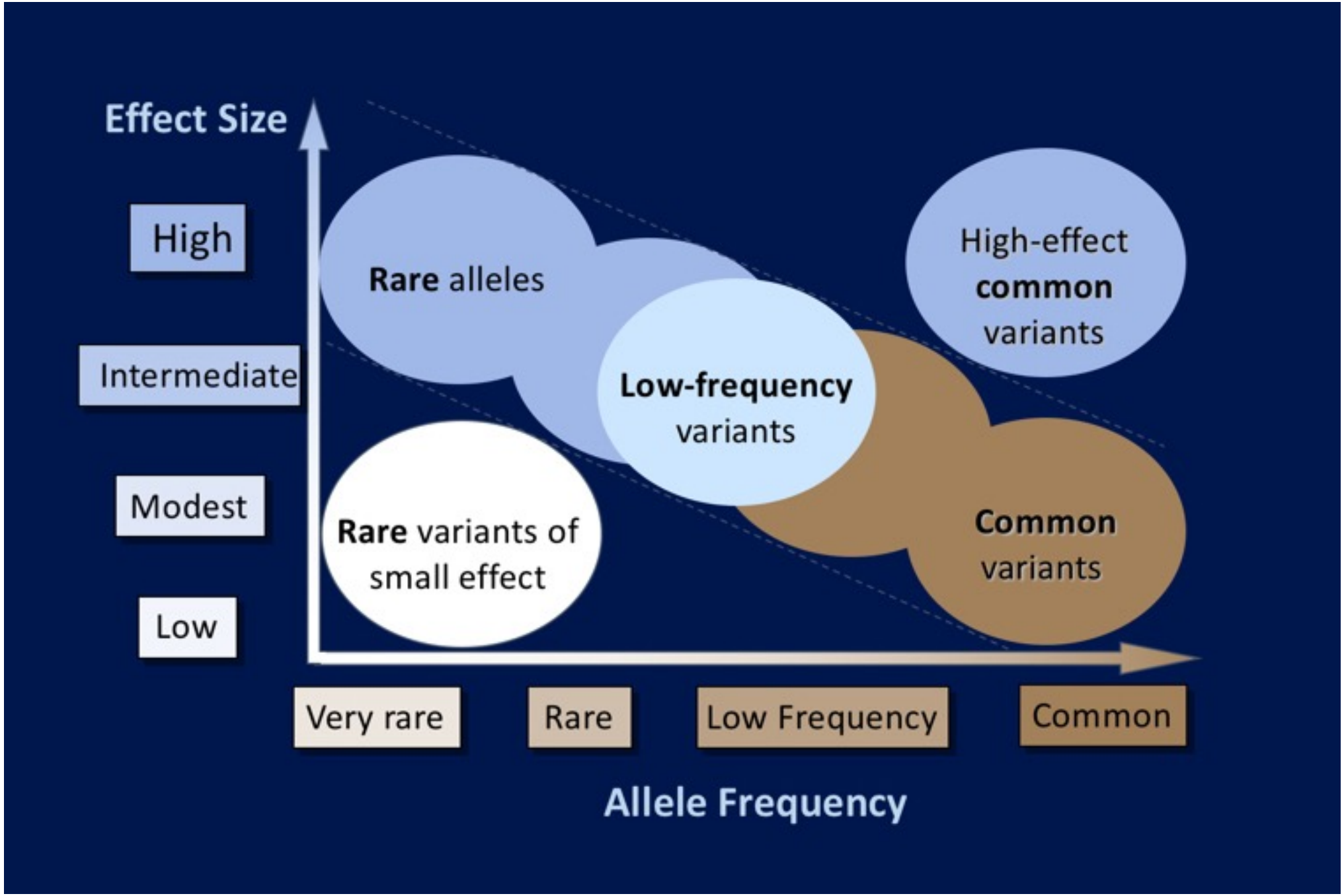
Missing heritability: when and why does GWAS fail?

GWAS findings fail to account for all heritability

<u>Some potential causes:</u>

- Dominance effects
- Low frequncy variants responsible for trait variation
- Allelic heterogeneity
- Untagged structural variants responsible for trait variation
- Uncontrolled environmental confounders
- Condition-dependent effects (GxG, GxE)

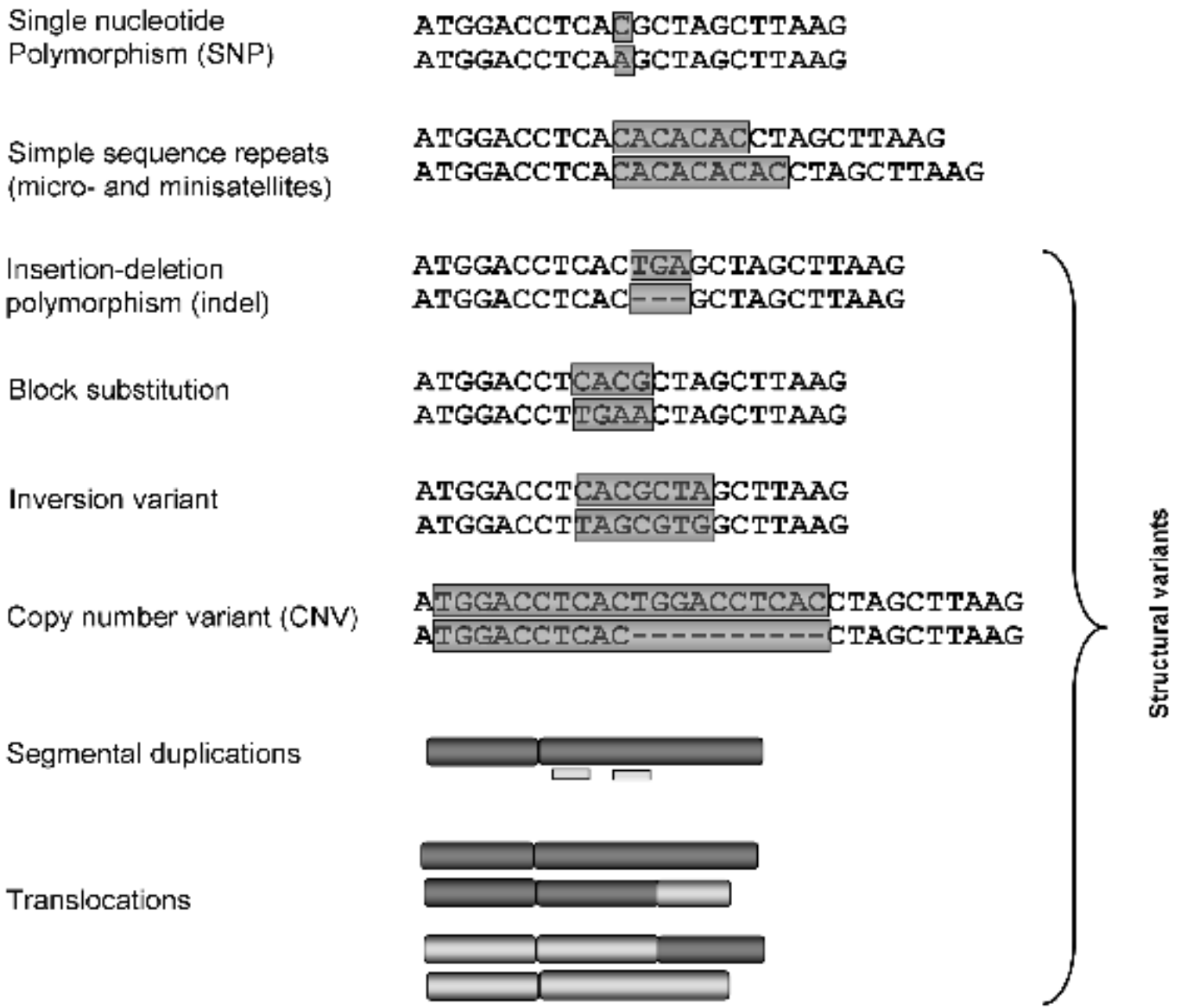
https://www.nature.com/articles/456018a



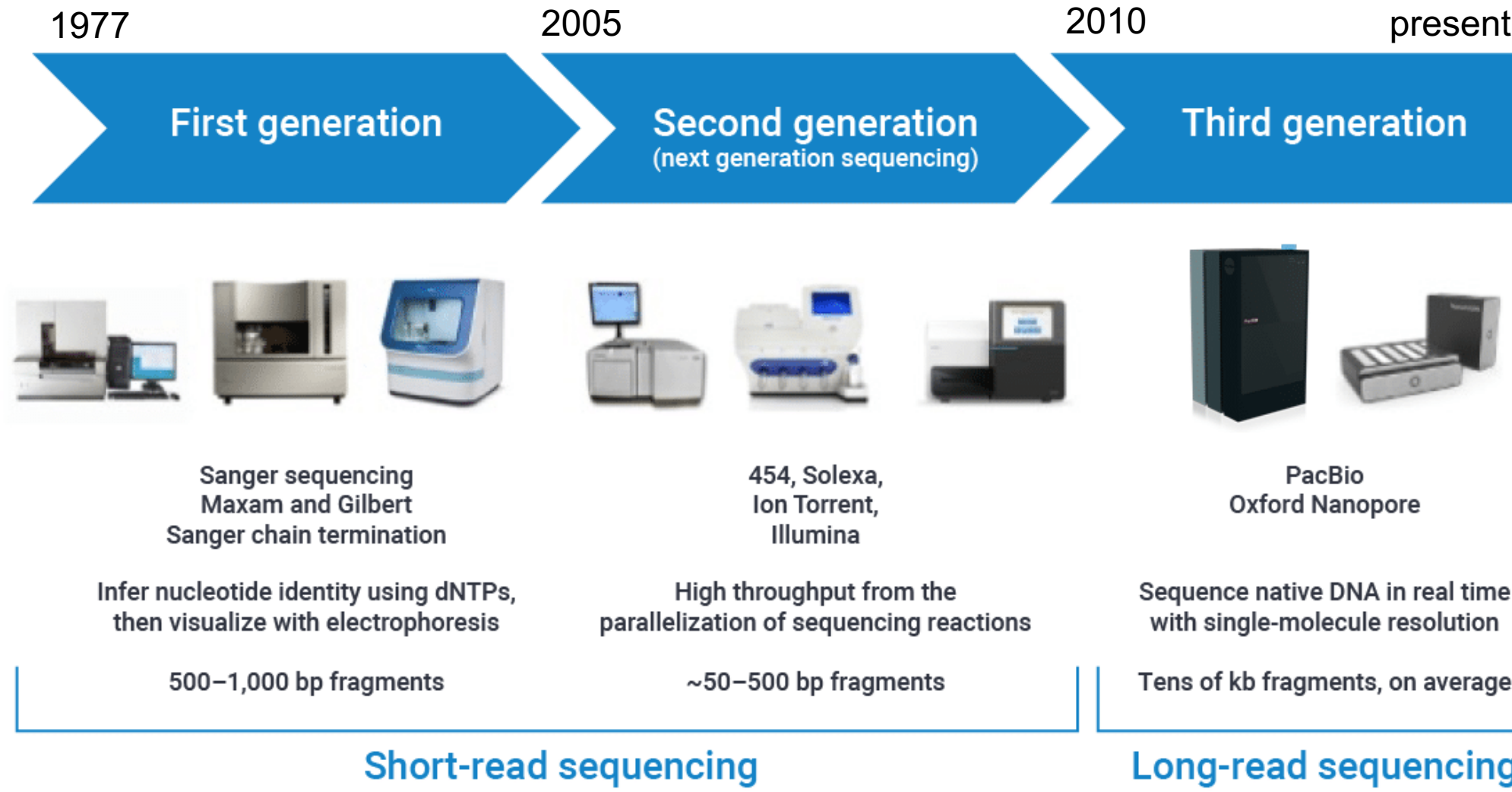
https://www.nature.com/articles/nature08494

SNPs only represent one type of variant

- Many potential types of variants are structural variants
- Many of these are difficult to assay accurately in short-read sequencing data
- Since structural variants affect larger genomic regions, they may have relatively high contributions to trait variation



Sequencing technology is improving, allowing us to assess variation more completely



Source: https://www.pacb.com/blog/the-evolution-of-dna-sequencing-

Potential solutions to the missing heritability problem

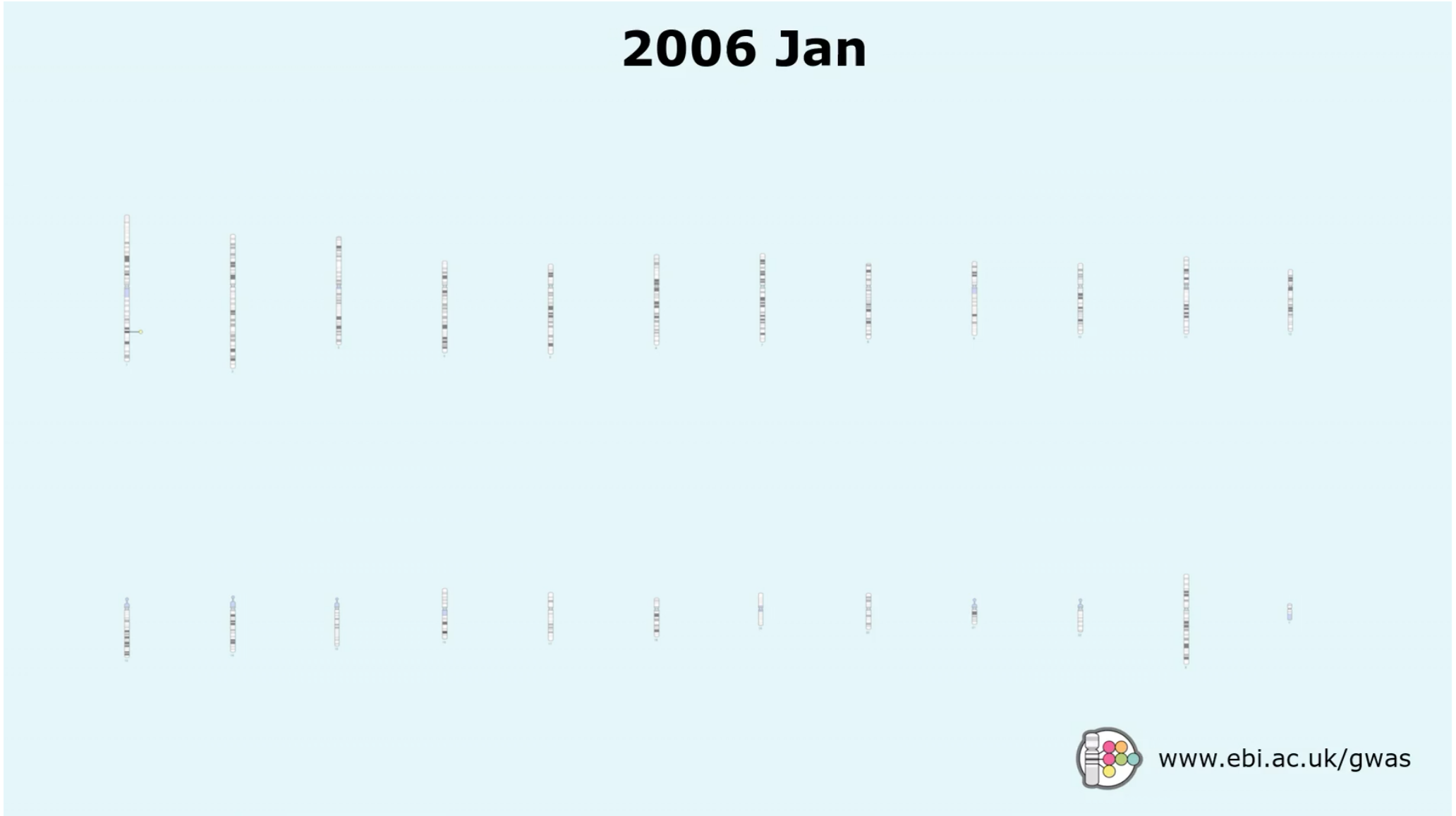
- Larger sample sizes to improve power for low frequency variants
- Burden tests to combine signals in cases of allelic heterogeneity
- Include structural variants in the analysis
- Collect more thorough information about study subjects during DNA sampling

WTCCC was the starting point

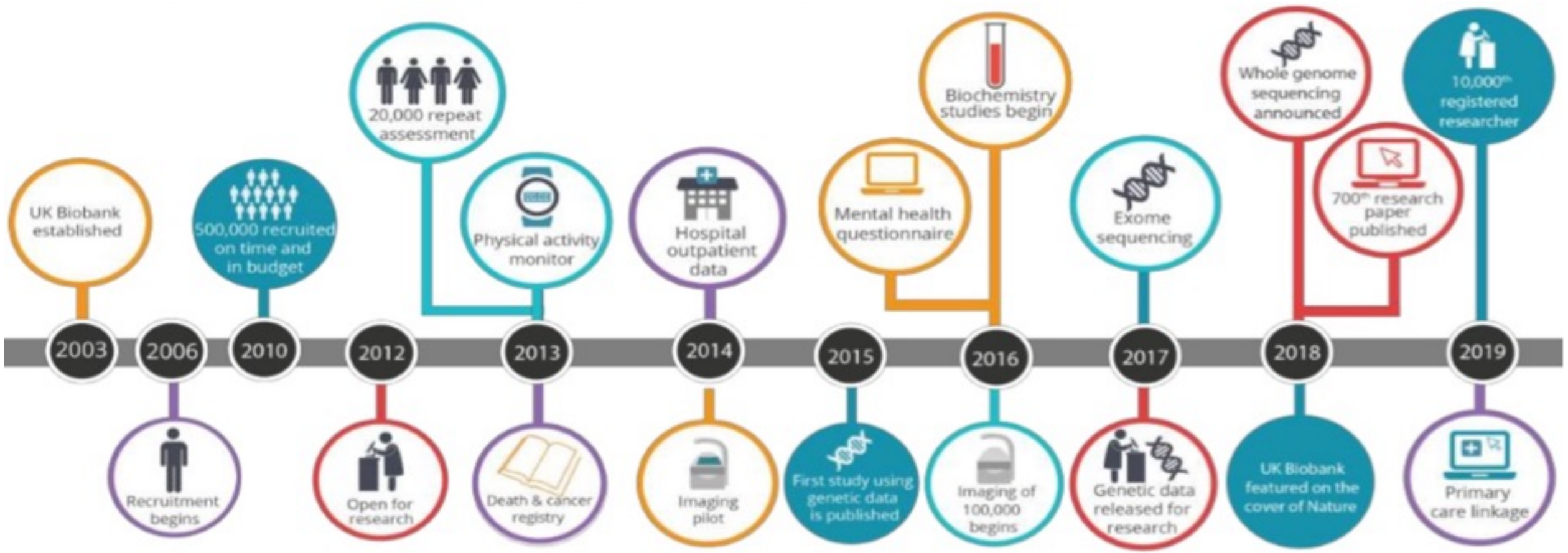
Where are we now?

https://www.ebi.ac.uk/training/online/courses/gwas-catalogue-exploring-snp-trait-associations/what-is-gwas-catalog/

GWAS through time

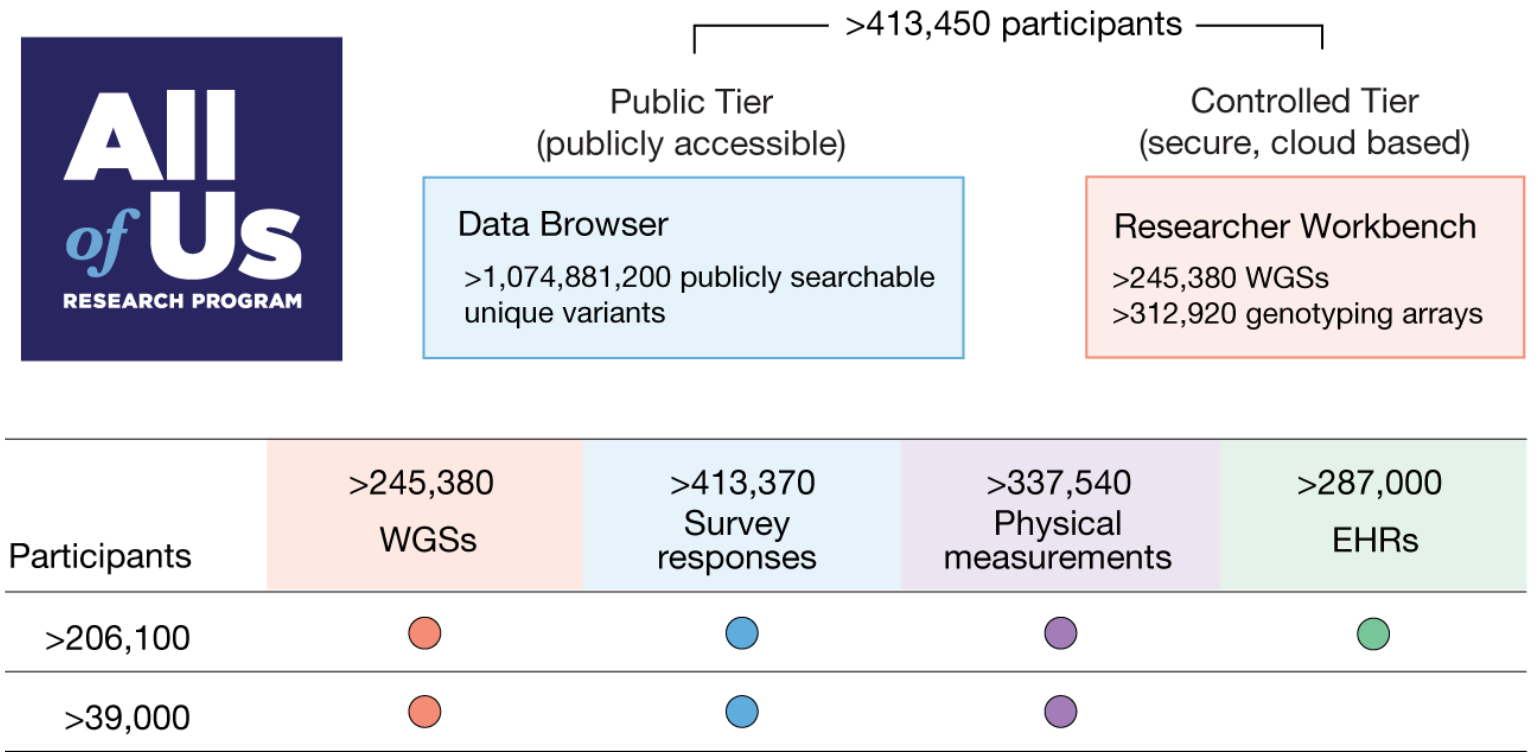


Complete sequences: UK Biobank



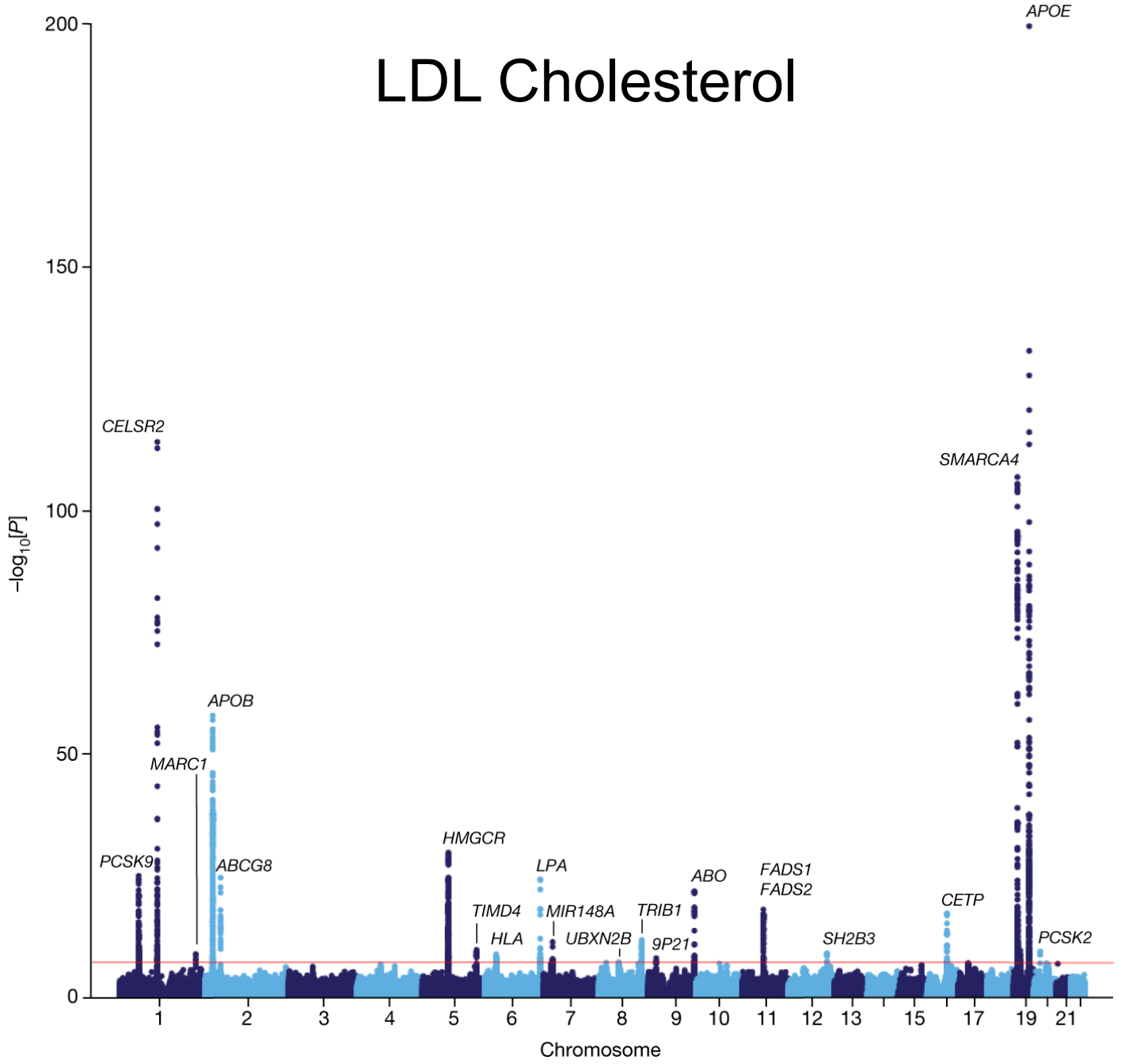
https://labs.icahn.mssm.edu/minervalab/resources/data-ark/uk-biobank/

US-based project *All of Us* aims to sequence 1M+



A major goal of the All of Us project is to increase diversity in GWAS since most of what we know about trait-variant associations is derived from European populations

Mapping in the *All of Us* panel

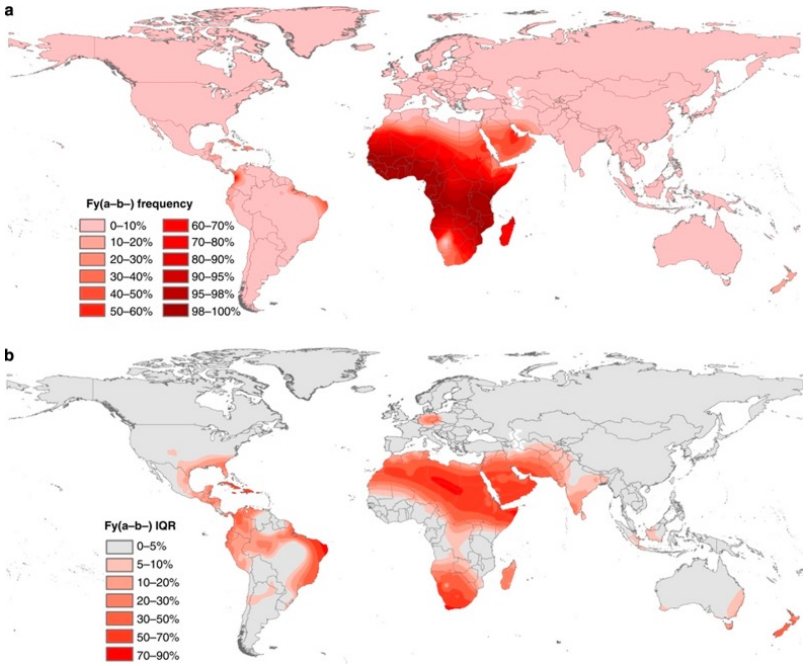


https://www.nature.com/articles/s41586-023-06957-x

Phenome-wide association

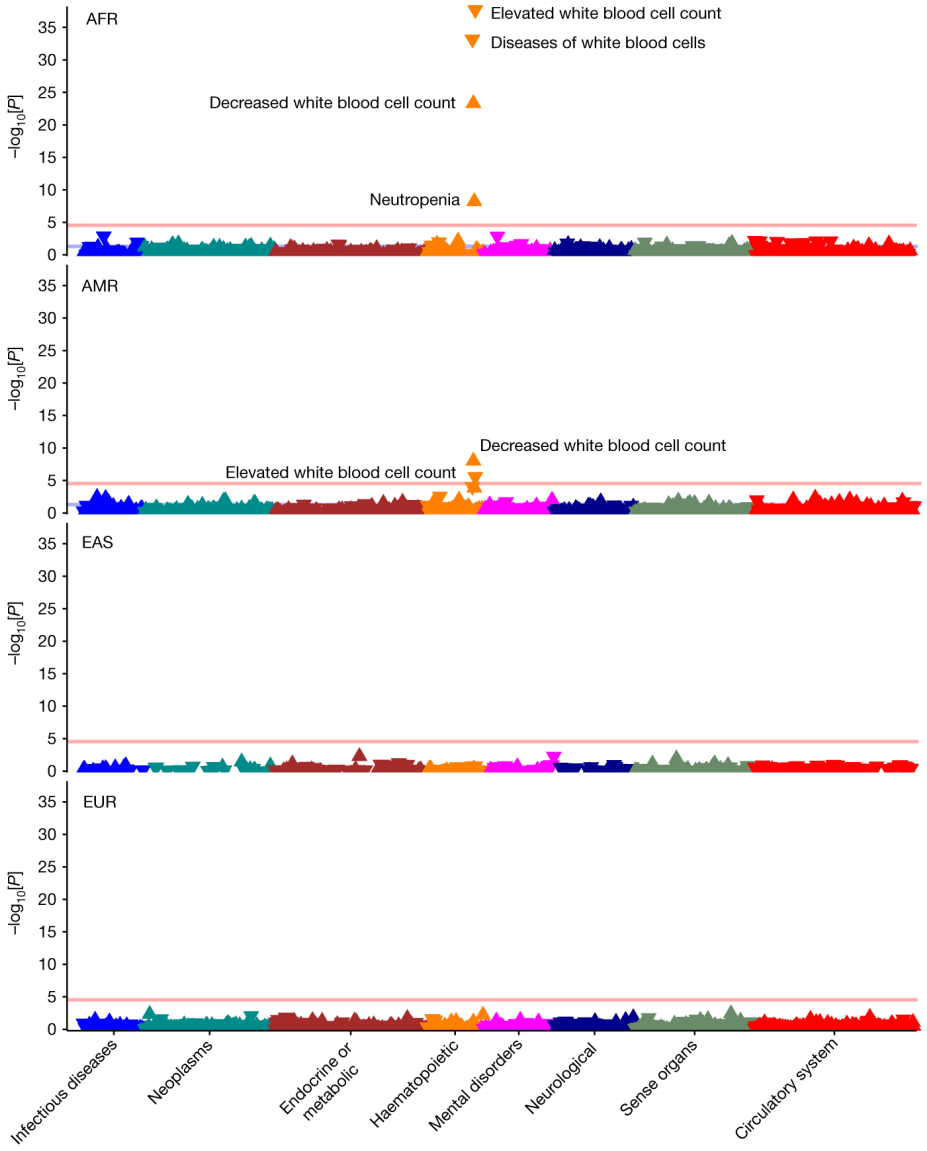
starts with a genetic variant and conduct association analysis across phenotypes

Duffy-negative alleles are at high frequency in African populations and confer resistance to vivax malaria (*Plasmodium vivax*)



Howes et al., Nature Communications, 2011

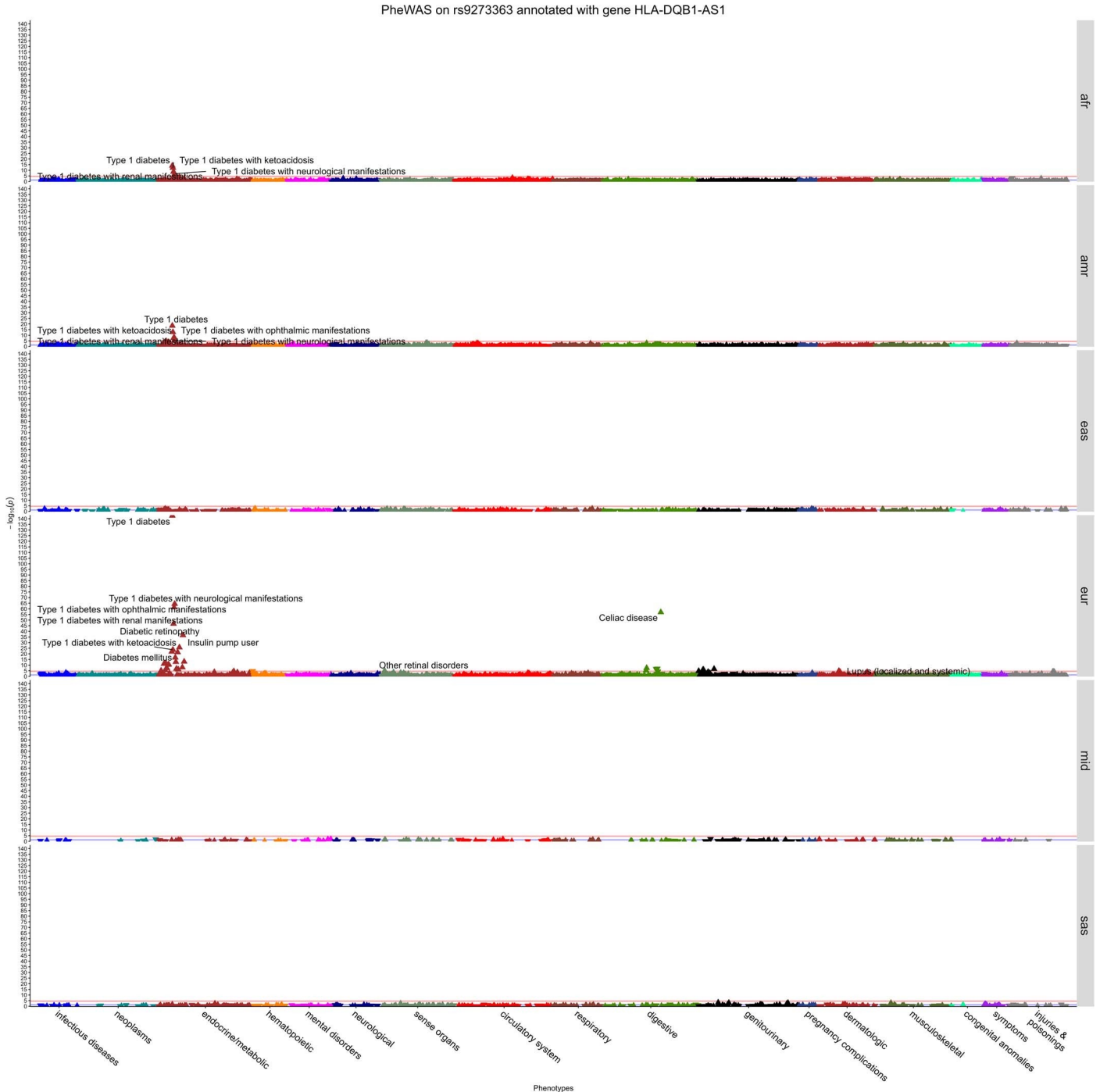
Phenome-wide association of Duffy blood group (*ACKR1*) identifies variation in individuals with African ancestry



Phenotype

HLA-DQB1 (rs9273363)

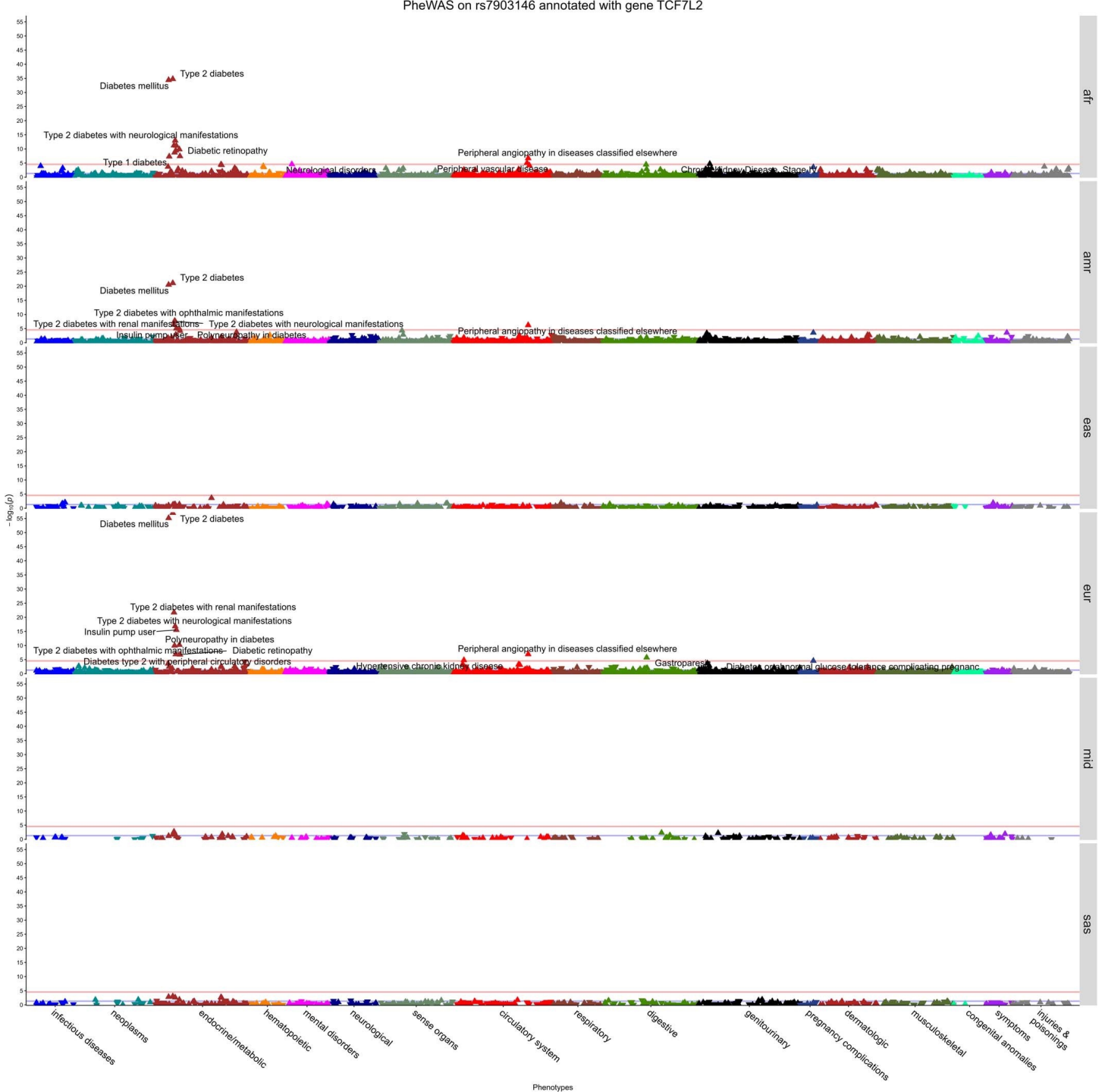
AFR: African ancestry
AMR: Latinx/admixed ancestry
EAS: East Asian ancestry
EUR: European ancestry
MID: Middle Eastern ancestry
SAS: South Asian ancestry



https://www.nature.com/articles/s41586-023-06957-x/figures/9

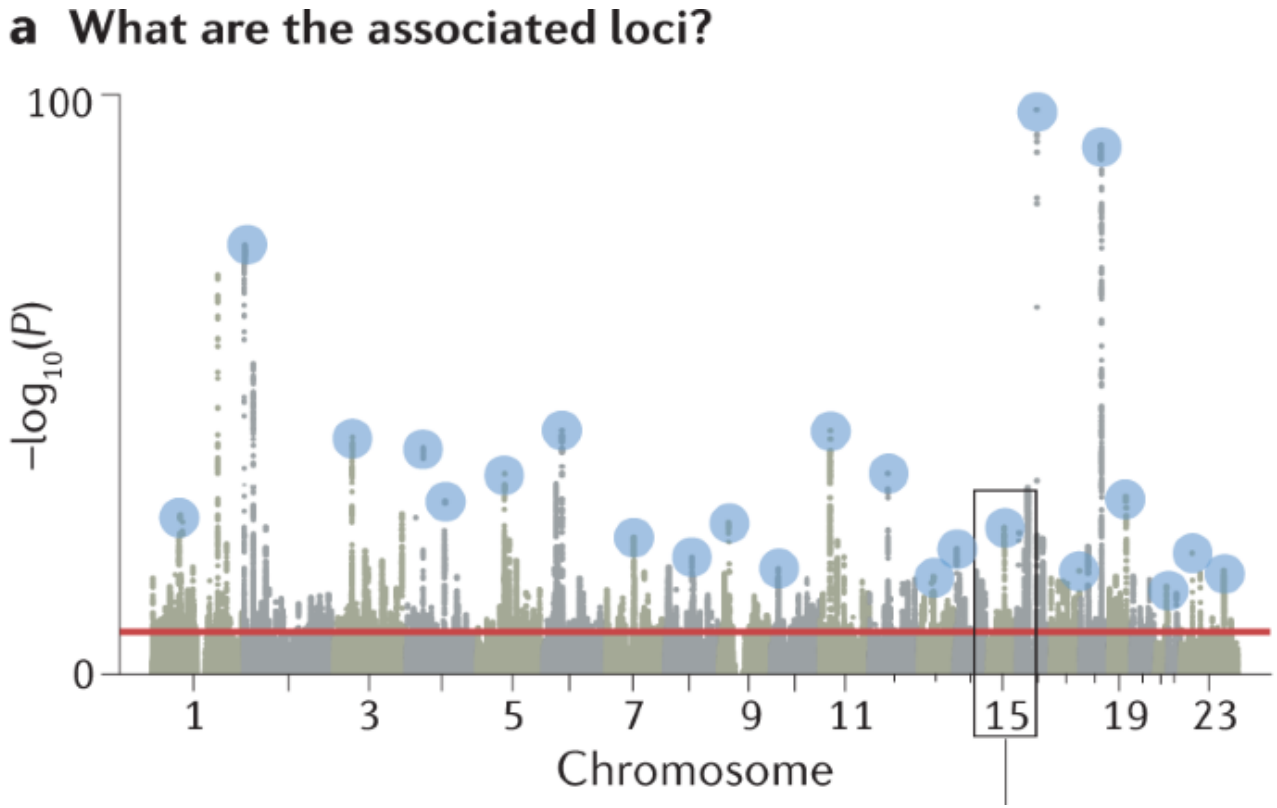
TCF7L2(rs7903146)

AFR: African ancestry
AMR: Latinx/admixed ancestry
EAS: East Asian ancestry
EUR: European ancestry
MID: Middle Eastern ancestry
SAS: South Asian ancestry



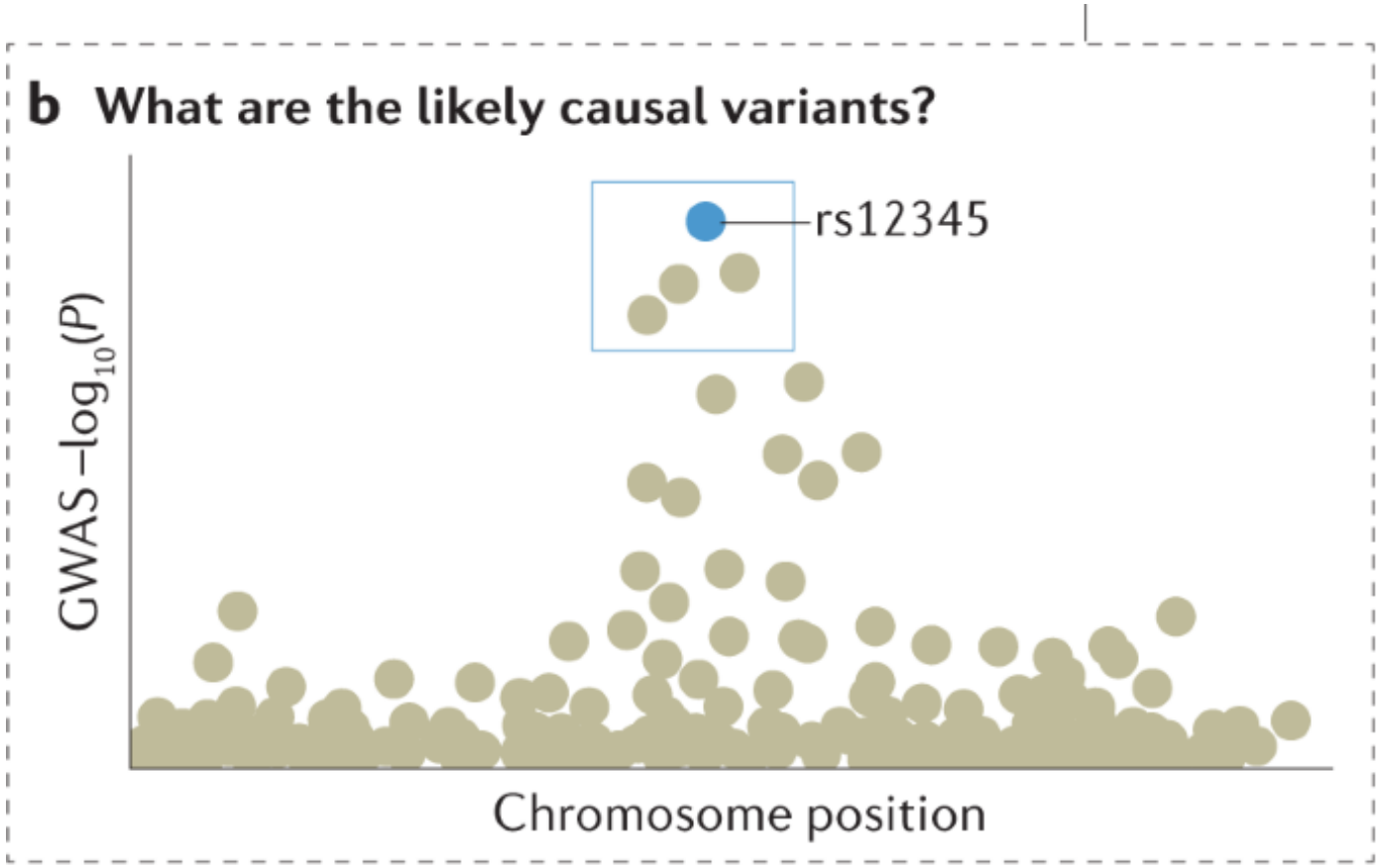
https://www.nature.com/articles/s41586-023-06957-x/figures/10

Gleaning biological understanding from GWAS (combined with other methods)



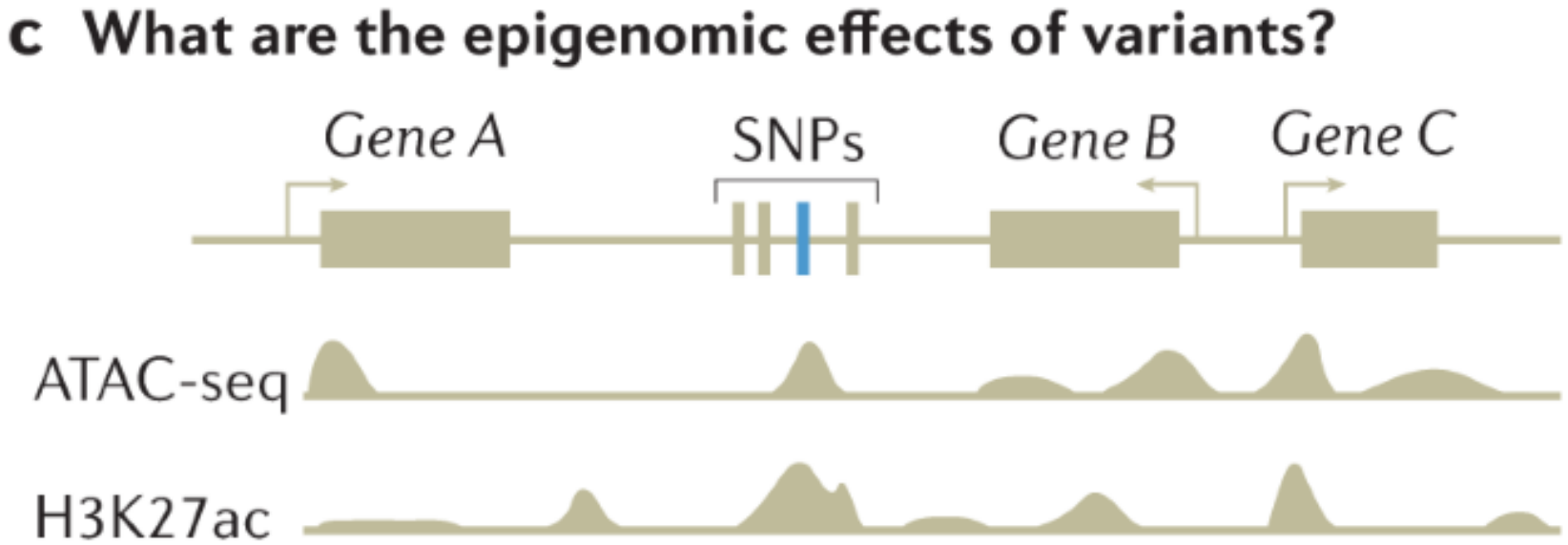
https://www.nature.com/articles/s43586-021-00056-9

Gleaning biological understanding from GWAS (combined with other methods)



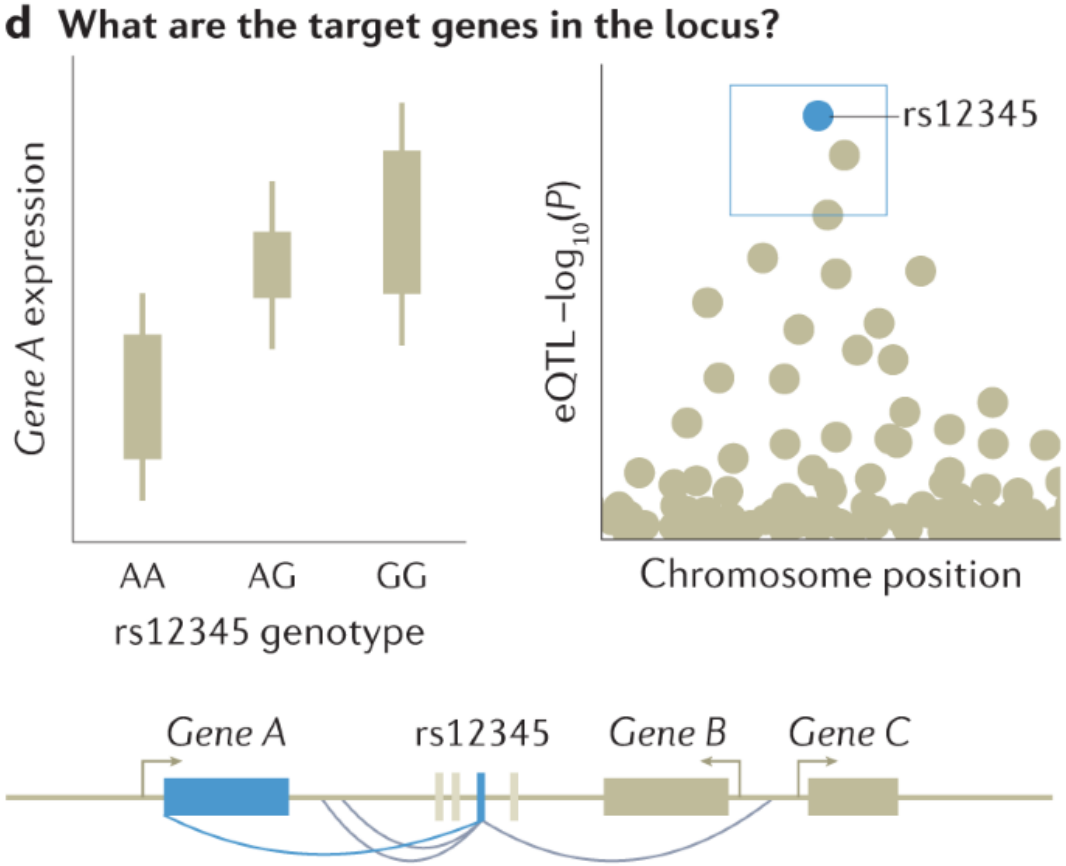
https://www.nature.com/articles/s43586-021-00056-9

Gleaning biological understanding from GWAS (combined with other methods)



https://www.nature.com/articles/s43586-021-00056-9

Gleaning biological understanding from GWAS (combined with other methods)



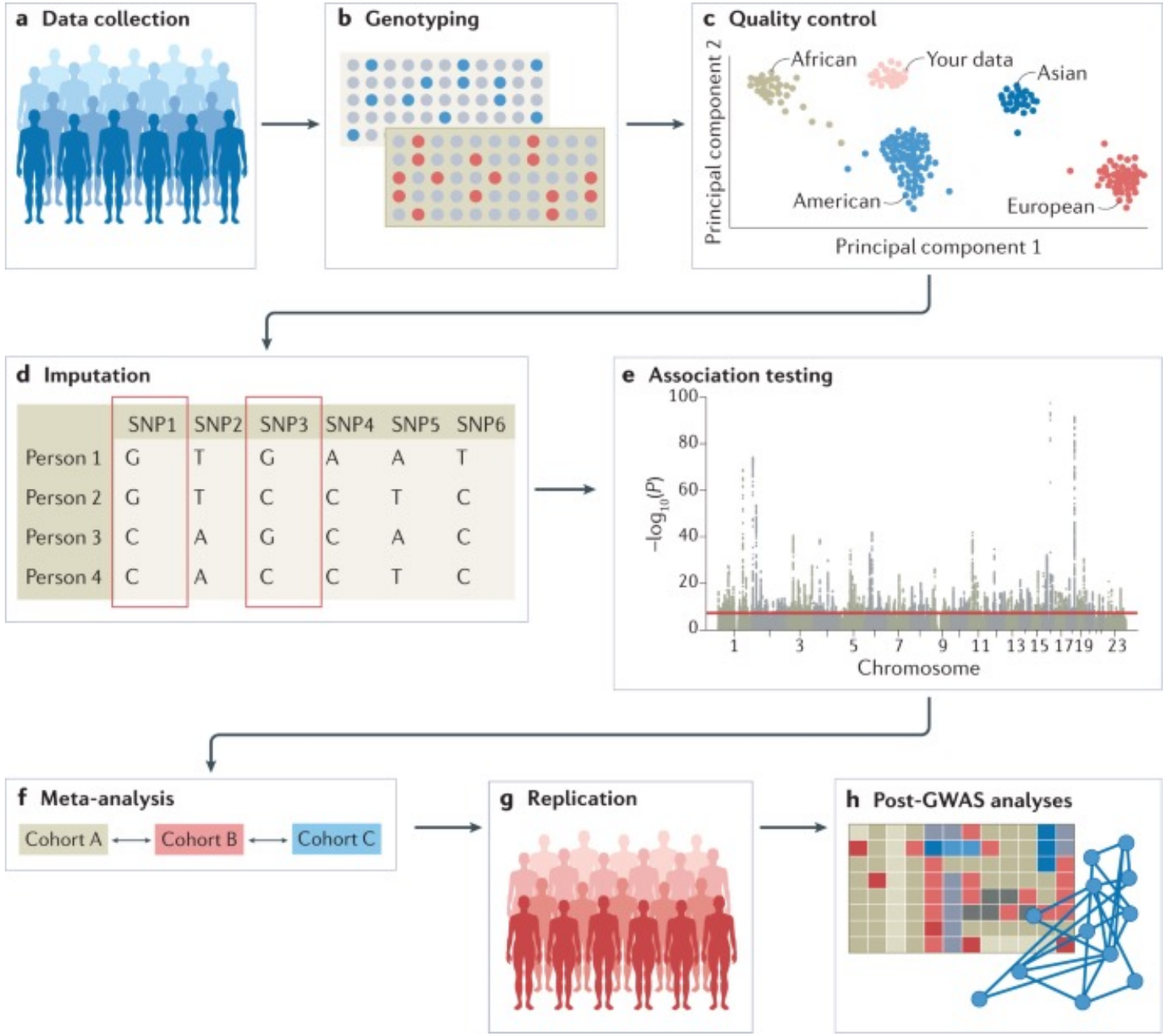
https://www.nature.com/articles/s43586-021-00056-9

Gleaning biological understanding from GWAS (combined with other methods)

e What are the affected pathways?

https://www.nature.com/articles/s43586-021-00056-9

Review: Schematic of a GWAS pipeline



https://www.nature.com/articles/s43586-021-00056-9