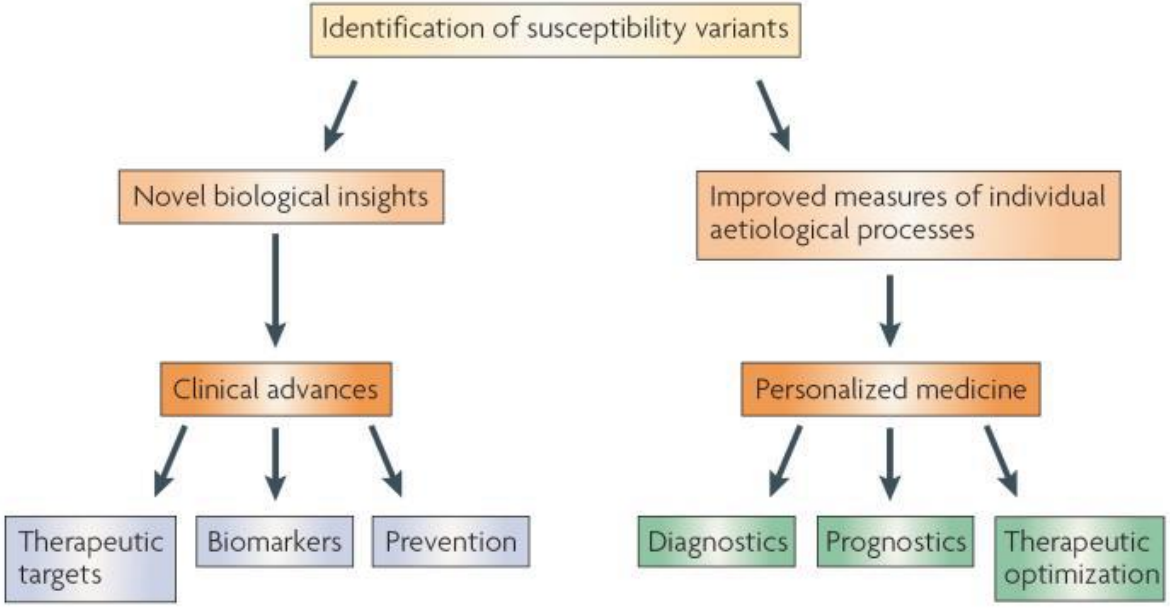
Genome-Wide Association Studies (GWAS)

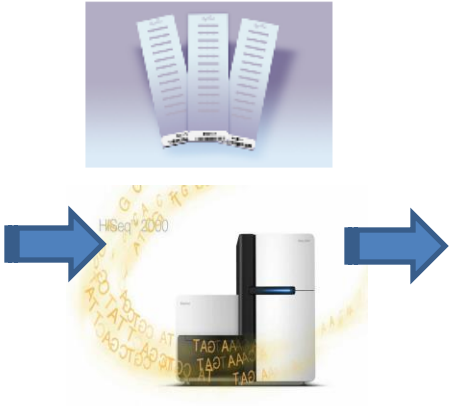
Purpose of Genetic Association Studies

- Determine if there is a genetic component contributing to a phenotype (i.e., disease) (**heritability**)
- Identify the genetic region/gene/polymorphism causing the disease (**association hit**)
- Determine the **effect size** of the genetic component



Genome wide association studies (GWAS)

- High-throughput approach scanning marker across the genome - linking genotype to phenotype
- Relies on dense sets of genetic markers - usually SNPs and SNP tags for other variation (via linkage disequilibrium (LD) or correlation between markers)
- Usually comparison of variation between affected (cases) and unaffected individuals (controls).
- Goal: Identify markers with significant associations to disease



SNP chips or sequencing with millions of SNPs

..ACTC**G**ACGATTTACG**G**TACTTAGGAGCATA**C**GCTAC..
..ACTC**T**ACGATTTACG**G**TACTTAGGAGCATA**C**GCTAC..
..ACTG**T**ACGATTTACG**A**TACTTAGGAGCATA**T**GCTAC..
..ACTG**T**ACGATTTACG**G**TACTTAGGAGCATA**T**GCTAC..
..ACTG**T**ACGATTTACG**G**TACTTAGGAGCATA**T**GCTAC..
..ACTG**T**ACGATTTACG**A**TACTTAGGAGCATA**G**GCTAC..
..ACTG**T**ACGATTTACG**A**TACTTAGGAGCATA**G**GCTAC..
..ACTG**T**ACGATTTACG**G**TACTTAGGAGCATA**T**GCTAC..
..ACTG**T**ACGATTTACG**A**TACTTAGGAGCATA**G**GCTAC..

SNPs may have 2, 3 or 4 alleles (most are biallelic)

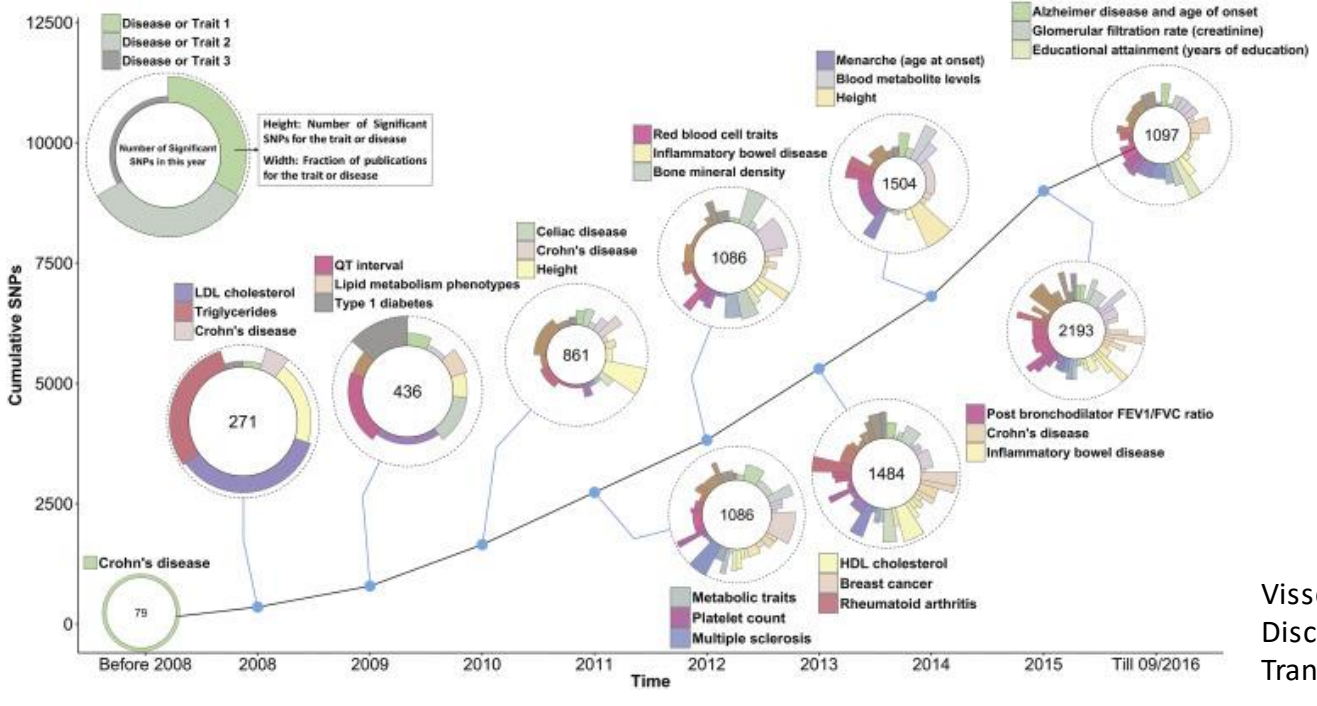
Lots of Success

May 2021 ($p \leq 5X10^{-8}$)

- >5700 studies
- >3,300 traits
- >70,000 trait associations

(Uffelmann et al., 2021)

NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

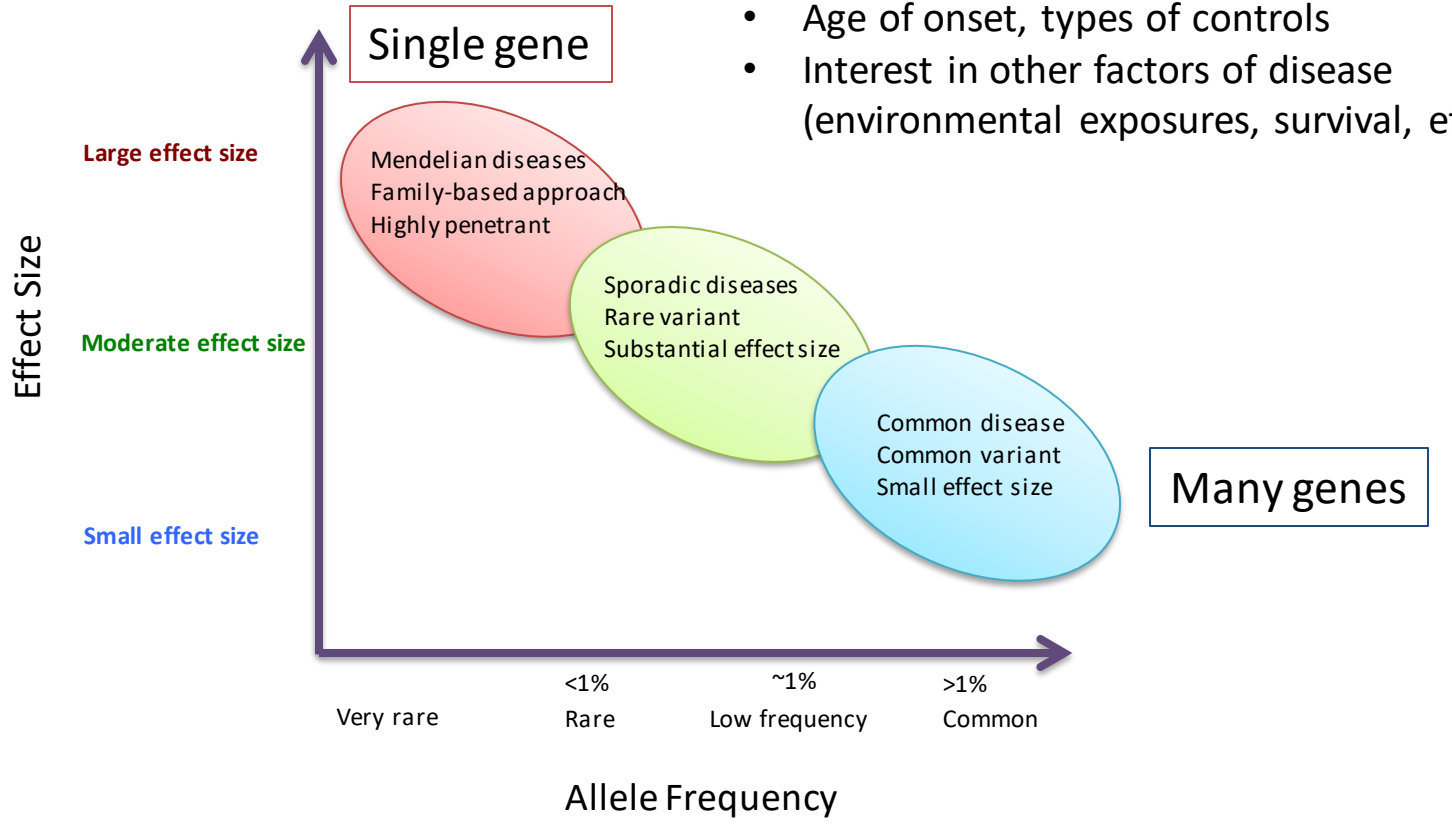


Visscher et al; 2017, 10 Years of GWAS Discovery: Biology, Function, and Translation

Variant Identification

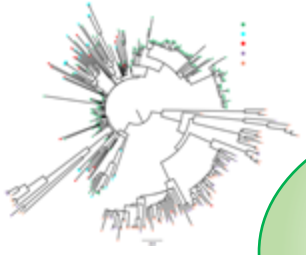
Study design is key

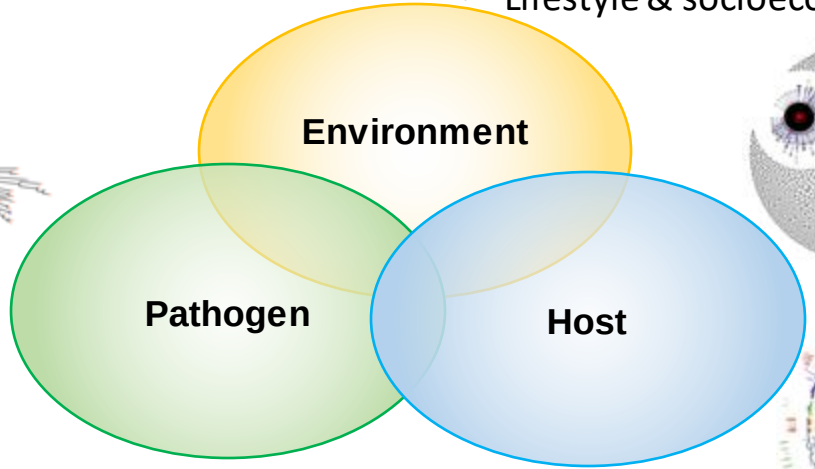
- Sample size?
- Age of onset, types of controls
- Interest in other factors of disease (environmental exposures, survival, effect size)

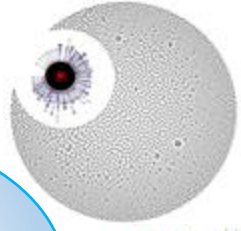


Multifactorial determinants of pathogenesis & clinical outcome

- Co-infection
- Immunosuppression
- Lifestyle & socioeconomic factors







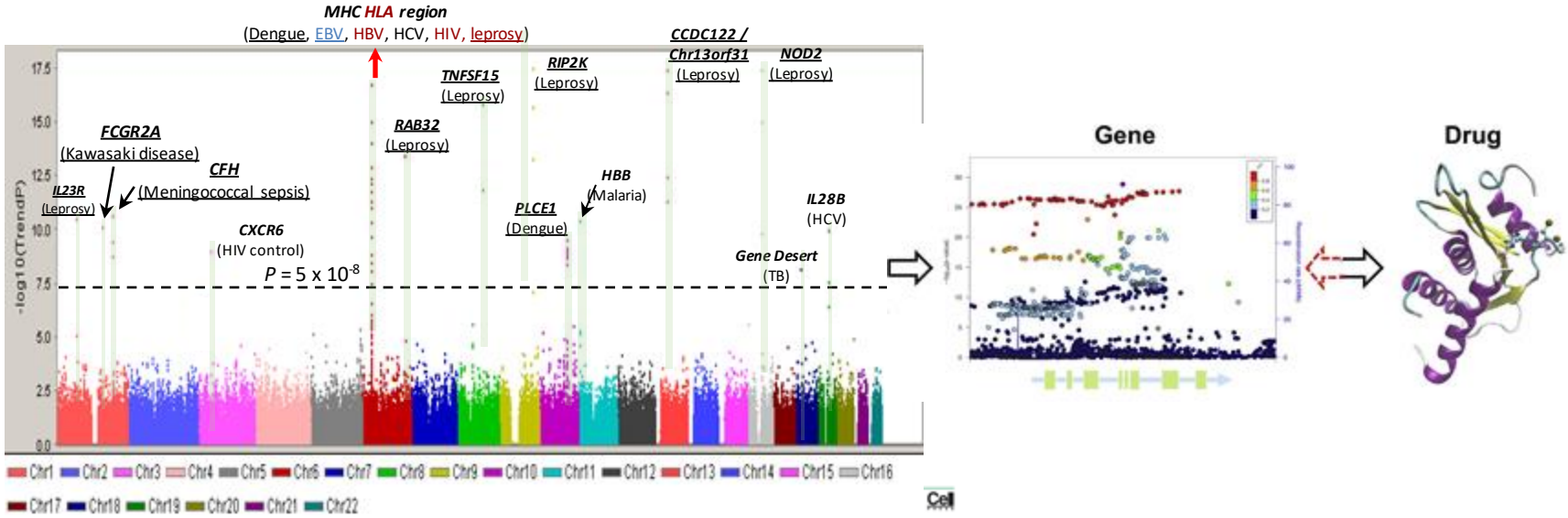
- What is the extent of pathogen genome diversity?
- Is there a link between pathogen genotype and phenotype?
- What are the transmission patterns?

- **Does host genetics contribute to susceptibility to infection/disease outcome?**
- **How do host genetic variants influence pathogen biological function?**

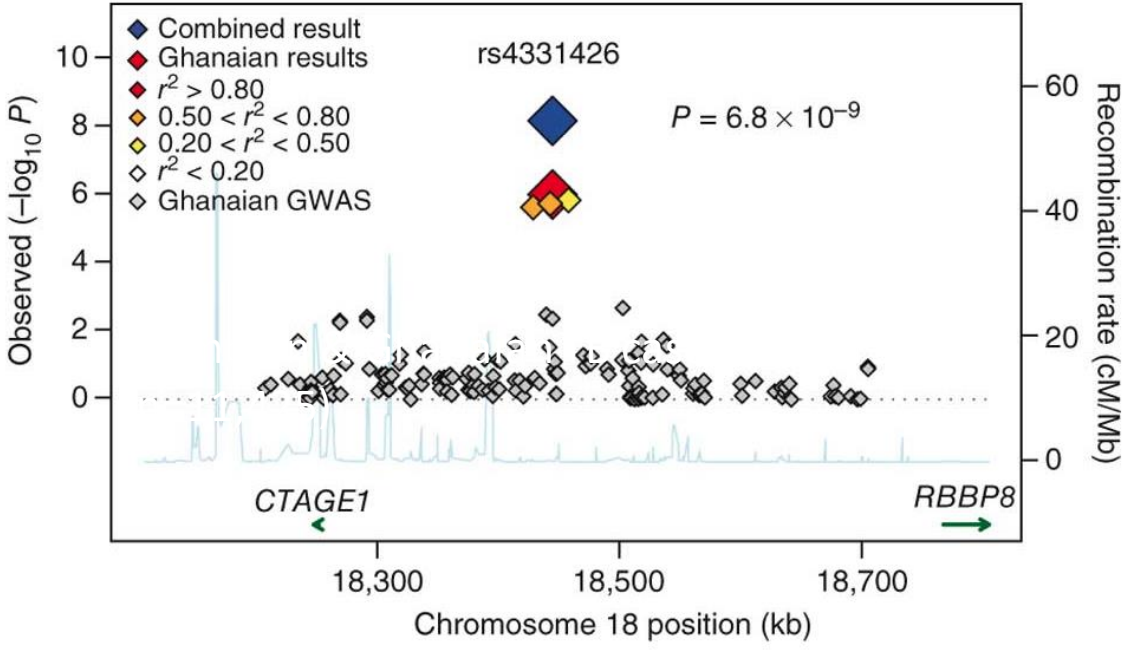
GWAS of Infectious diseases

Phenotypes Studied:

- Case-Control study: Susceptibility, severity, pathogen clearance, response to vaccination
- Quantitative trait: Antibody response, viral load, cell count



For TB host susceptibility , GWAS hits have not been replicated



Gambian & Ghanaian TB case-control study (n=11,425)

~500,000 SNPs across genome – Affymetrix 500K chip (Thye et al, 2010)

GWAS Workflow

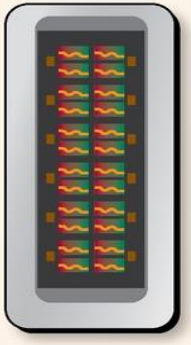
Sample Collection & Phenotype Determination → Genotyping or Whole genome sequencing → Quality Control → GWAS → Post-GWAS

Phenotypes: drug efficacy/toxicity

Case

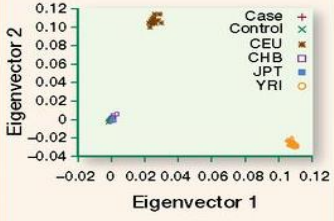
Control

Genotype with chip that contains probes for hundreds of thousands SNPs



Sample QC

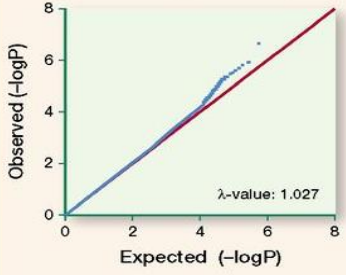
1. Sample quality
2. Relatedness
3. Population stratification



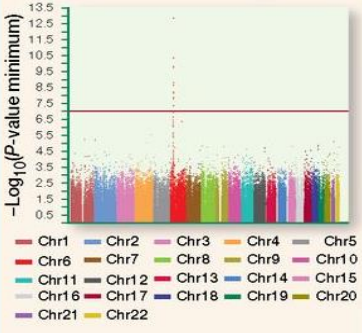
SNP QC

1. Genotype quality
2. Deviation from normal distribution
3. Allele frequency

Quantile–quantile plot



Manhattan plot



Validation study

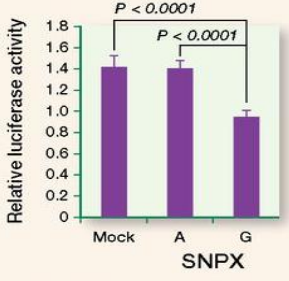
Replicate findings with independent sample set

1. Meta-analysis

Cohort	SNPX	OR (95% CI)
Study A		1.16 (1.06–1.27)
Study B		1.12 (1.00–1.26)
Study C		1.25 (1.12–1.39)
Replication		1.21 (1.10–1.33)
Total		1.19 (1.13–1.25)

2. Functional analysis

a. EMSA or b. reporter assay



3. Other analysis

Gene-based analysis, pathway analysis, polygenic risk estimation, SNP–SNP interaction, etc.

Need for high quality data

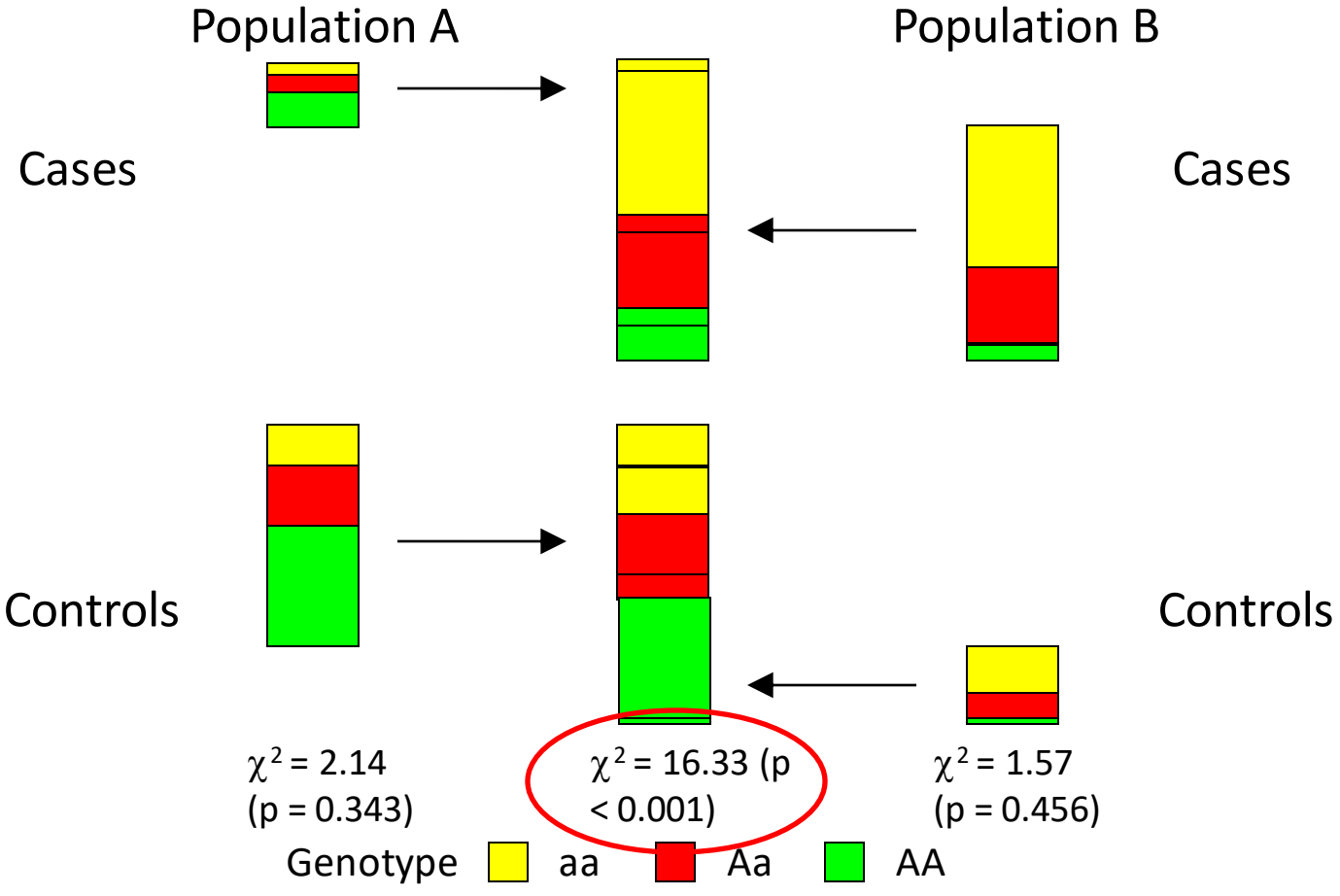
- Junk in -> junk out
- Many variants assayed $\Rightarrow$ errors and genotype/sequence miscalls occur
- If problematic samples are not identified and excluded, they can affect the results
- If SNPs with erroneous genotyping or sequencing not identified and excluded, they can produce false signals of associations
- **Perform QC on samples and SNPs**

Quality Checks

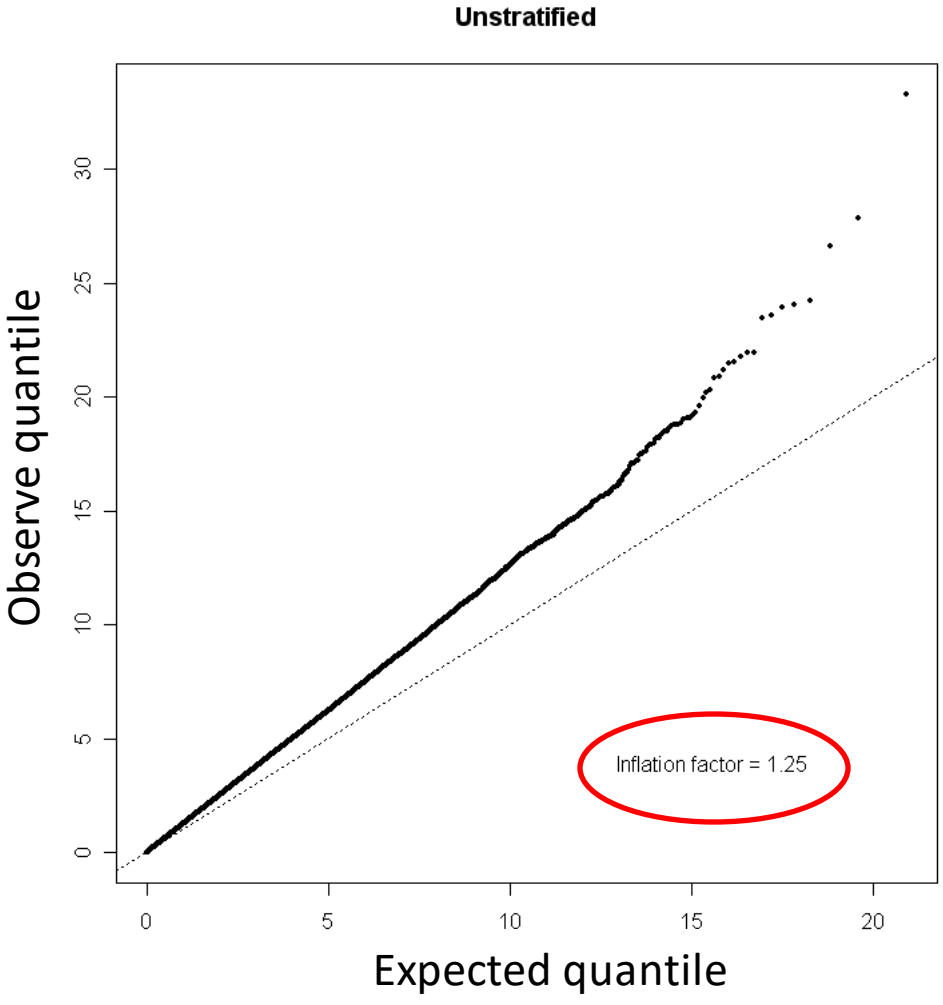
Variable	Comments
Genotyping Call Rate	Low call rate often correlates with error. Some low call rate SNPs or samples may still be good.
Genotyping Quality	Worse quality score (GenCall) correlates strongly with error rate
Sex concordance	Check expectations for X marker heterozygosity and Y marker positive results. Can estimate error rate.
Sample Relatedness	Check for related samples (expected or unexpected)
Mendelian Inheritance Errors	For trio/family data, can identify problem samples and families. Can estimate error rate.
Replicate concordance	Check for consistent genotype calls in duplicate samples
Batch effects	Check for genotyping call differences due to plate
Hardy-Weinberg Equilibrium	Violation across all sample groups may indicate error, but can also be a good test of association
Population Stratification	Check for population substructure using the genome-wide data

Effects of population stratification in an association study

… false positive association



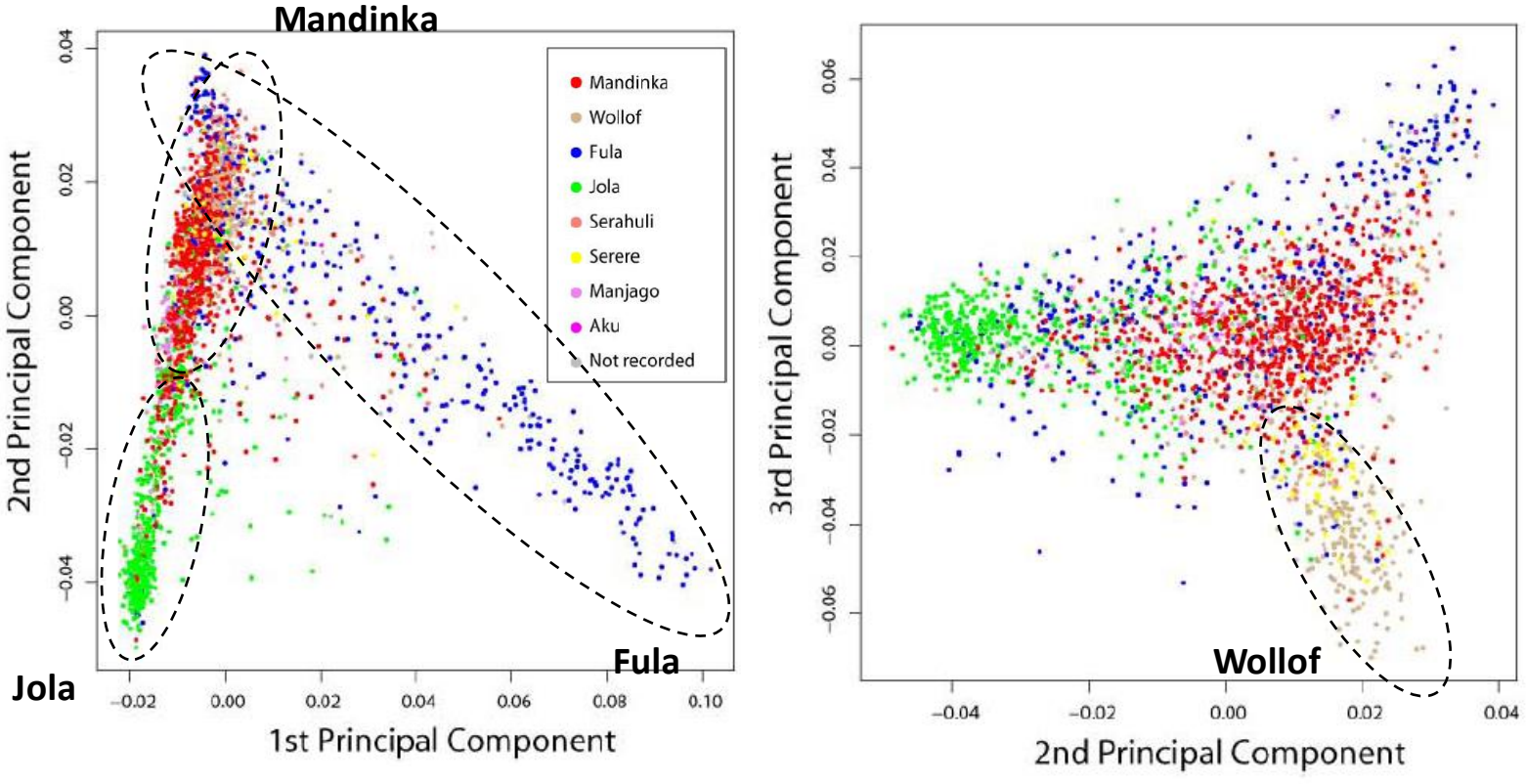
QQ-Plot for a GWAS affected by population structure



… lots of false positive associations

Principal component analysis can give insights into the population structure

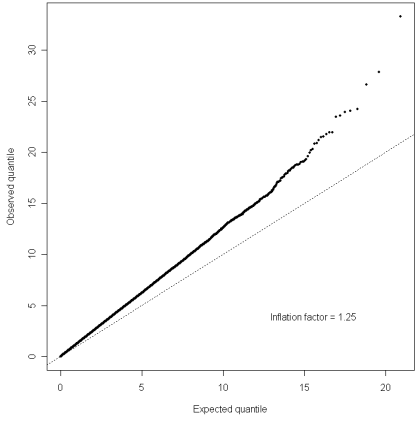
Jallow et al, 2010



We can adjust for these types of variables (surrogates for population structure) in association studies, or simply stratify by self-reported ethnic or language groups

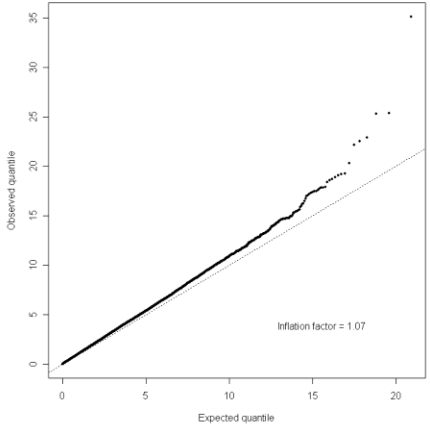
QQ-Plots after correcting for population structure

Unstratified



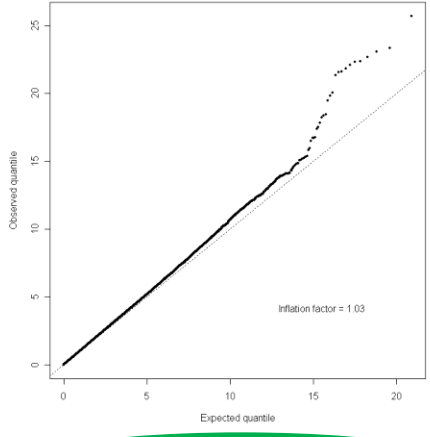
Inflation factor = 1.25

Stratified



Inflation factor = 1.07

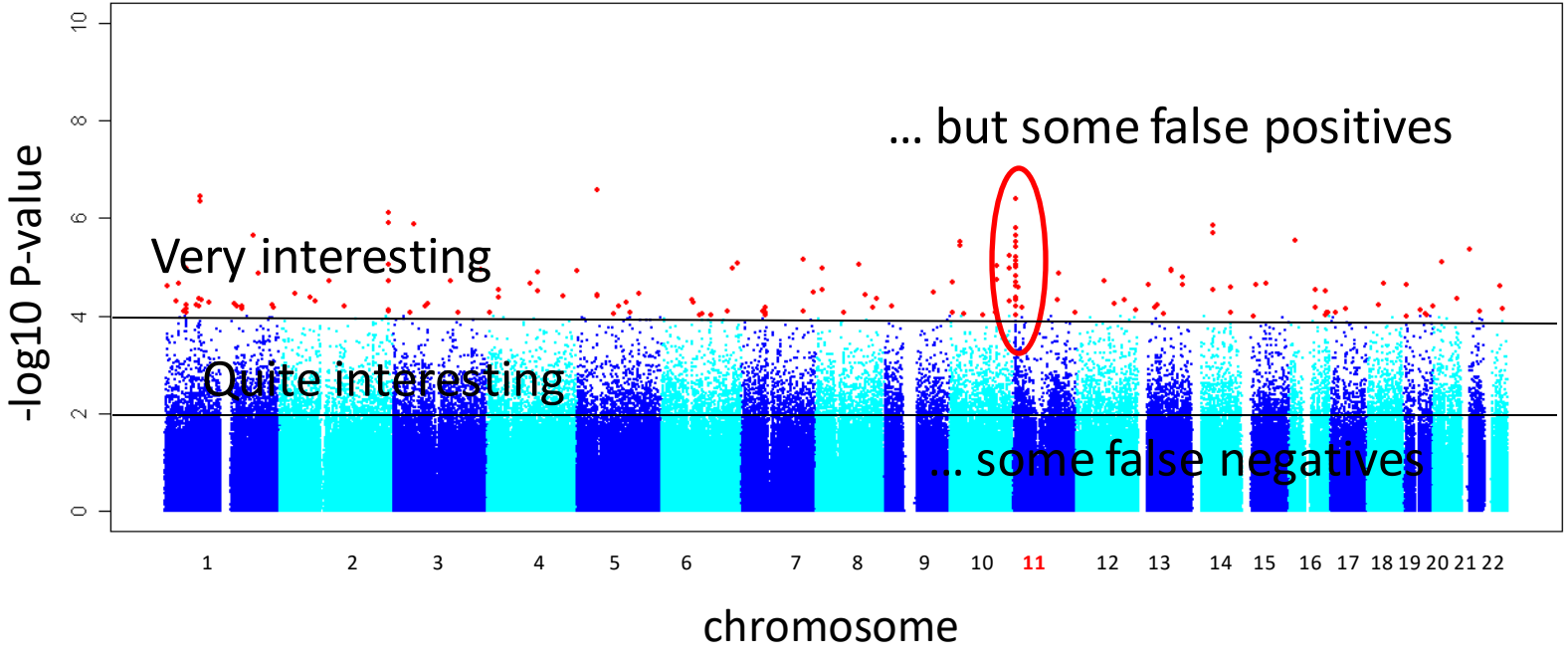
Principal components-corrected



Inflation factor = 1.03

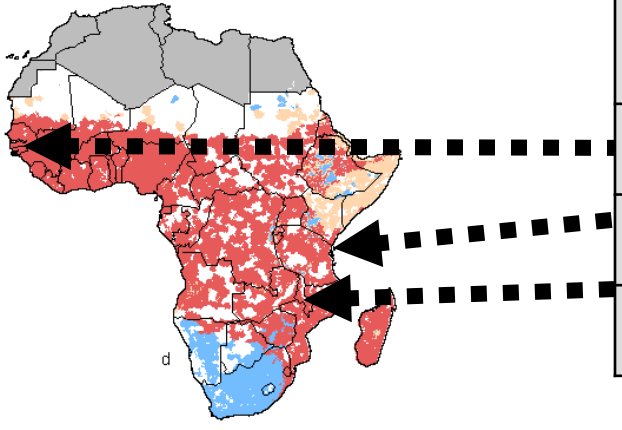
Association signals across the genome

Manhattan plot: Severe malaria GWAS (n=2,500)



~500,000 SNPs across genome – Affymetrix 500K chip (Jallow et al, 2009)

Sickle trait is the strongest known determinant of severe malaria risk



	Relative risk of severe malaria in children with HbS AS genotype
Gambia	0.11
Kenya	0.17
Malawi	0.08

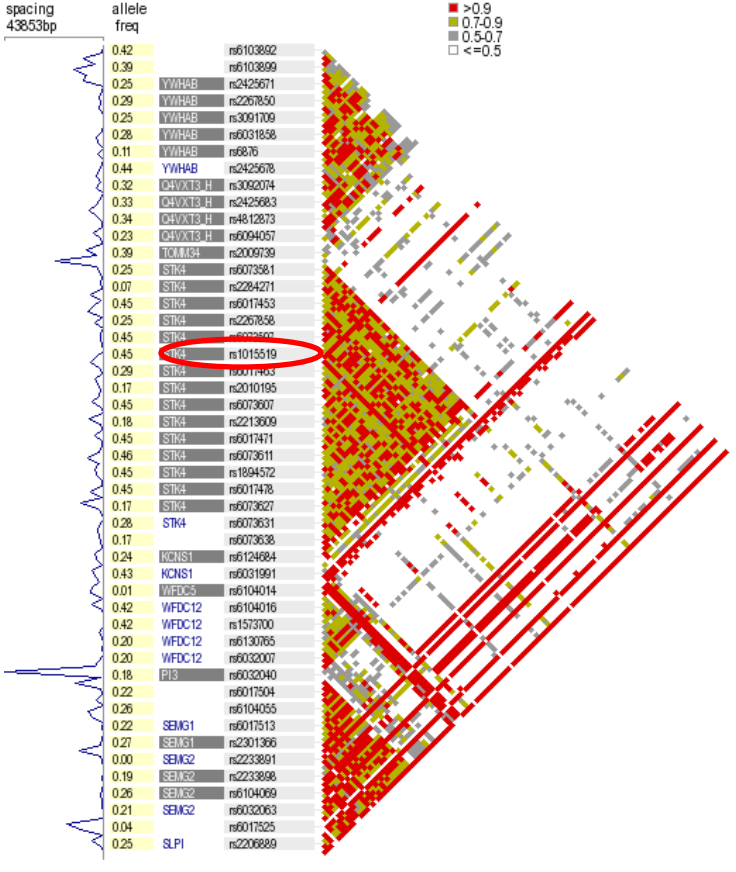
$P = 2 \times 10^{-31}$ ($n = 3630$)

- Genetic factors determine 25% of malaria risk in Kenyan children and sickle trait accounts for only 2% of total variation (Mackinnon et al, 2005)

Association Studies

Direct Association

Tests the genetic variant directly responsible for causing the disease.



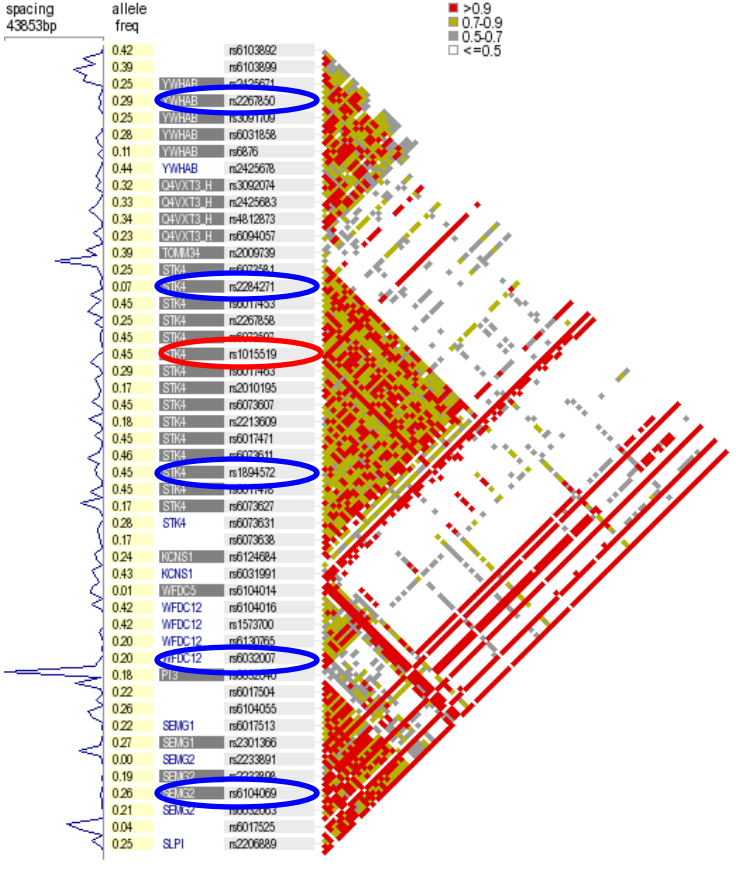
Association Studies

Direct Association

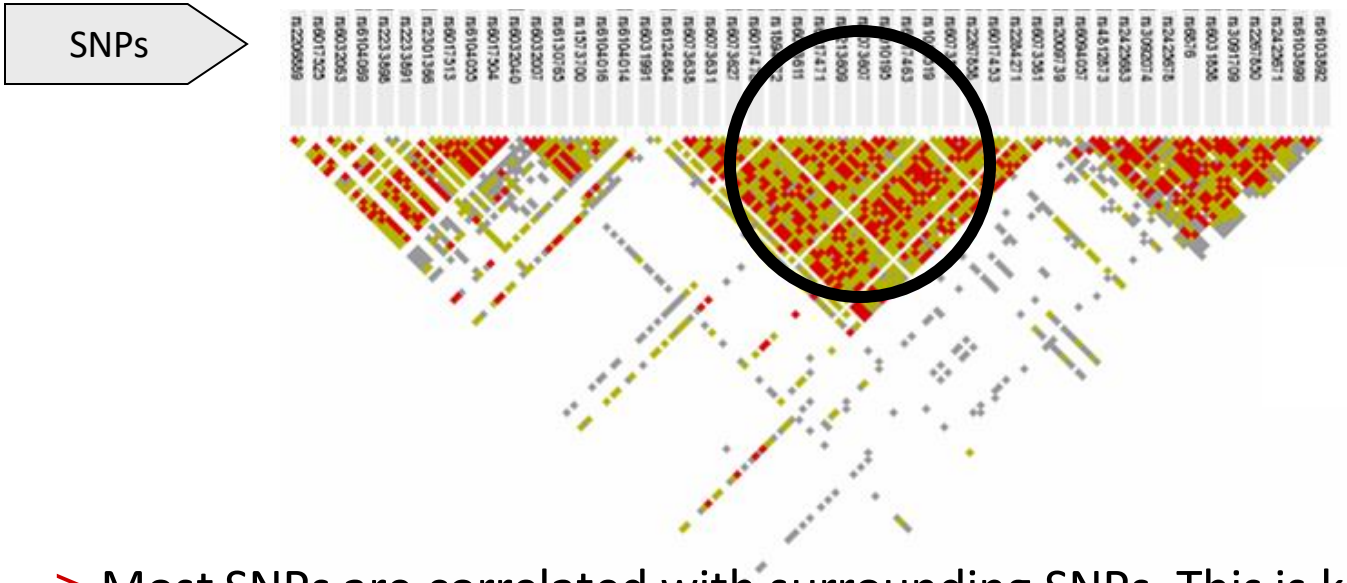
Tests the genetic variant directly responsible for causing the disease.

Indirect Association

Genetic variant tested is not directly responsible for the disease, but is located near to the disease-causing variant and thus ‘correlated’, or in linkage disequilibrium (LD).



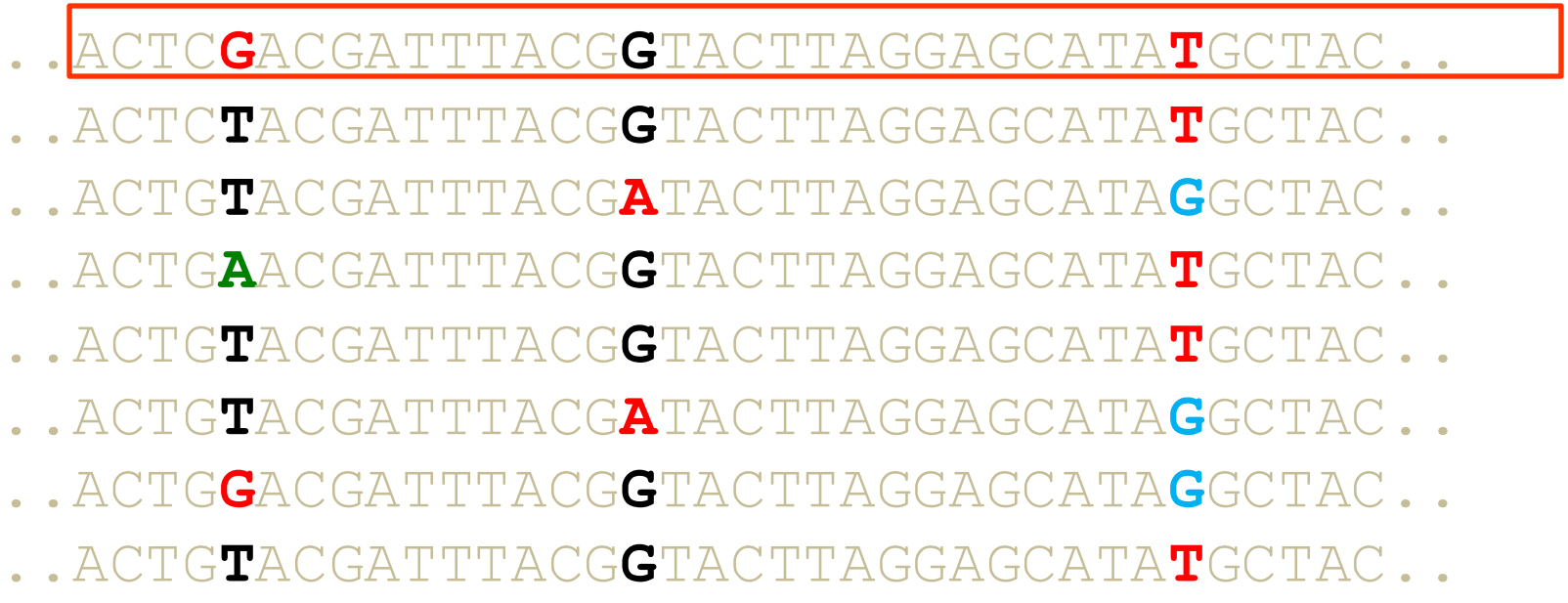
Each population has a distinct pattern of genome variation



> Most SNPs are correlated with surrounding SNPs. This is known as **linkage disequilibrium (LD)**

> Linkage disequilibrium reflects the common combinations of variants (haplotypes) that exist in the population

Haplotypes



- **A haplotype is an observed sequence of variants**
- Each population has its own pattern of common haplotypes
- By knowing the pattern of haplotypes within a population we may be able to impute genotype at an untyped position

Why is LD important in humans?

- >15 million genetic variants in the human genome, costly to genotype everything (pre-2012?)
- LD $\Rightarrow$ Reduced amount of genotyping required
- The availability of whole genome sequencing on large numbers of samples makes LD redundant
- Allows us to perform imputation to fill in gaps in arrays

Imputation

TCCGGACACCTTCTAAGG
TCTGGACACCTTCTAAGG
TCTGTACACAGGATTTCG
ACTGGACACAGGATTTGG
ACCGTCTTCCTTCTAACG
TCCGGACACCTTCTAAGG

Reference panel (e.g., 1000 genomes project):

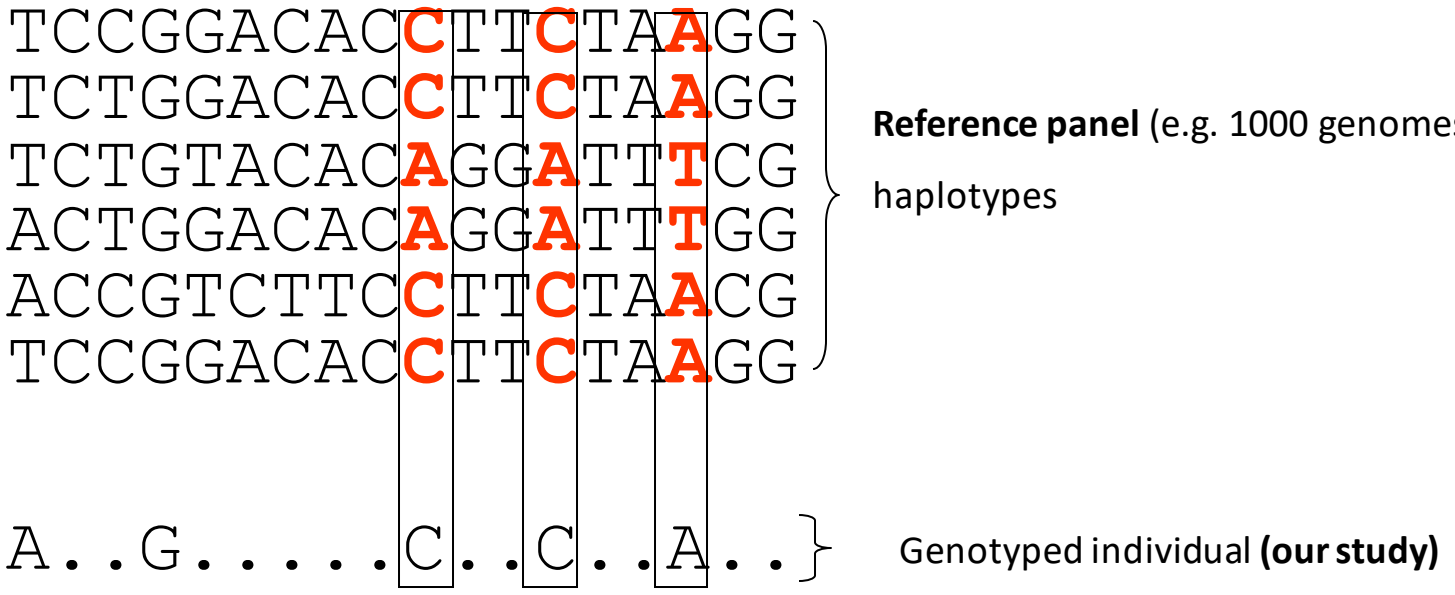
haplotypes

A..G.....C..C..A..

Genotyped individual **(our study)**

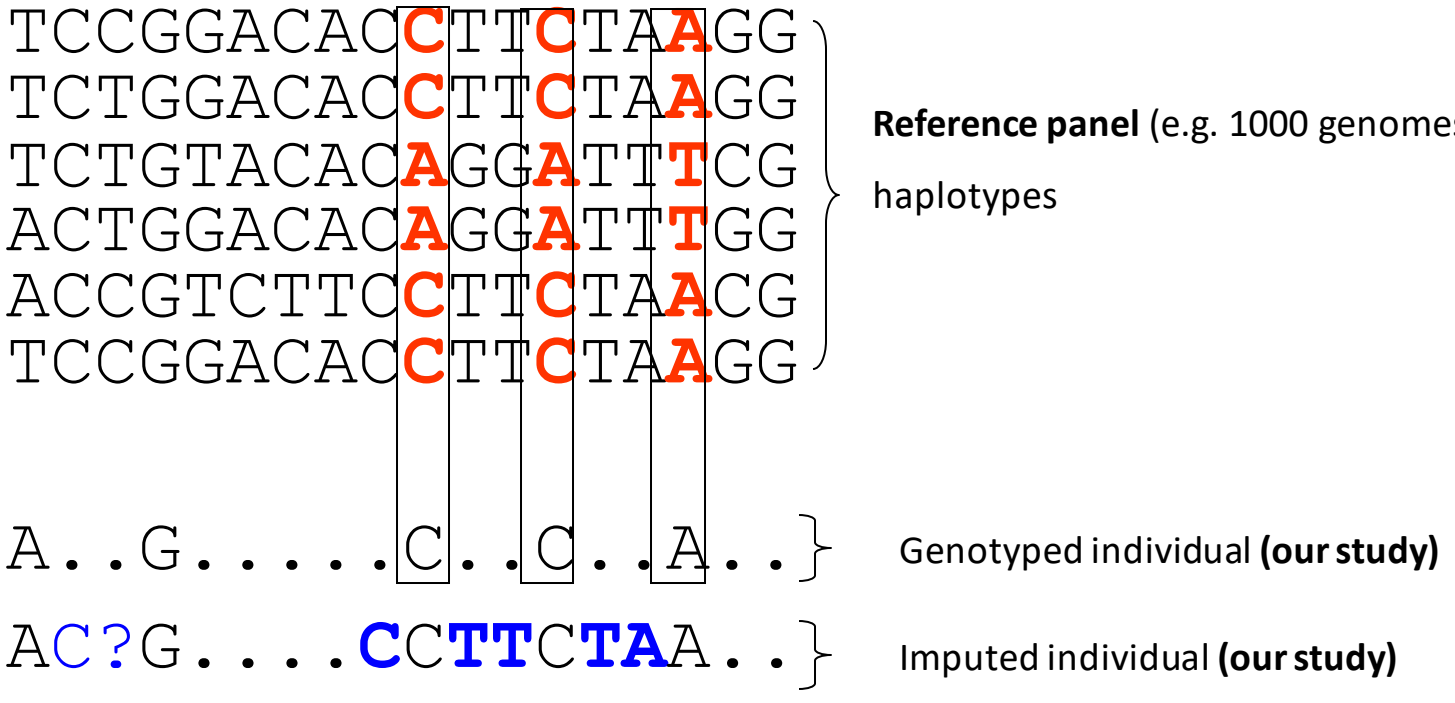
Using LD in the data to fill in the blanks.

Imputation



Using LD in the data to fill in the blanks.

Imputation

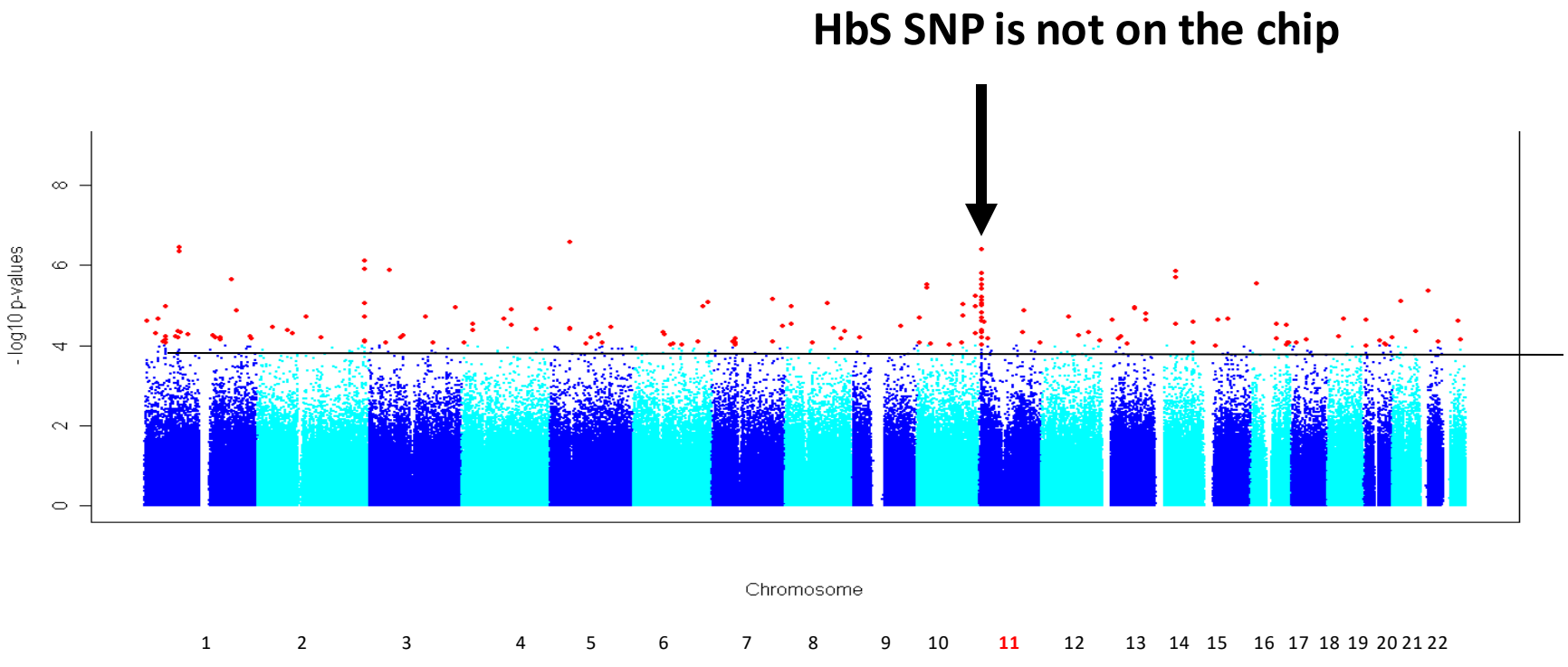


Using LD in the data to fill in the blanks.

Why impute?

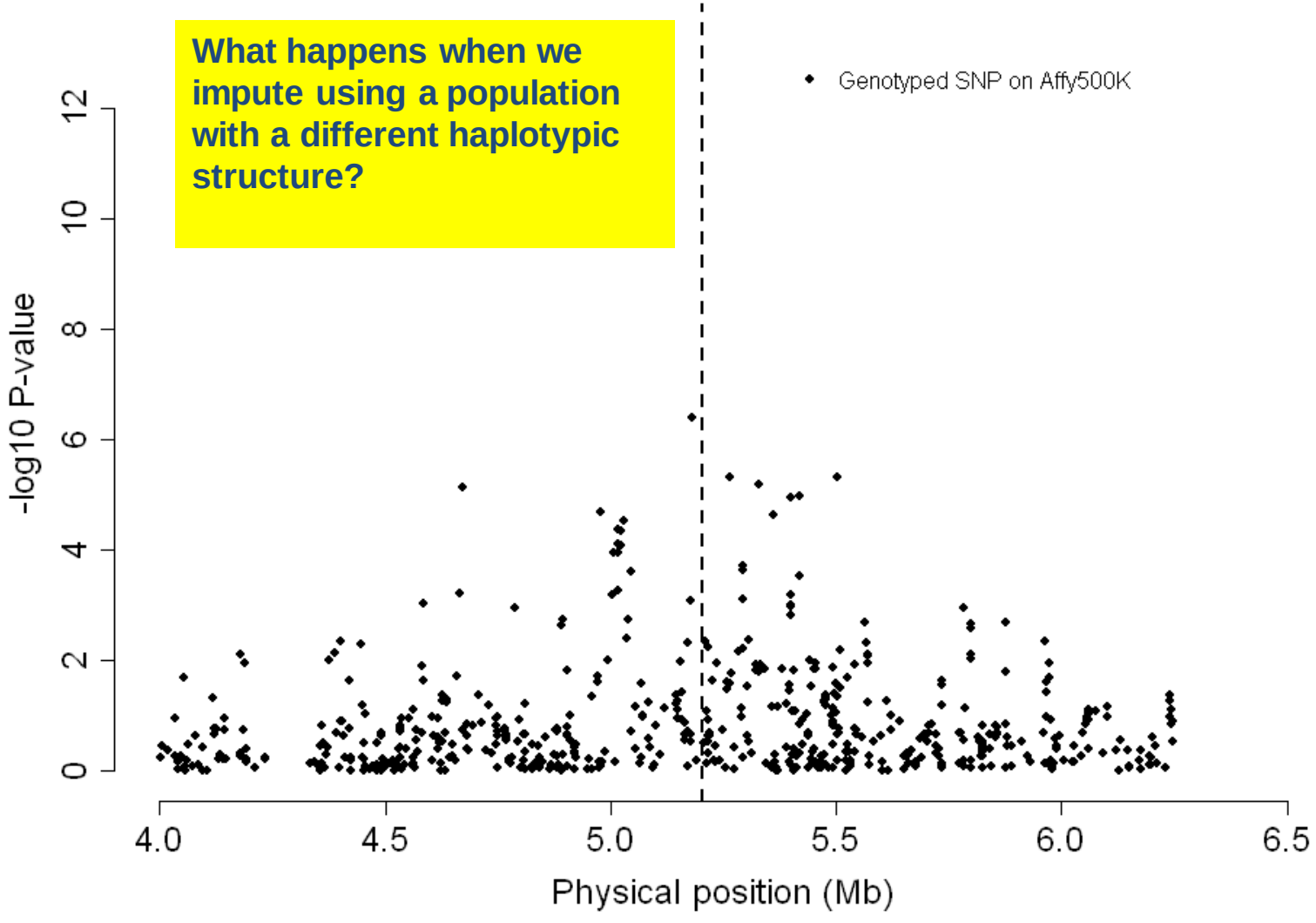
- To predict missing genotypes that have not been directly typed
  - **Increased power.** The reference panel is more likely to contain the causal variant (or a better tag) than a GWAS array.
  - **Fine-mapping.** Imputation provides a high-resolution overview of an association signal across a locus.
  - **Meta-analysis.** Imputation allows GWAS typed with different arrays to be combined up to variants in the reference panel.

Imputing to find the “causal” polymorphisms

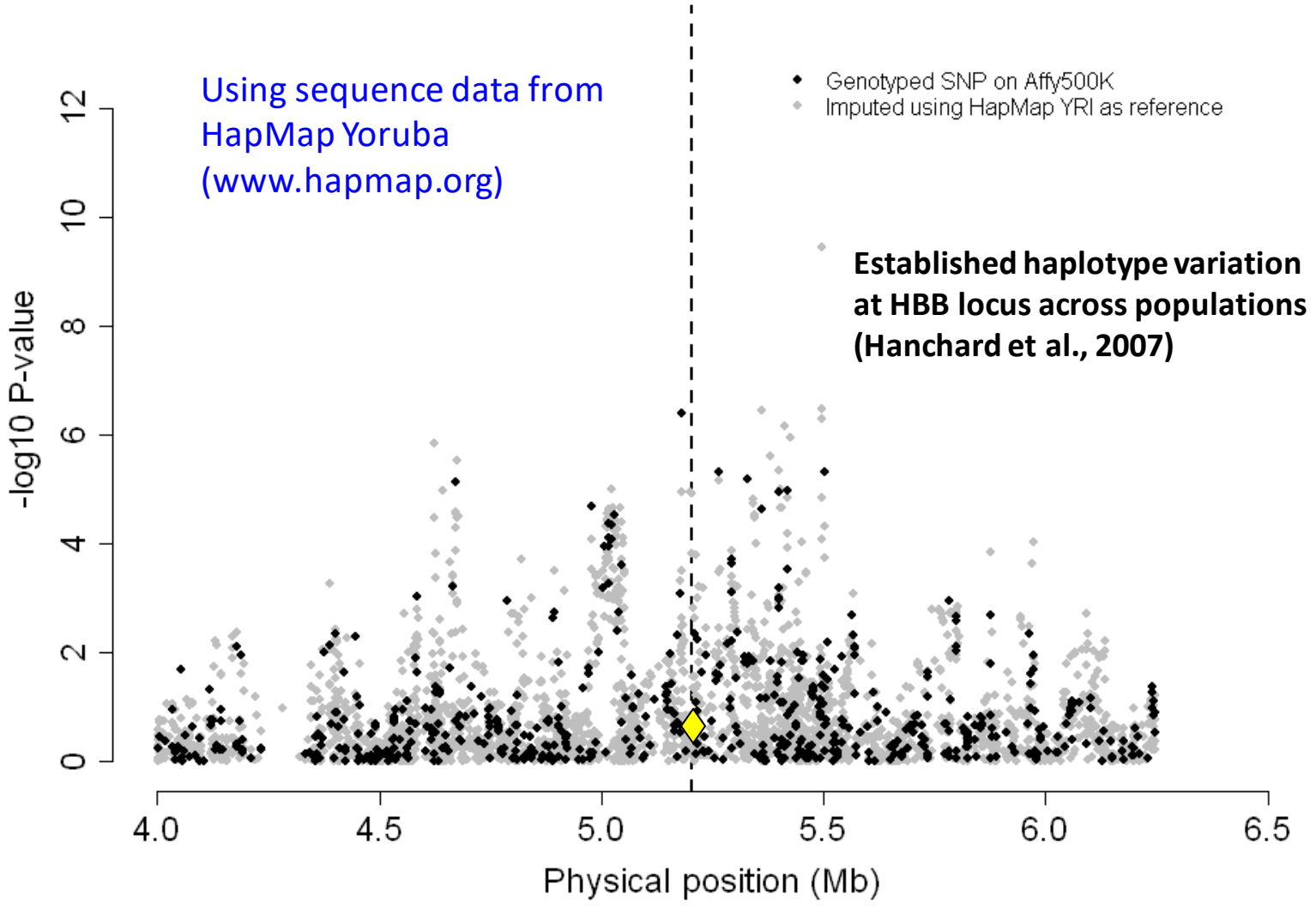


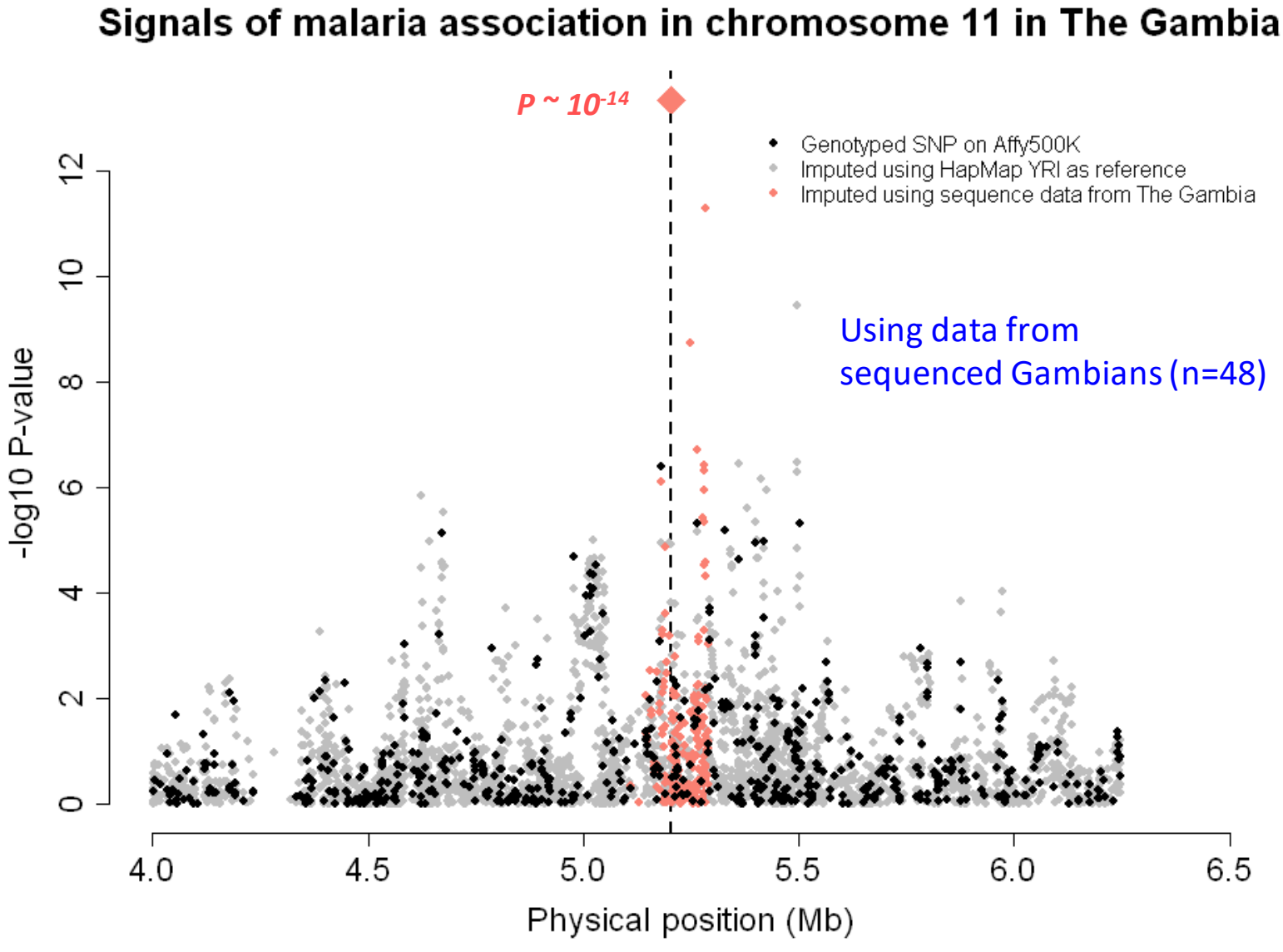
~500,000 SNPs across genome – Affymetrix 500K chip (Jallow et al, 2009)

Signals of malaria association in chromosome 11 in The Gambia



Signals of malaria association in chromosome 11 in The Gambia




Signals of malaria association in chromosome 11 in The Gambia
P ~ 10-14
Genotyped SNP on Affy500K
Imputed using HapMap YRI as reference
Imputed using sequence data from The Gambia
Using data from
sequenced Gambians (n=48)
-log10 P-value
0
2
4
6
8
10
12
4.0
4.5
5.0
5.5
6.0
6.5
Physical position (Mb)

Going beyond GWAS

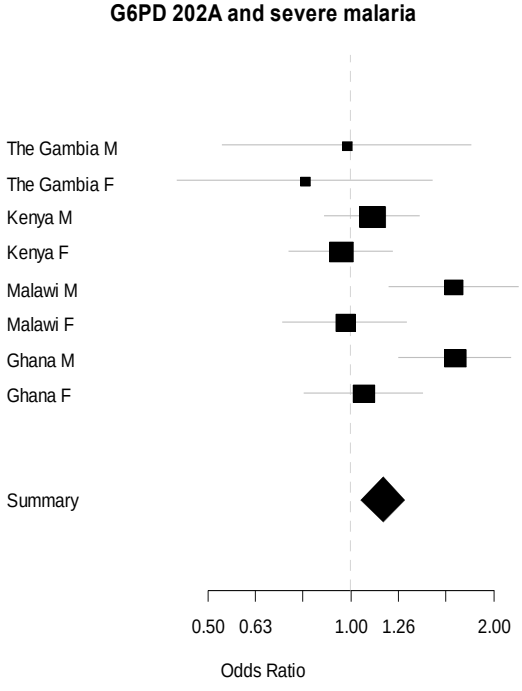
- Need to validate and confirm findings
  - Replication studies and meta analysis
- If using genotyping arrays, fine-mapping the causal variant
  - Targeted-resequencing
  - Transethnic mapping
- Functional studies

Replication

- Assay a small subset of SNPs that arose from GWAS scan
- Ideally within same population, but often unlikely
- Aim to replicate in other populations for a similarly defined phenotype
- Population structure:
  - Problematic, since we will not have genome-wide data to assess extent of confounding
  - Have to rely on informative surrogates if available (e.g. self-reported ethnicity, language, location)

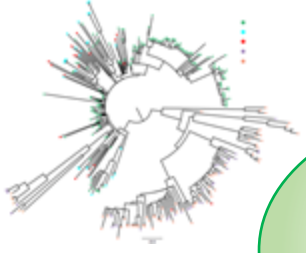
Meta-analysis

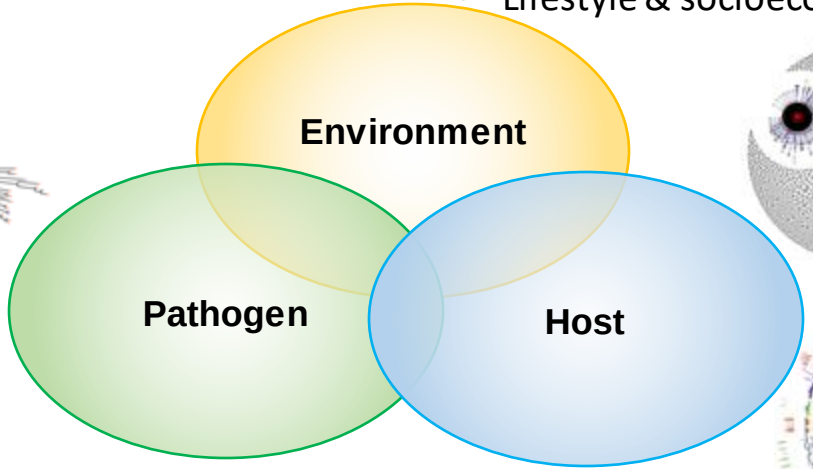
- Combine multiple genome-wide scans of the same phenotype
- Consistency of phenotypic definition is crucial
- Genome-wide pooling (less publication bias)
- Benefits:
  - Increased power to detect new loci add pathways, across allele frequencies
  - Reveal study heterogeneity (e.g., especially when combining populations of different ancestry)
  - Heterogeneity could be due to differences in study design, population structure, environmental exposures or pathogen diversity

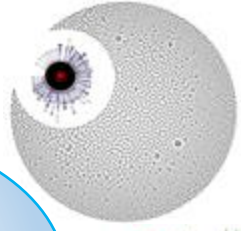


Host pathogen interactions

- Co-infection
- Immunosuppression
- Lifestyle & socioeconomic factors





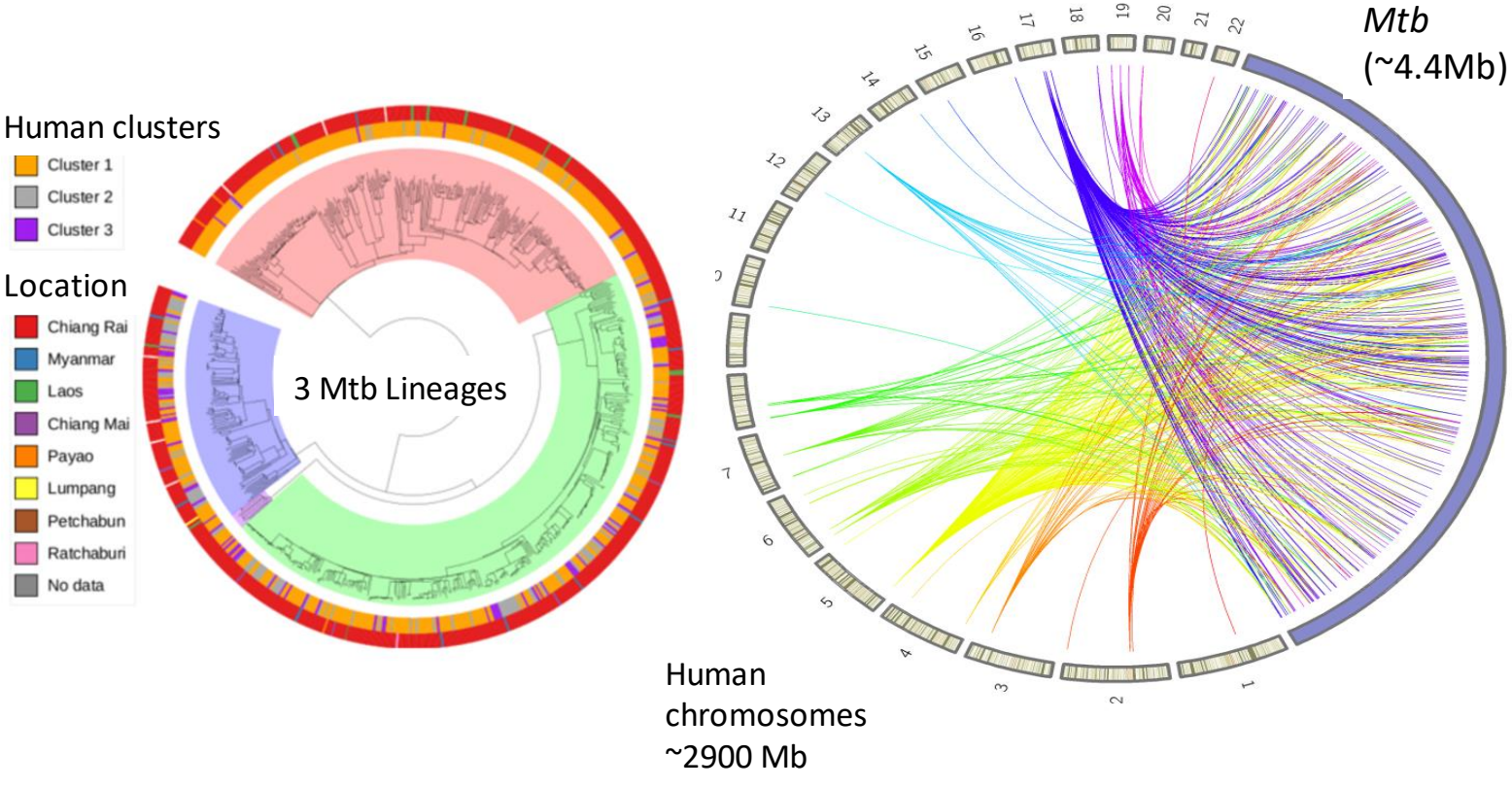


- What is the extent of pathogen genome diversity?
- Is there a link between pathogen genotype and phenotype?
- What are the transmission patterns?

- **Does host genetics contribute to susceptibility to infection/disease outcome?**
- **How do host genetic variants influence pathogen biological function?**

TB host - *M. tuberculosis* interactome through genome-to-genome analysis

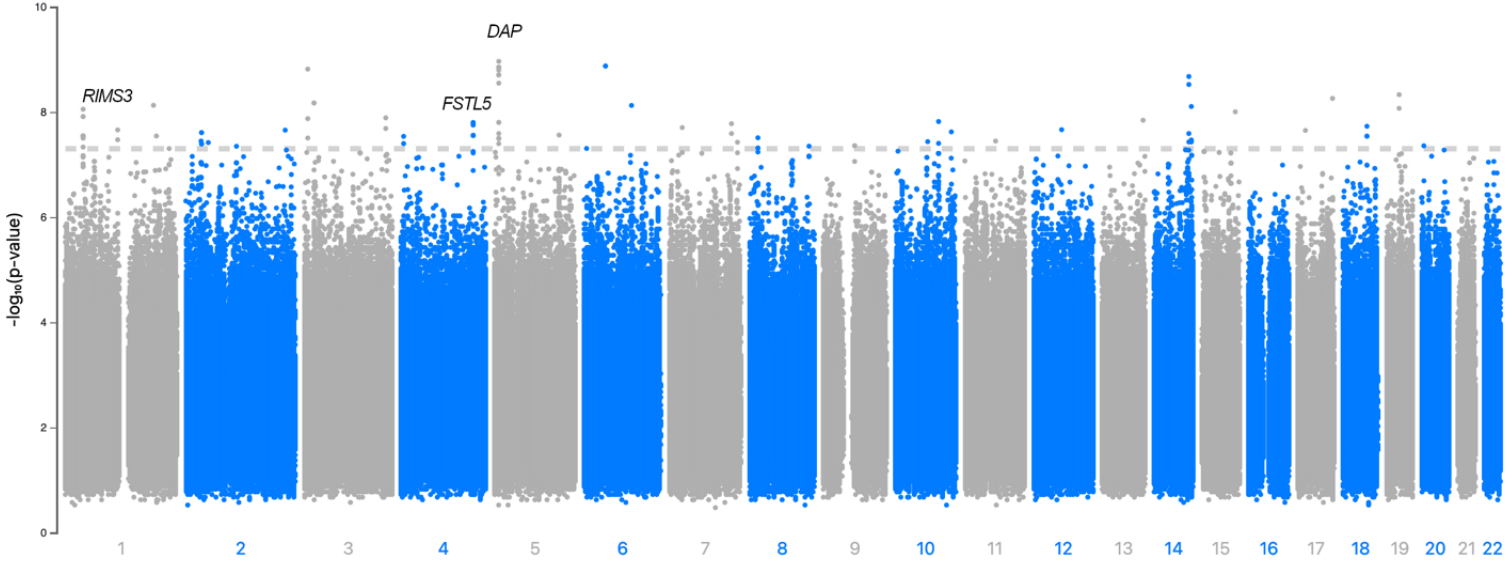
720 TB patients from Thailand, with paired human and Mtb DNA



Used a GWAS approach guided by a phylogenetic tree to identify 8 specific genetic interaction points ($P<5x10^{-8}$)

Phelan et al, Nature Comms, 2023

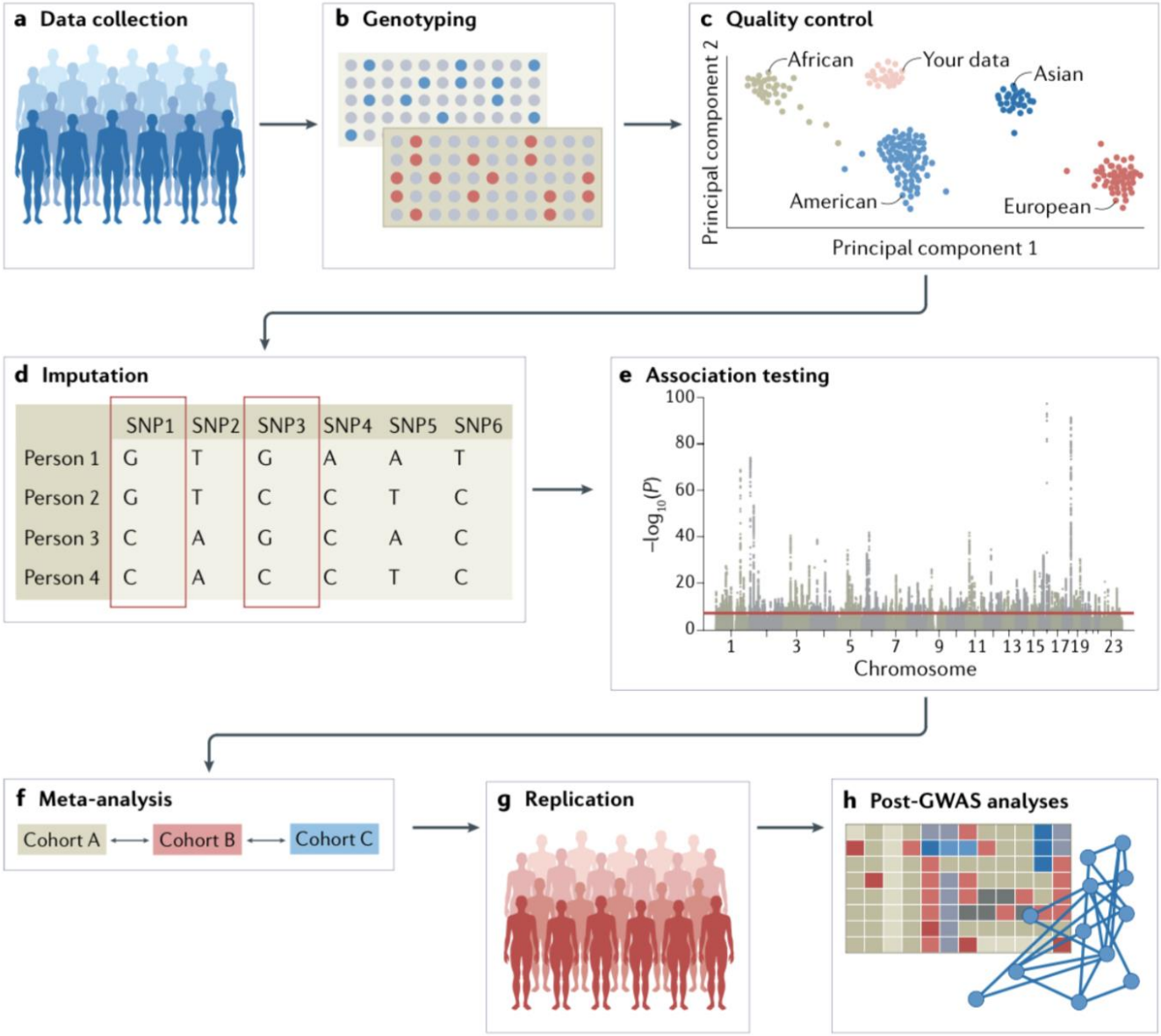
TB host - *M. tuberculosis* interactome through genome-to-genome analysis



Human loci *DAP* and *RIMS3* are both linked to the IFNγ cytokine and host immune system, as well as *FSTL5*, previously associated with susceptibility to TB.

Phelan et al, Nature Comms, 2023

GWAS Design and analysis - recap



Useful GWAS analysis tools

- SNP calling
  - Samtools : http://samtools.sourceforge.net
  - GATK : https://software.broadinstitute.org/gatk
  - OptiCall : https://opticall.bitbucket.io

- Data Imputation
  - Impute2: http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
  - Beagle: https://faculty.washington.edu/browning/beagle/beagle.html
  - Sanger Imputation Server: https://imputation.sanger.ac.uk

- Publically available datasets:
  - 1000 Genomes: http://www.internationalgenome.org/data
  - Exac: http://exac.broadinstitute.org
  - UK10K: https://www.uk10k.org
  - HRC: http://www.haplotype-reference-consortium.org
  - African Genome Variation Project: https://www.sanger.ac.uk/science/collaboration/african-genome-variation-project
  - UKBioBank: https://www.ukbiobank.ac.uk

- Analysis:
  - Plink: http://zzz.bwh.harvard.edu/plink/
  - SNPtest: https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html
  - GEMMA: http://www.xzlab.org/software.html
  - R Packages: https://cran.r-project.org/web/packages/SNPassoc/SNPassoc.pdf
  - GCTA: http://cnsgenomics.com/software/gcta/#Overview