Genome-Wide Association Studies (GWAS)

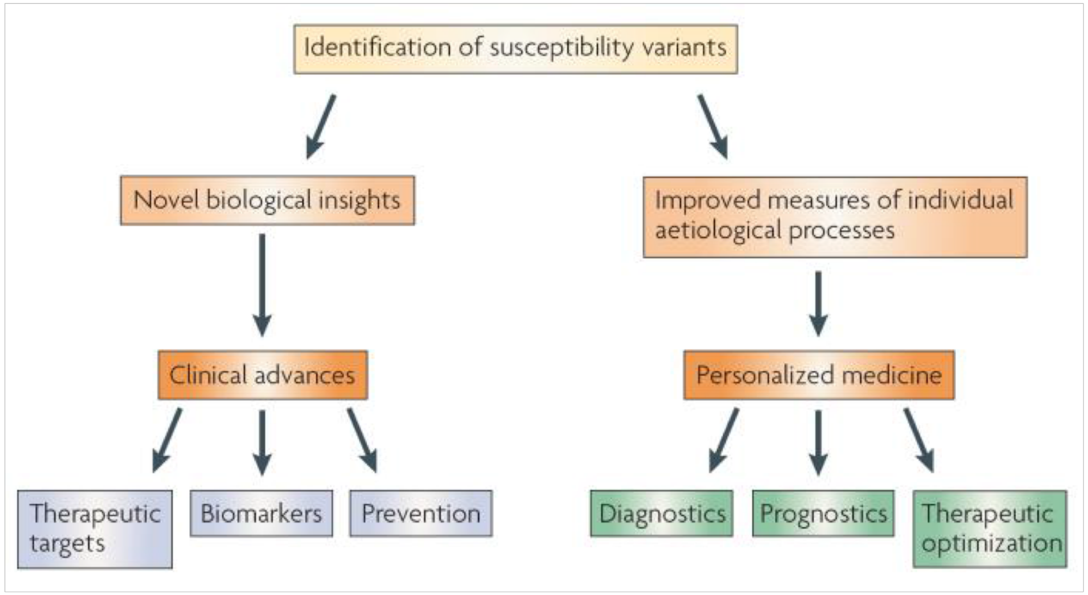
London School of Hygiene and Tropical Medicine

Outline

- GWAS overview – Utility & Successes
- GWAS Study Design
- Post GWAS Follow up
- Recap
- Case study

Purpose of Genetic Association Studies

- Determine if there is a genetic component contributing to phenotype (i.e. disease) under investigation (heritability)
- Identify the genetic region/gene/polymorphism causing the disease
- 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 LD)
- Usually comparison of variation between affected (cases) and unaffected individuals (controls).
- Goal: Identify markers with significant associations to disease

```
..ACTCGACGATTTACGGTACTTAGGAGCATACGCTAC..
..ACTCTACGATTTACGGTACTTAGGAGCATACGCTAC..
..ACTGTACGATTTACGATACTTAGGAGCATATGCTAC..
..ACTGTACGATTTACGGTACTTAGGAGCATATGCTAC..
..ACTGTACGATTTACGGTACTTAGGAGCATATGCTAC..
..ACTGTACGATTTACGATACTTAGGAGCATAGGCTAC..
..ACTGTACGATTTACGATACTTAGGAGCATAGGCTAC..
..ACTGTACGATTTACGGTACTTAGGAGCATATGCTAC..
..ACTGTACGATTTACGATACTTAGGAGCATAGGCTAC..
```

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

Lots of Success

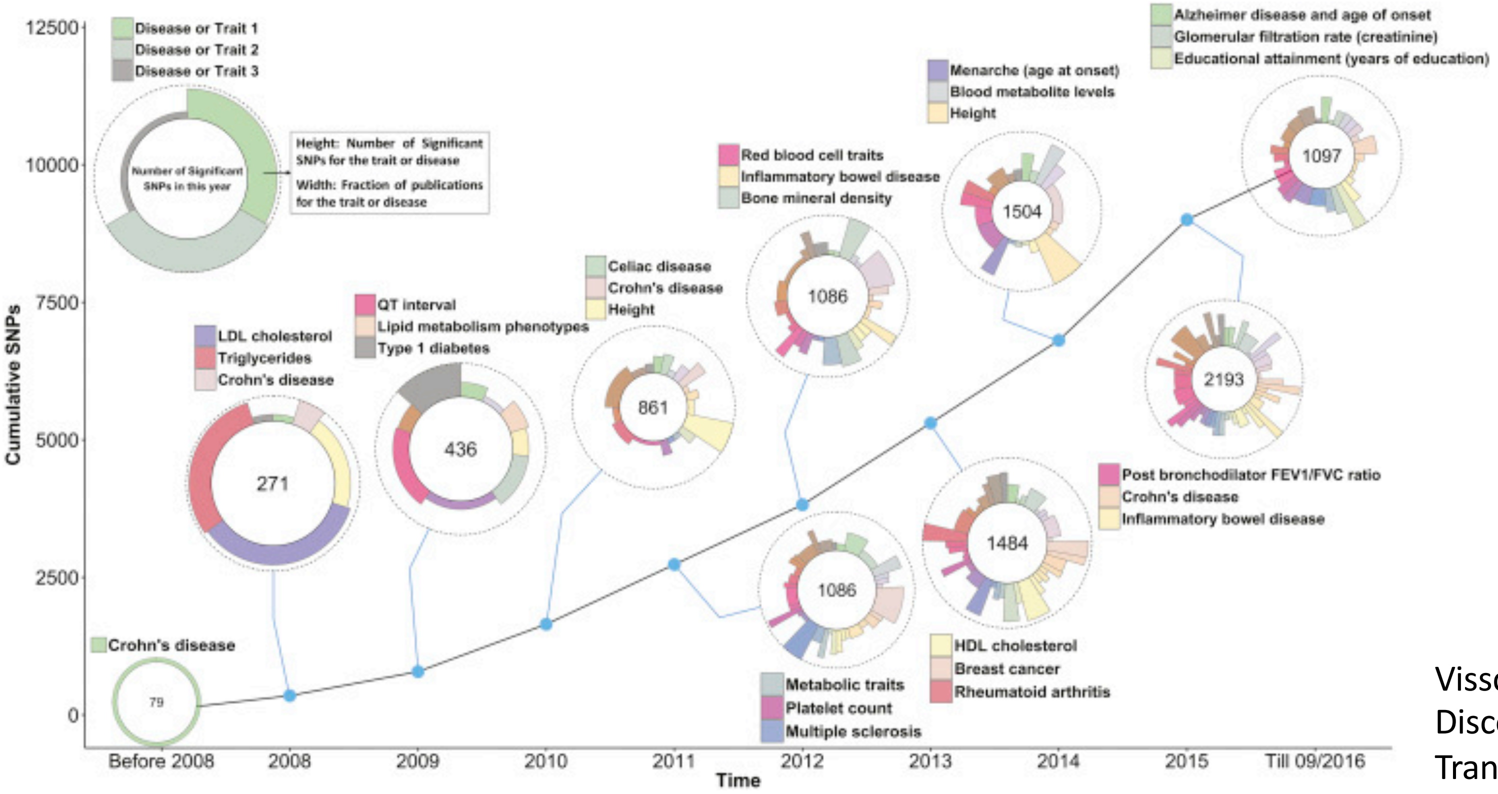
May 2018 ($p \leq 5X10^{-8}$)

- 69,000 trait associations
- >5000 studies
- 3378 publications

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

National Human Genome Research Institute

EMBL-EBI

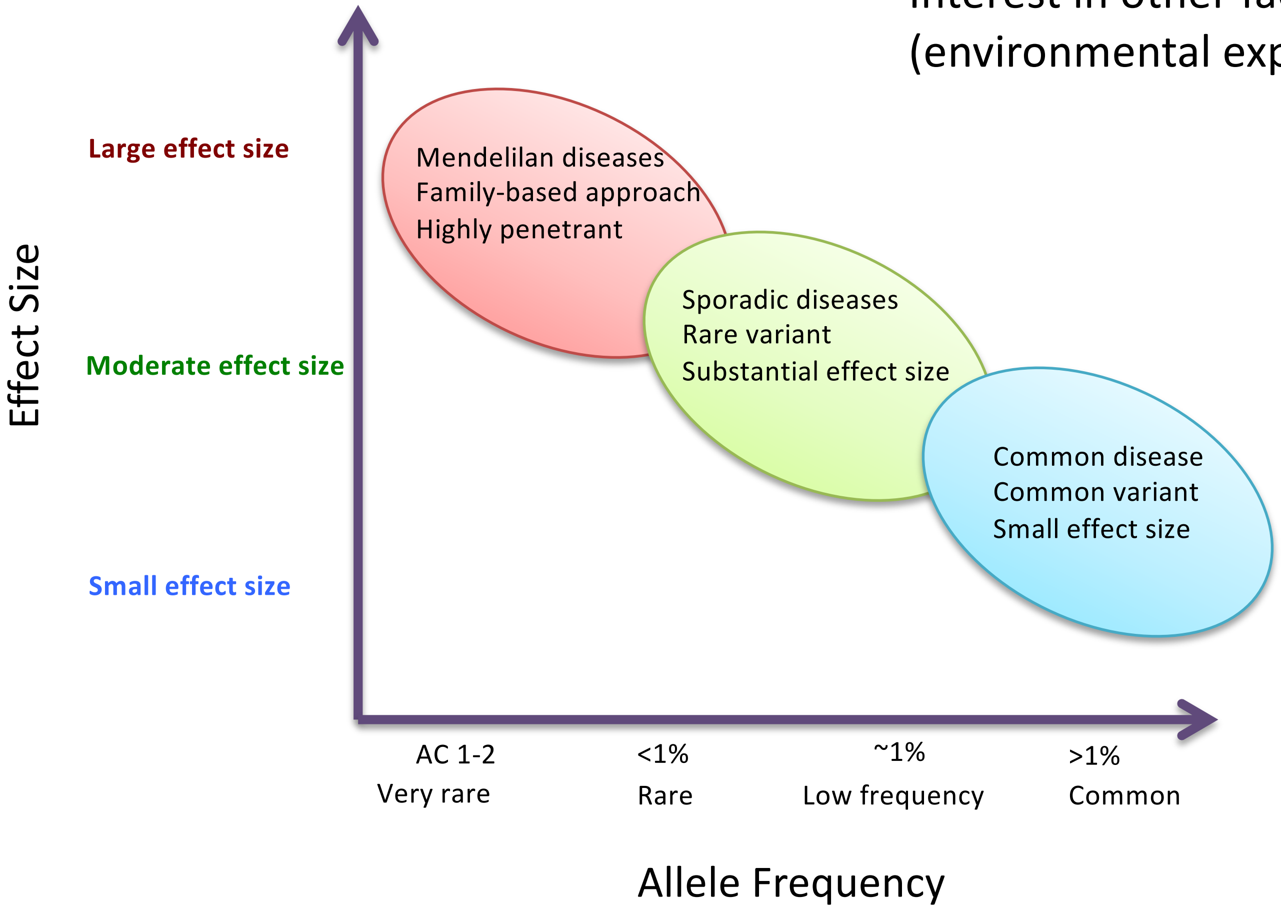


Visscher et al; 2017, 10 Years of GWAS Discovery: Biology, Function, and Translation

Variant Identification

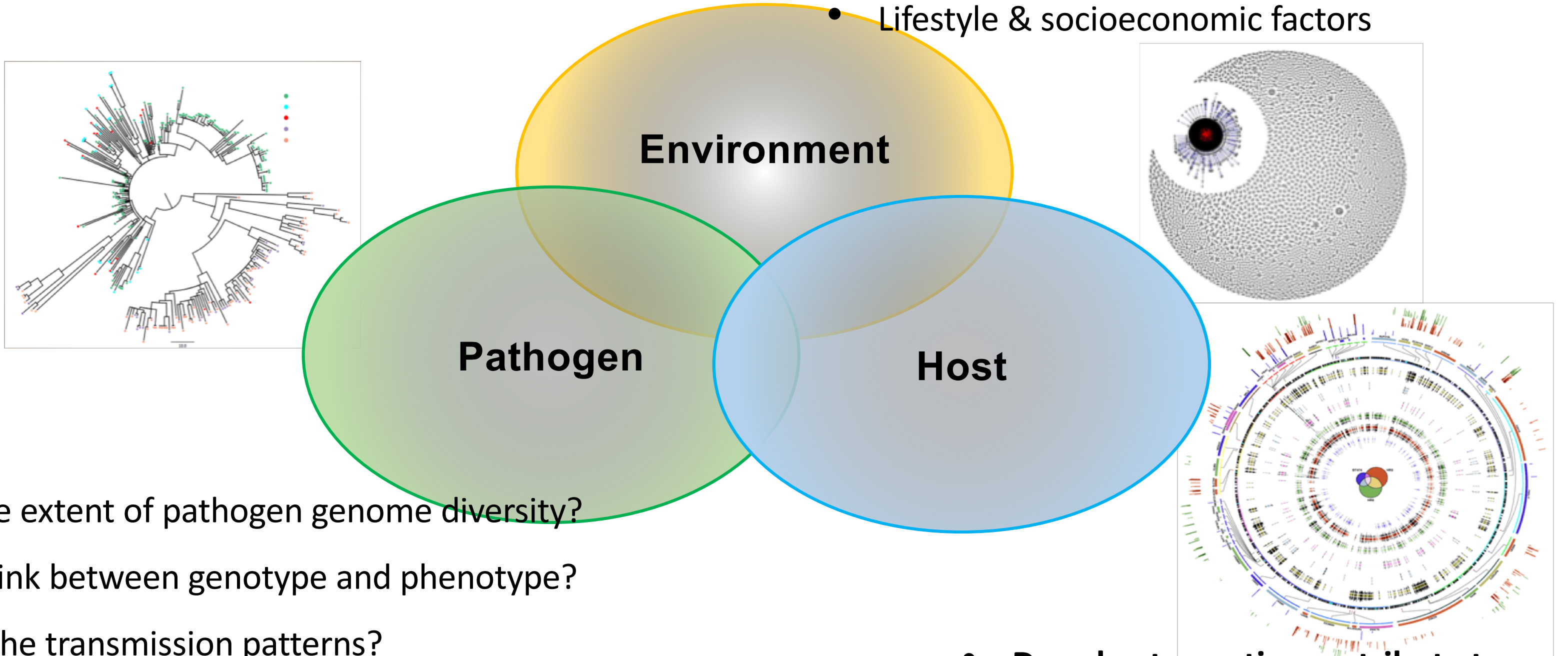
Study design is key

- Sample size?
- Interest in other factors of disease (environmental exposures, survival, effect size



Multifactorial determinants of pathogenesis & clinical outcome

- Co-infection
- Immunosuppression
- Lifestyle & socioeconomic factors



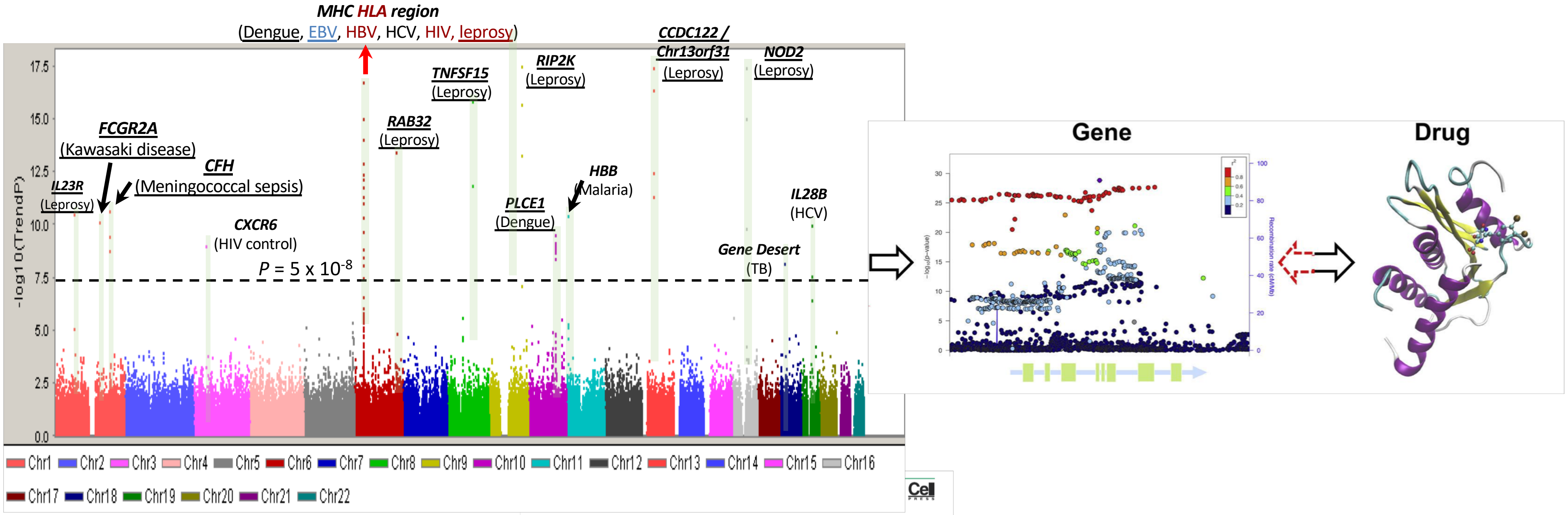
- What is the extent of pathogen genome diversity?
- Is there a link between genotype and phenotype?
- What are the transmission patterns?

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

GWAS of Infectious diseases

Phenotypes Studied:

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



Host–pathogen interactions revealed by human genome-wide surveys

Chiea Chuen Khor[1,2,3] and Martin L. Hibberd[1,2]

[1] Infectious Diseases, Genome Institute of Singapore, Singapore

GWAS Workflow

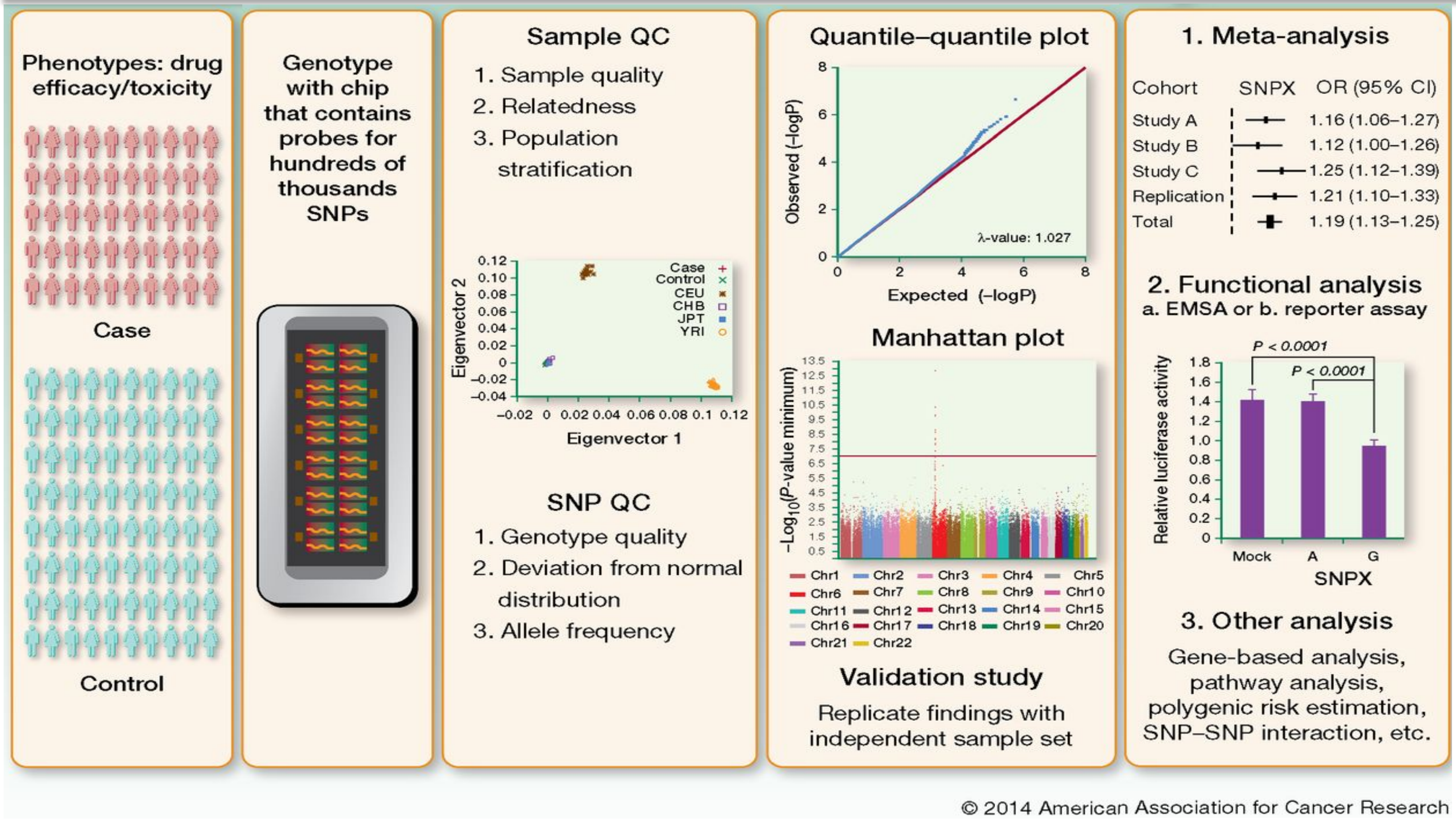
Sample Collection & Phenotype Determination

Genotyping or Whole genome sequencing

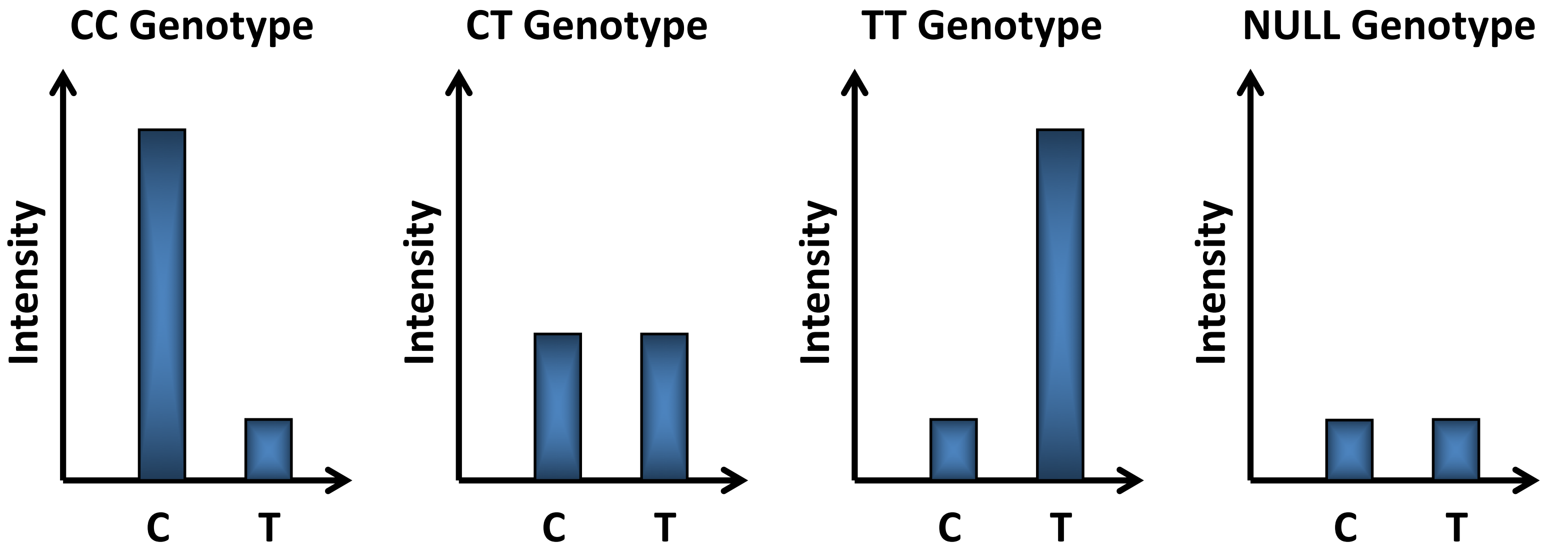
Quality Control

GWAS

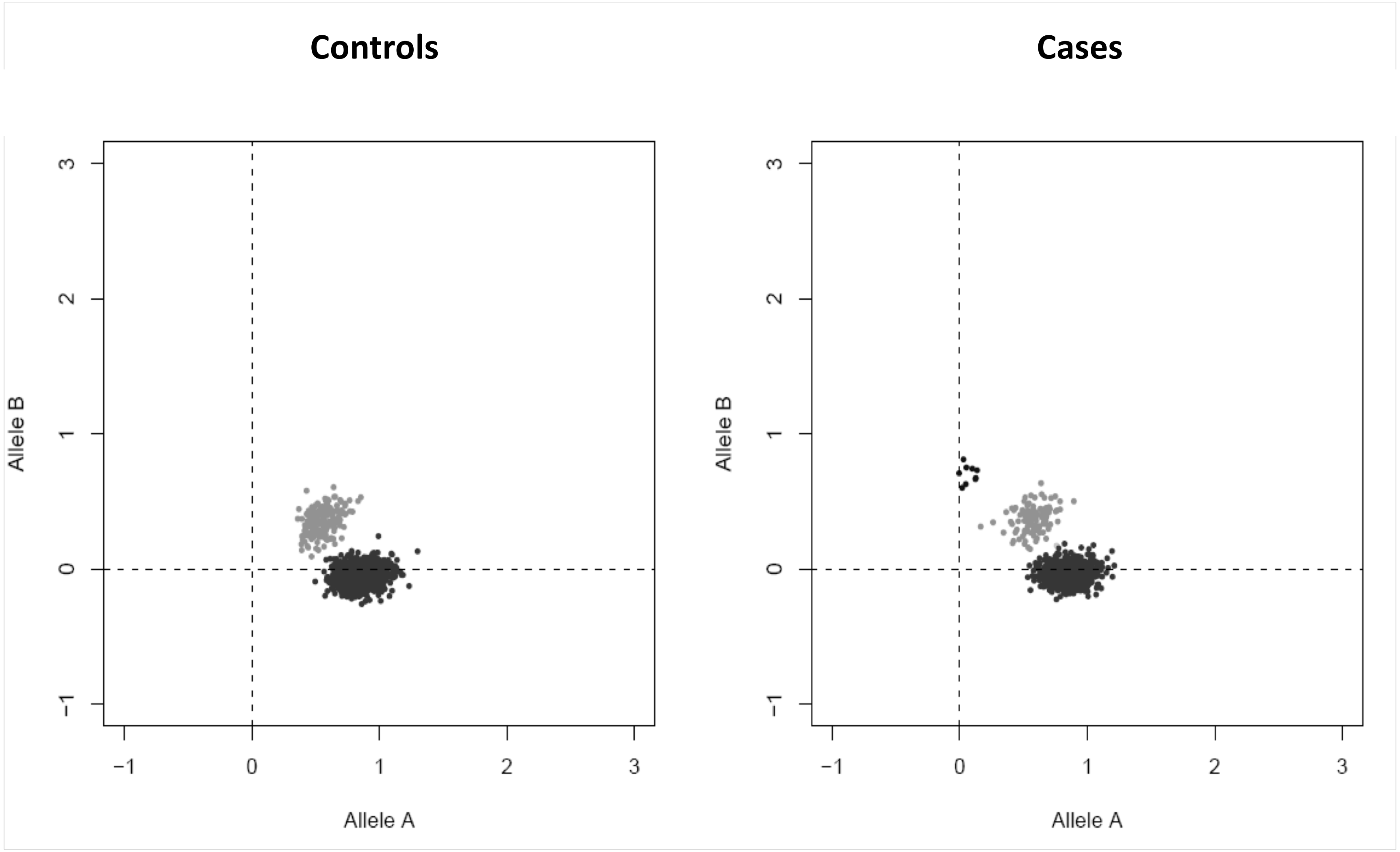
Post-GWAS



Raw data is not a genotypes, but Allelic hybridization



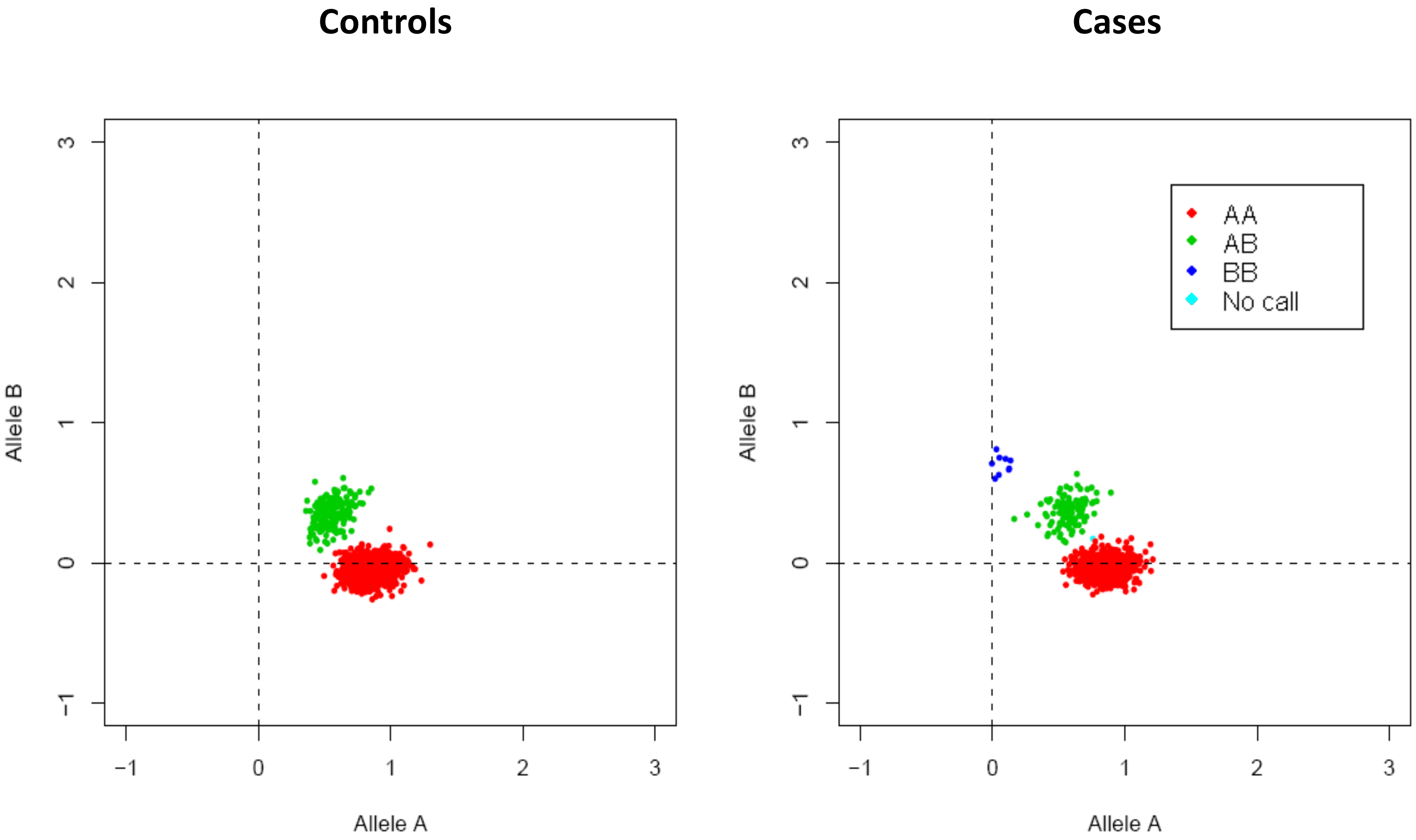
Genotype Calling



Controls

Cases

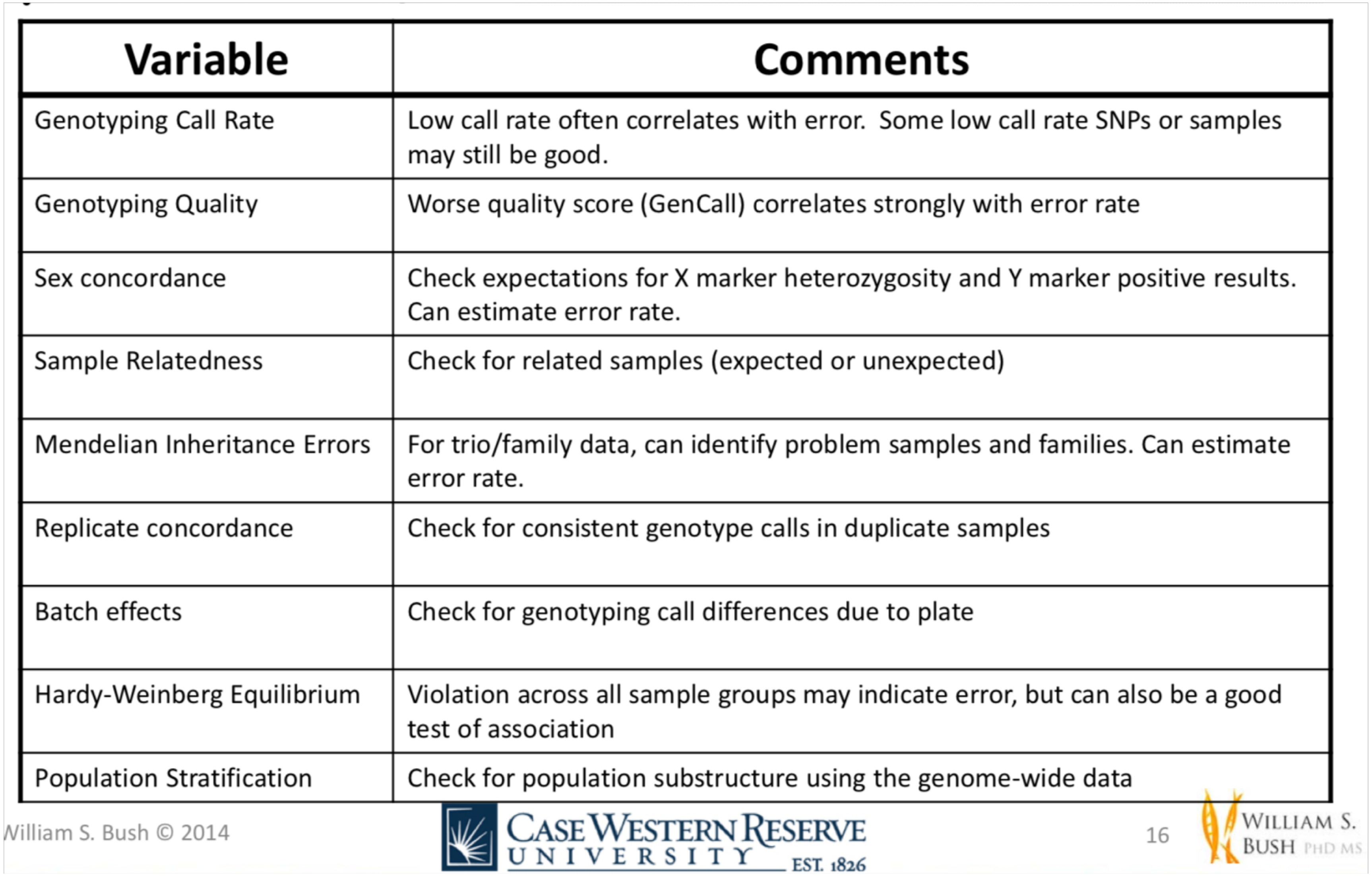
Genotype Calling



Need for high quality data

- Number of variants assayed ⇒ errors and genotype or sequence miscalls are bound to happen
- If problematic samples not identified and excluded, they can affect the results of the entire experiment
- If SNPs with erroneous genotyping or sequencing not identified and excluded, can produce false signals of associations
- QC samples and SNPs

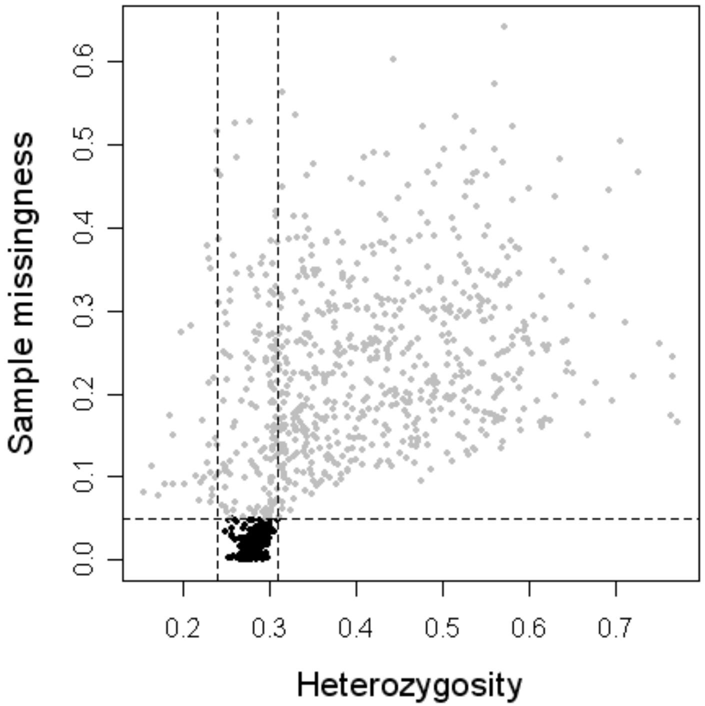
Quality Checks



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

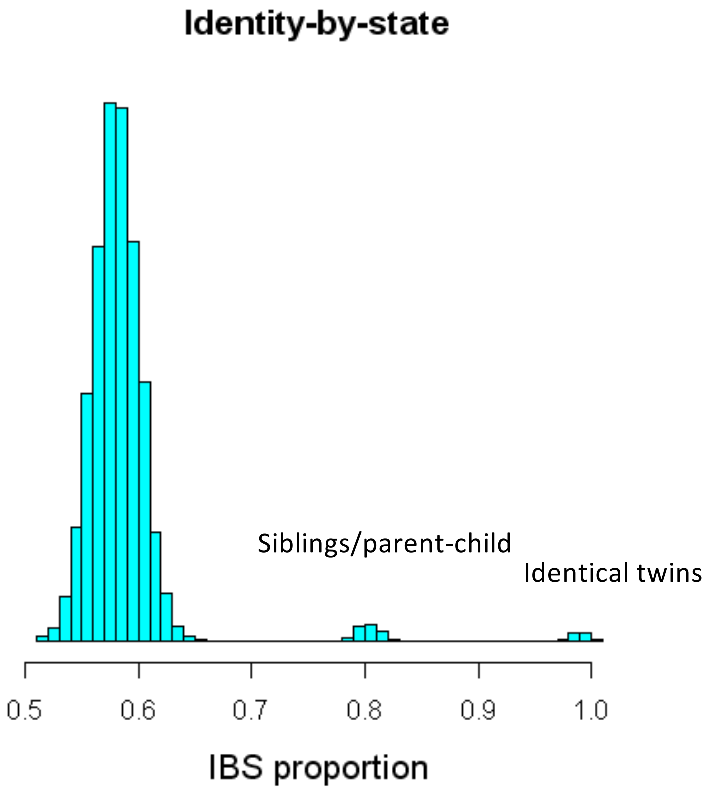
Sample QC

- Identify SNPs with high rates of missingness and heterozygosity
- Remove samples deviating from average
- Deviations could arise due to several reasons
  - Contamination of samples (high heterozygosity
  - Inbreeding (low heterozygosity)
  - Ancestral differences
  - Data quality / Poor genotype calling
- Heterozygotes more likely to be missing



Sample QC

- Identify related / duplicated samples
- Relatedness is a problem because of overrepresentation of selected alleles, which will bias any multivariate analysis (correlated data!); e.g. PCA or multivariate regression
  - Related samples need to be excluded or taken into account during subsequent analyses
- Related individuals will share more alleles IBS than expected by chance, with the degree of additional sharing proportional to the degree of relatedness.



Sample QC

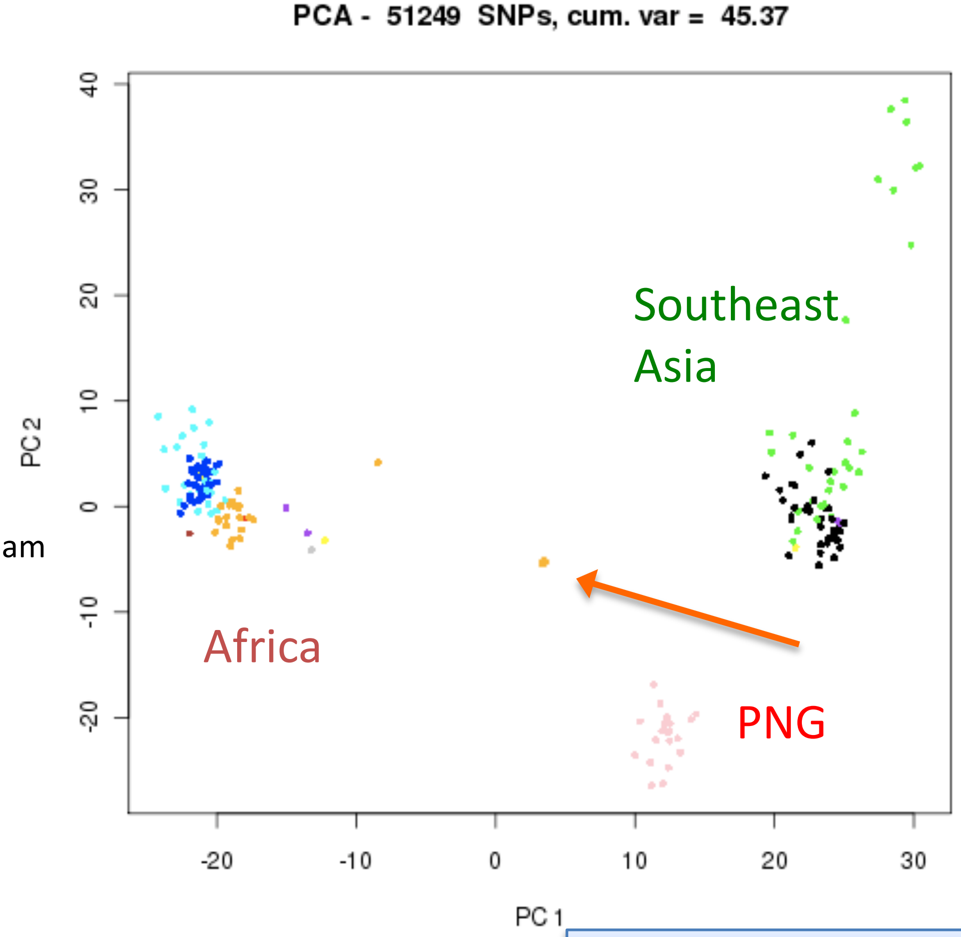
- Population substructure or stratification occurs when samples have different genetic ancestries

- Can lead to spurious associations due to differences in ancestry rather than true associations

- Imperative to check for population structure within samples

- Identify samples that are outliers
  - Can control for structure if identified, in downstream analysis

P. falciparum ~50k SNPs



Data from Ocholla et al. 2014

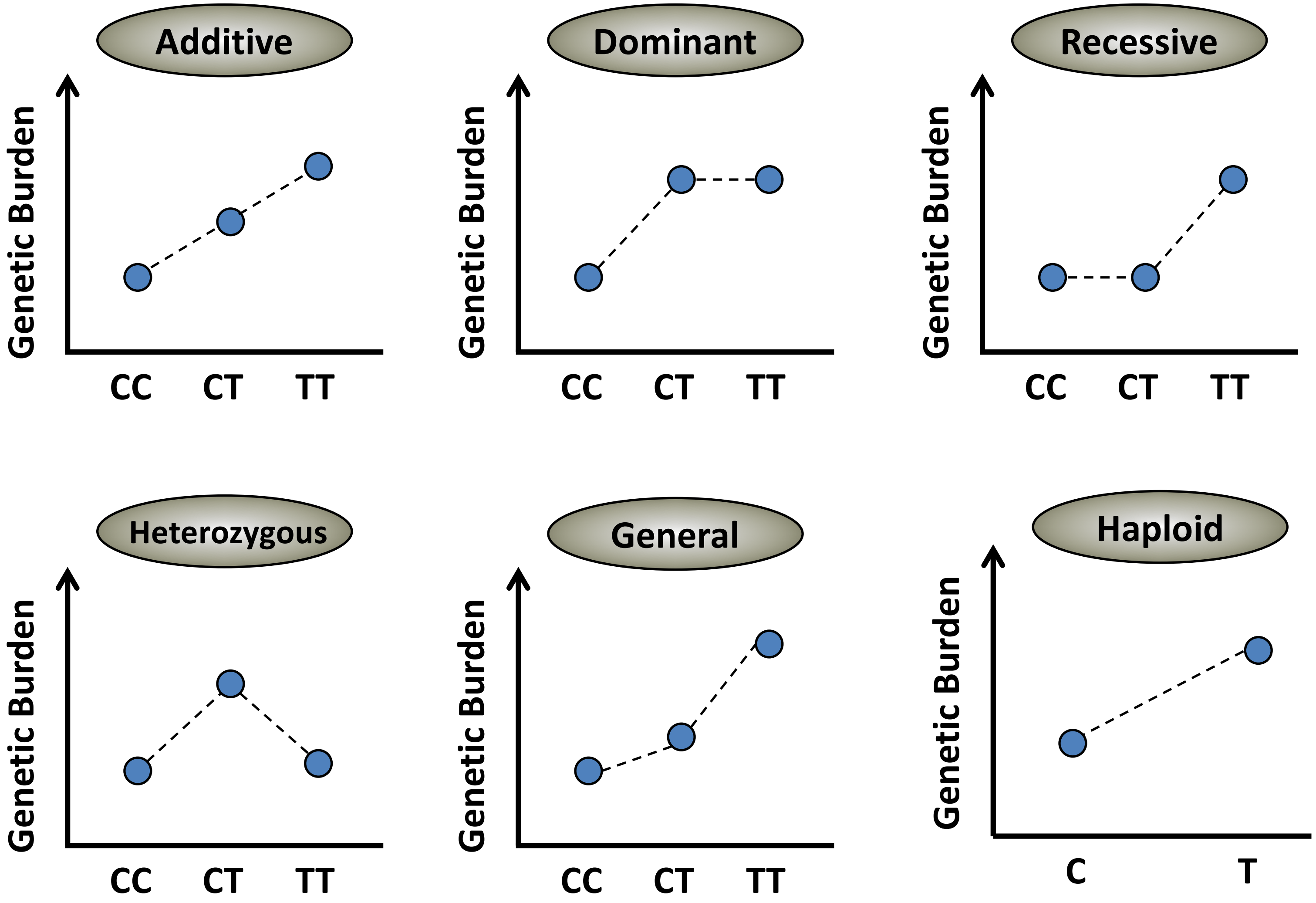
Testing for associations using regression models

Outcome	Example	Model
Continuous	IC50 levels	Linear regression
Binary	Malaria status	Logistic regression

- Using a single SNP in turn, but also can include
  - Interactions
  - Adjustment for confounders
  - Model building and risk prediction strategies
  - Diagnostic tools to assess model fit

- Also non-parametric approaches

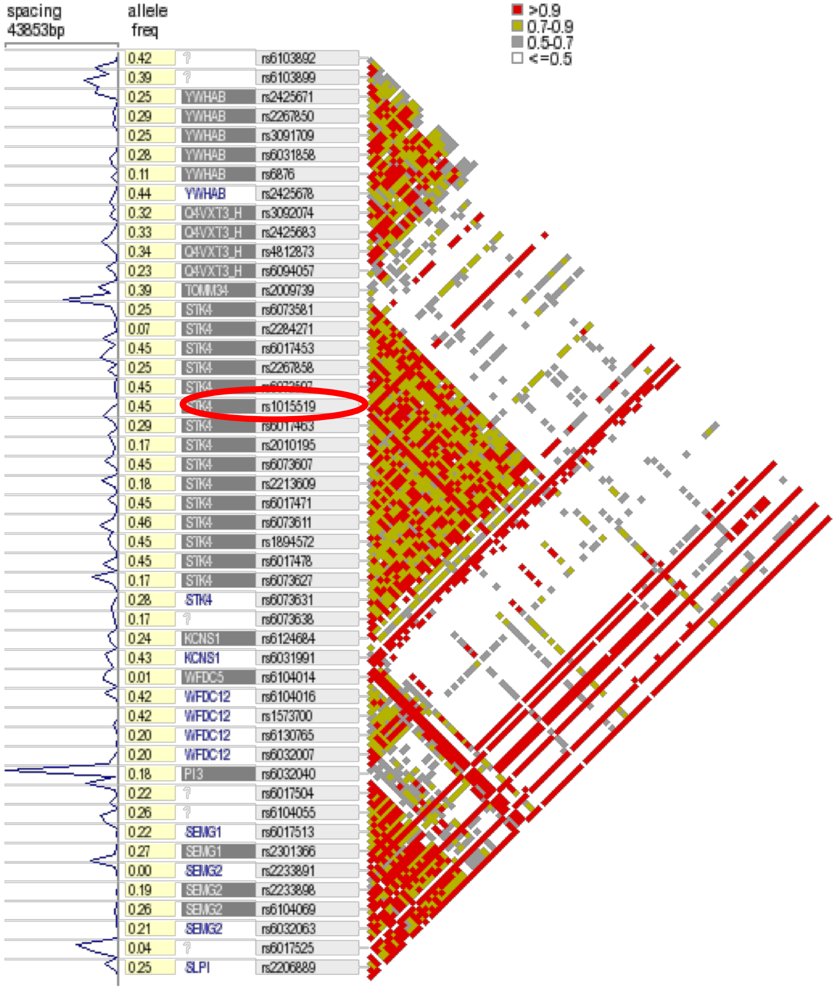
Genetic models tested in a regression framework



Association Studies

Direct Association

Tests the genetic variant directly responsible for causing the disease.



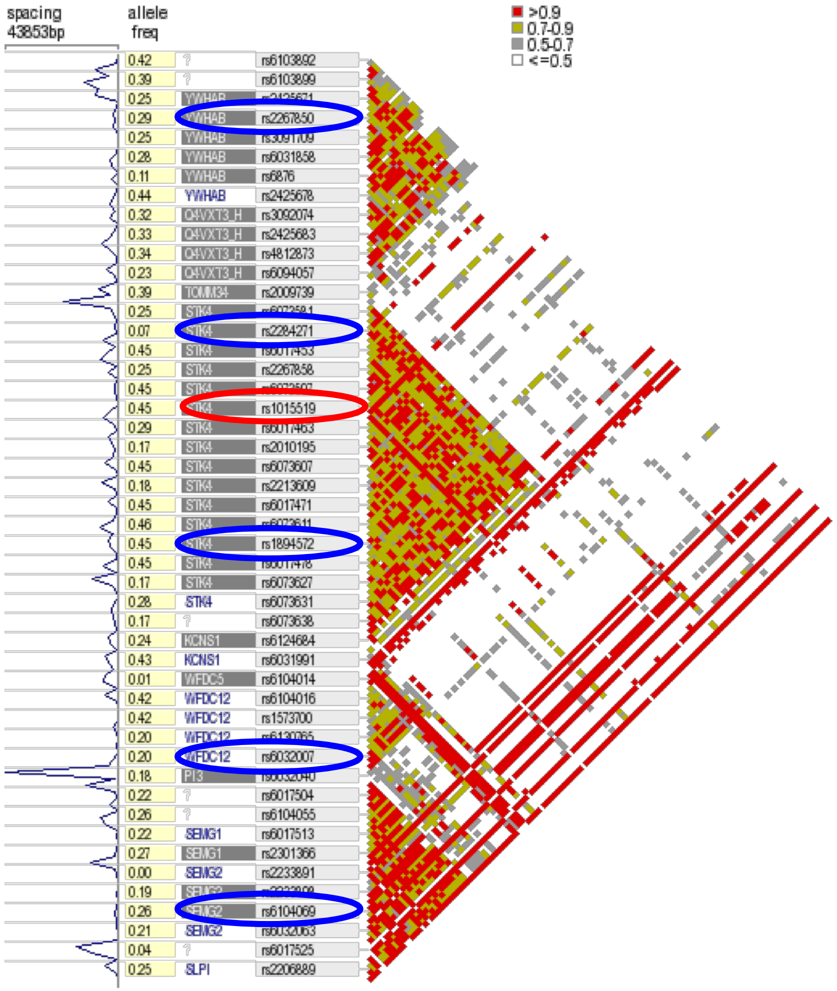
Association Studies

Direct Association

Tests the genetic variant directly responsible for causing the disease.

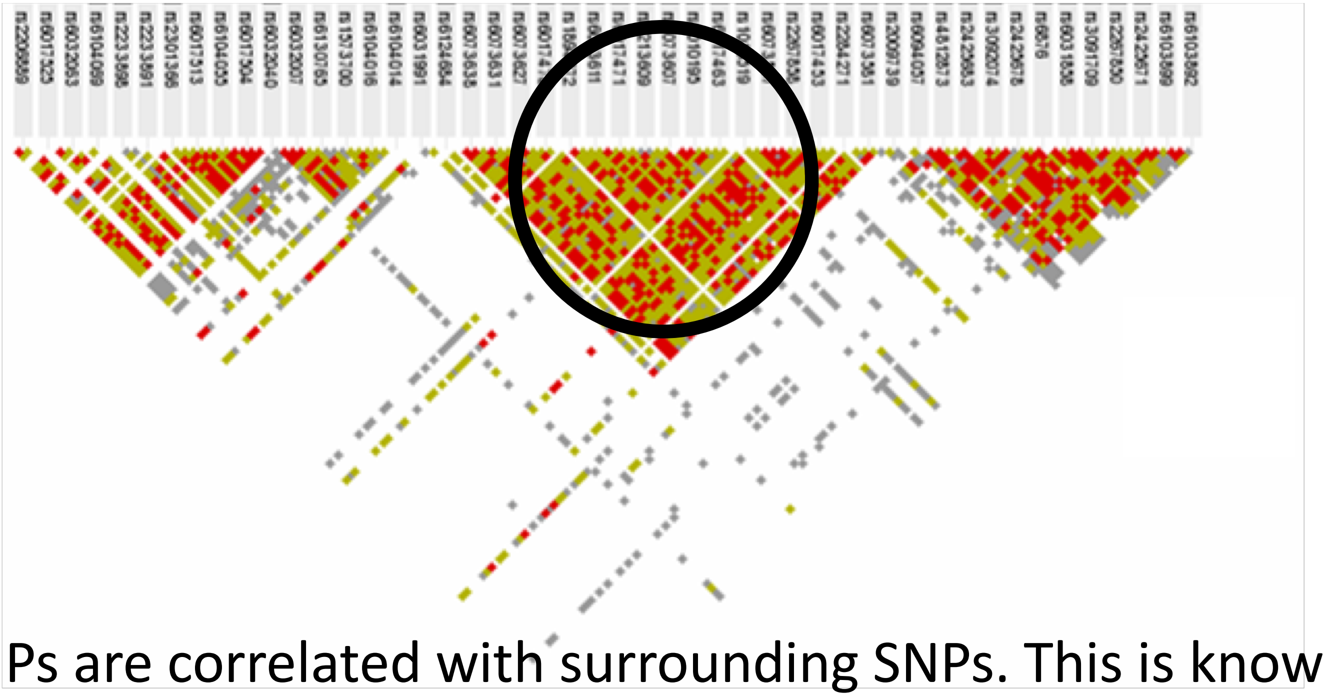
Indirect Association

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

SNPs



> 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

..ACTC**G**ACGATTTACG**G**TACTTAGGAGCATA**T**GCTAC..

..ACTC**T**ACGATTTACG**G**TACTTAGGAGCATA**T**GCTAC..

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

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

..ACTG**T**ACGATTTACG**G**TACTTAGGAGCATA**T**GCTAC..

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

..ACTG**G**ACGATTTACG**G**TACTTAGGAGCATA**G**GCTAC..

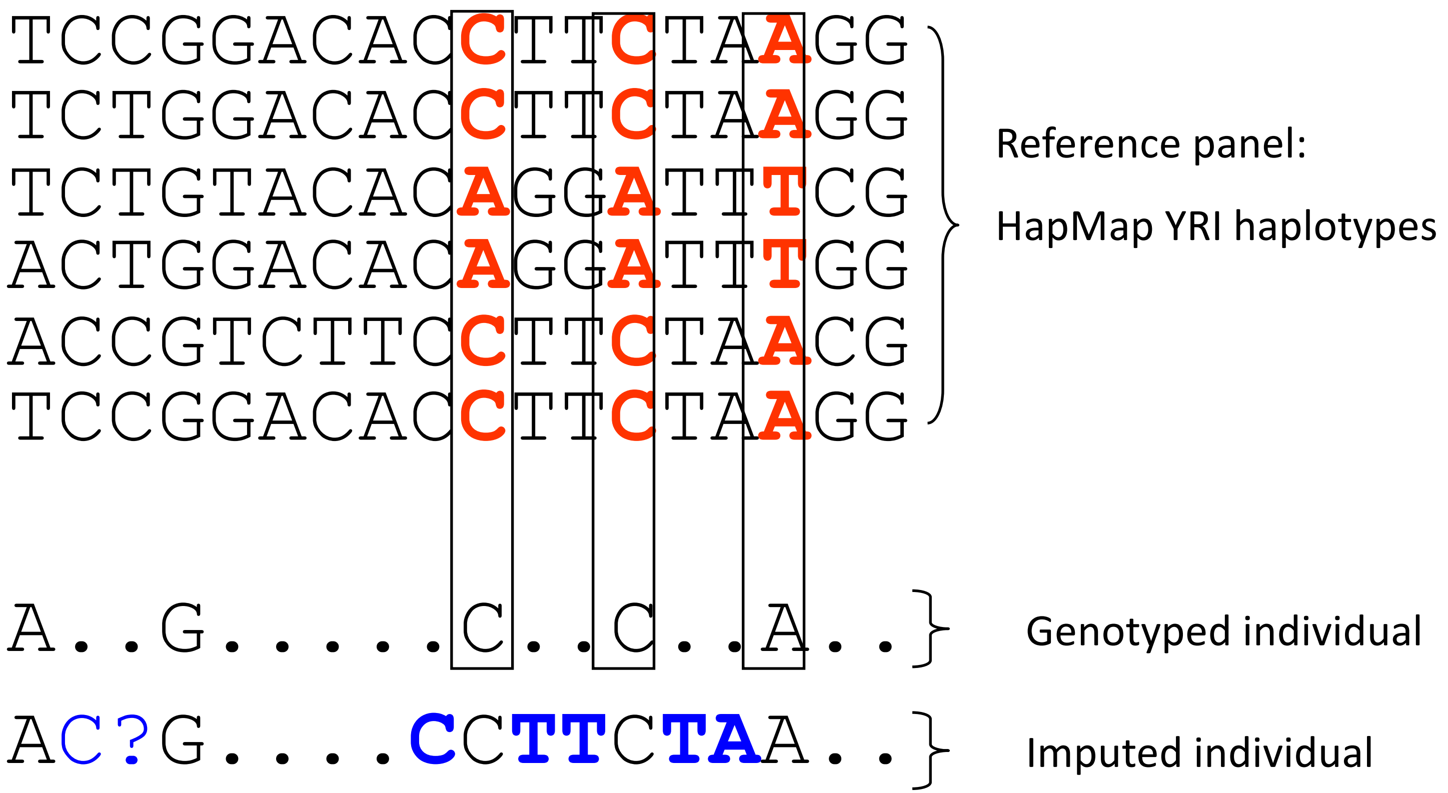
..ACTG**T**ACGATTTACG**G**TACTTAGGAGCATA**T**GCTAC..

- **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?

- ~10m 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

Imputation



TCCGGACACCTTCTAAGG
TCTGGACACCTTCTAAGG
TCTGTACACAGGATTTCG
ACTGGACACAGGATTTGG
ACCGTCTTCCTTCTAACG
TCCGGACACCTTCTAAGG

Reference panel:

HapMap YRI haplotypes

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

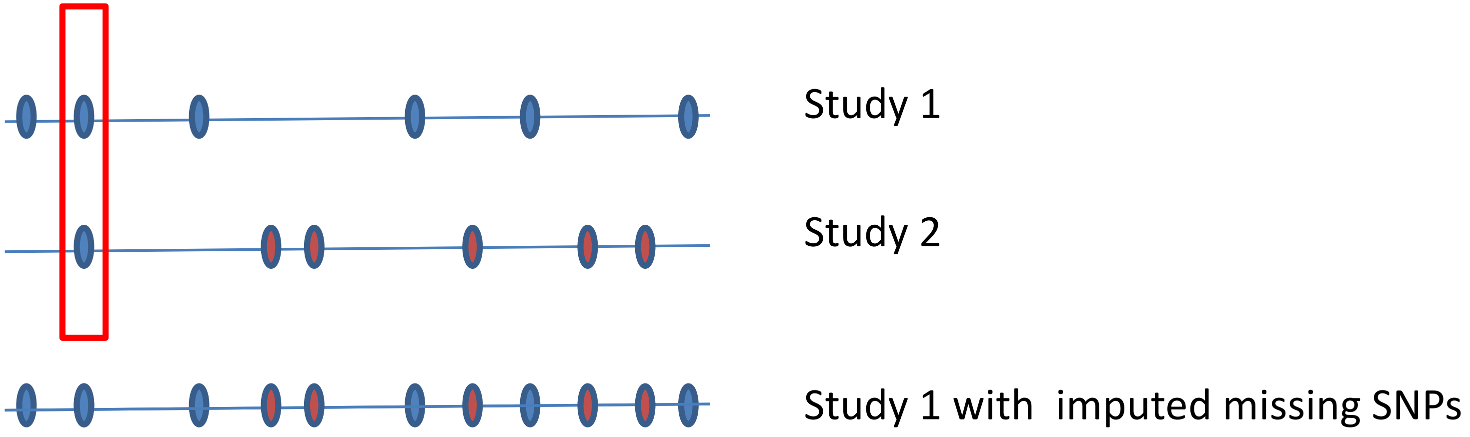
Genotyped individual

AC?G....CCTTCTAA..

Imputed individual

Using correlations or ‘recurring patterns’ in the data to fill in the blanks.

Imputation



- Imputation
  - Requires GWAS genotypes to be used as scaffold
  - Requires reference datasets (e.g. www.hapmap.org; www.1000genomes.org) where the LD (correlation) between SNPs is known and allows imputation of genotypes for variants not typed on a given array. Increasingly these could include reference datasets generated by whole-genome sequencing of subsets of individuals from the populations included in the study
  - There is specialist software to facilitate imputation as well as meta-analysis

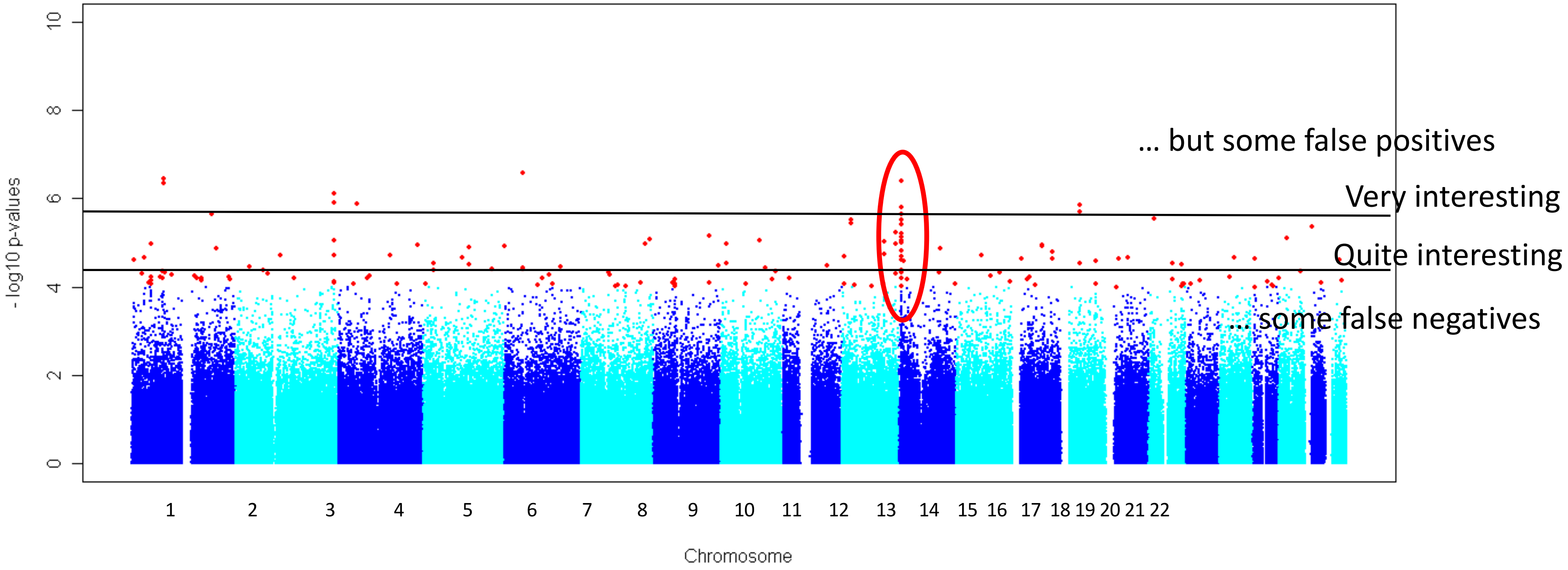
Why impute?

- To predict missing genotypes that haven't 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.

What if the LD structure in the imputed population is different to the reference?

Association signals across the genome

Manhattan plot: Severe malaria GWAS (n=1,000 cases, 1,500 controls)



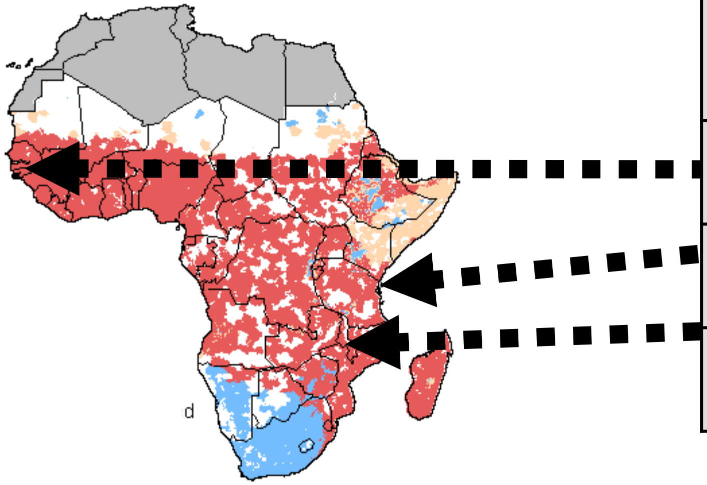
… but some false positives

Very interesting

Quite interesting

… some false negatives

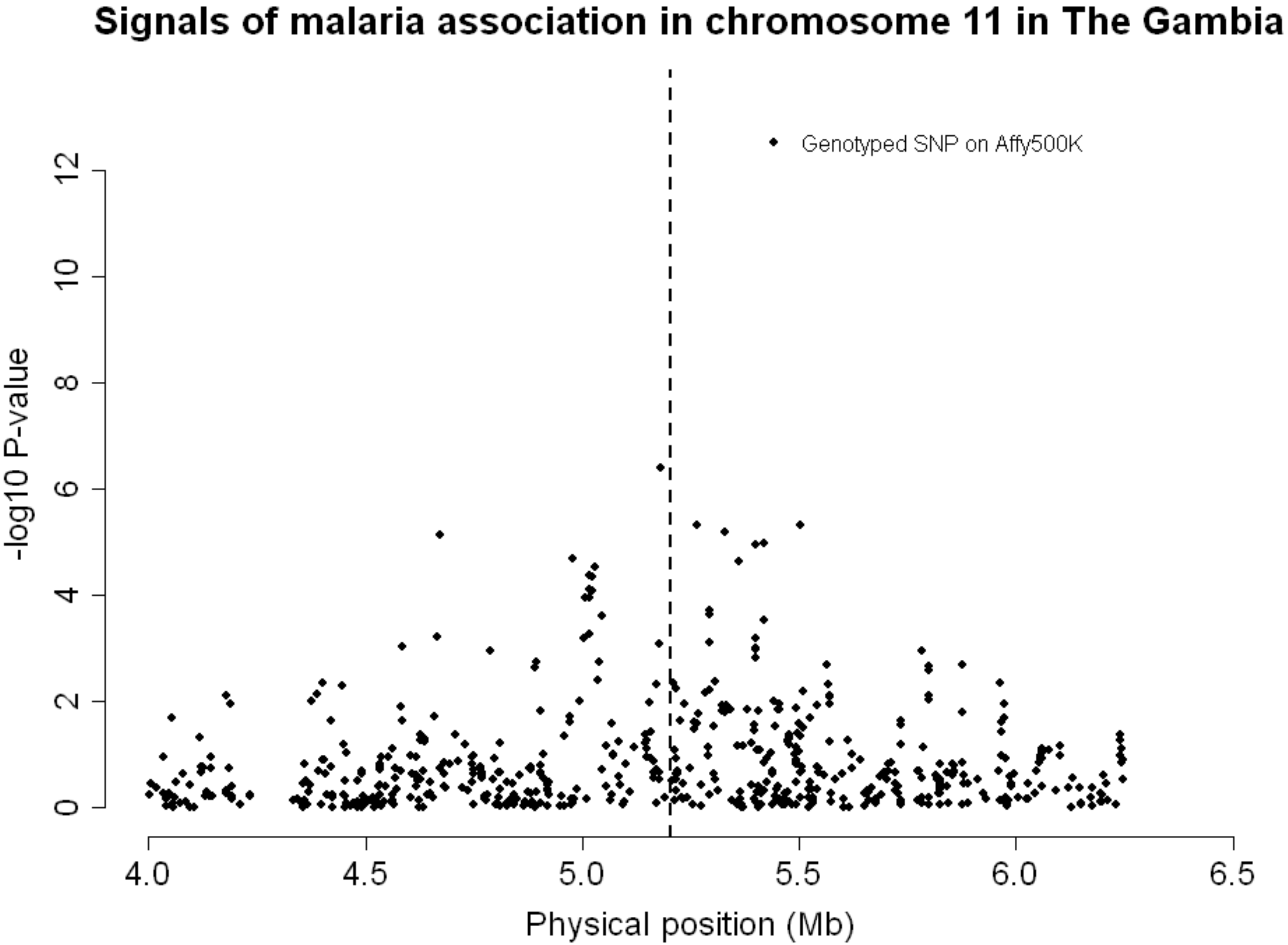
Sickle trait is the strongest known determinant of severe malaria risk

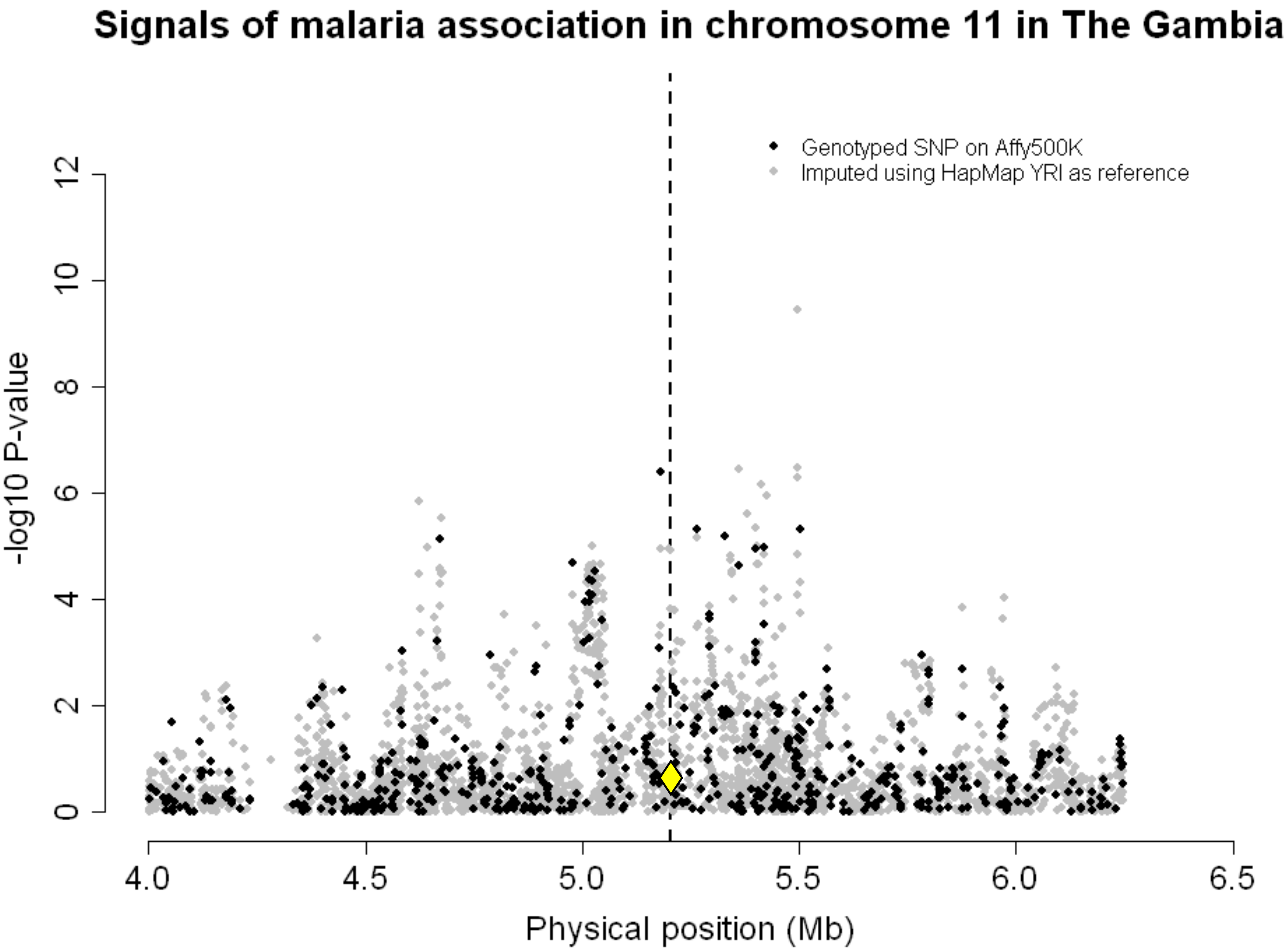


	Relative risk of severe malaria in children with HbS AS genotype
Gambia	0.11
Kenya	0.17
Malawi	0.08

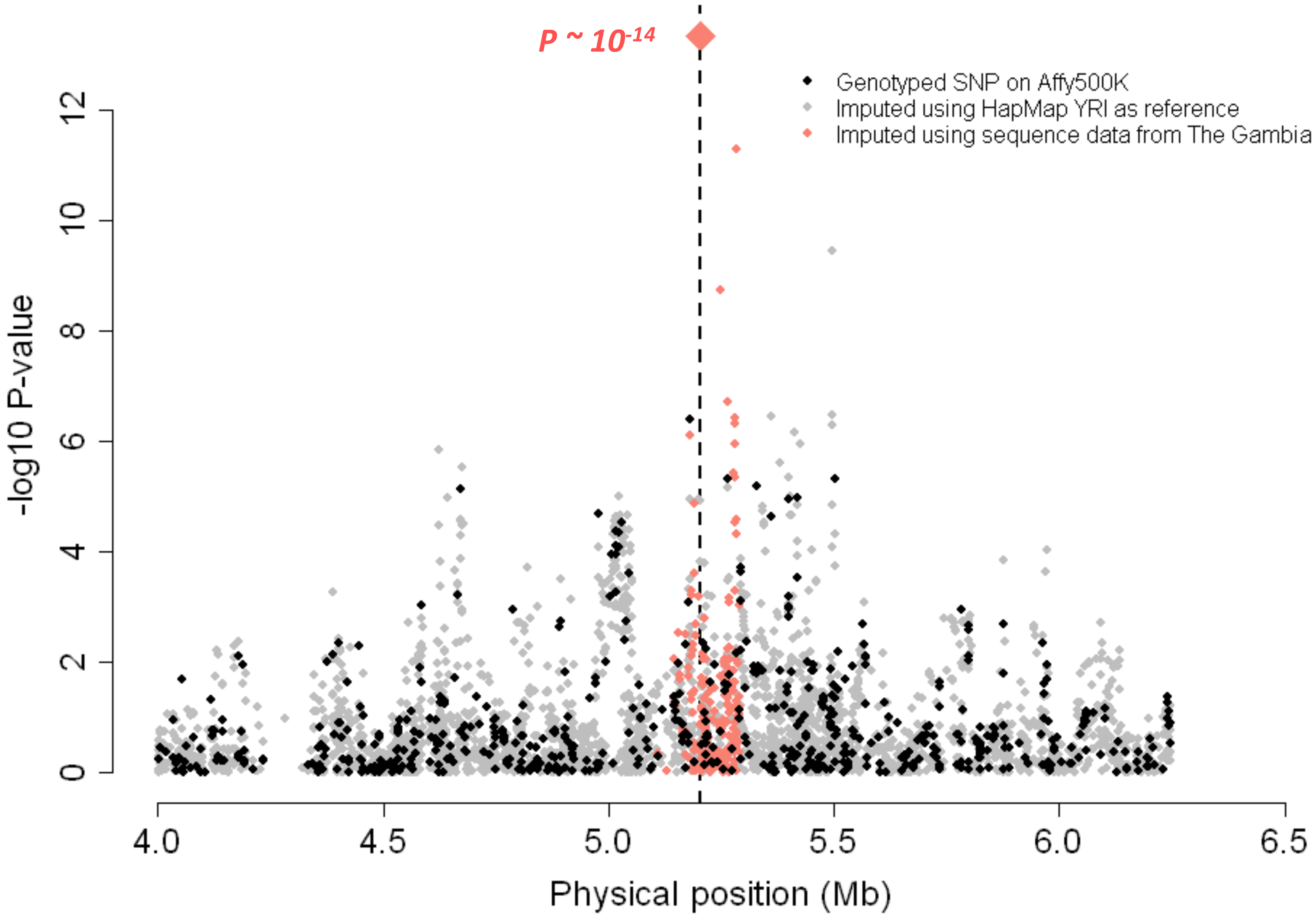
$P = 2 \times 10^{-31}$ (n = 3630)

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


Signals of malaria association in chromosome 11 in The Gambia
● Genotyped SNP on Affy500K
-log10 P-value
0
2
4
6
8
10
12
Physical position (Mb)
4.0
4.5
5.0
5.5
6.0
6.5


Signals of malaria association in chromosome 11 in The Gambia
● Genotyped SNP on Affy500K
● Imputed using HapMap YRI as reference
-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)

Signals of malaria association in chromosome 11 in The Gambia



Going beyond GWAS

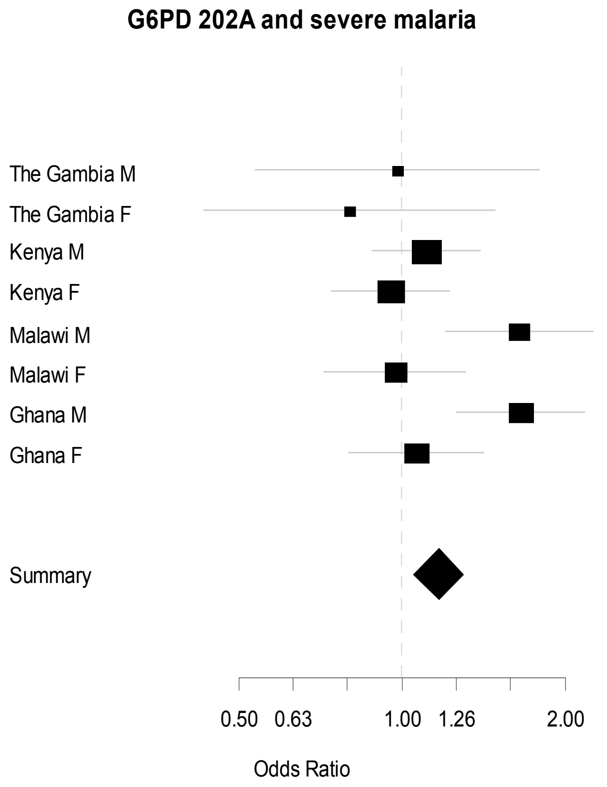
- Need to validate and confirm findings
  - Replication studies and meta analysis
  - In silico annotation tools
- If using genotyping arrays, fine-mapping the causal variant
  - Targeted-resequencing
  - Transethnic mapping
- Functional studies

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, given expectation of marginal genetic effects
- Genome-wide pooling, publication bias less of an issue
- Summary stats can be used for analysis



Benefits of GWAS Meta-Analysis

- Increased sample sizes for many disease and continuous trait consortia
  - increased power to detect new loci
  - new pathways and important biological insights gained
  - greater power to detect even smaller effect sizes and greater coverage of allele frequency spectrum
- Power of large collaborations/consortia
  - Design better powered replication and fine-mapping experiments

Heterogeneity

- Results from meta-analysis of various studies may suggest between study heterogeneity (e.g. especially when combining populations of different ancestry)

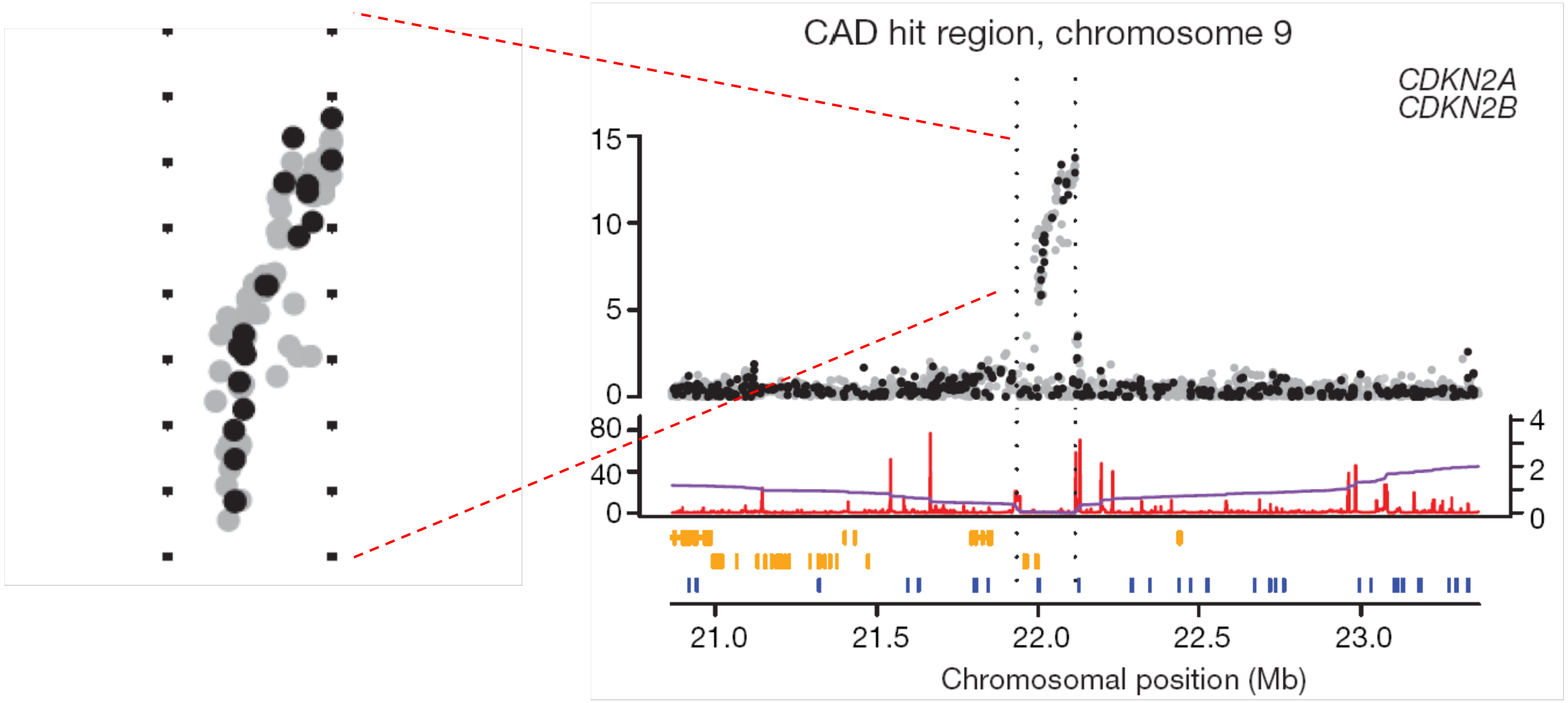
- How to interpret heterogeneity?
  - Differences in study design
  - Differences in population structure
  - Differences in environmental exposures
  - False-positive?

Need to study diverse populations

- Most GWAS have been done in populations of European ancestry

Hindrance of long LD

- Long LD is valuable at the stage of hunting for associations
- But long LD is a hindrance at fine-mapping – potentially lots of hits



- < LD in African populations lead to > difficulty to detect signals in initial scan, but easier to fine-map causal variants

Using GWAS To Study Infectious Disease Traits In Africa

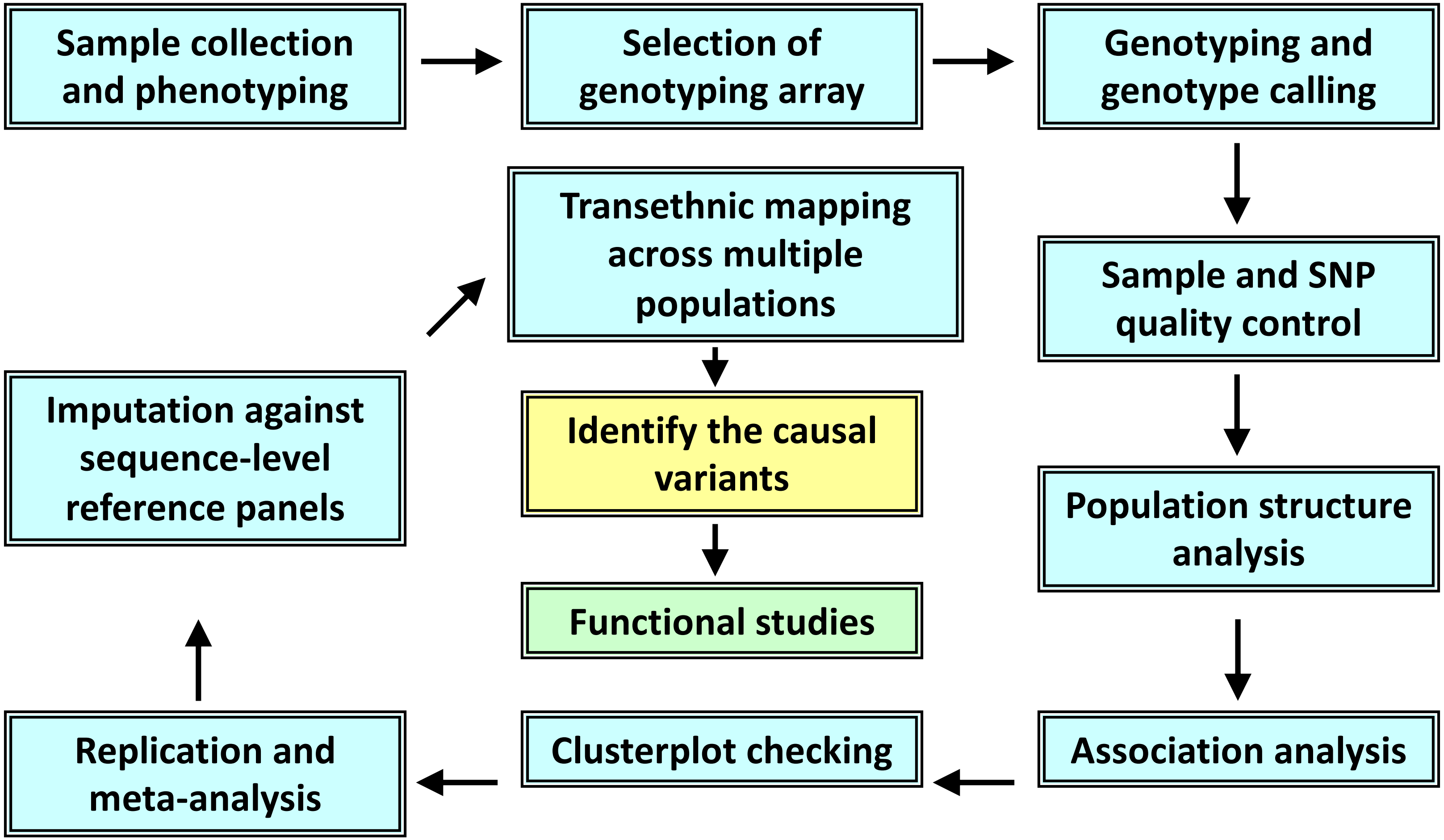
Benefits

- High prevalence of infection
- Identification of functionally relevant loci
- Fine mapping of causal variants

Challenges

- High genomic diversity
- Pathogen genetic variation
- Lack of African genetic data & resources

GWAS pipeline recap



Sample collection and phenotyping

Selection of genotyping array

Genotyping and genotype calling

Sample and SNP quality control

Population structure analysis

Association analysis

Clusterplot checking

Replication and meta-analysis

Imputation against sequence-level reference panels

Transethnic mapping across multiple populations

Identify the causal variants

Functional studies

Acknowledgements: YY Teo, S Campino, D Kwiatkowski, Wellcome Trust.

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

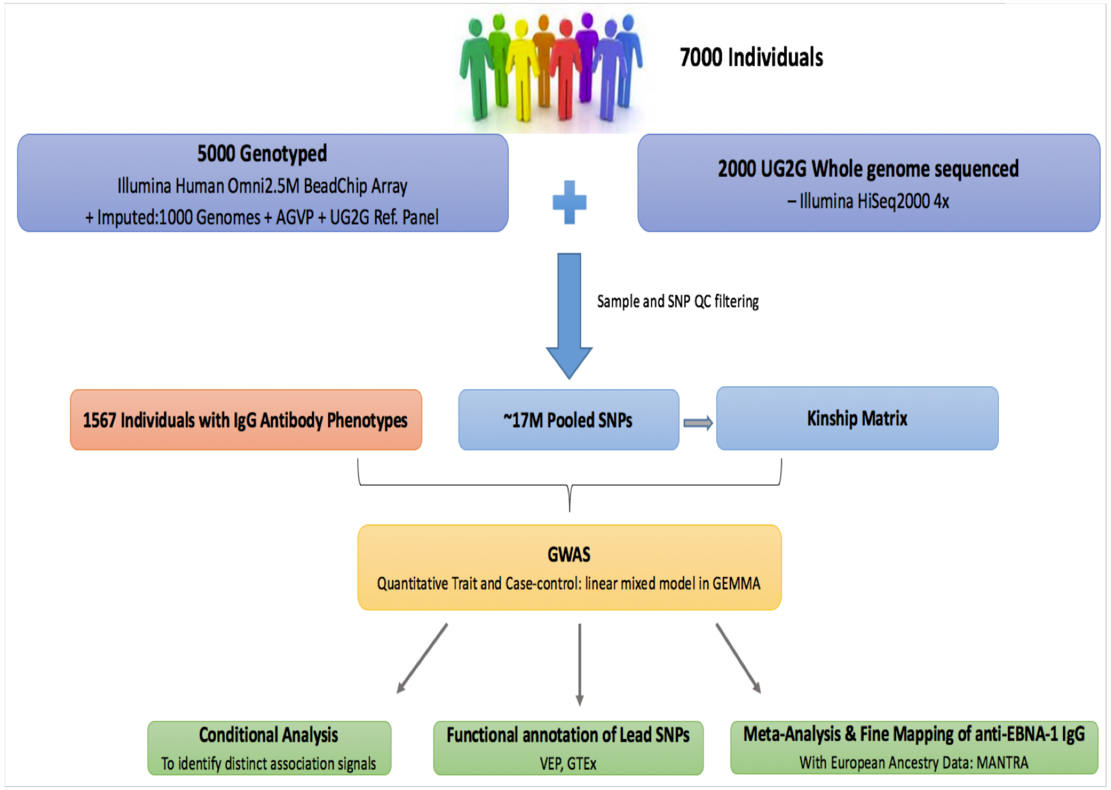
Case in Point - GWAS of EBV in an African population

Glob Health Epidemiol Genom. 2017 Nov 27;2:e18. doi: 10.1017/gheg.2017.16. eCollection 2017.

Whole-genome association study of antibody response to Epstein-Barr virus in an African population: a pilot.

Sallah N[1,2], Carstensen T[1,3], Wakeham K[4,5], Bagni R[6], Labo N[7], Pollard MO[1,3], Gurdasani D[1,3], Ekoru K[1,3], Pomilla C[1,3], Young EH[1,3], Fatumo S[1,3,8], Asiki G[4], Kamali A[4], Sandhu M[1,3], Kellam P[2], Whitby D[7], Barroso I[1], Newton R[4].

Genome-wide association workflow for EBV serological traits in the Uganda GPC



7000 Individuals

5000 Genotyped
Illumina Human Omni2.5M BeadChip Array
+ Imputed:1000 Genomes + AGVP + UG2G Ref. Panel

+

2000 UG2G Whole genome sequenced
– Illumina HiSeq2000 4x

Sample and SNP QC filtering

1567 Individuals with IgG Antibody Phenotypes

~17M Pooled SNPs

Kinship Matrix

GWAS
Quantitative Trait and Case-control: linear mixed model in GEMMA

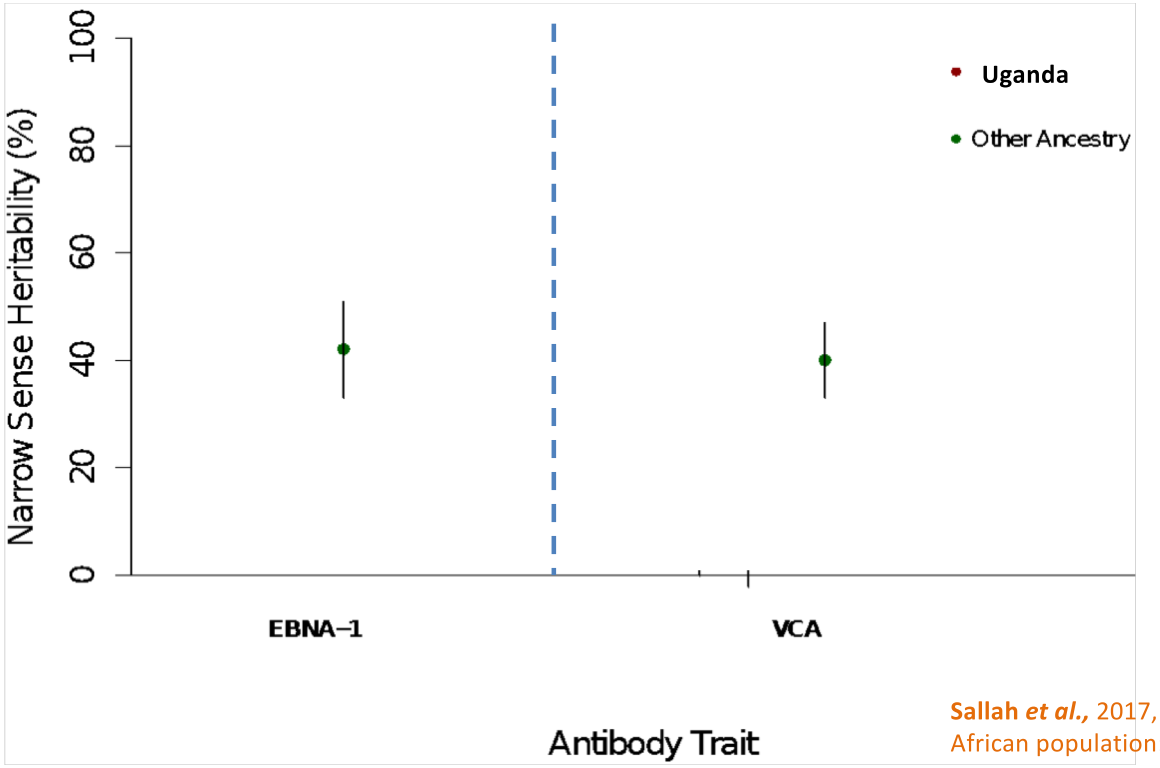
Conditional Analysis
To identify distinct association signals

Functional annotation of Lead SNPs
VEP, GTEx

Meta-Analysis & Fine Mapping of anti-EBNA-1 IgG
With European Ancestry Data: MANTRA

Heritability of EBV IgG antibody response traits

- **Proportion of variation in antibody responses due to host genetics**
- **h^2 based on genotype data using FaST-LMM (Heckerman, et al. 2016)**
- **Adjustment for environmental correlation using GPS data**

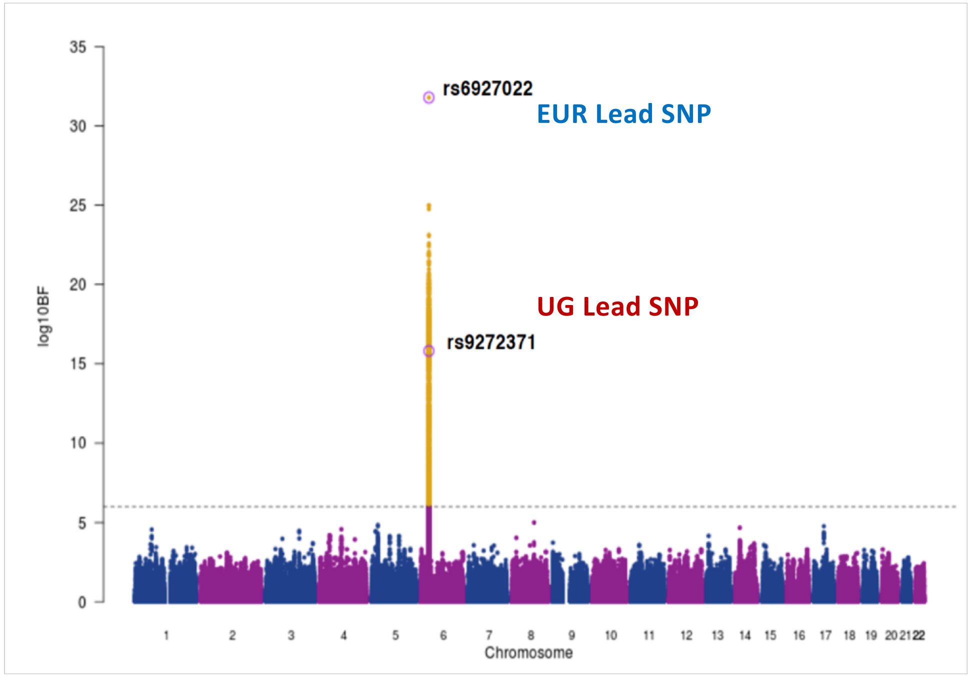


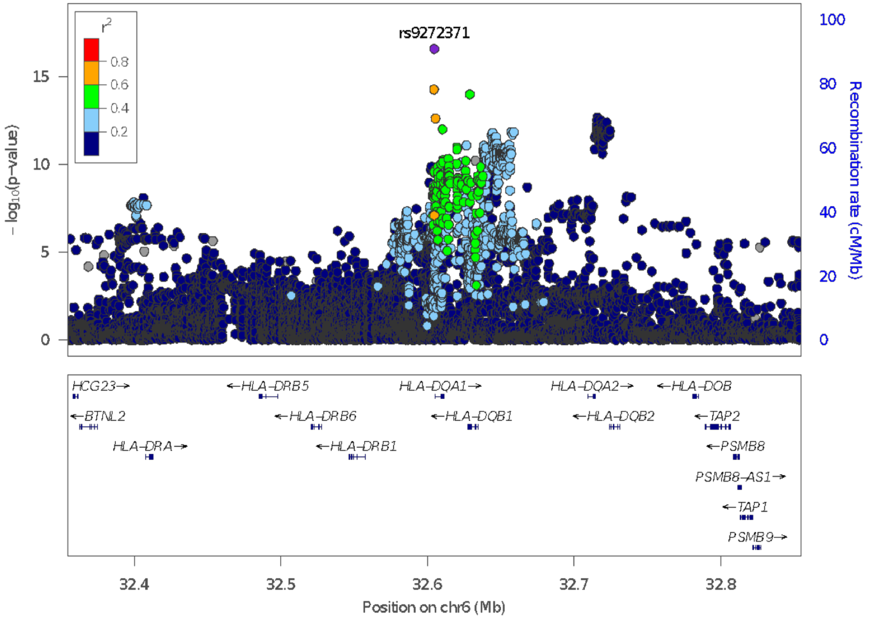
Lower h^2 estimates in Ugandan population

- Differences in in environmental variation
- Differences in gene-environment interactions
- Differences in variants or allele frequencies/effect sizes contributing to phenotypic variation

Sallah *et al.*, 2017, Whole-genome association study of antibody response to Epstein-Barr virus in an African population: A pilot. *Global Health, Epidemiology and Genomics, 2*. doi:10.1017/gheg.2017.16

Distinct association signals in the *HLA* class II region for anti-EBNA-1 IgG response

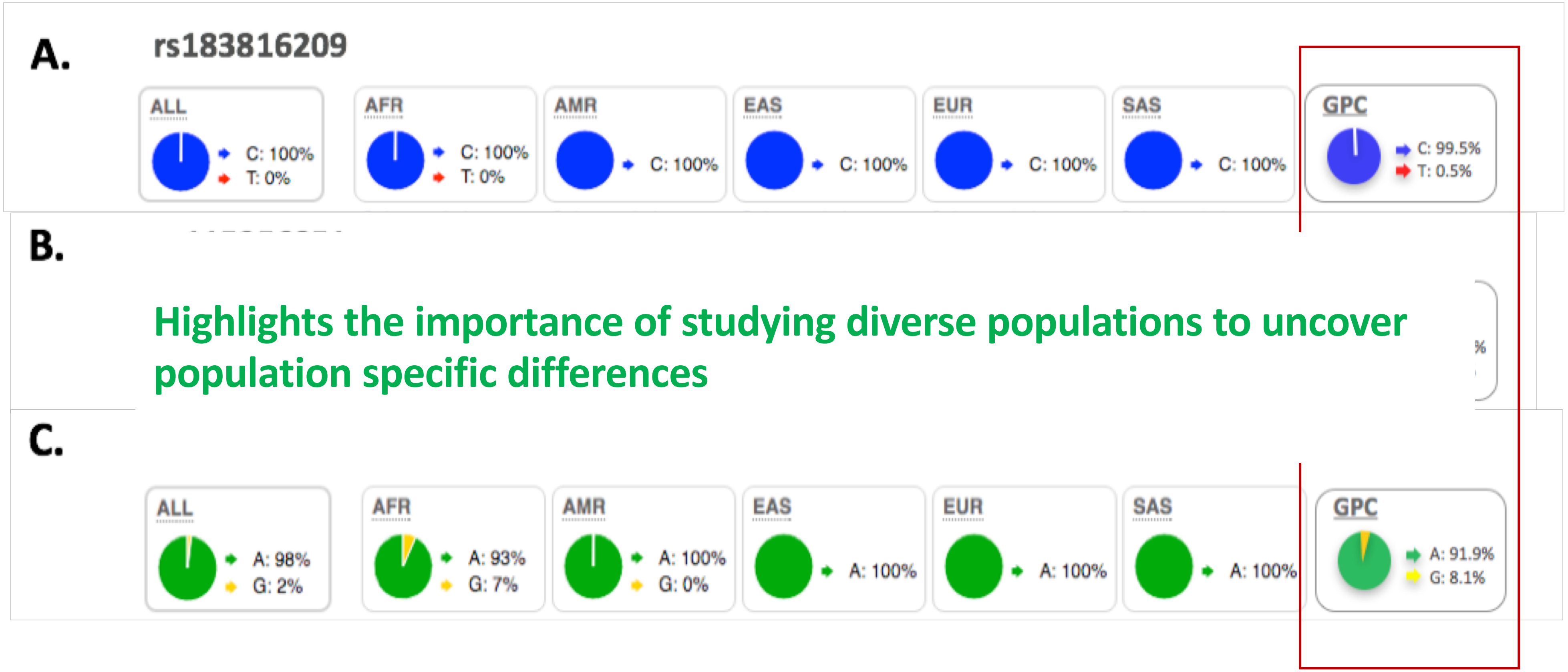




Further analysis shows single signal in Eu and 2 signals in African population

Sallah *et al.*, 2017, Whole-genome association study of antibody response to Epstein-Barr virus in an African population: A pilot. *Global Health, Epidemiology and Genomics, 2*. doi:10.1017/gheg.2017.16

Other variants identified that are African-Specific



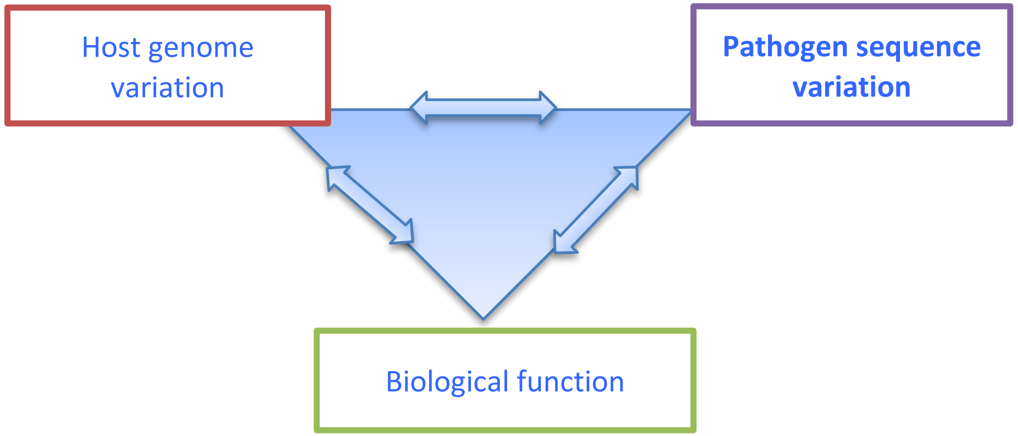
Highlights the importance of studying diverse populations to uncover population specific differences

Sallah *et al.*, 2017, Whole-genome association study of antibody response to Epstein-Barr virus in an African population: A pilot. *Global Health, Epidemiology and Genomics, 2*. doi:10.1017/gheg.2017.16

Future Perspectives

- More data & resources from Africa & other diverse populations needed to leverage GWAS findings to uncover meaningful biological insights

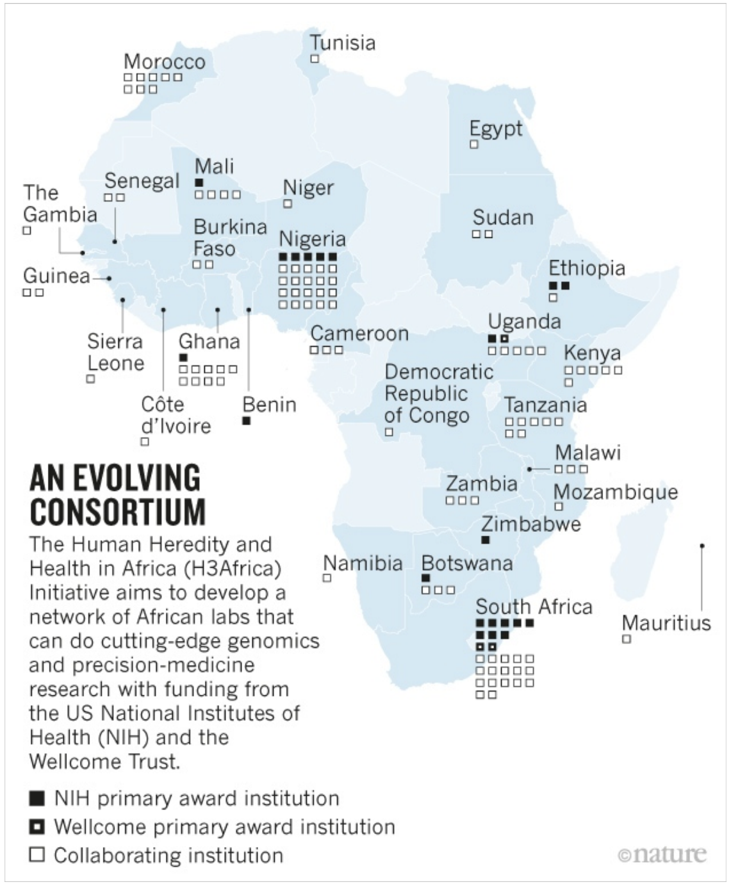
- Large cohorts allow comprehensive analysis of infection – with host and pathogen genomes isolated from the same individuals



Host genome variation

Pathogen sequence variation

Biological function



AN EVOLVING CONSORTIUM

The Human Heredity and Health in Africa (H3Africa) Initiative aims to develop a network of African labs that can do cutting-edge genomics and precision-medicine research with funding from the US National Institutes of Health (NIH) and the Wellcome Trust.

■ NIH primary award institution
▣ Wellcome primary award institution
□ Collaborating institution

Neneh Sallah