Genome-wide Association Studies

Part 1: Cleaning

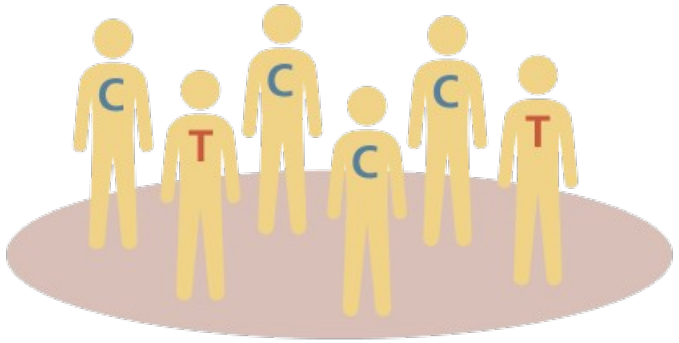
What is a GWAS?

- "***A genome-wide association study is defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as height, blood pressure or weight), or the presence or absence of a disease or condition (such as cancer, diabetes or schizophrenia).***"
- Overall goal: identify correlations between a phenotype and whole genome genotypes
- Some specific goals
    - Identify statistical connections between individual genetic variants (or regions) in the genome and the phenotype
    - Generate insights on genetic architecture of phenotype
    - Build statistical models to predict phenotype from genotype

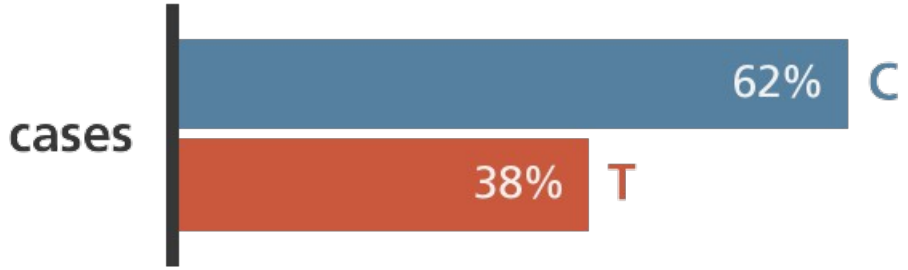
GWAS Methodology

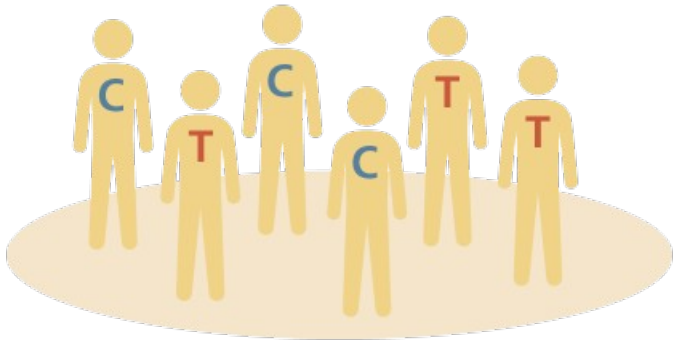
- Collect large samples of individuals (same size typically in the 10s or 100s of thousands)
- Measure each individuals for genetic variation throughout the genome
- Association testing – statistically associate genetic variants with the phenotype (simple regression models)

Single Genotype Example



cases (n=1,000)
people with heart disease

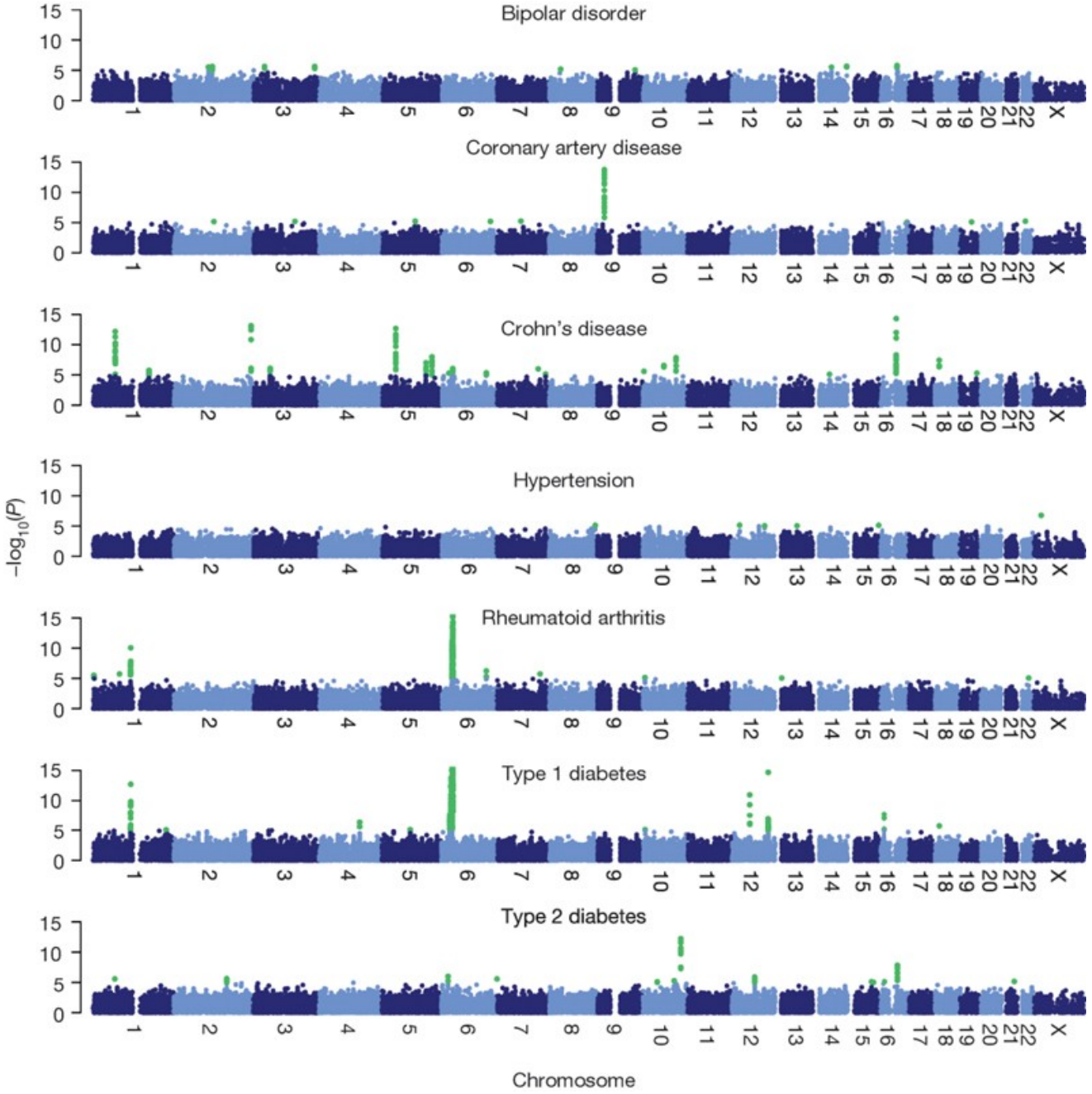




controls (n=1,000)
people without heart disease

WTCCC

- First large scale GWAS (2007)
- 14,000 cases over 7 diseases
- 3,000 shared controls

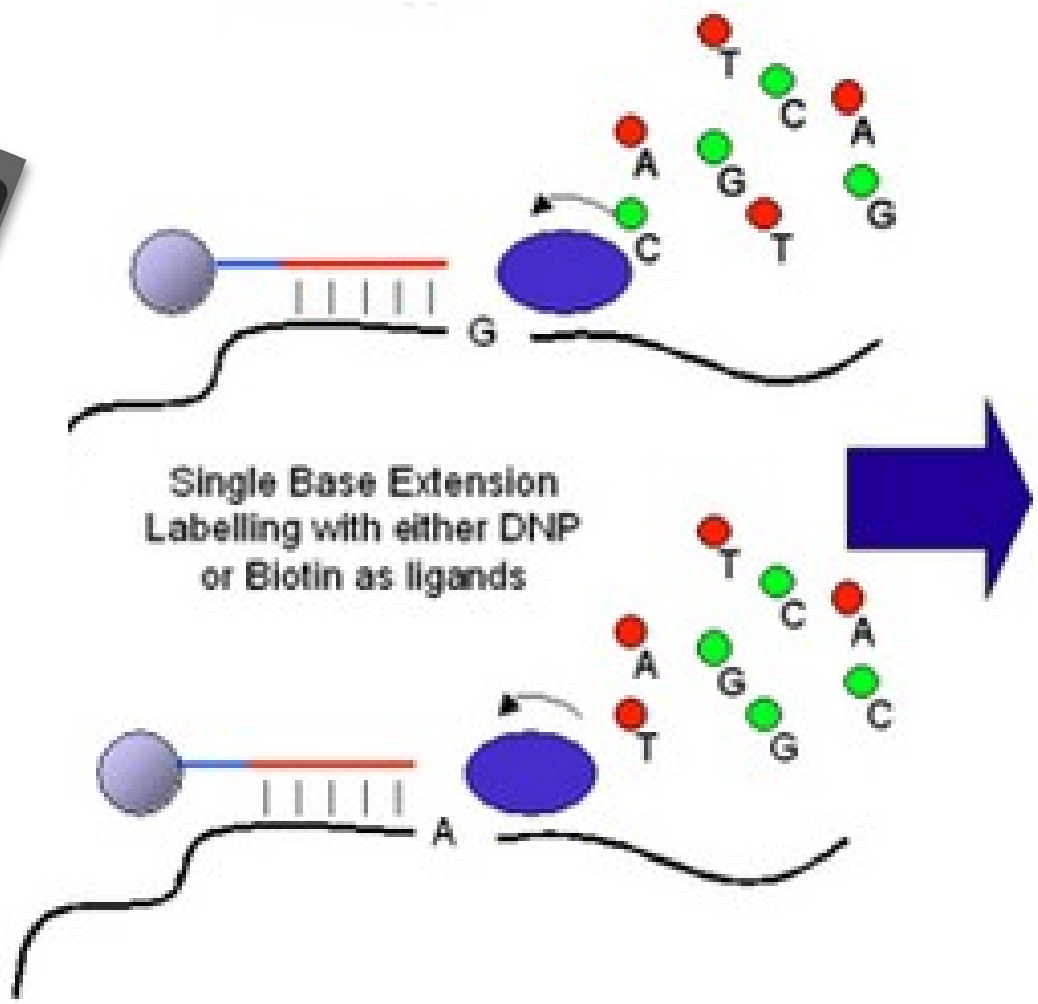


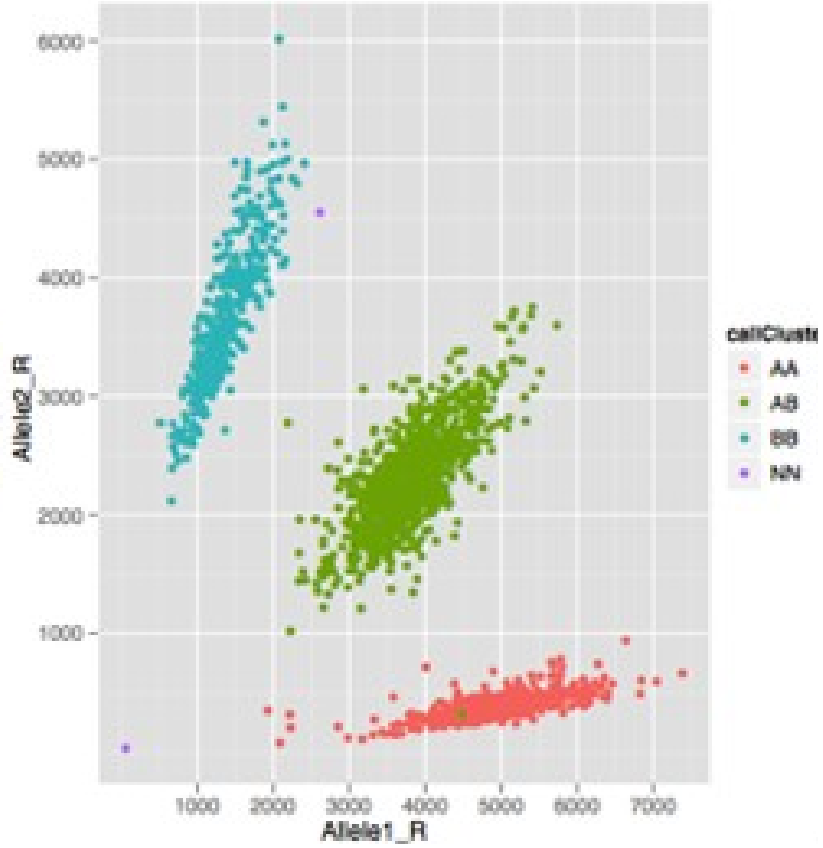
GWAS is Mostly About Preparation

- LOTS can go wrong!
- A GWAS will preform 100s of thousands or millions of statistical tests and take the most significant results
- Any deviation from underlying assumptions can results in a many false positive result
- **Most of the time in GWAS is spent in preparing the data to avoid this pitfall**

SNP Arrays


illumina


Single Base Extension
Labelling with either DNP
or Biotin as ligands
G
A
T
C
A
G
T
C
A
G
T
C
A
T
G
G
C
A


6000
5000
4000
3000
2000
1000
Allele2_R
1000
2000
3000
4000
5000
6000
7000
Allele1_R
AA
AB
BB
NN

Whole Genome Sequencing

Preparing Genotype Data

- We divide the cleaning of genotype data into two steps
- 1) removing any individuals with poor quality data
- 2) removing SNP markers that have substandard genotyping performance

- Performing the per-individual steps first prevents individuals with poor quality genotypes having an undue influence on the removal of SNP markers in the later step.

Per Individual Quality Control

- There are five basic steps to removing “bad” individuals

    1) removal of individuals with excess missing genotypes
    2) removal of individuals with outlying homozygosity values
    3) remove of samples showing a discordant sex
    4) removal of related or duplicate samples, and
    5) removal of ancestry outliers

Excess Missing Genotypes

- For a SNP array, large numbers of missing SNP calls for an individual indicate intensity measures are failing to fall in any genotype clusters
- For sequencing data, large numbers of missing SNP calls indicate low number of reads covering regions of the genome

- Samples with a high missingness rate also tend to have higher genotyping error in the genotypes that are called.
- Remove any sample with “high” (~5%) missingness from further analysis

- Particularly important when using a case-control design – cases and controls often recruited separately → batch effects

Outlying Homozygosity

- The proportion of homozygous (or inversely heterozygous) genotypes across an individual's genome (excluding sex chromosomes) can detect several issues with genotyping
- Average heterozygosity correlates with genotype missingness such that samples with high missingness tend to have lower average heterozygosity (can also reflect inbreeding)
- Sample contamination, where multiple samples are accidentally genotyped on a singe array, results in high average heterozygosity.

- The average heterozygosity depends on population and genotyping platform → need to determine your own threshold!

Discordant Sex

- Determining whether an individual is (genetically) male or female is straightforward from genotype data!

- Males only have a single copy of the X chromosome so can not be heterozygous
- The small pseudo-autosomal region at the end of the chromosome may show heterozygosity
- Some small number of heterozygotes may be attributable to genotyping errors

Related/Duplicate Samples

- Even distantly related samples can bias GWAS results **if not properly accounted for**.
- e.g. if we have two related cases in a case-control analysis, their genotypes being on average more similar to each other than the rest of the cohort will provide a slight bias to the estimate of the allele frequency in cases and its associated standard error
- Even this small bias is important when considering the number of statistical tests being performed.

- Can detect related individuals by calculating Identity-by-State (IBS) or **Identity-by-Descent** (IBD)
  - IBS measures the average proportion of alleles shared by two individuals across the autosomal genome
  - IBD measures the proportion of the genome that is shared between two individuals
- IBD = 1 → Duplicate or monozygotic (identical twin)
  IBD = 0.5 → Parent/offspring, siblings
  IBD = 0.25 → Second degree relative
- For any pair with an IBD > 0.05, remove the one with the lowest genotyping rate

Outlying Ancestry

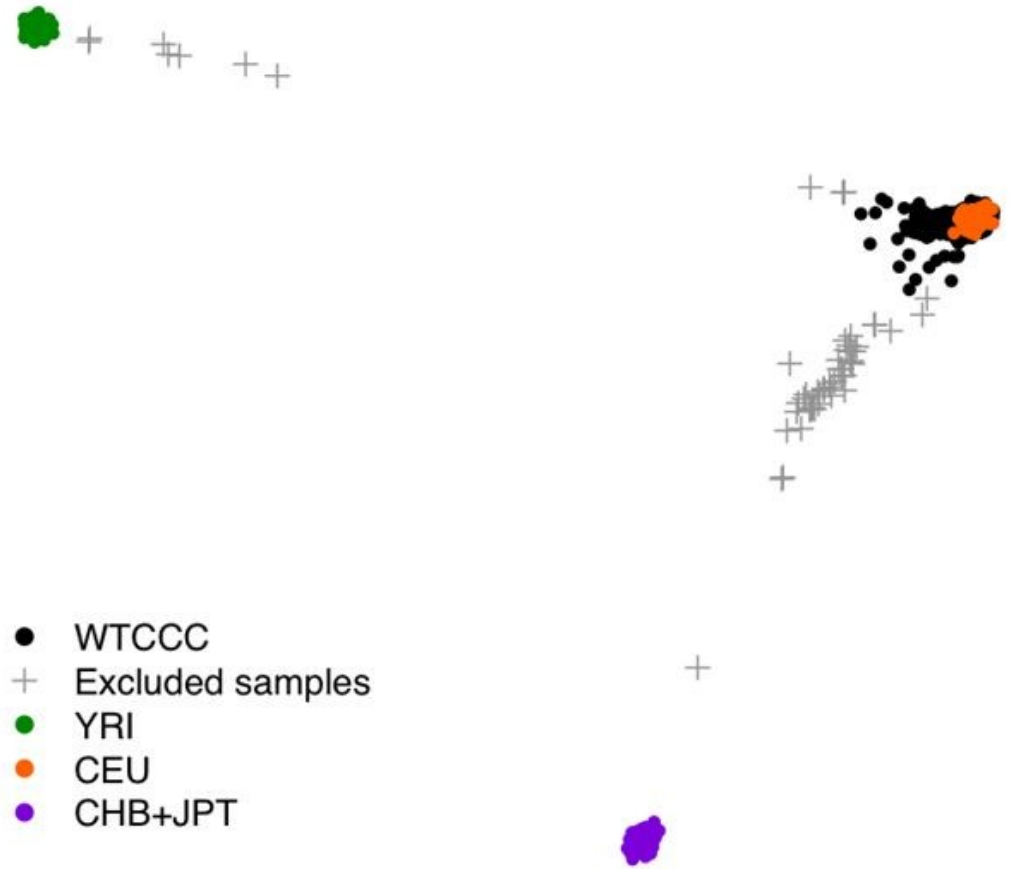
- **Population stratification** is the ***major source of bias*** in GWAS, as it is common for disease or quantitative traits to have different frequencies or distributions across populations

- Real example: Campbell et al. Demonstrating stratification in a European American population. *Nature Genetics* (2005) 37:868–72.
- Performed GWAS on two groups of individuals of European descent that were discordant for height and identified an association with the LCT (lactase) locus

	Height (Adult men)	Lactose Tolerance
Northern (Sweden)	5 ft 11 1/2 in	98%
Southern (Italy)	5 ft 9 1/2 in	$\sim$ 50%

Population Stratification

- Standard approach is to perform PCA on the genotypes of a diverse set of individuals
- Project your samples onto PCs
- Remove outliers



Per Marker Quality Control

- The second stage of genotype cleaning involves looking at individual SNPs to determine genotype accuracy.

- The optimal approach is to look at all cluster plots/sequence alignments individually… That will take a long time!
- Instead we rely on statistical measures on each SNP to detect bad quality data and remove it
  → SNP filtering is a short cut
  → The level of SNP filtering is a trade-off

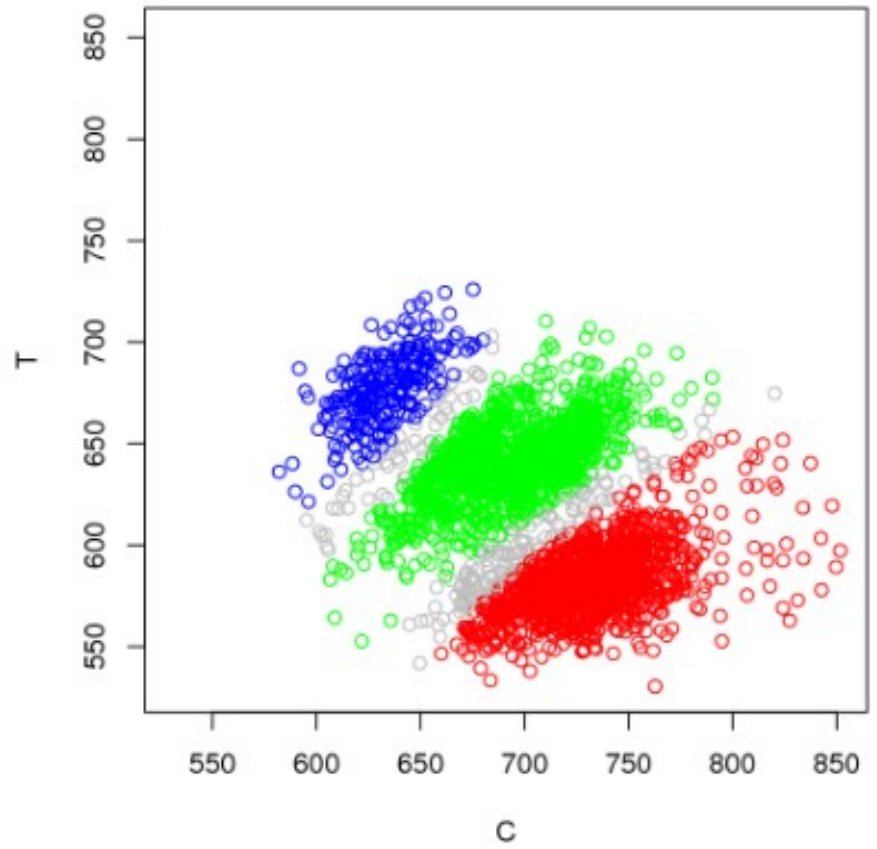
Per Marker Quality Control

- The four steps of marker quality control:

1) removal of SNPs with excess missing genotypes
2) removal of SNPs that deviate from Hardy-Weinberg equilibrium
3) removal of SNPs with low minor allele frequency
4) comparing allele frequency to known values

Excess Missing Genotypes

- Caused by poor separation of genotyping clusters on arrays, or low number of sequence reads over a region
- These conditions make the error rate in the non-missing genotypes higher
- Remove any SNP with > 5% missing data



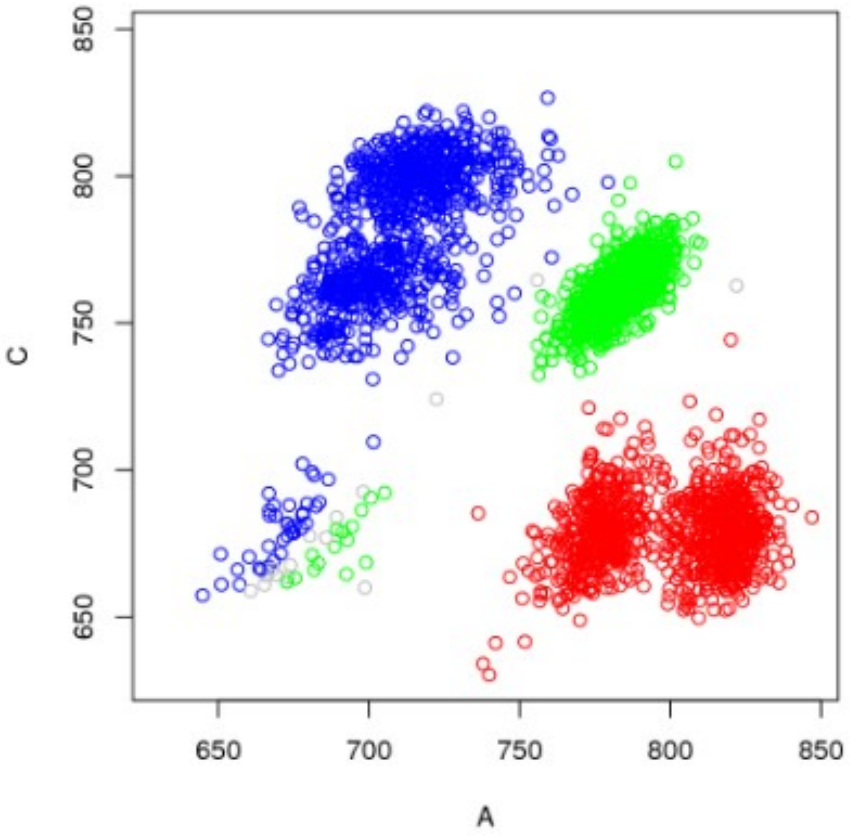
Excess Missing Genotypes

- **An additional check is particularly important for case-control studies!**

- Remove any SNPs that have different rates of missingness between cases and controls
- Missingness can be non-random with respect to the underlying genotype
- Differential missing genotype rates between cases and controls can lead to false positive results

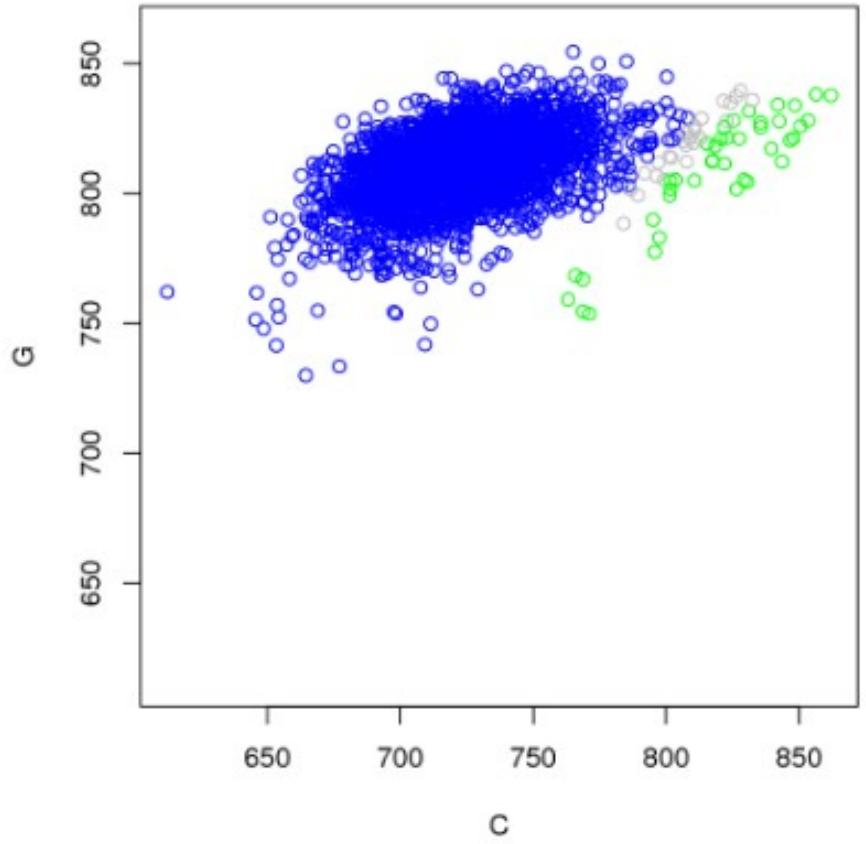
Deviation From Hardy-Weinberg Equilibrium

- Humans are a large population, so genotype frequencies tend to satisfy HWE
- Poor genotype calling can result in genotype frequencies deviating from Hardy-Weinberg equilibrium
- Poor cluster separation in arrays, lack of heterozygous calls in sequencing
- Remove SNPs that have a deviation from HWE $p < 10^{-6}$



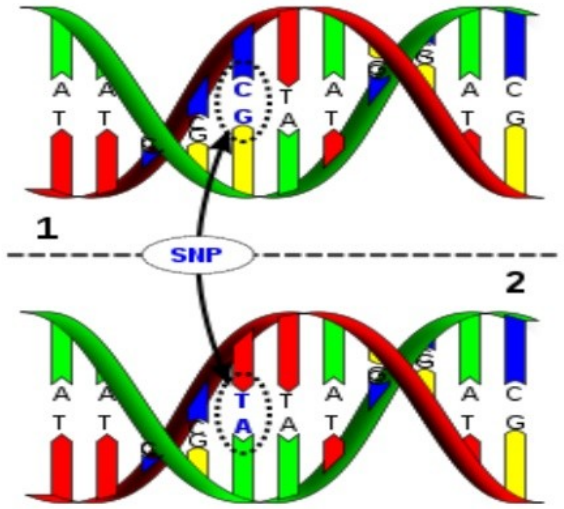
Low Minor Allele Frequency

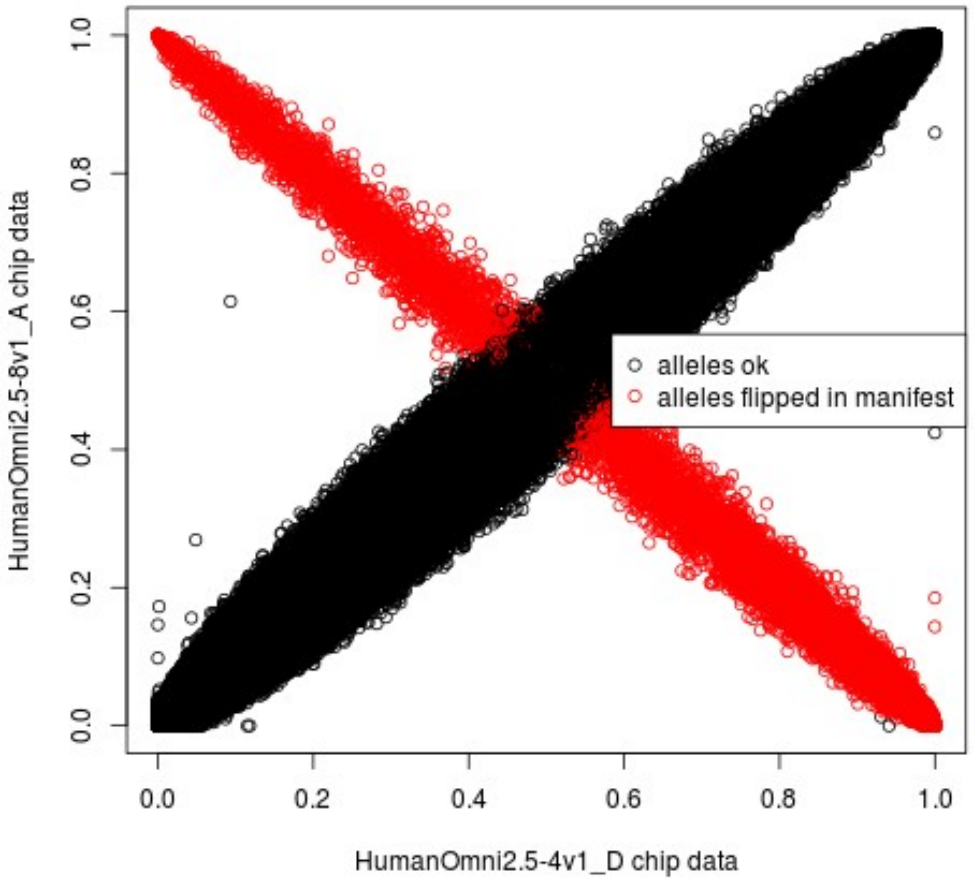
- For a SNP with minor allele frequency of q = 1%, the frequency of the minor homozygote is 0.01% under HWE
- We need 10,000 individuals if we expect to see 1 of the rare homozygous genotypes
- SNP calling algorithms for arrays want three clusters and may invent clusters when they can not find them
- A “reasonable” number of each genotype are required



Allele Frequency

- We have good allele frequency estimates for genetic variants in a range of populations
- Differences in allele frequency between populations can indicate poor quality genotypes
- Also can detect strand alignment issues





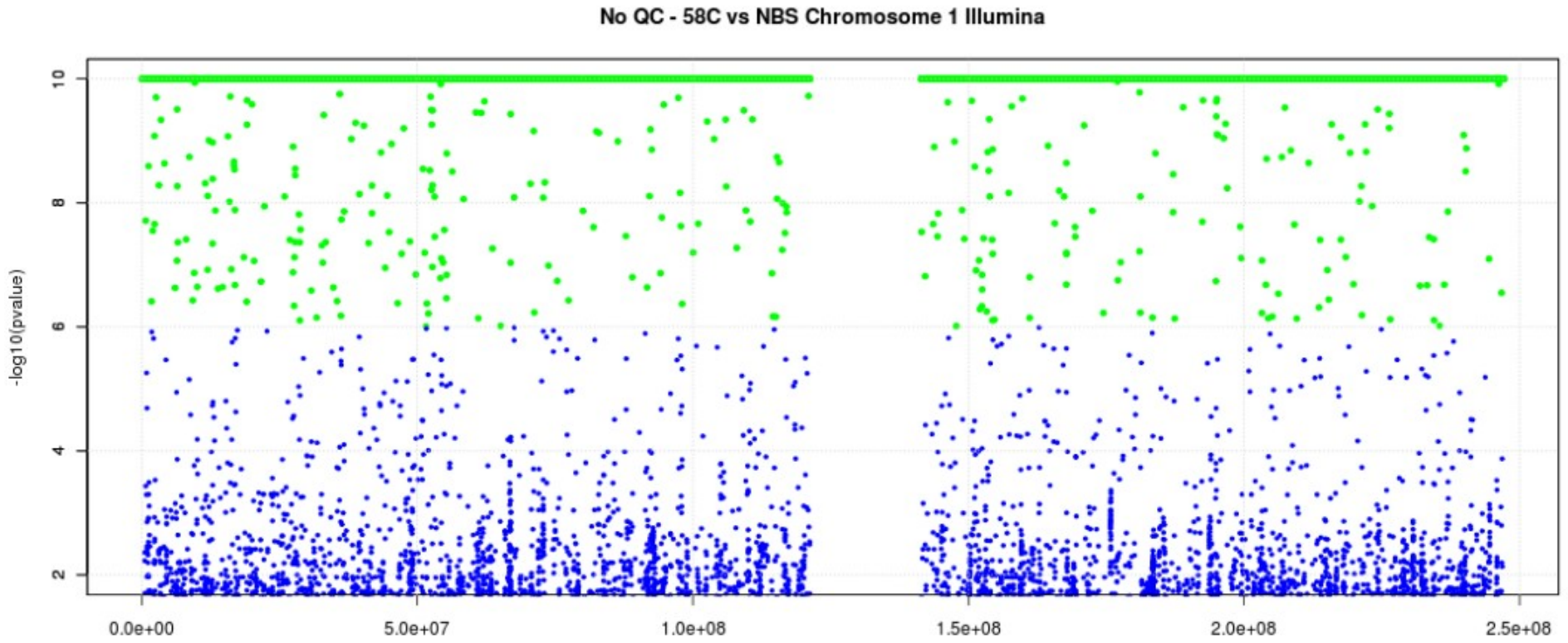
Example: Importance of Good Cleaning

- The WTCCC study used controls from two populations:

  - 1,500 from the 1958 British Birth Cohort (58C)
  - 1,500 from the National Blood Service (NBS)

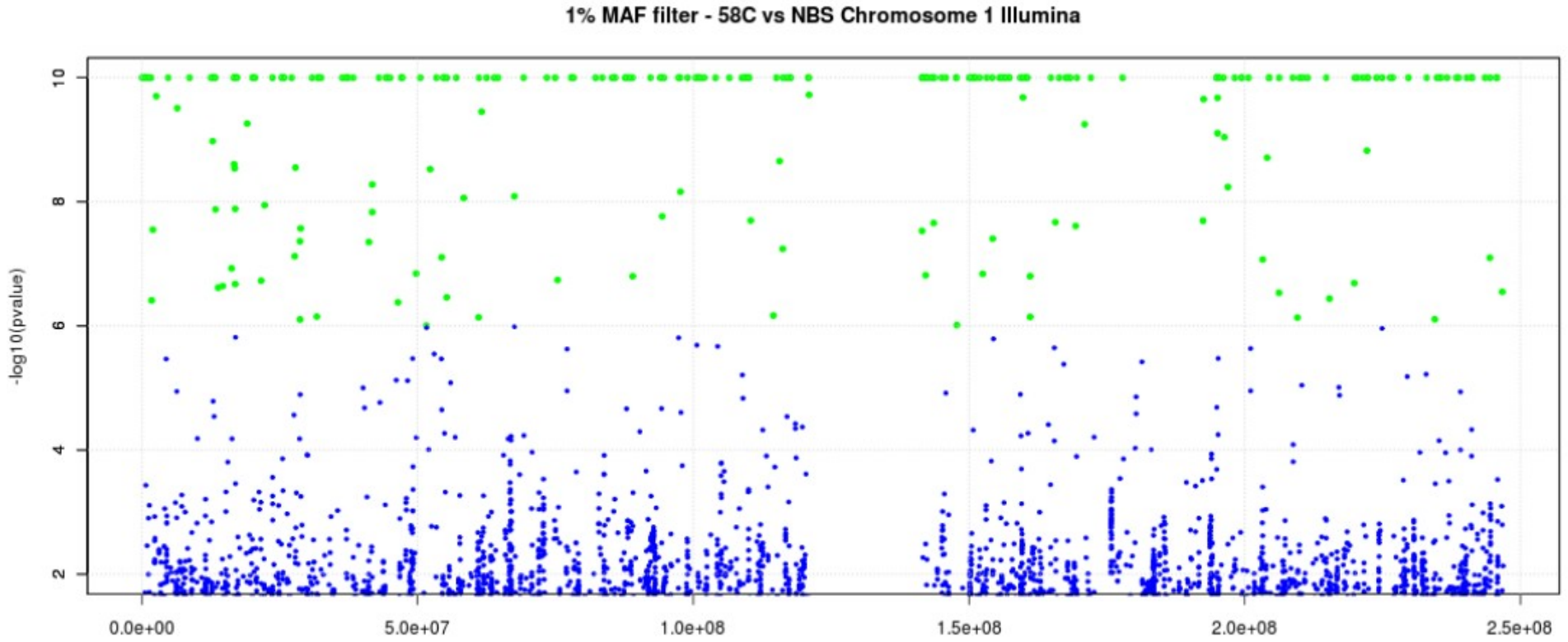
- Both these are unselected population cohorts, so performing a “case-control” study between these populations should find no significant differences

Importance of Good Cleaning



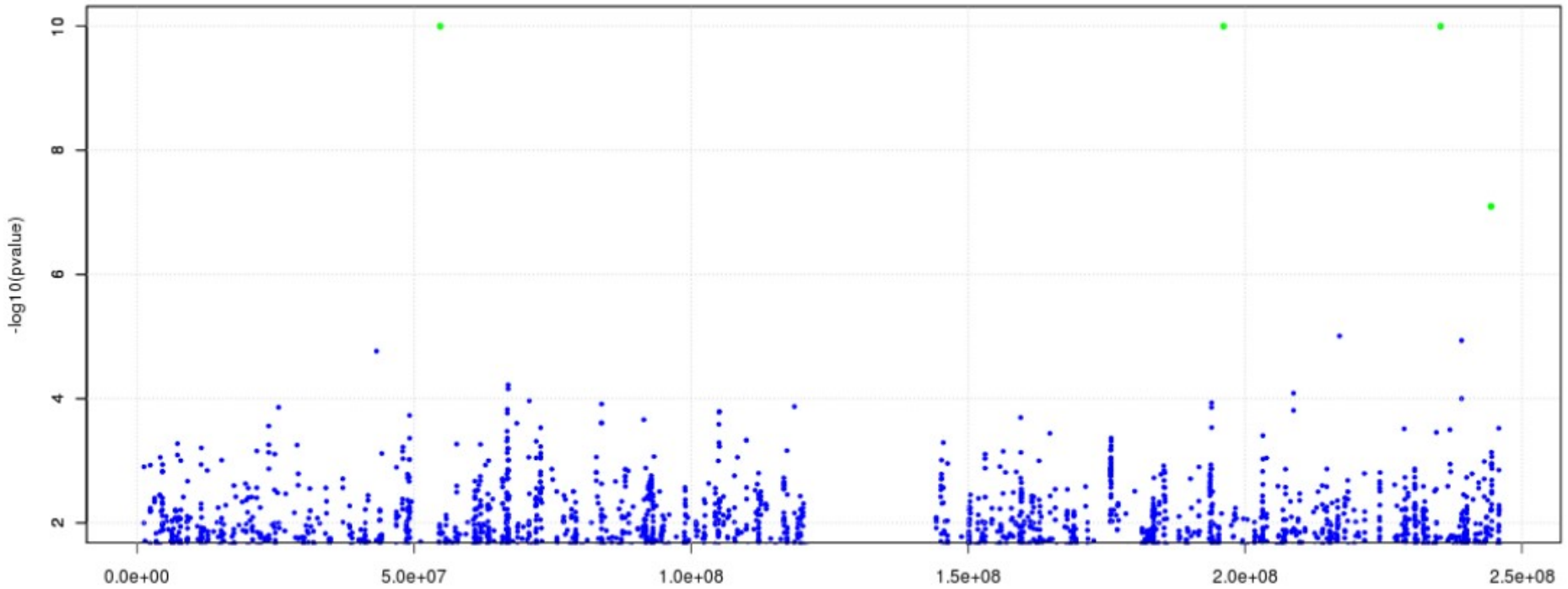
100% of SNPs

Importance of Good Cleaning



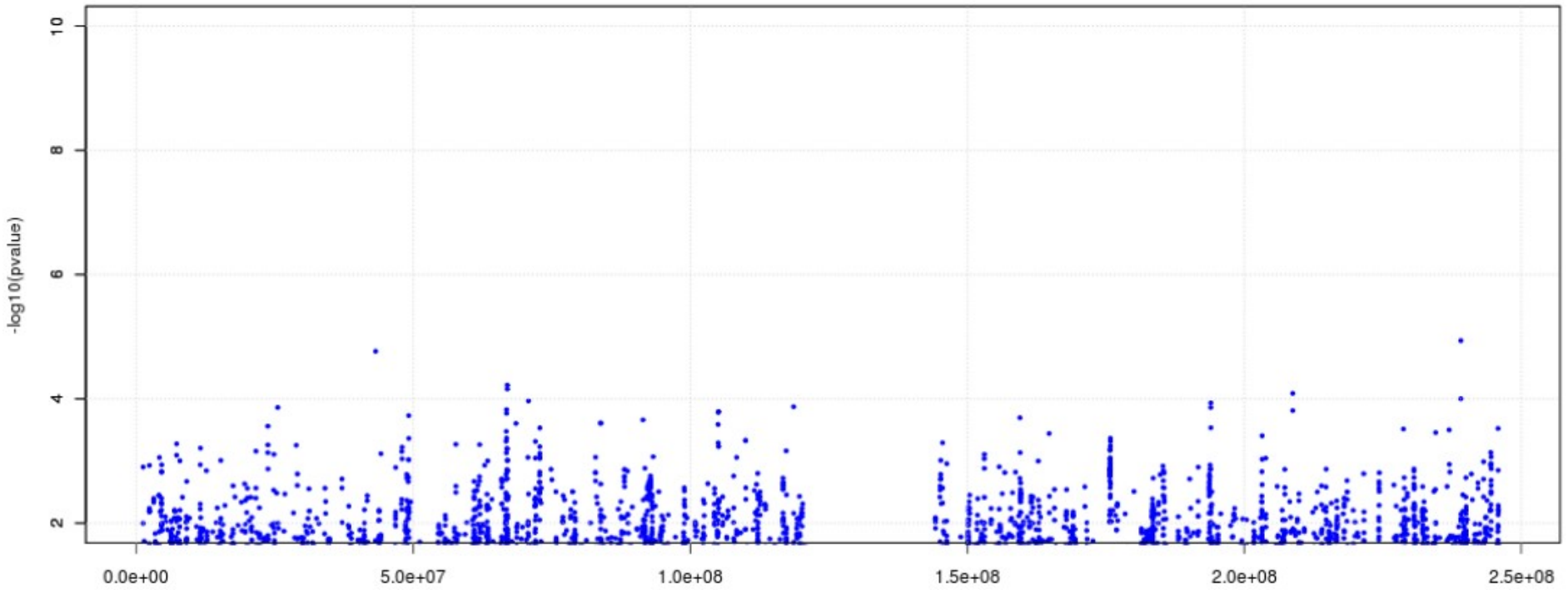
80.69% of SNPS
Filtering: MAF

Importance of Good Cleaning



78.36% of SNPs
Filtering: MAF + HWE

Importance of Good Cleaning



77.92% of SNPs
Filtering: MAF + HWE + Missingness

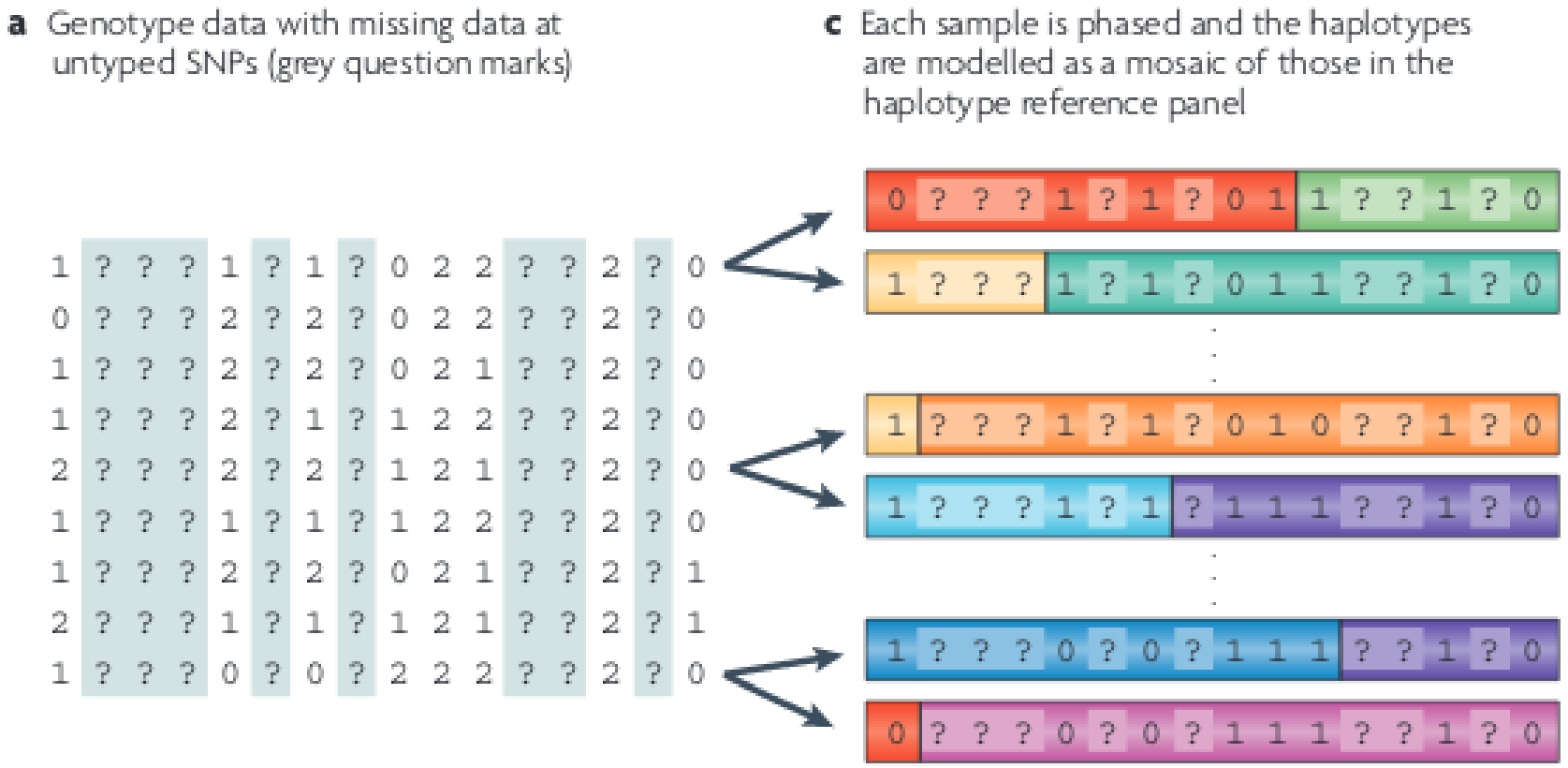
Imputation

- Genotype imputation is the process of predicting, or imputing, genotypes that are not known in a sample of individuals
- This is used to:

    - Fill missing genotypes for an individual at a SNP
    - Recover genotypes of SNPs removed during QC
    - Get genotypes at SNPs not measured on an array

- The imputed SNPs can be tested for association in the GWAS in the same way as actually genotyped SNPs
- This increase the power to detect associations

Imputation

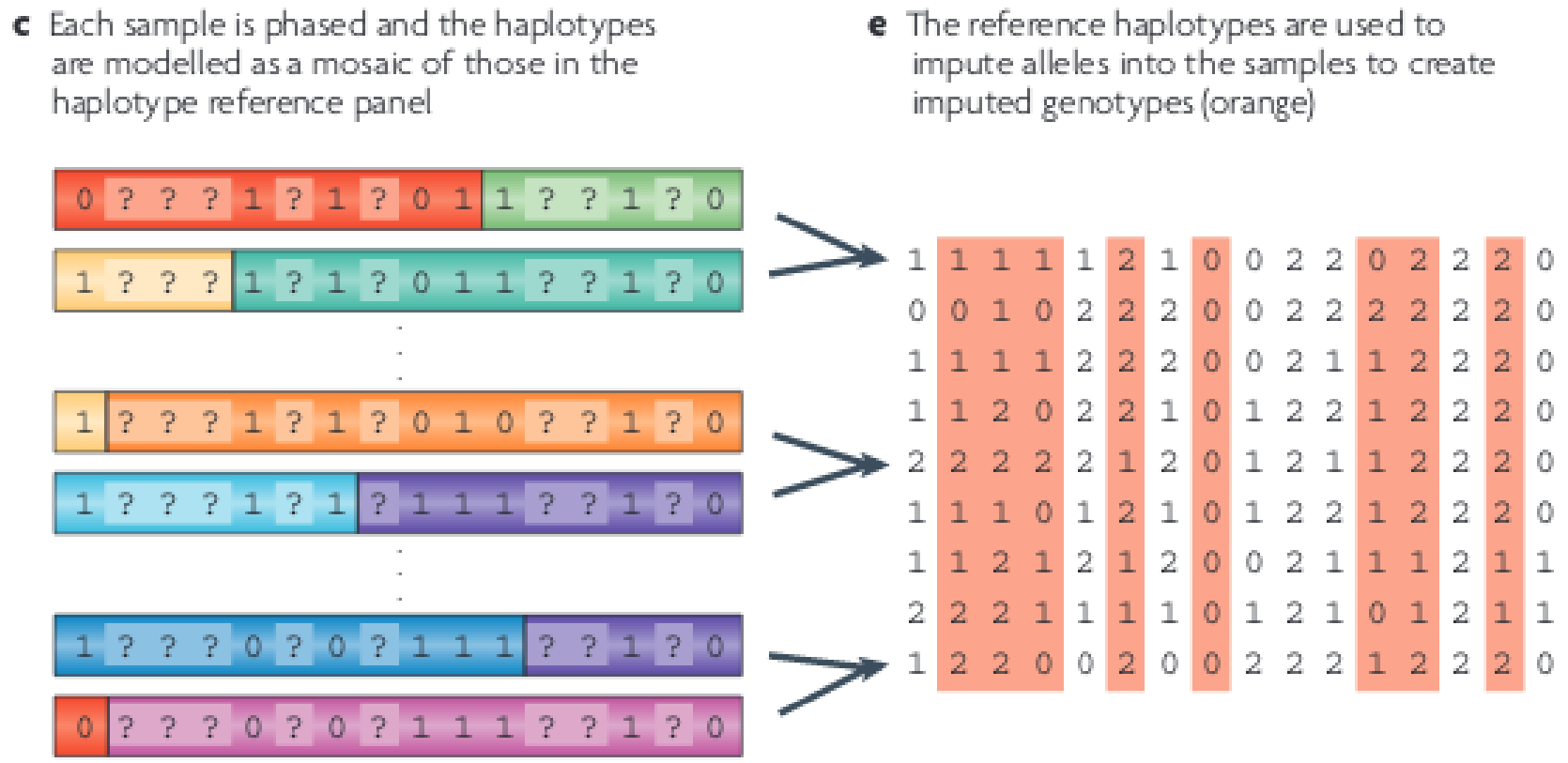

a Genotype data with missing data at untyped SNPs (grey question marks)
c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

Imputation

d Reference set of haplotypes, for example, HapMap

c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

Imputation


c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel
e The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)
0 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 0 1 0 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 1 1 1 ? ? 1 ? 0
1 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
0 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
1 1 1 1 1 2 1 0 0 2 2 0 2 2 2 0
0 0 1 0 2 2 2 0 0 2 2 2 2 2 2 0
1 1 1 1 2 2 2 0 0 2 1 1 2 2 2 0
1 1 2 0 2 2 1 0 1 2 2 1 2 2 2 0
2 2 2 2 2 1 2 0 1 2 1 1 2 2 2 0
1 1 1 0 1 2 1 0 1 2 2 1 2 2 2 0
1 1 2 1 2 1 2 0 0 2 1 1 1 2 1 1
2 2 2 1 1 1 1 0 1 2 1 0 1 2 1 1
1 2 2 0 0 2 0 0 2 2 2 1 2 2 2 0