Genome-Wide Association Studies (GWAS) – #2

Recap

A GWAS performs 100s of thousands or millions of statistical tests and takes the most significant results

Any deviation from underlying assumptions can results in a many false positive results

Most of the time in GWAS is spent in preparing the data to avoid this pitfall

Per Individual Quality Control

The 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

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 would tale a long time...

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 therefore a trade-off

Per Marker Quality Control

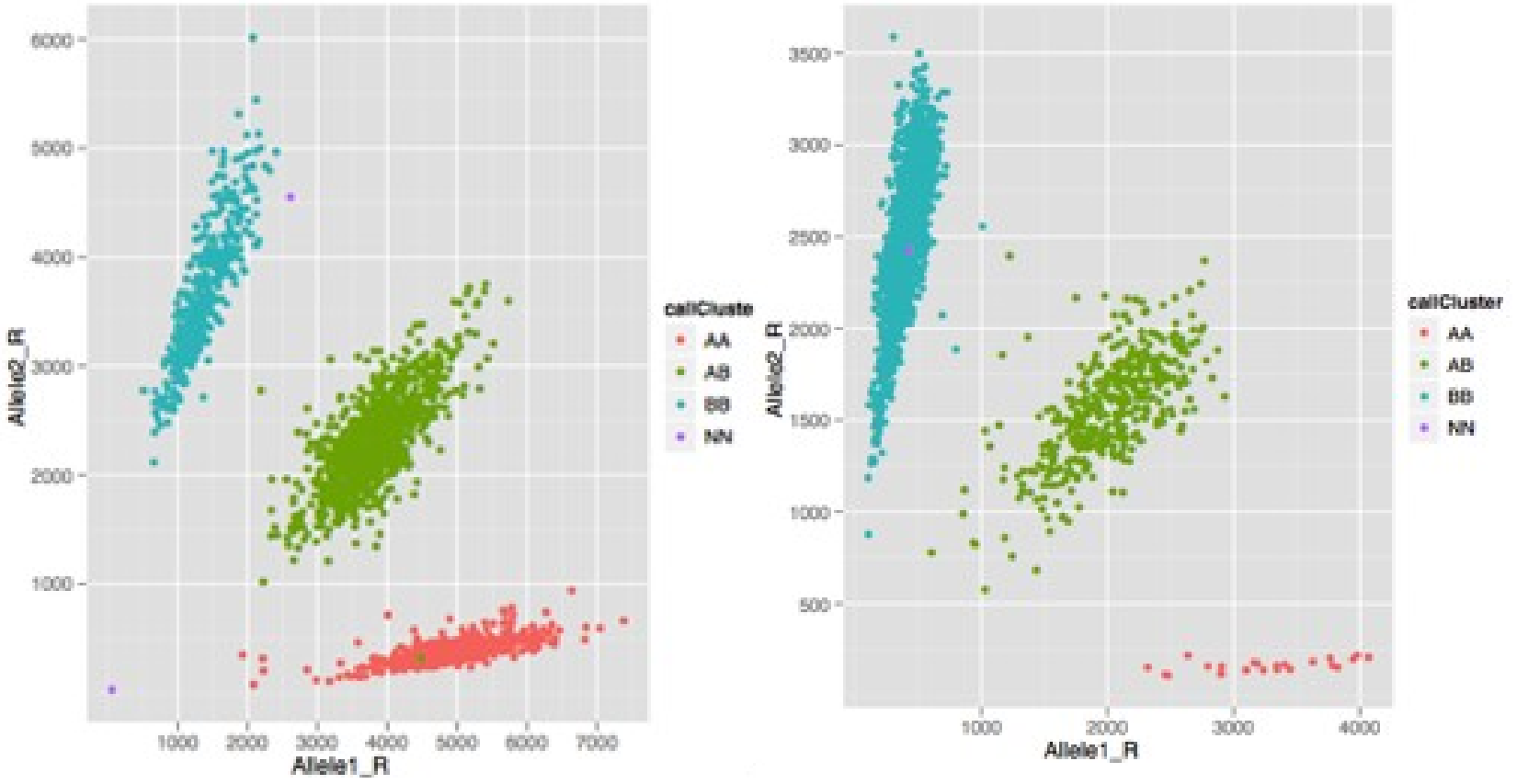
1) removal of SNPs with excess missing genotypes

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

SNP Arrays


Allele2_R
Allele1_R
6000
5000
4000
3000
2000
1000
1000
2000
3000
4000
5000
6000
7000
callCluste
AA
AB
BB
NN
3500
3000
2500
2000
1500
1000
500
callCluster

Excess Missing Genotypes

1) removal of SNPs with excess missing genotypes

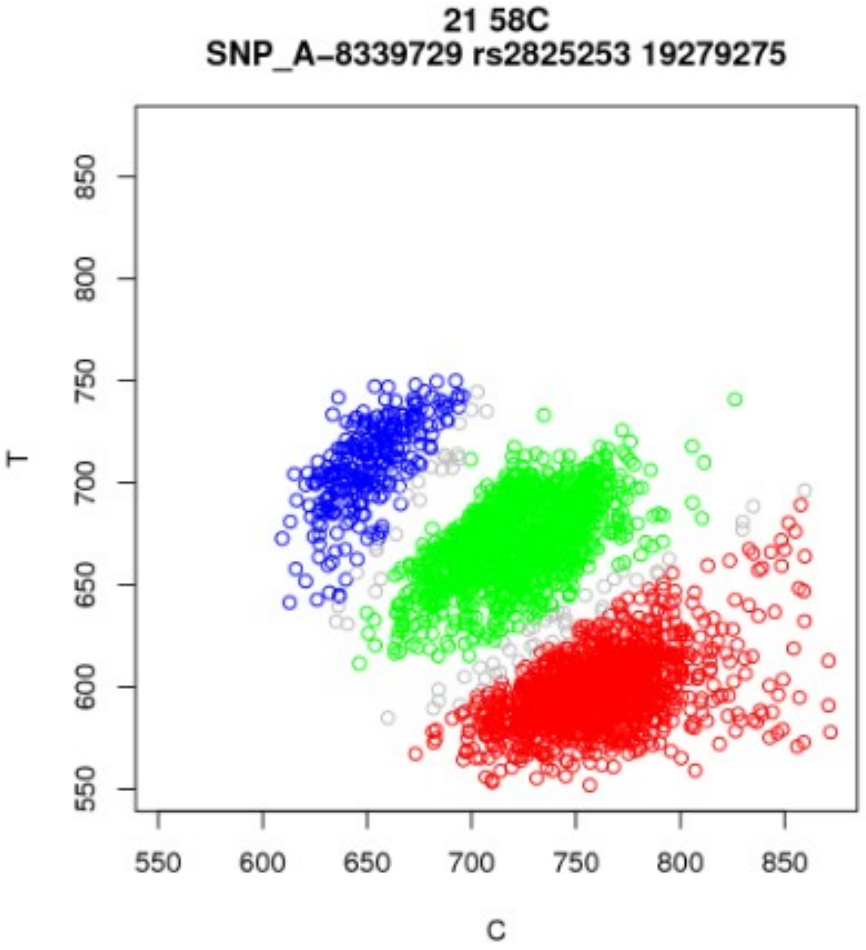
Caused by:

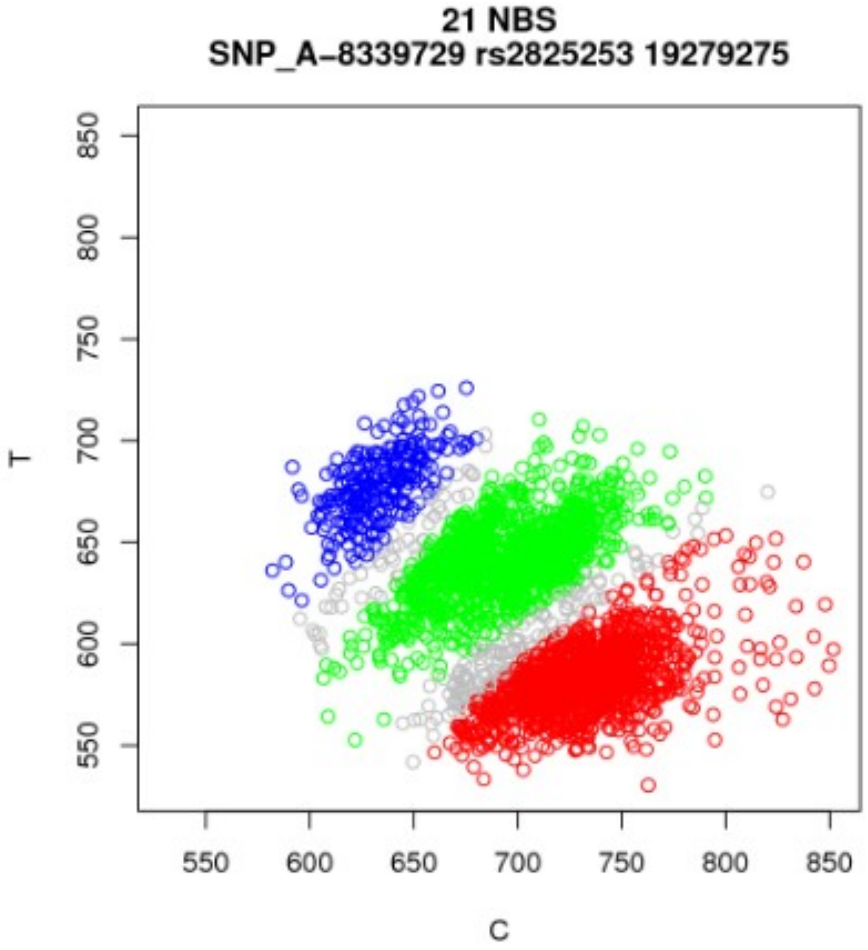
- Poor separation of genotyping clusters (arrays)
- Low number of sequence reads over a region (sequencing)

These conditions make the error rate in the non-missing genotypes higher

Remove any SNP with > 5% missing data

Excess Missing Genotypes





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 HWE

Refresher – Hardy-Weinberg Equilibrium

p = frequency of allele A
q = frequency of allele a

$p + q = 1$

$(p + q)^2 = 1$
$p^2 + 2pq + q^2 = 1$

p^2 = frequency of genotype AA
2pq = frequency of genotype Aa
q^2 = frequency of genotype aa

Deviation From HWE

Refresher – Hardy-Weinberg Equilibrium

Assumptions:

- Large population
- Random mating
- No mutation
- Migration ~ 0
- Natural selection does not affect the locus

Deviation From HWE

2) removal of SNPs that deviate from Hardy-Weinberg equilibrium

Poor genotype calling can result in genotype frequencies deviating from Hardy-Weinberg equilibrium.

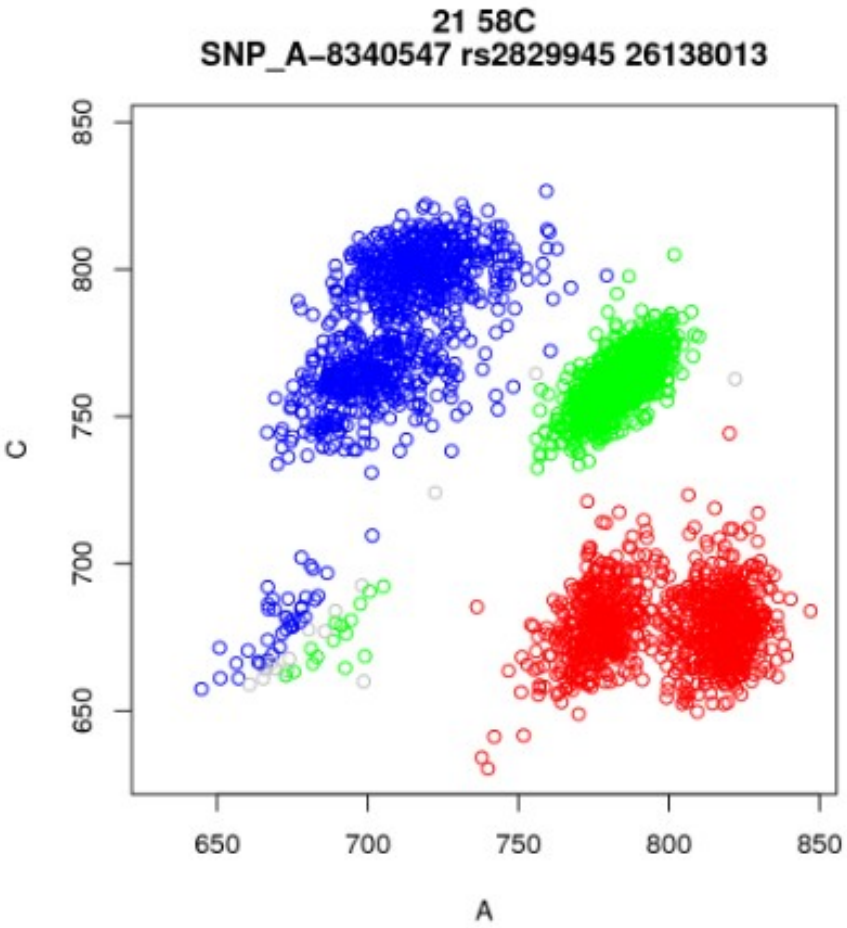
Arrays:

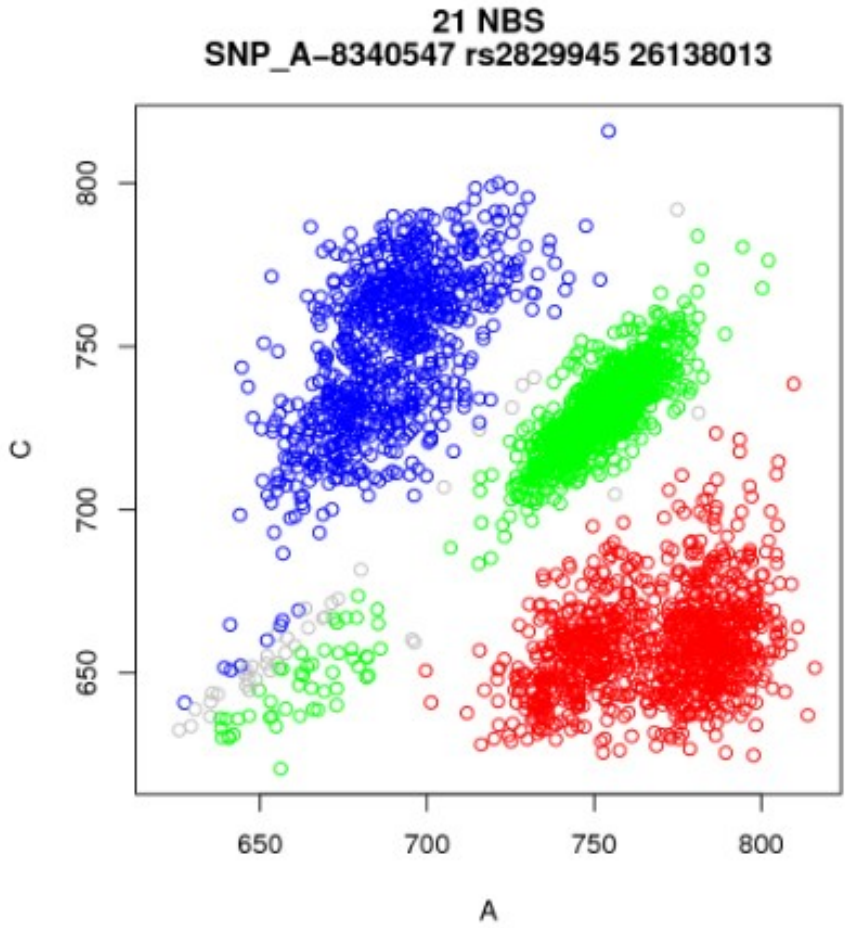
- poor cluster separation
- structural variation (copy number variation)

Sequencing:

- lack of heterozygous calls

Deviation From HWE





Deviation From HWE

Remove SNPs that have a deviation from HWE $p < 10^{-6}$

NOTE:
The following assumption may be violated for disease loci:
- Natural selection does not affect the locus

→ only test for deviation from HWE in controls

Low Minor Allele Frequency

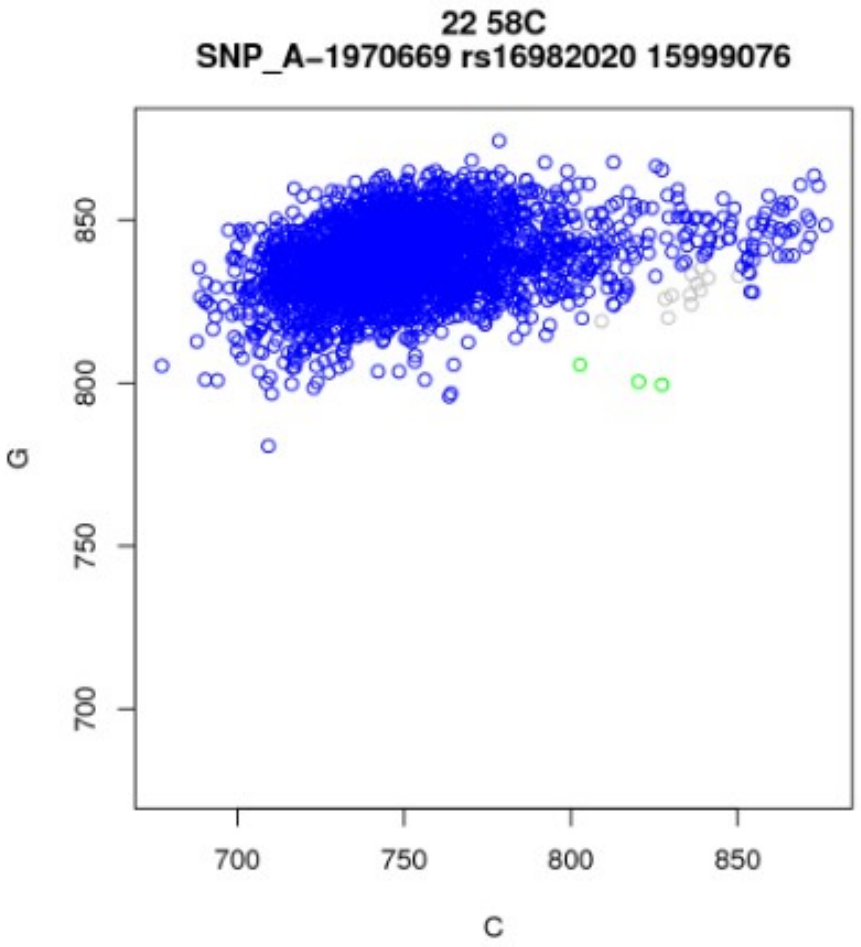
3) removal of SNPs with low minor allele frequency

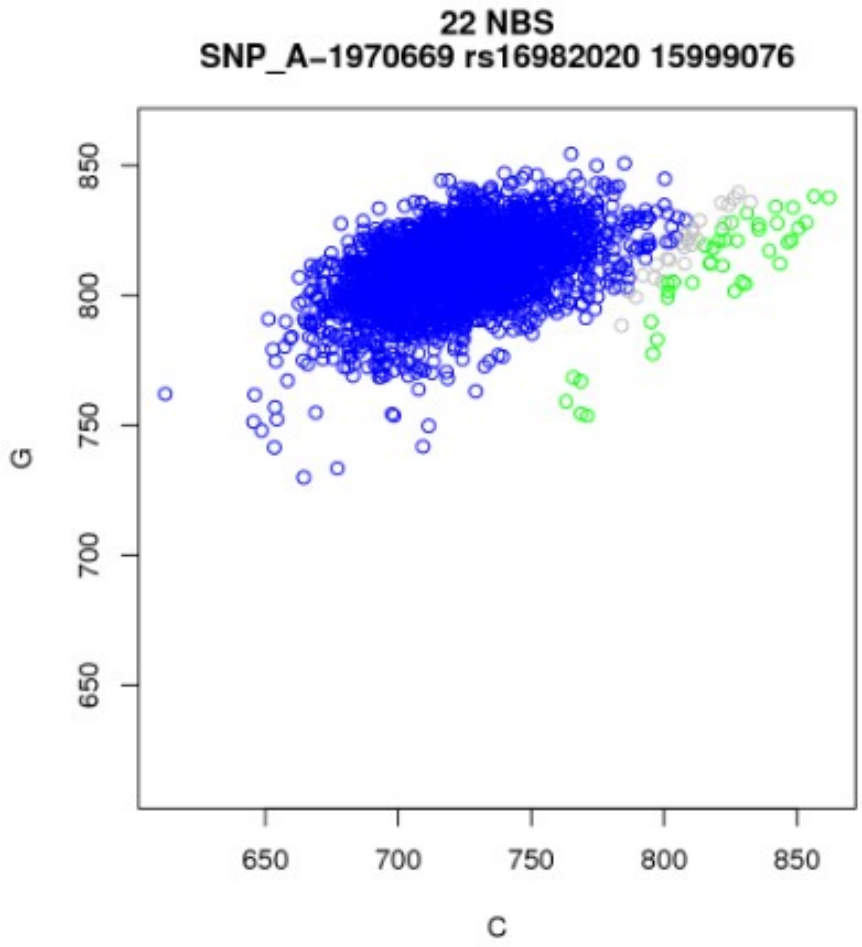
For a SNP with minor allele frequency of q = 1%

- p(AA) = $(0.99)^2$ = 98.01%
- p(Aa) = 2*0.99*0.01 = 1.98%
- p(aa) = $(0.01)^2$ = 0.01%

We need 10,000 individuals if we expect to see 1 of the rare homozygous genotypes

Low Minor Allele Frequency





Low Minor Allele Frequency

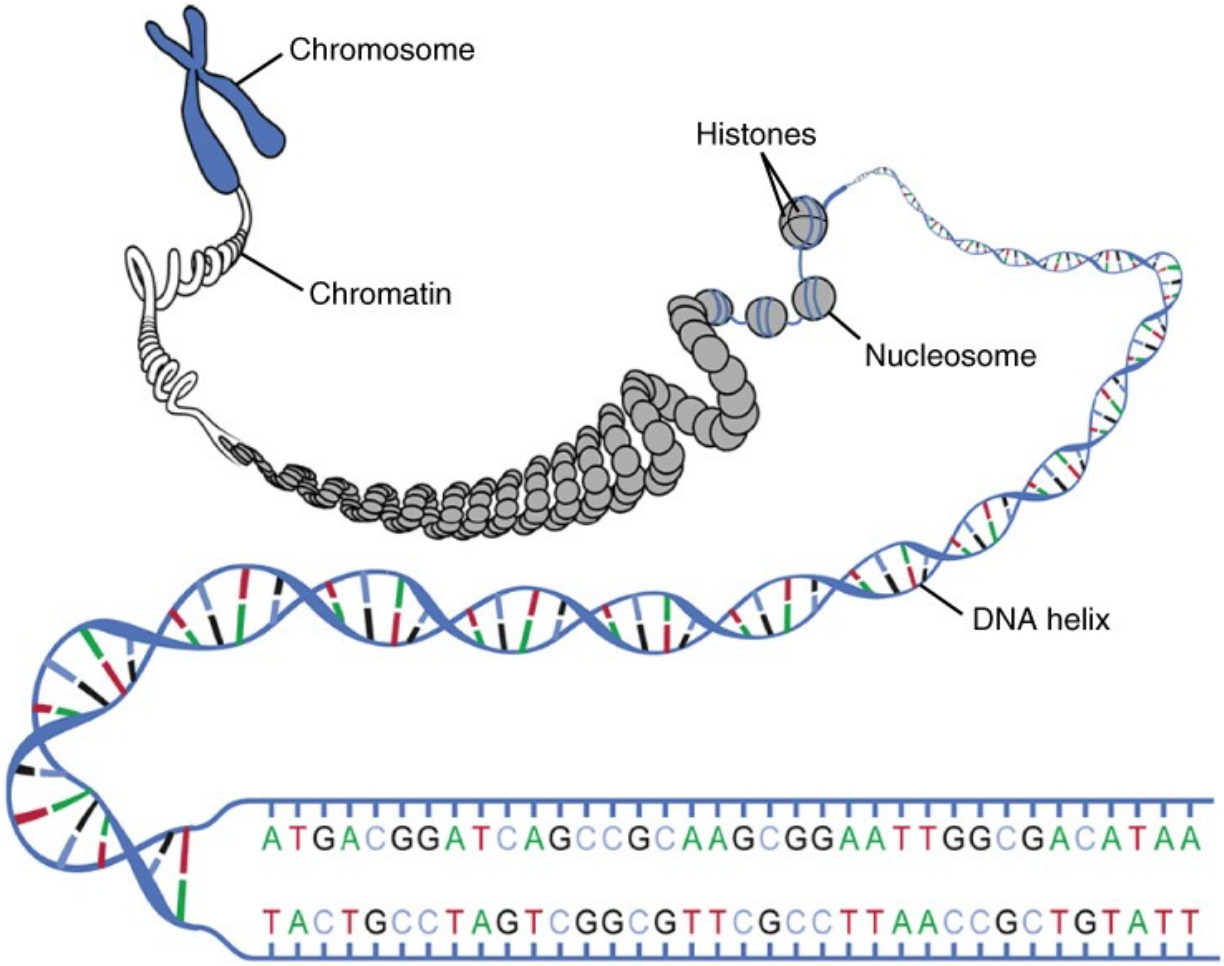
SNP Arrays:
- SNP calling algorithms want three clusters and may invent clusters when they can not find them

- A “reasonable” number of each genotype are required

- Can work around this with population panels, but there are large limitations

Remove SNP with MAF < 1% or 5% depending on your sample size

Strand Alignment



Allele Frequency

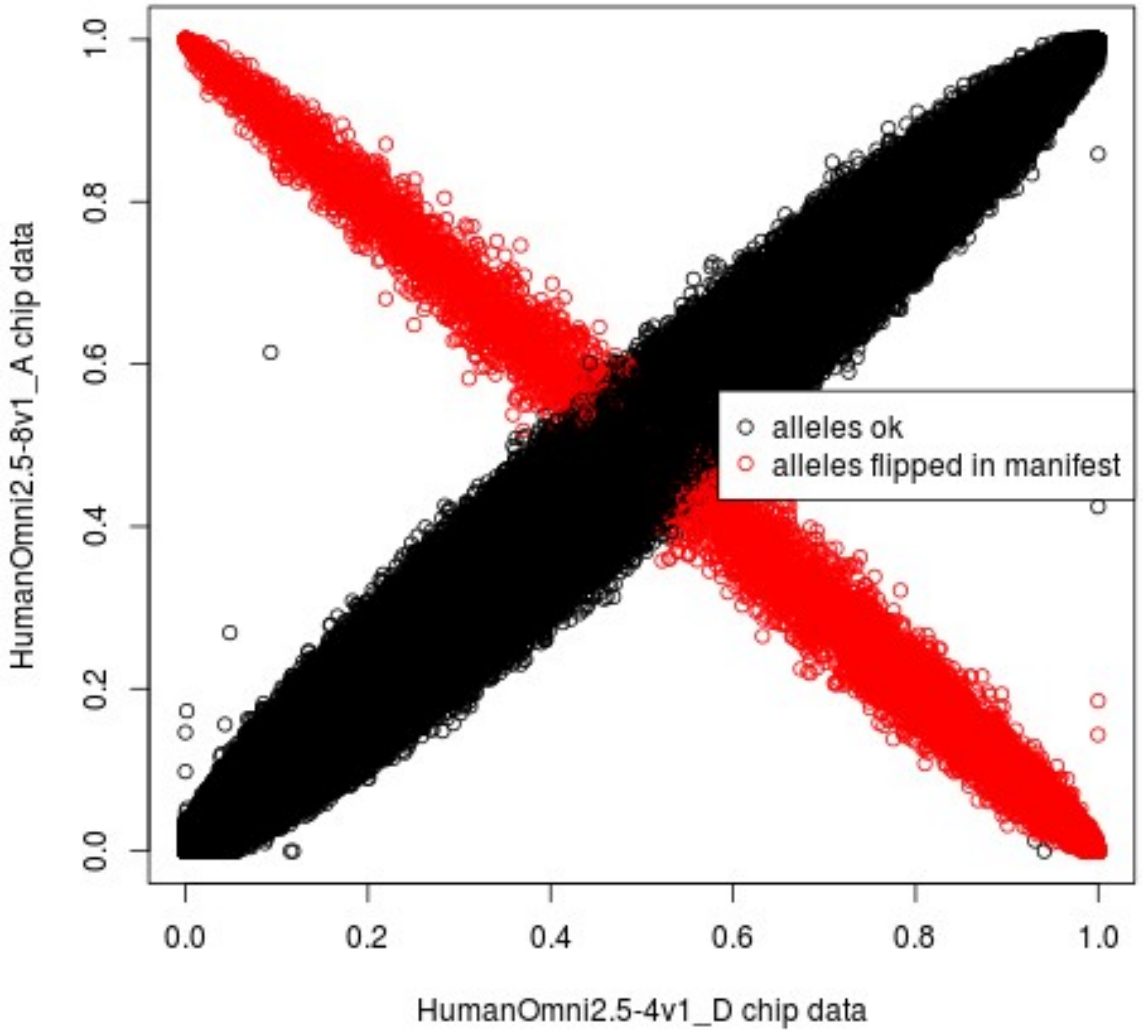
4) comparing allele frequency to known values

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
- failure to generate clusters

Also can detect strand alignment issues

Allele Frequency



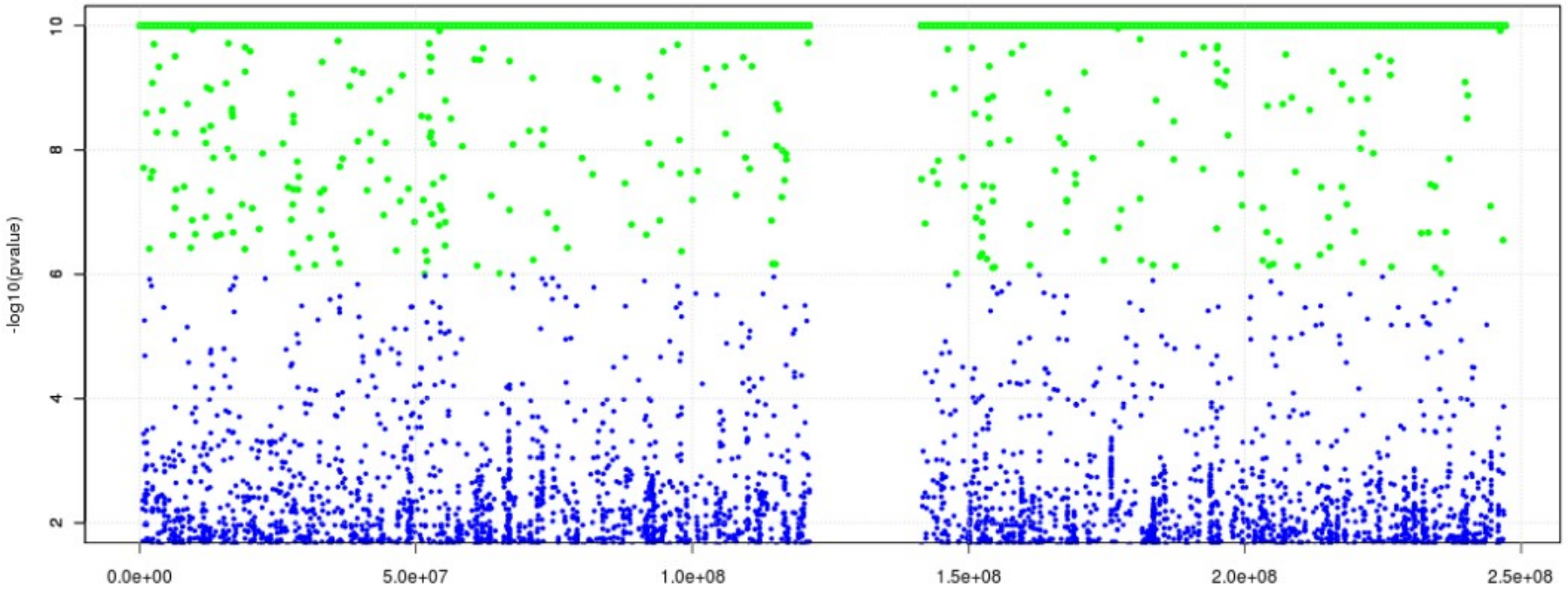
Importance of Good Cleaning

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)

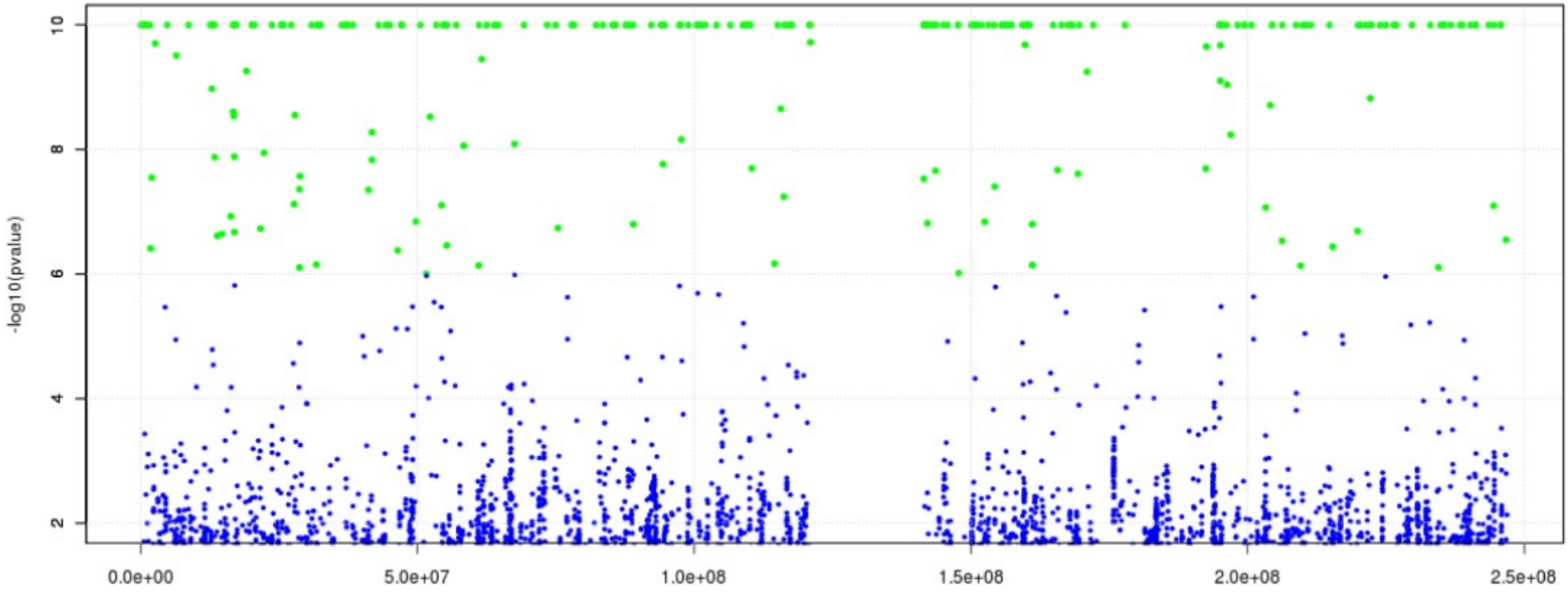
Both these are unselected population cohorts, so performing a “case-control” study between these populations ***should*** find no significant differences

Importance of Good Cleaning



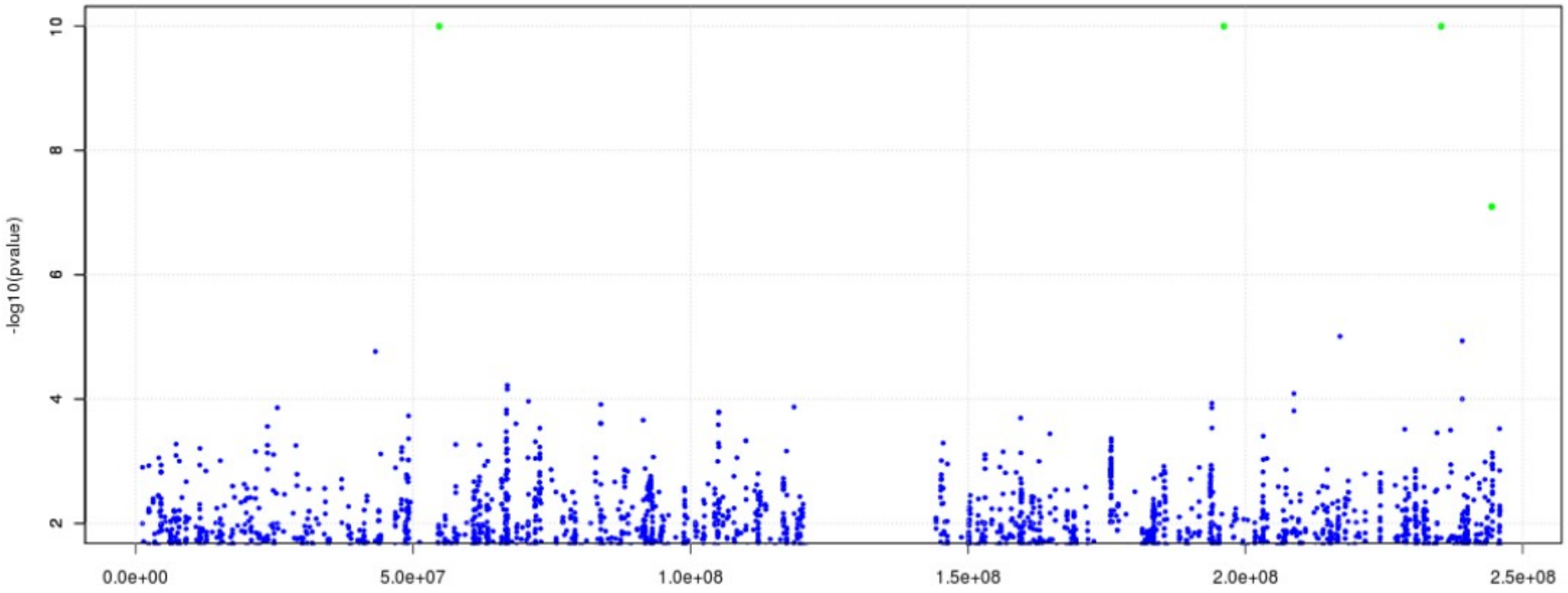
100% of SNPs

Importance of Good Cleaning



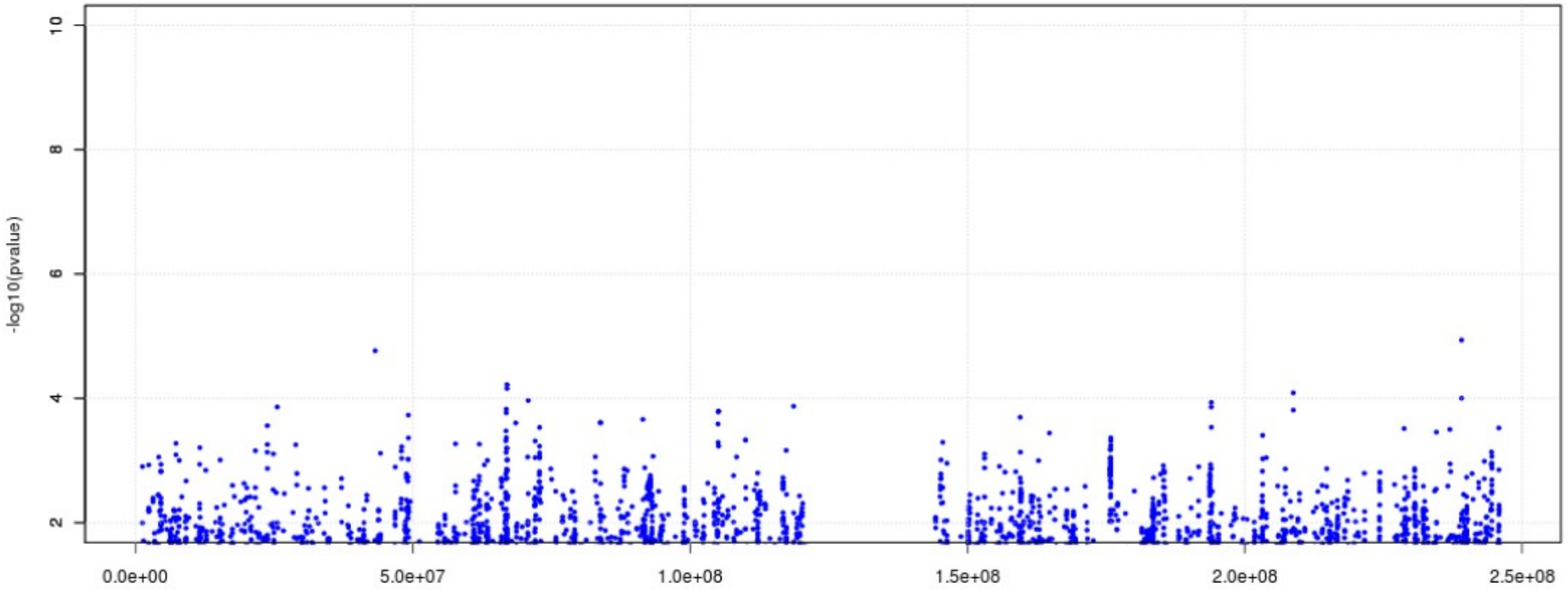
80.69% of SNPS
Filtering: MAF

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 actually genotyped SNPs are

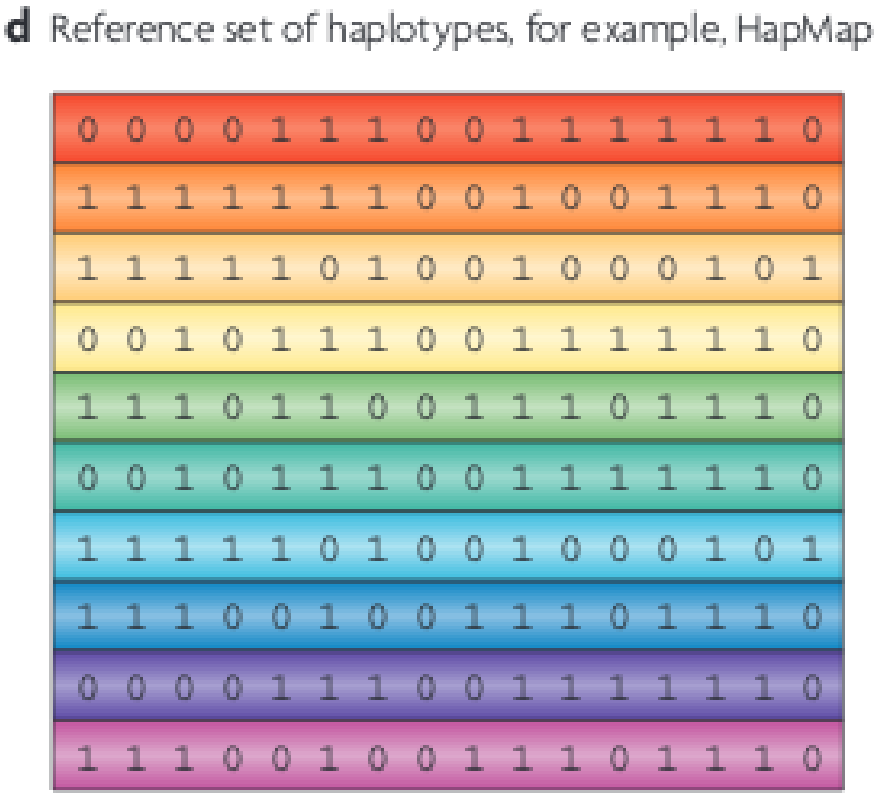
This increase the power to detect associations

Reference Panels

Sets of dencely genotyped individuals

Need to cover the genetic variation in the population being imputed
→ match ancestry

Some gain by including additional ancestries in the reference population to capture rarer haplotypes

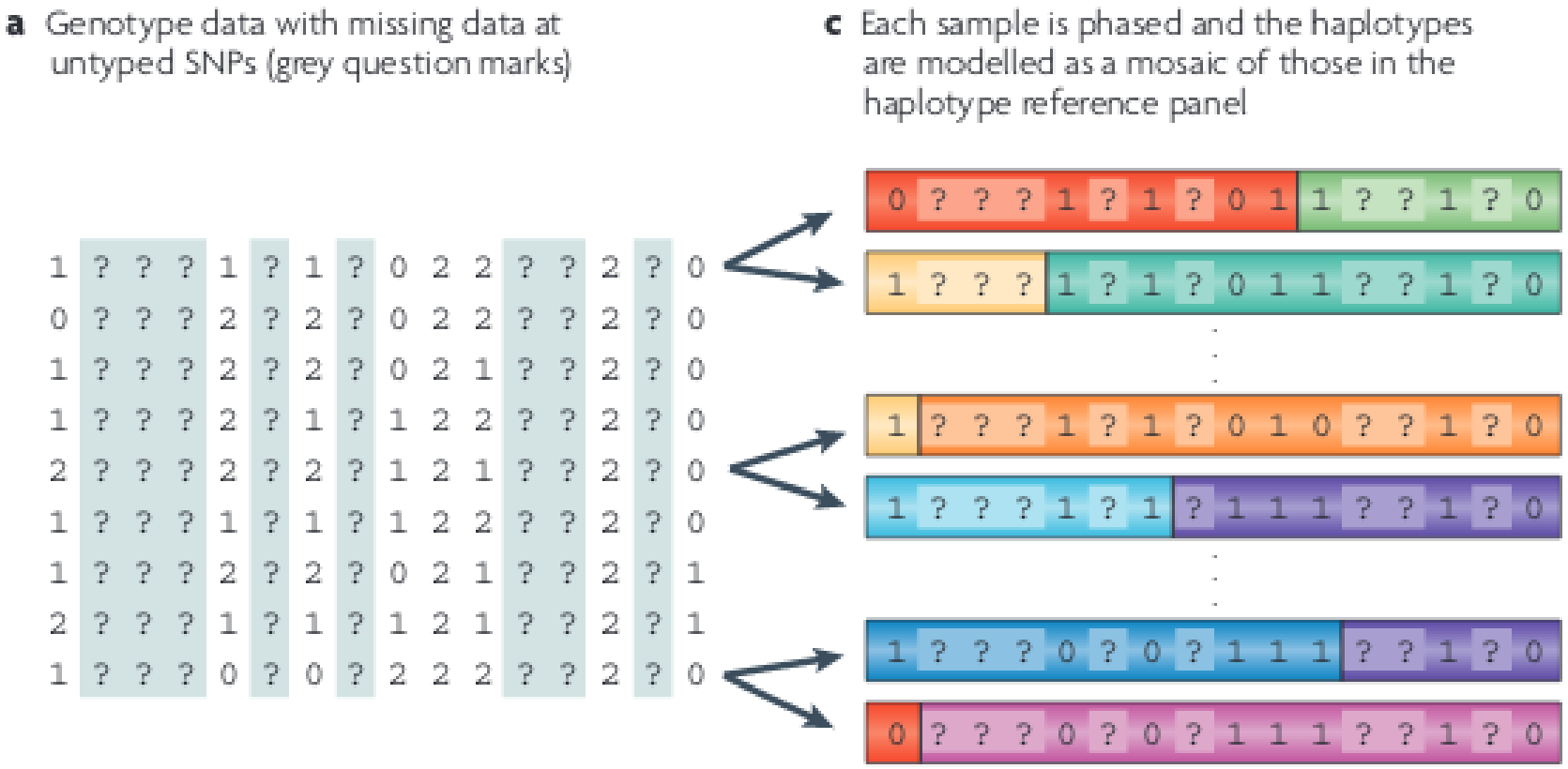


Reference Panels

A number of widely used reference panels are availabe:

Haplotype Reference Consortium (release 1.1)	32,470	40M
African Genome Resources	4,956	93M
1000 Genomes Phase 3	2,504	85M
UK10K	3,781	24M

Imputation

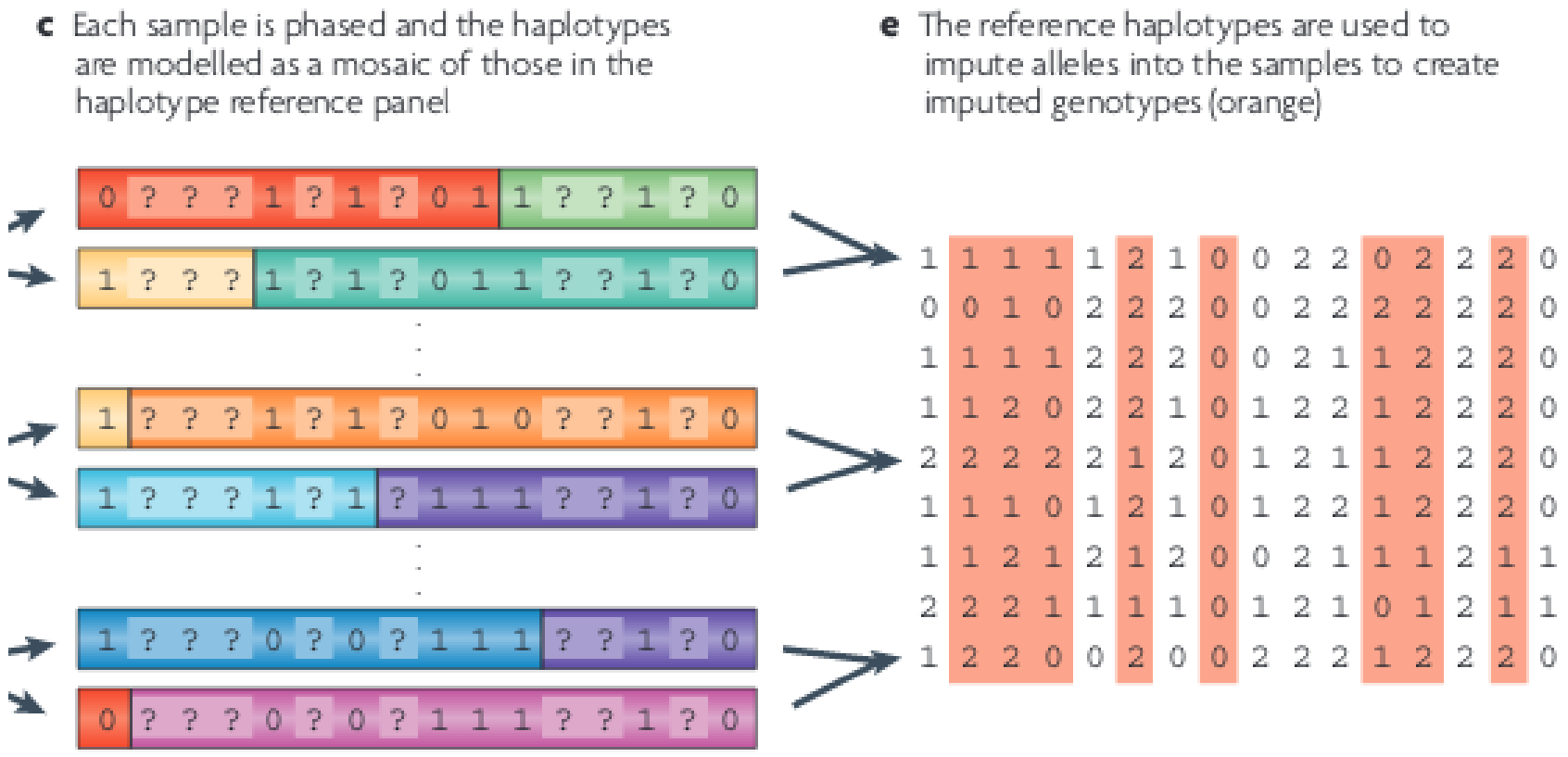

a Genotype data with missing data at untyped SNPs (grey question marks)
c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel
1 ? ? ? 1 ? 1 ? 0 2 2 ? ? 2 ? 0
0 ? ? ? 2 ? 2 ? 0 2 2 ? ? 2 ? 0
1 ? ? ? 2 ? 2 ? 0 2 1 ? ? 2 ? 0
1 ? ? ? 2 ? 1 ? 1 2 2 ? ? 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
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

Reference Panels

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)

Imputation

Output of imputation is not genotypes, but genotype probabilities:

P(AA) = 0.92
P(Aa) = 0.07
P(aa) = 0.01

Can either use probabilies directly in GWAS analysis of convert to “best guess” genotype

Also given a measure of imputation accuracy - “info score”
→ common to only include SNPs of “high” imputation accuracy in final analysis (info > 0.8 or 0.3)

And onto GWAS!

Next time...