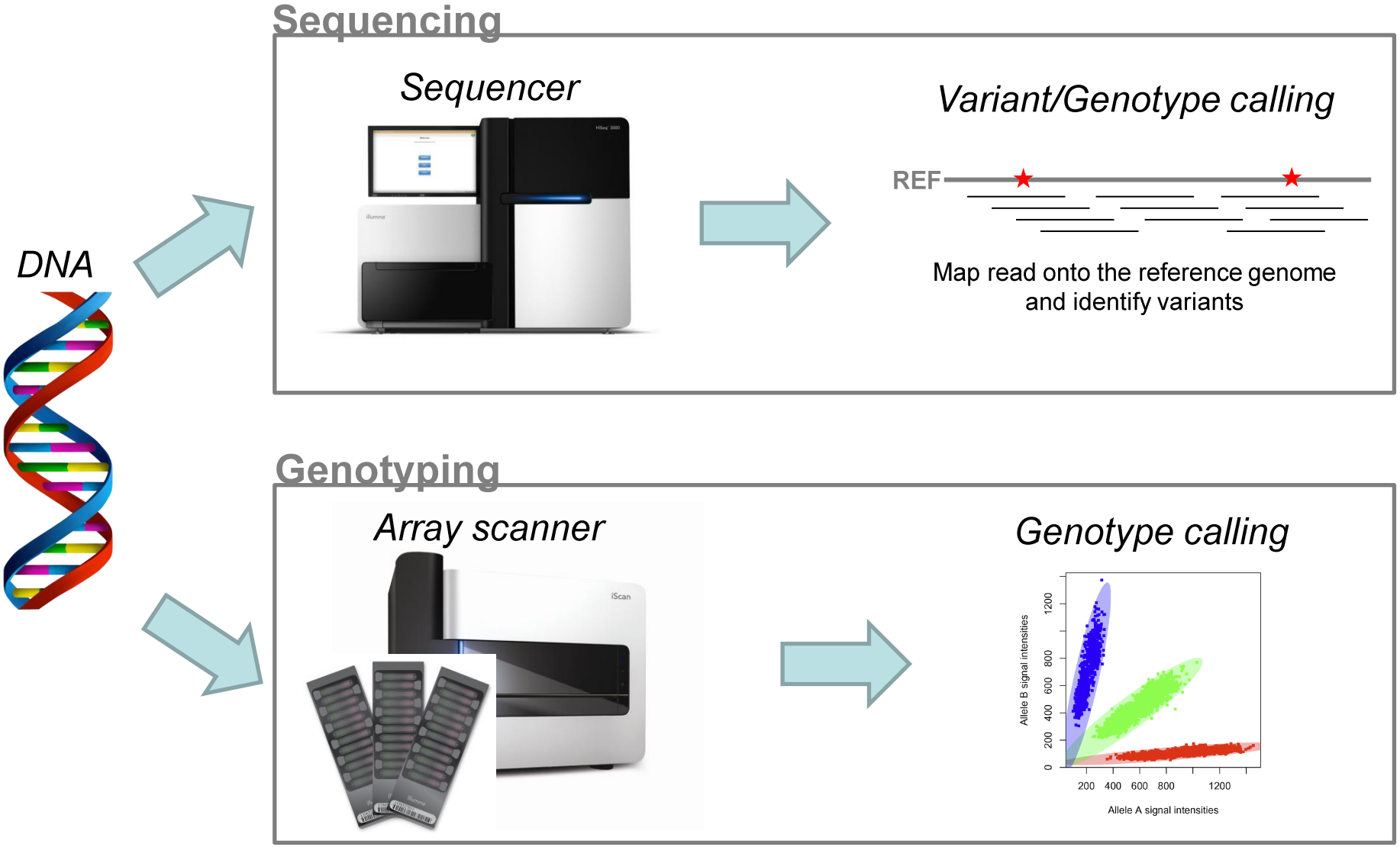
Data QC and Exploratory Data Analysis

Olivier Delaneau
University of Geneva
21/04/2018

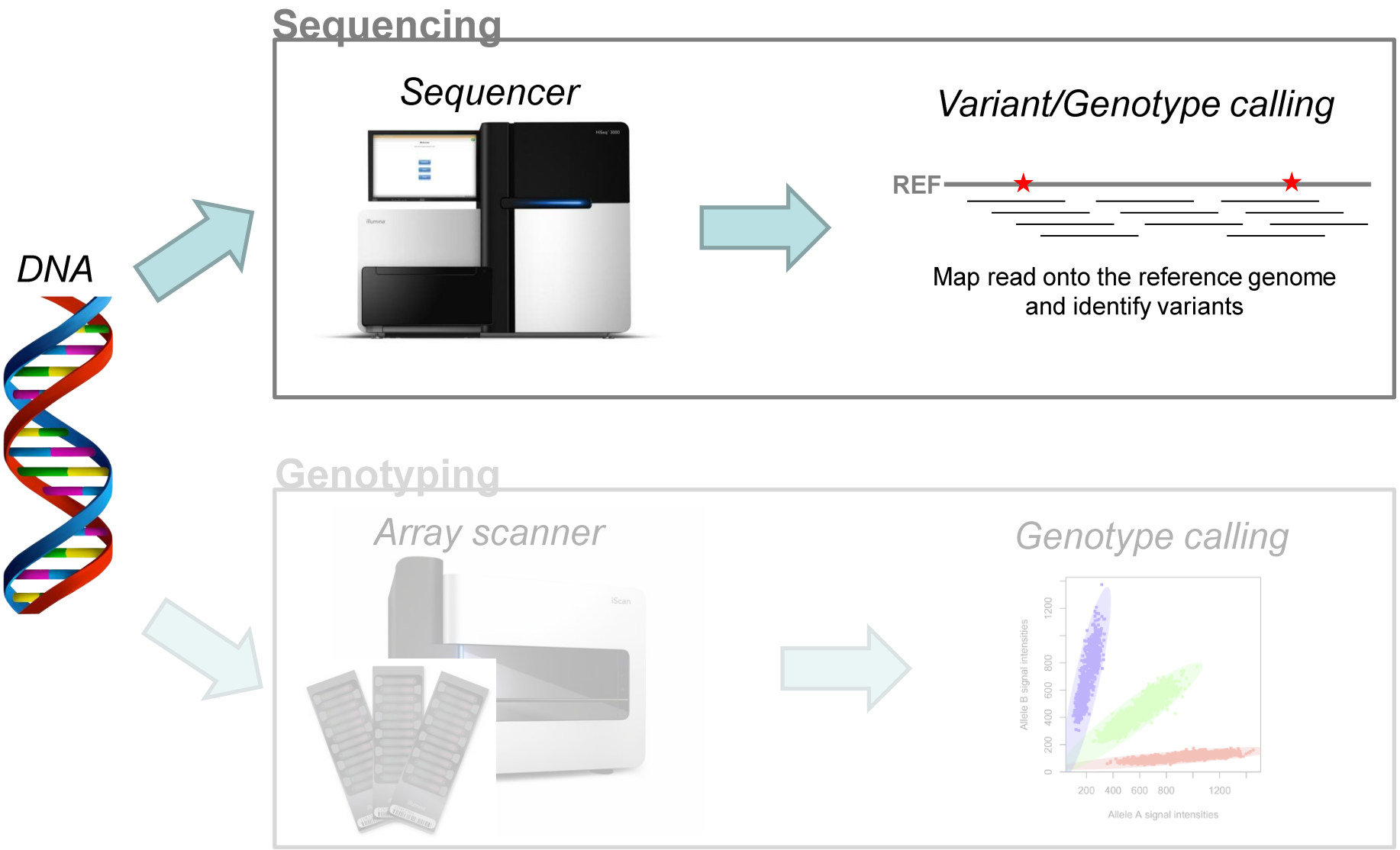
Outline

- Methods for genotyping
- File format for genotype data and standard data management tools
- QC at the genetic variant level
- QC at the sample level
- Population structure
- Information on the practical

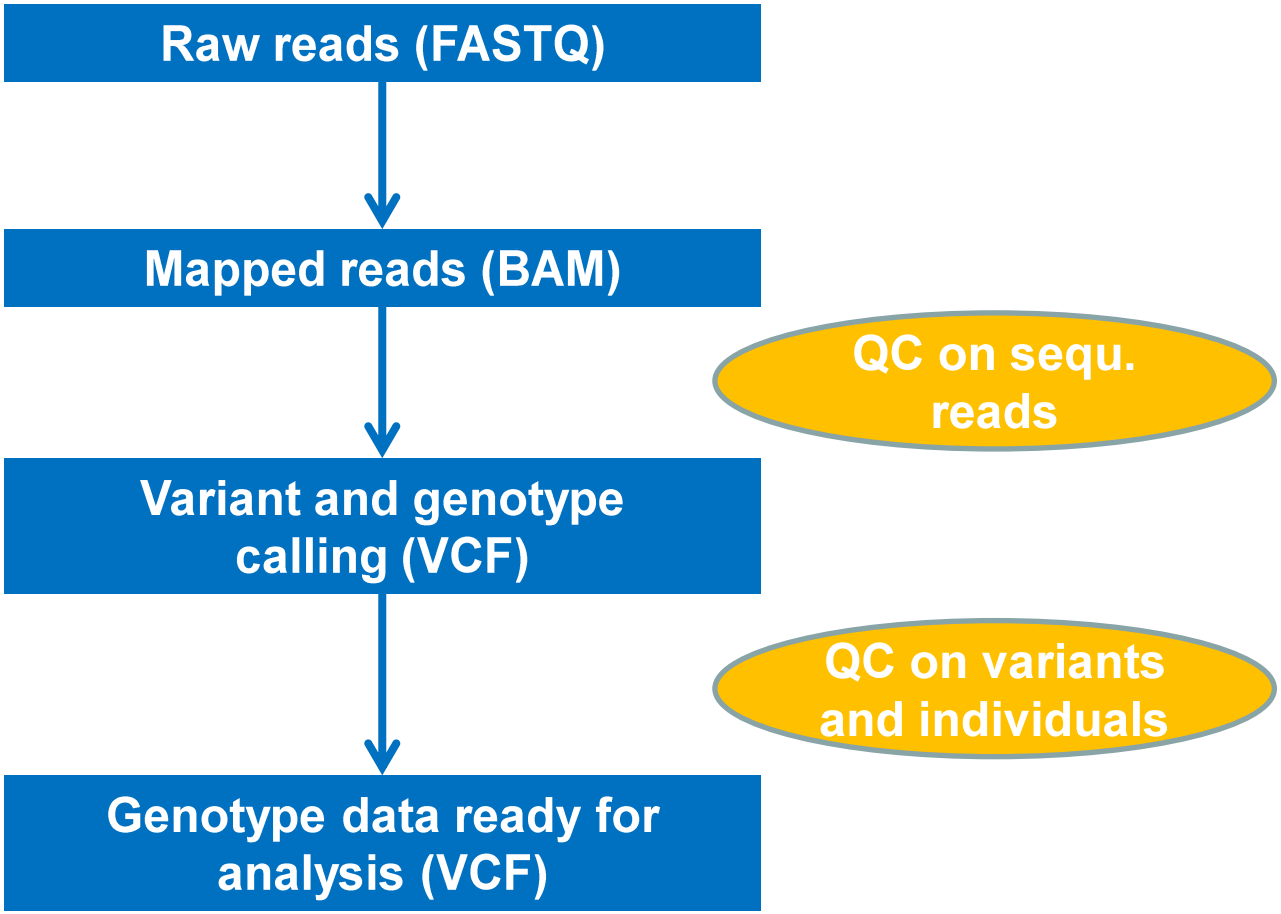
How to get genotype data?



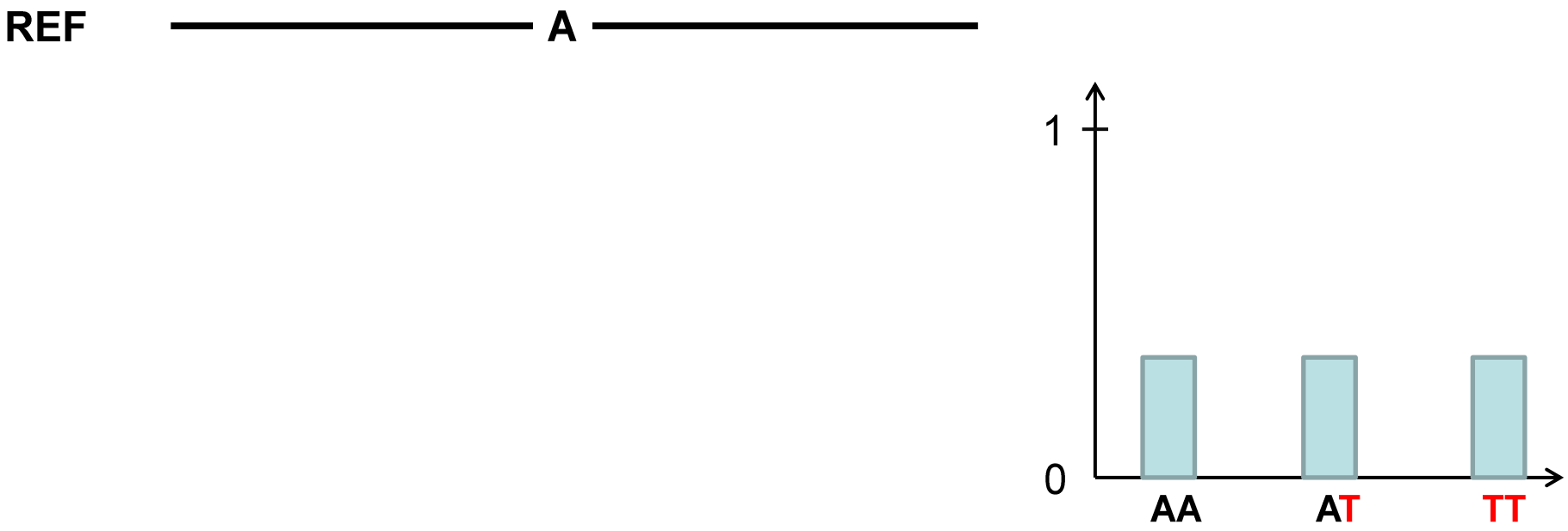
How to get genotype data?



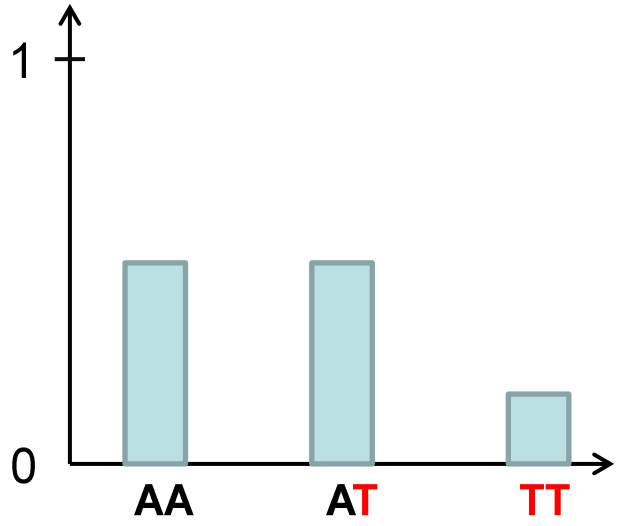
Calling genetic variations



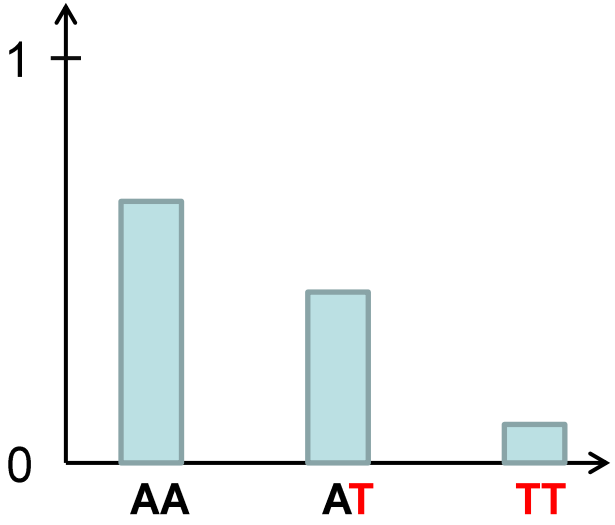
Computing genotype likelihoods



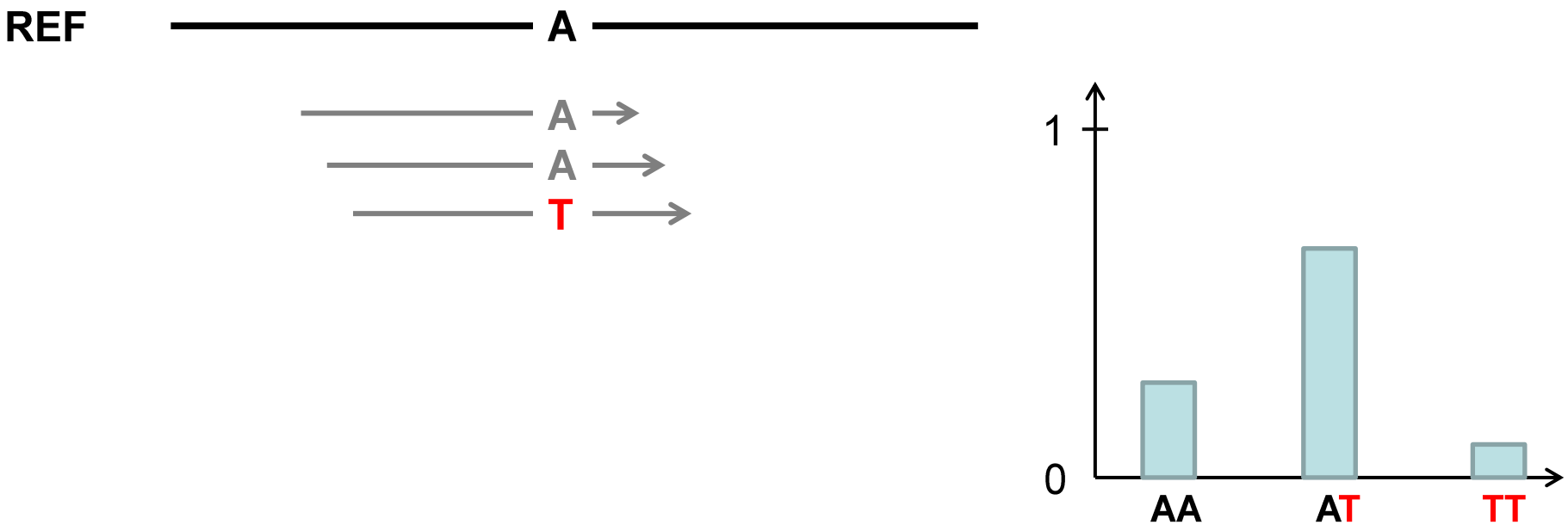
Computing genotype likelihoods



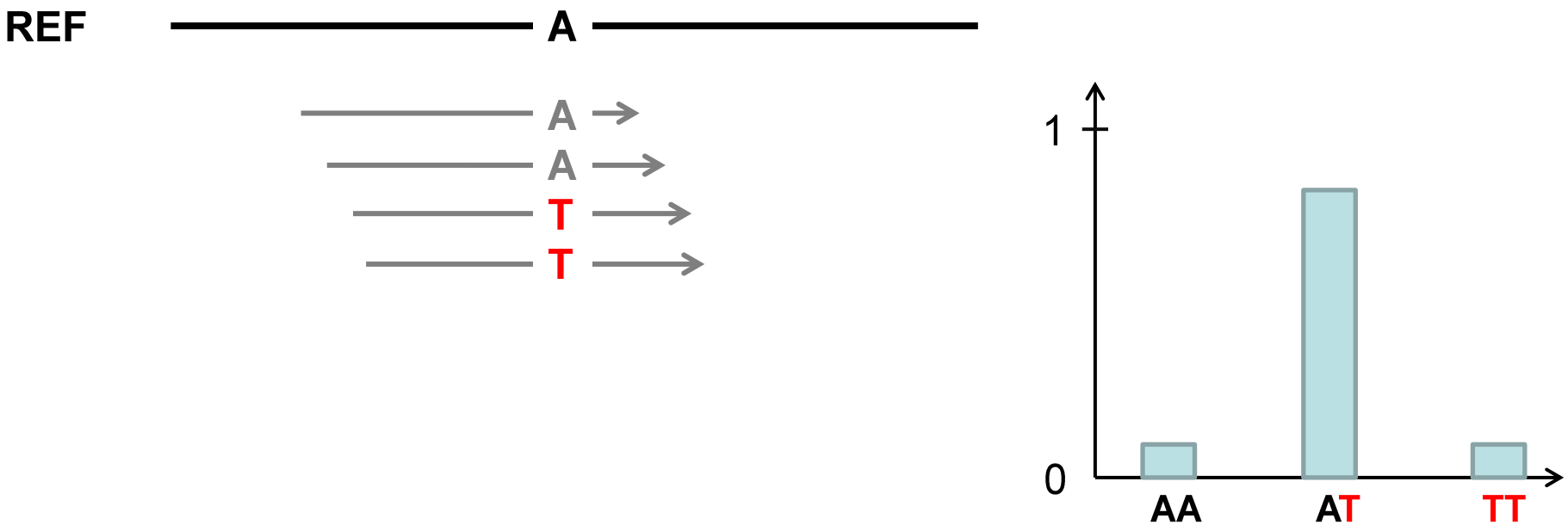
Computing genotype likelihoods



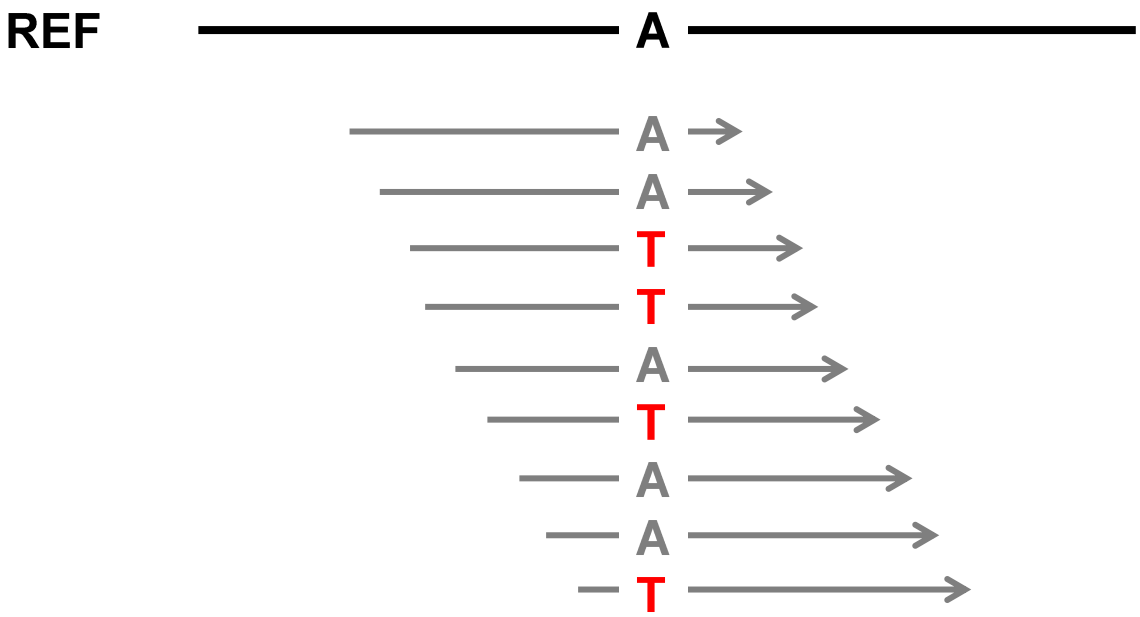
Computing genotype likelihoods

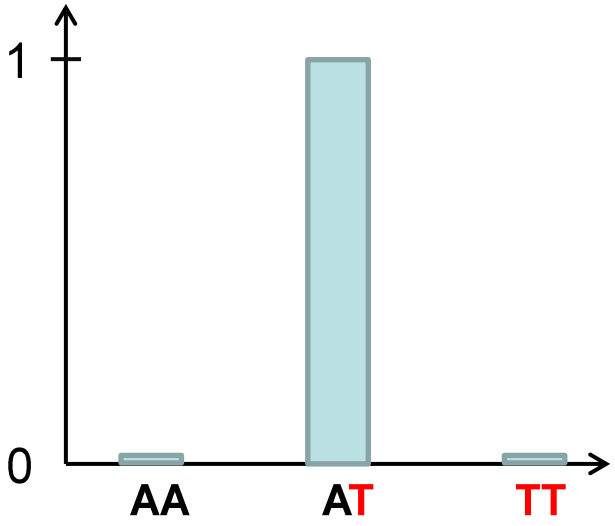


Computing genotype likelihoods

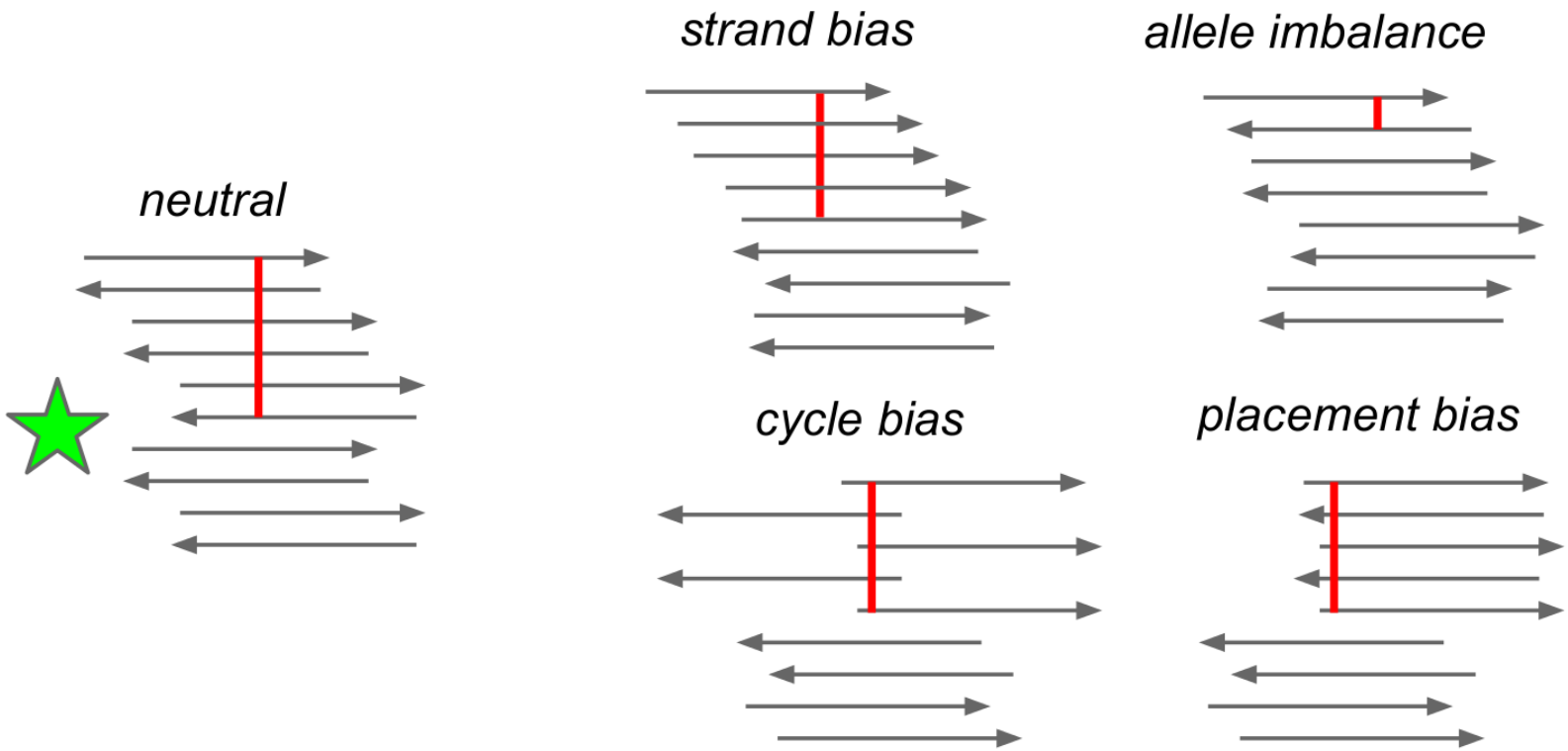


Computing genotype likelihoods





Accounting for sequencing biases



From Erik Garrison

These biases are either modeled in the genotype likelihoods (e.g. *FreeBayes*) or variants can be filtered out after calling (e.g. *samtools*) by statistical testing.

Refining genotype calls

$$P(R \mid G)$$

Genotype likelihood

Refining genotype calls

$$P(G \mid R) \sim P(R \mid G) \times P(G)$$

Genotype probability

Genotype likelihood

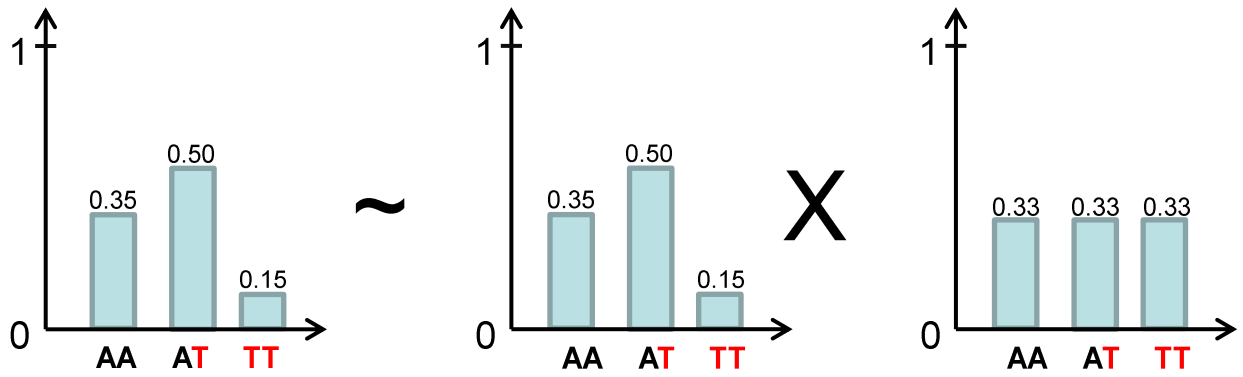
Prior (Uniform, Frequency or LD)

Refining genotype calls

$$P(G \mid R) \sim P(R \mid G) \times P(G)$$

Genotype probability — *Genotype likelihood* — *Prior (Uniform, Frequency or LD)*

<u>Uniform prior:</u>

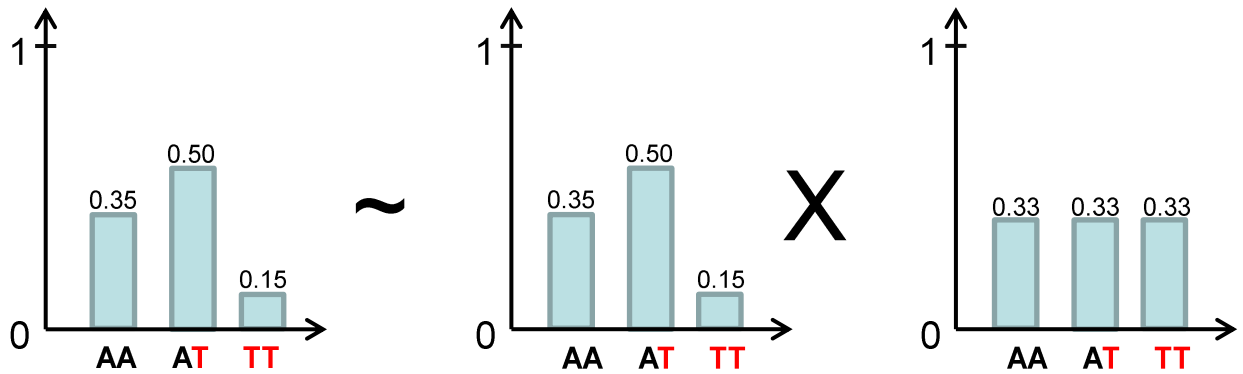


Refining genotype calls

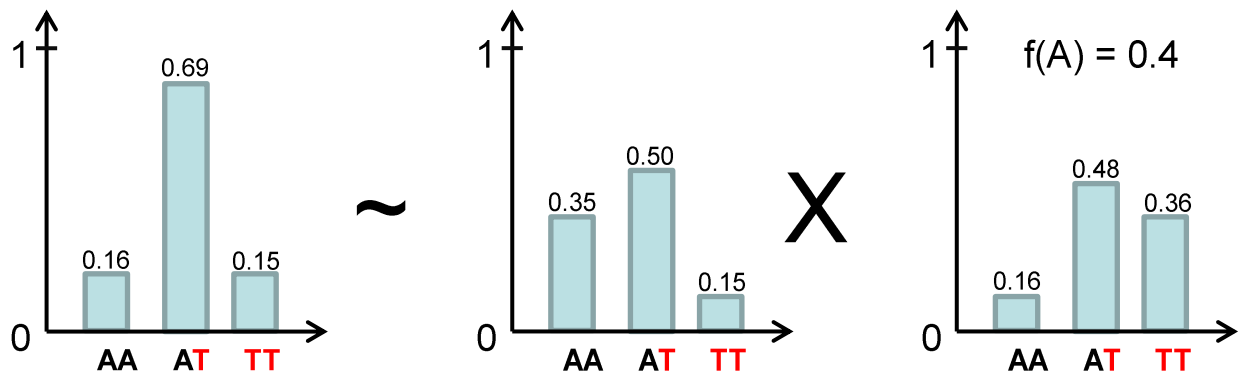
$$P(G \mid R) \sim P(R \mid G) \times P(G)$$

Genotype probability — *Genotype likelihood* — *Prior (Uniform, Frequency or LD)*

<u>Uniform prior:</u>

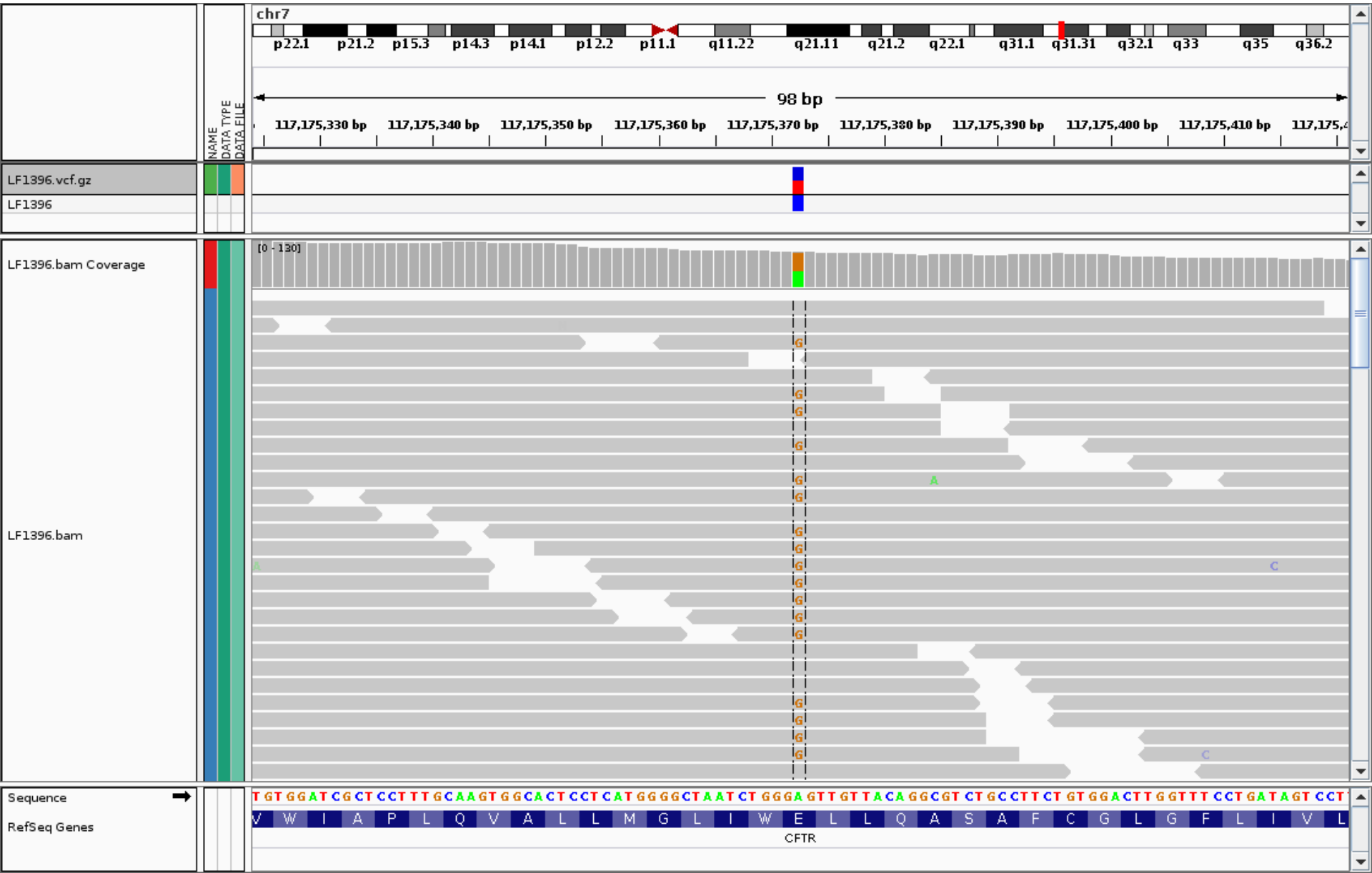


<u>Frequency prior:</u>
(samtools)

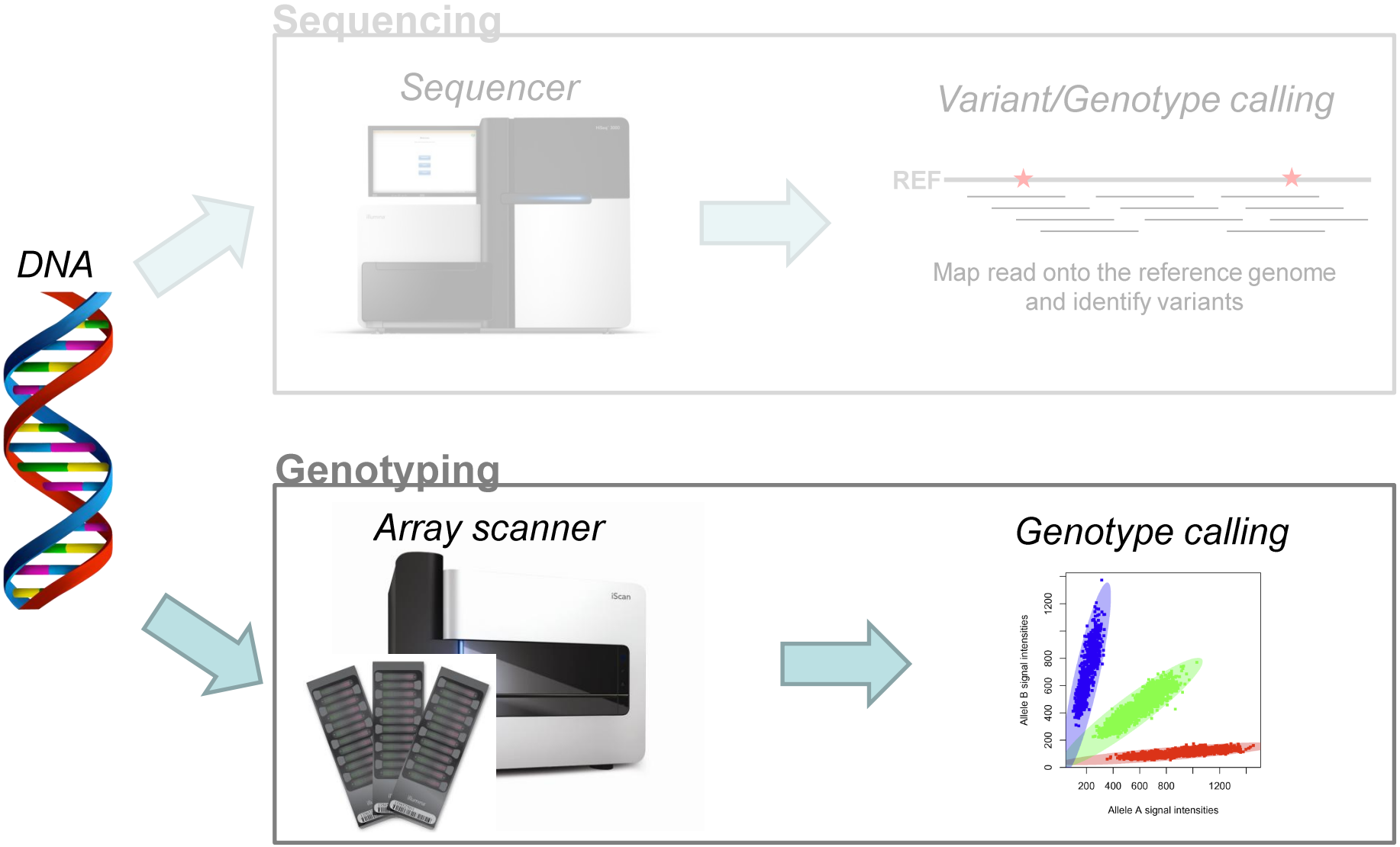


Integrative Genomics Viewer

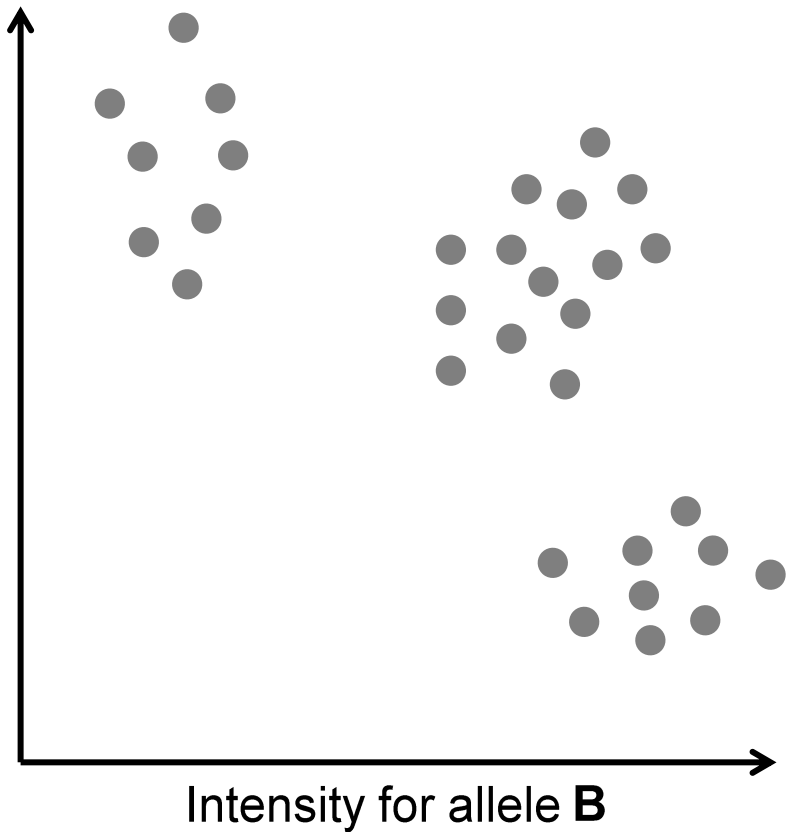
http://software.broadinstitute.org/software/igv/



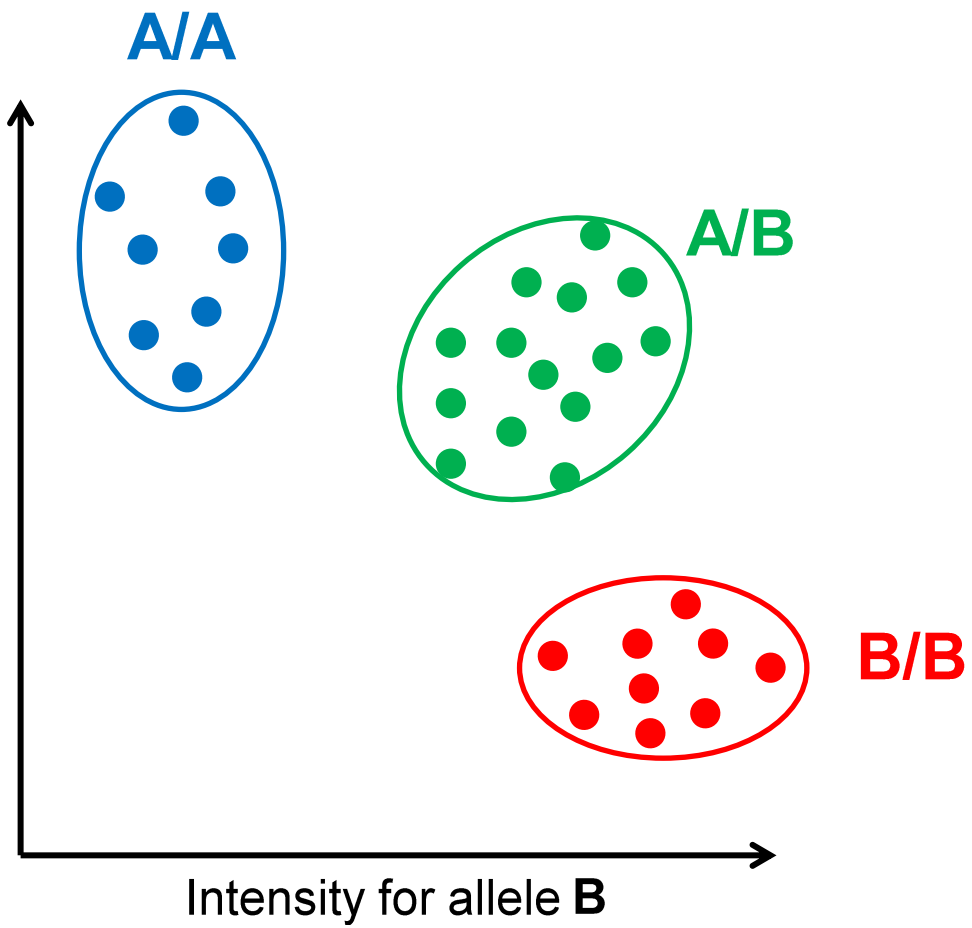
How to get genotype data?



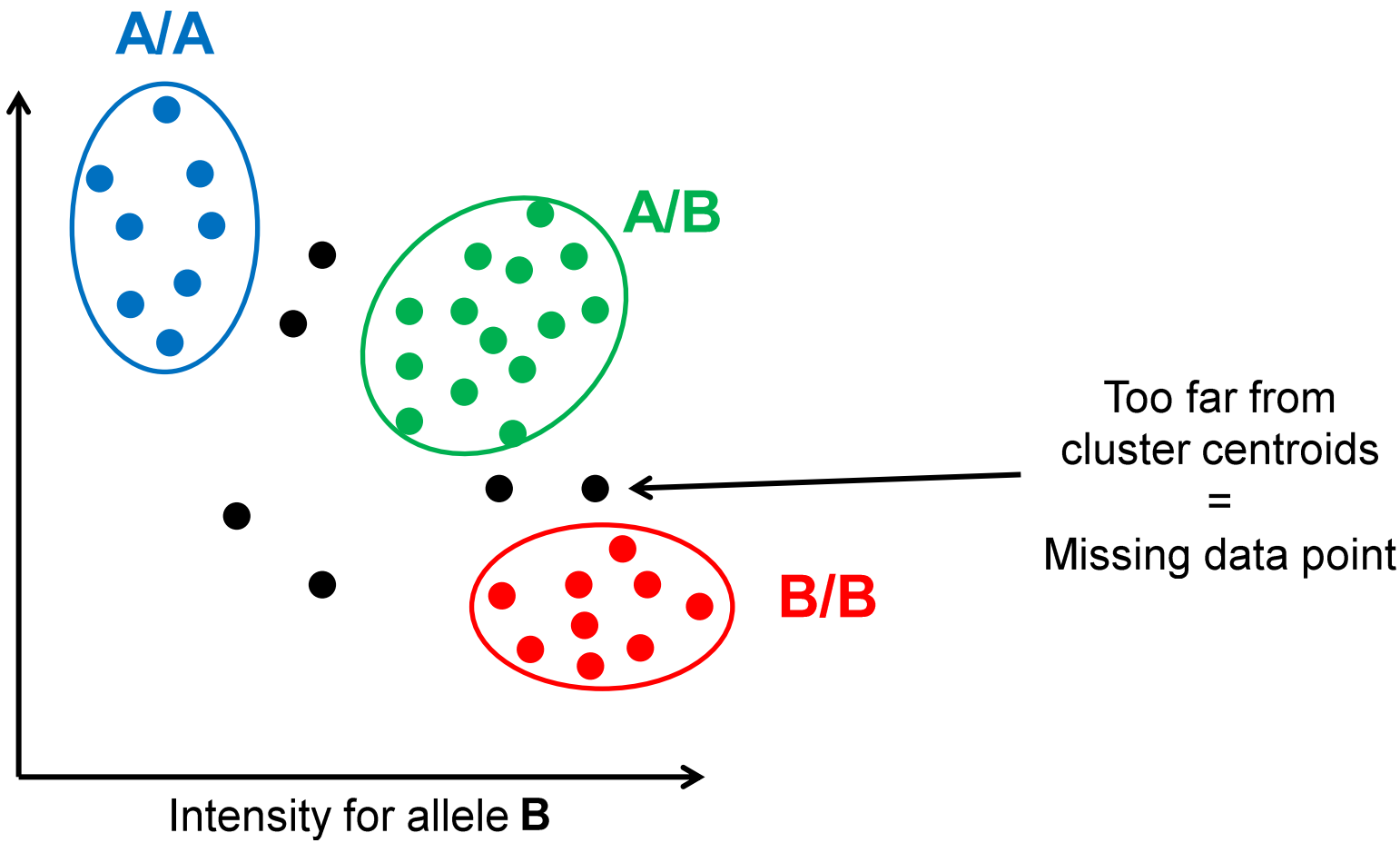
Genotype calling from chips



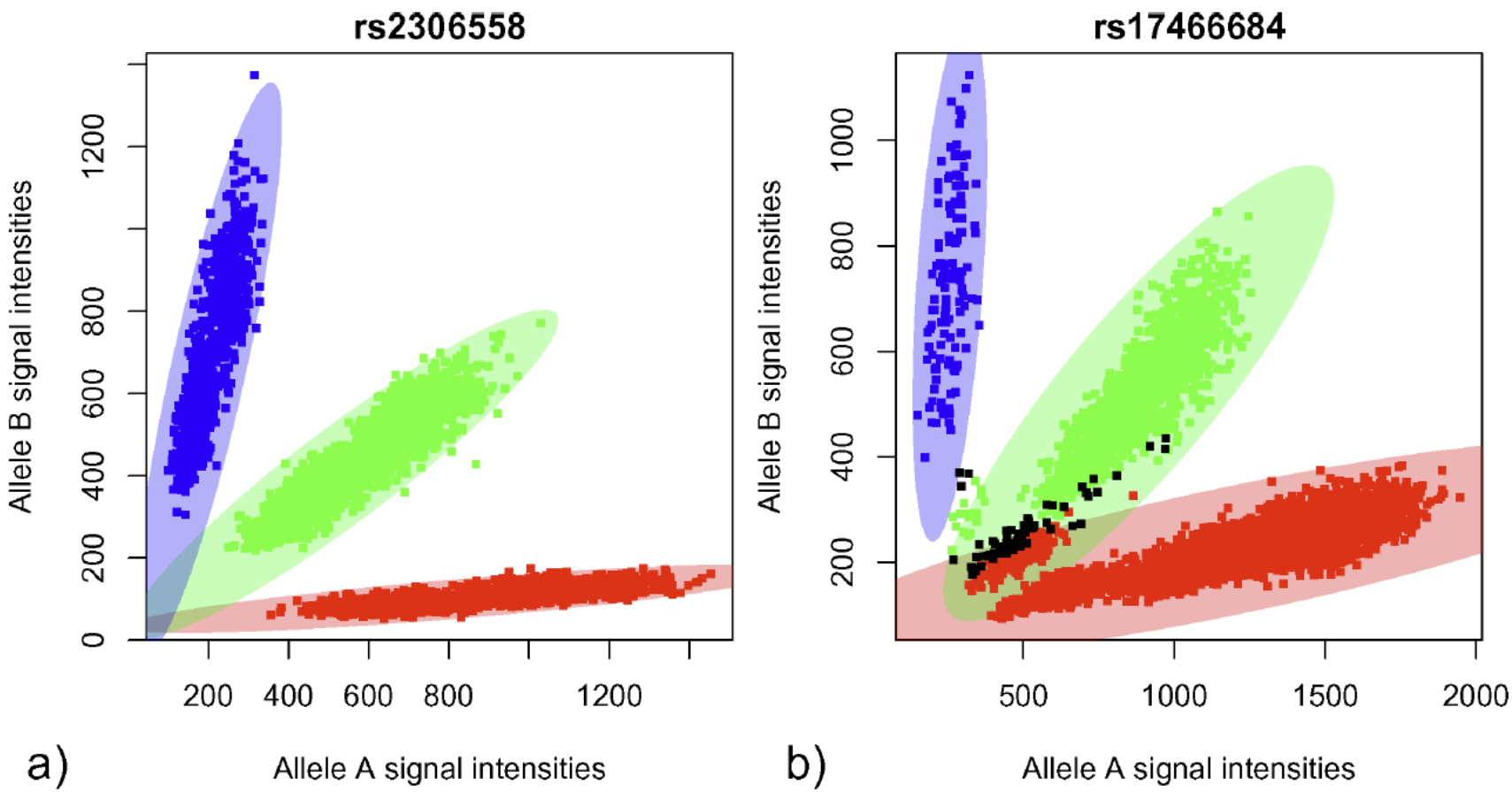
Genotype calling from chips



Genotype calling from chips

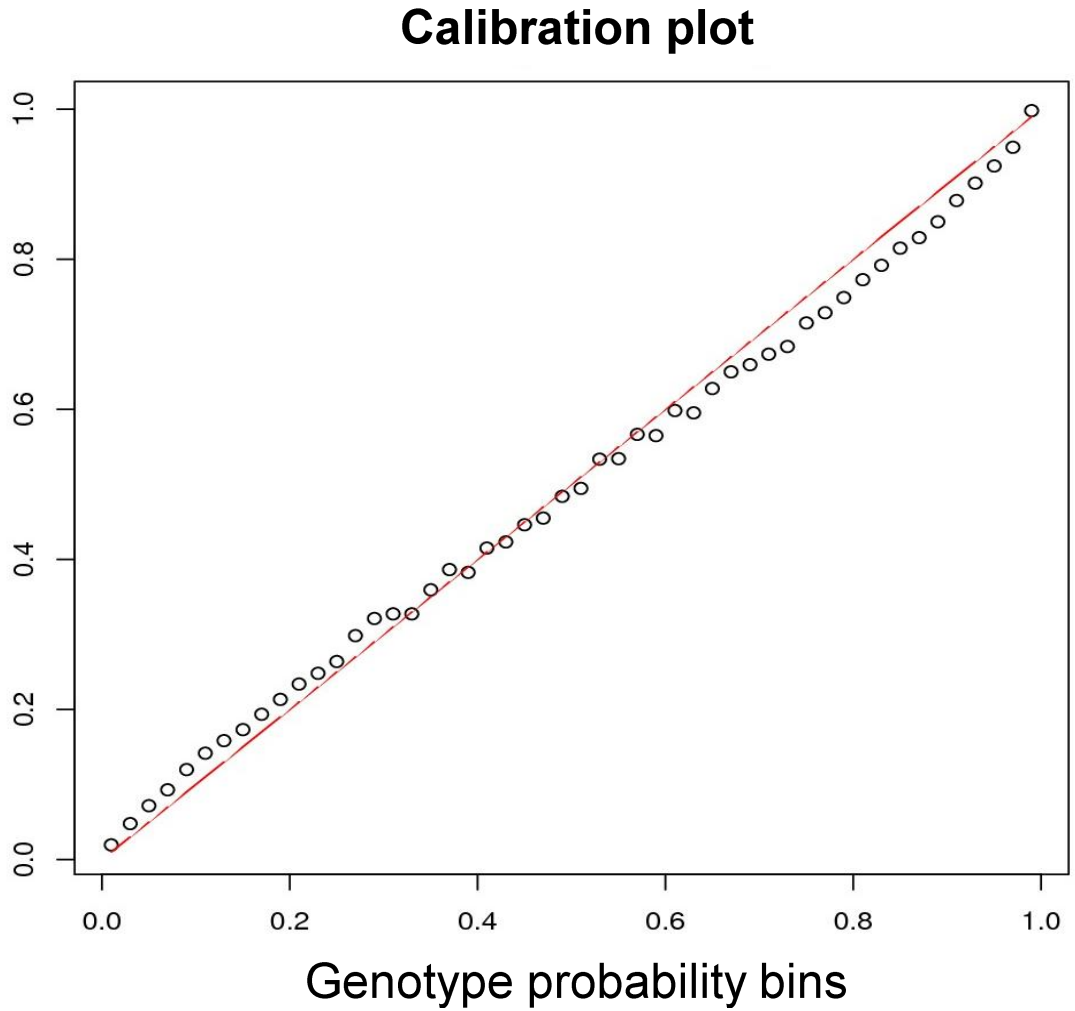


Genotype calling from chips



Approaches to call genotypes from intensities (GenomeStudio, APCA or Chiamo)
Automated approaches to call genotypes at millions of SNPs prior to the analysis.
Any finding need to be checked manually and carefully.

What does genotype probability mean?



We usually set as missing all genotypes with low prob. The threshold we use depends on the study design.

Can we afford some incorrect calls? Two scenarios:

1. Interest in global patterns, a value of 0.95 is okay,
2. Interest in some particular variants, a value 0.999 is more adapted.

Variant Call Format

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##simulateGenotypeData=1.1
##source=simulateGenotypeDataFrom1000G
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##contig=<ID=20>
#CHROM  POS     ID        REF     ALT     QUAL    FILTER  INFO    FORMAT  HG01124 HG01363 HG01500 HG01501 HG01503
20      61651   rs76553454      C       A       100     PASS    .       GT      0/0     0/0     0/0     0/0
20      63231   rs6076506       T       G       100     PASS    .       GT      0/1     0/0     0/0     0/0
20      63244   rs6139074       A       C       100     PASS    .       GT      0/0     0/0     0/0     0/0
20      63799   rs1418258       C       T       100     PASS    .       GT      0/0     0/0     0/0     0/0
20      68749   rs6086616       T       C       100     PASS    .       GT      0/1     1/1     0/1     1/1
20      69094   rs6039403       G       A       100     PASS    .       GT      0/0     0/0     1/1     0/0
20      69408   rs17685809      C       T       100     PASS    .       GT      0/0     0/0     0/0     0/0
20      71093   rs6040395       G       A       100     PASS    .       GT      0/0     0/0     0/0     0/0
20      74347   rs6135141       G       A       100     PASS    .       GT      0/0     0/0     0/1     0/1
20      75254   rs892665        C       A       100     PASS    .       GT
20      80655   rs2196239       A       G       100     PASS    .       GT
20      83611   rs114000219     C       A       100     PASS    .       GT
20      87112   rs34383360      G       A       100     PASS    .       GT
20      87416   rs1935386       A       C       100     PASS    .       GT      0/1     1/1     1/1     0/1
20      88155   rs78720032      T       C       100     PASS    .       GT      0/0     0/0     0/0     0/0
20      90814   rs75507632      T       C       100     PASS    .       GT      0/0     0/0     0/0     0/0
20      90984   rs1077784       A       G       100     PASS    .       GT      0/0     0/1
20      91088   rs2002370       C       T       100     PASS    .       GT      0/1     1/1
20      91346   rs11698455      G       A       100     PASS    .       GT      0/0     0/1
20      91508   rs6051659       G       A       100     PASS    .       GT      0/1     1/1
20      92366   rs13039134      A       G       100     PASS    .       GT      0/1     0/1     1/1     0/1
20      96931   rs6052070       A       G       100     PASS    .       GT      0/0     0/1     0/0     0/0
20      97122   rs6515824       C       T       100     PASS    .       GT      0/0     0/1     0/0     0/0
20      98930   rs6116135       G       A       100     PASS    .       GT      0/0     0/1     ./.     0/0
20      100505  rs6037772       T       C       100     PASS    .       GT      0/0     0/0     0/0     0/0
```

HEADER

Sample IDs

HG01124 has genotype **TC** at rs6086616

Missing genotype

Chromosome

Position

Variant

Reference

Alternative

Quality

Filter

Annotation

Genotype field

Genotypes

Basic operations on VCF

- BCFtools (www.htslib.org)
  - Efficient data management
  - Excellent API to develop new tools
- VCFtools (vcftools.sourceforge.net)
  - Standard data analysis (frequency, LD, etc…)
- PLINK1.9 (www.cog-genomics.org/plink2)
  - Efficient implementation of all PLINK functionalities (e.g. association testing)

The goal of **Q**uality **C**ontrol

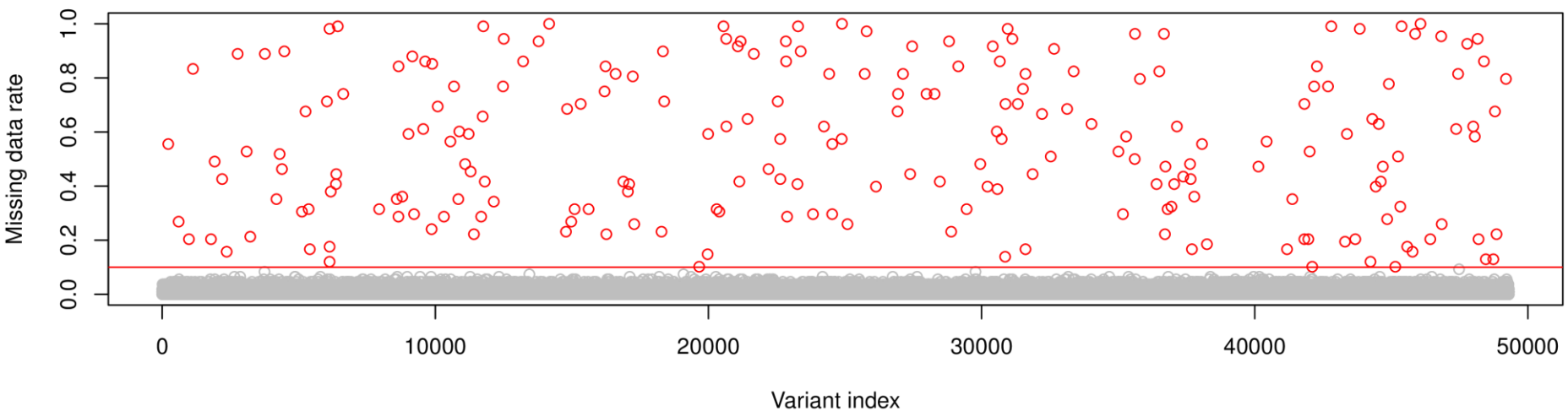
- Errors in genotype calling can introduce systematic bias in downstream analysis and can increase the chance of false discoveries
- Assess data quality to remove sub-standard genotypes, individuals and variants from subsequent analysis
- To do so, there are standardized QC procedures implemented in well-established software packages

Filter variants with low call rates

Apply threshold to genotype probabilities in order to call genotype, otherwise treated as missing. The choice of calling threshold will impact results:

- Too low: include poor quality genotypes.
- Too high: unnecessarily remove high quality genotypes, or may introduce bias by preferentially calling specific genotypes.

Variants with poor call rates (i.e. high missing data rates) likely result from poor genotype calling. They need to be removed from the data.

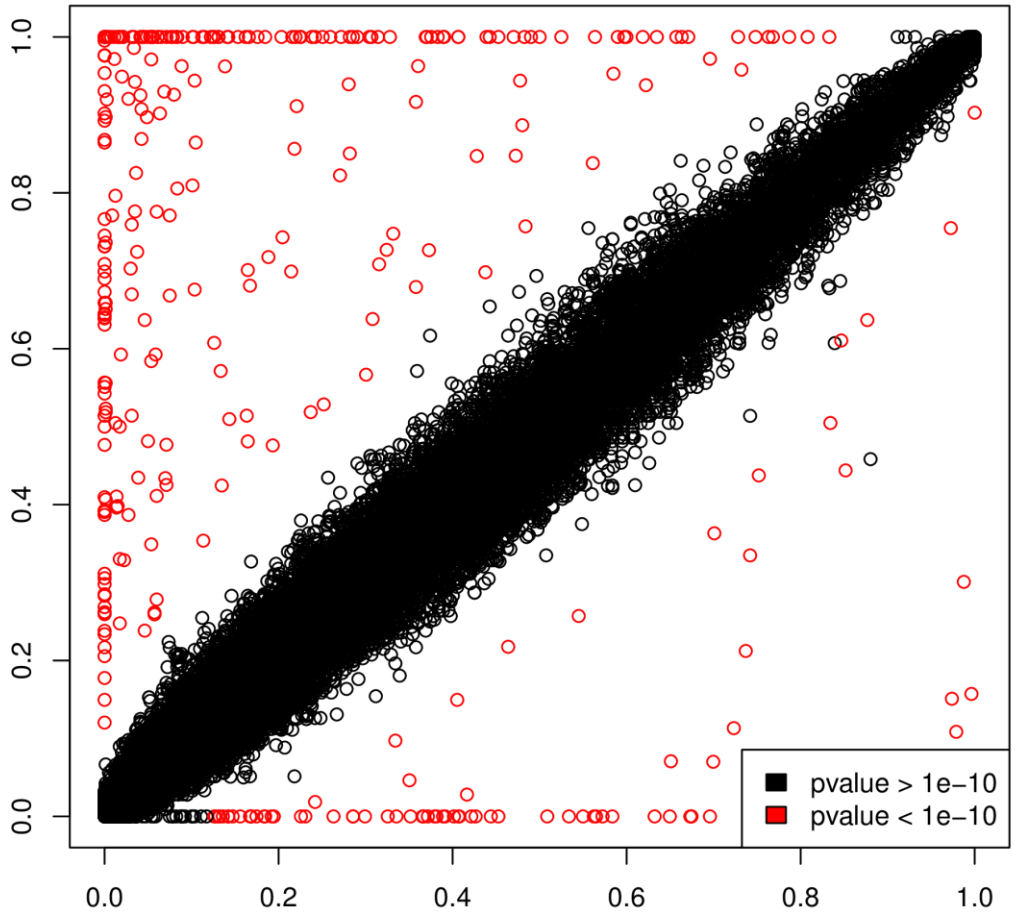


Check variant frequencies

To detect genotyping errors, one can also use large and high quality reference panels such as 1000 Genomes.

We compare the variant frequencies in our data to the frequencies in 1000 Genomes from a relevant population.

Then, one can test for massive differences and remove the corresponding variants.

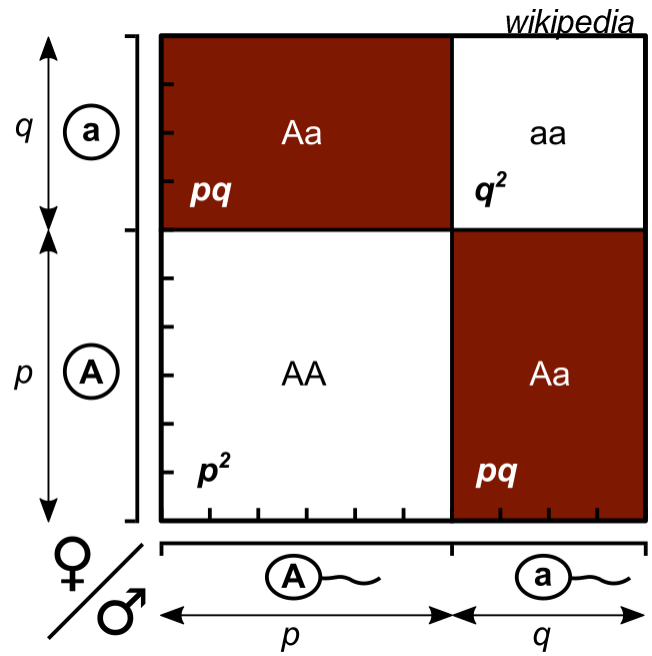


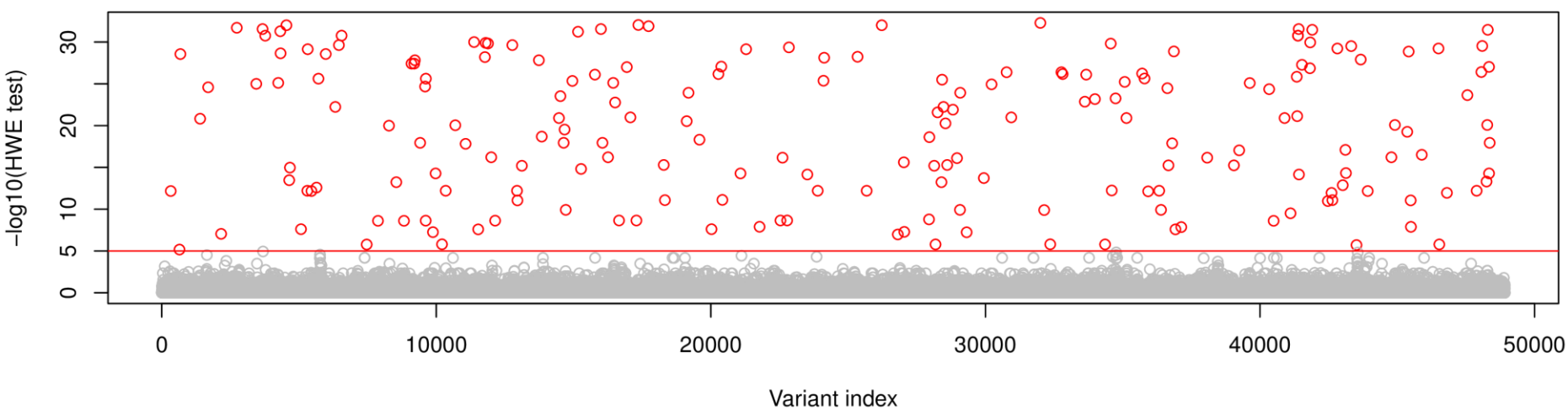
ALT frequency in 1000 Genomes

Check **H**ardy **W**einberg **E**quilibrium

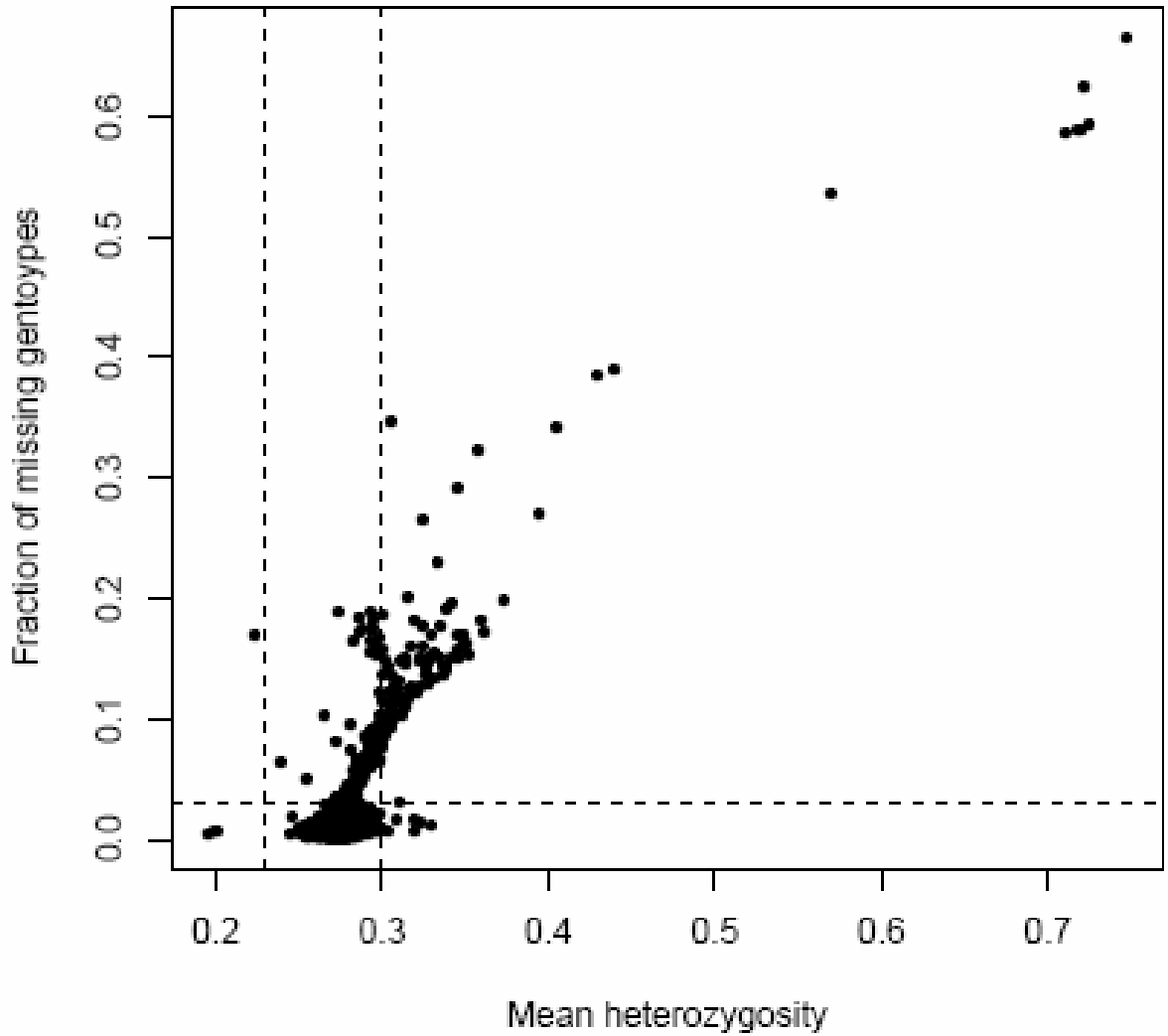
Under the assumption of random mating between individuals in a population, we have the following relationships between allele frequencies and genotype frequencies.

This can also be used to assess the quality of the genotype data by statistical testing (*Wigginton, Cutler and Abecasis, 2005)*.





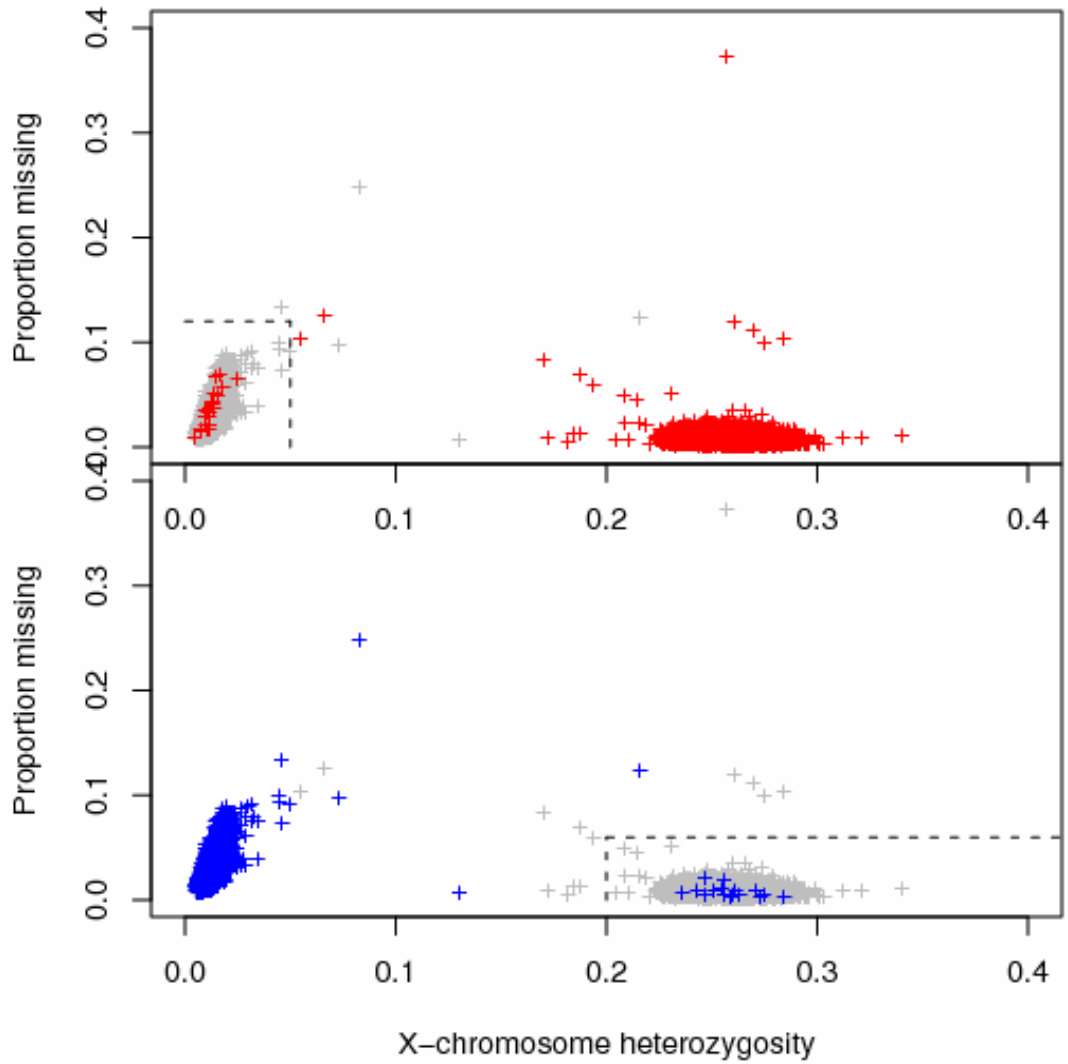
Filter samples with low call rate



Sex check with chromosome X

Each individual plotted twice according to reported sex: females in red and males in blue.

Should these samples be removed from the study or the sex corrected based on heterozygosity? May impact on results if sex is adjusted for in the analysis or if sex specific analyses are to be undertaken.



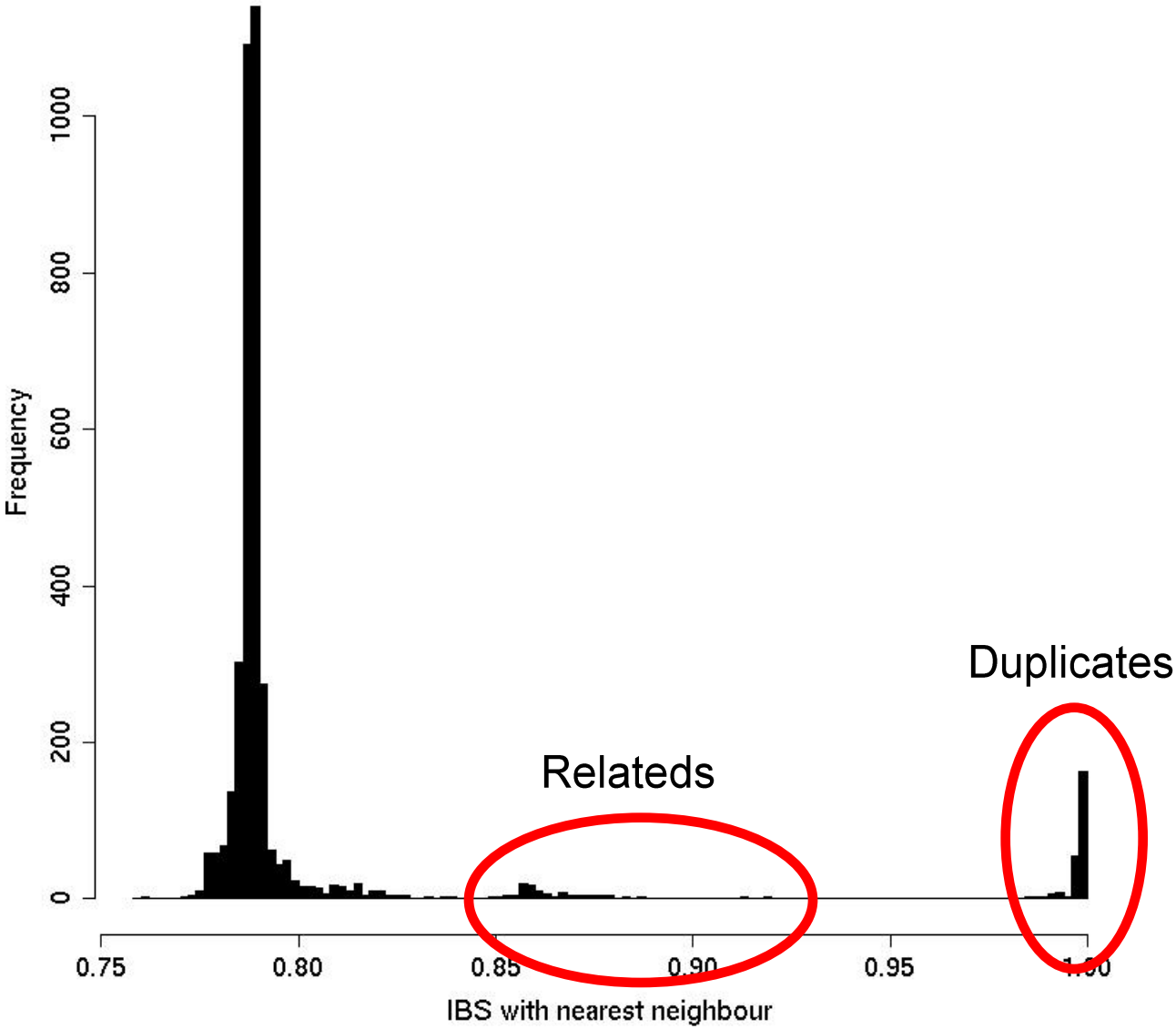
Check relatedness

- Over *M* markers, the IBS between the *i*th and *j*th individuals is given by

$$IBS_{ij} = 1 - \frac{1}{2M}\sum_{k}\left|G_{ik} - G_{jk}\right|$$

  where G_{ik} and G_{jk} denote the number of minor alleles (0, 1 or 2) carried by the *i*th and *j*th individuals at variant *k*.
- Identical samples will share IBS near to 100% (allowing for genotyping errors).
- Related individuals will share higher IBS than unrelated individuals.

Check relatedness

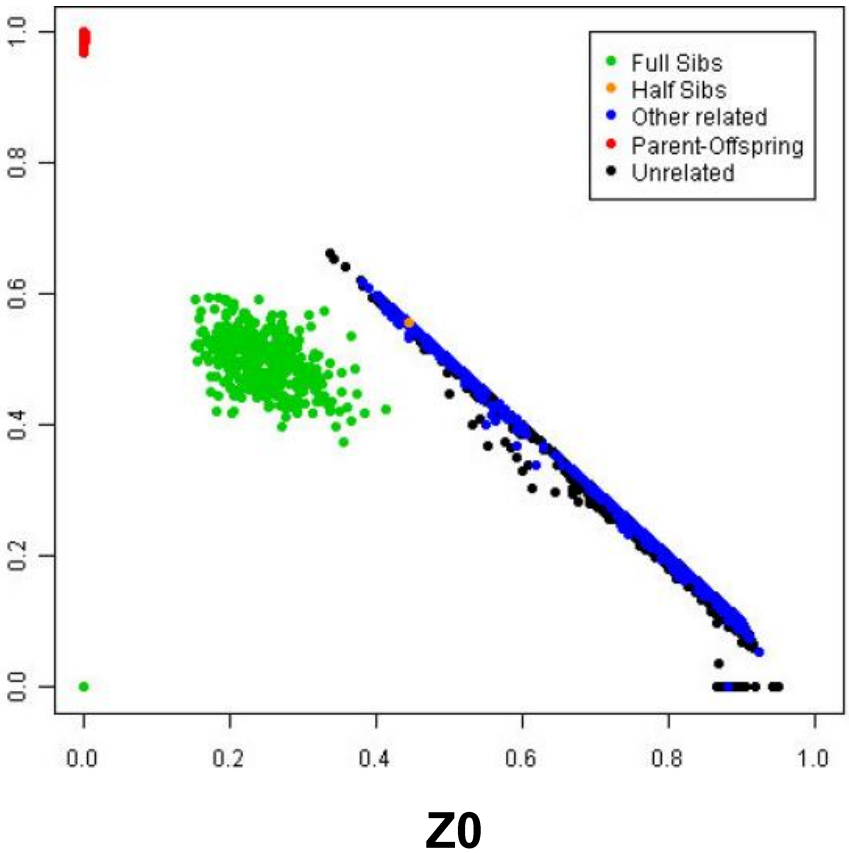


Alternative checks for relatedness

IBS can be used to estimate the proportion of the genome at which a pair of individuals share 0, 1 or 2 chromosomes IBD "identical by descent" (denoted z0, z1, or z2).

Once relative spotted, two options:

1. Accounting for relatedness in downstream analysis,
2. Remove close relatives.



Population structure

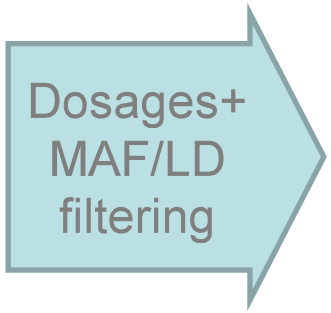
- PCA captures widespread sub-structure in allele frequencies.
  - Top PCs reflect genetic variations due to ancestry.
  - Plotting top PCs including 1000 genomes samples can be used to identify population outliers.
  - Top PCs can be used as covariates in downstream analysis to adjust for population stratification.

- Alternative methods exist such as MDS for instance as implemented in PLINK.

Population stratification

Individuals

0/0	0/0	…	0/1
0/1	0/1	…	1/1
0/0	1/1	…	1/1
…	…	…	…
0/0	0/0	0/0	0/1



Individuals

0	0	…	1
1	1	…	2
0	2	…	2
…	…	…	…
0	0	0	1

Population stratification

Individuals

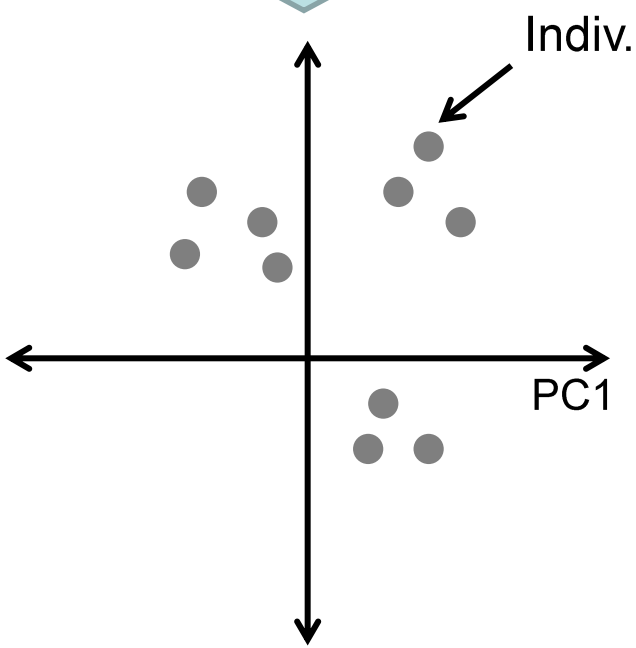
0/0	0/0	…	0/1
0/1	0/1	…	1/1
0/0	1/1	…	1/1
…	…	…	…
0/0	0/0	0/0	0/1

Dosages+ MAF/LD filtering

Individuals

0	0	…	1
1	1	…	2
0	2	…	2
…	…	…	…
0	0	0	1

PCA



Population stratification

Individuals

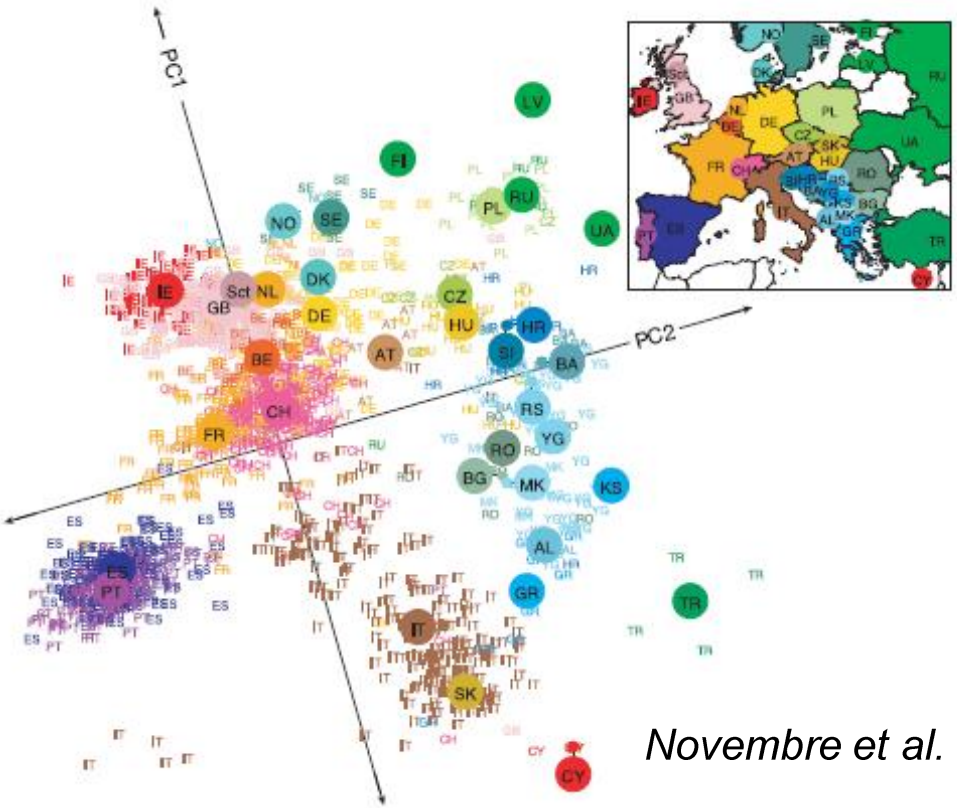
0/0	0/0	…	0/1
0/1	0/1	…	1/1
0/0	1/1	…	1/1
…	…	…	…
0/0	0/0	0/0	0/1

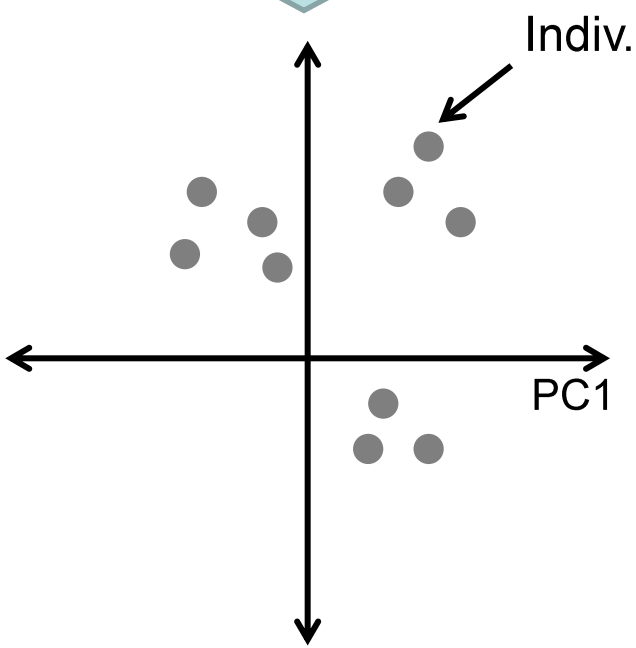
Dosages+ MAF/LD filtering

Individuals

0	0	…	1
1	1	…	2
0	2	…	2
…	…	…	…
0	0	0	1

PCA

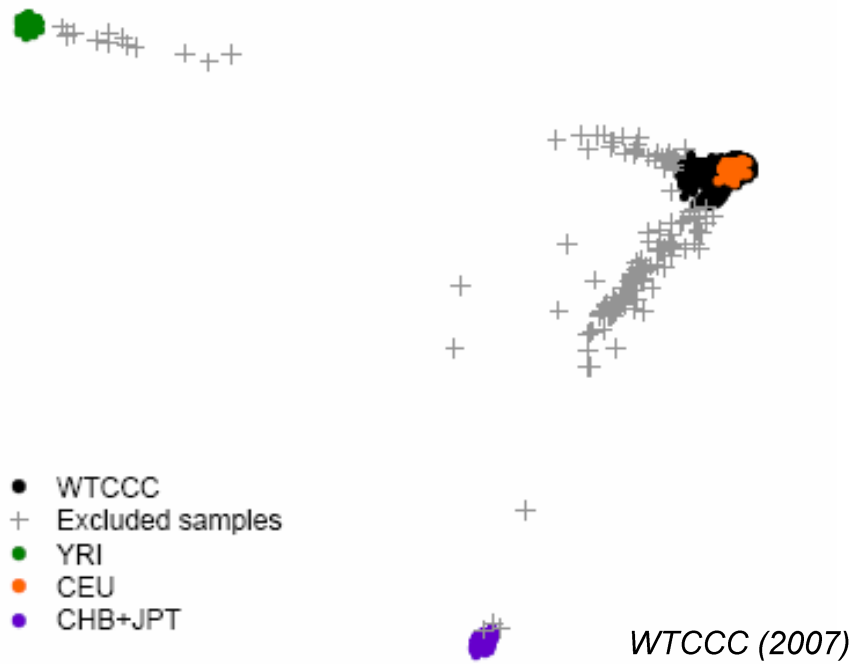




Novembre et al. (2008).

What do I do with my PCs?

1. Remove outliers from analysis



WTCCC (2007)

2. Account for structure
- By defining ethnic groups
- By using the PC loadings as covariates

Summary

- QC criteria are subjective and vary from one study to another
- Variant QC filters should eliminate the worst quality markers without “throwing the baby out with the bathwater”
- Sample QC filters should not be so stringent as to remove the majority of the analysis cohort
- Any finding should be followed up with visual inspection of the raw data (i.e. sequencing reads or cluster plots)

Practical

- The goal of the practical of this afternoon is to go from raw to ready-to-analyze data
- We will use multiple standard QC steps both at the variant and sample levels
- The resulting data set will be used as starting material for the afternoon practical
- Look at ***/home/delaneau***, everything you need is there.