Data generation and pre-processing

Workshop outline

- Introduction
- **Omic data – genetic, epigenetic and metabolite**
- Prediction – methodology
- Prediction – demonstration
- Examples and case studies

Outline

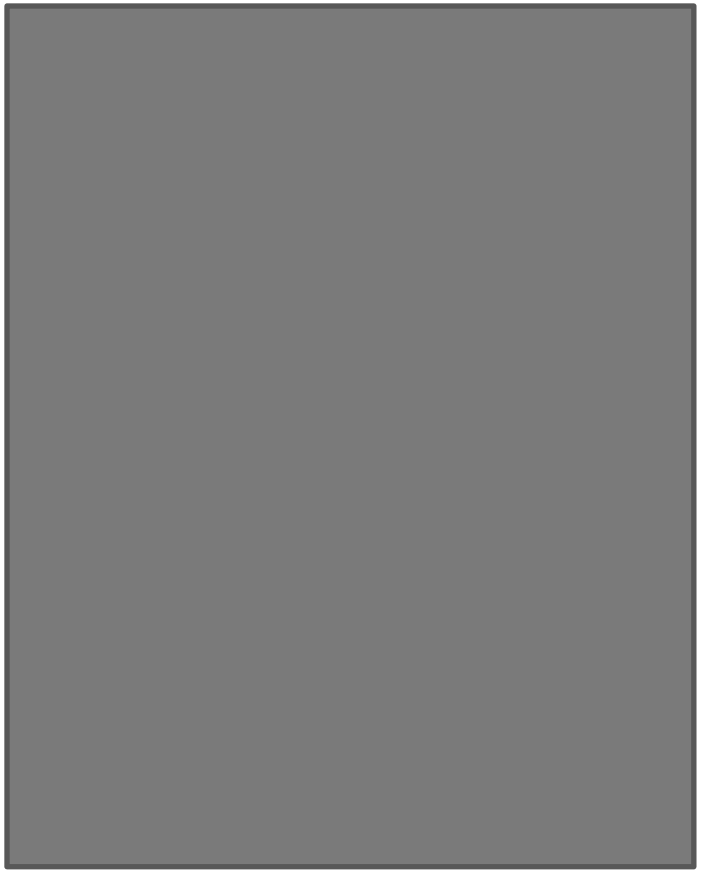
- Where the data comes from
- What the data looks like
- Steps for pre-processing
- We'll go through this for genetic, DNA methylation and metabolomic data!

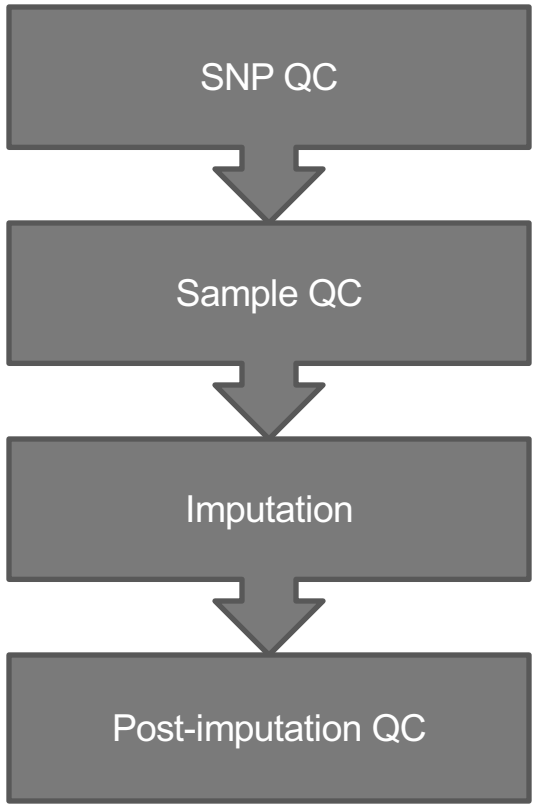
What is a genome-wide association study?

- Asking the question:
- What is the association between genetic variants and a trait?
- Looks at genetic variants across the genome (hence "genome-wide"!)

Genotyping or sequencing

Genotyping	Sequencing
Look at known genetic variants (300,000-4 million SNPs covered)	Looks at all forms of variation
Cheaper	More expensive
Common loci	Good for rare loci
Easy to analyse	More complex to analyse
Less computationally expensive	Computationally challenging




SNP QC
Sample QC
Imputation
Post-imputation QC

What your data will look like

MAP

CHR	SNP ID	Position (morgans)	position (bp)
1	rs123456	0	1234555
1	rs234567	0	1237793

BIM

CHR	SNP ID	Position (morgans)	position (bp)	A1	A2
1	rs123456	0	1234555	C	A
1	rs234567	0	1237793	G	A

Binary

FAM

Family ID	Individual ID	Paternal ID	Maternal ID	Sex	Pheno
FAM001	1	0	0	1	1
FAM002	2	0	0	2	1

PED

Family ID	Individual ID	Paternal ID	Maternal ID	Sex	Pheno	rs123456	rs123456	rs234567	rs234567
FAM001	1	0	0	1	1	A	A	G	G
FAM002	2	0	0	2	1	A	A	A	G

BED

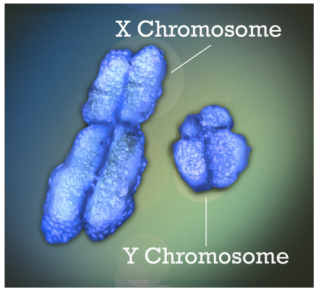
SNP QC

Missing data can arise from inaccuracy of identifying genotype – remove SNPs with excessive missingness!

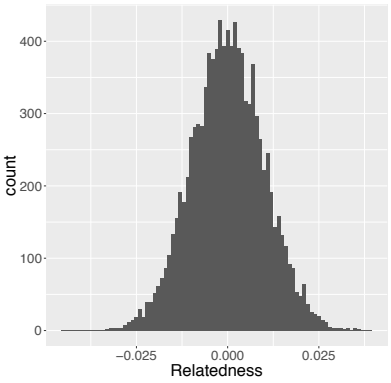
Rare alleles are generally excluded due to power, also questions about accuracy (MAF < 0.01)

Hardy-Weinberg equilibrium (HWE) → Assumed if a SNP not in HWE then it's due to a genotyping error!

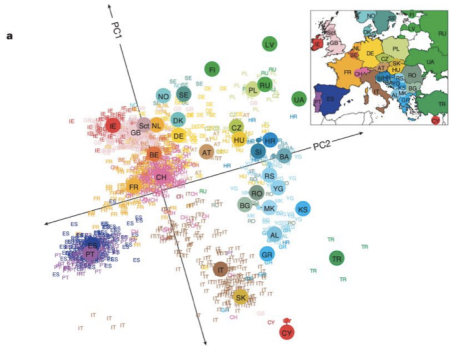
Sample QC



Should be able to identify sex from genotype → Some SNPs present on the X chromosomes are not present on the Y!

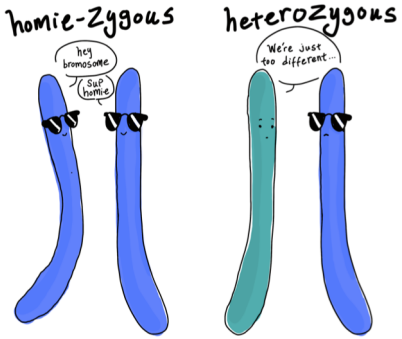


Duplicated and related individuals will bias results



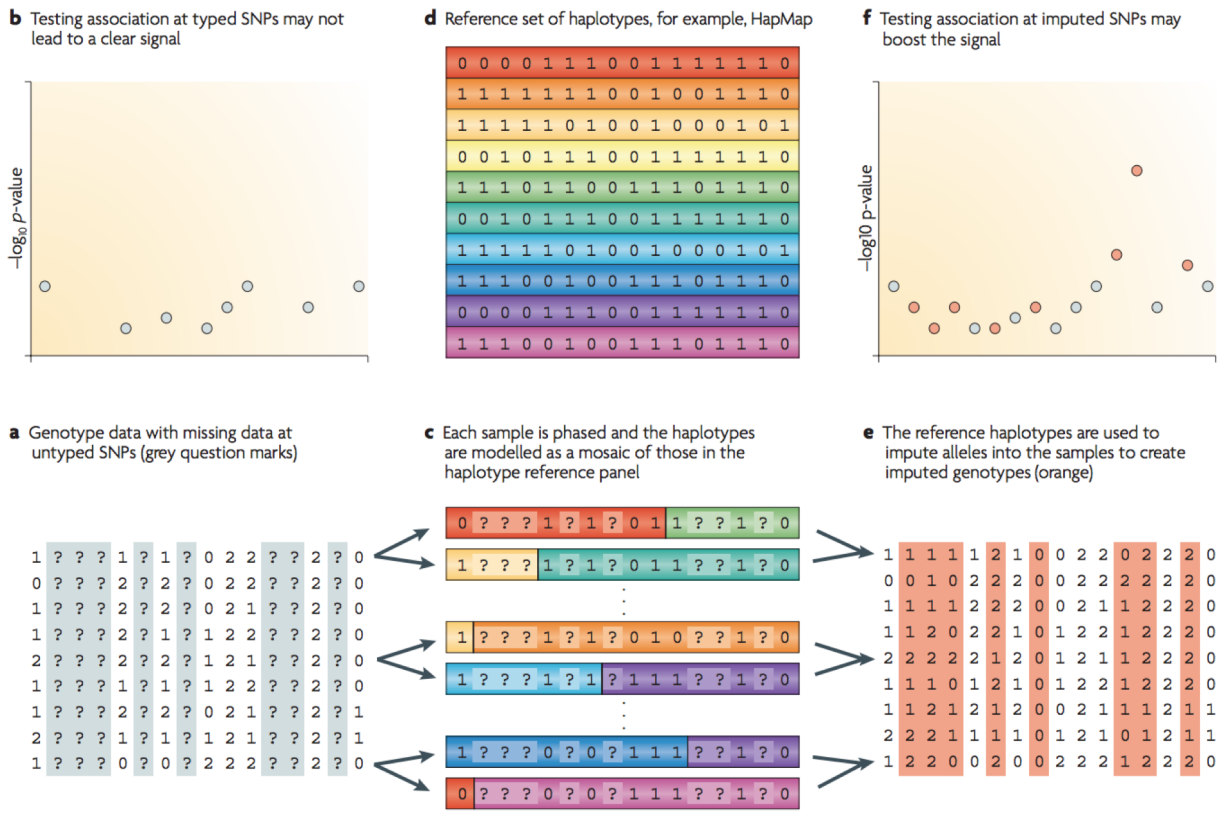
J. Novembre *et al.*, "Genes mirror geography within Europe," *Nature*,

Ancestry needs to be considered



Heterozygosity gives an indication of sample quality
Also, missingness again!

Imputation



J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies," *Nat. Rev. Genet.*

Imputation

- Choose a reference haplotype (e.g. HapMap, 1kG, HRC) → important to consider the ancestry of your sample!!
- Phase the data (also known as haplotype estimation) using one of various statistical methods
- Impute using one of the two cloud-based servers

- Post imputation: assess accuracy using concordance or info score

What's next?

- Left with:
  - ~10 million genetic variants
  - Some samples removed
  - Data in one of few formats (e.g. BED, BIM, FAM)
- Well curated software to do the analysis e.g. PLINK, SNPTEST
- Can easily convert the data into a format that can be read into R

- Guide to genome-wide association tests (including pre-processing!):
  - doi: 10.1002/mpr.1608