HelmholtzZentrum münchen
German Research Center for Environmental Health

GWAS

Tools and best practices for performing genome-wide scans

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Prof. Eleftheria Zeggini, Lorraine Southam and Dr. Konstantinos Hatzikotoulas
Institute of Translational Genomics

Presentation outline

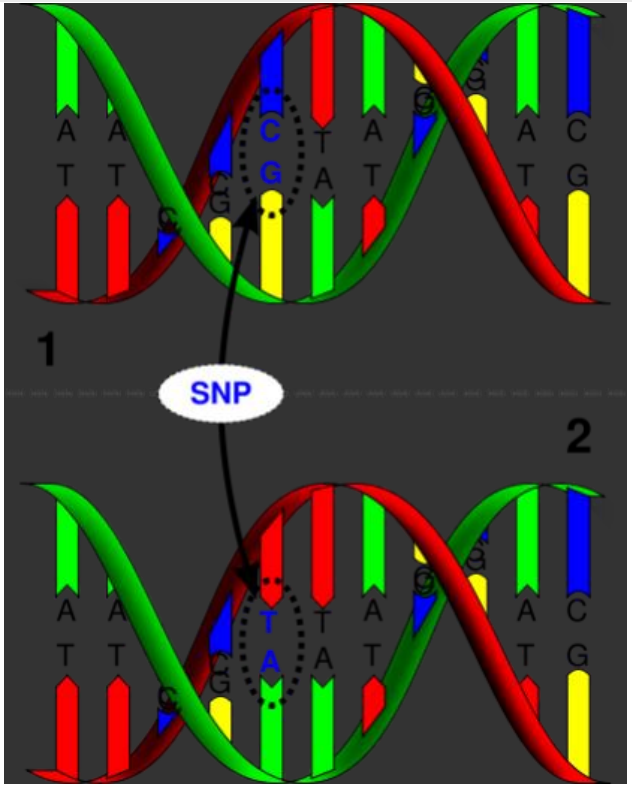
- **Some key terms and concepts** [Kostas]
- **Brief introduction to genome-wide association analysis** [Loz]
  - What input data do we need to undertake a GWAS?
  - What are our analysis options and how do we analyse?
  - How to visualise the results
  - Post GWAS analysis
  - Putting GWAS into context (with an example from ITG)
- **Overview of practical** [Loz]
- **Introduction to practical 1** [Kostas]
- **Introduction to practical 2** [Loz]

Some key terms and concepts

Single nucleotide polymorphism (SNP)
Minor allele frequesncy (MAF)
Genotypes and alleles
Genotype frequency

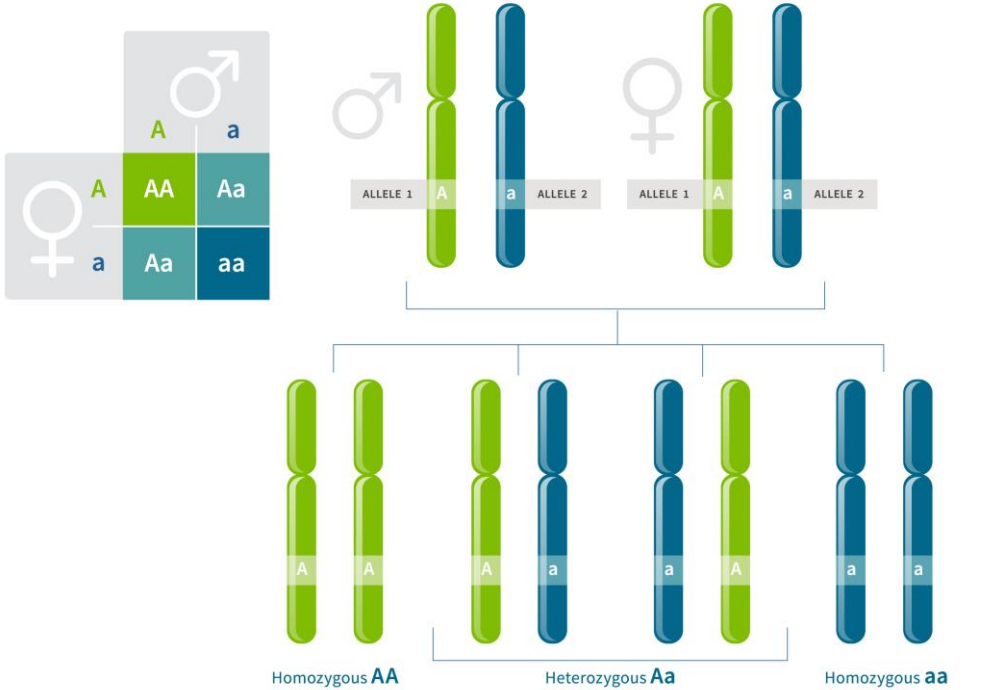
Before we start …

- A **Single Nucleotide Polymorphism (SNP)** is a single base pair at which more than one nucleotide is observed.
- The **Minor Allele Frequency (MAF)** is the relative frequency in a relevant population of the minor (2nd most common) allele.
- For biallelic SNPs, if the MAF of T allele is ***q*** then the frequency of the C allele is ***p=1-q***.



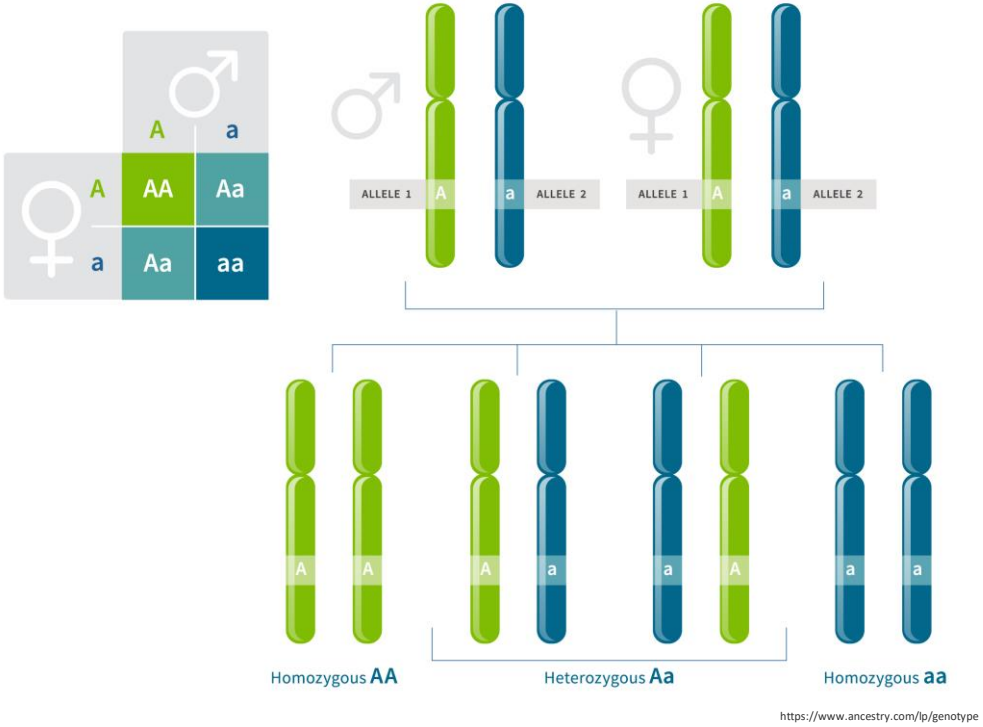
Before we start …

- At a given position in the DNA (or genetic locus), the pair of alleles from the two chromosomes makes up the **genotype** at that position.
- SNP genotypes are **usually encoded as 0, 1 or 2**, based on the number of copies of non-reference alleles.
  - genotype TT is coded as 0 (homozygous non-reference)
  - genotype CT is coded as 1 (heterozygous)
  - genotype CC is coded as 2 (homozygous reference)



Before we start …

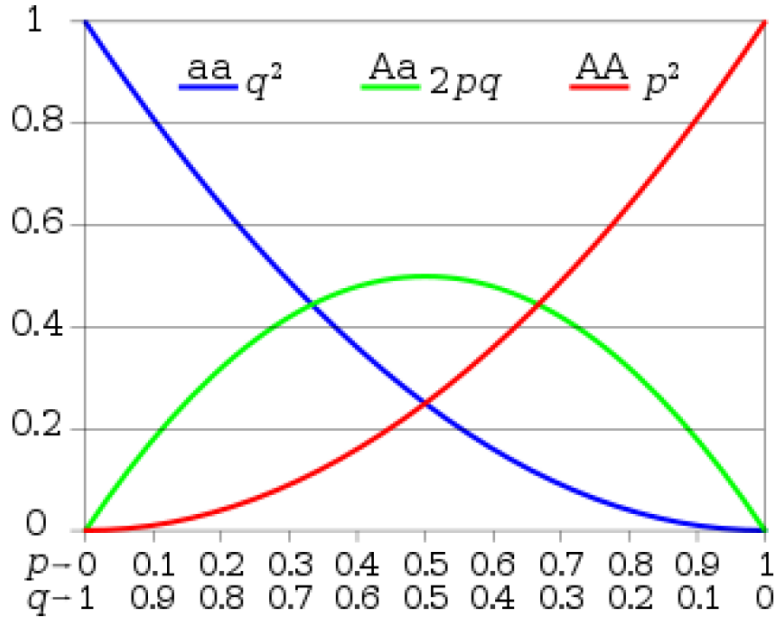
- SNP genotypes are usually **encoded as 0, 1 or 2**, based on the number of copies of non-reference alleles.
  1. genotype TT is coded as 0 (homozygous non-reference)
  2. genotype CT is coded as 1 (heterozygous)
  3. genotype CC is coded as 2 (homozygous reference)
- **Genotypes frequency**:
  - For 1. = q^2
  - For 2.= 2pq
  - For 3.= p^2



Before we start …

- Quick quiz

If we see that 40% of all alleles are a, what is the proportion of aa, Aa, AA?

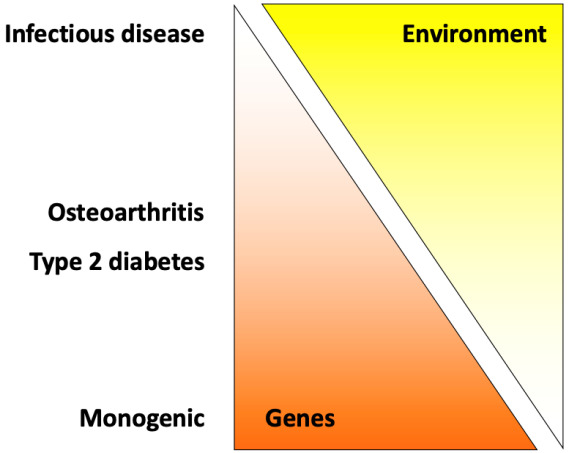


Brief introduction to GWAS

- What input data do we need to undertake a GWAS?

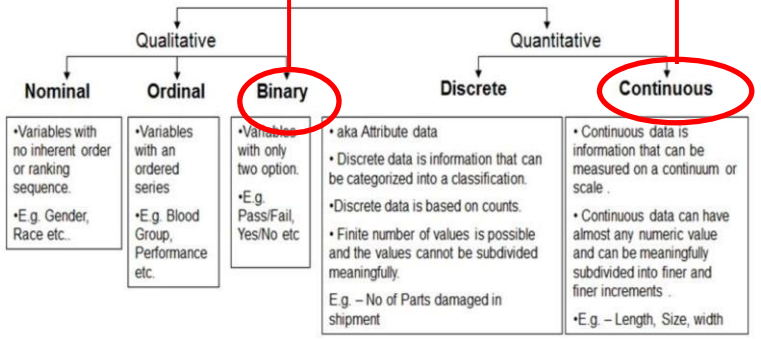
- **Trait of interest (phenotype)**
- Individuals
- Genotypes

Genes vs environment



Case:control
OA GWAS
T2D GWAS

Quantitative
Height GWAS
Blood cholesterol GWAS



Now we need some individuals …

Brief introduction to GWAS

- What input data do we need to undertake a GWAS?

- Trait of interest (phenotype)
- **Individuals**
- Genotypes

Disease specific

arcOGEN

German Chronic Kidney Disease (GCKD)

Special populations

Hellenic Isolated Cohorts (HELIC)
Orkney Complex Disease Study (ORCADES)
SardiNIA Project
Amish

Biobanks

National Joint Registry
www.njrcentre.org.uk
Working for patients, driving forward quality

NAKO GESUNDHEITS-STUDIE

The Estonian Biobank
UNIVERSITY OF TARTU
estonian genome center

CARTAGENE

generation scotland

CHINA KADOORIE BIOBANK
中国慢性病前瞻性研究

biobank uk
Enabling scientific discoveries that improve human health

Consortia

Genetics of Osteoarthritis

GIANT CONSORTIUM

MAGIC
Meta-Analyses of Glucose and Insulin-related traits Consortium

Things to consider:
Ethnicity
Relatedness
Sample size

Individuals need to be genotyped….

Brief introduction to GWAS

- What input data do we need to undertake a GWAS?

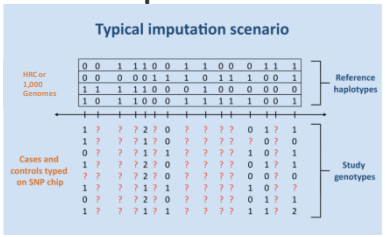
- Trait of interest (phenotype)
- Individuals
- **Genotypes**

WGS → ~3 billion bases in reference genome and >20,000 coding genes

WES → ~40 million bases

Array → 0.5 -5.4 million bases

\+ Imputation →



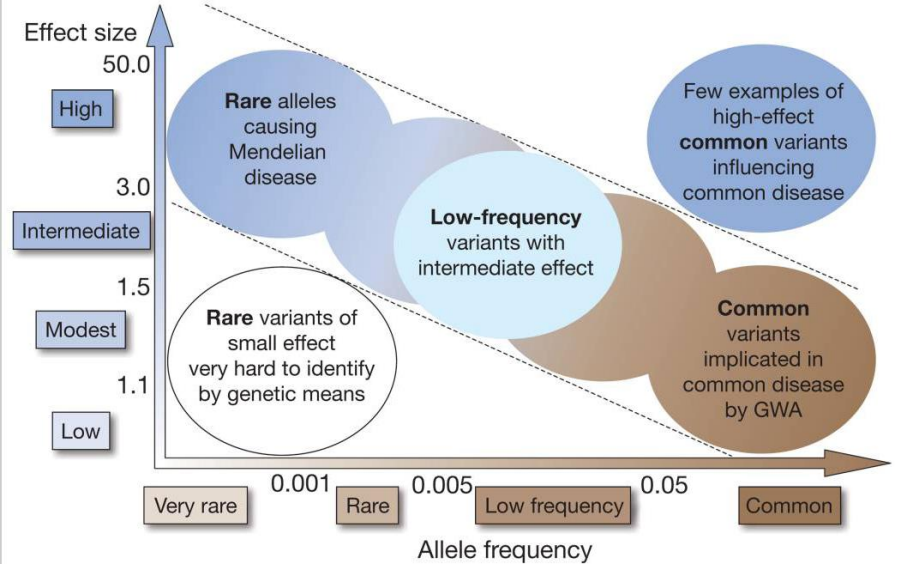
~40-90 million probability based

↓

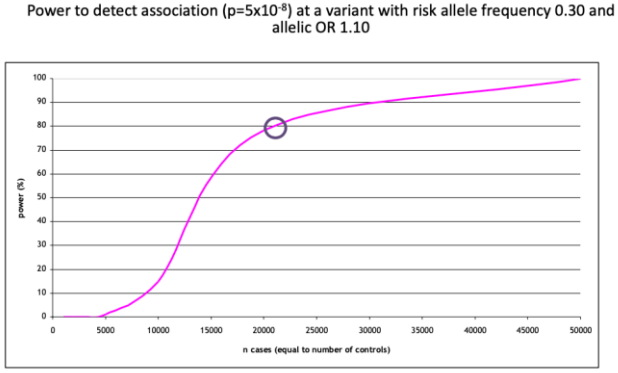
~12-20 million good quality

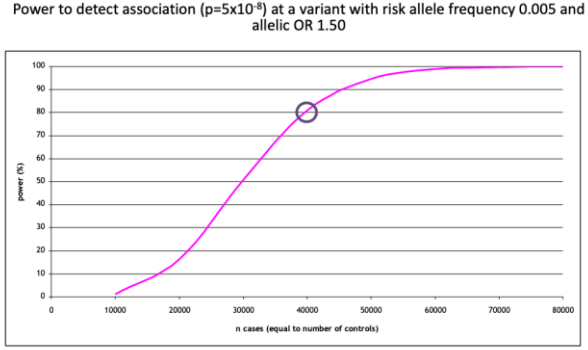
Brief introduction to GWAS

What do we see tend to in GWAS?



Sample size matters





PMID: 19812666

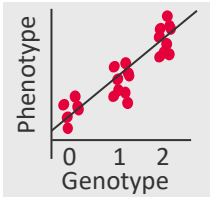
Brief introduction to GWAS

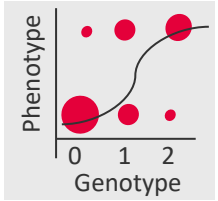
- What are our analysis options and how do we analyse?

Input: Individuals with genotypes and phenotype(s)

Perform an association test e.g.:

- Linear regression (quantitative)
- Logistic regression (case:control)

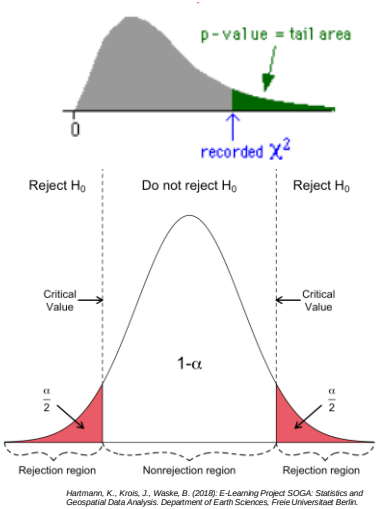




P-value
Beta / Odds ratio
SE / Confidence interval

Null/Alternative hypothesis (H_O /H_A)

Test statistic → P-value the probability of a more 'extreme' test statistic.
If the P-value [0>P<1] is more significant than our threshold we reject H_O



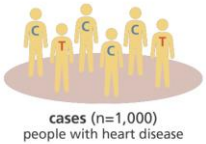
Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin.

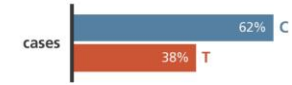
- One test: P-value≤0.05 (1 in 20 chance of a false positive)
- Genome-wide: one test per independent variant and phenotype
- But all variants are not independent, we need to account for LD
- Bonferroni correction: 0.05/1 million independent common variants genome-wide
- **P-value ≤5x10^{-8} for GWAS**

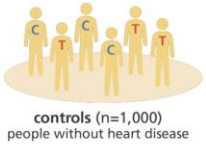
Brief introduction to GWAS

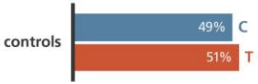
- What are our analysis options and how do we analyse?

Case:control









	Cases	Controls
T	**380**	510
C	**620**	490

- Estimated effect: odds ratio (OR)
  "how much more likely are you to be a case if you carry the risk allele?"
  per genotype g and for a disease Y, calculate the odds $O = \frac{p_{Y=1|g}}{1-p_{Y=1|g}}$

$$OR_{C/T} = \frac{620 \times 510}{490 \times 380} = 1.69$$

$$\text{OR} = \frac{\text{(n) exposed cases / (n) unexposed cases}}{\text{(n) exposed non-cases / (n) unexposed non-cases}} = \frac{\text{(n) exposed cases x (n) unexposed non-cases}}{\text{(n) exposed non-cases x (n) unexposed cases}}$$

Dominant

Marker allele	Affected	Unaffected
DD+Dd	$n_{2A} + n_{1A}$	$n_{2U} + n_{1U}$
dd	n_{0A}	n_{0U}

Recessive

Marker allele	Affected	Unaffected
DD	n_{2A}	n_{2U}
Dd+dd	$n_{1A} + n_{0A}$	$n_{1U} + n_{0U}$

$$OR = \frac{n_{affected\ carriers} \times n_{healthy\ non-carriers}}{n_{healthy\ carriers} \times n_{affected\ non-carriers}}$$

Additive

Marker genotype	Affected	Unaffected
DD	n_{2A}	n_{2U}
Dd	n_{1A}	n_{1U}
dd	n_{0A}	n_{0U}

$$\text{OR} = \frac{(2 \times n_{2A} + n_{1A}) \times (2 \times n_{0U} + n_{1U})}{(2 \times n_{0A} + n_{1A}) \times (2 \times n_{2U} + n_{1U})}$$

Allelic odds-ratio

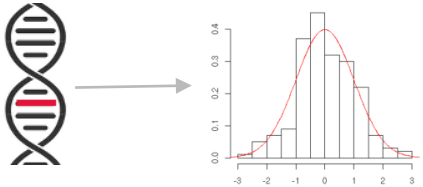
- Output: OR and 95% confidence interval of the OR
- Test: is it significantly different from 1?
- Tests: Fisher's exact test or Chi-squared
- In case of dosages or covariates: logistic regression

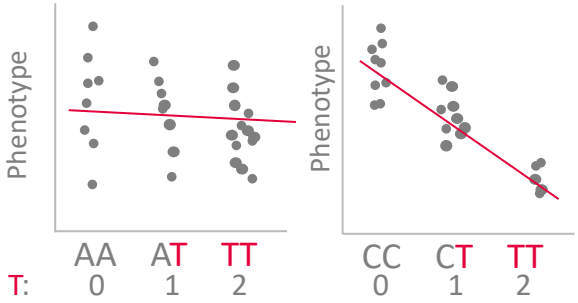
P-values, odds ratio + confidence interval

Brief introduction to GWAS

- What are our analysis options and how do we analyse?

Quantitative





Effect allele / non-effect allele

Additive effect

A linear regression model is defined as

$$y = x\beta_1 + \beta_0 + \varepsilon$$

Data:

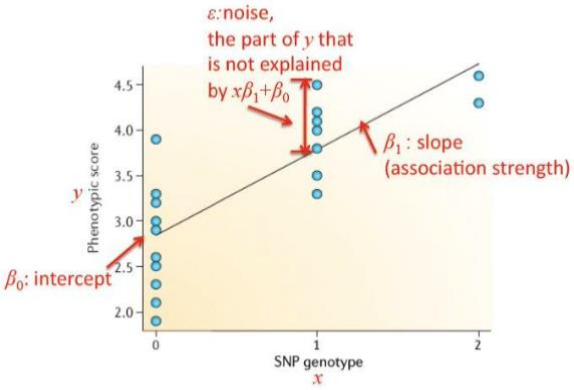
- y: a continuous trait
- x: SNP genotype at a given locus

Parameters:

- β_1: regression coefficient, represents the strength of association between x and y
- β_0: intercept term (is 0 or ignored)
- ε: noise or the part of y that is not explained by x (e.g., environmental effect)

Assumptions:

- The individuals in the study are not related
- The phenotype y has a normal distribution



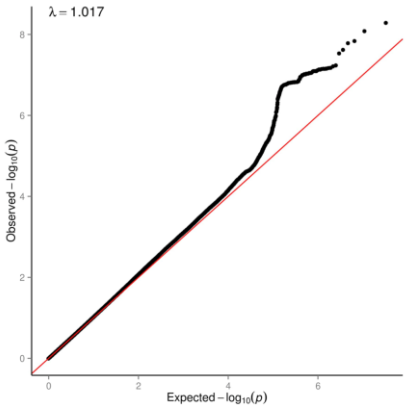
P-values, beta + standard error

Brief introduction to GWAS

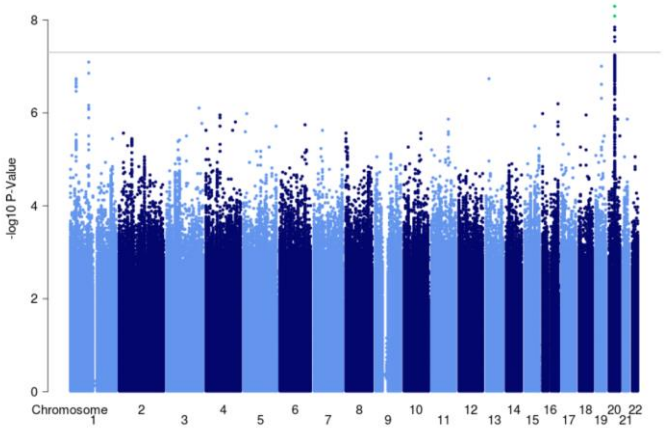
- How to visualise the results

Genome-wide significance
P-value = 5 x 10^{-8}

QQ-Plot

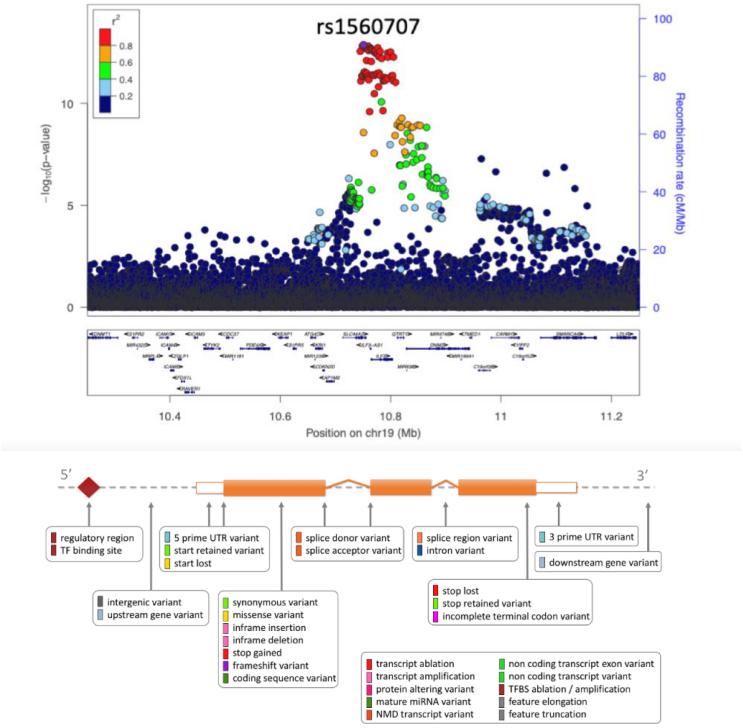


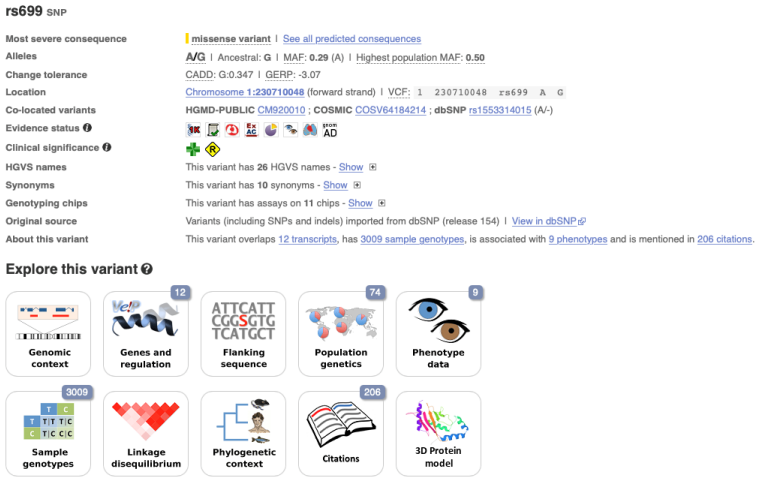
Manhattan Plot



Brief introduction to GWAS

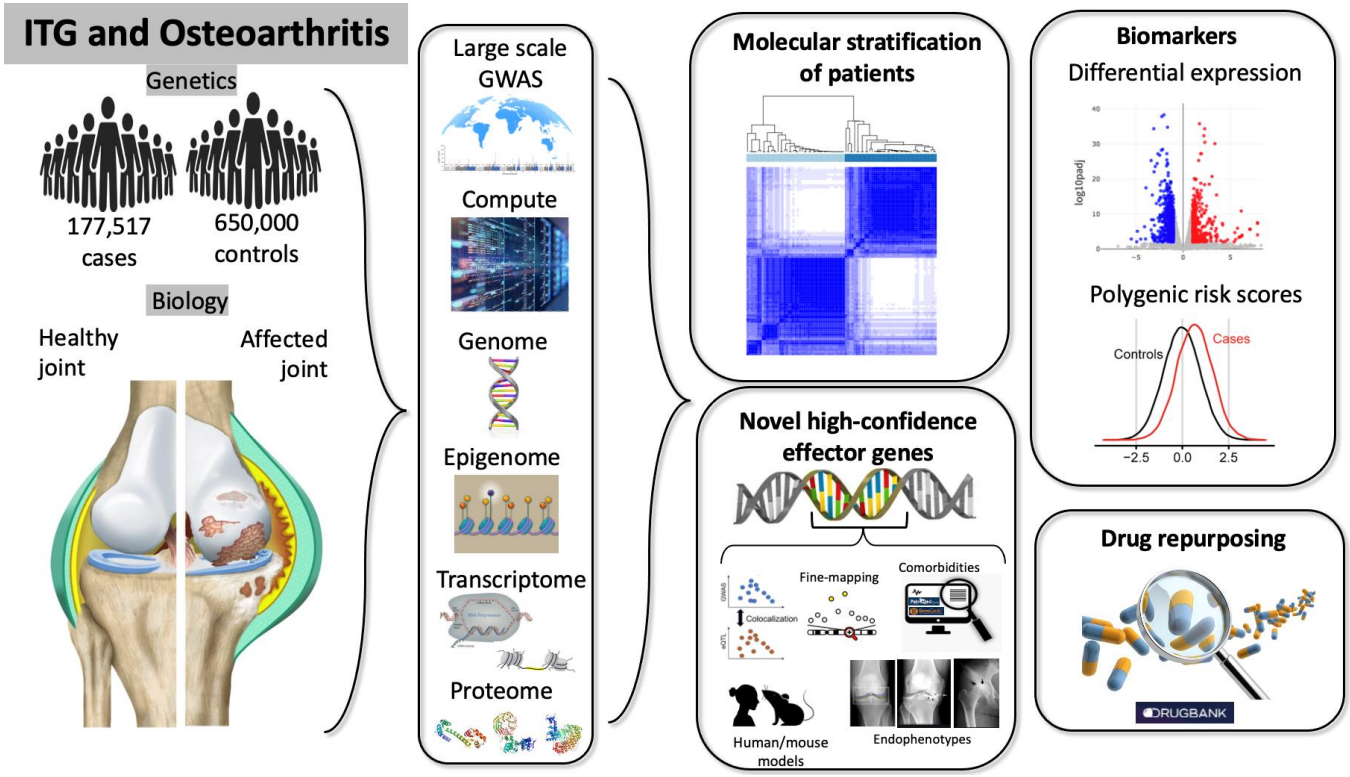
- Post GWAS analysis





Brief introduction to GWAS

- Putting GWAS into context (with an example from ITG)

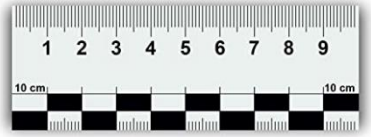


Overview of the practical

Small dataset

Genotypes

Phenotype



Practical 1

Genotyping quality control

Phenotype quality control

Practical 2.1

Perform a GWAS

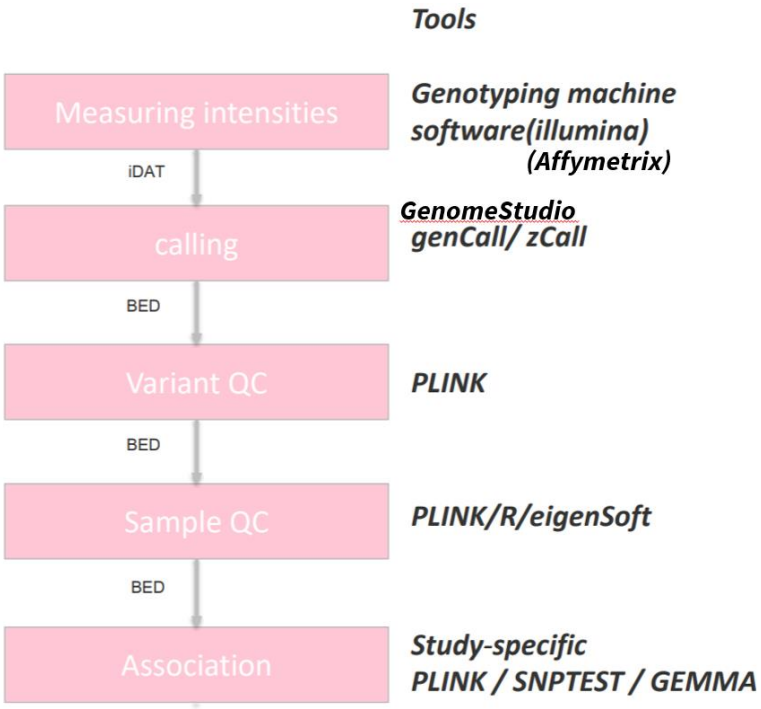
Practical 2.2

Inspect results

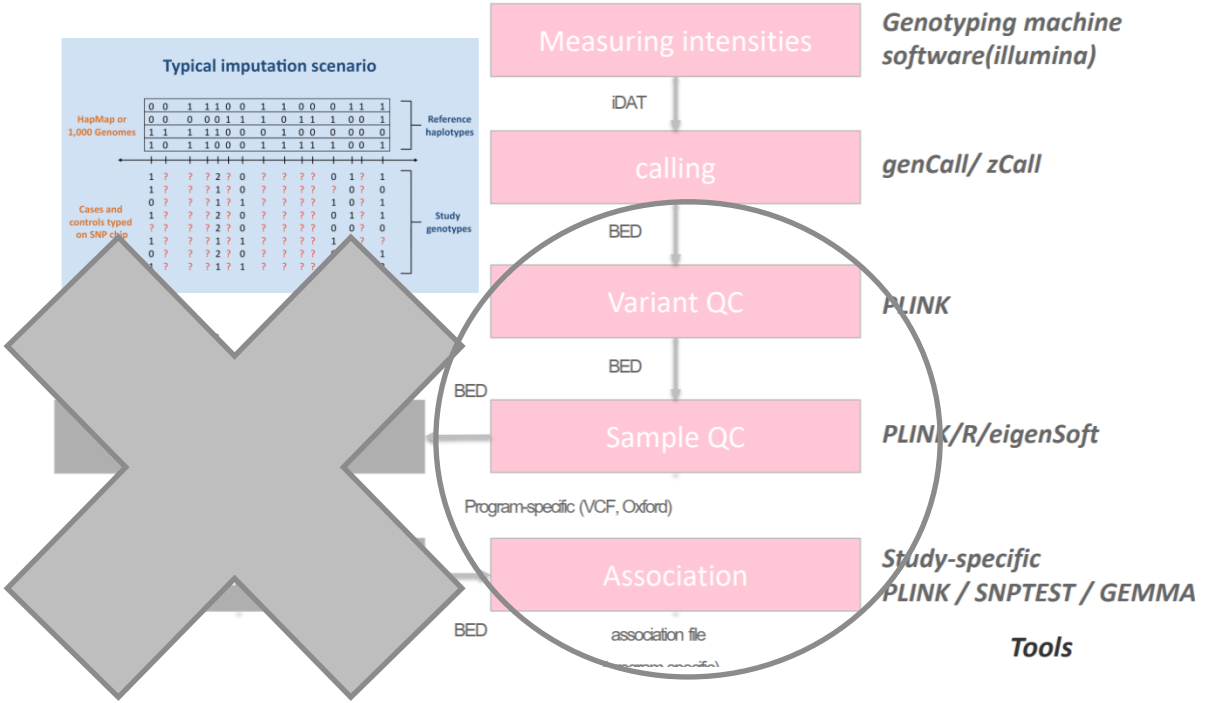
Practical 2.3

Introduction to practical 1

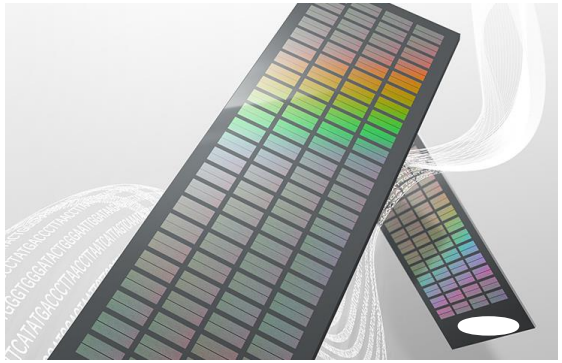
The GWAS analysis pipeline



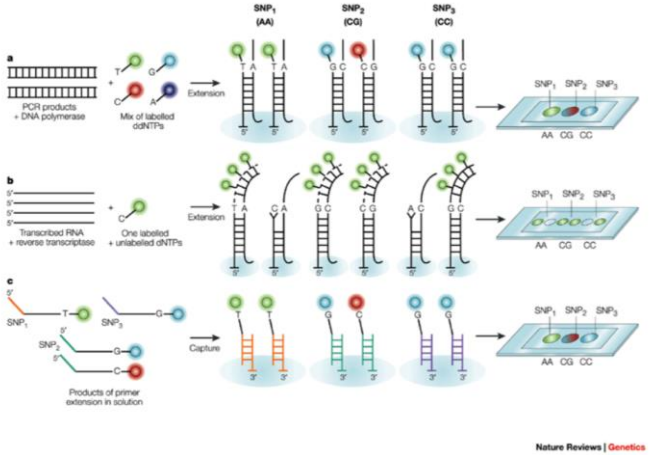
The (imputed) GWAS analysis pipeline

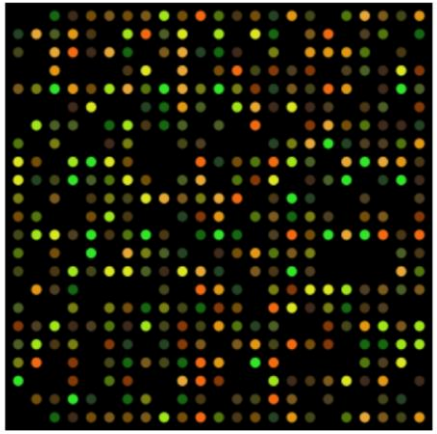


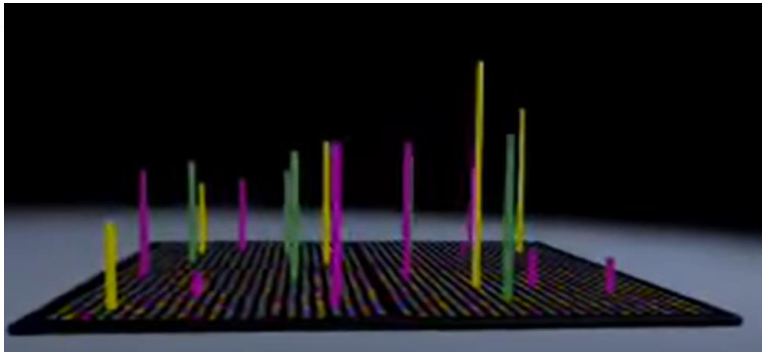
Did you say intensities?



Did you say intensities?

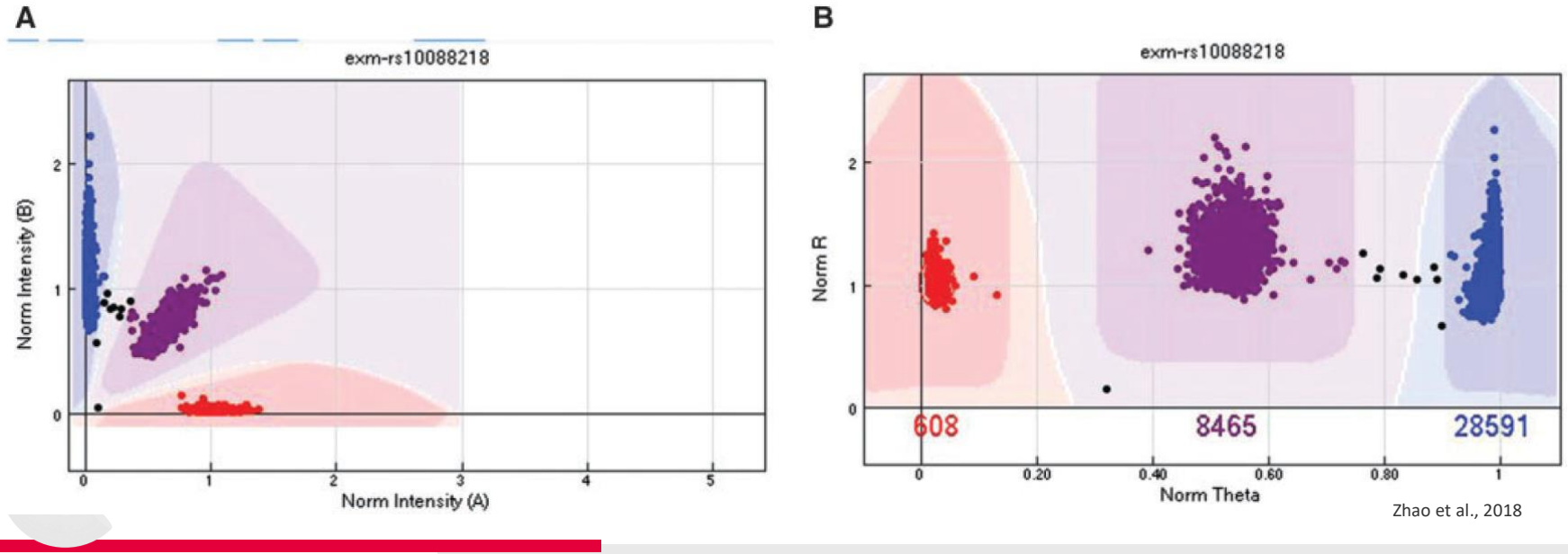






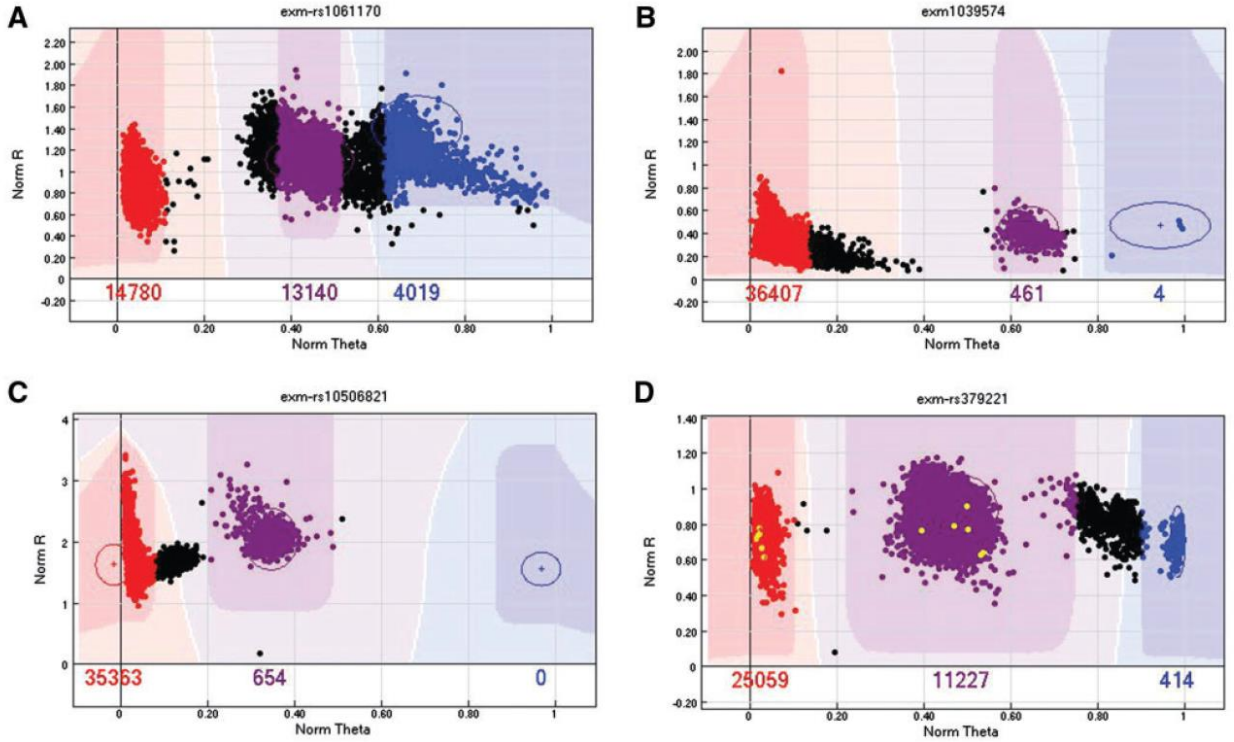
https://www.youtube.com/watch?v=AhnTT6-Jgcg

Intensities: the good …



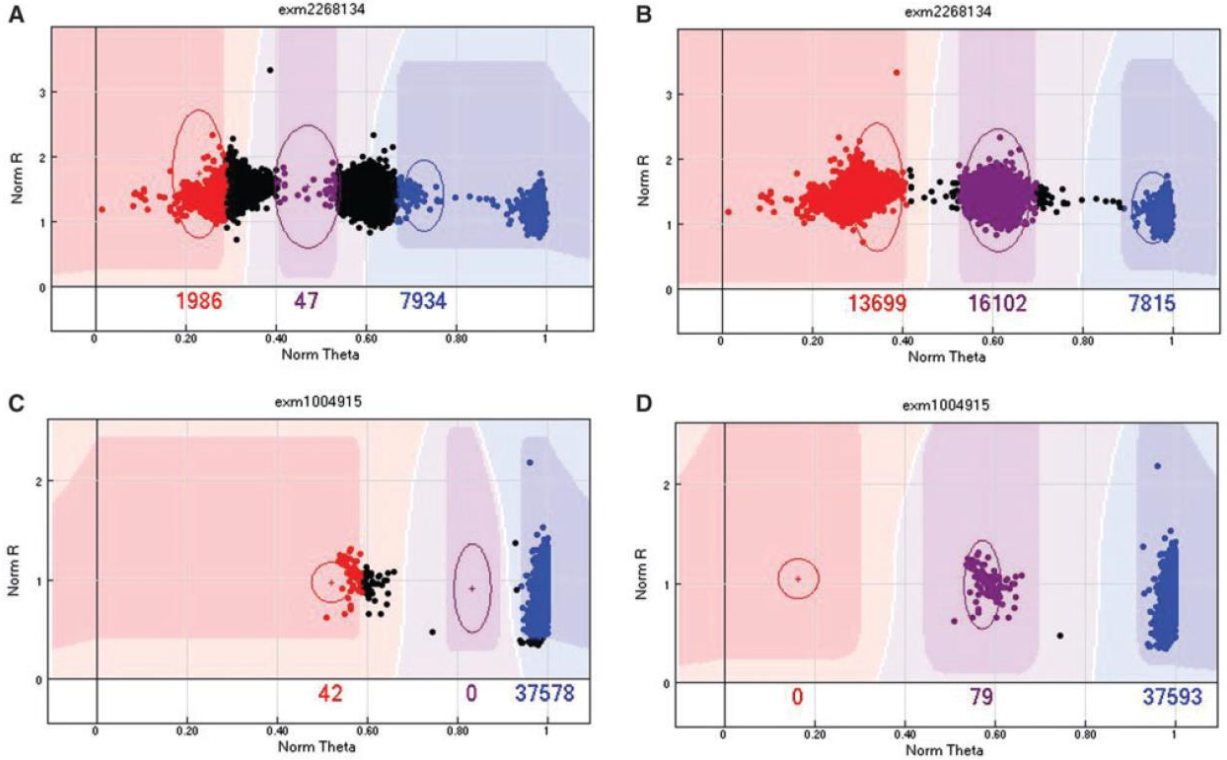
Zhao et al., 2018

Intensities: the bad …



Zhao et al., 2018

Intensities: the ugly …



Zhao et al., 2018

Genotyping data storage

- Which data types do we need?

$$\text{phenotype} \sim \beta \times \text{genotype} + \text{covariates} + \text{structure} + \varepsilon$$

$$\begin{bmatrix} pheno_0 \\ \vdots \\ pheno_n \end{bmatrix} \quad \begin{bmatrix} A/T \\ \vdots \\ T/T \end{bmatrix} \quad \begin{bmatrix} male \\ \vdots \\ female \end{bmatrix} \quad \begin{bmatrix} 22\ years \\ \vdots \\ 65\ years \end{bmatrix} \quad \begin{bmatrix} r_{00} & \cdots & r_{0n} \\ \vdots & r_{ij} & \vdots \\ r_{n0} & \cdots & r_{nn} \end{bmatrix}$$

These stay constant (they describe the samples)

This one changes

Genotyping data storage: PLINK

- Can either be text-format files or binary files.

*.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

*.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

```
10101111 10101111 10100010 10111011 10101000 10000000
00101011 00100000 10101000 10001011 00000011 11111111
11111111 11111111 11111111 11111110 11111111 11111111
11111111 11111110 11111110 11111110 11101111 11111111
```

Legend			
FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)

Marees et al., 2017

Genotyping data storage: PLINK

*.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

- **ped**(igree) file has **6+2*n***, providing:
  1. Family ID
  2. Individual ID
  3. Paternal ID (0 if father not in dataset)
  4. Maternal ID (0 if mother not in dataset)
  5. Sex (1=Male, 2=Female, 0 or -9=missing)
  6. Phenotype (here 2 or 1, corresponding to case and control)
  7. 2 alleles for each SNP (0 = missing)

- **map**(ing) file has **4 columns**, providing:
  1. Chromosome
  2. SNP Name
  3. Genetic distance (in morgans)
  4. Base-pair position (bp unit)

Genotyping data storage: PLINK

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

*.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

- **fam**(ily) file consists of the first six columns of ped file
- The **bed** (binary pedigree) file is a matrix of 0s, 1s, 2s or NAs stored in binary format.
- PLINK uses the following two-bit coding of genotypes:
  - 00 = A1/A1 (Homozygous non-reference)
  - 01 = A1/A2 (Heterozygous)
  - 11 = A2/A2 (Homozygous reference)
  - 10 = 0/0 (Missing)
- **bim** (binary mapping) file is the .map file plus two columns, providing the A1 and A2 alleles

Genotyping data storage: PLINK

- What is left?

Matrix file
(program-specific)

$$\begin{bmatrix} r_{00} & \cdots & r_{0n} \\ \vdots & r_{ij} & \vdots \\ r_{n0} & \cdots & r_{nn} \end{bmatrix}$$

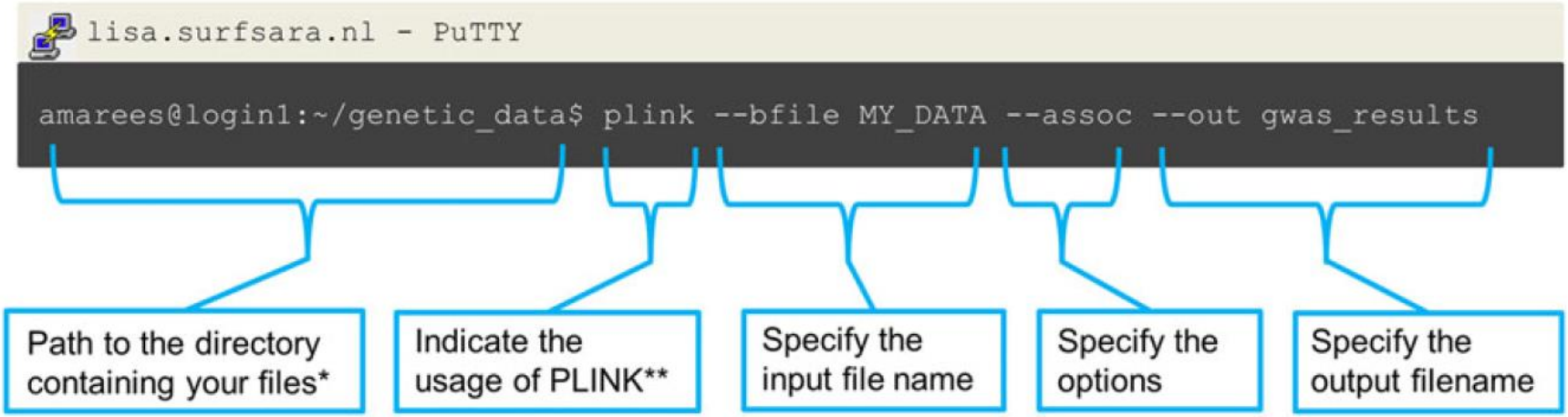
Covariate file

FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

Phenotype files have 2 + M columns: Family ID, Individual ID, then value for each of M phenotypes

Marees et al., 2017

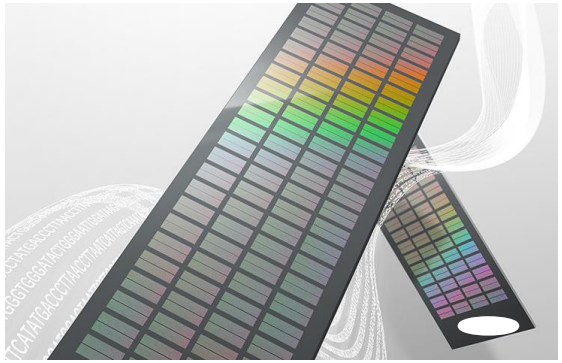
Genotyping data : PLINK common operations



https://www.cog-genomics.org/plink/1.9/index
https://www.cog-genomics.org/plink/2.0/index

Marees et al., 2017

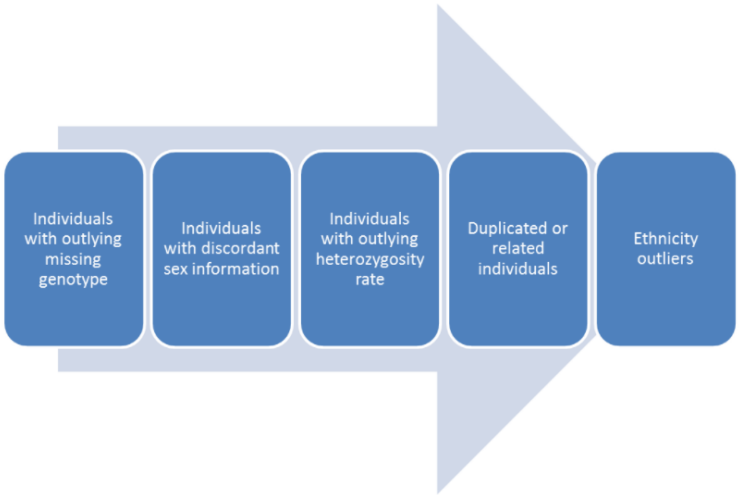
Why Quality Control?



Why Quality Control?

- The QC protocol of a GWAS is usually split into two broad categories.

“Sample QC”



“Variant QC”

1. Identification of variants with an **excessive missing genotype**
2. Identification of variants demonstrating a significant deviation from **Hardy-Weinberg equilibrium (HWE)**
3. Removal of all makers with a **very low minor allele frequency**
4. Removal of all makers with **cluster separation score <0.4**
5. **Differential missingness** (case/control studies)

(A generalized) Quality Control process

- **Missingness**

1. Per sample missingness

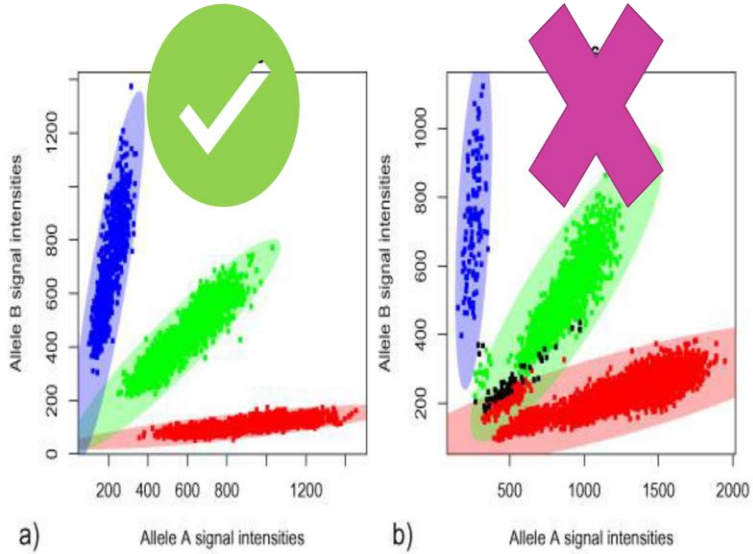
➢ % missing for a sample across your variants

2. Per SNP missingness

➢ % missing for a particular variant among your samples

Quality control step	PLINK summary commands	PLINK filtering commands
Missingness	`--missing`	`--geno`, `--mind`

➢ Low genotyping call rate indicates issues with sample DNA (eg low concentration).



(A generalized) Quality Control process

- **Discordant Sex Check**

➢ Men have only one copy of the X chromosome

➢ All X chromosome data is expected to be homozygous.

Example Alleles	Female genotypes possible	Male genotypes possible
A,C	A/A, A/C, C,C	A/A or C/C

➢ X chromosome homozygosity estimate for males (F statistic or inbreeding coefficient) is 1.

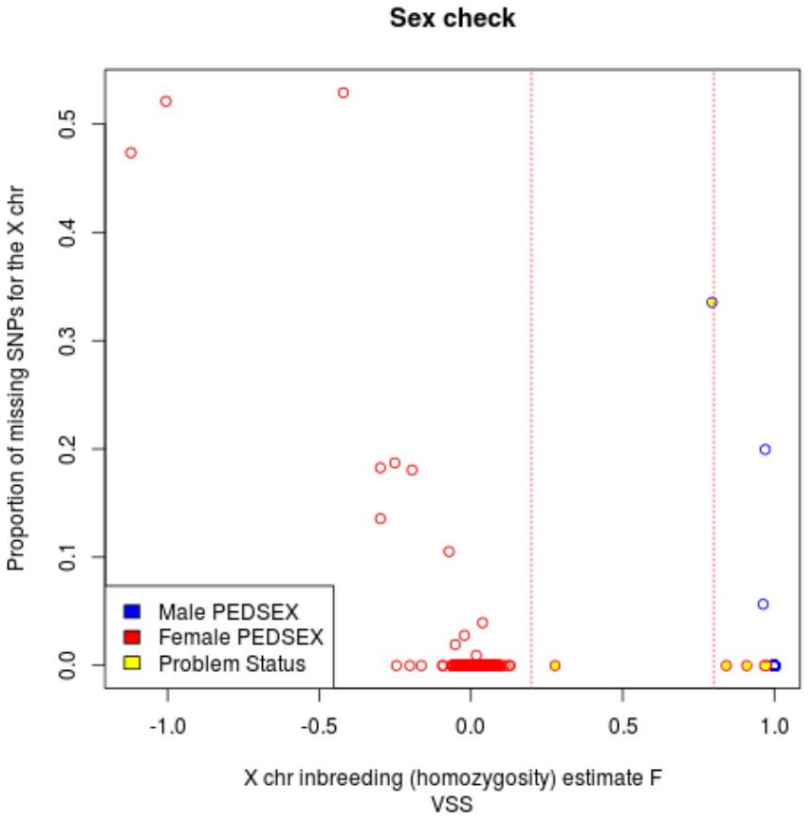
➢ In Plink

`--check-sex`	Check sexes by looking at chrX

➢ Male (1) : XHE >0.80

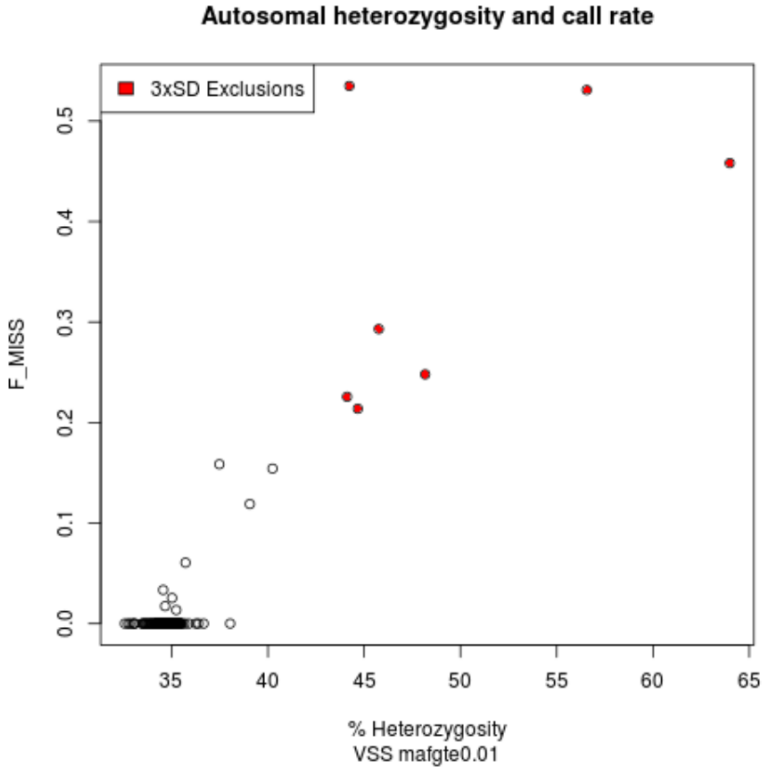
➢ Female (2) : XHE <0.20

➢ No sex (0) : 0.20 <XHE <0.80



(A generalized) Quality Control process

- **Heterozygosity rate**
- The proportion of heterozygous genotypes (per sample)
- Various ways of calculating the rate
- PLINK: (<observed hom. count> - <expected count>) / (<total observations> - <expected count>))
- --het (gives back and F estimate)
- <custom scripts>
- Excess heterozygosity -> Possible sample contamination
- Less than expected heterozygosity –> Possibly inbreeding

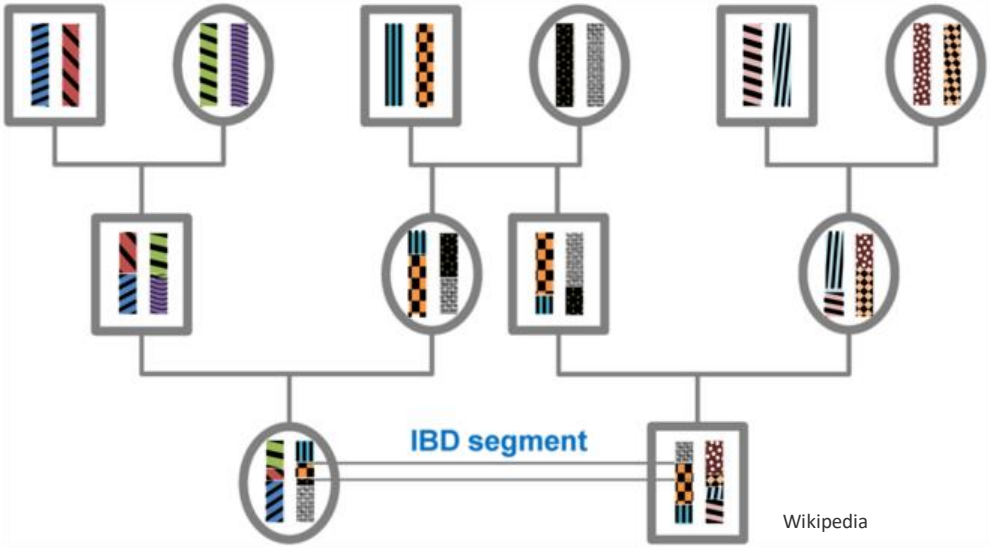


(A generalized) Quality Control process

- **Duplicated or related individuals**
- A basic assumption of GWAS: unrelated individuals
  - Either exclude or account for it
- The presence can introduce a bias: genotypes in families to be over-represented

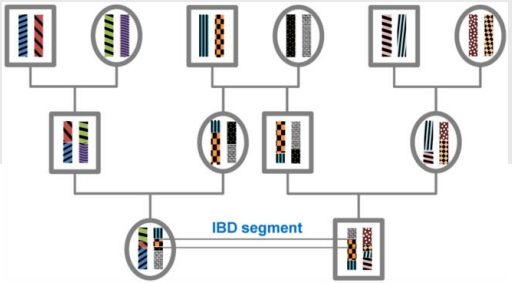
(A generalized) Quality Control process

- **Duplicated or related individuals**
- Calculated metrics:
  - **Identity by state (IBS)**: A DNA segment is identical by state (IBS) in two or more individuals if they have identical nucleotide sequences in this segment.
  - **Identity by Descent (IBD)**: An IBS segment is identical by descent (IBD) in two or more individuals if they have inherited it from a common ancestor without recombination (that is, the segment has the same ancestral origin in these individuals).



Wikipedia

(A generalized) Quality Control process



- **Duplicated or related individuals**
- PLINK calculates identity by descent (IBD) of all sample
- Approximates the percentage IBD overall, representing pairs

```
--genome
```

  - Zero alleles IBD (z0)
  - One allele IBD (z1)
  - Two alleles IBD (z2)
- PI_HAT (the proportion IBD, defined as P(IBD = 2) +

Use an independent SNP set before running this command:

1) removing regions of extended Linkage Disequilibrium (LD)

and

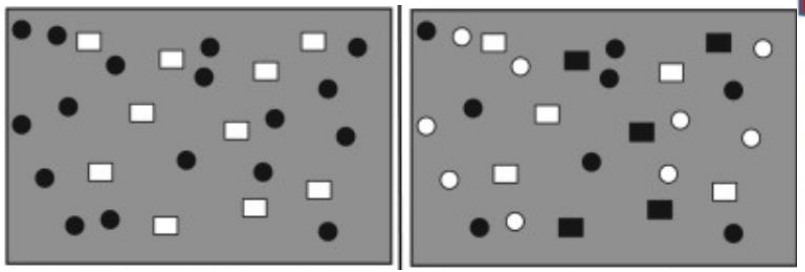
2) pruning the remaining regions so that no pair of SNPs within a given window is correlated

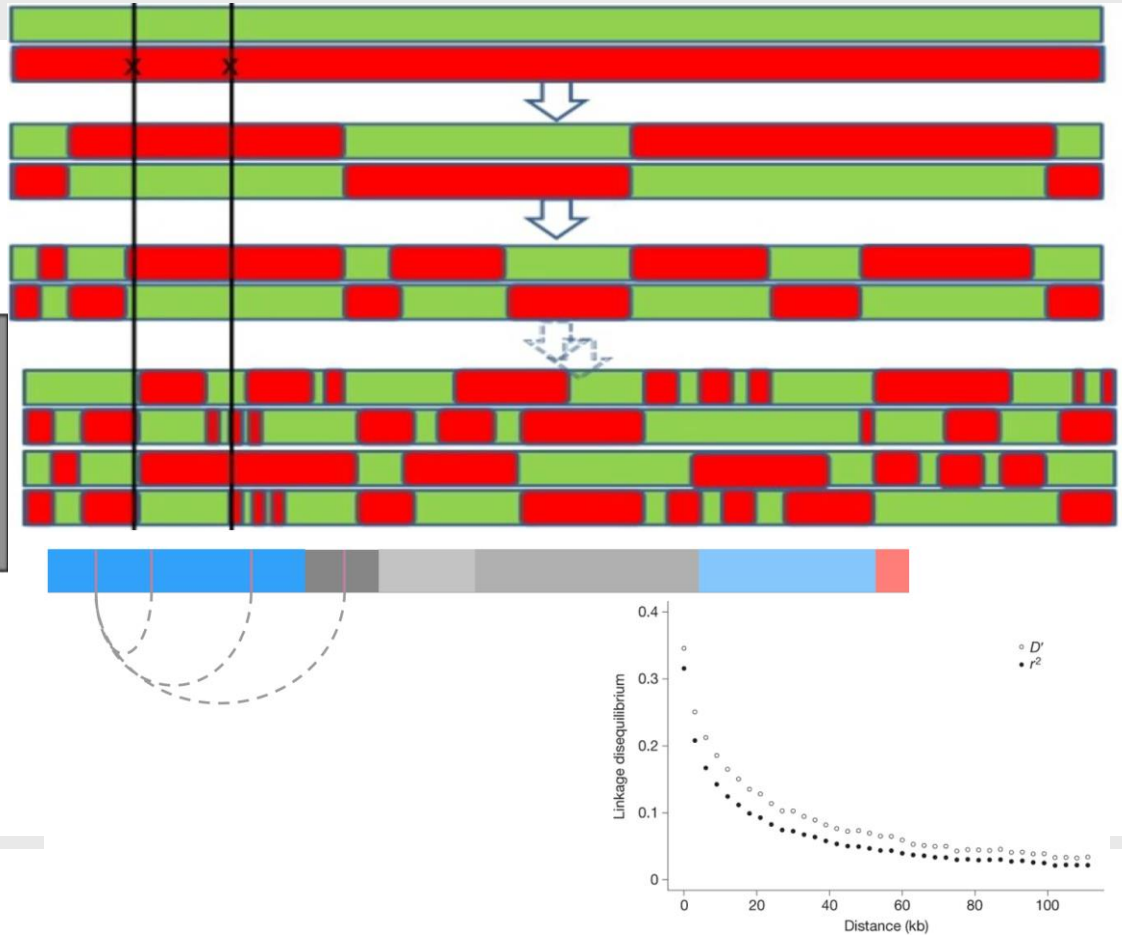
Relationship type	z0	z1	z2	
Unrelated	1	0	0	0
Monozygotic (MZ) twin	0	0	1	1
Full siblings	0.25	0.5	0.25	0.5
Half siblings	0.5	0.5	0	0.25
Parent-offspring	0	1	0	0.5

Slifer, 2018

(A generalized) Quality Control process

- **Linkage disequilibrium (LD)**
- Is the non-random association of alleles at different loci in a given population.



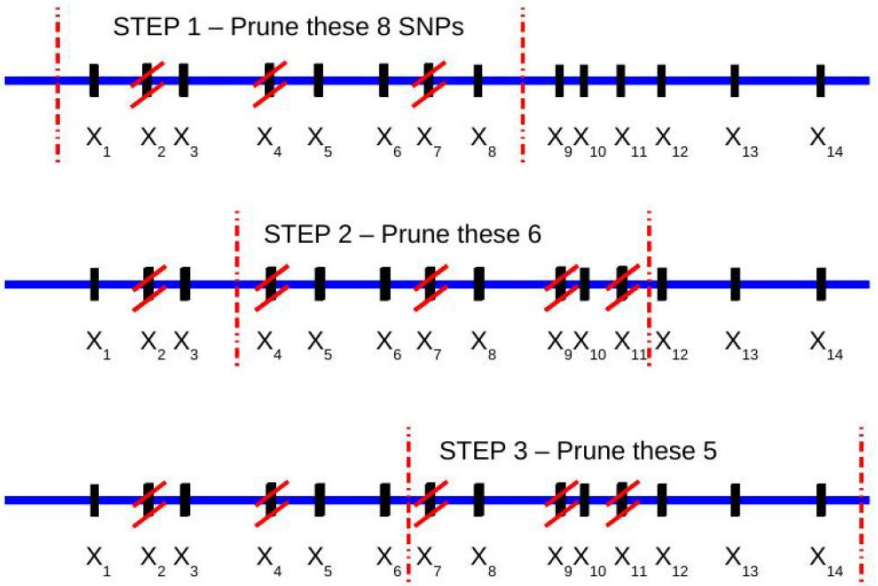


(A generalized) Quality Control process

- **PLINK: LD-based SNP prunning**

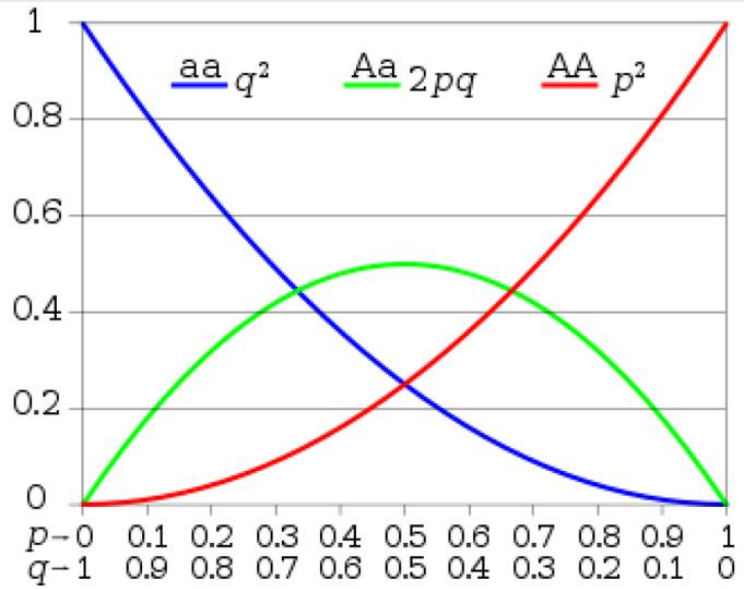
```
plink --indep-pairwise <window> <step> <rsq> --bfile <data> --out <output>
```

```
plink --indep-pairwise 8 3 <rsq> --bfile <data> --out <output>
```



(A generalized) Quality Control process

- **The Hardy–Weinberg (dis)equilibrium (HWE) law:**
- Assumes:
  - An indefinitely large population
  - With no selection
  - No mutation
  - No genetic drift
  - Etc..
- The genotype and the allele frequencies are constant over generations.
- Significant deviations indicate genotyping errors
- PLINK:

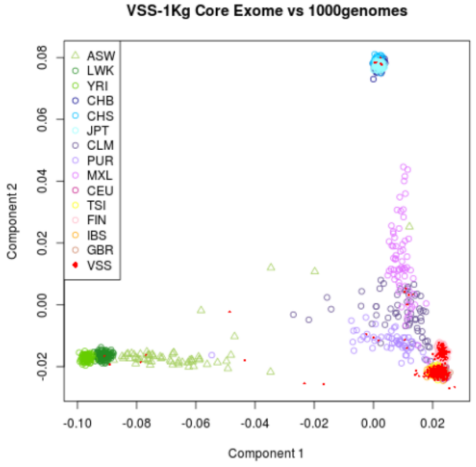


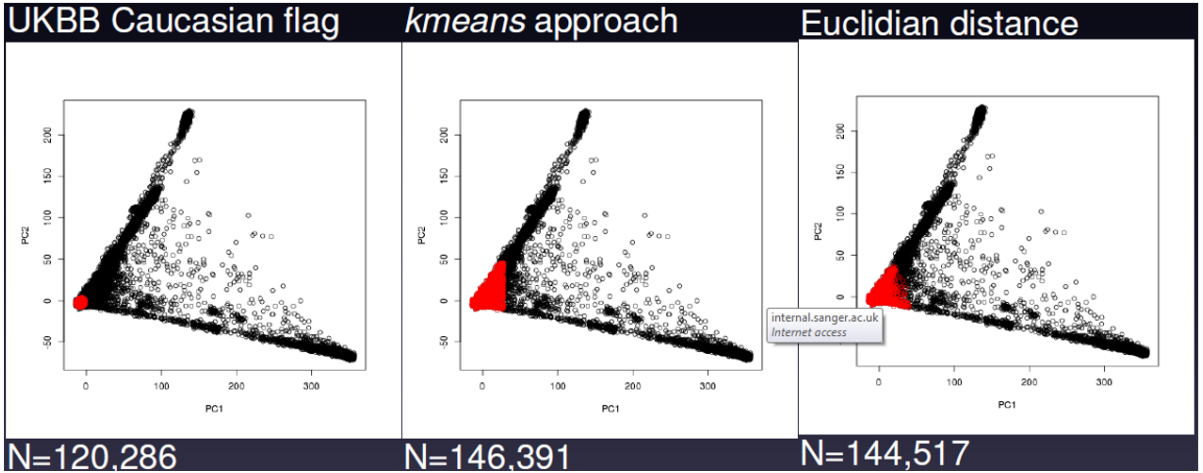
Quality control step	PLINK summary commands	PLINK filtering commands
Hardy-Weinberg equilibrium check	`--hardy`	`--hwe`

- Less strict case threshold avoids discarding disease-associated

(A generalized) Quality Control process

- **Population structure**
- Occurs when samples have different genetic ancestries
- Allele frequencies can differ between subpopulations and can lead to spurious associations due to differences in ancestry rather than true associations
- PLINK: Merge with a population of known ethnic structure (e.g., HapMap/1KG data) and identify outliers through dimension reduction analyses such as Principal Component Analysis and/or MultiDimensional Scaling (MDS).



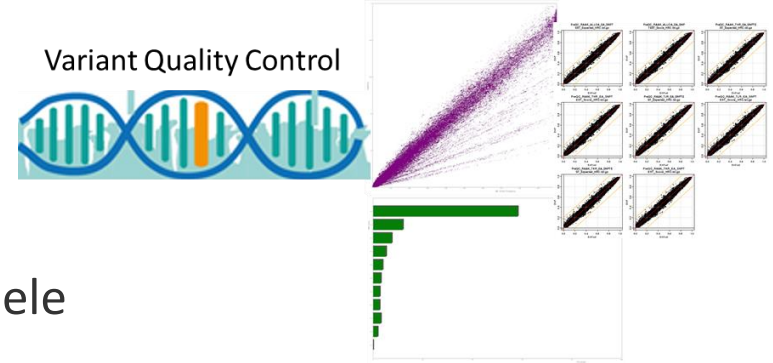


(A generalized) Quality Control process

- **Variant QC**

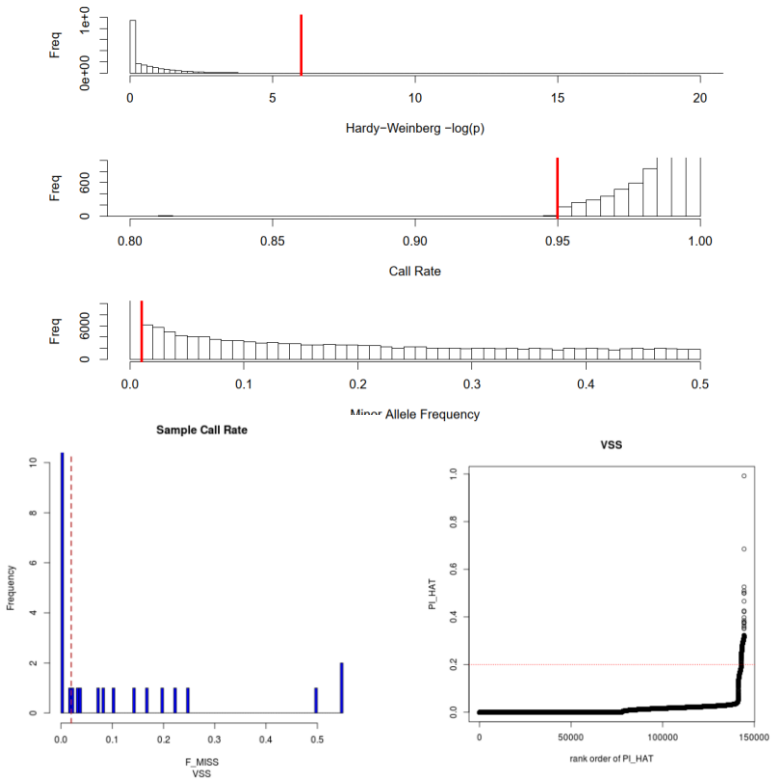
It consists of (at least) four steps:

1. Identification of variants with an excessive missing genotype
2. Identification of variants demonstrating a significant deviation from Hardy-Weinberg equilibrium (HWE)
3. Removal of all makers with a very low minor allele frequency
4. Removal of all makers with cluster separation score



(A generalized) Quality Control process

- **Where to draw the line?**



Genotyping data : PLINK common operations

Sample management

`--keep [file]`	Keep samples in file
`--remove [file]`	Remove samples in file

SNP management

`--extract [file]`	Keep SNPs in file
`--exclude [file]`	Remove SNPs in file

Extracting regions

`--chr [name]`	Extract data on specified chromosome
`--from-bp [pos]`	From specified position
`--to-bp [pos]`	To specified position

Genotyping data : PLINK common operations

Variant QC

`--maf [threshold]`	Keep variants with MAF>threshold
`--hwe midp [threshold]`	Keep variants with HWE p>threshold

Sample QC

`--missing`	Compute per-sample and per-variant missingness
`--check-sex`	Check sexes by looking at chrX
`--genome`	Compute relatedness, check for duplicates

Genotyping data : PLINK common operations

- What is the command for :
  - Excluding SNPs that are missing in a large proportion of the subjects (<0.90).
  - Excluding individuals who have high rates of genotype missingness (<0.85).
  - Keeping autosomal SNPs.
  - Extracting the top 20 principal components.
  - The association between SNPs and a binary/quantitative outcome.

Introduction to practical 2

Quantitative phenotype quality control

Association analysis

Visualising and annotation of results

Quantitative Phenotype Quality Control

We will use R

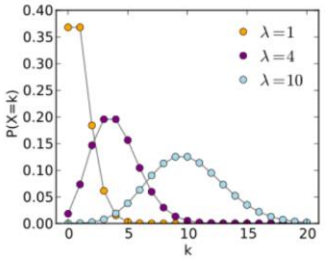
- Inspect the data and look at the distribution
- Convert to normal distribution
- Write out the transformed phenotype

Inspect the data
Are the values sensible are there any missing values?
What does the distribution look like ?
Are there outliers?

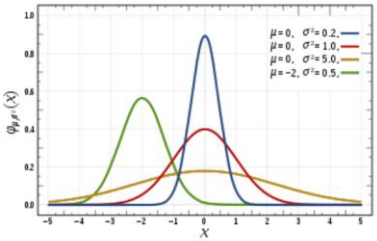
ID	T1	T2	T3
3456	1.4	M	2
5689	1.3		1
1658	1.6	F	1
5698	1.4	M	2
6589	1.3	F	NA
5611	15	NA	1

Quantitative Phenotype Quality Control

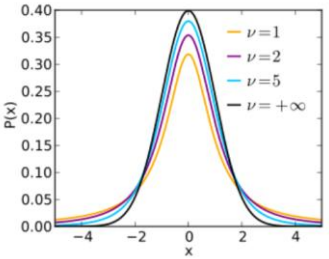
There are 2 broad types of distribution:
Those followed by random variables (real world) + those followed by test statistics



$X \sim Poisson(\lambda)$

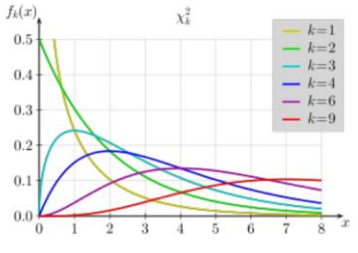


$X \sim \mathcal{N}(\mu, \sigma^2)$



$X \sim T(\nu)$

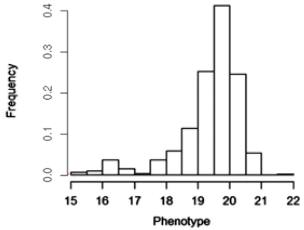
T-test

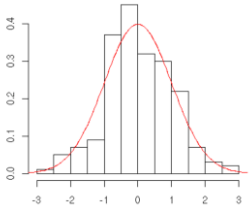


$X \sim \chi^2(k)$

Chi-square (square of standard normal distribution)

In R we can apply a transformation to make our phenotype have a normal distribution





Natural log
Log_2
Log_{10}
Cube root
Square root
Inverse normal

Association Analysis

We will use plink

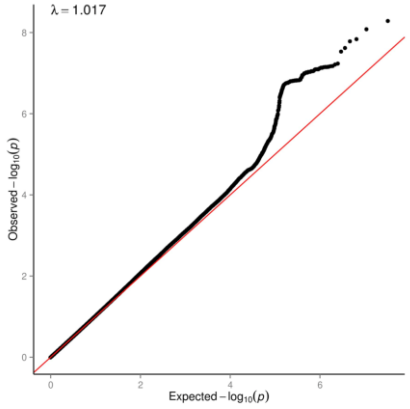
Genotypes can be included using **--bfile** and phenotypes using **--pheno**

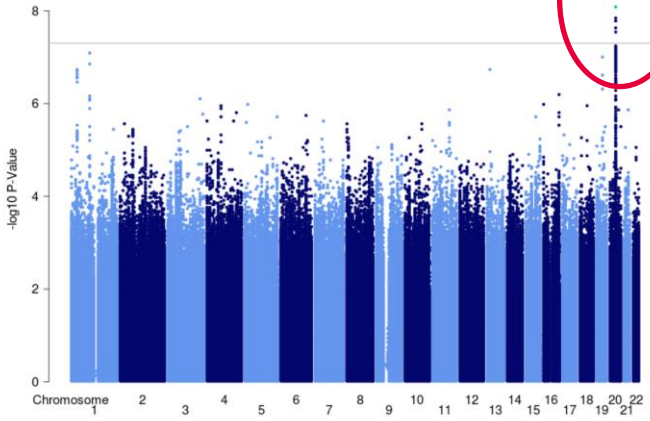
Binary trait: **--assoc** or **--logistic**
Quantitative trait: **--assoc** or **--linear**

Faster no covariates allowed
Slower can include covariates

Visualising and annotation of results

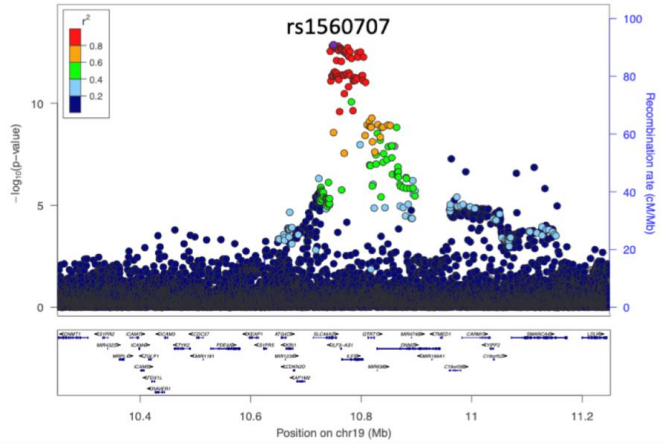
We will use R and the library **qqman**





$Log_{10}(5x10^{-8}) = 7.3$

We will query **Ensembl** to get the genes in the associated region and in R we will combine these with our data to create our own version of a regional plot (without LD annotated).



Thank-you!

Any questions and then let's get started!