

# Volos Summer School

Current tools and best practices for  
performing genome-wide scans

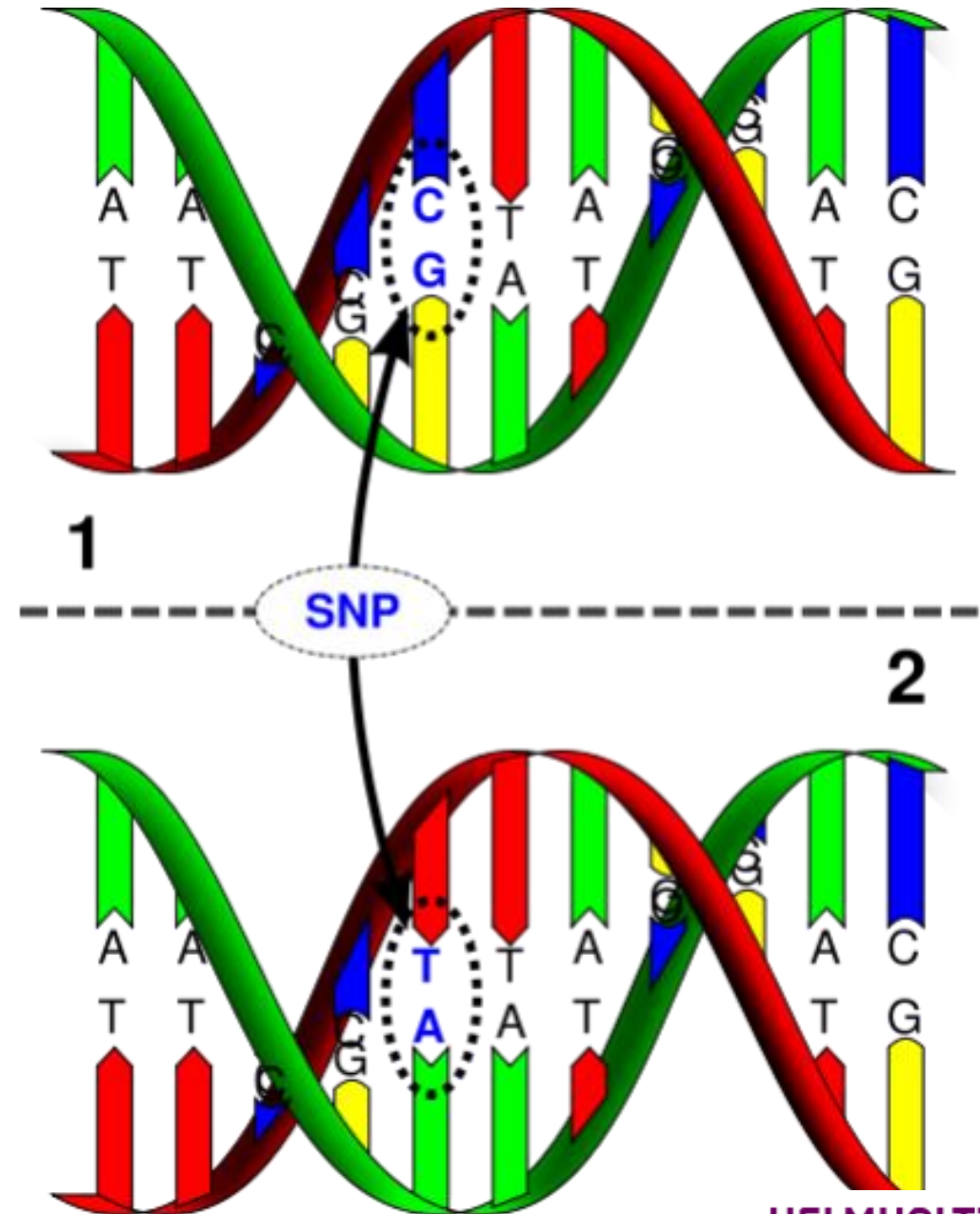
Institute of  
Translational Genomics



Konstantinos Hatzikotoulas

# 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  $p=1-q$ .

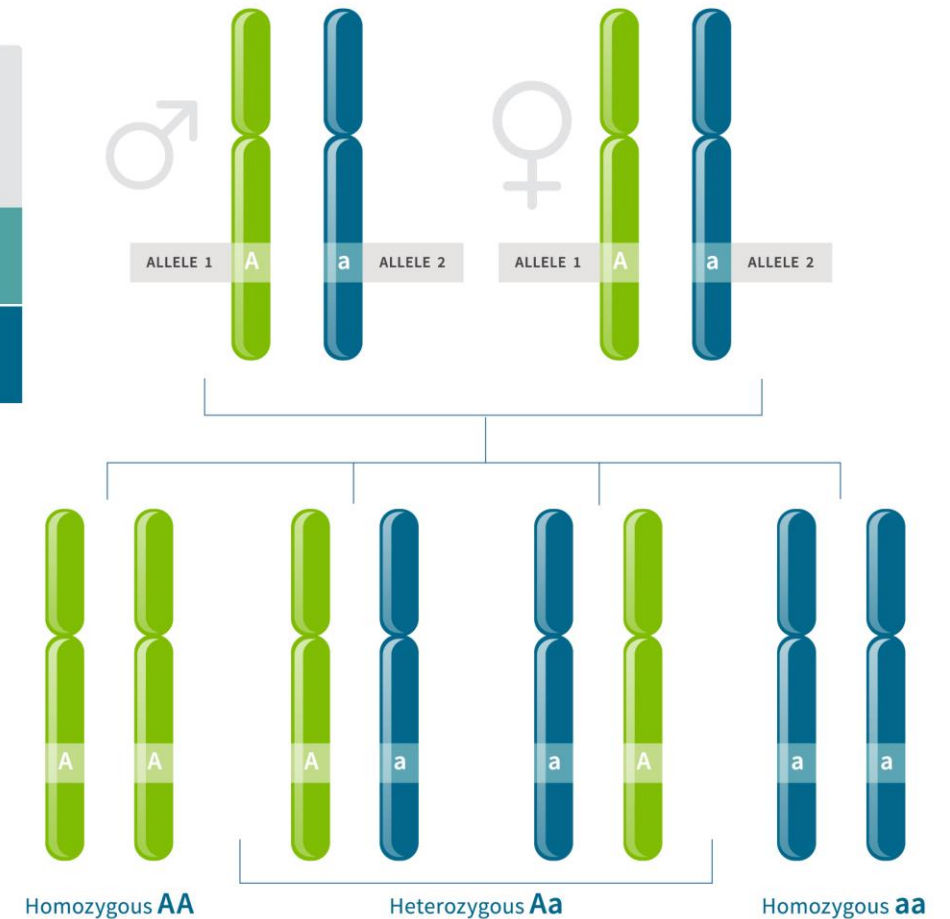
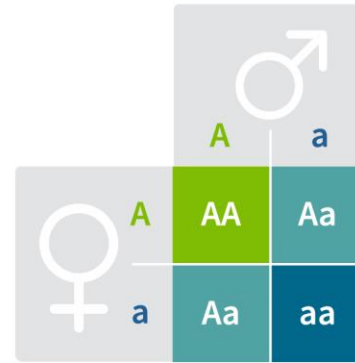


# Before we start

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 TT is coded as 2 (homozygous reference)



<https://www.ancestry.com/lp/genotype>



# Before we start

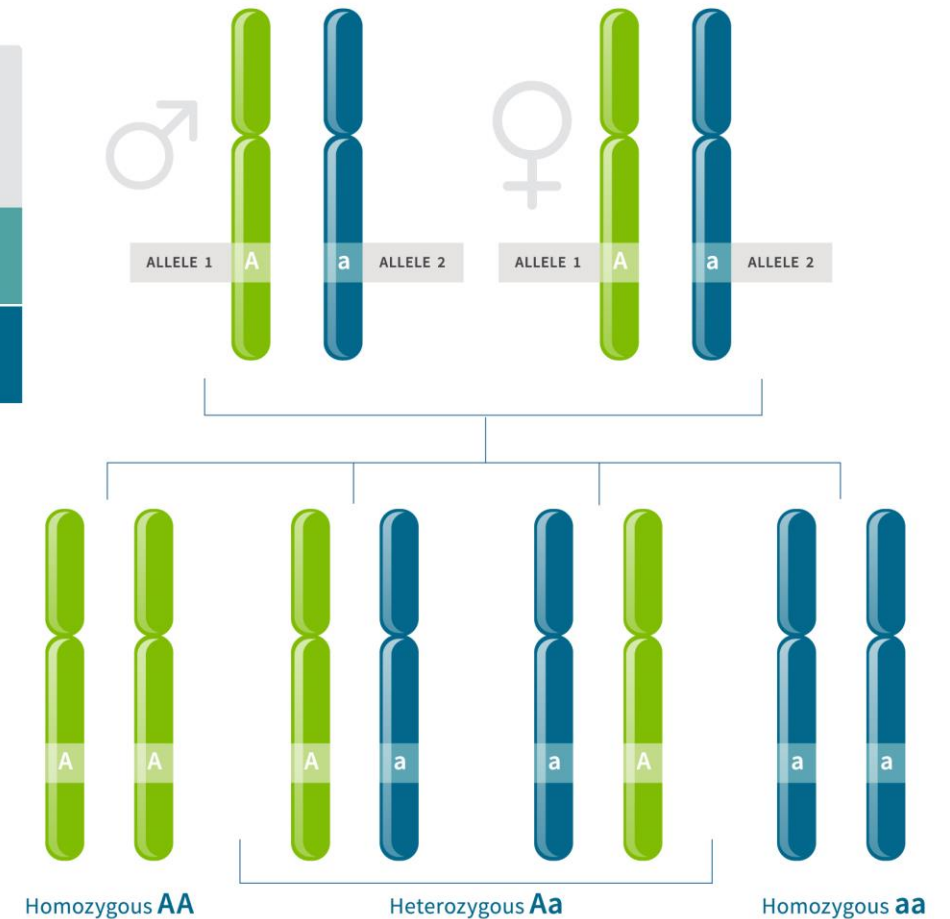
SNP genotypes are usually **encoded as 0, 1 or 2**, based on the number of copies of non-reference alleles.

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:

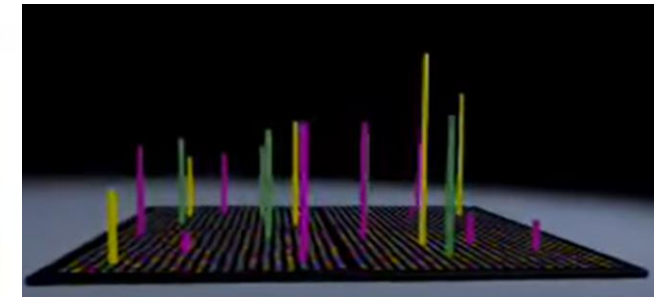
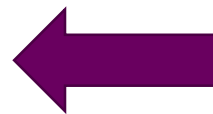
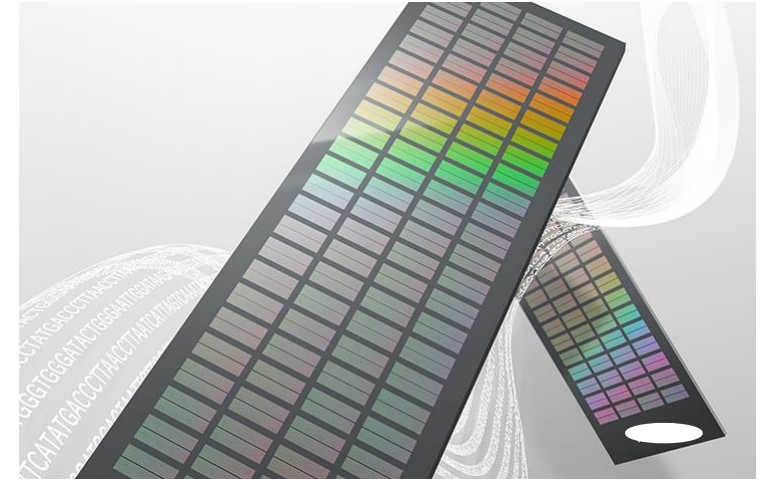
- For 1. =  $q^2$
- For 2. =  $2pq$
- For 3. =  $p^2$

		♂	
		A	a
♀	A	AA	Aa
	a	Aa	aa

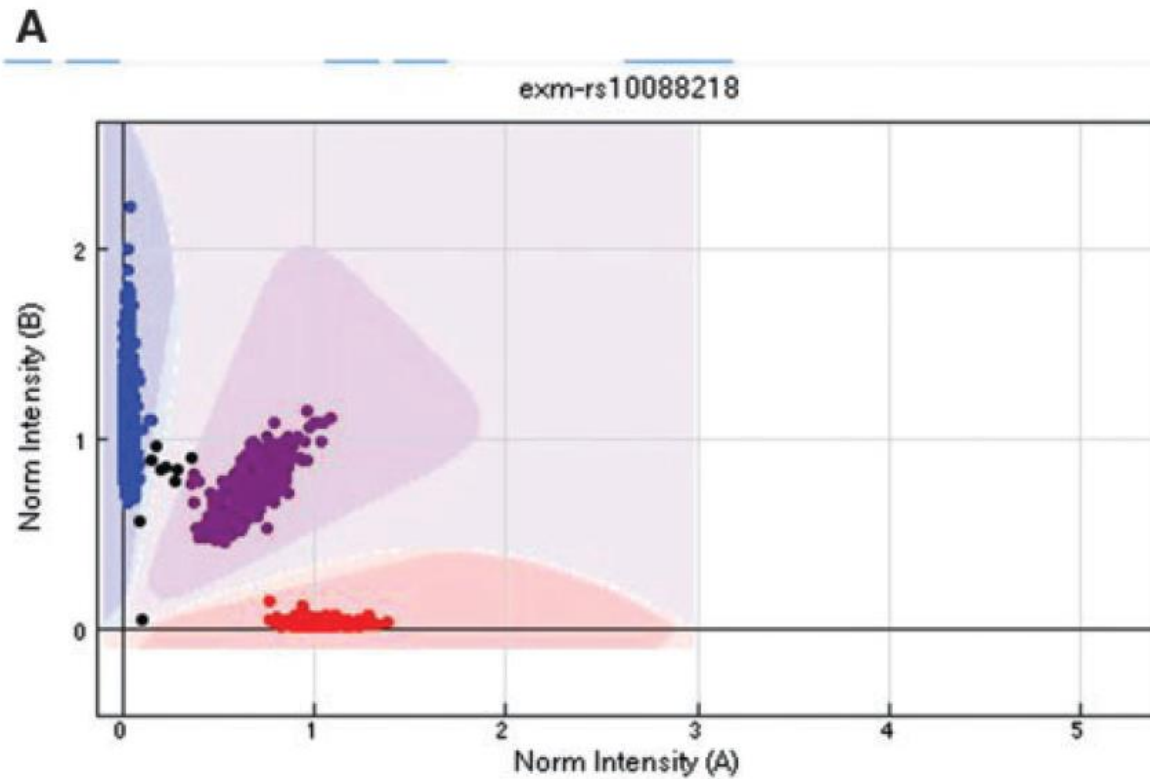


<https://www.ancestry.com/lp/genotype>

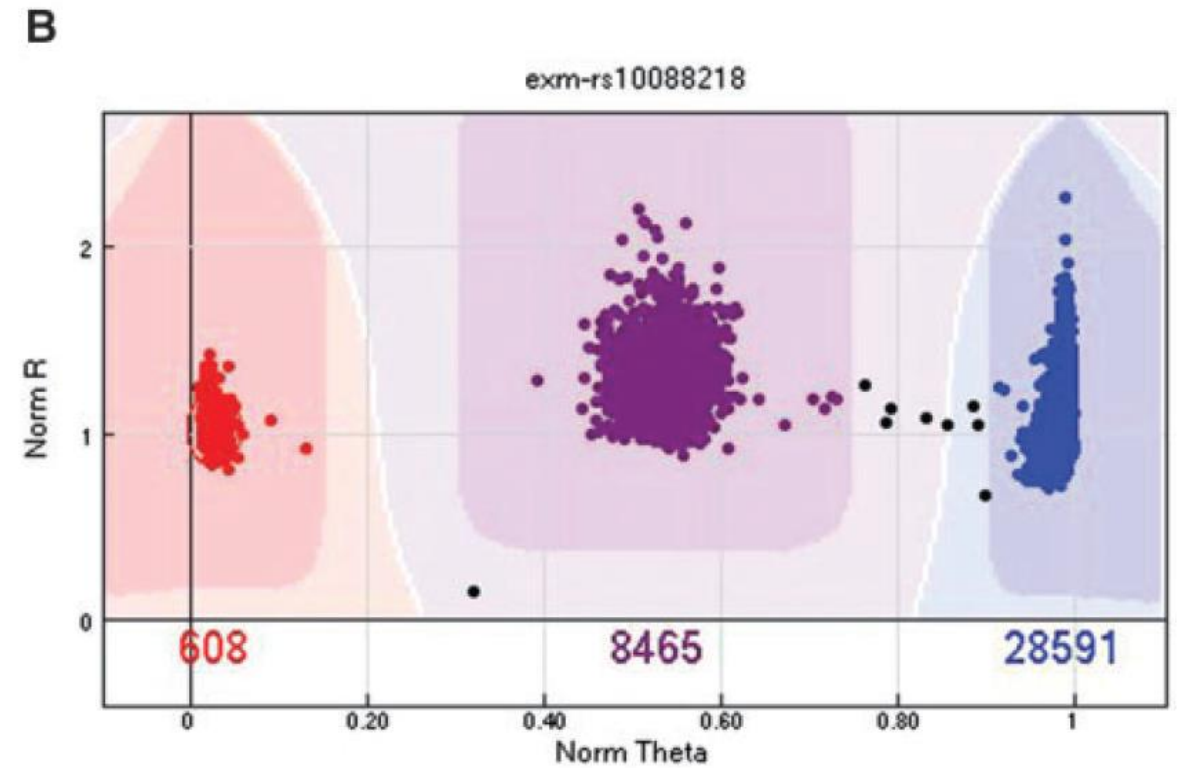
# Did you say intensities?



# Intensities: the good ...

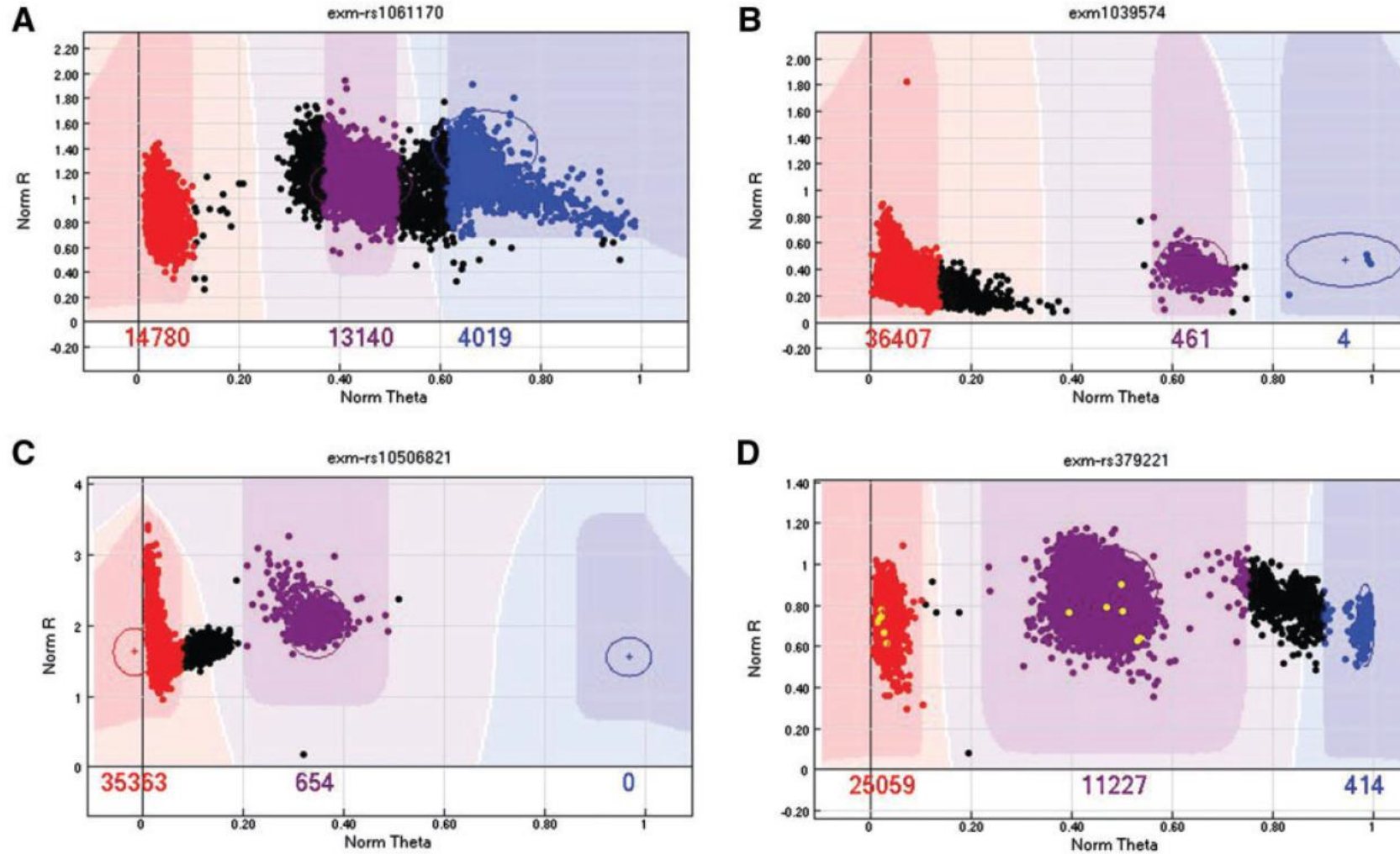


Cartesian coordinates



Polar coordinates

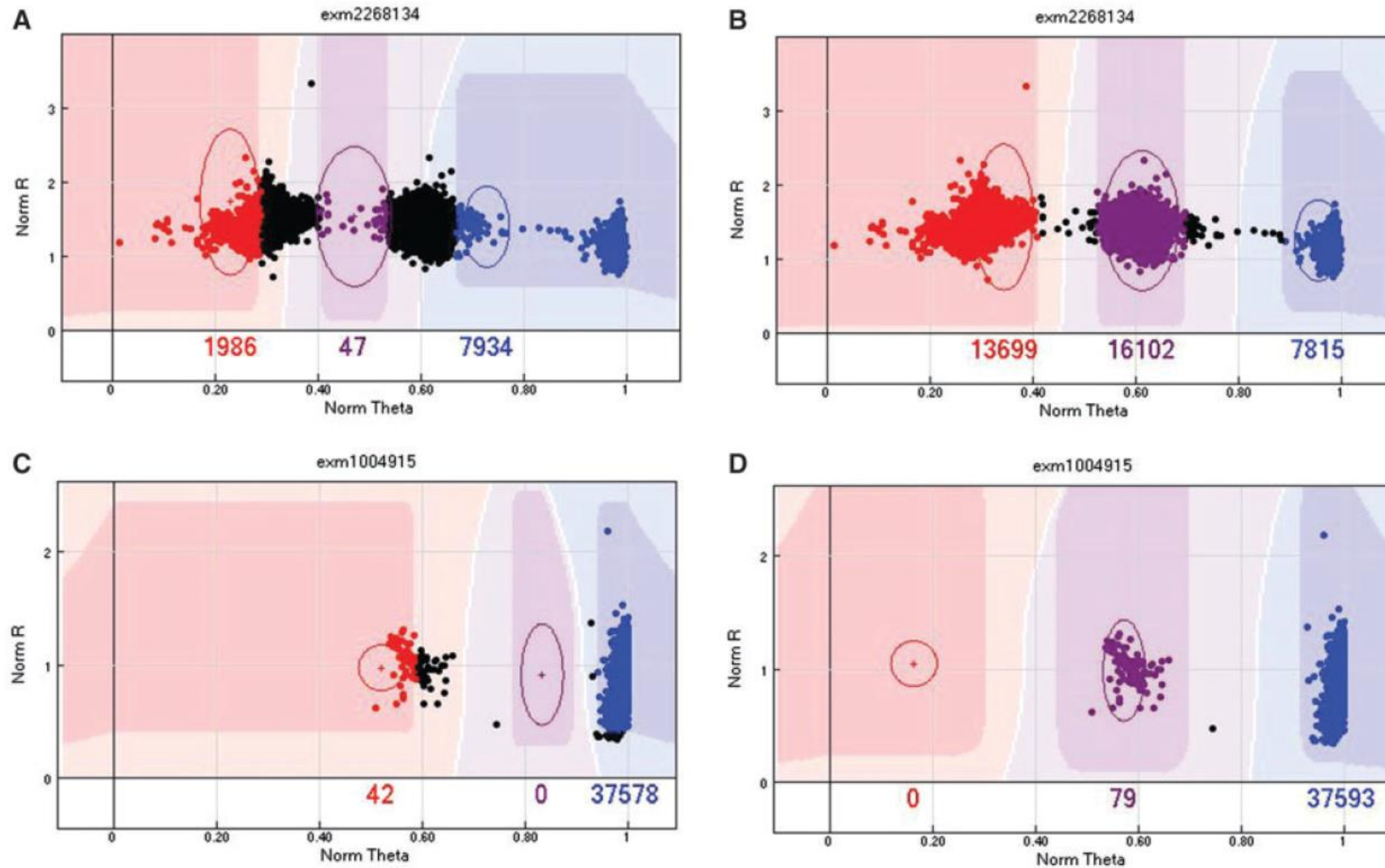
# Intensities: the bad ...



Zhao et al., 2018



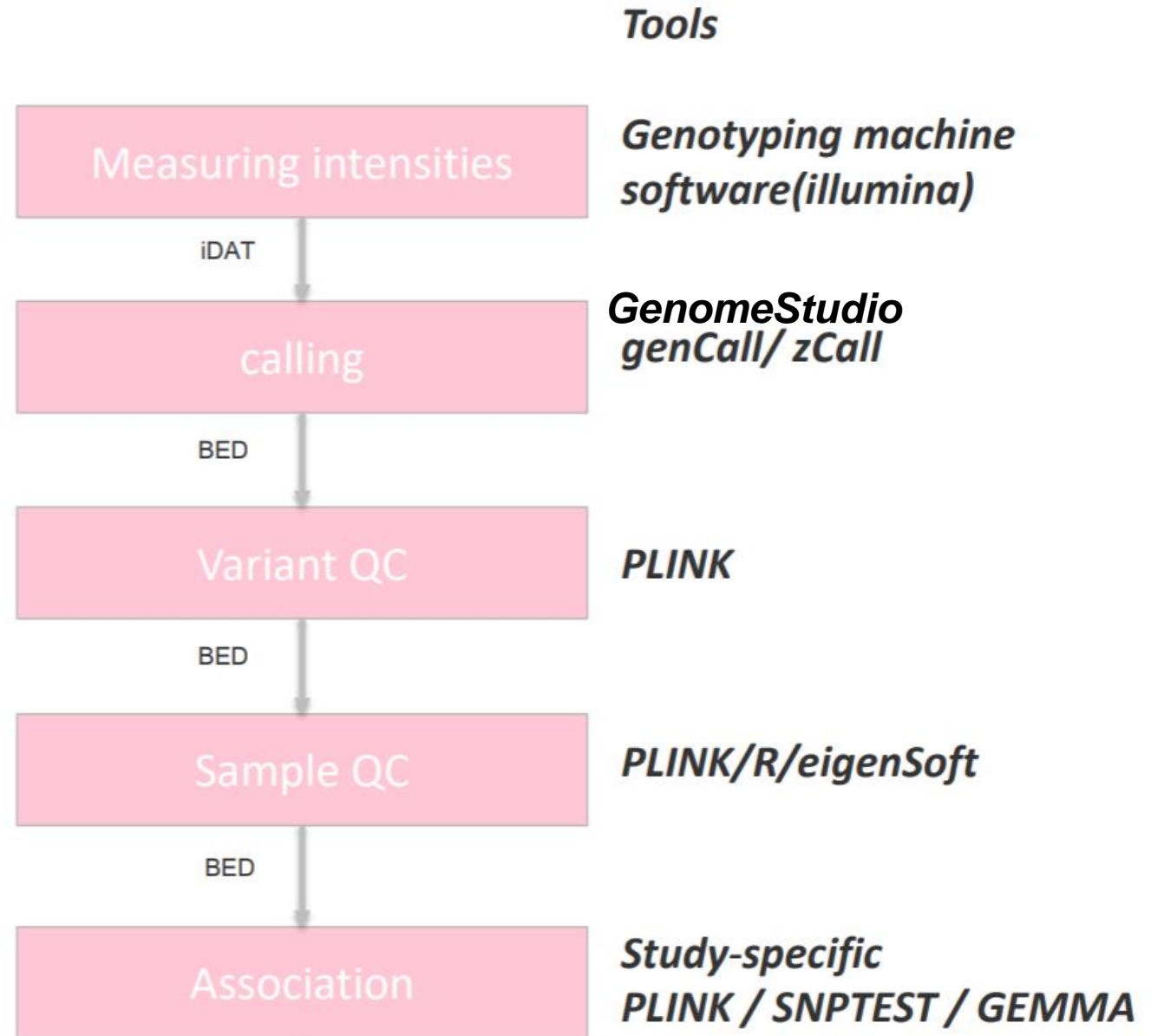
# Intensities: the ugly ...



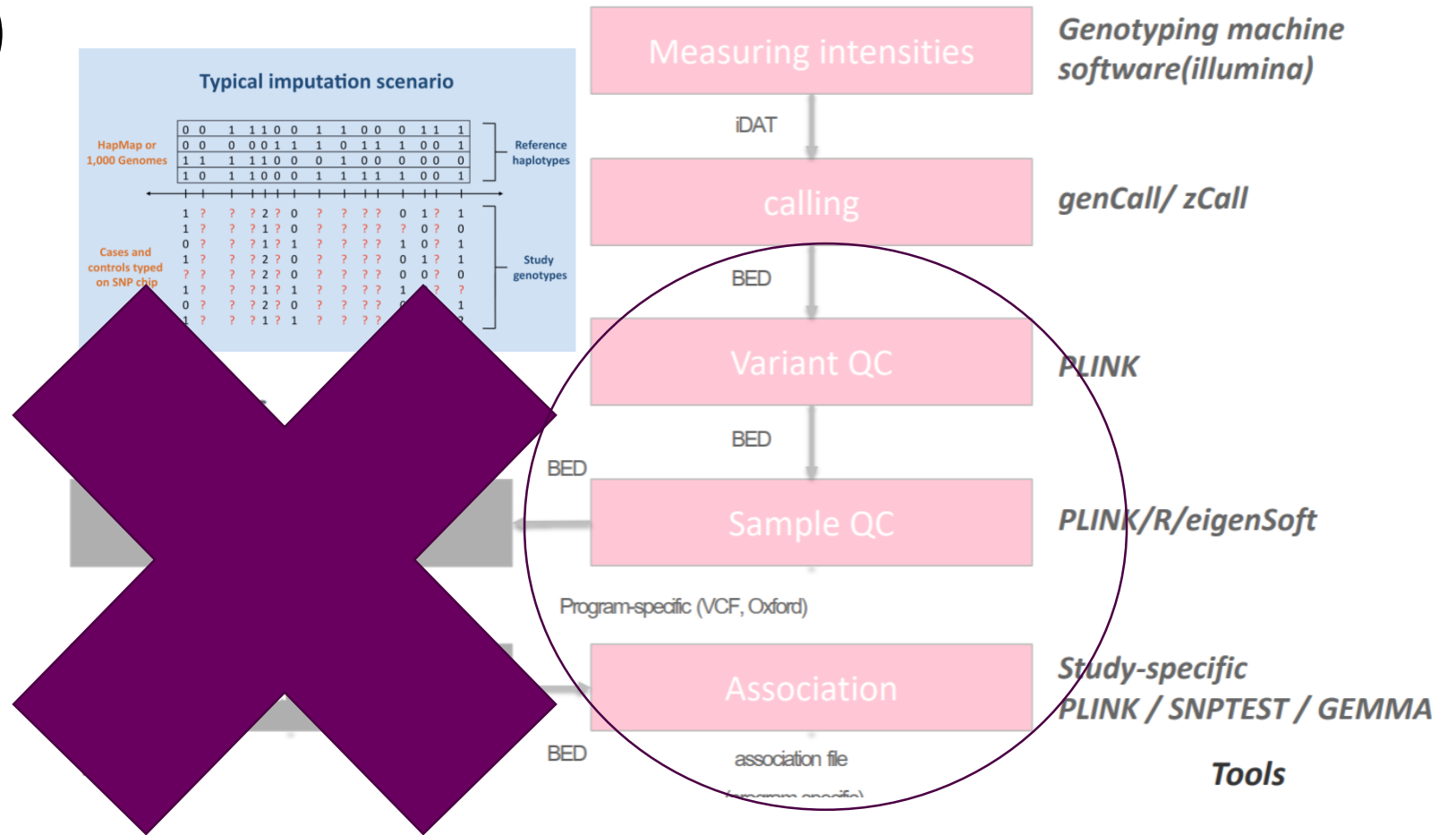
Zhao et al., 2018



# The GWAS analysis pipeline



# The (imputed) GWAS analysis pipeline



# Genotyping data storage

Which data types do we need?

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

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


*These stay constant (they describe the samples)*

*This one changes*

# Genotyping data storage: PLINK

Can either be text-format files or binary files.

*.ped									*.map			
FID	IID	PID	MID	Sex	P	rs1	rs2	rs3	Chr	SNP	GD	BPP
1	1	0	0	2	1	CT	AG	AA	1	rs1	0	870000
2	2	0	0	1	0	CC	AA	AC	1	rs2	0	880000
3	3	0	0	1	1	CC	AA	AC	1	rs3	0	890000

*.fam						*.bed		*.bim					
FID	IID	PID	MID	Sex	P	Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)		Chr	SNP	GD	BPP	Allele 1	Allele 2
1	1	0	0	2	1			1	rs1	0	870000	C	T
2	2	0	0	1	0			1	rs2	0	880000	A	G
3	3	0	0	1	1			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 (morgan)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)





# Genotyping data storage: PLINK

*.ped									*.map			
FID	IID	PID	MID	Sex	P	rs1	rs2	rs3	Chr	SNP	GD	BPP
1	1	0	0	2	1	CT	AG	AA	1	rs1	0	870000
2	2	0	0	1	0	CC	AA	AC	1	rs2	0	880000
3	3	0	0	1	1	CC	AA	AC	1	rs3	0	890000

ped(igree) file has **6+2n**, 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						*.bed		*.bim					
FID	IID	PID	MID	Sex	P	Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)		Chr	SNP	GD	BPP	Allele 1	Allele 2
1	1	0	0	2	1			1	rs1	0	870000	C	T
2	2	0	0	1	0			1	rs2	0	880000	A	G
3	3	0	0	1	1			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.

- **bim** (binary mapping) file is the .map file plus two columns, providing the A1 and A2 alleles

- 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)



# Genotyping data storage: PLINK

What is left?

Matrix file  
(program-specific)

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

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



# Genotyping data : PLINK common operations

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

 lisa.surfsara.nl - PuTTY

```
amarees@login1:~/genetic_data$ plink --bfile MY_DATA --assoc --out gwas_results
```

Path to the directory  
containing your files\*

Indicate the  
usage of PLINK\*\*

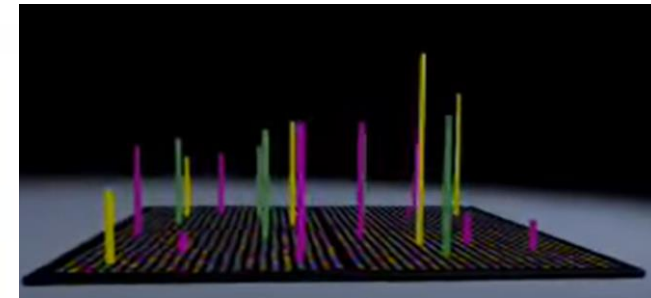
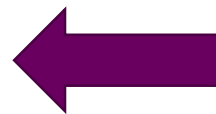
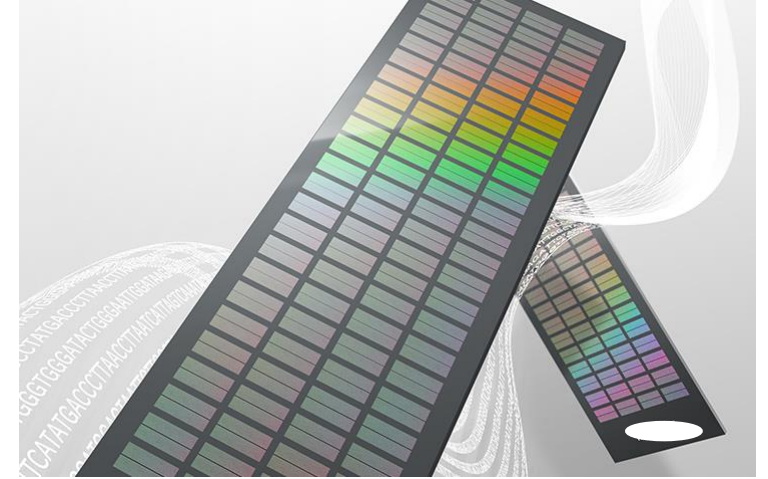
Specify the  
input file name

Specify the  
options

Specify the  
output filename



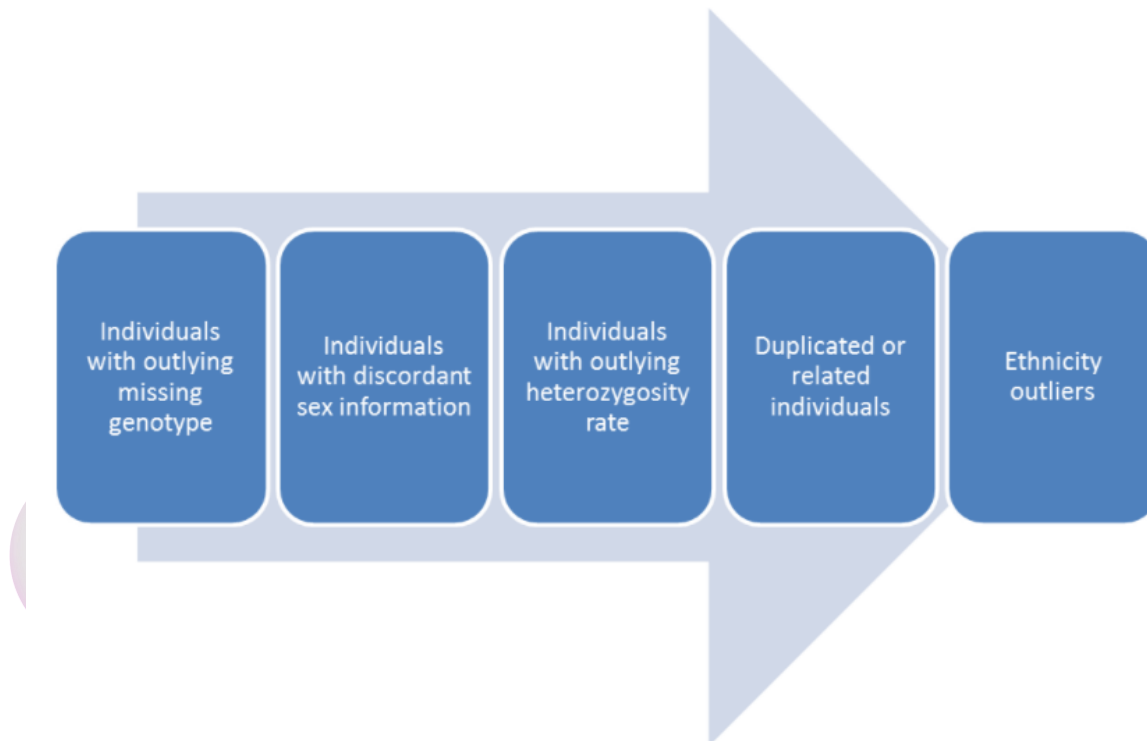
# Why Quality Control?



# 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)

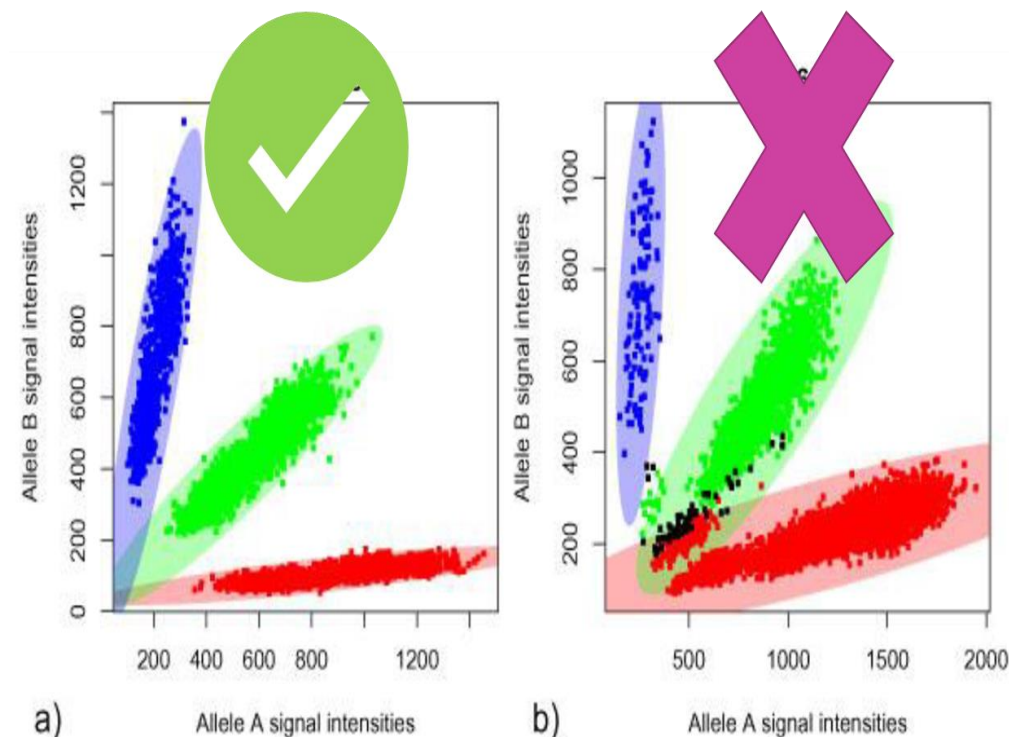
# (A generalized) Quality Control process

## 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

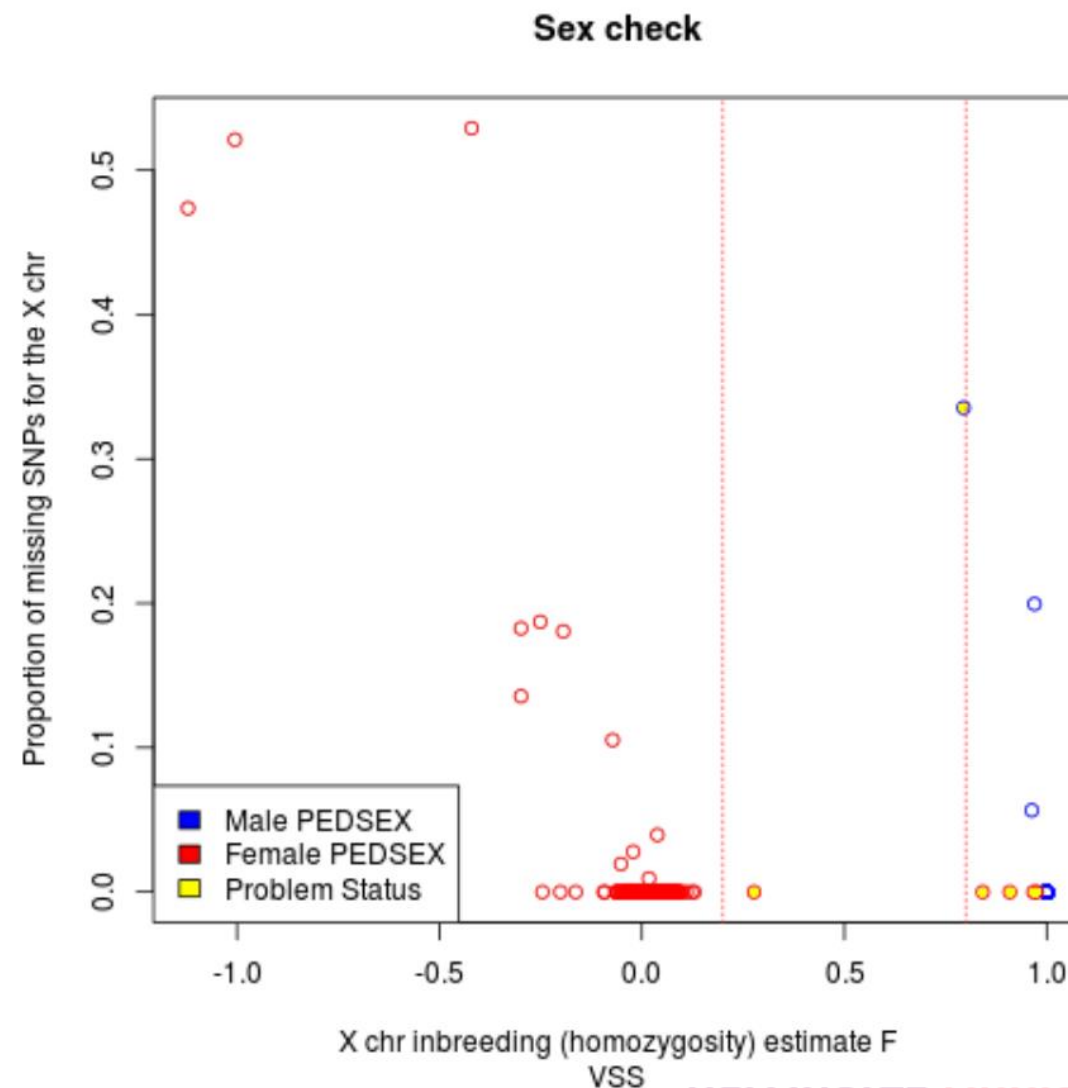
`--check-sex`

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

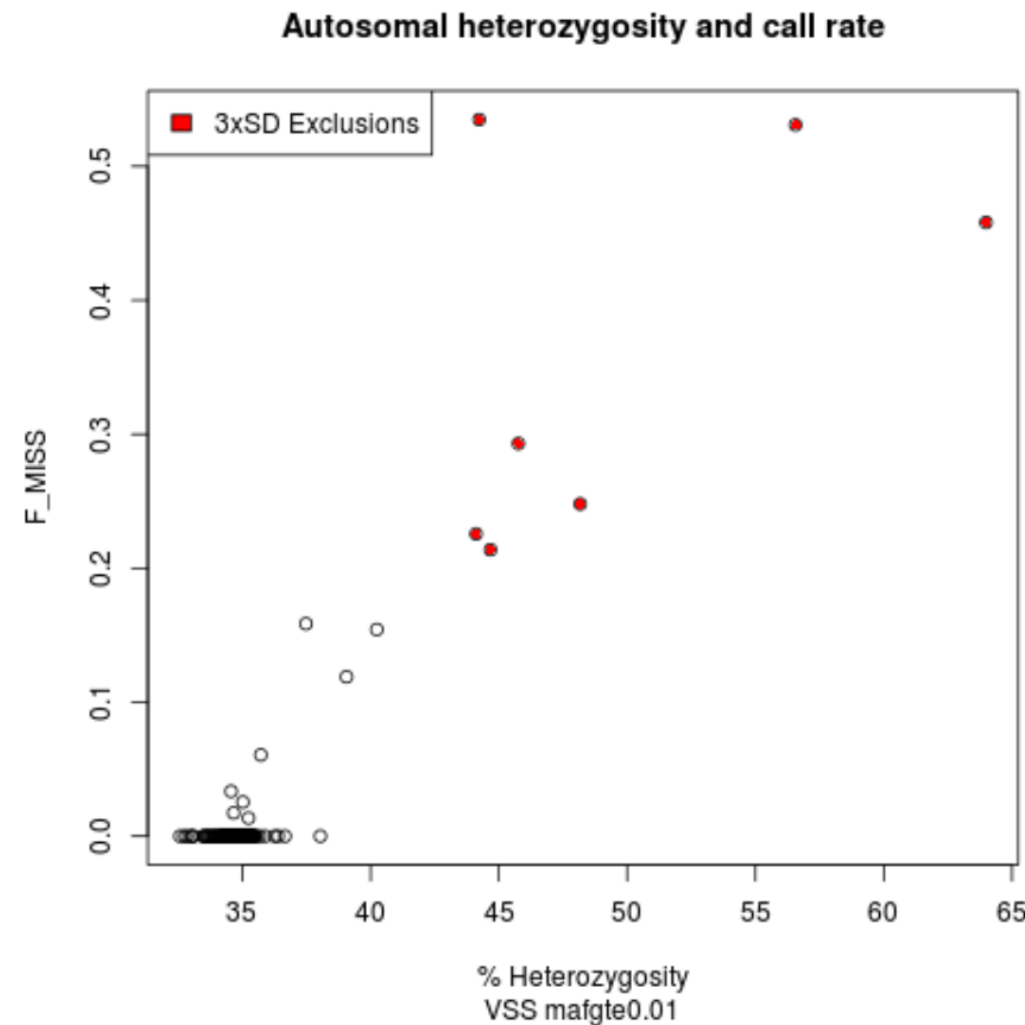
- The proportion of heterozygous genotypes (per sample)
- Various ways of calculating the rate

PLINK: ( $\text{<observed hom. count>} - \text{<expected count>}$ ) / ( $\text{<total observations>} - \text{<expected count>}$ )

--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



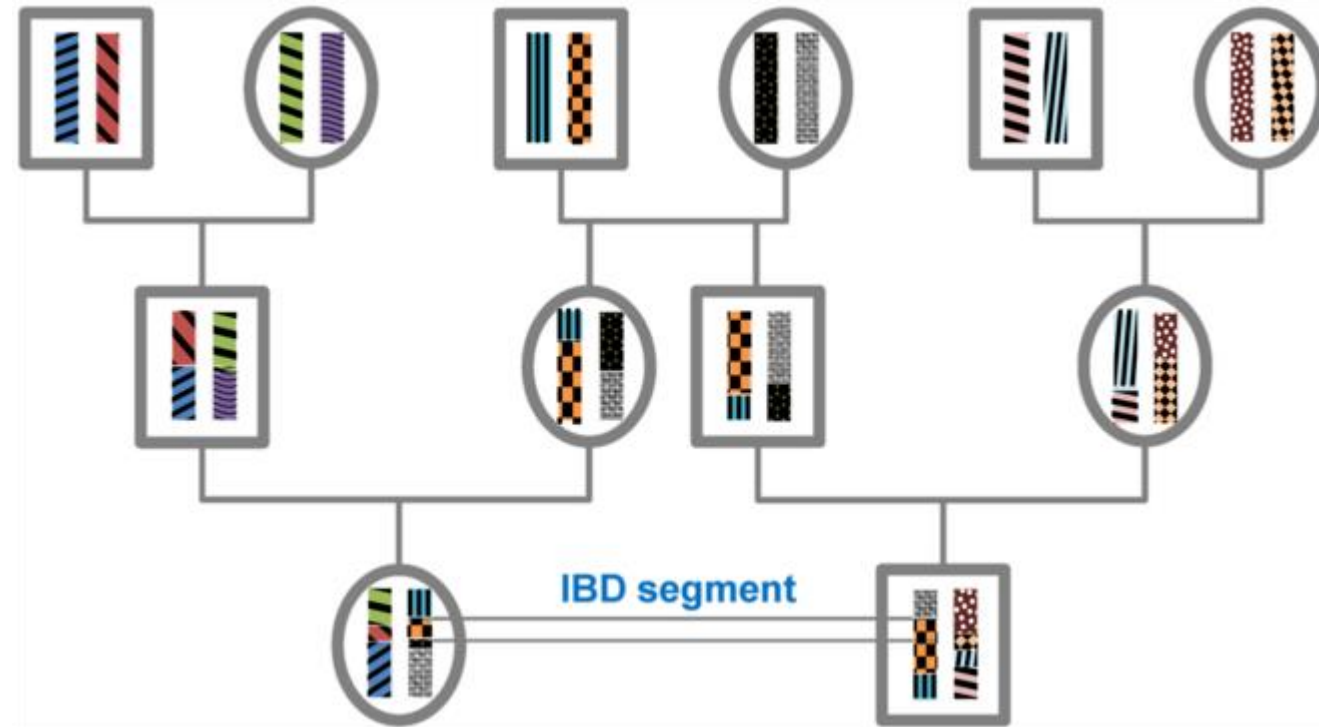
# (A generalized) Quality Control process



## 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.



# (A generalized) Quality Control process

## Duplicated or related individuals

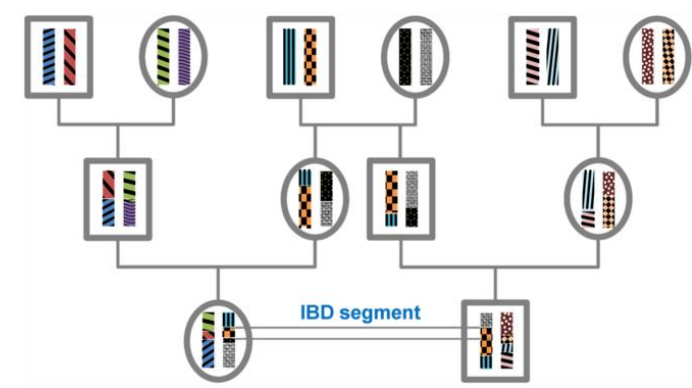
PLINK calculates identity by descent (IBD) of all sample

Approximates the percentage IBD overall, representing pairs as s

- Zero alleles IBD (z0)
- One allele IBD (z1)
- Two alleles IBD (z2)

PI\_HAT (the proportion IBD, defined as  $P(\text{IBD} = 2) + 0.5 * P(\text{IBD} = 1)$ )

Relationship type	z0	z1	z2	PI_HAT
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



--genome

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.

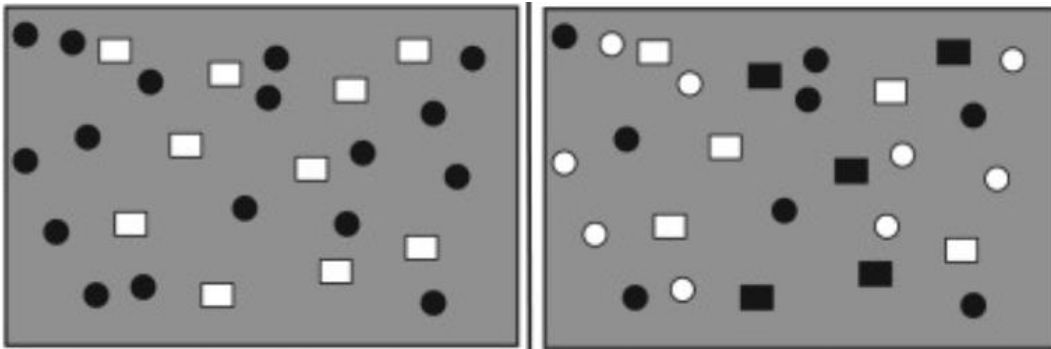




# (A generalized) Quality Control process

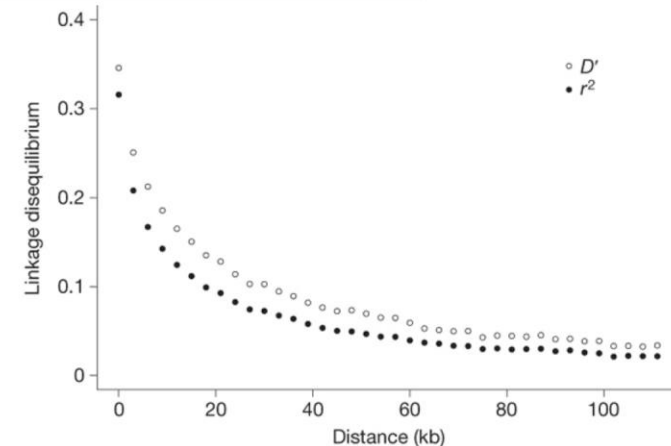
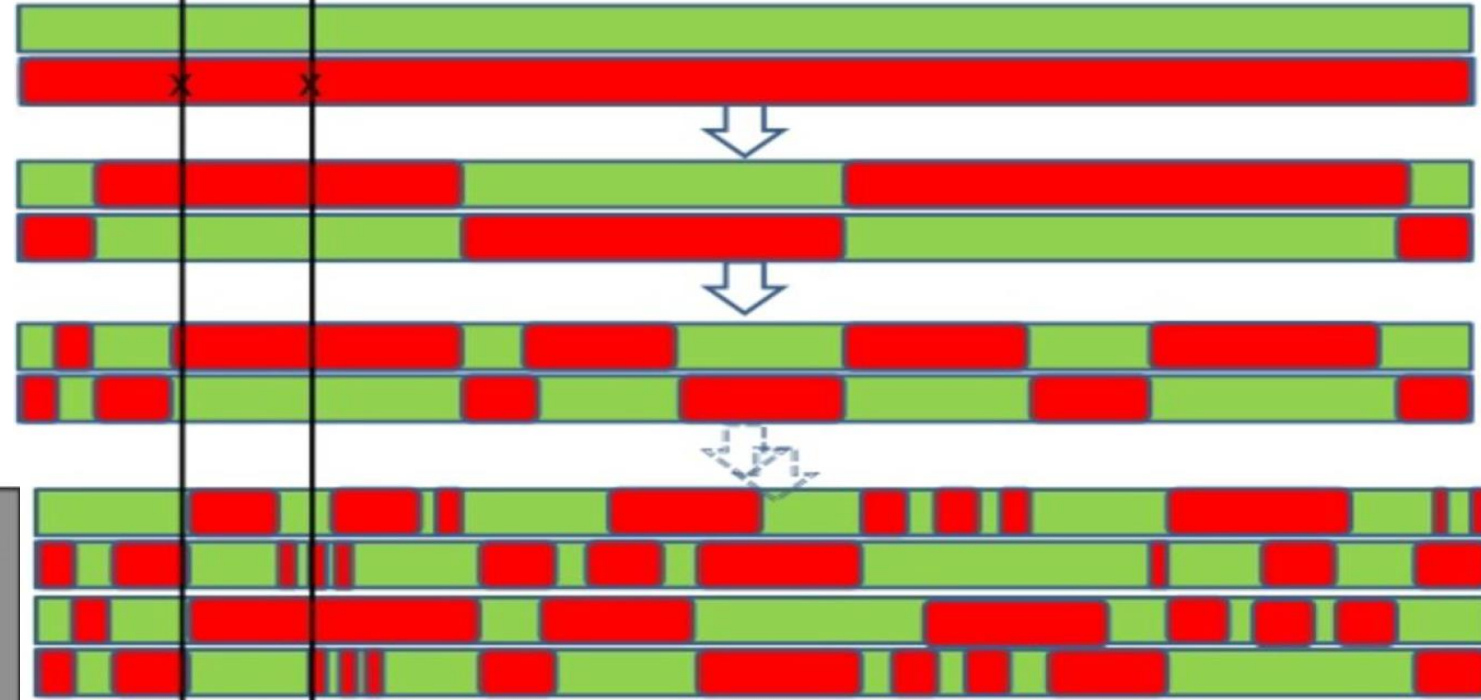
## Linkage disequilibrium (LD)

Is the non-random association of alleles at different loci in a given population.



In GWAS we (mainly) use correlation coefficient between pairs of loci,  $r^2$

$r^2=1$  is perfect LD

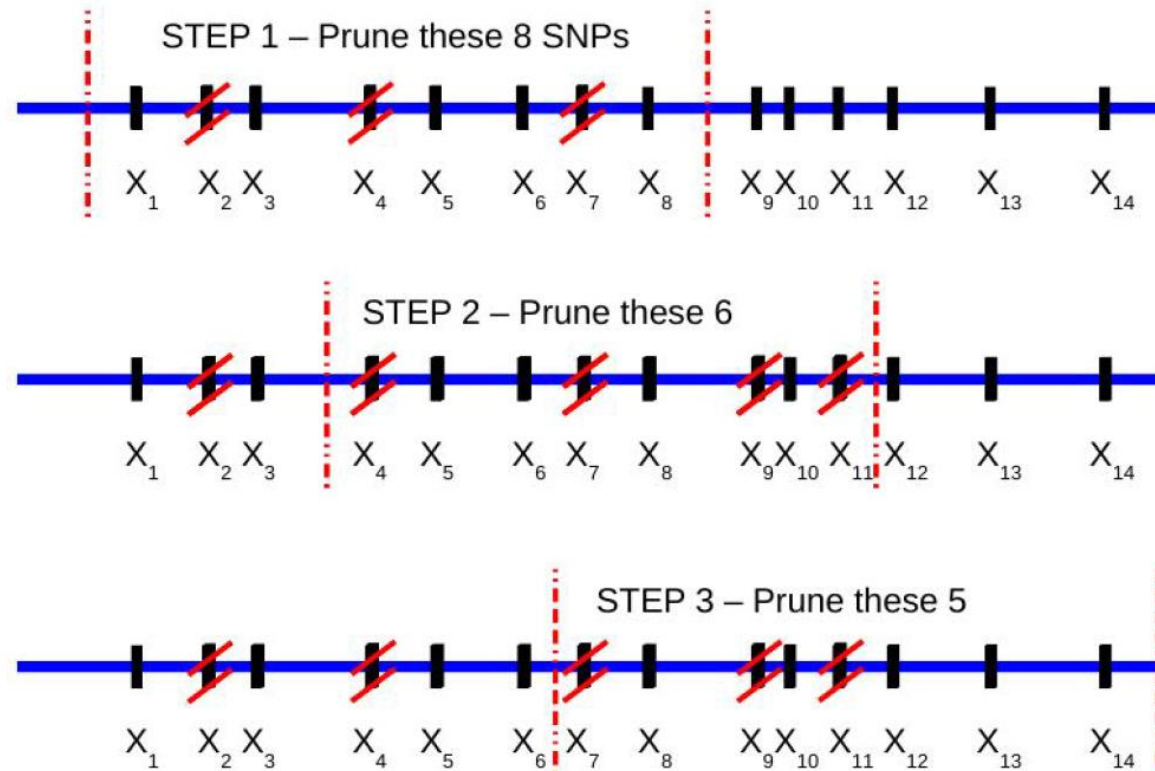


# (A generalized) Quality Control process

## PLINK: LD-based SNP pruning

```
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

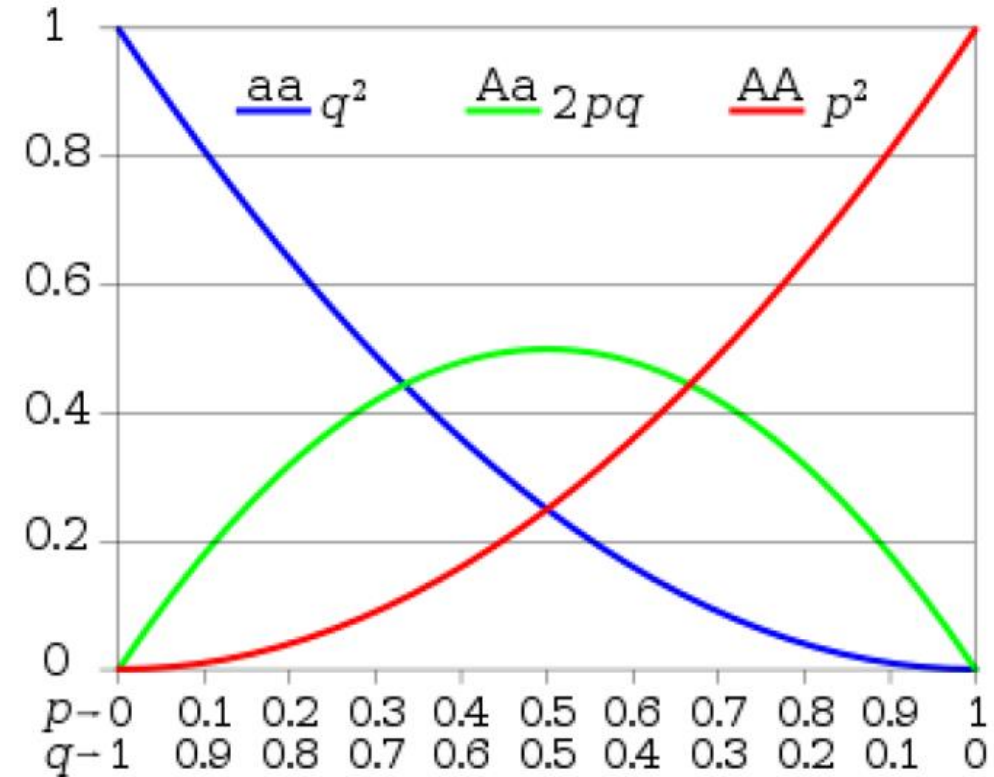
## The Hardy–Weinberg (dis)equilibrium (HWE) law:

The genotype and the allele frequencies are constant over generations.

Assumes:

- An indefinitely large population
- With no selection, no Mutation, no Migration .....

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 SNPs

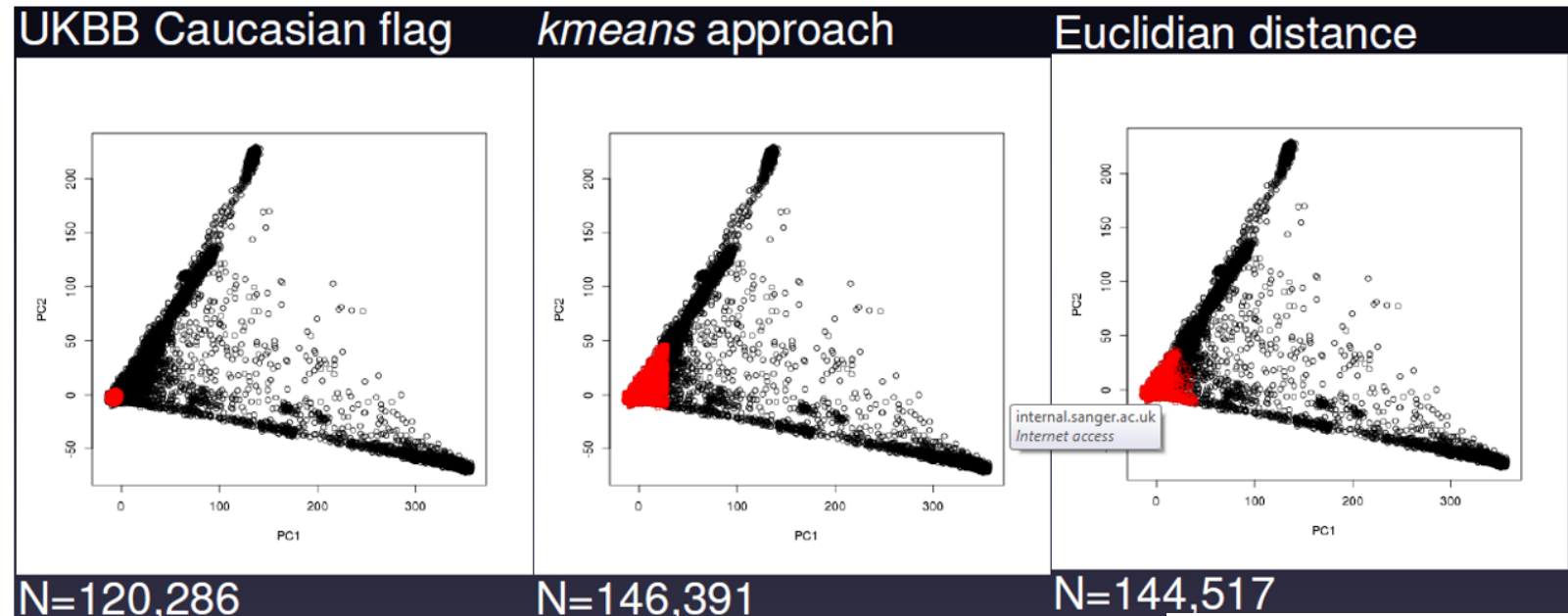
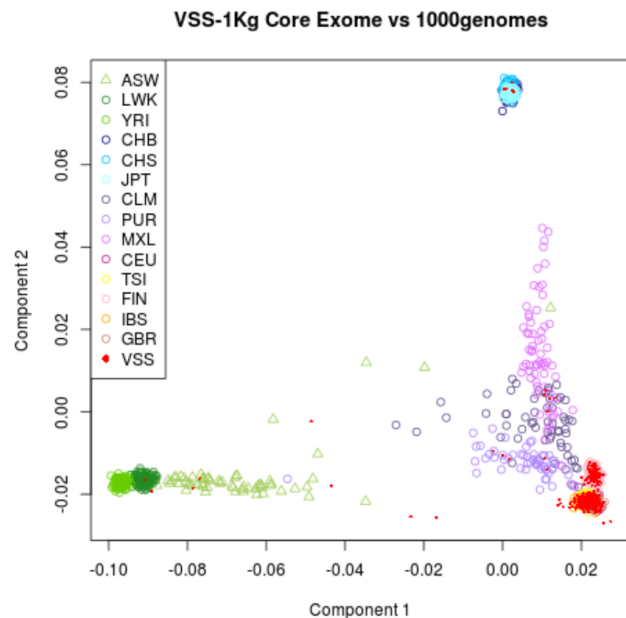
# (A generalized) Quality Control process

## 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).



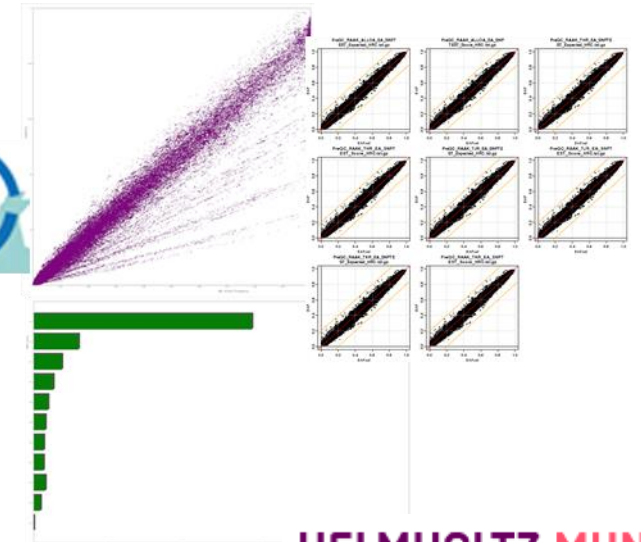
# (A generalized) Quality Control process

## 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

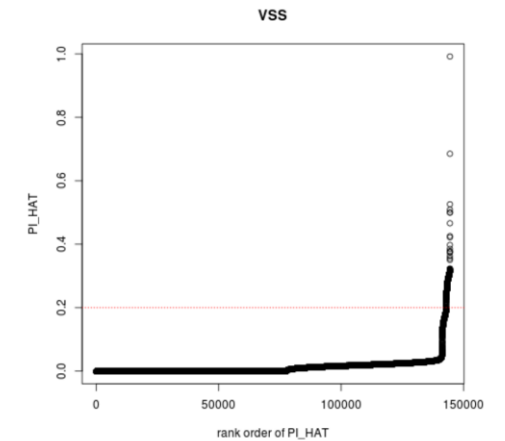
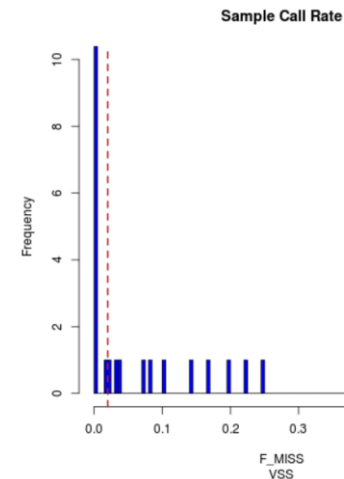
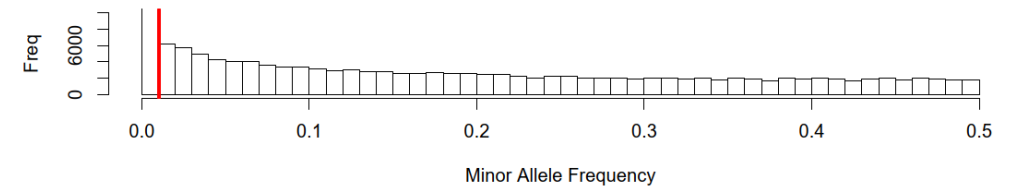
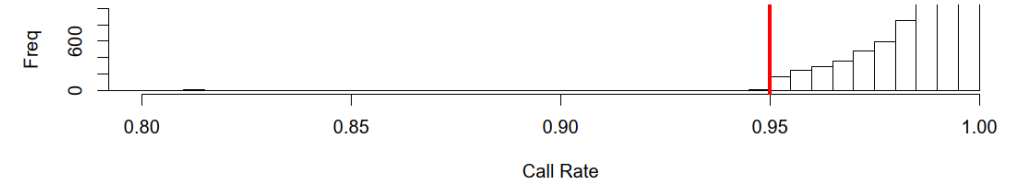
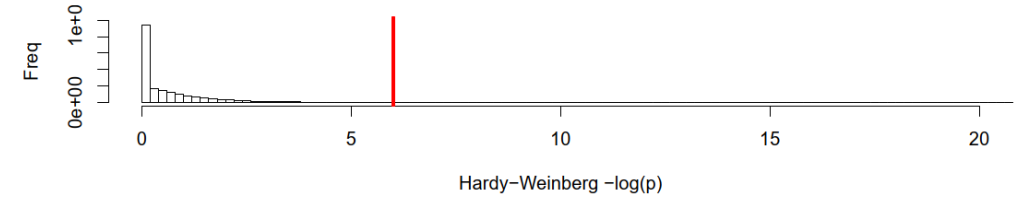
Variant Quality Control





# (A generalized) Quality Control process

Where to draw  
the line?



# Genotyping data : PLINK common operations

## Sample management

<code>--keep [file]</code>	Keep samples in file
<code>--remove [file]</code>	Remove samples in file

## SNP management

<code>--extract [file]</code>	Keep SNPs in file
<code>--exclude [file]</code>	Remove SNPs in file

## Extracting regions

<code>--chr [name]</code>	Extract data on specified chromosome
<code>--from-bp [pos]</code>	From specified position
<code>--to-bp [pos]</code>	To specified position

# Genotyping data : PLINK common operations

## Variant QC

<code>--maf [threshold]</code>	Keep variants with $MAF > \text{threshold}$
<code>--hwe midp [threshold]</code>	Keep variants with HWE $p > \text{threshold}$

## Sample QC

<code>--missing</code>	Compute per-sample and per-variant missingness
<code>--check-sex</code>	Check sexes by looking at chrX
<code>--genome</code>	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.





Thank you.