

GWAS

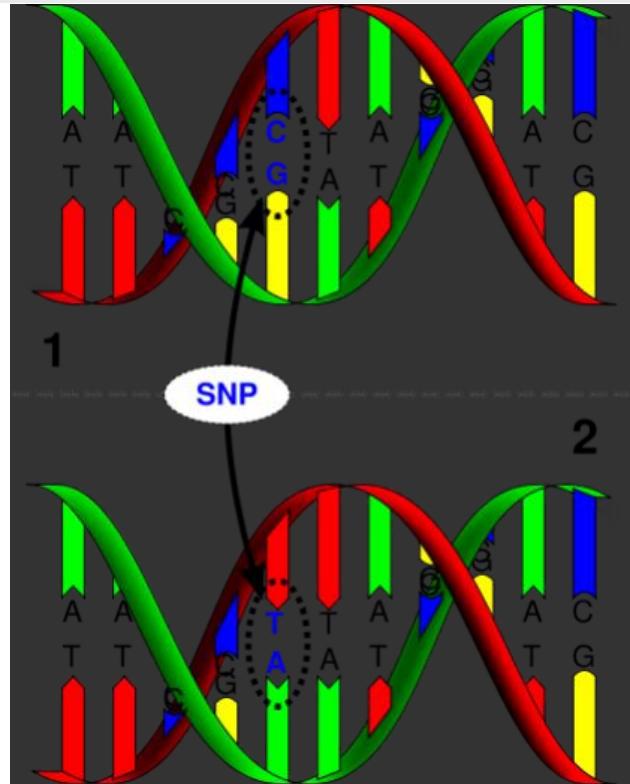
Tools and best practices for performing genome-wide scans

Presentation outline

- Some key terms and concepts
- Brief introduction to genome-wide association analysis
- The GWAS analysis pipeline
- Phenotype QC & transformation
- Association analysis
- Visualising and annotation of results

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 $p=1-q$.

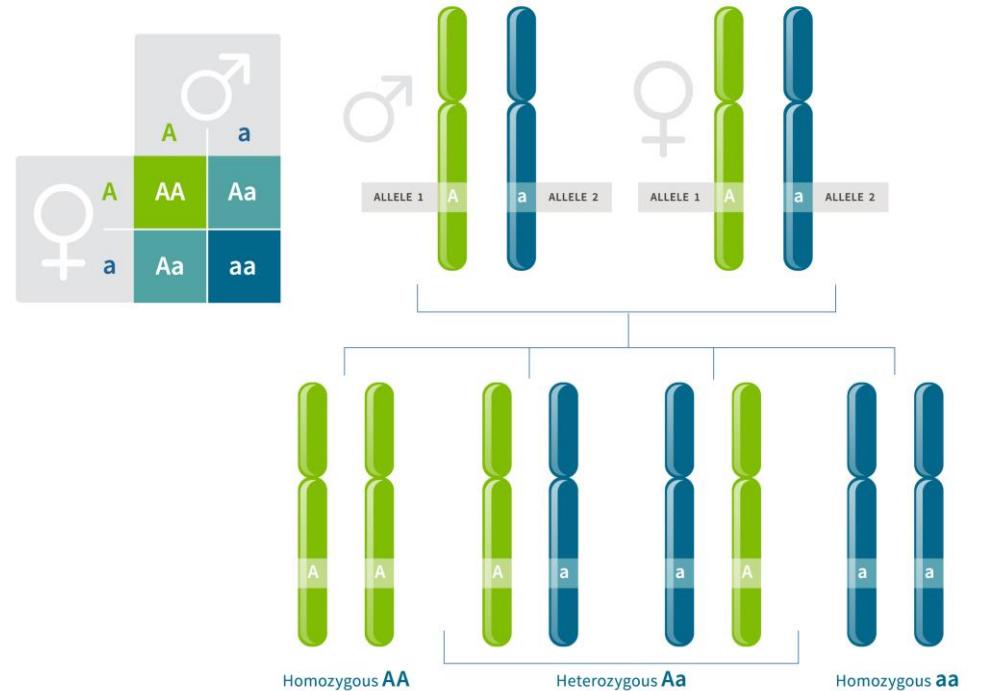


Reference Allele

Non-
Reference
Allele

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)



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

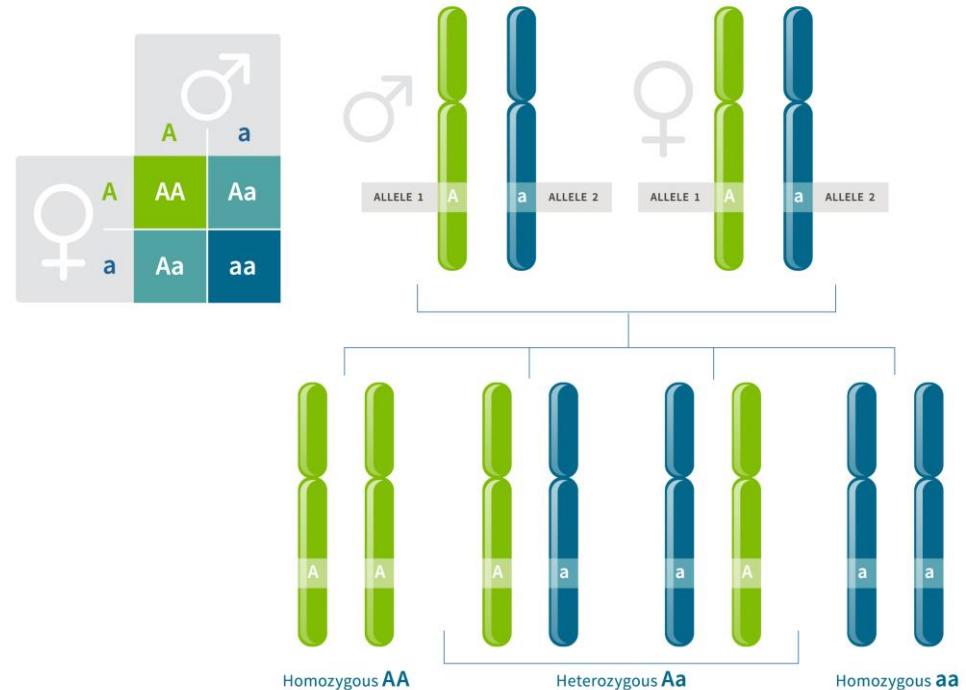
Before we start ...

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

- genotype TT is coded as 0 (homozygous non-reference)
- genotype CT is coded as 1 (heterozygous)
- genotype CC is coded as 2 (homozygous reference)

- Genotypes frequency:**

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



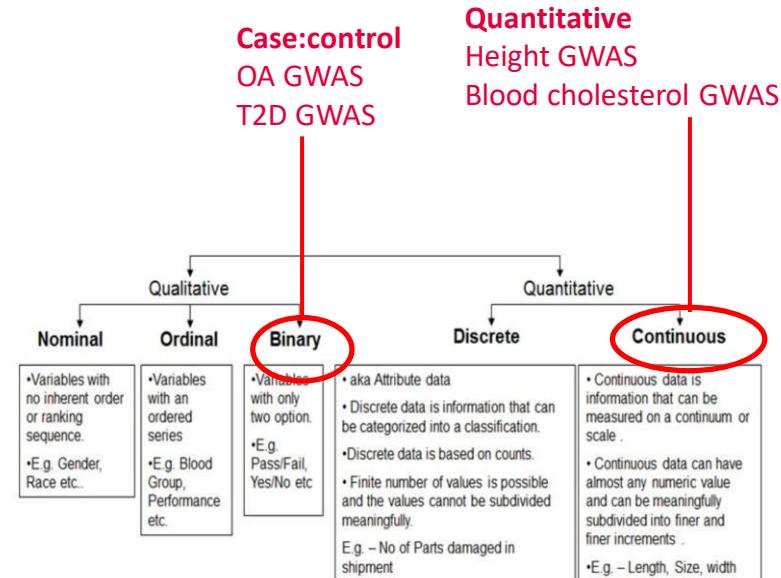
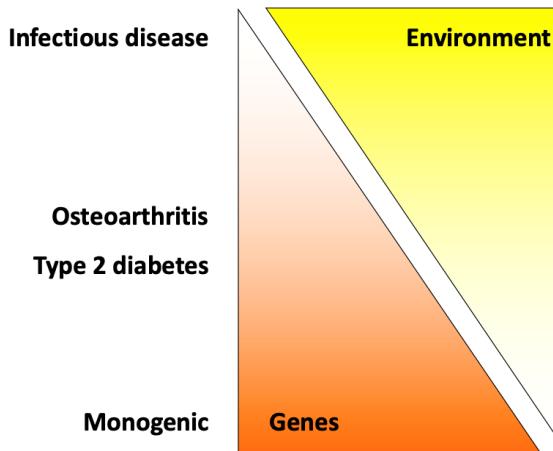
<https://www.ancestry.com/lp/genotype>

Brief introduction to GWAS

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

- Trait of interest (phenotype)
- Individuals
- Genotypes

Genes vs environment



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

CARTOGENE NAKO GESUNDHEITS-STUDIE

BIOBANK JAPAN BioBank Japan(BBJ)

Founded by Japan Agency for Medical Research and Development
BioBank Japan Project for Genomic and Clinical Research

South Asia Biobank

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4945521/pdf/40064_2016_Article_2723.pdf

biobank
Enabling scientific discoveries that improve human health

Consortia



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

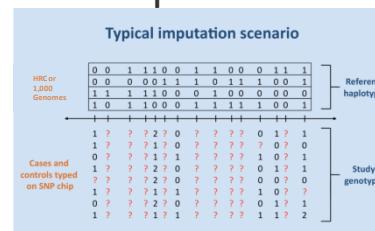
WES → ~40 million bases

Array → 0.24 -4.5 million bases



Infinium Asian
Screening Array

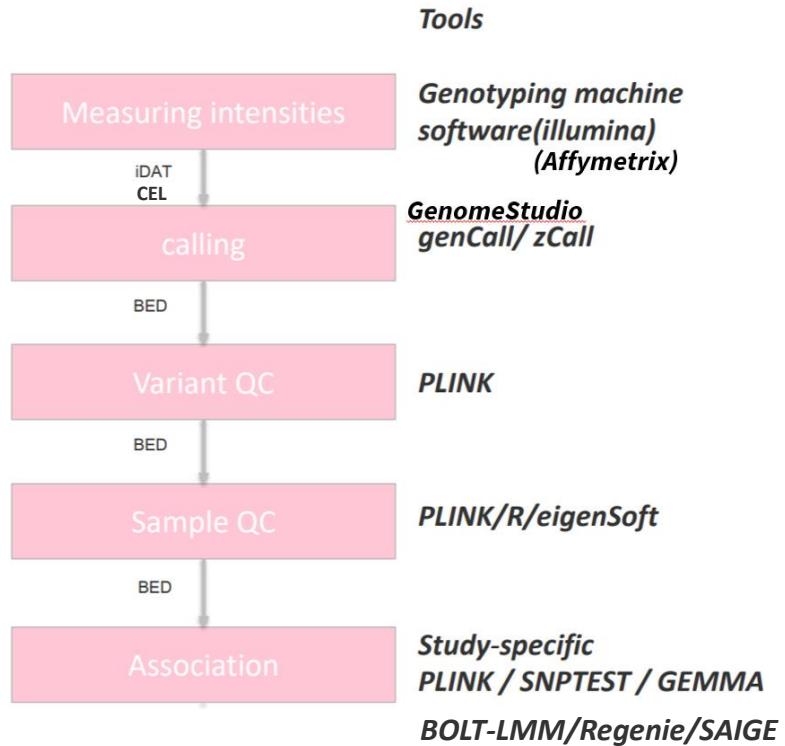
+ Imputation



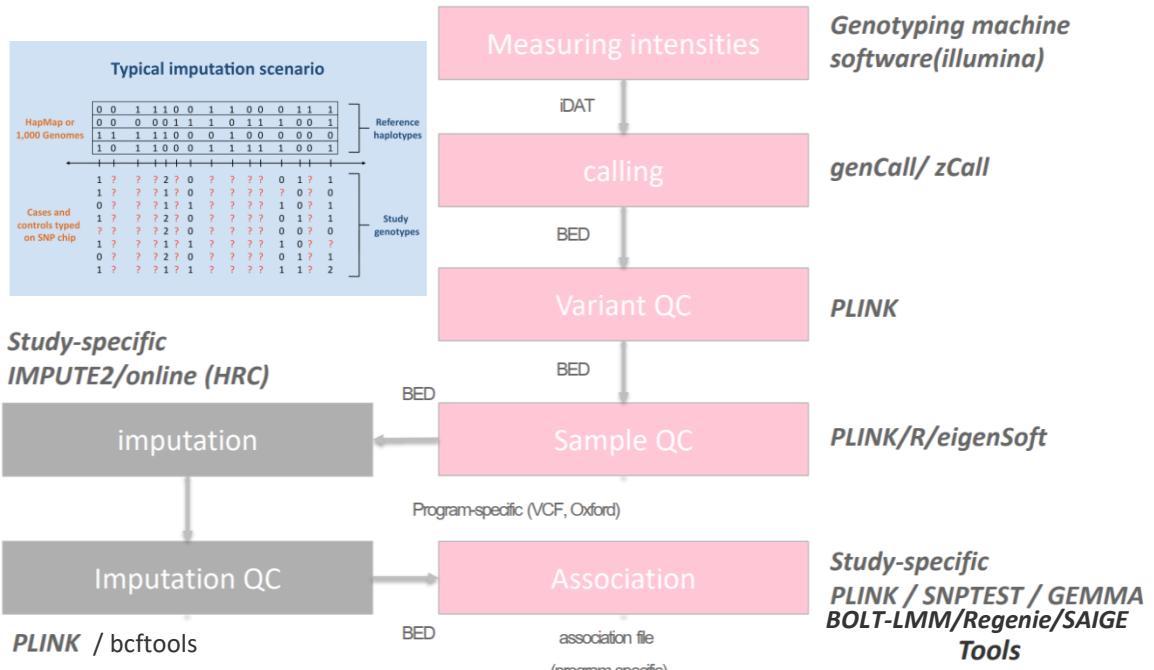
→ ~40-90 million probability based

↓
~12-20 million good quality

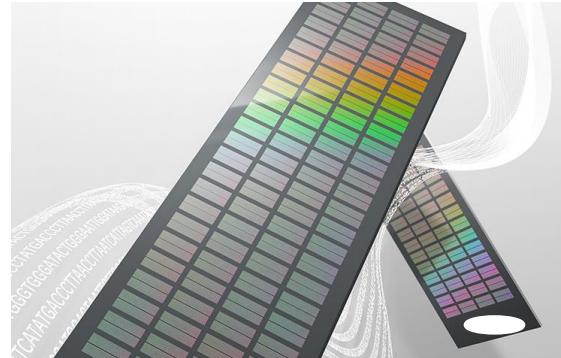
The GWAS analysis pipeline



The (imputed) GWAS analysis pipeline

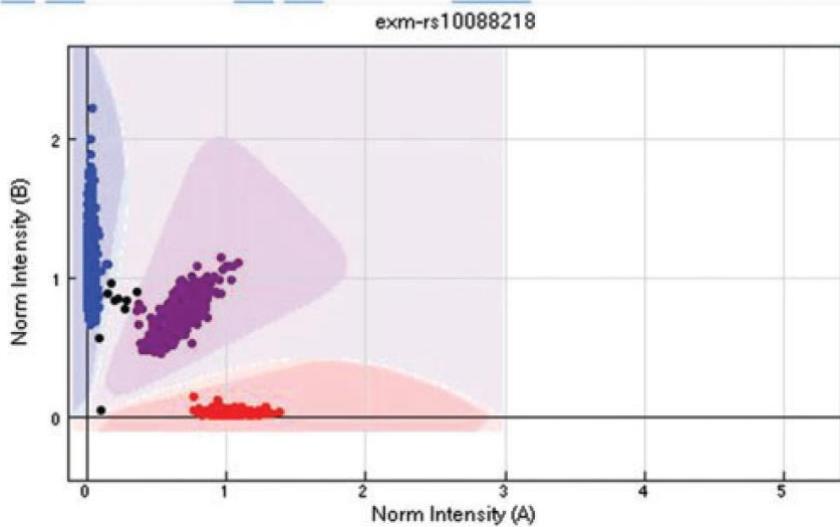


Did you say intensities?

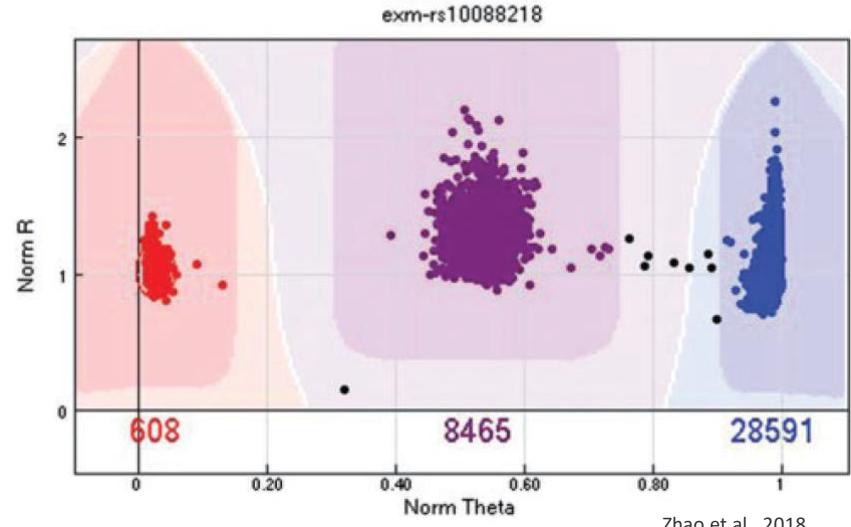


Intensities: the good ...

A

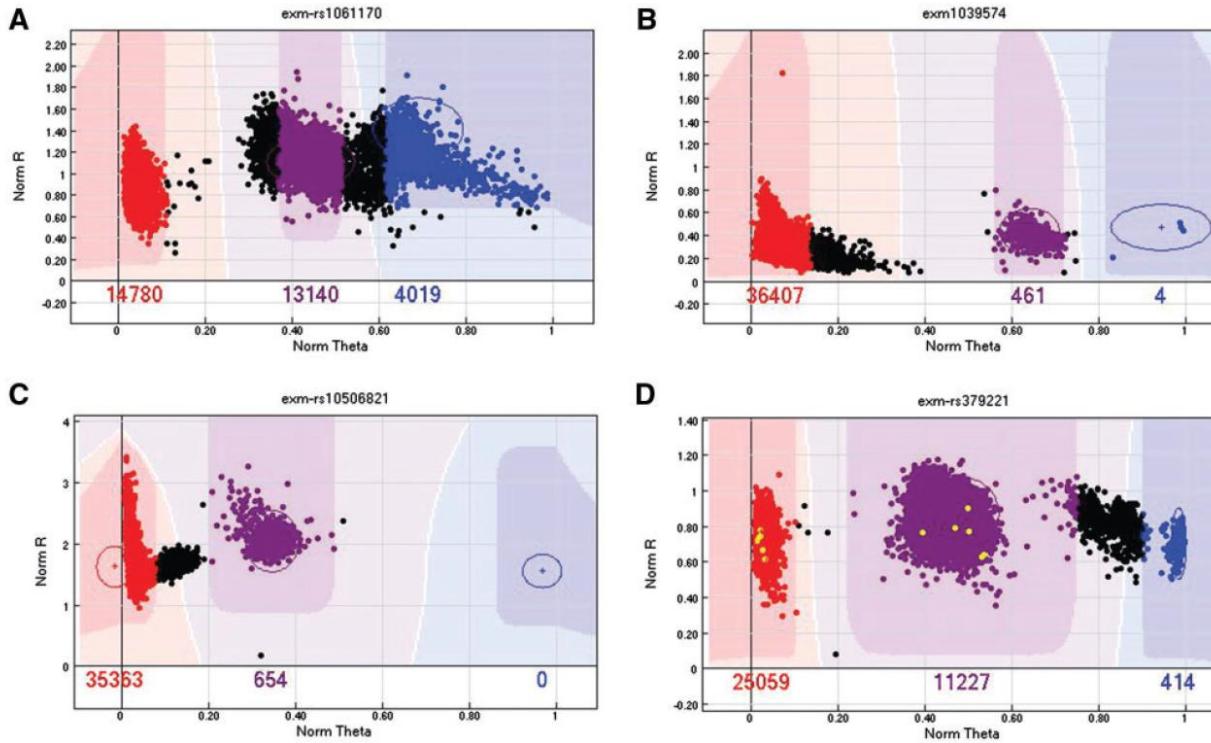


B



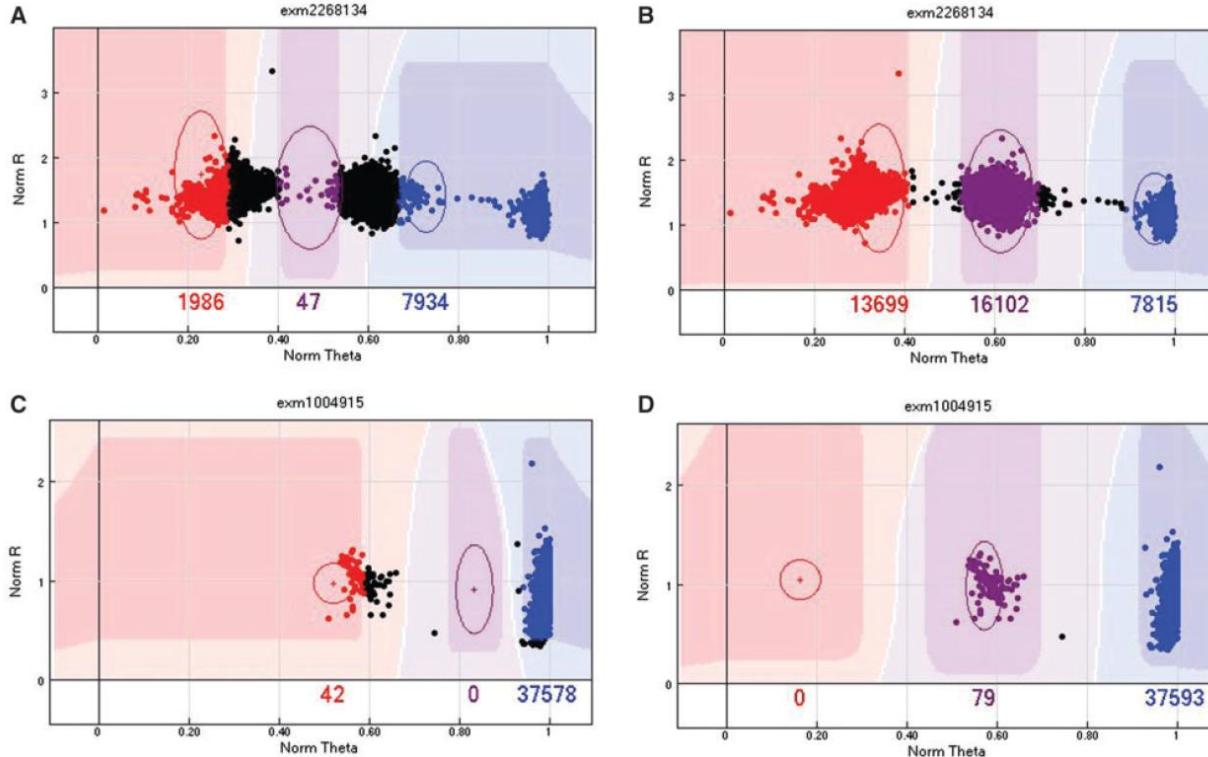
Zhao et al., 2018

Intensities: the bad ...



Zhao et al., 2018

Intensities: the ugly ...



Zhao et al., 2018

Genotyping data storage

- Which data types do we need?

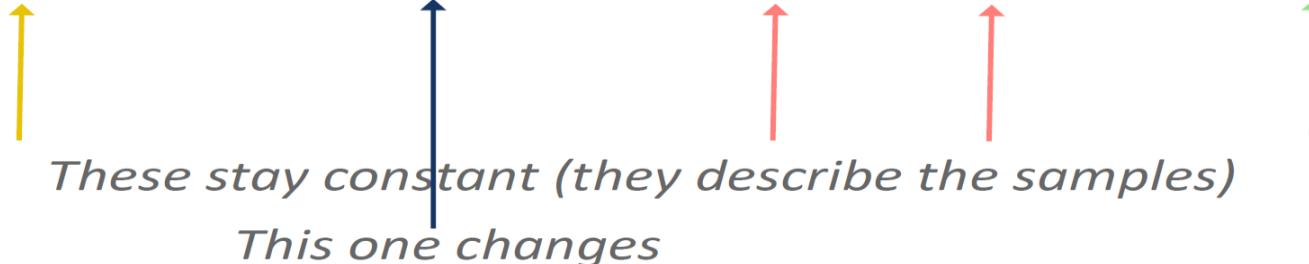
phenotype $\sim \beta \times$ genotype + covariates + structure + ϵ

$$\begin{bmatrix} pheno_0 \\ \vdots \\ pheno_n \end{bmatrix}$$

$$\begin{bmatrix} A/T \\ \vdots \\ T/T \end{bmatrix}$$

$$\begin{bmatrix} male \\ \vdots \\ female \end{bmatrix} \begin{bmatrix} 22 \text{ years} \\ \vdots \\ 65 \text{ years} \end{bmatrix}$$

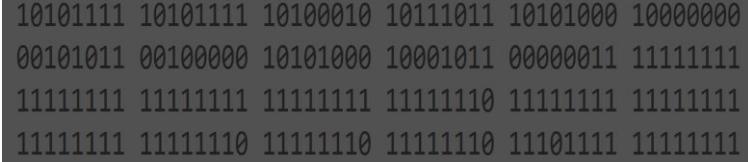
$$\begin{bmatrix} r_{00} & \dots & r_{0n} \\ \vdots & & \vdots \\ r_{n0} & \dots & r_{nn} \end{bmatrix}$$



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

*.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



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)

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

Marees et al., 2017

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.
- 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)

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

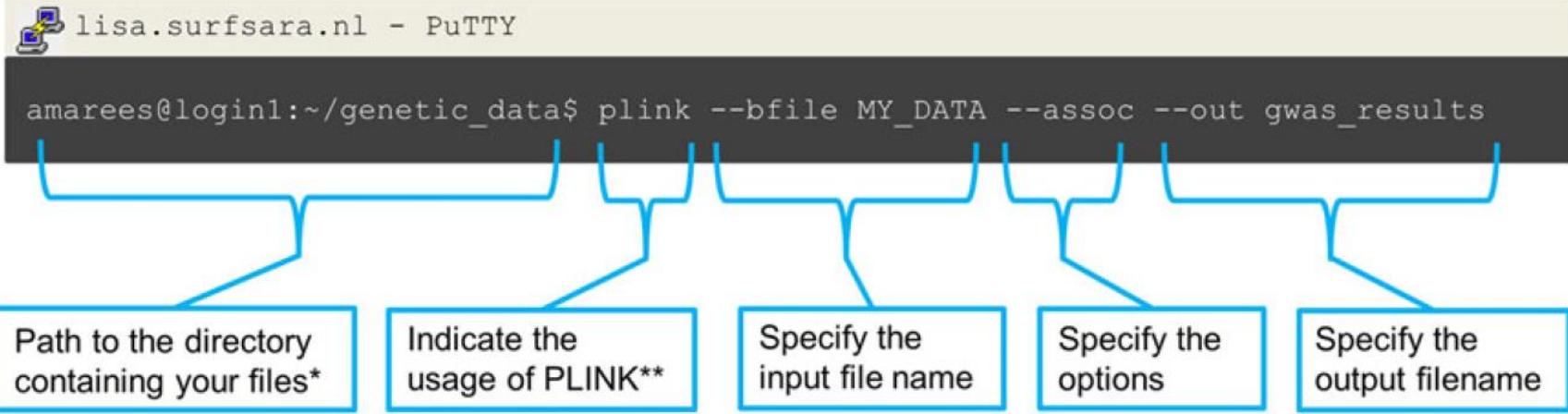
Covariate file

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

Marees et al., 2017

Genotyping data : PLINK common operations



```
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

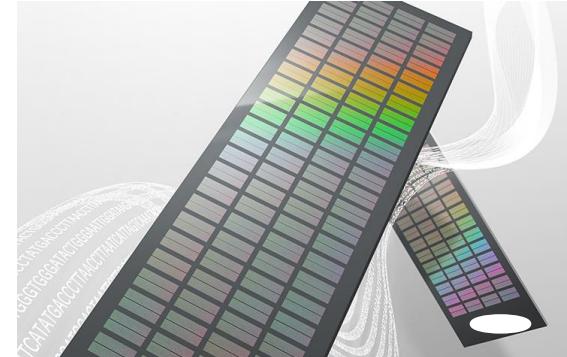
Specify the options

Specify the output filename

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

Marees et al., 2017

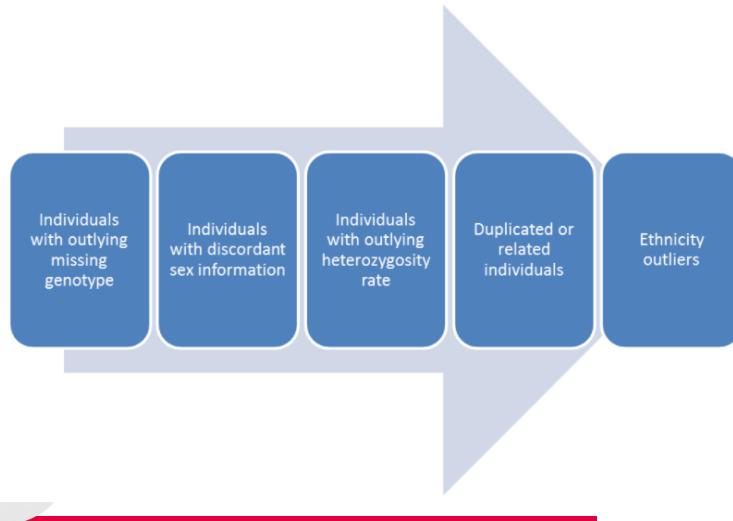
Why Quality Control?



Why Quality Control?

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

“Sample QC”



“Variant QC”

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

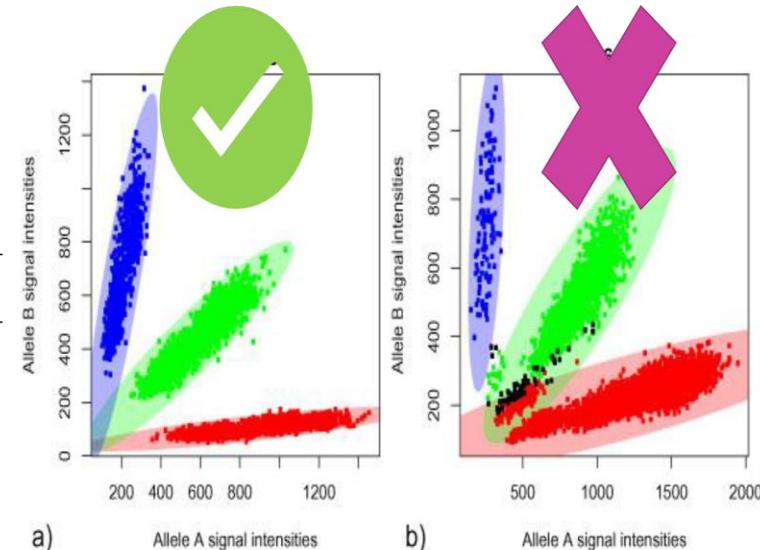
(A generalized) Quality Control process

• 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

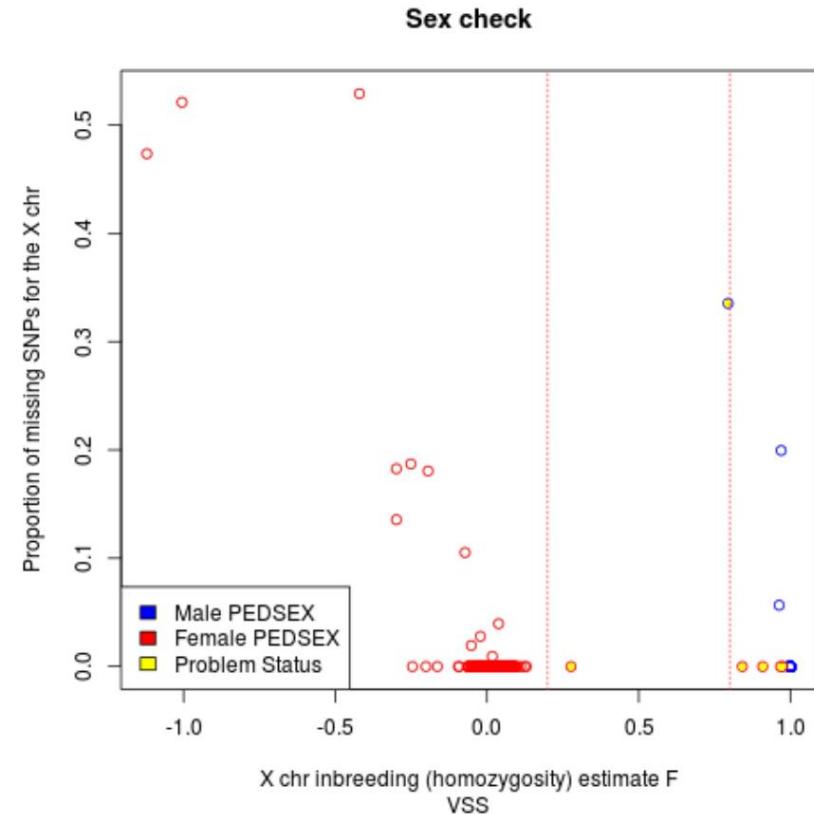
Alleles	Female genotypes possible	Male genotypes possible
A,C	A/A, A/C, C,C	A/A or C/C

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

- In Plink

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

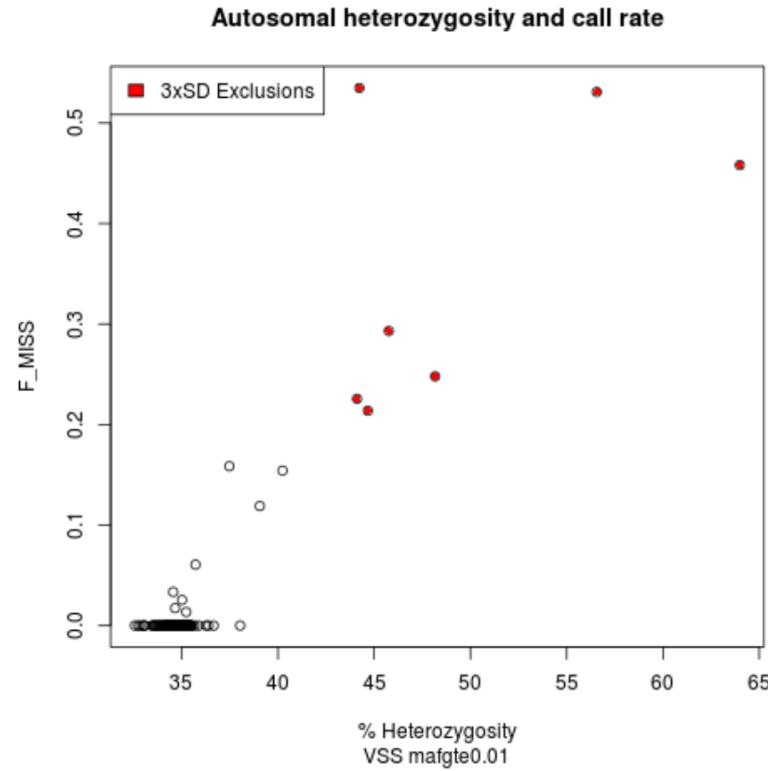
- Male (1) : $XHE > 0.80$
- Female (2) : $XHE < 0.20$
- No sex (0) : $0.20 < XHE < 0.80$



(A generalized) Quality Control process

• Heterozygosity rate

- The proportion of heterozygous genotypes (per sample)
- Various ways of calculating the rate
- PLINK: ($\langle \text{observed hom. count} \rangle - \langle \text{expected count} \rangle$) / ($\langle \text{total observations} \rangle - \langle \text{expected count} \rangle$)
- --het (gives back and F estimate)
- <custom scripts>
- Excess heterozygosity → Possible sample contamination
- Less than expected heterozygosity → Possibly inbreeding



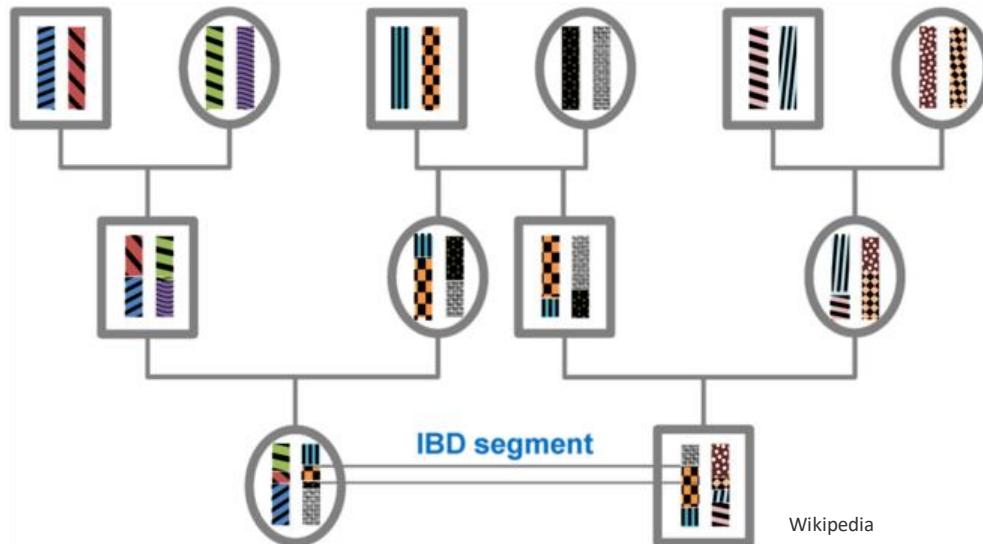


(A generalized) Quality Control process

- 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

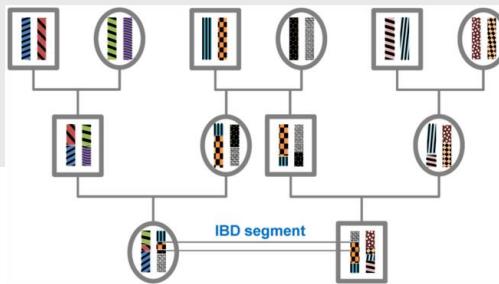
(A generalized) Quality Control process

- **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
 - Zero alleles IBD (z_0)
 - One allele IBD (z_1)
 - Two alleles IBD (z_2)
- PI_HAT (the proportion IBD, defined as $P(\text{IBD} = 2) + 0.5P(\text{IBD} = 1)$)

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



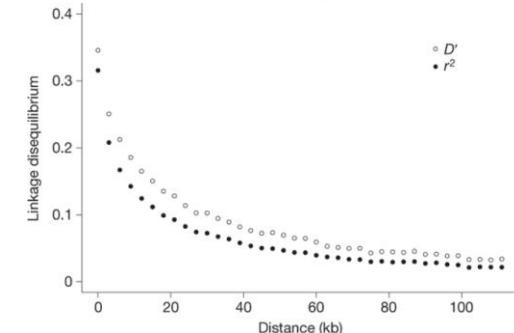
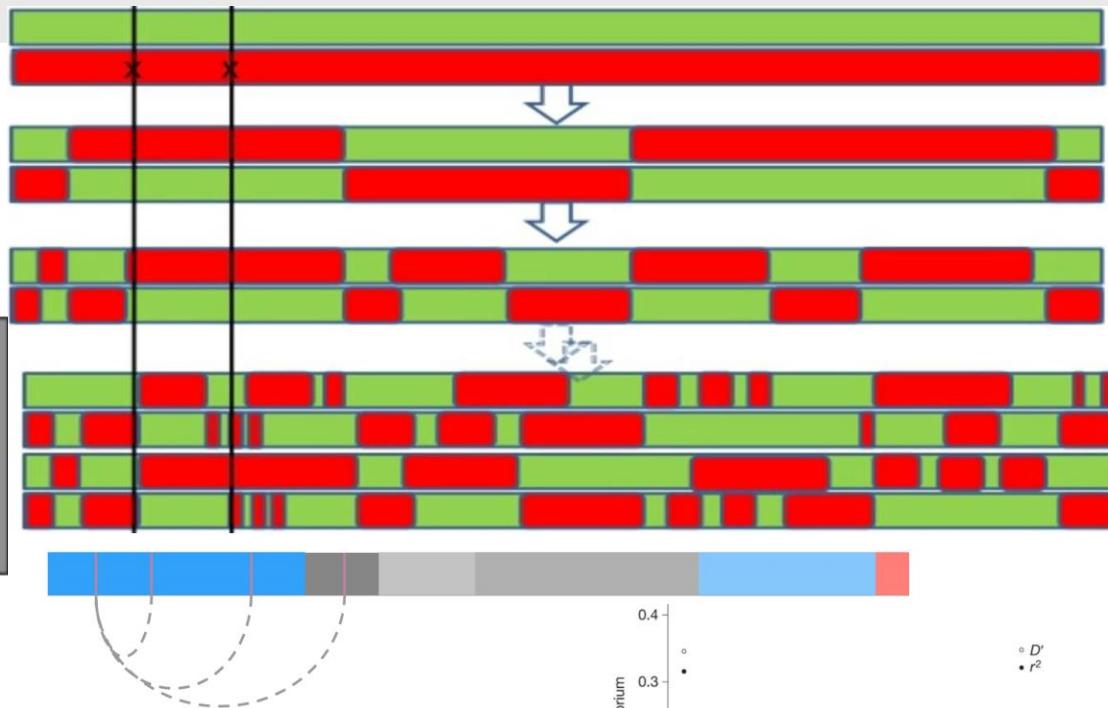
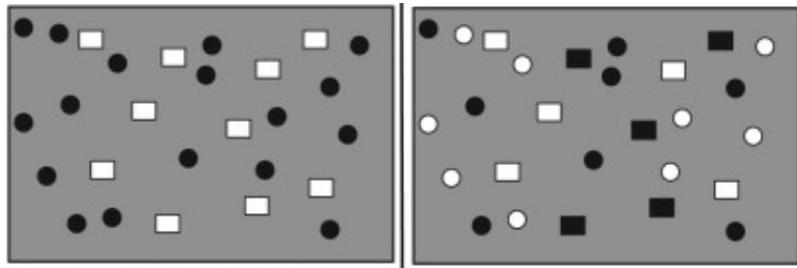
Relationship type	z_0	z_1	z_2	
Unrelated	1	0	0	0
Monozygotic (MZ) twin	0	0	1	1
Full siblings	0.25	0.5	0.25	0.5
Half siblings	0.5	0.5	0	0.25
Parent-offspring	0	1	0	0.5

Slifer, 2018

(A generalized) Quality Control process

- **Linkage disequilibrium (LD)**

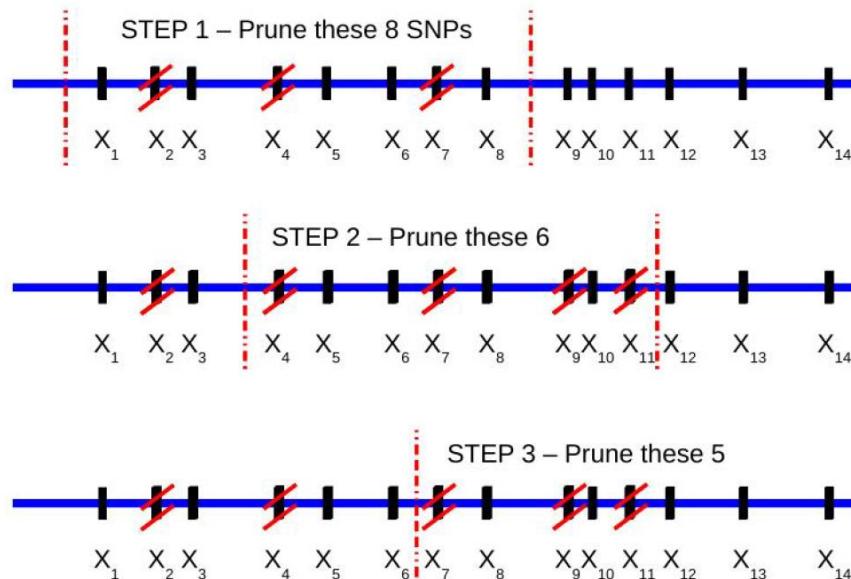
- Is the non-random association of alleles at different loci in a given population.



(A generalized) Quality Control process

- PLINK: LD-based SNP pruning

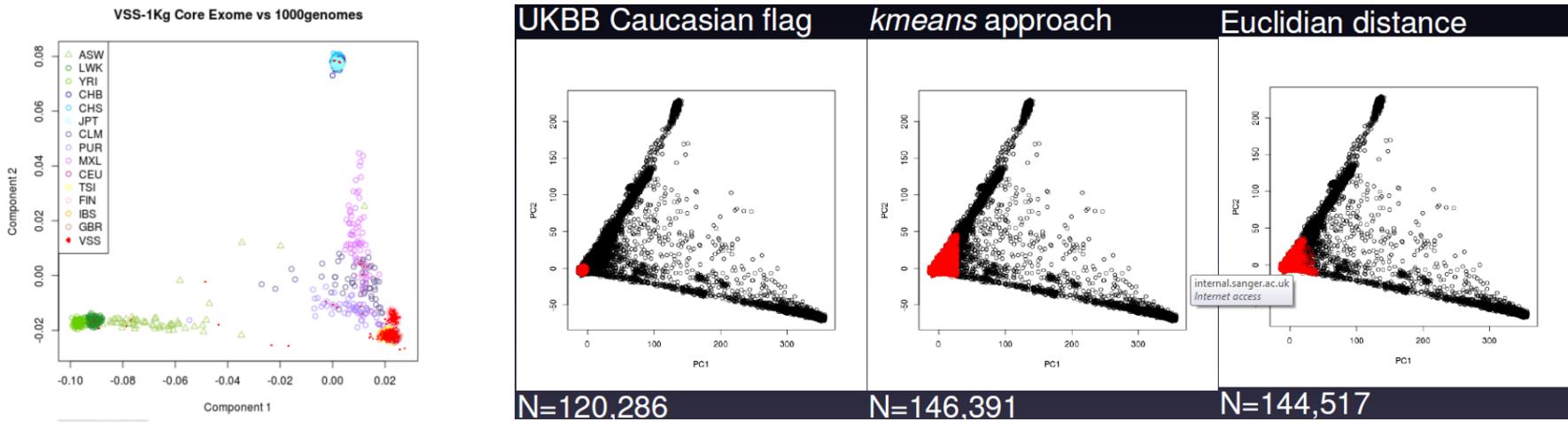
```
plink --indep-pairwise <window> <step> <rsq> --bfile <data> --out <output>  
plink --indep-pairwise 8 3 <rsq> --bfile <data> --out <output>
```



(A generalized) Quality Control process

• 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).

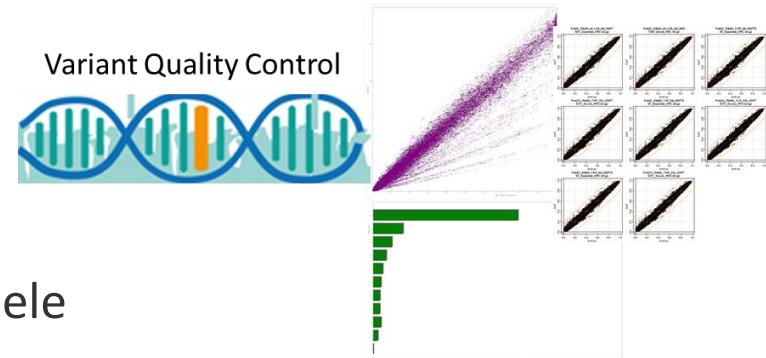


(A generalized) Quality Control process

- **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



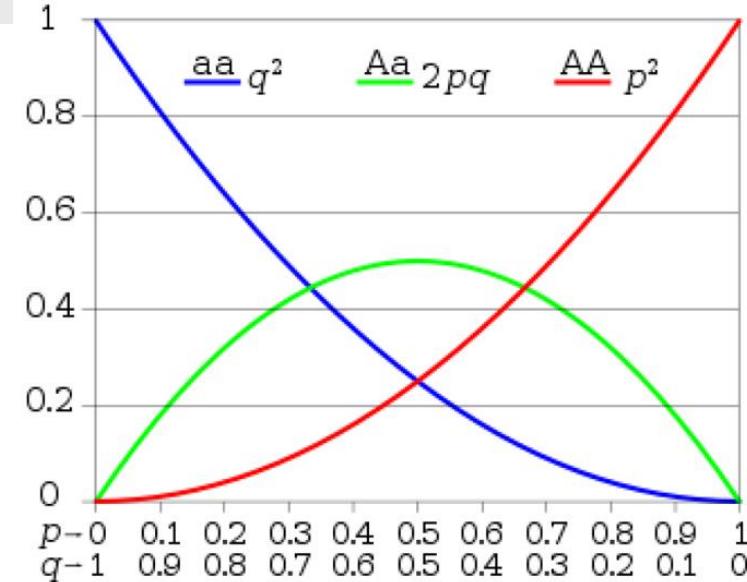
(A generalized) Quality Control process

- 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
----------------------	------------------------	--------------------------

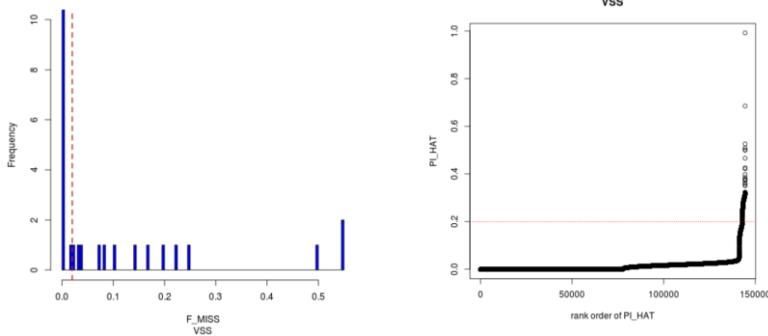
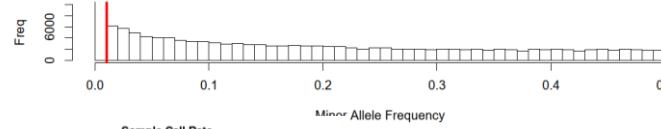
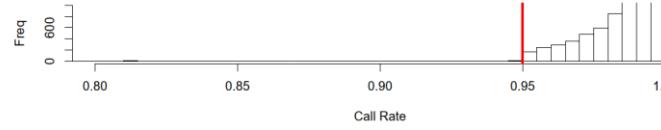
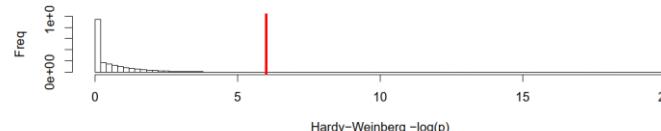
Hardy-Weinberg equilibrium check	--hardy	--hwe
----------------------------------	---------	-------

- Less strict case threshold avoids discarding disease-associated



(A generalized) Quality Control process

- Where to draw the line?



HelmholtzZentrum münchen
German Research Center for Environmental Health

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

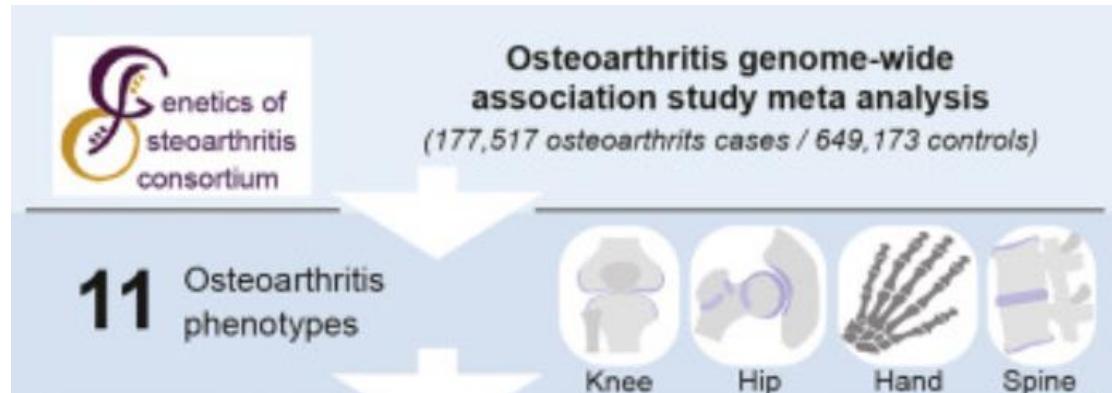
Sample QC

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



Qualitative Phenotype Quality Control

- Carefully select the phenotype definition
- Inspect the data

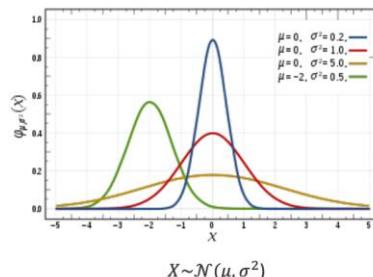
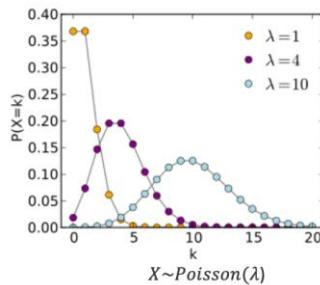


Quantitative Phenotype Quality Control

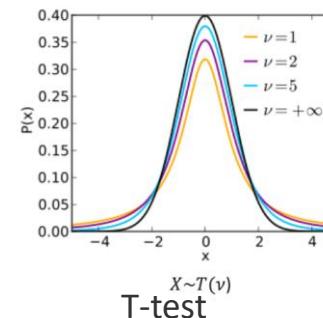
- Inspect the data and look at the distribution
- Convert to normal distribution
- Write out the transformed phenotype
- Tool: R

There are 2 broad types of distribution:

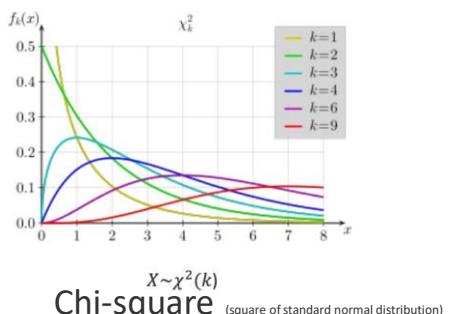
Those followed by random variables (real world) + those followed by test statistics



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



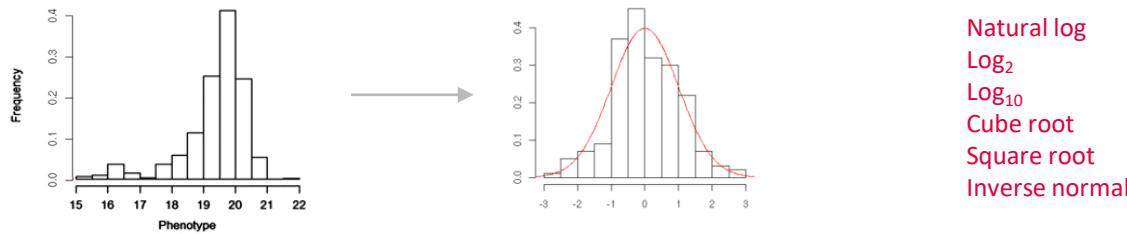
T-test



Chi-square
(square of standard normal distribution)

Quantitative Phenotype Quality Control

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



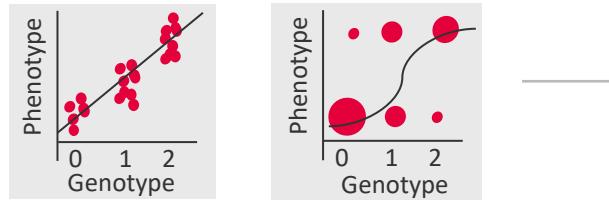
Association Analysis

- 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)



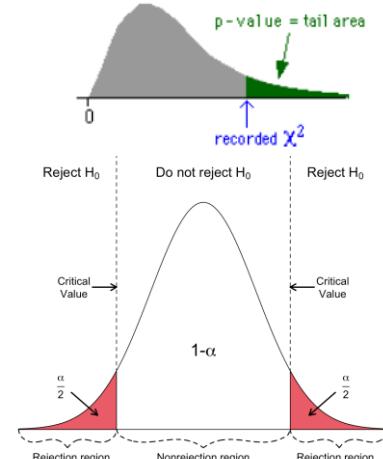
P-value
Beta / Odds ratio
SE / Confidence interval

Null/Alternative hypothesis (H_0 / 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_0

- 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: $0.05/1 \text{ million}$ independent common variants genome-wide
- **P-value $\leq 5 \times 10^{-8}$ for GWAS**



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

Association Analysis

Plink

- Genotypes can be included using **--bfile** and phenotypes using **--pheno**
- Binary trait: **--assoc** or **--logistic**
- Quantitative trait: **--assoc** or **--linear**

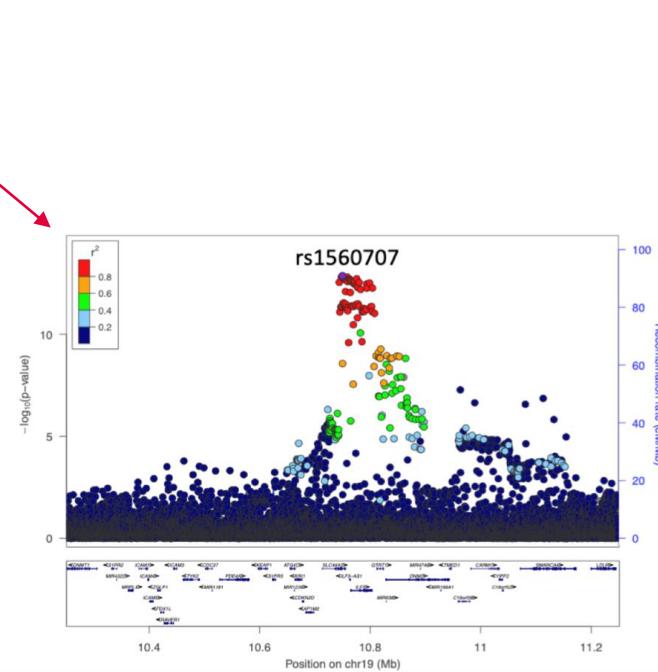
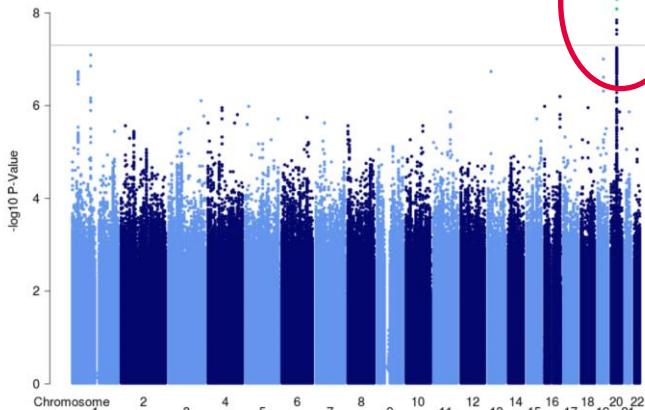
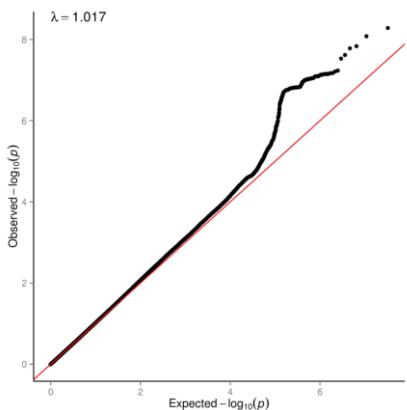
BOLT-LMM

- Imputed SNPs in dosage, 2-dosage, IMPUTE2 and BGEN format and phenotypes using **--phenoUseFam** or **--phenoFile**
- Mixed model association options
- For highly unbalanced case-control, BOLT-LMM analysis is inappropriate

SAIGE & Regenie

- PLINK (bed, bim, fam), BGEN, VCF, BCF, SAV and PLINK2 pgen/pvar/psam genetic data formats
- handles case-control imbalance of binary traits through Firth or Saddle point approximation (SPA) correction

Visualising and annotation of results

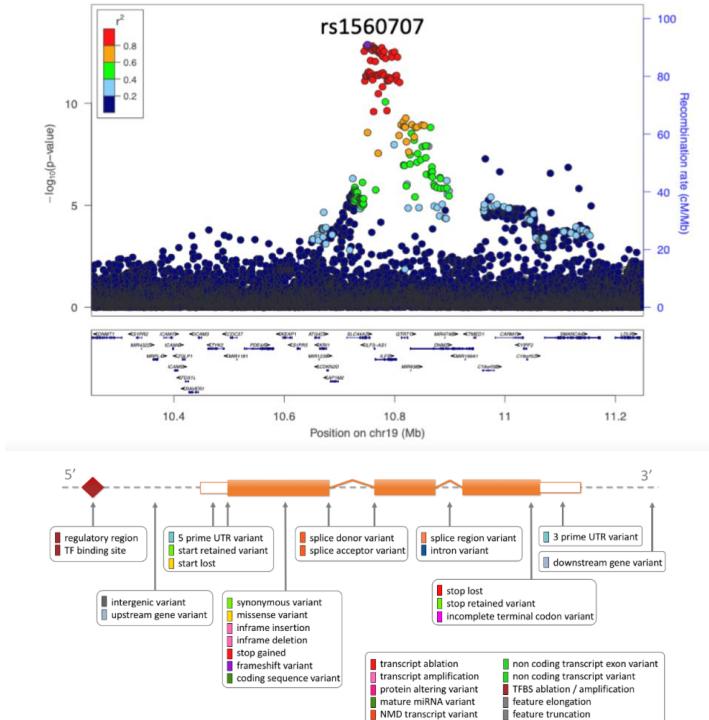


Tool: R and the library **qqman**

$$\log_{10}(5 \times 10^{-8}) = 7.3$$

Brief introduction to GWAS

- Post GWAS analysis



rs699 SNP

Most severe consequence

missense variant | See all predicted consequences

Alleles

AG | Ancestral: G | MAF: 0.29 (A) | Highest population MAF: 0.50

Change tolerance

CADD: 0.347 | GERP: -3.07

Location

Chromosome 1:230710048 (forward strand) | VCF: 1 230710048 rs699 A G HGMD-PUBLIC CM92010 ; COSMIC COSV64184214 ; dbSNP rs1553314015 (A→G)

Co-located variants

SNVs, indels, structural variants, CNVs, PAVs, SVs, AD

Evidence status

This variant has 26 HGVS names - Show

Clinical significance

This variant has 10 synonyms - Show

HGVS names

This variant has assays on 11 chips - Show

Synonyms

Variants (including SNPs and indels) imported from dbSNP (release 154) | View in dbSNP

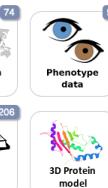
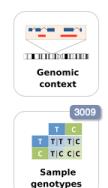
Genotyping chips

This variant overlaps 12 transcripts, has 3009 sample_genotypes, is associated with 9 phenotypes and is mentioned in 206 citations.

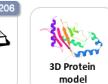
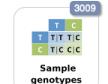
Original source

About this variant

Explore this variant



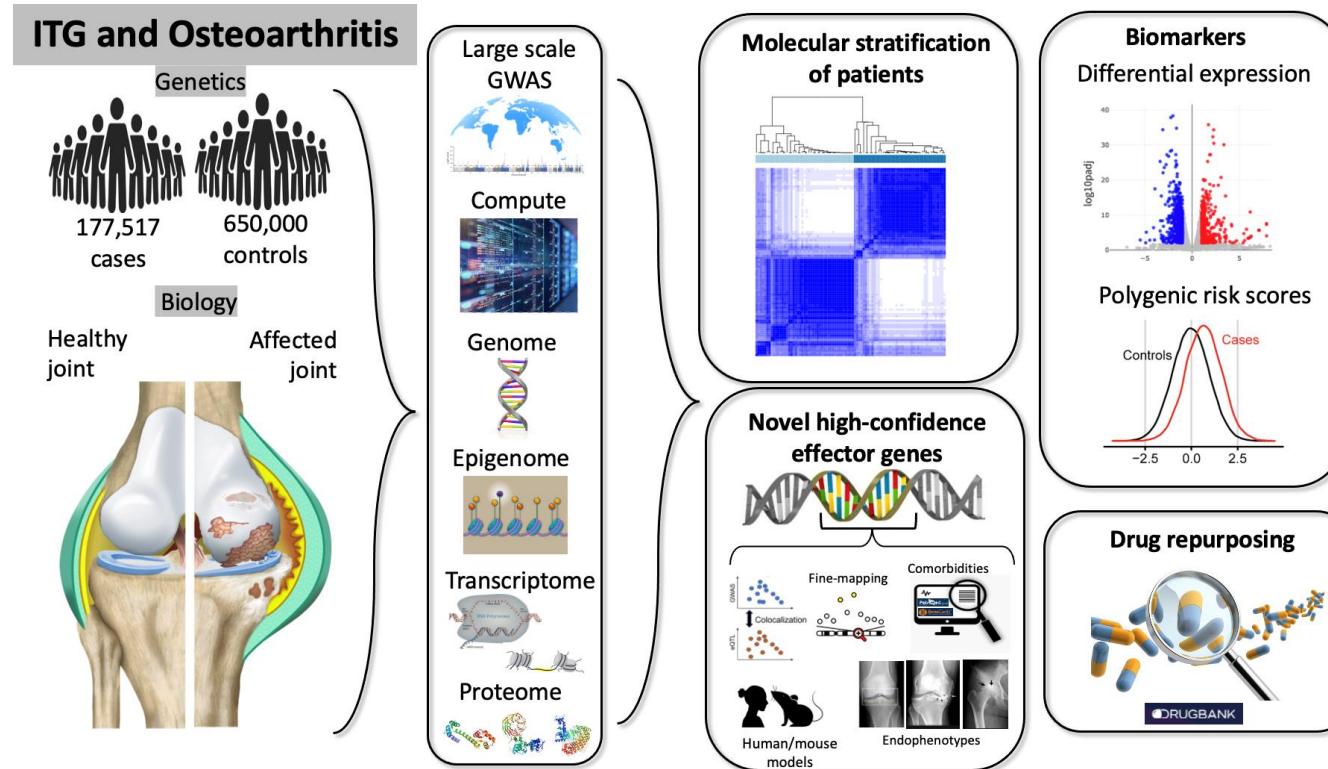
Phenotype data



3D Protein model

Brief introduction to GWAS

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



Thank-you!

Any questions?