生物資訊與分析

P2P: pre-GWAS to post-GWAS

陳佳煒 (Jia-Wei Chen)

2025/04/01


ru
e
i
b
mr
fu
g/k
s
m
f
c
r
fi
t
p
h
a
prs
c
m
h
c

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

A tool for genetic data analysis, focusing on association testing and quality control.

https://www.r-project.org/

A statistical programming language widely used for data analysis, visualization, and bioinformatics.

https://www.python.org/

A versatile coding language for data science, automation, and machine learning.

https://github.com/saigegit/SAIGE

A mixed-model method for GWAS, designed for binary traits while accounting for relatedness.

https://rgcgithub.github.io/regenie/

A two-step GWAS tool optimized for large datasets, reducing computational burden with ridge regression.

https://annovar.openbioinformatics.org/en/latest/

A bioinformatics tool for variant annotation, functional interpretation, and filtering, aiding genomic studies by linking mutations to biological insights.

https://cncr.nl/research/magma/

A tool for gene and gene-set/pathway analysis

https://fuma.ctglab.nl/

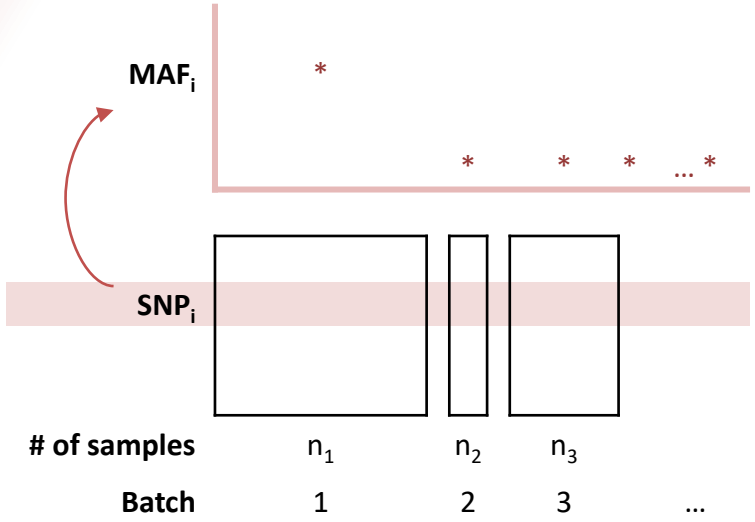
A tool for SNP-to-gene mapping, functional annotation, and pathway analysis.

Batch SNP check

Why?

- A batch effect can be arose from differences in platforms, protocols, sample compositions, DNA qualities, etc.

- Potentially leading to spurious associations or confounding in GWAS



How?

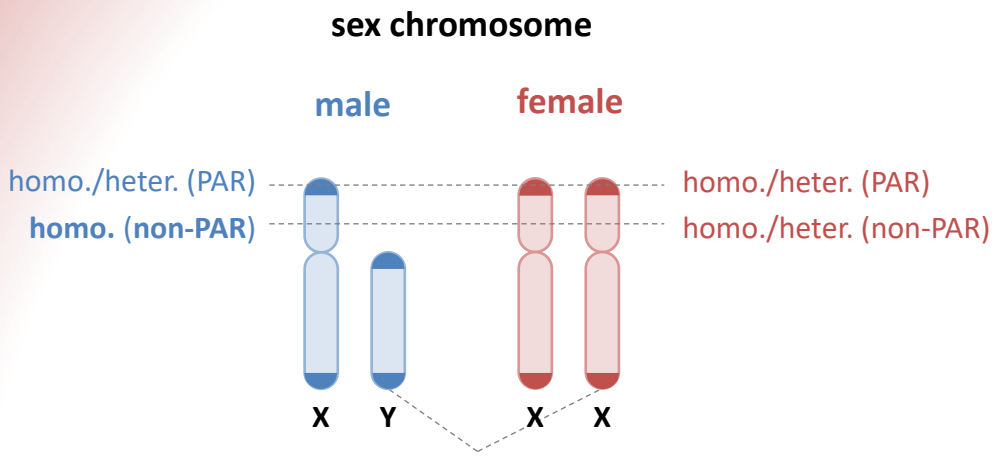
- Compare AFs of SNPs to ones in public database (e.g., 1000 genomes project) to identify the SNPs with weird Afs

- Instead of excluding the SNPs, marking the genotype calls of samples from problematic batches as missing for such SNPs

Sex check

Why?

- May have chromosome anomaly or structural variation
- May be a covariate in the subsequent analysis



PAR (Pseudoautosomal region)
homologous sequences of nucleotides
on the X and Y chromosomes

How?

- Using either **homozygosity rate** or **inbreeding coefficient** of X chromosome to check for the gender

	Male	Female
Homozygosity rate	≥ 0.9	< 0.9
Inbreeding coefficient (F)	> 0.8	< 0.2

- If EHR data is accessed, directly comparing EHR gender to genetic gender

Genotype call-rate check

plink --missing

Why?

- Low DNA quality or concentration often have below-average call rates & genotype accuracy

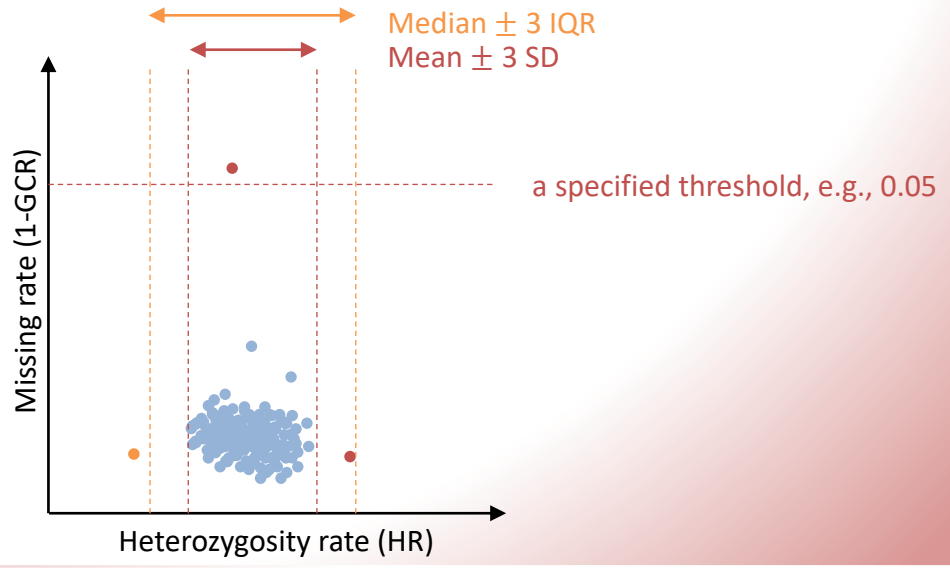
How?

- Calculate genotype call-rate (GCR) for **autosomes**

N = # of markers

N_i = # of nonmissing genotypes for individual i

$GCR_i = \frac{N_i}{N}$



Heterozygosity rate check

plink --het

Why?

- An excessive of heterozygote genotypes, which may be indicative of DNA sample contamination

- A reduced proportion of heterozygote genotypes, which may be indicative of inbreeding

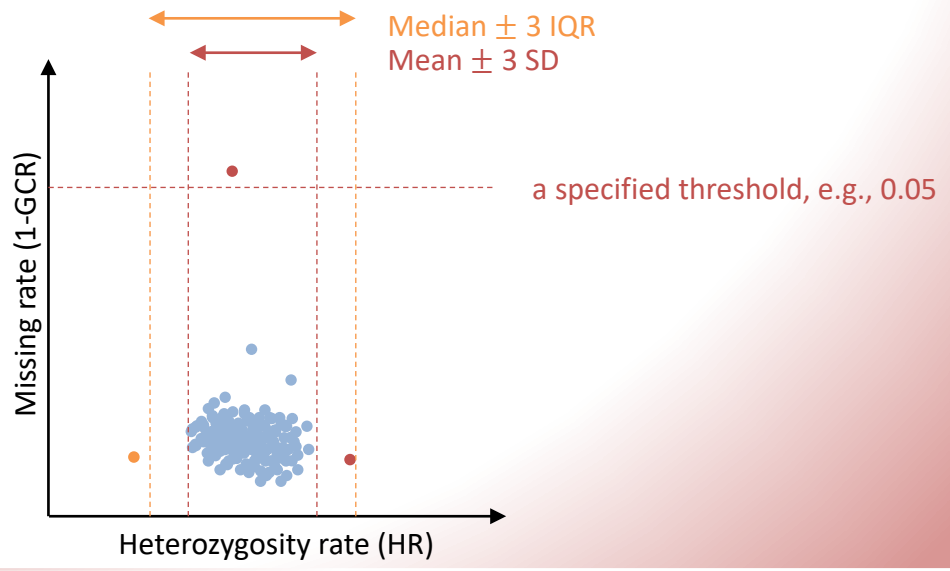
How?

- Calculate heterozygosity rate (HR) for **autosomes**

N_i = # of nonmissing genotypes for individual i

O_i = # of homozygous genotypes for individual i

$HR_i = \frac{N_i - O_i}{N_i}$



Cryptic relatedness check

plink --indep-pairwise 50 5 0.2
plink –genome / plink --king-cutoff 0.0442

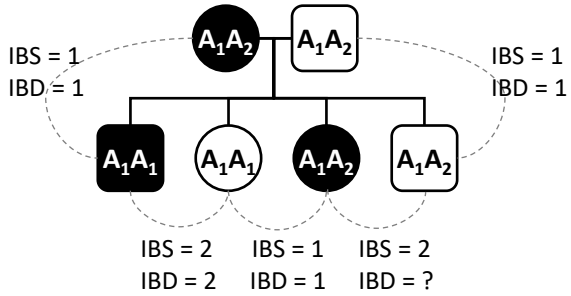
Why?

- An association study will be interfered by the relatedness, e.g., violation of assumption for linear/logistic models, bias AF estimation, etc.

- A family-based study should take the relatedness into account

How?

- Usually, using **independent** SNPs (pair correlation $r^2 < 0.2$) to calculate the relatedness of individuals by either **proportion IBD** or **kinship coefficient** to check for the relatedness



	Proportion IBD	Kinship coeff.
duplicate/MZ twin	1	> 0.354
1st degree	0.5	[0.177, 0.354]
2nd degree	0.25	[0.0884, 0.177]
halfway of 2nd & 3rd degrees	> 0.1875	
3rd degree	0.125	[0.0442, 0.0884]

Divergent ancestry check

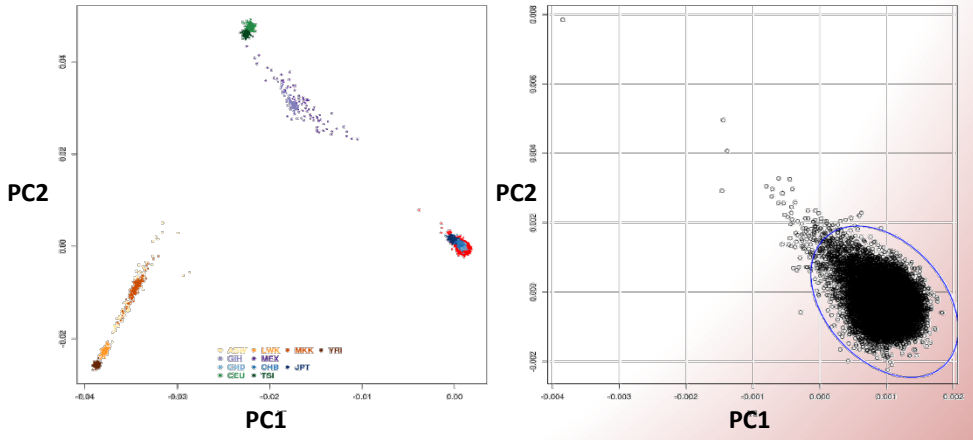
plink2 --pca approx

Why?

- Confounding to case/control groups, i.e. identified markers may not associated with disease but differently distributed in ancestries (Hamer, D. , & Sirota, L. (2000)

How?

- Merge study genotypes to HapMap3 or 1000 genomes data
- Exclude ambiguous SNPs (A/T or C/G polymorphic)
- Prune highly correlated SNPs
- PCA
- Exclude individuals out of 99.9% confidence band of data



Divergent ancestry check

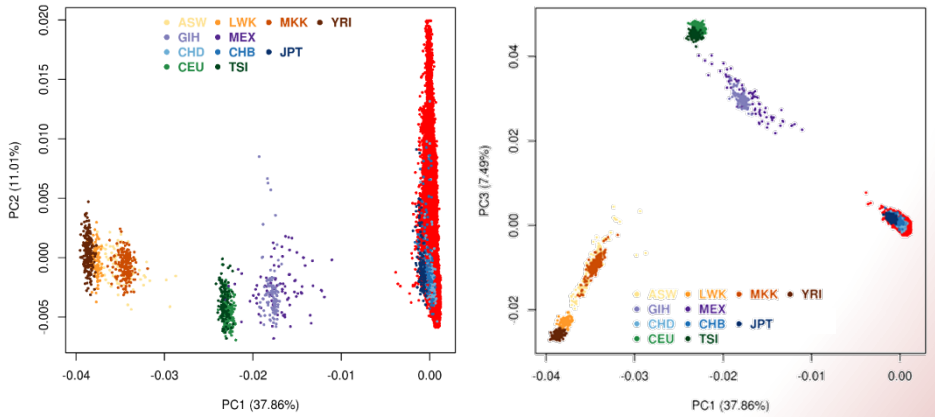
plink2 --pca approx

Why?

- Confounding to case/control groups, i.e. identified markers may not associated with disease but differently distributed in ancestries (Hamer, D. , & Sirota, L. (2000)

Note!

- When sample size of target data is much larger than that of ancestry data, the first PCs may be dominant by the variations of target data



Genotype call-rate check

plink --geno 0.05
plink --test-missing

Why?

- Low call-rates may indicate technical issues or poor-quality data

How?

- Exclude markers with lower call rate, say 95% (sometimes, 99%)

- For case/control study, we can further

1. exclude markers with a large difference of call rates between cases & controls

2. exclude markers with a call rate less than 95% in either cases or controls

Minor allele frequency check

plink --maf 0.01

Why?

- Rare variants (with very low MAF) may have limited statistical power to detect associations and could also prone to genotype errors

How?

- Exclude markers with lower MAF, say 0.01 (sometimes, 0.05)

Individual-based AF

Genotype call	AF_A	AF_B	MAF
AA	1	0	0
AB	0.5	0.5	0.5
BB	0	1	0

Population-based AF

		Allele 2	
		A	**B**
Allele 1	**A**	n_{AA}	n_{AB}
	B	n_{BA}	n_{BB}

N

$$\left.\begin{aligned} AF_A &= \frac{n_{AA}\times 2 + n_{AB}\times 1 + n_{BA}\times 1 + n_{BB}\times 0}{2N} \\ AF_B &= \frac{n_{AA}\times 0 + n_{AB}\times 1 + n_{BA}\times 1 + n_{BB}\times 2}{2N} \end{aligned}\right\} MAF = \min(\, AF_A, AF_B \,),$$

We may also interested in what the **minor/major allele** is

Hardy-Weinberg eq. check

plink --hwe 1e-8

Why?

- Deviation from expected genotype frequencies (based on observed allele frequencies) may indicate genotyping errors, population stratification, etc.

How?

- Exclude markers with a p-value of Hardy-Weinberg equilibrium (HWE) test less than a threshold (e.g., Bonferroni's level)

1. In case/control study, the HWE test is applied for control individuals

2. In quantitative-trait study, the HWE test is applied for all individuals

		Allele 2		
		A	**B**	
Allele 1	**A**	$p_{AA} = p_A^2$	$p_{AB} = p_A p_B$	p_A
	B	$p_{BA} = p_B p_A$	$p_{BB} = p_B^2$	p_B
		p_A	p_B	1

Hardy-Weinberg principle

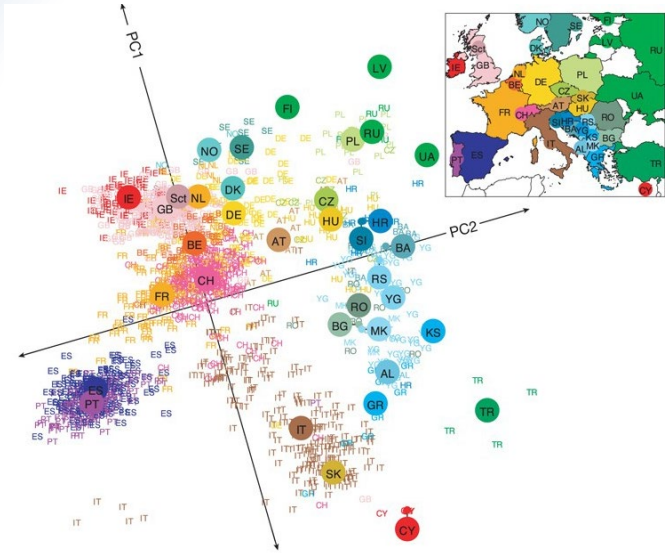
The genetic variation (allele/genotype frequencies) in a population will remain **constant** from one generation to the next in the **absence of disturbing factors** (e.g., selection, mutation and migration)

Subpopulation structure

plink --indep-pairwise / plink --clump
plink2 --pca approx biallelic-var-wts

Why?

- Confounding to case/control groups, i.e. identified markers may not associated with disease but differently distributed in ancestries

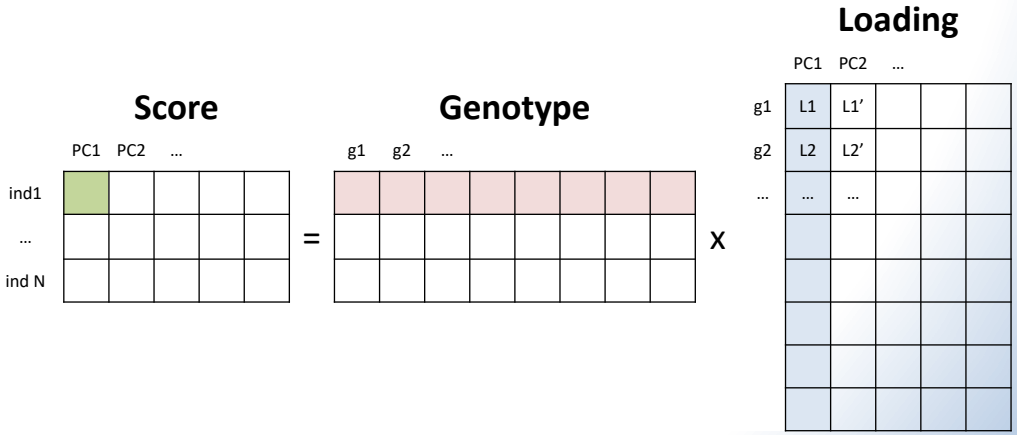


Novembre, J., Johnson, T., Bryc, K. *et al* (2008)

How?

- Identify independent SNPs through pruning, clumping, removing long-range LD (LRLD) region, or by combining methods

- PCA

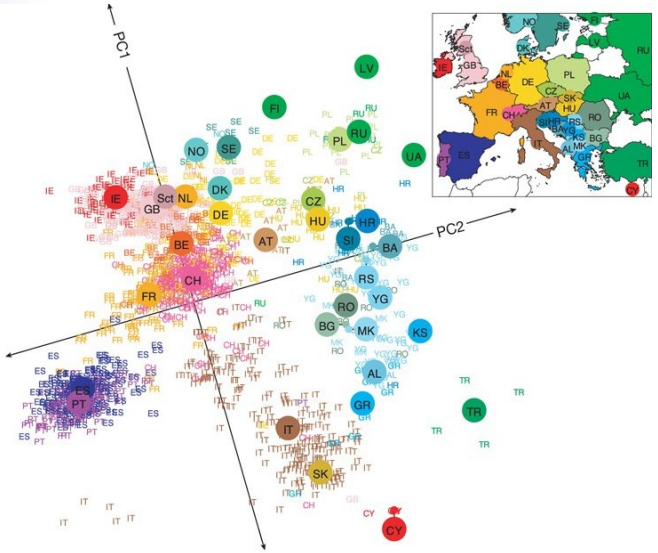


Subpopulation structure

plink --indep-pairwise / plink --clump
plink2 --pca approx biallelic-var-wts

Why?

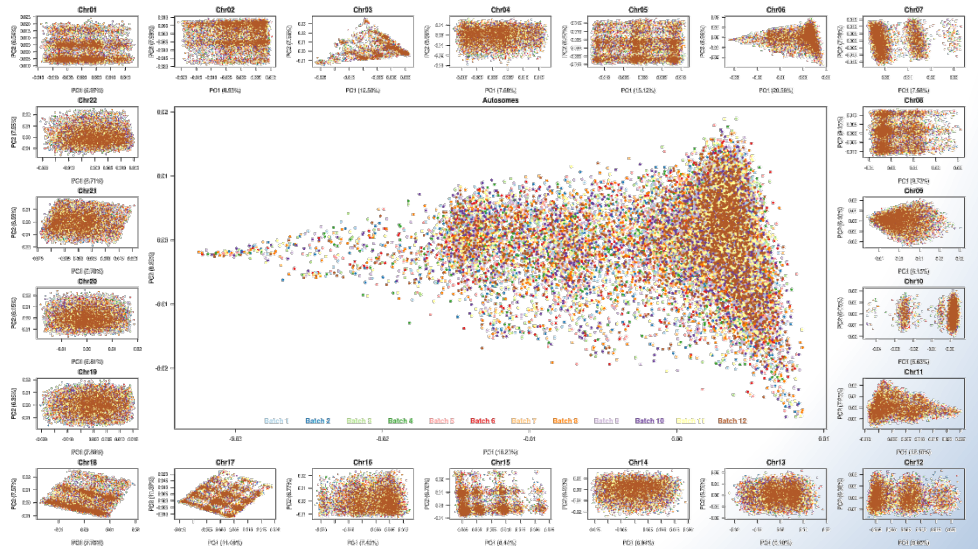
- Confounding to case/control groups, i.e. identified markers may not associated with disease but differently distributed in ancestries



Novembre, J., Johnson, T., Bryc, K. *et al* (2008)

Note!

- Without identifying independent SNPs, the first several PCs may be dominated by SNPs in MHC region (on chr. 6)



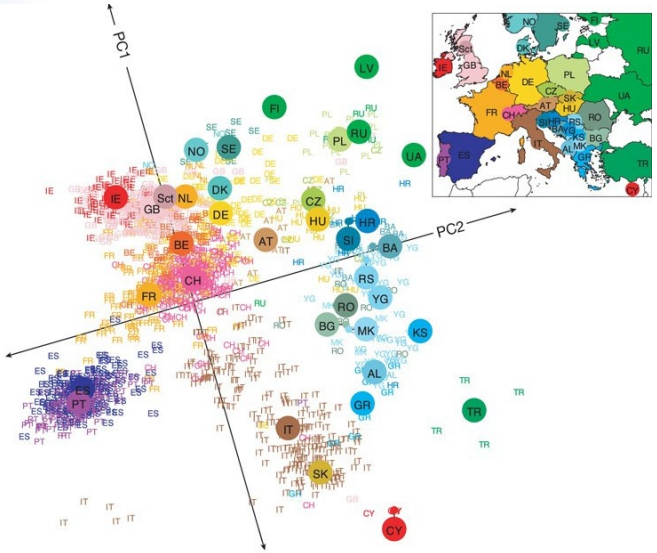
(score plot)

Subpopulation structure

plink --indep-pairwise / plink --clump
plink2 --pca approx biallelic-var-wts

Why?

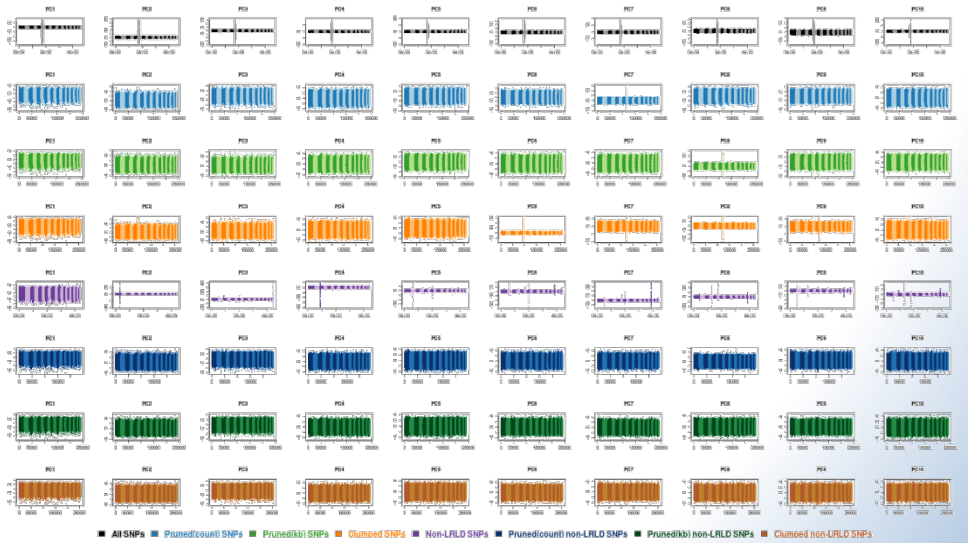
- Confounding to case/control groups, i.e. identified markers may not associated with disease but differently distributed in ancestries



Novembre, J., Johnson, T., Bryc, K. *et al* (2008)

Note!

- Without identifying independent SNPs, the first several PCs may be dominated by SNPs in MHC region (on chr. 6)



(loading plot)

Fixed-effects model

plink2 --glm no-x-sex hide-covar cols=… --pheno --covar --covar-variance-standardize

When?

- Assuming that each SNP has the same impact on a trait, regardless of an individual's background, ancestry, or other subgroup characteristics

Note!

- Binary trait

1. Strict disease group definitions are necessary, but sample size may still be the ultimate challenge in GWAS
2. Firth logistic regression can deal with the issue of rare events or complete separation in traditional logistic regression

- Quantitative trait

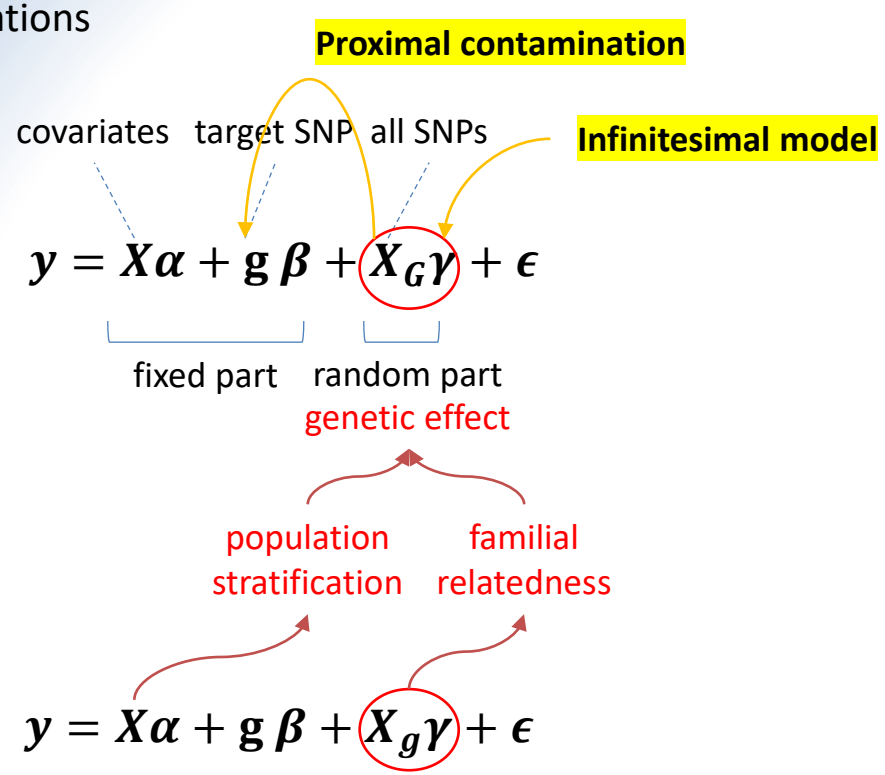
1. Inverse normal transformation (INT):

$$\mathrm{INT}(x) = \Phi^{-1}\left(\frac{rank(x)-c}{n-2c+1}\right)\begin{cases} c = \frac{3}{8}\ (Blom, 1958) \\ c = \frac{1}{3}\ (Tukey, 1962) \\ c = \frac{1}{2}\ (Bliss, 1967) \end{cases}$$

Mixed-effects model

When?

- Considering about fixed effects and random effects, and it is useful when dealing with structured or related populations



Note!

- Binary trait

1. Strict disease group definitions are necessary, but sample size may still be the ultimate challenge in GWAS
2. Firth logistic regression can deal with the issue of rare events or complete separation in traditional logistic regression

- Quantitative trait

1. Inverse normal transformation (INT):

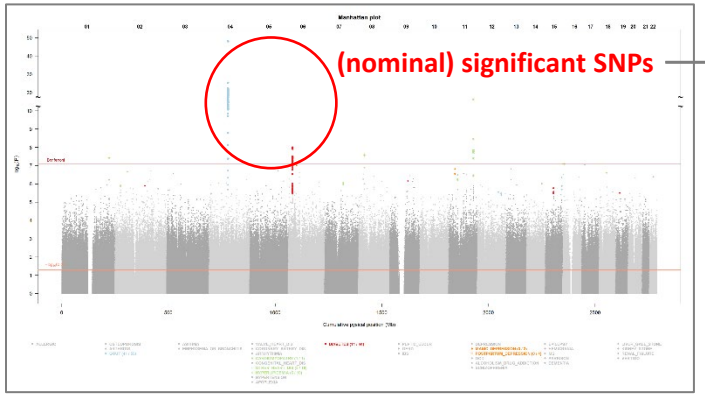
$$\mathrm{INT}(x) = \Phi^{-1}\left(\frac{rank(x)-c}{n-2c+1}\right)\begin{cases} c = \frac{3}{8}\ (Blom, 1958) \\ c = \frac{1}{3}\ (Tukey, 1962) \\ c = \frac{1}{2}\ (Bliss, 1967) \end{cases}$$

SAIGE and REGENIE include the of INT, but BOLT-LMM doesn't

After GWAS

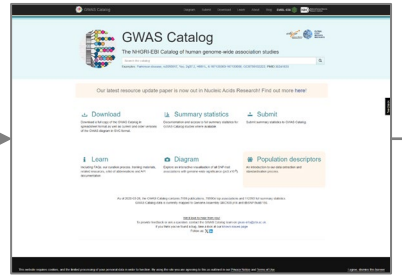
1. Expect **consecutive** SNPs with smaller p-values rather than scattered SNPs with smaller p-value across multiple chromosomes

2. **Significant** SNPs are identified using Bonferroni correction or FDR, while **nominally significant** SNPs are determined by a predefined threshold



Manhattan plot

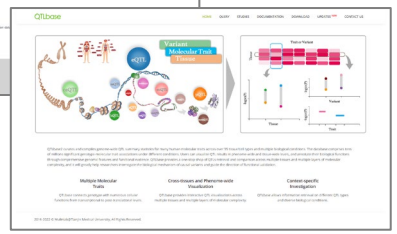
GWAS Catalog https://www.ebi.ac.uk/gwas/



Confirm by the previous findings

GTEx https://www.gtexportal.org/home/

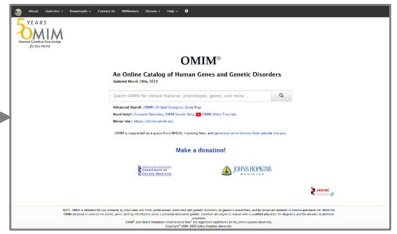
QTLbase http://www.mulinlab.org/qtlbase/index.html



Search genes of eQTL or proteins of pQTL

Explanation of relations between identified genes and phenotypes

OMIM https://omim.org/



Functional mapping

What?

- A bioinformatics approach used to link genetic variants or biological sequences to their functional roles.

Which?

- https://fuma.ctglab.nl/
- Functionally interpret GWAS results by

SNP2GENE (Maps SNPs to genes and provides functional annotation)
GENE2FUNC (Analyzes gene expression patterns and functional enrichment)
Cell Type (Identifies cell types relevant to GWAS findings)

Note!

- Summary statistic files should be **.zip**
- The **GRCh38** option is not functional
- Without providing information about chromosomes and physical positions, using **rsid** to determine genomic location
- Processing is not possible if no SNPs with p-values **< 1e-5**

Functional mapping

What?

- A bioinformatics approach used to link genetic variants or biological sequences to their functional roles.

Which?

- https://annovar.openbioinformatics.org/en/latest/

- A bioinformatics tool for variant annotation, functional interpretation, and filtering, aiding genomic studies by linking mutations to biological insights.

Note!

- Download reference database (e.g., refGene, 1000g, exac03, avsnp, dbnsfp, clinvar, icgc, cosmic, gwasCatalog, gtex, etc.) by **annotate_variation.pl**

- Annotate the reference database information to findings (significant SNPs) by **table_annovar.pl**

Gene-/pathway-based analysis

What?

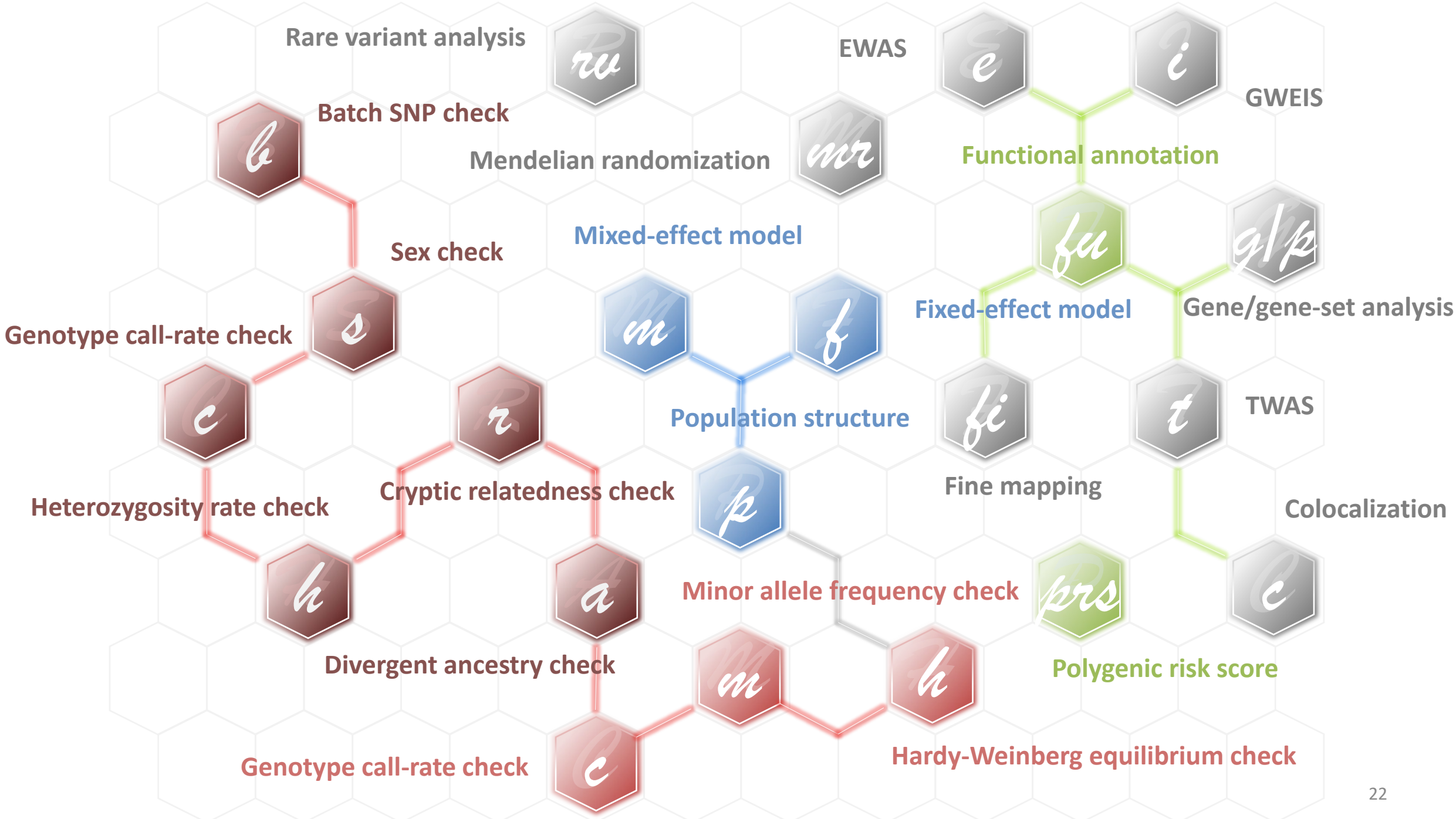
- Considering about **joint effects** of SNPs on a gene or in a gene set/pathway to improve statistical power and biological interpretation.

Which?

- https://cncr.nl/research/magma/
- A bioinformatics tool for gene and gene-set analysis by aggregating SNP associations at the gene level, accounting for linkage disequilibrium and interactions, and further extending to gene-set/pathway level.

Note!

- Download reference data of 1000 genomes project provided by MAGMA
- Download gene location files provided by MAGMA
- Download gene-set files provided by MSigDB (need to register)


Rare variant analysis
EWAS
GWEIS
Batch SNP check
Mendelian randomization
Functional annotation
Mixed-effect model
Sex check
Fixed-effect model
Gene/gene-set analysis
Genotype call-rate check
Population structure
TWAS
Fine mapping
Heterozygosity rate check
Cryptic relatedness check
Colocalization
Minor allele frequency check
Divergent ancestry check
Polygenic risk score
Genotype call-rate check
Hardy-Weinberg equilibrium check

Thanks for your attention!!
<(_ _)>