

GWAS: Data and Preprocessing



Samuele Soraggi, PhD

Sandbox data scientist

Bioinformatics Research Center (Aarhus)

UNIVERSITY OF
COPENHAGEN



Content

- Why QC
- PLINK Software
- Data format
- QC steps and Concepts
- Exercise Presentation

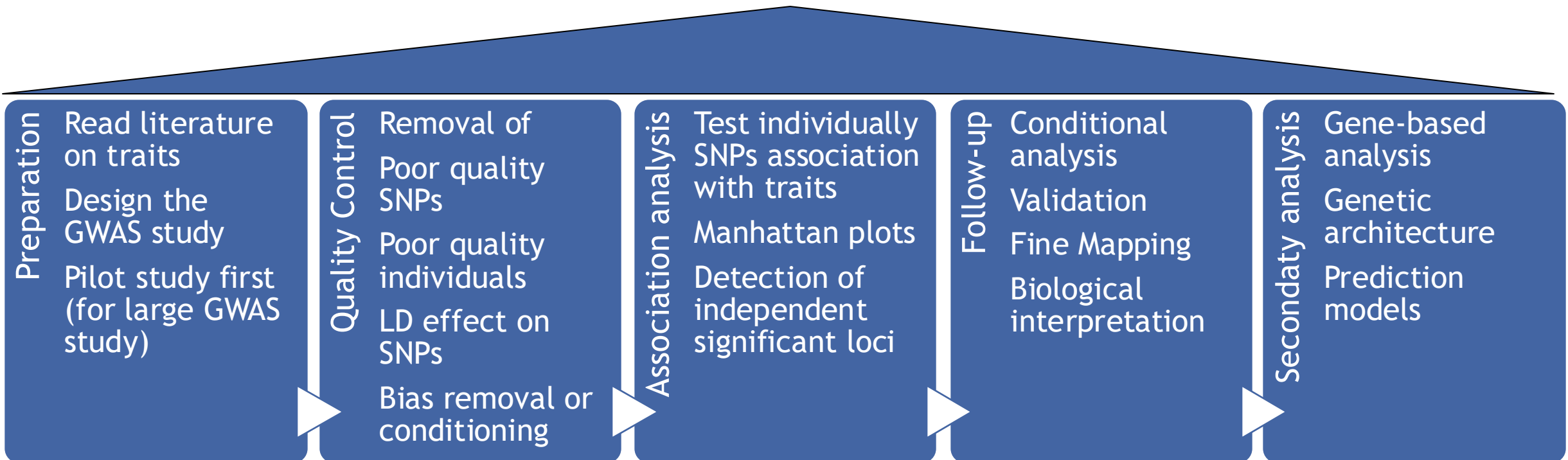
A GWAS workflow

Objectives

Detection of causal variants (risk loci, fine mapping)

Predict traits (personalized medicine)

Understand genetic architecture (SNP heritability, Nr casual loci)



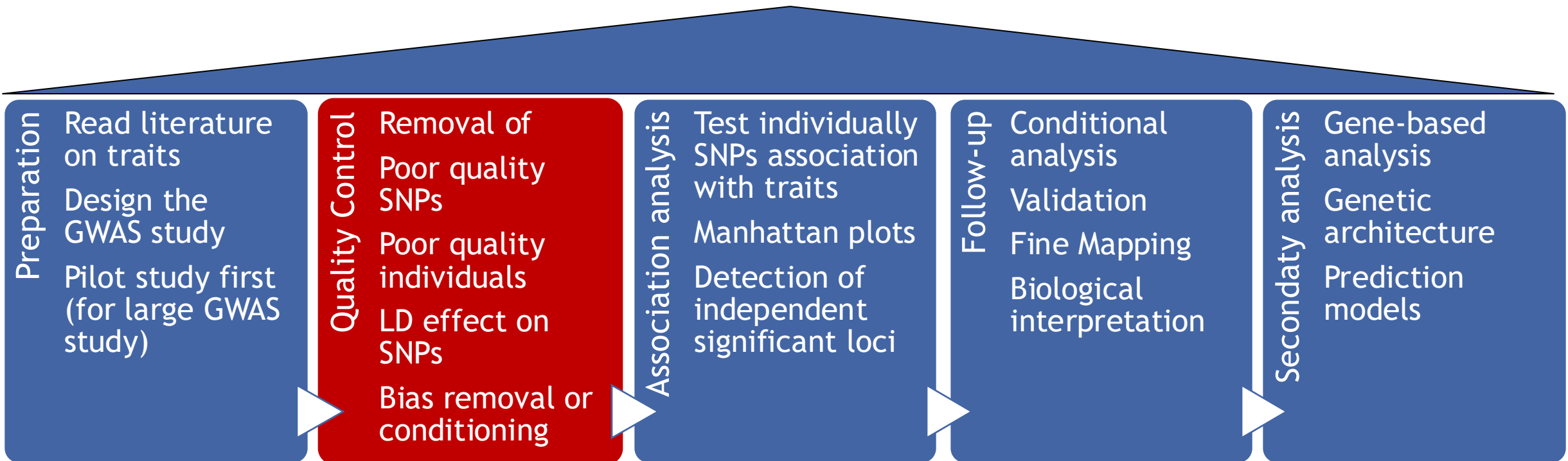
A GWAS workflow

Objectives

Detection of causal variants (risk loci, fine mapping)

Predict traits (personalized medicine)

Understand genetic architecture (SNP heritability, Nr casual loci)



Bad QC

Retraction: Genome-Wide Association Study Identifies Common Genetic Variants Associated With Salivary Gland Carcinoma and its Subtypes by Xu L, Tang H, Chen DW, El-Naggar AK, Wei P, Sturgis EM

► [Article notes](#) ► [Copyright and License information](#)

PMCID: PMC5991497 NIHMSID: NIHMS971275 PMID: [29406587](#)

False positive findings during genome-wide association studies with imputation: influence of allele frequency and imputation accuracy

Zhihui Zhang ^{1 2}, Xiangjun Xiao ¹, Wen Zhou ¹, Dakai Zhu ¹, Christopher I Amos ^{1 2}

Affiliations + expand

PMID: 34368847 PMCID: [PMC8682785](#) DOI: [10.1093/hmg/ddab203](#)

 [Sign in](#)

News | Published: 21 July 2011

Paper on genetics of longevity retracted

[Heidi Ledford](#)

[Nature](#) (2011) | [Cite this article](#)

479 Accesses | 1 Citations | 112 Altmetric | [Metrics](#)

Technical problems mar study of centenarians.

A prominent paper that claimed to reveal the genetic factors that help people live to 100 or older has been retracted¹, a year after it was first released.



Software: PLINK and PLINK2

www.cog-genomics.org/plink/

PLINK 1.9 home	plink2-users	GitHub	File formats	PLINK 1.9 index	PLINK 2.0
-----------------------	---------------------	---------------	---------------------	------------------------	------------------

Introduction, downloads
S: 22 Oct 2024 (b.7.7)
D: 22 Oct 2024
Recent version history
What's new?
Future development
Limitations
Note to testers
[Jump to search box]
General usage
Getting started
Citation instructions

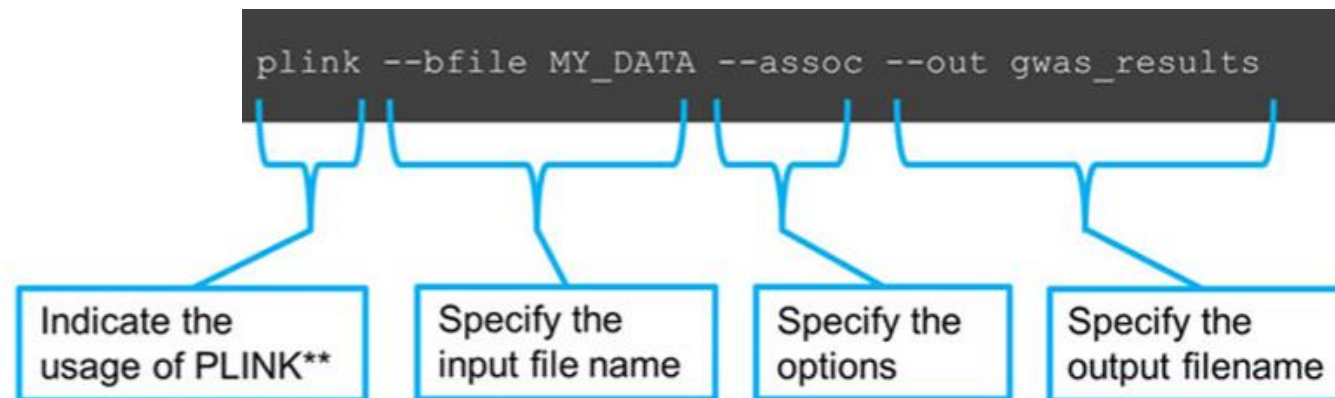
PLINK 1.9 beta

PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

PLINK 1.9 is a comprehensive update to the [original version developed by Shaun Purcell](#). It was developed by [Christopher Chang](#) with support from the [NIH-NIDDK's](#) Laboratory of Biological Modeling, the [Purcell Lab](#), and others. ([What's new?](#)) ([Credits.](#)) ([Methods paper.](#)) (Usage questions should be sent to the [plink2-users Google group](#), not Christopher's email.)

Software: PLINK

- The go-to software for all the **standard GWAS analysis** tasks
 - Made for position-based SNP data
 - Has its own **fileset** which is almost the GWAS standard
 - **Optimized** for thousands of samples
 - Performs **preprocessing** and whole genome **association** analysis
 - **On command line!** Has a lot of options
 - You can work directly in R with **snpStats**, but it is not optimized for large datasets
- Other softwares **integrate with pLinks's** inputs and outputs
- pLink outputs can be easily **visualized in R or python**



Data: Storing SNP information

We can represent SNP data as a large matrix

$$X = \begin{bmatrix} AA & CG & TT & \dots & GG \\ AG & CG & AT & \dots & CG \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ GG & CG & NA & \dots & CC \end{bmatrix} \begin{array}{l} \leftarrow \text{Individual 1} \\ \leftarrow \text{Individual 2} \\ \vdots \\ \leftarrow \text{Individual n} \end{array}$$

SNP 1 SNP 2 SNP 3 SNP m

Though usually we prefer to use allele counts

$$X = \begin{bmatrix} 0 & 1 & 2 & \dots & 2 \\ 1 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 2 & 1 & NA & \dots & 0 \end{bmatrix} \begin{array}{l} \leftarrow \text{Individual 1} \\ \leftarrow \text{Individual 2} \\ \vdots \\ \leftarrow \text{Individual n} \end{array}$$

SNP 1 SNP 2 SNP 3 SNP m

Data formats: the PLINK binary files

Plink saves in **binary format**

0, 1, 2, NA \rightarrow 00, 01, 11, 10

Why?

- 4 SNPs take only one byte (2 bits x 4 SNPs)
- **SNPs can be accessed instantly at any position without reading all the file**
- A normal text file must be read up to any position!

The SNP files are made of 3 formats:

.bim

.fam

.bed

Data formats: the PLINK binary files

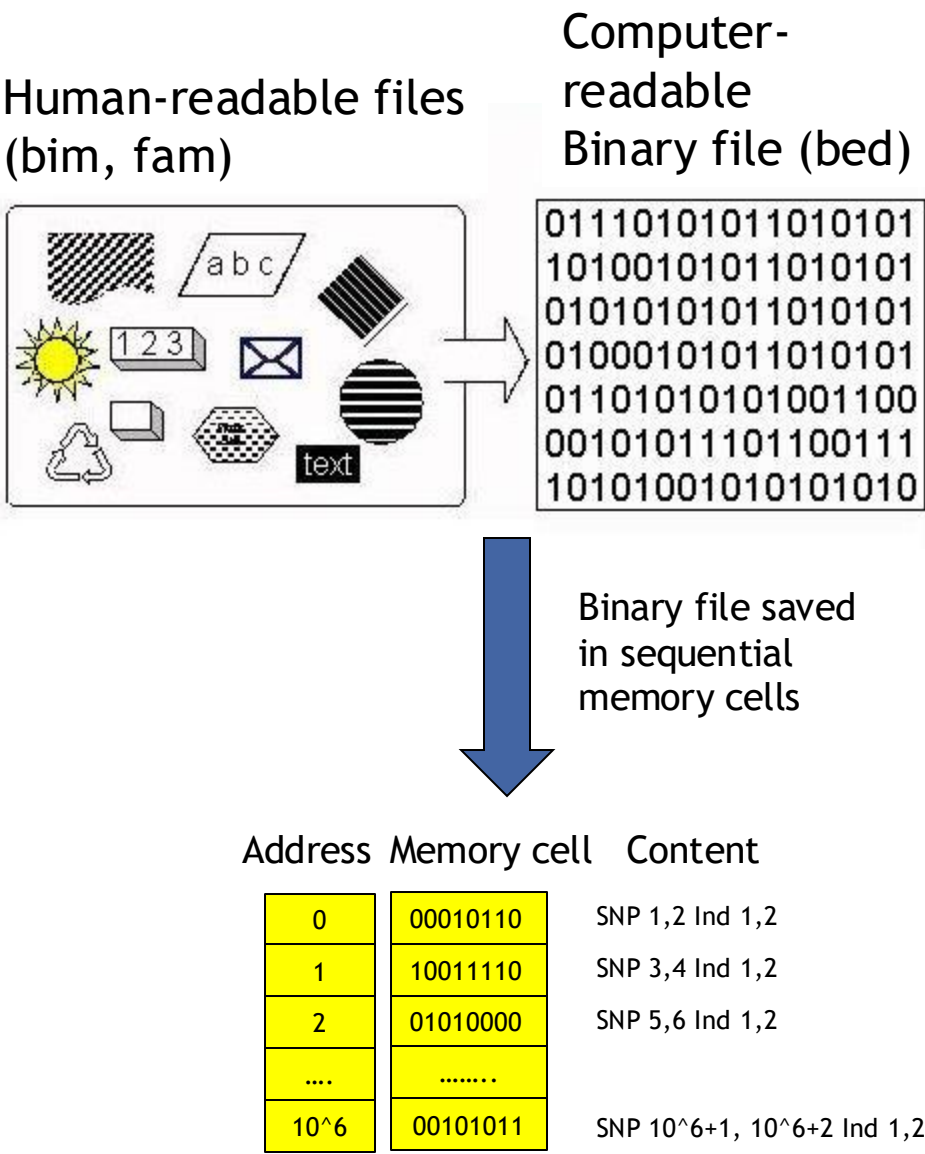
Plink saves in **binary format**
0,1, 2, NA → 00, 01, 11, 10

Why?

- 4 SNPs take only one byte (2 bits x 4 SNPs)
- **SNPs can be accessed instantly at any position without reading all the file**
- A normal text file must be read up to any position!

The SNP files is made of 3 formats:

.bim .fam .bed



Data formats: the PLINK binary files

*.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

Binary version of the info of fam and bim files					
---	--	--	--	--	--

*.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

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

Usually created during the analysis. Covariates can be e.g. from PCA, ...

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)



File Edit View Run Kernel Git Tabs Settings Help

+

+

+

work_ARM / Notebook

Name

exercises

Images

lib

references

general-intro.ipynb

GWAS2-DataColl...

GWAS3-QualityC...

GWAS4-QualityC...

GWAS5-Associat...

GWAS5b-Popula...

GWAS6-PRSAAnal...

GWAS7-PRSIIPy...

Interrupt Kernel

Restart Kernel...

Restart Kernel and Clear Outputs of All Cells...

Restart Kernel and Run up to Selected Cell...

Restart Kernel and Run All Cells...

Restart Kernel and Debug...

Reconnect to Kernel

Shut Down Kernel

Shut Down All Kernels...

Change Kernel...

CPU: 40% | Mem: 802 MB | Disk: 811.10 / 1023.50 GB

intro.ipynb


How to make the notebooks work


bash command line programming languages, where R is used for statistical analysis of the output from

g languages, you need to **choose a kernel every time we shift from one language to another**. A kernel

contains a programming language and the necessary packages to run the course material. To choose a kernel, go to the menu at the top of the page, select Kernel --> Change Kernel, and then select the preferred one.

- We will shift between two kernels, and along the notebook, you will see a picture indicating when to change the kernel. The two pictures are shown below:

 Choose the Bash kernel

 Choose the R kernel

- You can run the code in each cell (grey background) by clicking the run cell sign in the toolbar, or simply by pressing **Shift + Enter**. When the code is done running, a small green check mark will appear on the left side.
- You need to **run the cells sequentially** to execute the analysis. Please do not run a cell until the one above is done running, and do not skip any cells.
- Textual descriptions accompany the code to help you understand what is happening. Please try not to focus on understanding the code itself in too much detail, but rather focus on the explanations and commands' output.
- You can create new code cells by pressing **+** in the Menu bar above or by pressing **B** after selecting a cell.

Warning

- If a cell fails to run, verify the kernel in use, shown in the top-right corner.
- You don't know the answer to the exercises? You can use Generative AI to help with the code. It can assist you when the tutorial alone isn't enough or if you want to explore additional concepts beyond the exercise.

[]:

↑ ↓ ↶ ↷ ↵

Read the general intro carefully!

Would you like to get notified about official Jupyter news?
[Open privacy policy](#) Yes No

Simple 0 1 R | Idle Disk: 811.10 / 1023.50 GB | CPU: 0.00 % | Mem: 802.22 MB Mode: Command Ln 1, Col 1 general-intro.ipynb 1

 ~ 30min



GWAS2-DataCollection.ipynb



- Get familiar with PLINK file formats (.bim, .bed, .fam)
- Mice dataset



Choose the Bash kernel

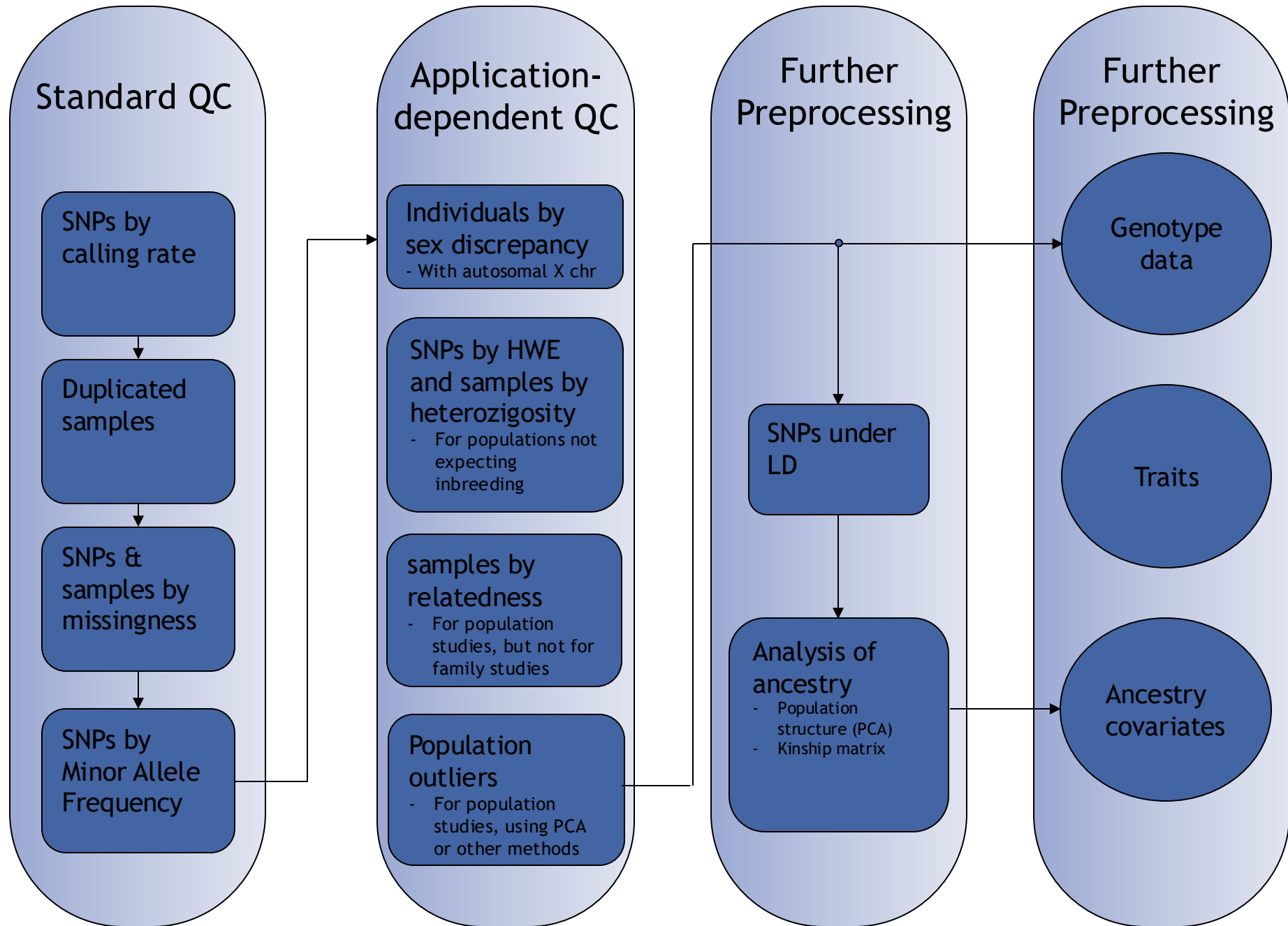


Choose the R-GWAS kernel

Solutions

- Problems/Issues/Comments?
 - Does it matter in which order you perform filtering at the individual- and variant-level?
- Mice dataset
 - Q1.** There are 1940 individuals and 2984 SNPs. Data from chromosomes 1-4. Cannot look for chromosome X, thus, inbreeding
 - Q2.** No parents' information.
 - Q3.** No information in the fam file
 - Q4.** The minimum value is -4.13 and the maximum is 3.60.

QC - what to filter?



QC - technical criteria



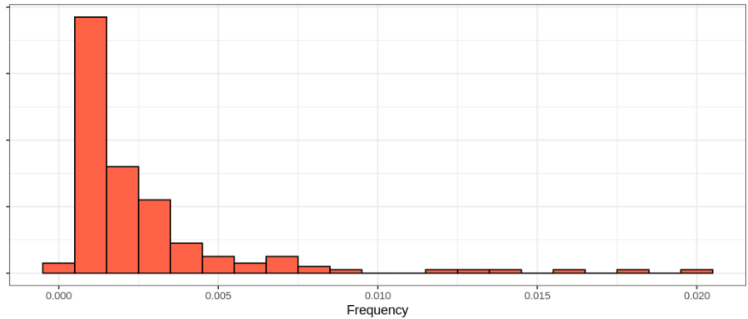
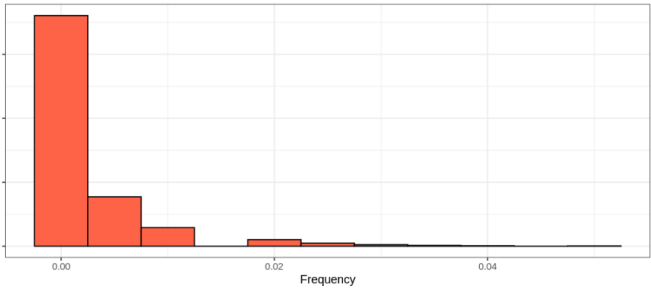
QC - technical criteria

SNP missingness

		/	SNP1	SNP2	SNP3	SNP4	SNP5
Sample duplication	IND1	22	00	11	12	22	
	IND2	22	00	11	12	22	
Sample missingness	IND3	11	12	11	22	21	
	IND4	00	00	11	11	00	
	IND5	22	00	11	22	22	

Note: duplication can originate from

- **Plate-ing errors,**
- duplicate samples from one of a pair of **genotyping chips,**
- **sample contamination.**

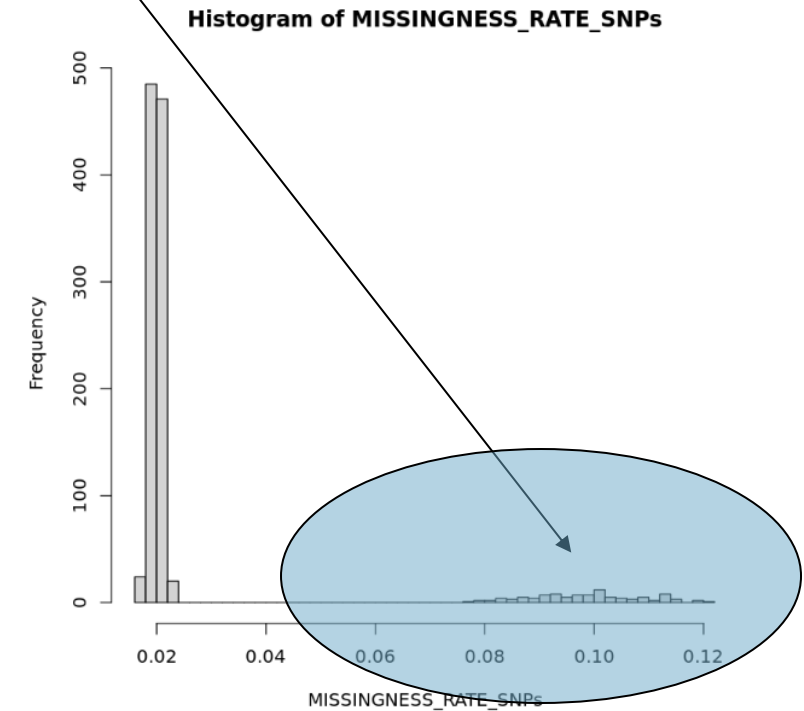
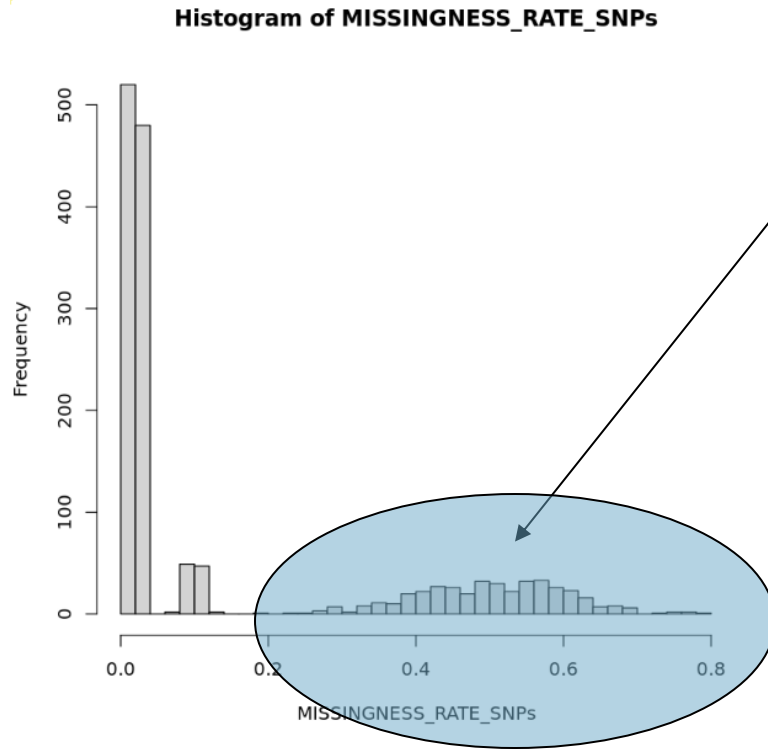


QC - technical criteria

1- Remove SNPs with very high missingness

2 - Filter samples based on missingness

3 - refilter SNPs



QC - gender (mis)match



female

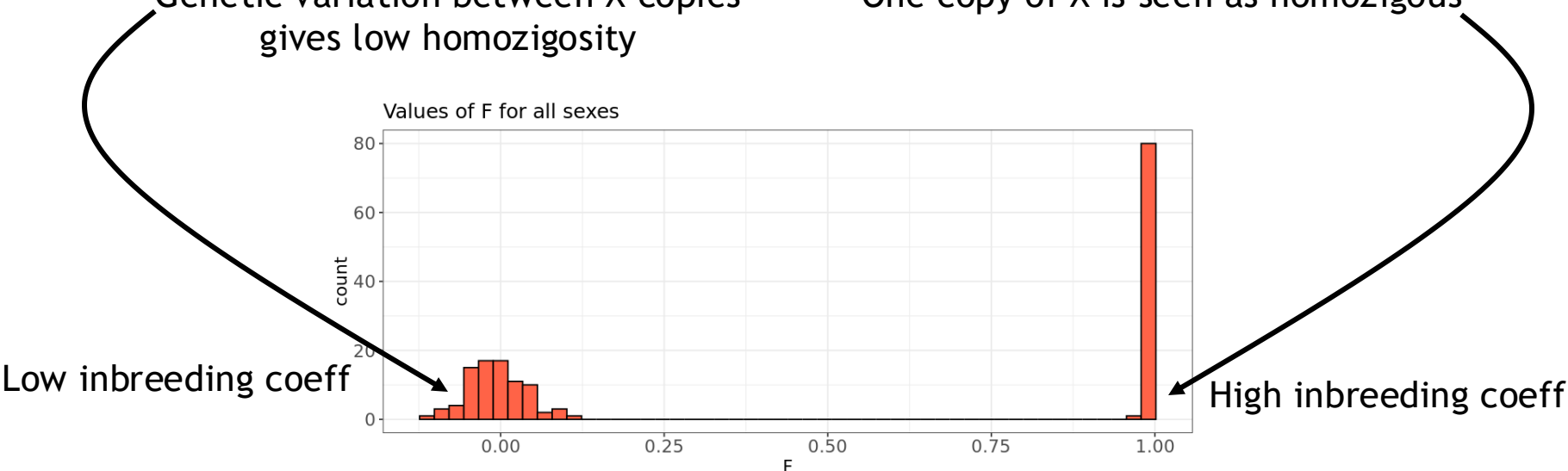


male

Inbreeding coefficient:
prob. of getting identical
alleles from two common
ancestors

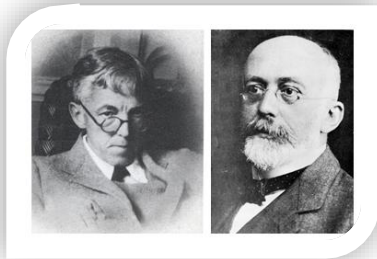
Genetic variation between X copies
gives low homozygosity

One copy of X is seen as homozygous



QC - Hardy Weinberg Equilibrium (HWE)

The ideal HWE world



“ Genetic variation in a population remains constant if

- Random mating
- No gene flow, mutation or selection
- Same allele frequencies across sexes
- Infinite size population
- Non-crossing generations

”

Never really holds because the above are not all satisfied in nature over time

Useful as a baseline for scientists to test against

QC - Hardy Weinberg Equilibrium

HWE for filtering

Freq dominant allele \swarrow p + q = 1 \nwarrow Freq recessive allele

$$p^2 + 2pq + q^2 = 1$$

How to check if a SNP is under HWE or not?

- Count observed Genotypes from the SNP

(#AA #Aa #aa)

- Calculate observed allele freq

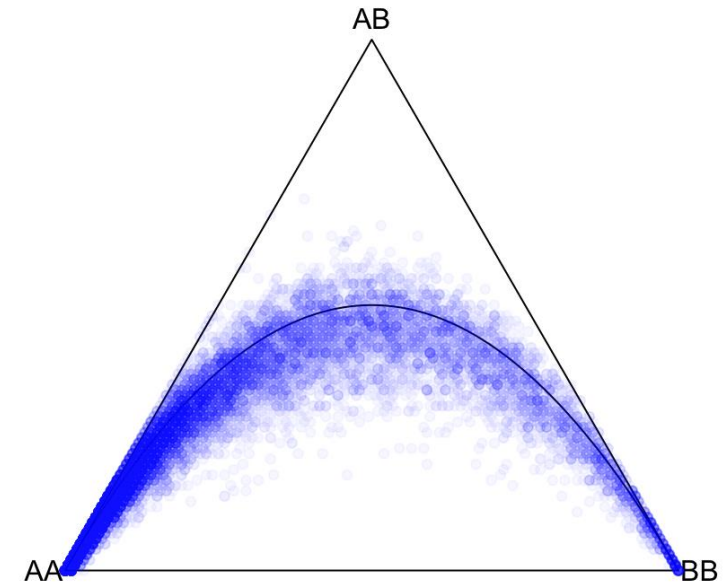
$$p = \frac{2\#AA + \#Aa}{2n} \quad \text{and} \quad q = 1 - p$$

- Calculate expected allele freq

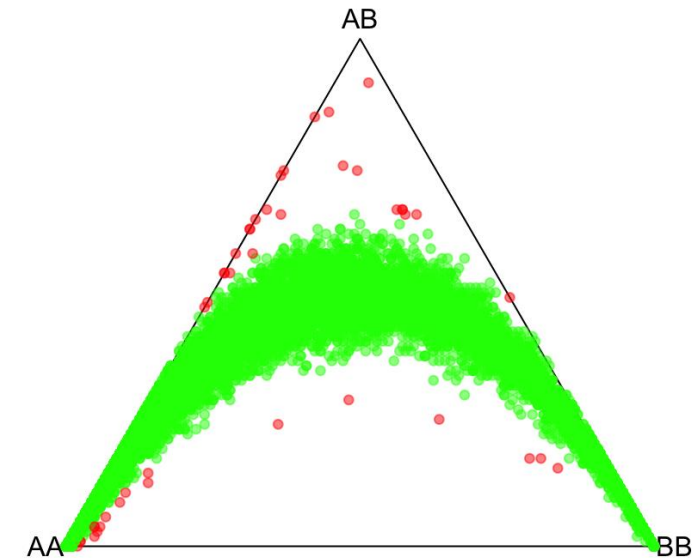
$$[E_{AA}, E_{Aa}, E_{aa}] = [np^2, 2npq, nq^2]$$

- Chi-square p-value of expected vs observed

$$\chi^2 = \frac{(\#AA - E_{AA})^2}{E_{AA}} + \frac{(\#Aa - E_{Aa})^2}{E_{Aa}} + \frac{(\#aa - E_{aa})^2}{E_{aa}}$$



Samples on the ideal HWE line



Samples departing from HWE

QC - Hardy Weinberg Equilibrium

HWE for filtering

Freq dominant allele \swarrow p + q = 1 \nwarrow Freq recessive allele

$$p^2 + 2pq + q^2 = 1$$

How to check if a SNP is under HWE or not?

- Count observed Genotypes from the SNP

(#AA #Aa #aa)

- Calculate observed allele freq

$$p = \frac{2\#AA + \#Aa}{2n} \quad \text{and} \quad q = 1 - p$$

- Calculate expected allele freq

$$[E_{AA}, E_{Aa}, E_{aa}] = [np^2, 2npq, nq^2]$$

- Chi-square p-value of expected vs observed

$$\chi^2 = \frac{(\#AA - E_{AA})^2}{E_{AA}} + \frac{(\#Aa - E_{Aa})^2}{E_{Aa}} + \frac{(\#aa - E_{aa})^2}{E_{aa}}$$

SNP with extremely small p-value:

- Allele calling errors
- Mating not random
- Strong selection
-

Note:

- HWE baseline can be **asking for too much** (illness groups are e.g. under selection)
- usual p-values (1e-5 to 1e-15) not effective on very large sample size ([Greer et al, 2024](#))
- Departure from HWE can hide other factors or be due to many other causes ([Pearman et al, 2015](#))

QC - Hardy Weinberg Equilibrium

HWE for filtering

SNP with extremely small p-value:

- Allele calling errors
- Mating not random
- Strong selection
-

Note:

- HWE baseline is just asking for too much
- usual p-values ($1e-5$ to $1e-15$) not effective on very large sample size ([Greer et al, 2024](#))
- Departure from HWE can hide other factors or be due to many other causes ([Pearman et al, 2015](#))

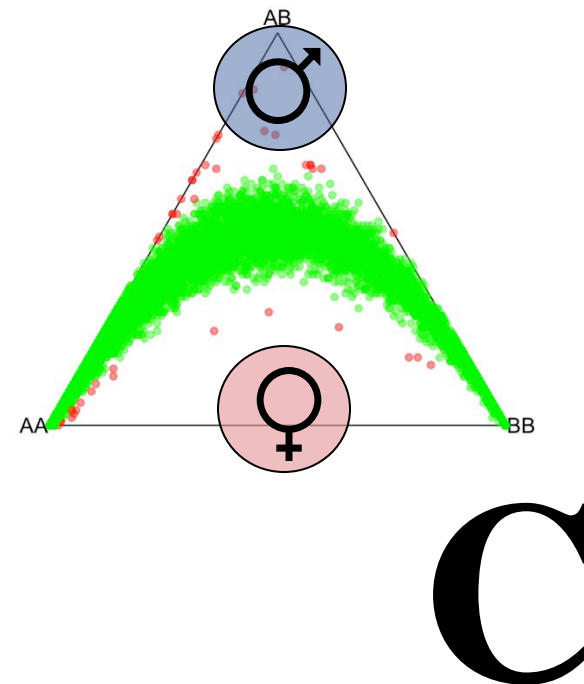
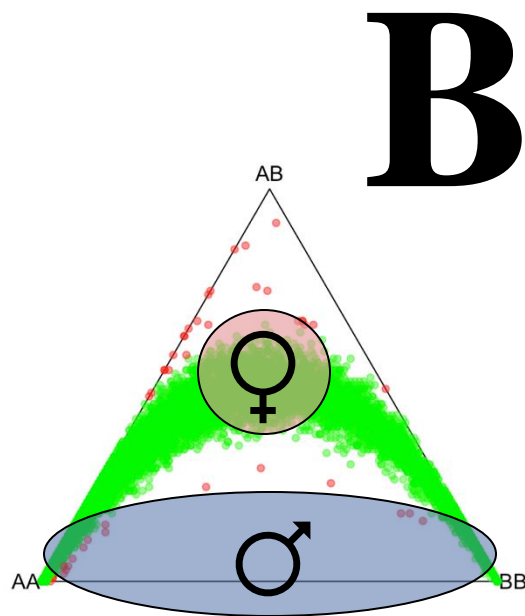
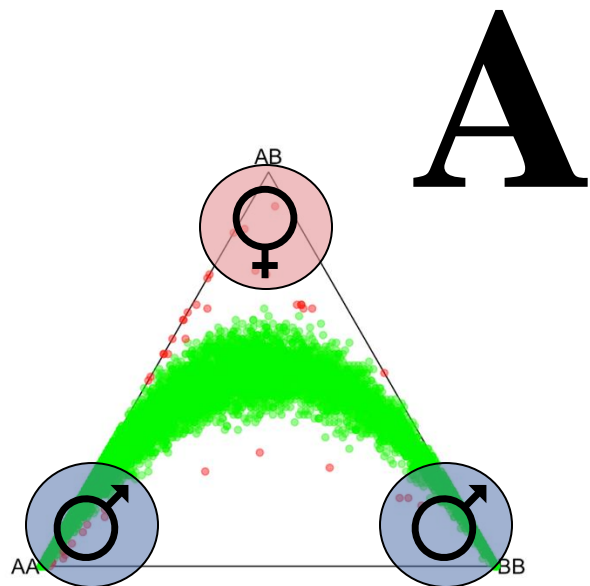
Consider:

- Calculating p-values of HWE and combining it with MAF values before filtering
- Remember your pop structure: admixed/diverse populations depart from HWE
- Selection affects disease SNPs
 - Case-Control studies: HWE filtering only on Controls to avoid removing disease SNPs

QC - Hardy Weinberg Equilibrium

HWE for filtering

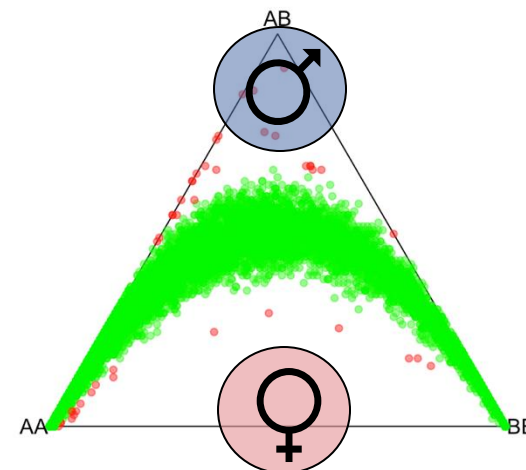
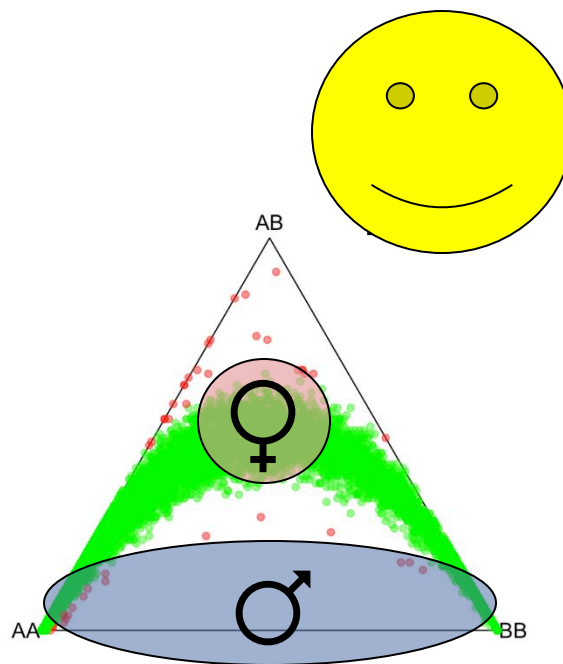
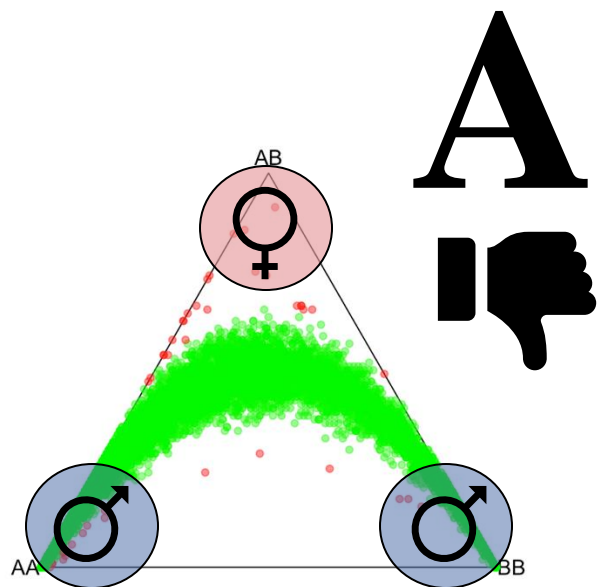
If you were calculating HWE only on chromosome X, where would you find males and females in the triangle?



QC - Hardy Weinberg Equilibrium

HWE for filtering

If you were calculating HWE only on chromosome X, where would you find males and females in the triangle?





GWAS3-QualityControlA.ipynb



Quality control: initial steps

- Individual Missingness
- Sex discrepancy
- Minor allele frequency
- Hardy-Weinberg Equilibrium (HWE)
- Heterozygosity Rate



Choose the Bash kernel



Choose the R-GWAS kernel

Recap, important aspects

- What happens if I don't remove rare variants?

General: increase the # snps, so, # tests -> harder to get significant results (also penalising the power on common variants)

Rare variants: Large effect size, no statistical power (too low frequency)

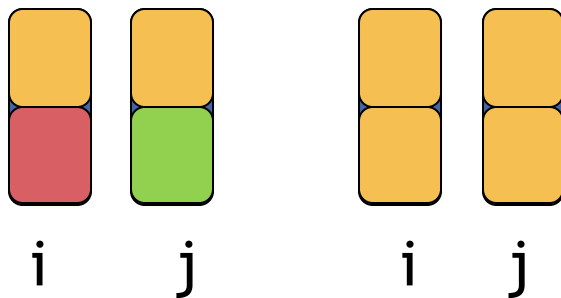
- What if I am interested in rare variants? How can we gain statistical power?

Burden tests (weighted sum inside a gene or specified region)

QC - relatedness

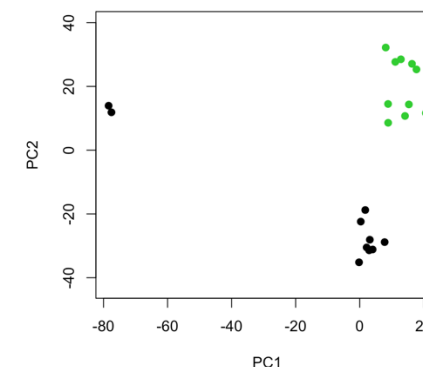
Pi-hat approximation in PLINK
(genomes i,j)

$$\hat{\pi}_{ij} = \frac{1}{2} \text{IBD1}_{ij} + \text{IBD2}_{ij}$$

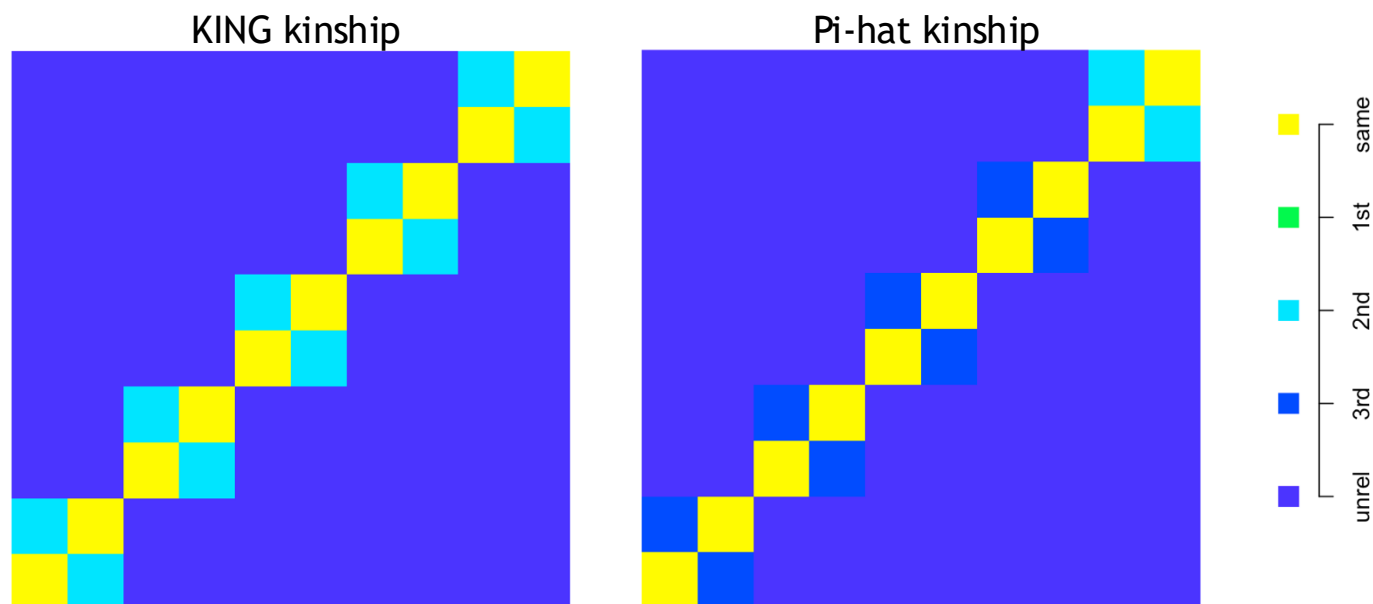


Bad at finding family
relatedness in presence of
population structure.
KING estimator is more robust.

Right: Pca of data simulated with
Population and family structure



Below: Pi-hat misses almost
all half-siblings from the data

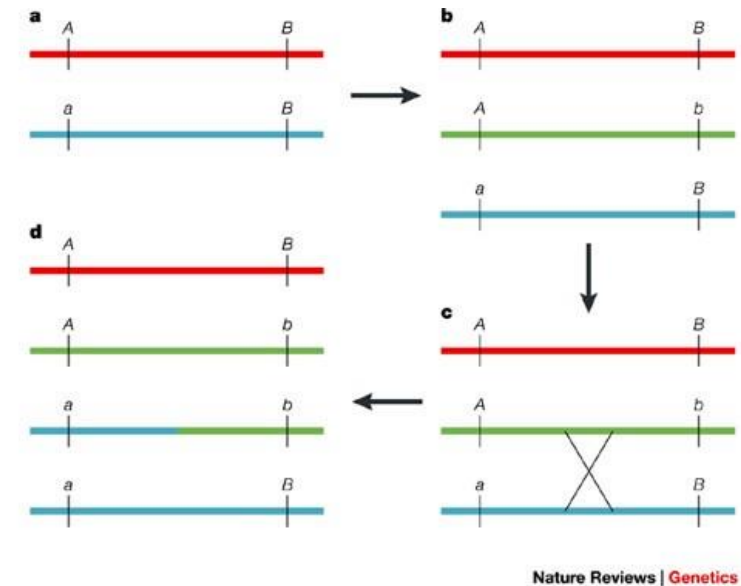


Modified from Matti Pirinen's course

QC - Linkage Disequilibrium (LD)

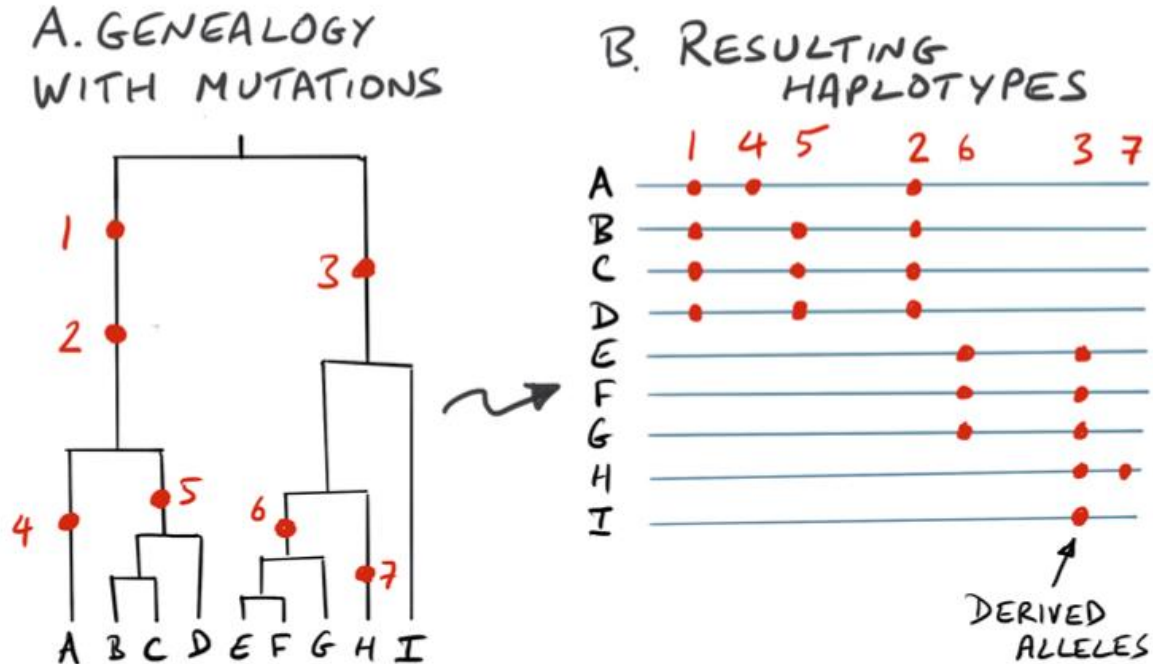
Linkage equilibrium: haplotype frequencies in a population have the same value that they would have if the genes at each locus were **combined at random**.

Linkage disequilibrium: **Non-random association** of alleles at different loci in a given population



- Recombination is the main destructive force of LD
- Physically close SNPs retain high LD for longer under recombination force

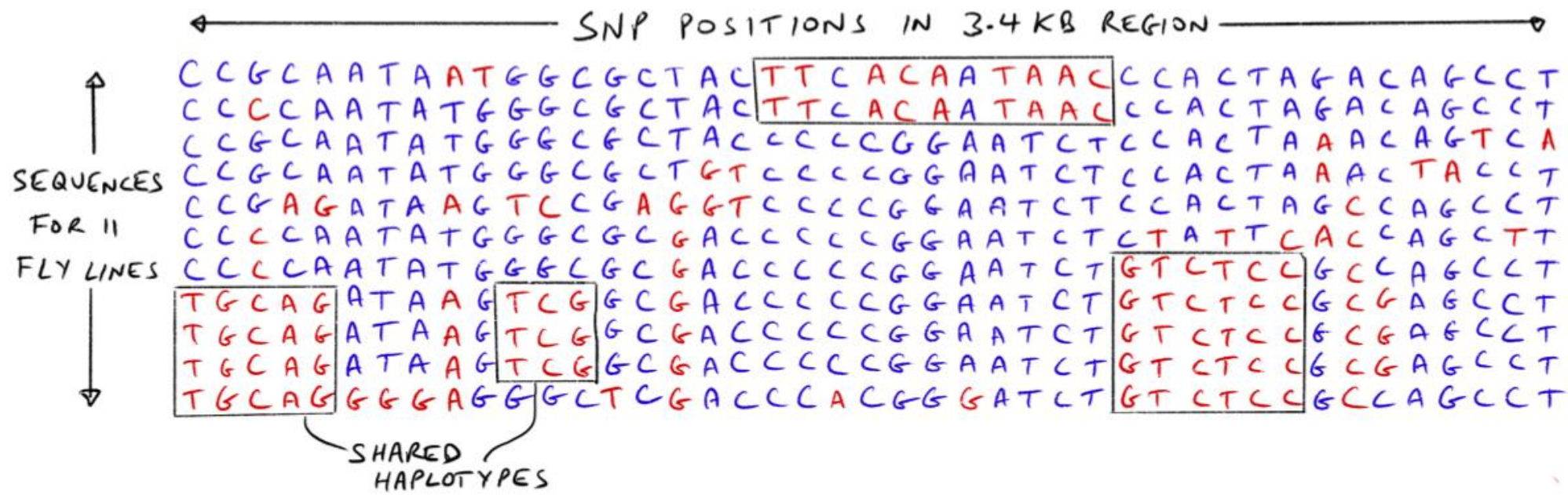
QC - Linkage Disequilibrium (LD)



Phylogeny without
any recombination
is in full LD

Reflected in the
haplotype structure

QC - Linkage Disequilibrium (LD)



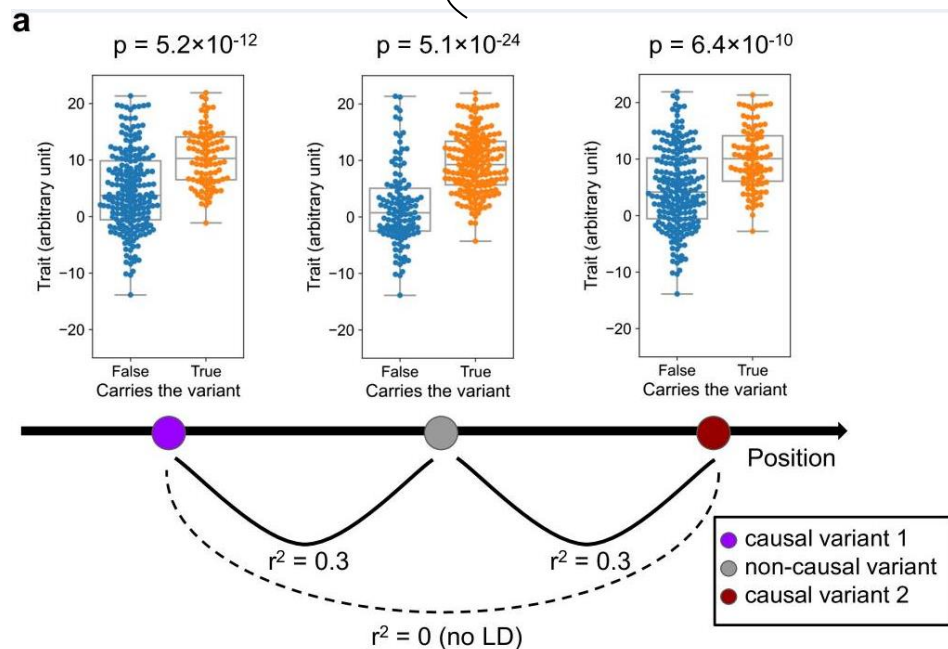
LD shows in blocks of shared haplotypes across samples

QC - Linkage Disequilibrium

Causality and colocalization

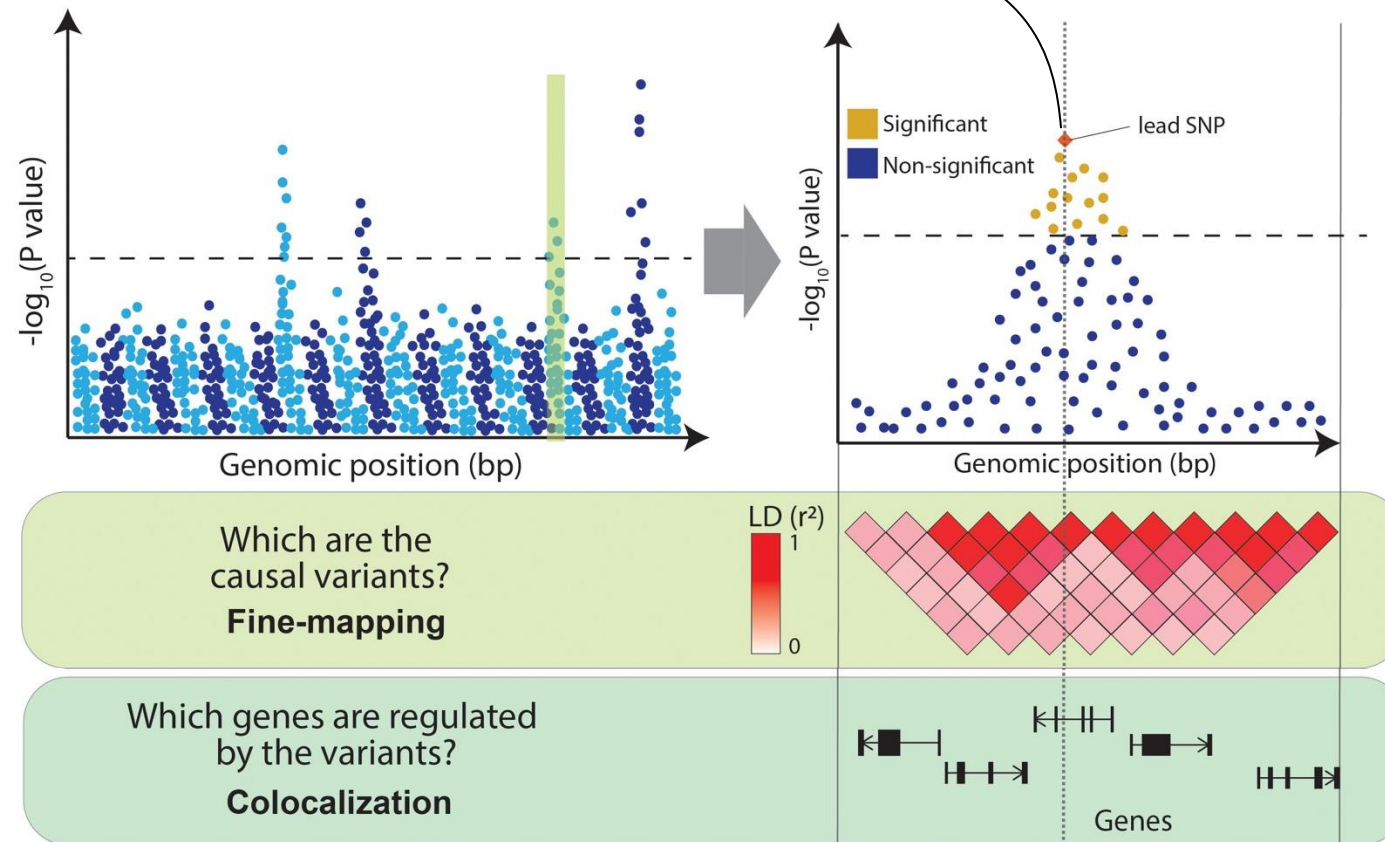
Why is LD important?

The most significant SNP in GWAS is often not the functional variant but in LD with it.



PLINK note:

PLINK uses r-squared (squared correlation of loci) and not D' for LD values



 ~ 1h



GWAS4-QualityControlB.ipynb



- Relatedness (ONLY)



Choose the Bash kernel



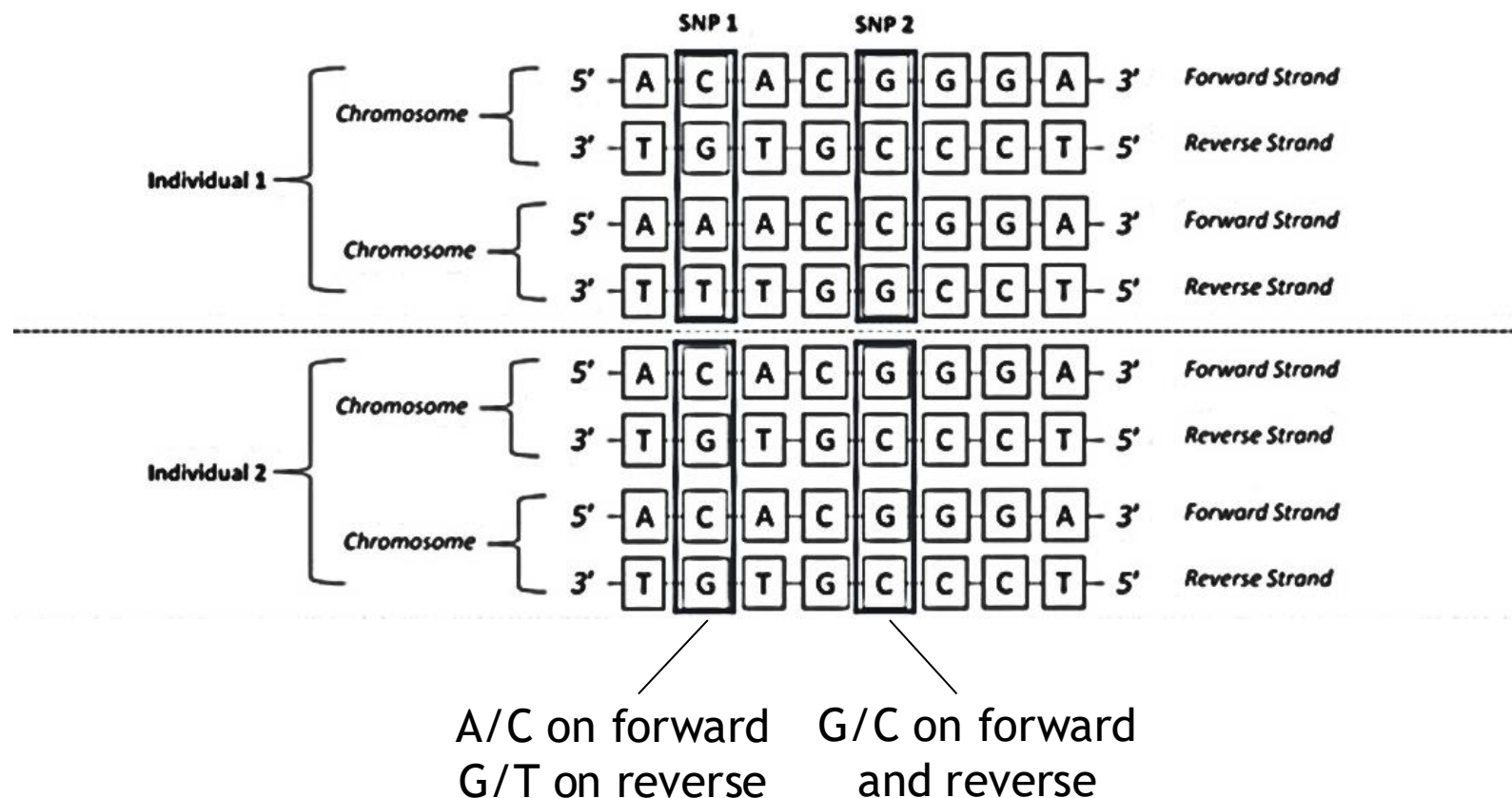
Choose the R-GWAS kernel

Solutions

- Problems/Issues/Comments?

QC - alignment 1: strands

Needed for data integration from various sources



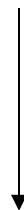
SNP1: flip all G/T's to C/A's in the SNP data
SNP2: unknown genotyping strand

QC - alignment 2: SNPs from different reference builds

Needed for data integration from various sources

Example: <https://www.ncbi.nlm.nih.gov/snp/?term=rs4149056>

☐ rs4149056 [*Homo sapiens*]
1.
Variant type: SNV
Alleles: T>A,C [Show Flanks]
Chromosome: 12:21178615 (GRCh38)
12:21331549 (GRCh37)



Two datasets aligned respectively to reference builds 37 and 38 will need proper merging by moving the location of one build to the other so that they match

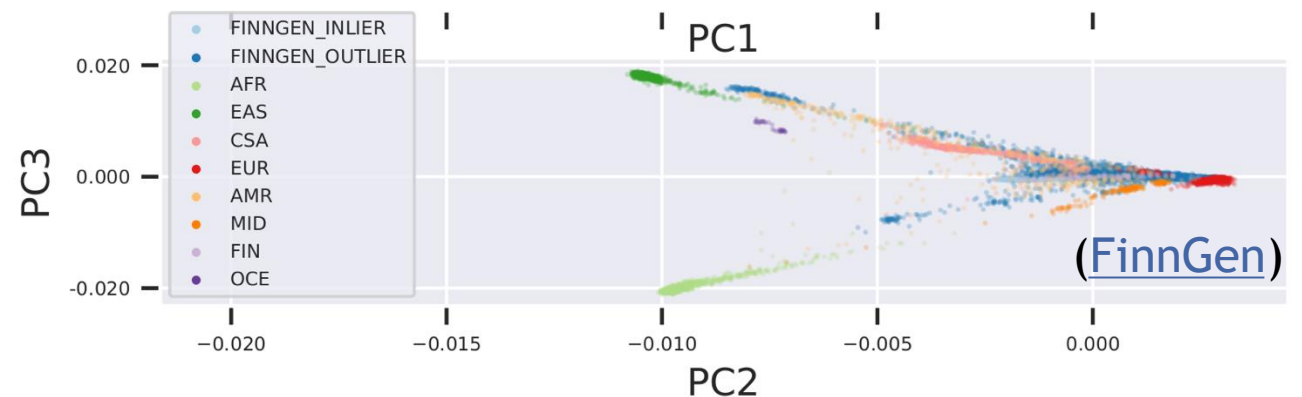
(Novembre et al, 2008)

(Novembre et al, 2008)

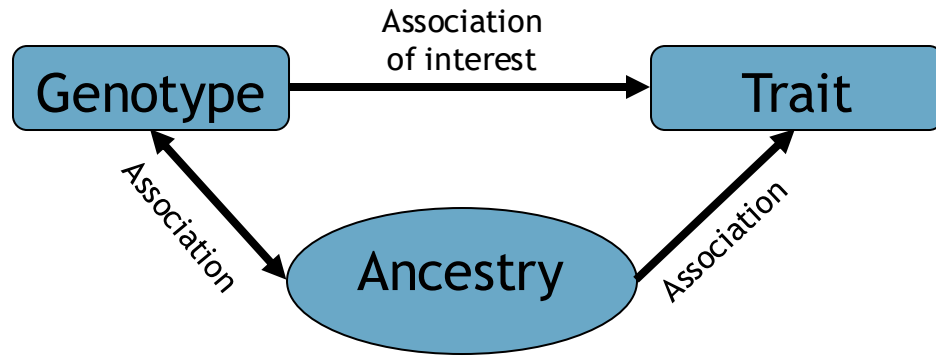


FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

- Coordinates embedded in the PCA axes are fundamental covariates for better GWAS studies
- PCA can be used to find (ethnic) outliers
- You can anchor new data to a PCA of known ethnic data to see its population structure

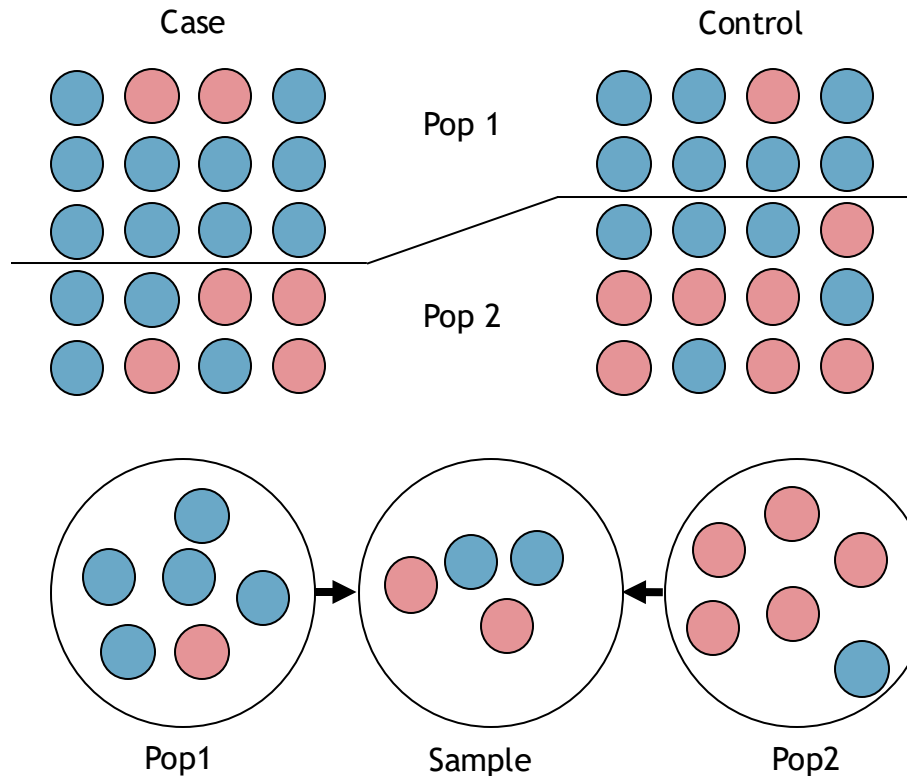


QC - PCA - population covariates?



How a population (or other) covariate acts on traits:

- Influence on SNPs & trait
 - MAF changes if pop structure is not modelled
 - In all type of studies
- Modeling only $\text{SNP} \rightarrow \text{Trait}$ is not enough



More on pop structure in modeling associations in later slides of the course

 ~ 1.5h



GWAS4-QualityControlB.ipynb



- Strand/Mapping correction
- Data merging
- Anchoring
- Population covariates
- Mice dataset



Choose the Bash kernel



Choose the R-GWAS kernel

Solutions

- Problems/Issues/Comments?
 - Why don't we use the covariates from MDS on HapMap-CEU data anchored by 1000 Genomes data in the GWAS analysis?
 - MDS vs. PCA
- Mice dataset

Wrap up

