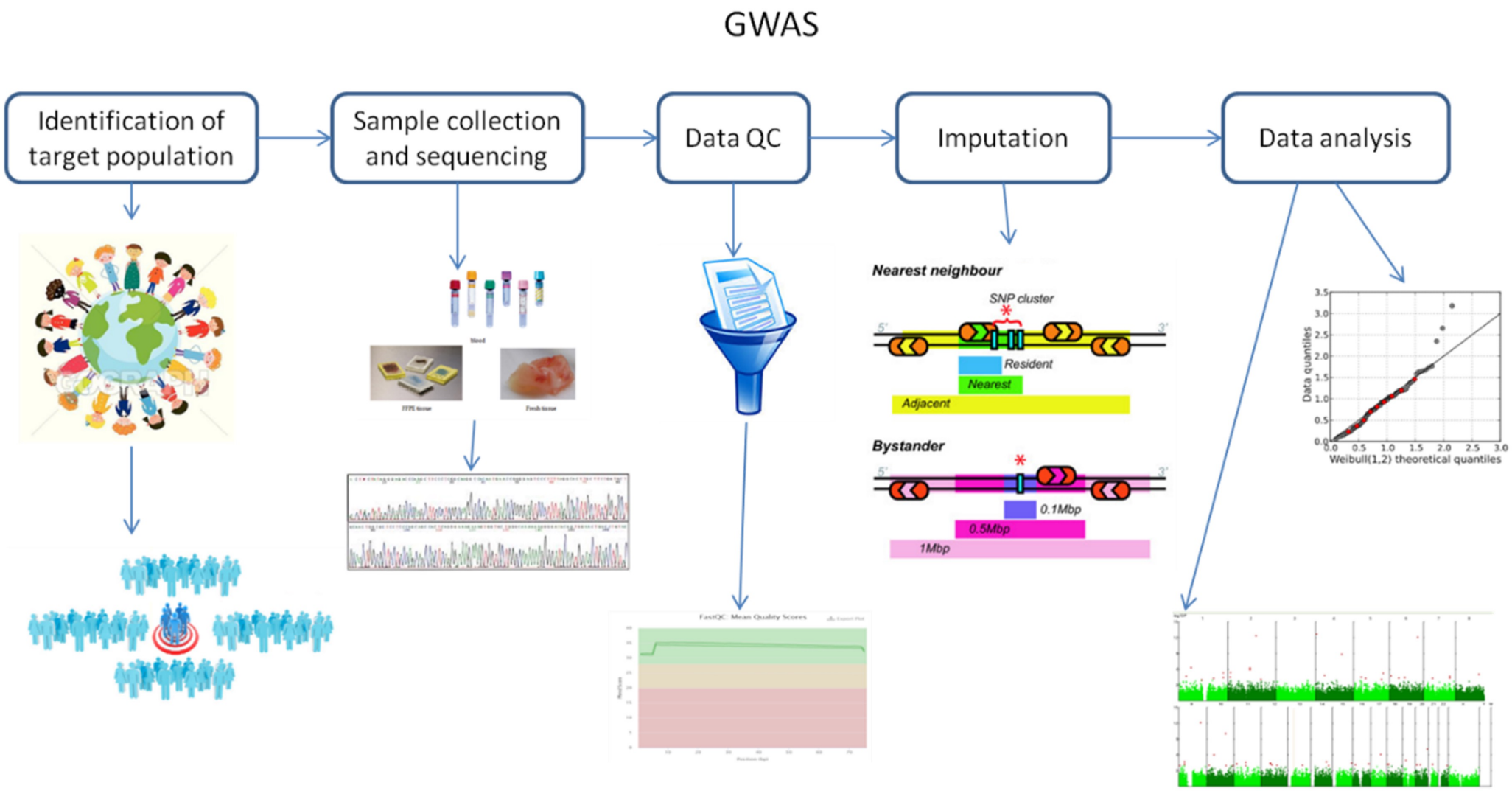
GWAS tutorial

Teaching fellows: Carla Marquez-Luna and Kumar Veerapen

UCT training

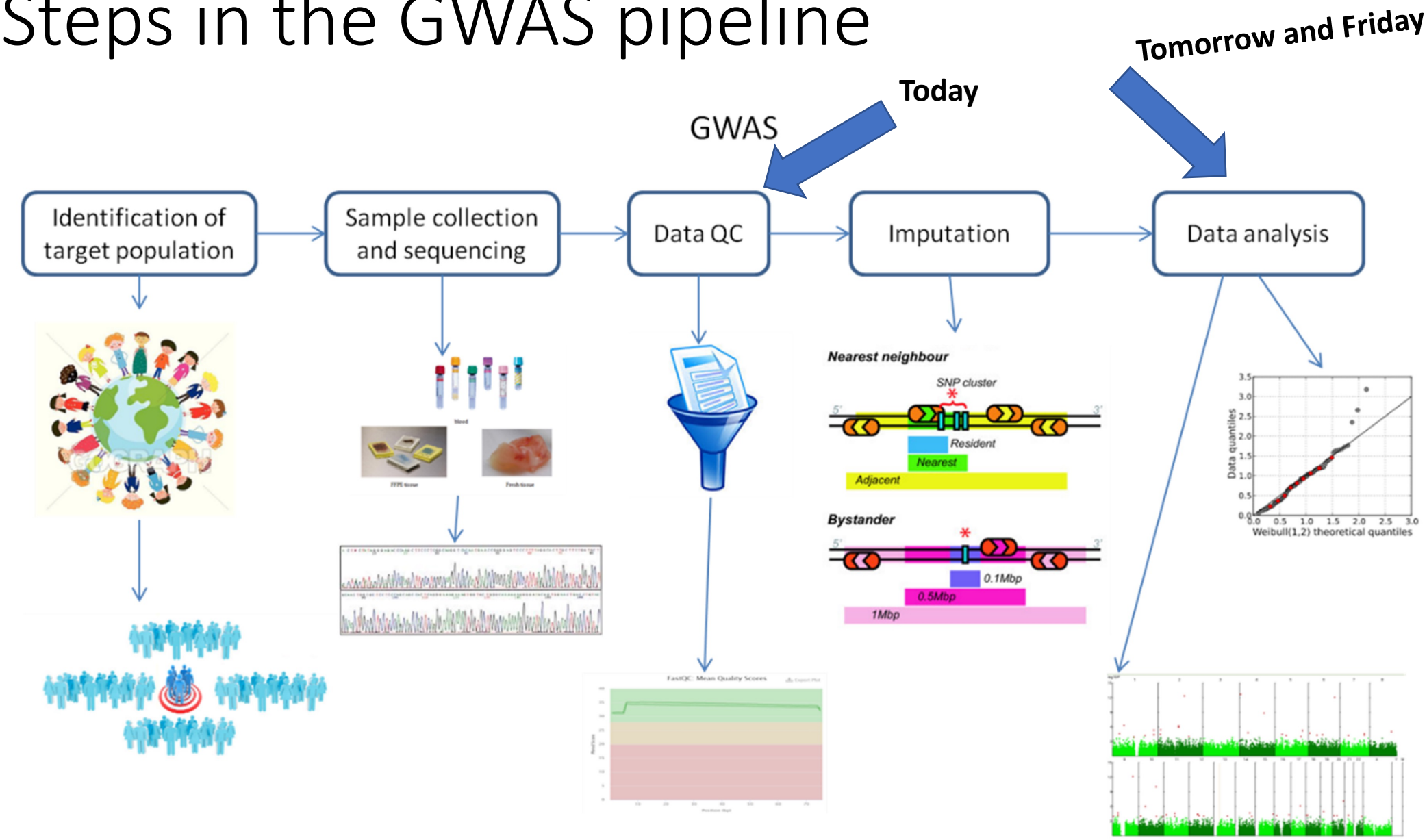
April 2022

Steps in the GWAS pipeline



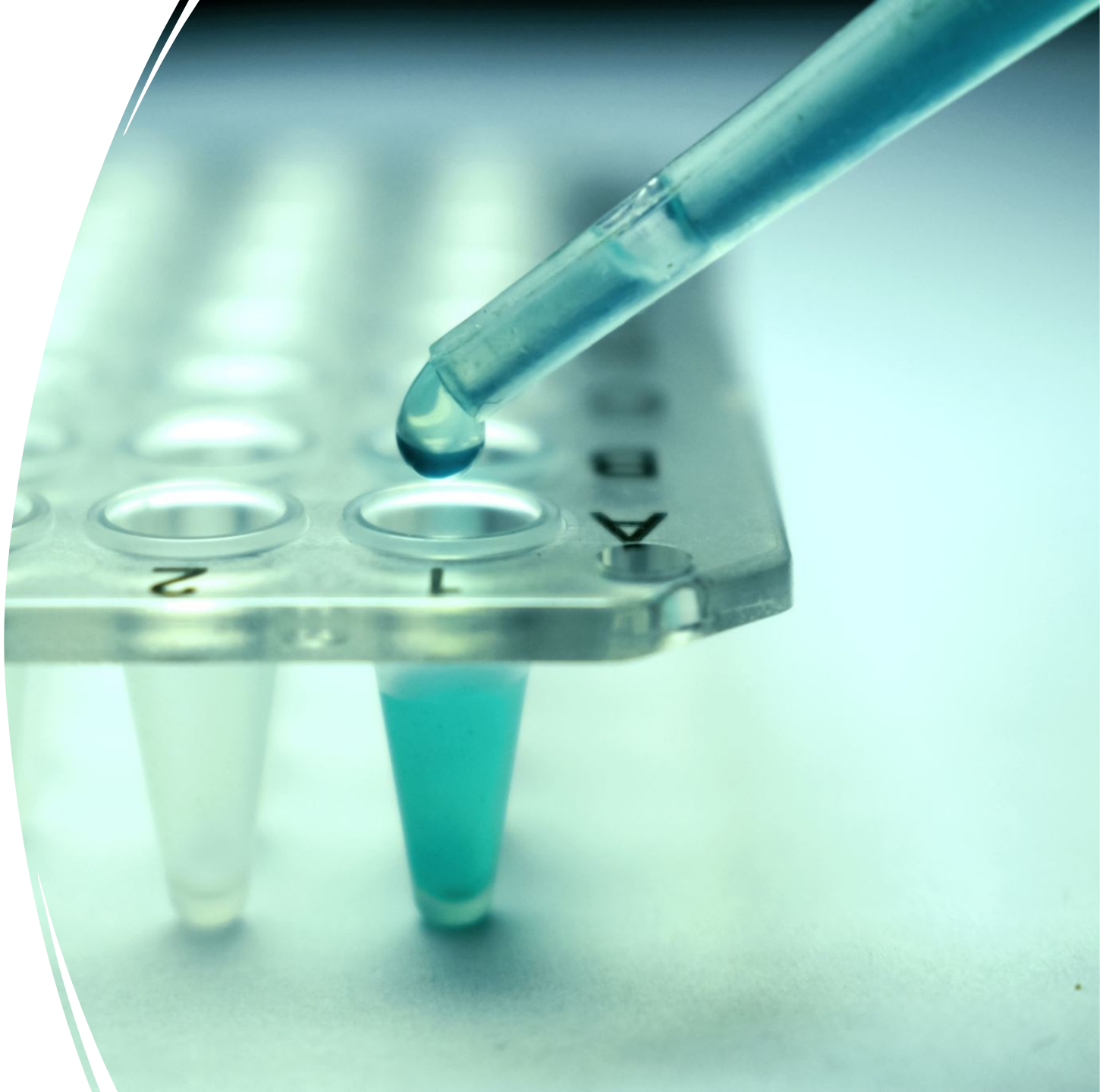
Jurj et al, Int. J. Mol. Sci. 2020

Steps in the GWAS pipeline



Jurj et al, Int. J. Mol. Sci. 2020

Quality control



QC for raw genotype data

Subjects

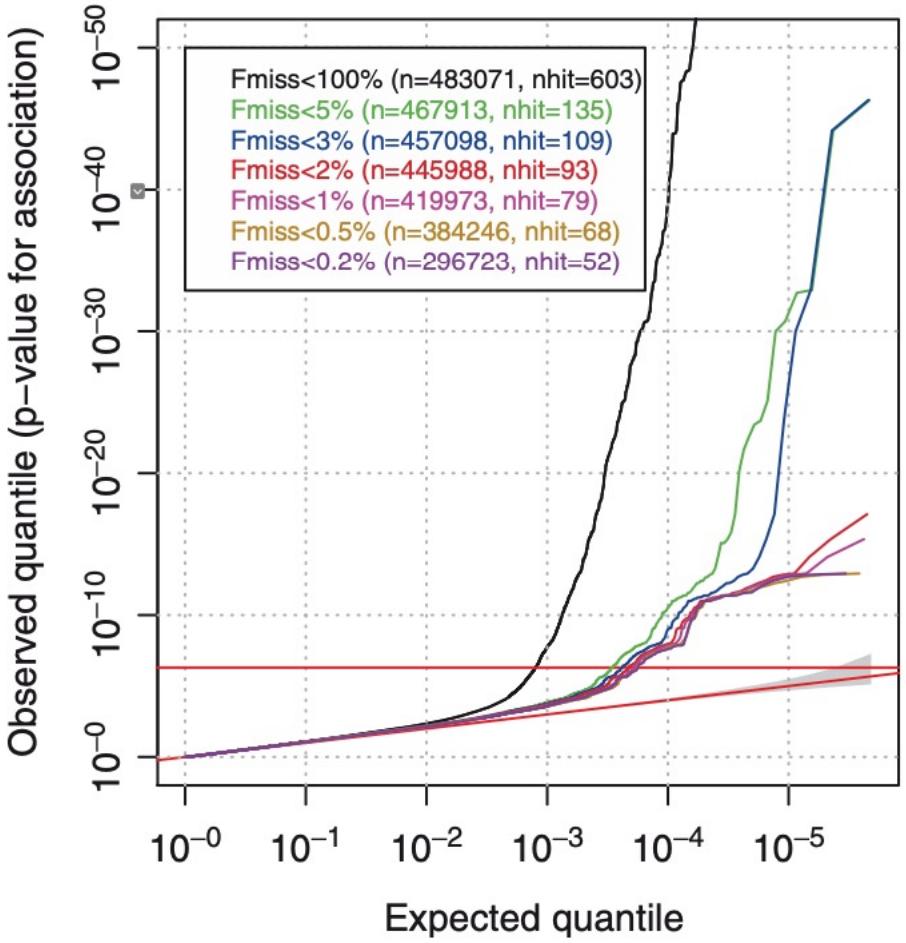
- Sex discordance [possible data mislabeling or contamination]
- Missing genotype and heterozygosity rate [sample quality / contamination]
- Duplication [confounding]
- Ancestry and cryptic relatedness [confounding]

Variants (SNPs)

- Missing genotype rate (cases and controls separately and together) [quality]
- Deviation from Hardy-Weinberg equilibrium (HWE) [genotype error/selection (controls only)]
- Minor allele frequency [power]

SNP QC - Missingness

- It arises because missingness is not randomly distributed among genotypes but rather is over-represented in, say, AA and AB calls.
- False positives will arise if DNA quality differs with phenotype, leading to differences in the frequency of called genotypes
- False negatives will arise if the informative missingness signal acts in the opposite direction to the real signal, or by reducing power via a reduced sample size for nonmissing values.



QQ plot applied after different missingness thresholds
The pool of hits contains both real and false positive signals.

Michael E. Weale, Genetic Variation: Methods and Protocols, Methods in Molecular Biology, vol. 628, 2010

Sample QC - Gender Check

- Most likely due to labeling error
- Usually, it is impossible to tell whether this is a labelling error involving just the gender label or whether it is a more serious labelling error linking the wrong DNA sample to the wrong clinical record.
- What do we do? We usually remove them.
- We look at the heterozygocity rate based on the X chromosome. Too many heterozygous SNPs tend to be male, to many homozygous SNPs tend to be female.

Note: The function to apply this filter (--check-sex) is found in plink 1.9. For the purposes of this practice, we will not apply this filter to our data.

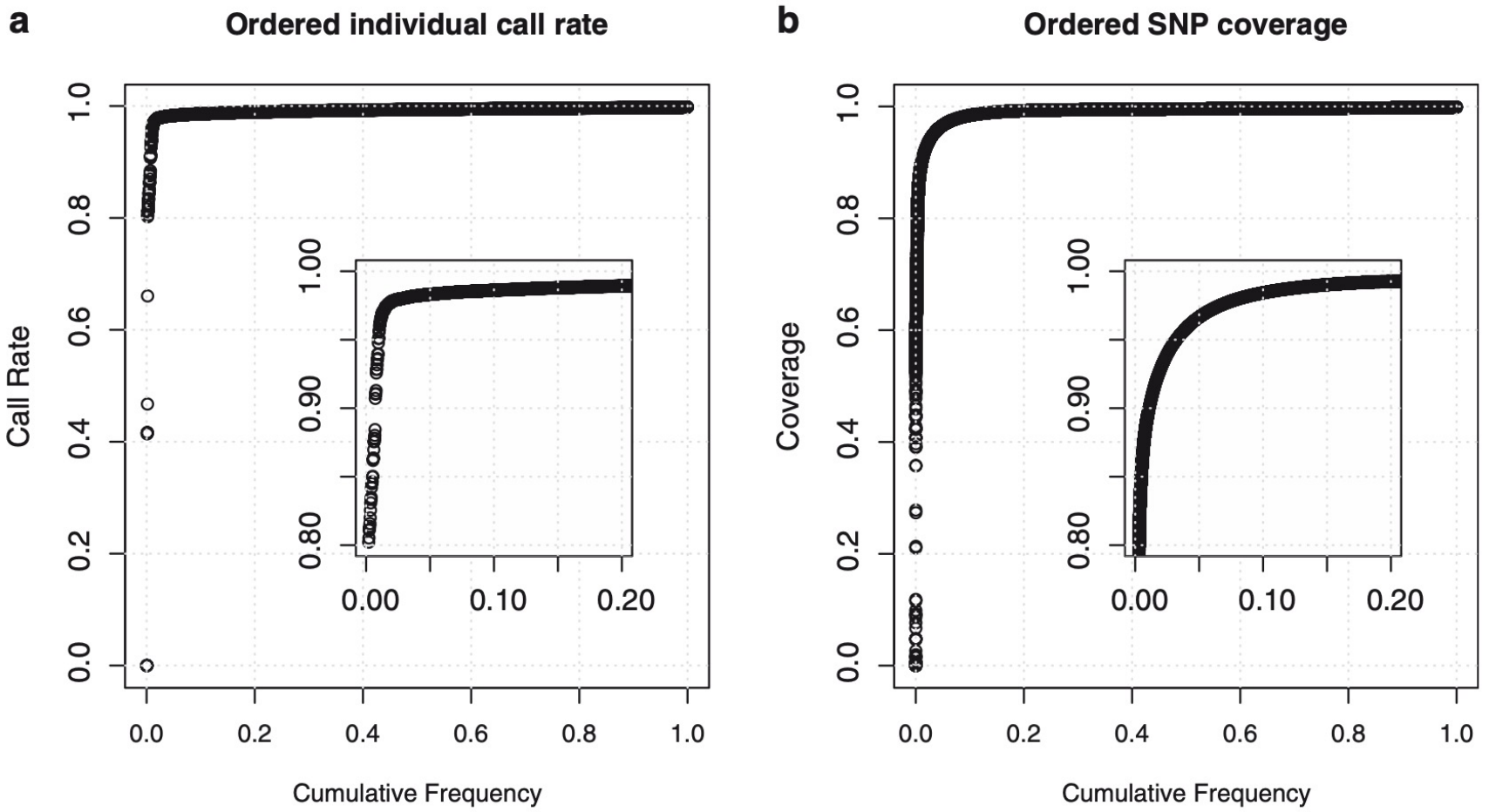


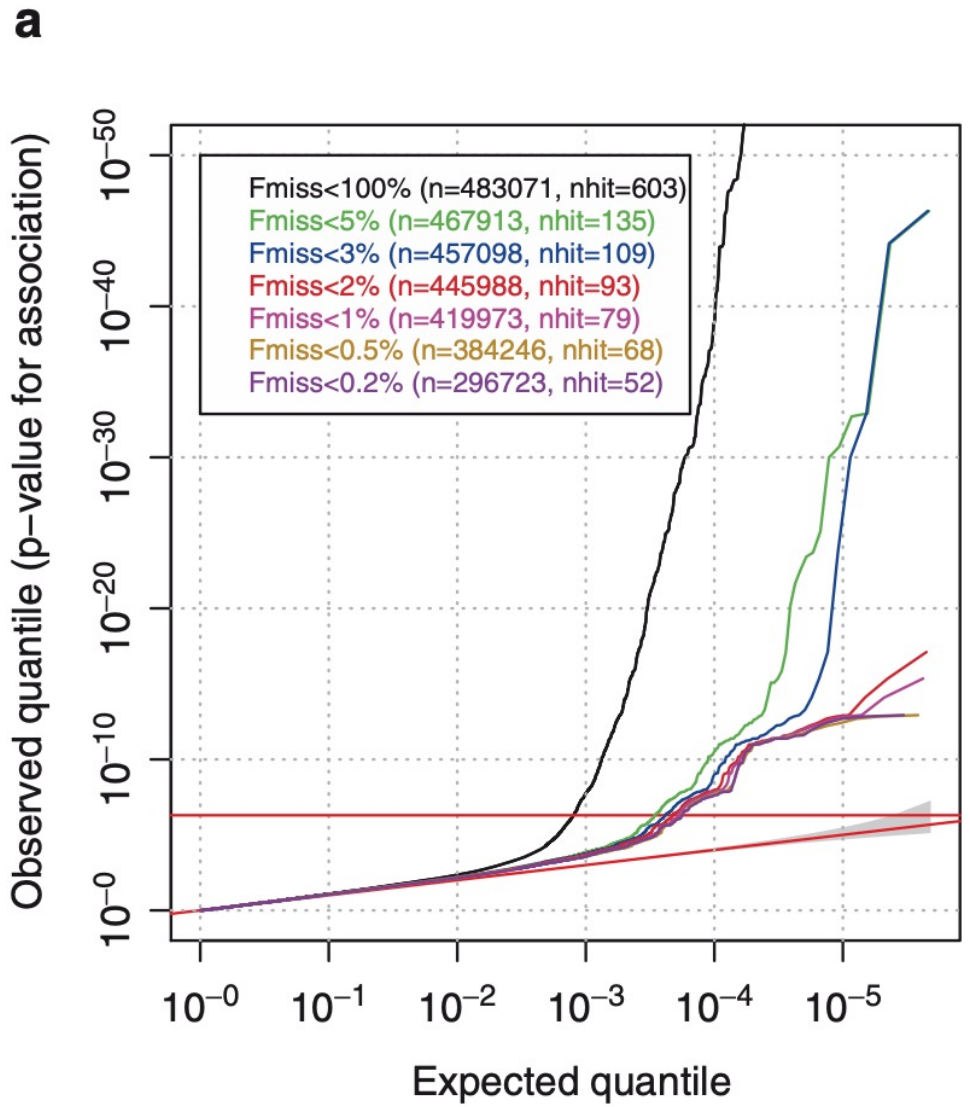
Fig. 2. Plots of one minus missingness ("call rate" for individuals and "coverage" for SNPs) against cumulative frequency (i.e., data points ordered from lowest to highest), using data (58C + NBS + CD cohorts) from (18). Insets show a zoom-in of the top left hand corner. R-code for plotting this figure is available at www.kcl.ac.uk/mmg/gwascode

Michael E. Weale, Genetic Variation: Methods and Protocols, Methods in Molecular Biology, vol. 628, 2010

SNP - QC

- **SNP missingness** is the complement of individual missingness. It is a must-have QC step due to the strong correlation of missingness with SNP quality and the impact of informative missingness on both false positive and false negative signals of association.

- For this example we can conclude Fmiss<0.02 appears to have little additional effect on the Q–Q plot, compared to setting Fmiss = 0.02, so this setting appears a good choice.



Michael E. Weale, Genetic Variation: Methods and Protocols, Methods in Molecular Biology, vol. 628, 2010

SNP - QC

- **Minor Allele Frequency:** Power to detect an association signal decreases with decreasing MAF. There seems to be little point in including SNPs below a certain MAF in your analysis, because you will never be able to detect an association signal with them and you will increase the number of tests performed.
- **Hardy-Weinberg Equilibrium:** Departure from Hardy-Weinberg equilibrium (HWE) can indicate problems with genotype calling. A common scenario is when two of the three clouds representing the three genotypes on an allele intensity plot overlap with one another, leading the genotype calling algorithm to mistake them for one single genotype. This almost always leads to a very large departure from HWE.

Remove duplicates and related individuals

- Cryptic relatedness occurs when, for unforeseen reasons, pairs or groups of individuals are more closely related to each other than the population average – thus indicating they are close family members.
- Sample duplicates can also be treated as an extreme case of cryptic relatedness.
- Individuals that are closely related induce a correlation structure that will upset downstream association analyses unless properly accounted for. This may introduce false positive and/or false negative results, depending on the situation.

Note: the decision to remove related individuals is up to the investigator. There are association methods called linear mixed models that can deal with cryptic relatedness. It will depend on which statistical method is more suitable for our study.

- The degree of relatedness is inferred from the estimated kinship coefficient using KING's criteria
- The *x*-axis shows the proportion of zero identity-by-state (IBS0), defined as the proportion of SNPs at which one sample carries the minor homozygote and the other sample – the major homozygote, so that they share no alleles.)
- Parent-child and full sibling pairs have the same expected kinship coefficient but can be distinguished by their IBS0 fraction, defined as the proportion of SNPs at which two samples have no alleles in common.

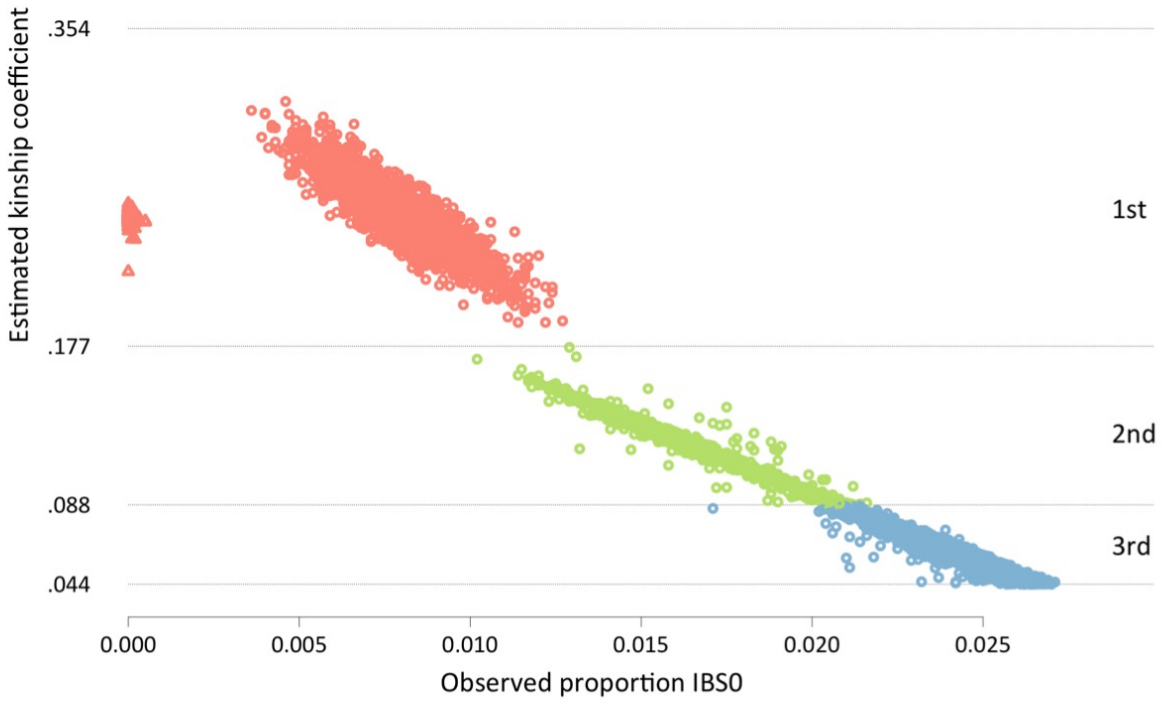


Figure 5 Close relationships for ~150,000 UK Biobank participants genotyped in the

Figure: Bycroft et al

Table 1. Relationship inference criteria based on estimating kinship coefficients (ϕ) and probability of zero IBD sharing (π_0)

Relationship	ϕ	Inference criteria	π_0	Inference criteria
Monozygotic twin	$\frac{1}{2}$	$> \frac{1}{2^{3/2}}$	0	< 0.1
Parent–offspring	$\frac{1}{4}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$	0	< 0.1
Full sib	$\frac{1}{4}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$	$\frac{1}{4}$	(0.1,0.365)
2nd Degree	$\frac{1}{8}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$	$\frac{1}{2}$	$(0.365, 1-\frac{1}{2^{3/2}})$
3rd Degree	$\frac{1}{16}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}})$	$\frac{3}{4}$	$(1-\frac{1}{2^{3/2}}, 1-\frac{1}{2^{5/2}})$
Unrelated	0	$< \frac{1}{2^{9/2}}$	1	$> 1-\frac{1}{2^{5/2}}$

*Inference criteria = Confidence interval

Manichaikul et al, Bioinformatics, 2010

Note that KING kinship coefficients are scaled such that duplicate samples have kinship 0.5, not 1. First-degree relations (parent-child, full siblings) correspond to ~0.25, second-degree relations correspond to ~0.125, etc. It is conventional to use a cutoff of ~0.354 (the geometric mean of 0.5 and 0.25) to screen for monozygotic twins and duplicate samples, ~0.177 to add first-degree relations, etc.

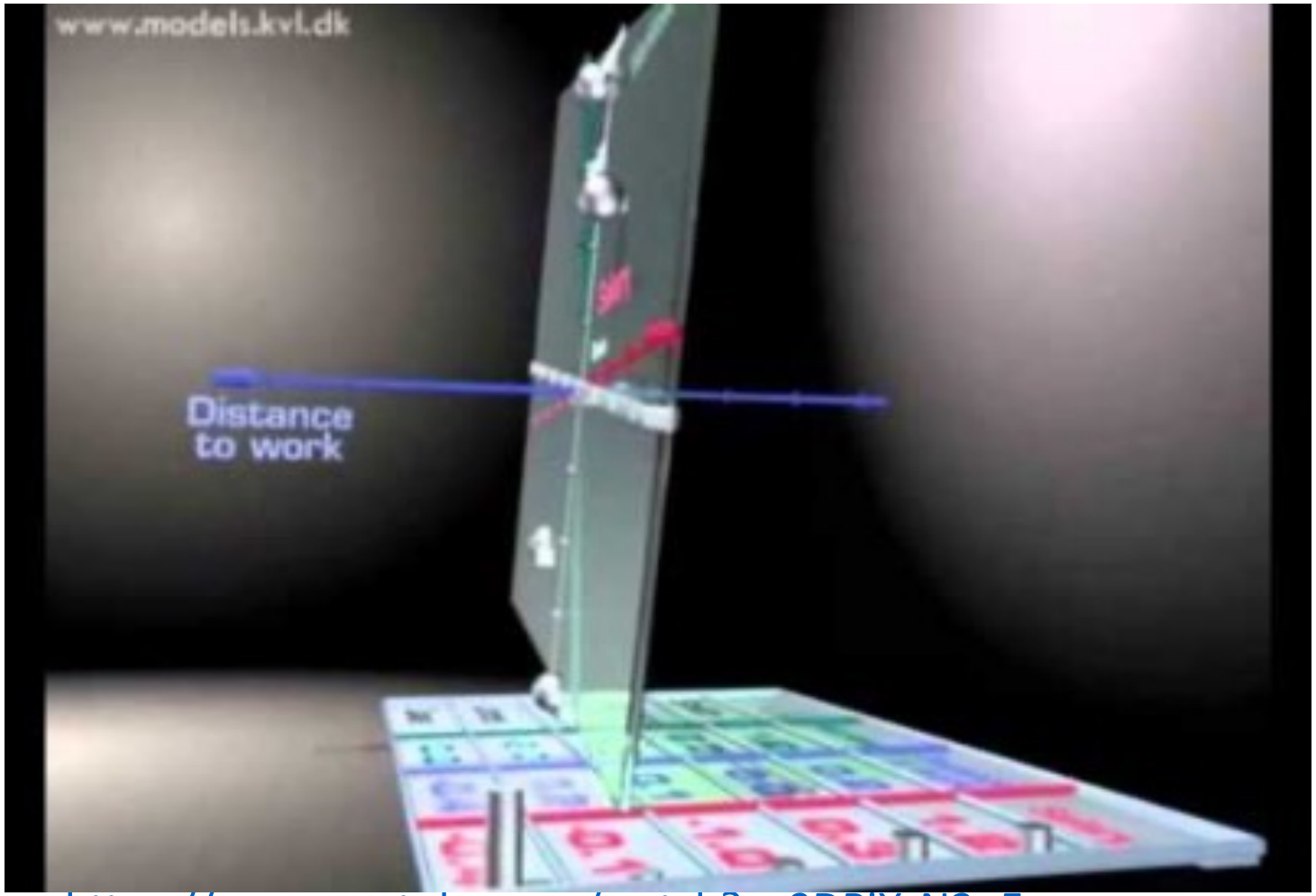
```
--make-king [{square | square0 | triangle}] [{zs | bin | bin4}]
--make-king-table ['zs'] ['counts'] ['rel-check'] ['cols='<col. set descrip.>]
--king-table-filter <min. kinship coefficient>
--king-table-subset <.kin0 file> [min. kinship coefficient]
```

--make-king writes KING-robust coefficients in matrix form to `plink2.king`[.zst] or `plink2.king.bin`, while **--make-king-table** writes them in table form to `plink2.kin0`[.zst]. (See above for matrix-output options.)

- Only autosomes are included in this computation.
- Pedigree information is currently ignored; the between-family estimator is used for all pairs.
- For multiallelic variants, REF allele counts are used.
- --make-king jobs with the 'square0' or 'triangle' output shapes and all --make-king-table jobs can be subdivided with --parallel.

Principal component analysis

PCA example



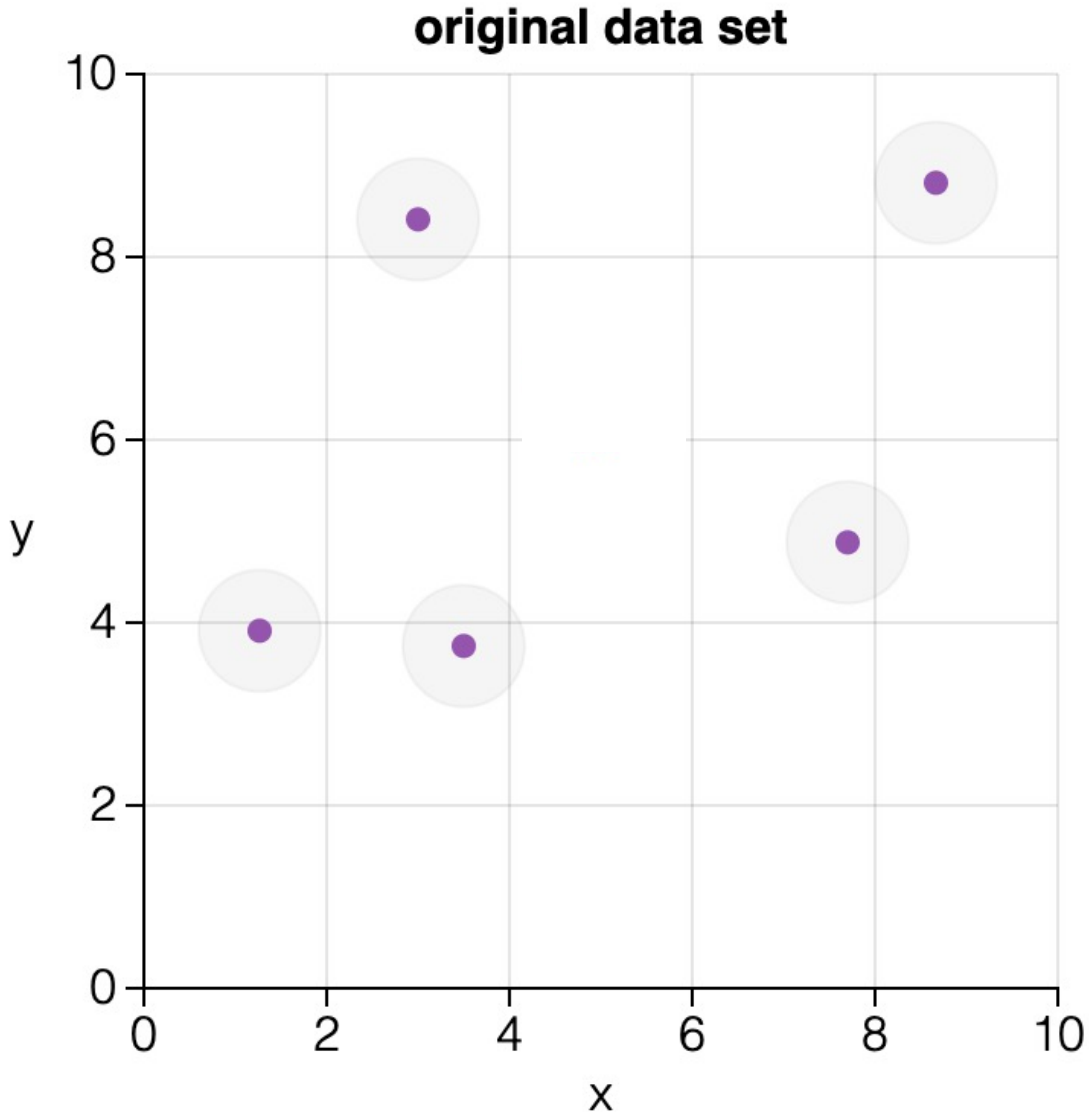
https://www.youtube.com/watch?v=9DPiXrN2pEg

PCA example

PCA is a method that allows us visualize and identify patterns from datasets with multiple variables (i.e. genotype data).

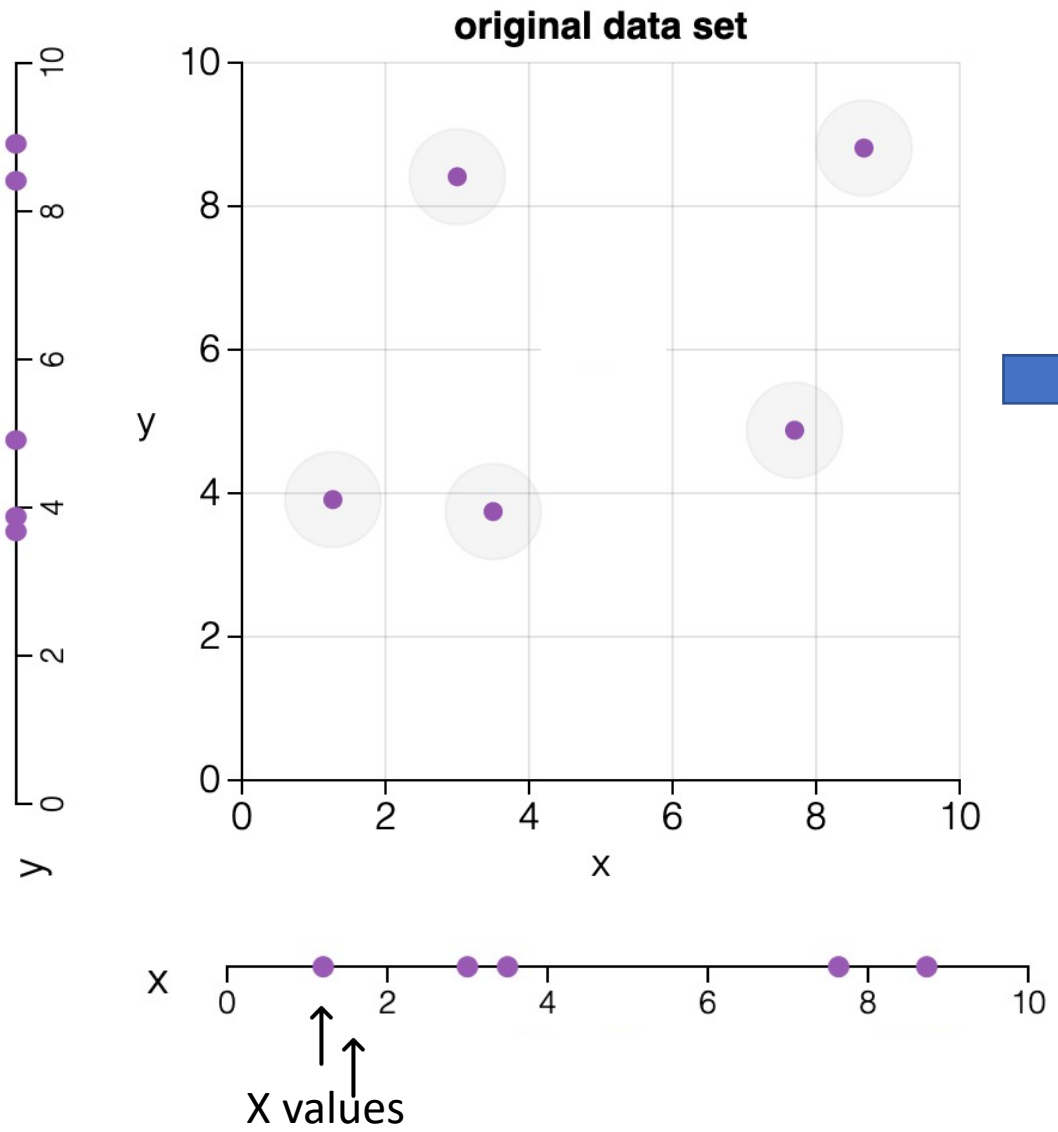
We can go from a K-dimensional variable space to a more manageable space, for example 2-dimensional.

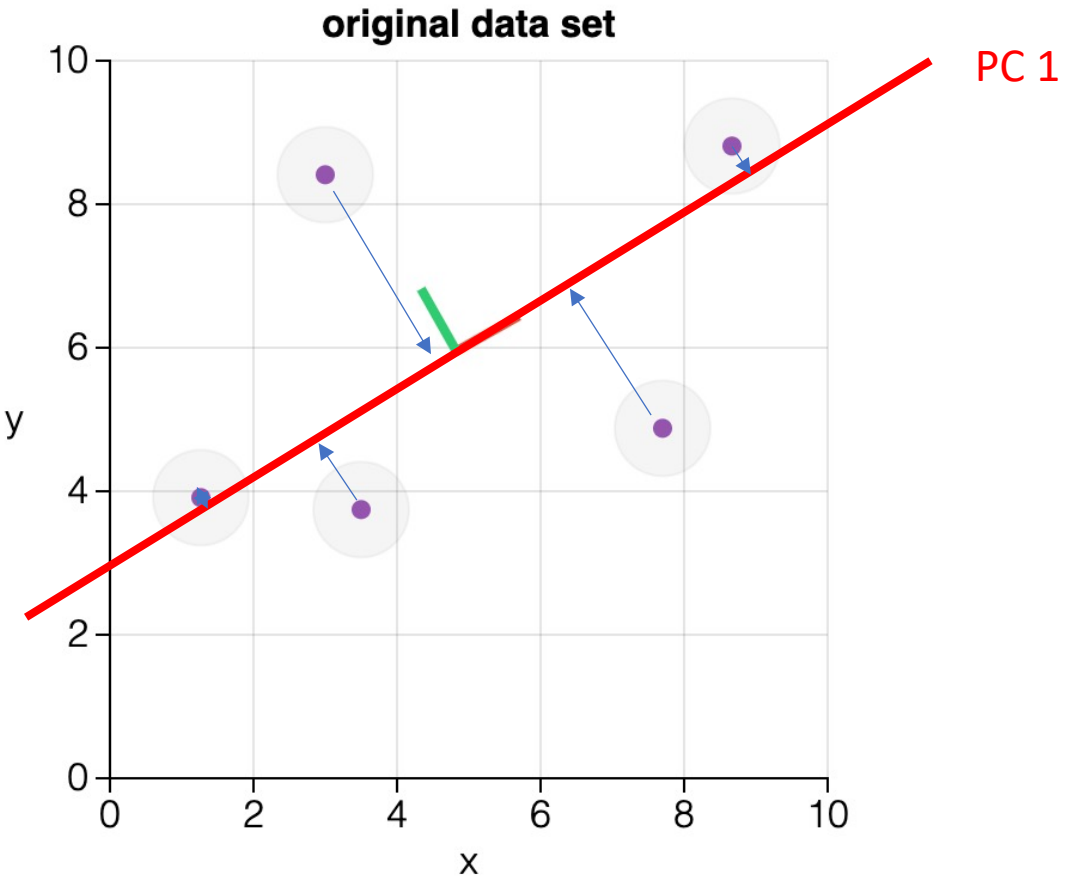
In this example we have 2-dimensional data. For example, we could have a dataset were each individual has recorded its BMI and height.
We want to go from 2-dimensions to 1-dimension.



Example from: https://setosa.io/ev/principal-component-analysis/

PCA example

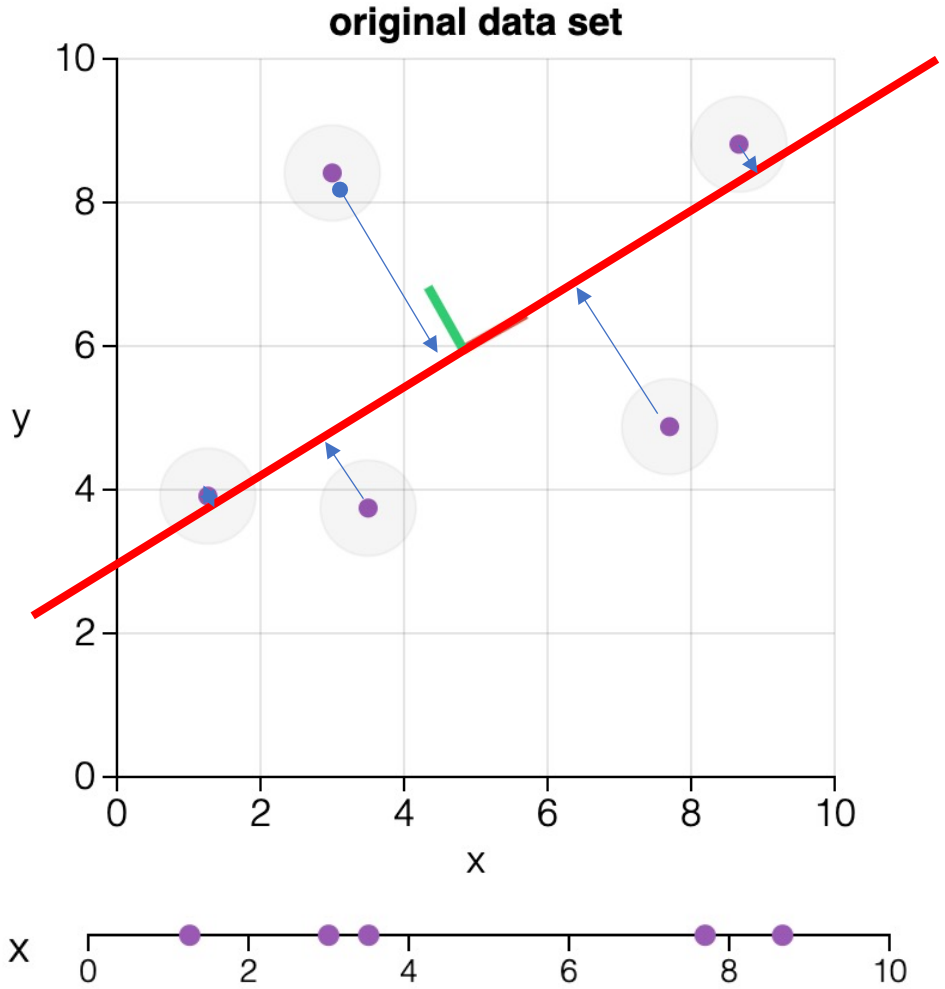




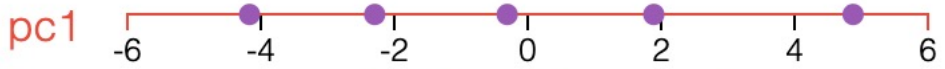
We find the first PC by finding the line that captures the most variance. This is done through using linear regression.

Example from: https://setosa.io/ev/principal-component-analysis/

PCA example



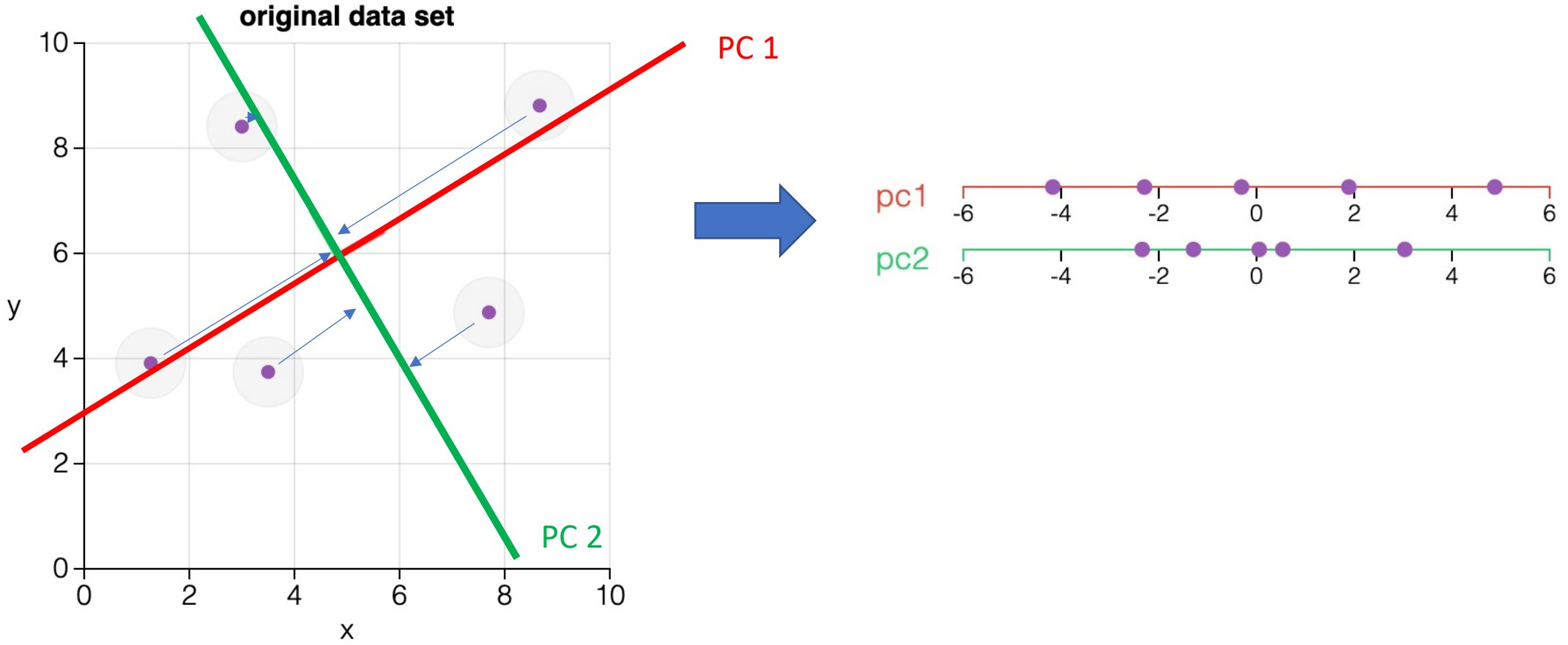
PC 1



The principal component 1 is a combination of height and weight that is optimally chosen to capture the most variation.

Example from: https://setosa.io/ev/principal-component-analysis/

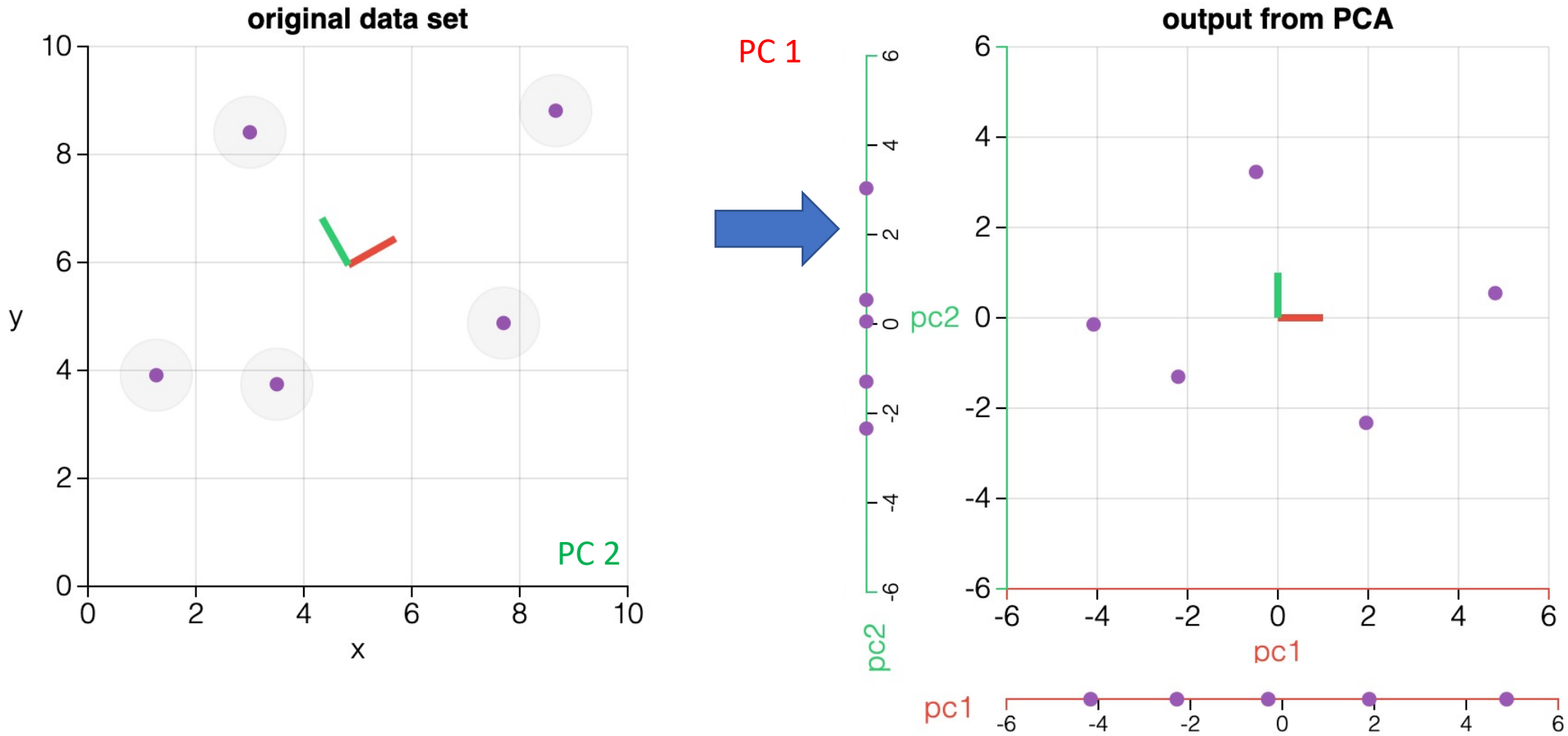
PCA example



Visit this webpage to play around with this example. The webpage is interactive!

Example from: https://setosa.io/ev/principal-component-analysis/

PCA example



Visit this webpage to play around with this example. The webpage is interactive!

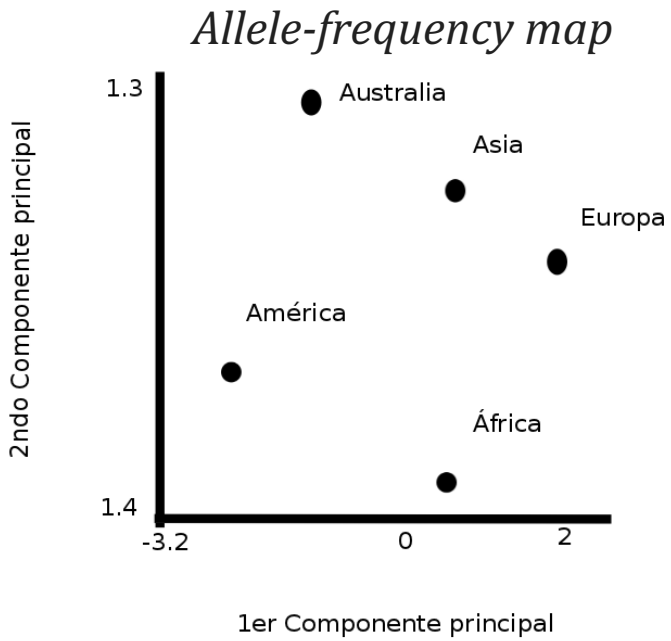
Example from: https://setosa.io/ev/principal-component-analysis/

Principal component analysis

	Population				
Gene	Africa	Asia	Europa	America	Australia
RH*D	20	15	36	2	0
ABO*O	69	60	65	90	76
FY*A	11	60	42	70	99
KM	34	19	8	35	30
DI*A	0	2	0	9	0

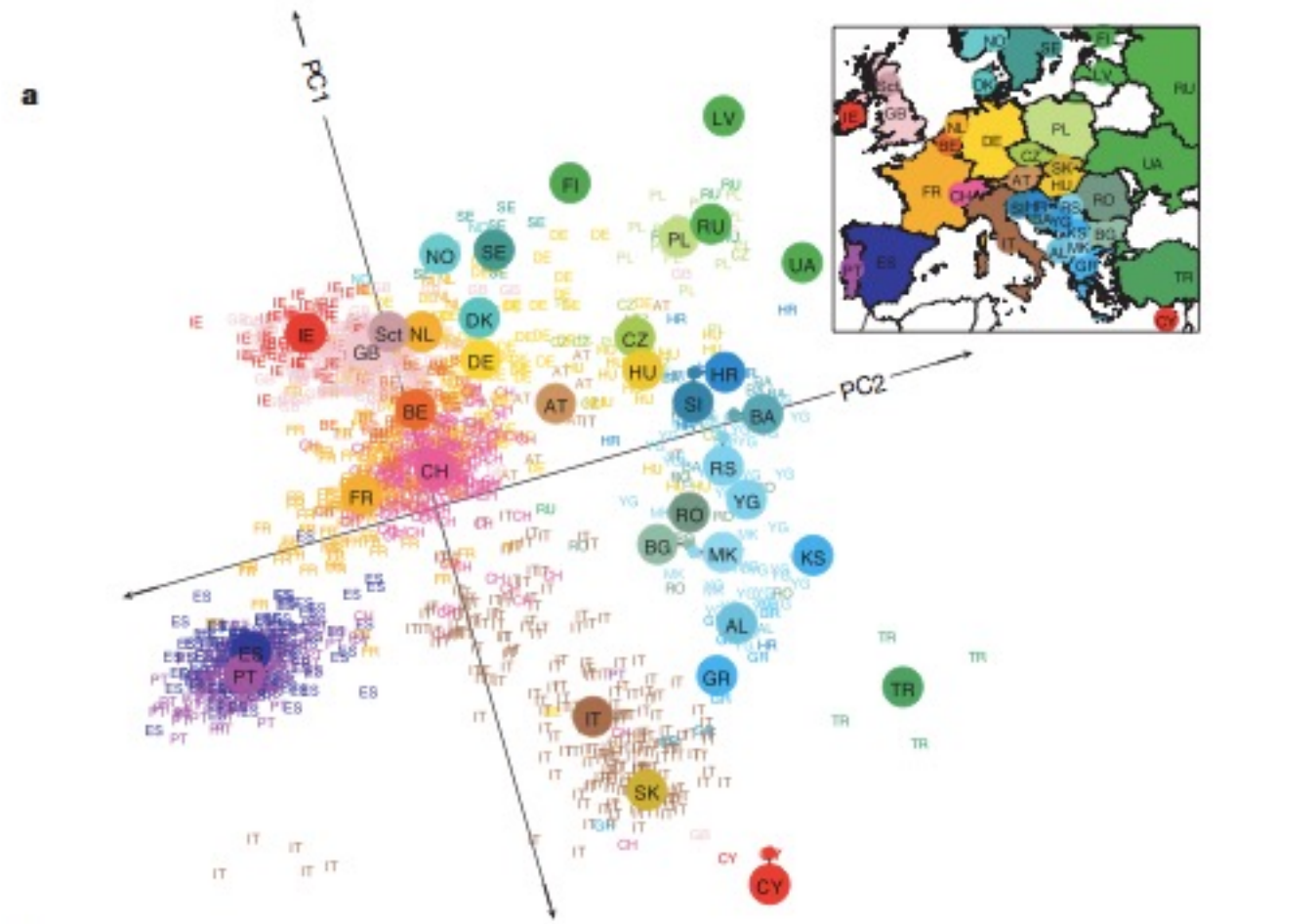
Principal components	Eigenvalues	% de variance	% de cumulative variance
1	3.195	63.9	63.9
2	1.365	27.3	91.2
3	0.276	5.5	96.7
4	0.164	3.3	100
5	0	0	100
Total	5		

Idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.



Cavalli-Sforza, Menozzi & Piazza. The History and Geography of Human Genes. 1994

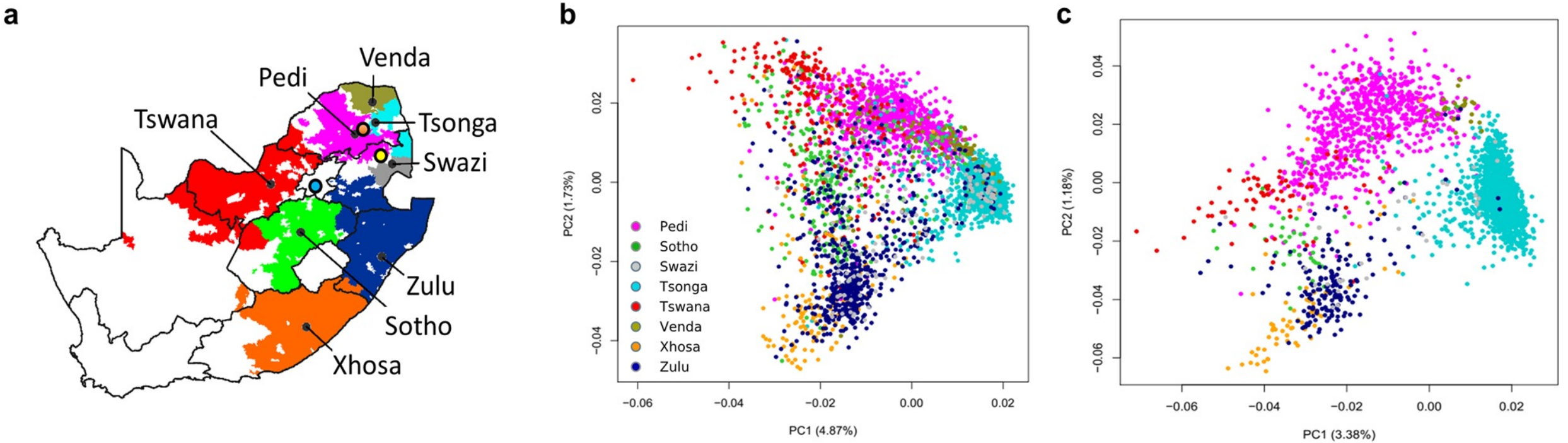
Principal component analysis can model population structure using only genetic data



Genes mirror geography within Europe

Novembre et al Nature 2008

Population structure and genetic affinities of South-Eastern Bantu-speaking (SEB) groups from South Africa correspond to both linguistic phylogeny and geographic distribution.



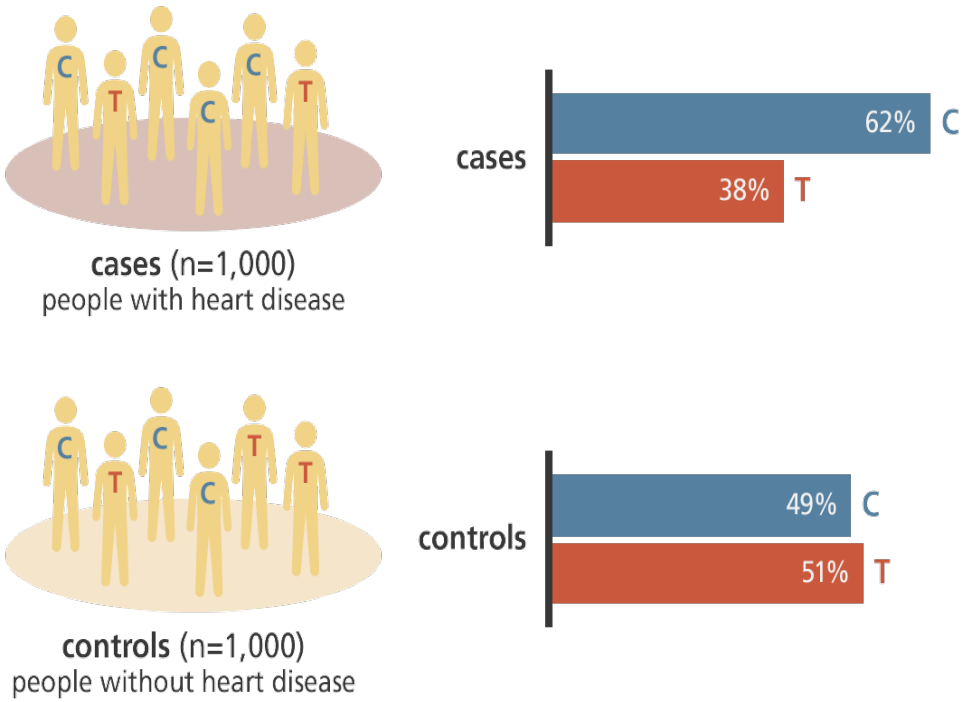
Sengupta, et al, Nat Comm 2021

*Study from University of the Witwatersrand, Johannesburg, South Africa

Association analysis

Association analysis

- Null Hypothesis: There is no difference in the frequency of allele C ($p_{controls}$) among diseased and the frequency of allele C among non-diseased.
  - H_0: $p_{cases}=p_{controls}$
- Alternative Hypothesis: We have enough evidence to conclude that the frequency of allele C ($p_{controls}$) among diseased is different than the frequency of allele C among non-diseased.
  - H1: $p_{cases} \neq p_{controls}$



Different types of association tests depending on the type of outcome

- Binary traits/disease status (i.e. schizophrenia)
  - Allele test (2x2 table - Chi-square test)
  - Armitage trend test (3x2 table - Chi-square test).
  - Logistic regression – Covariate adjustment
- Continuos traits (i.e. height).
  - Linear regression - Covariate adjustment

- Mixed Models. Complex statistical method that can model cryptic relatedness.

Multiple testing problem

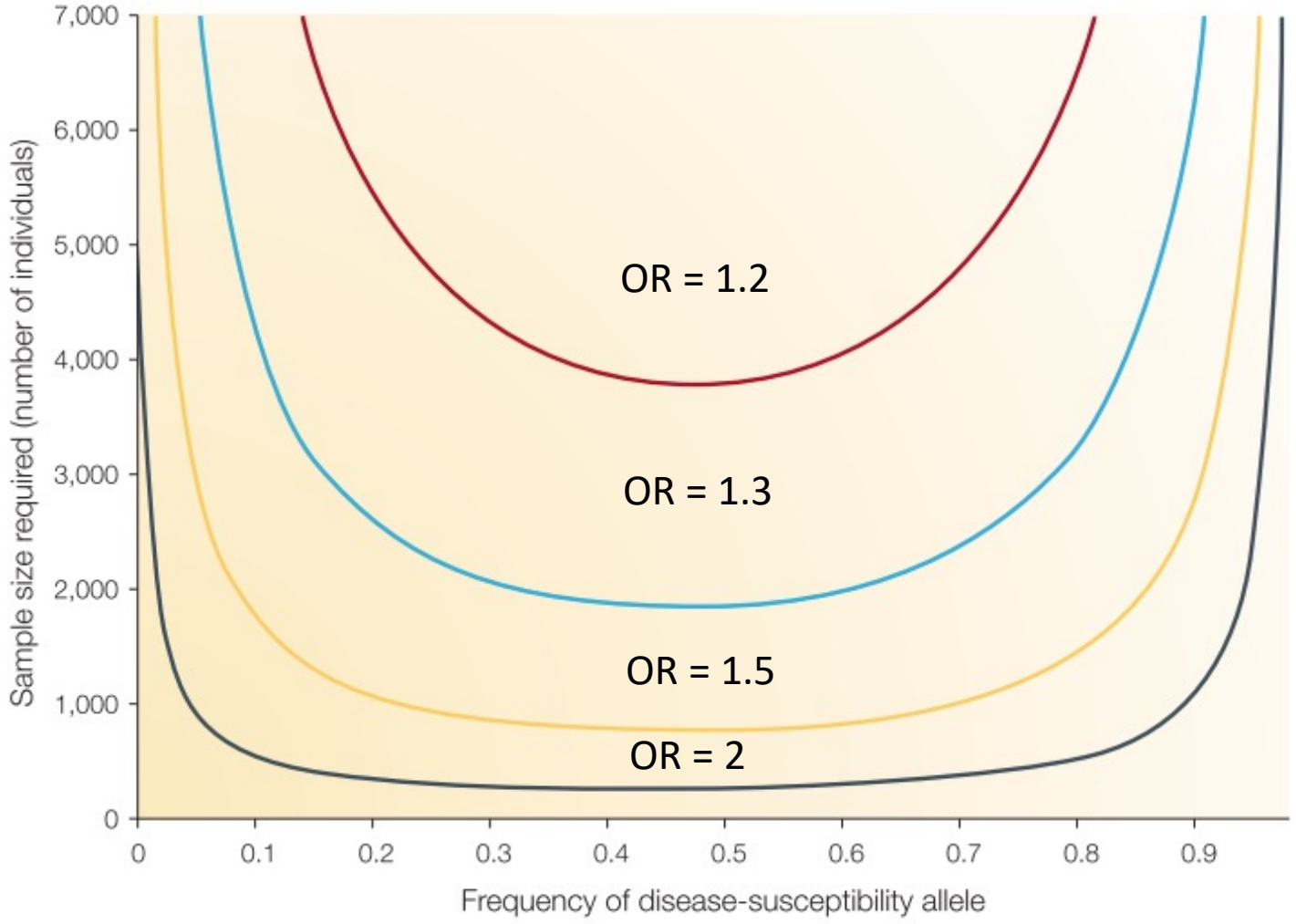
- To avoid false positives, we apply the Bonferroni correction:
  - We divide the standard significance level (0.05) by the number of independent comparisons we are making.
- The significance level in GWAs is: 5x10-8.
  - Estimated considering 1 million independent tests.

Pe'er et al 2008 Genet Epidem

Determining factors in the power to detect a causal variant.

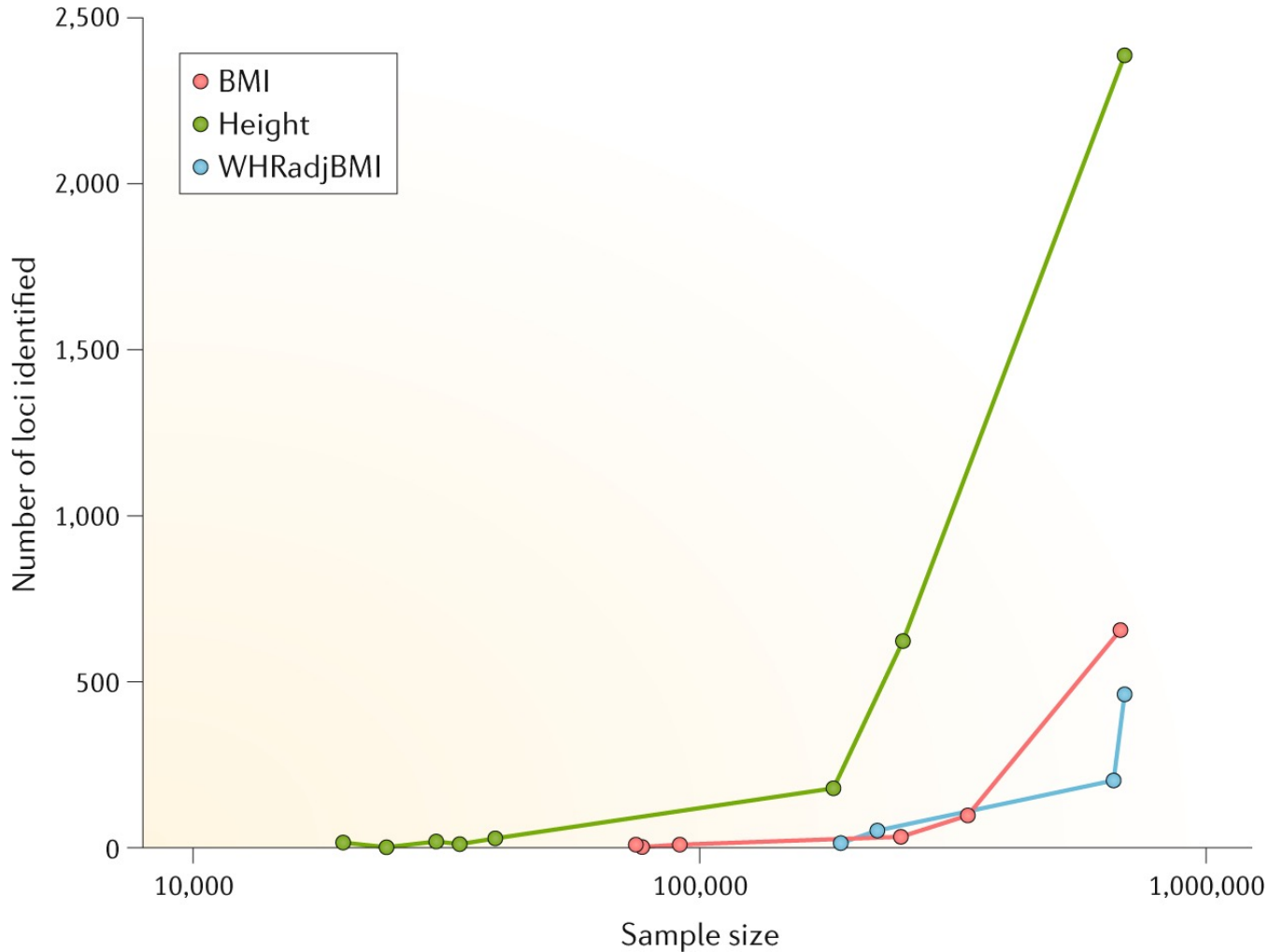
- Factors that we can not control:
  - Effect size (OR)
  - Allele frequency.
- Factors that we can control:
  - Sample size
  - Technology (chips with better coverage)

Relationship between sample size, effect size and allele frequency



Want et al Nat Rev 2005

Relationship between sample size and the number of independent loci discovered for BMI, height and WHRadjBMI

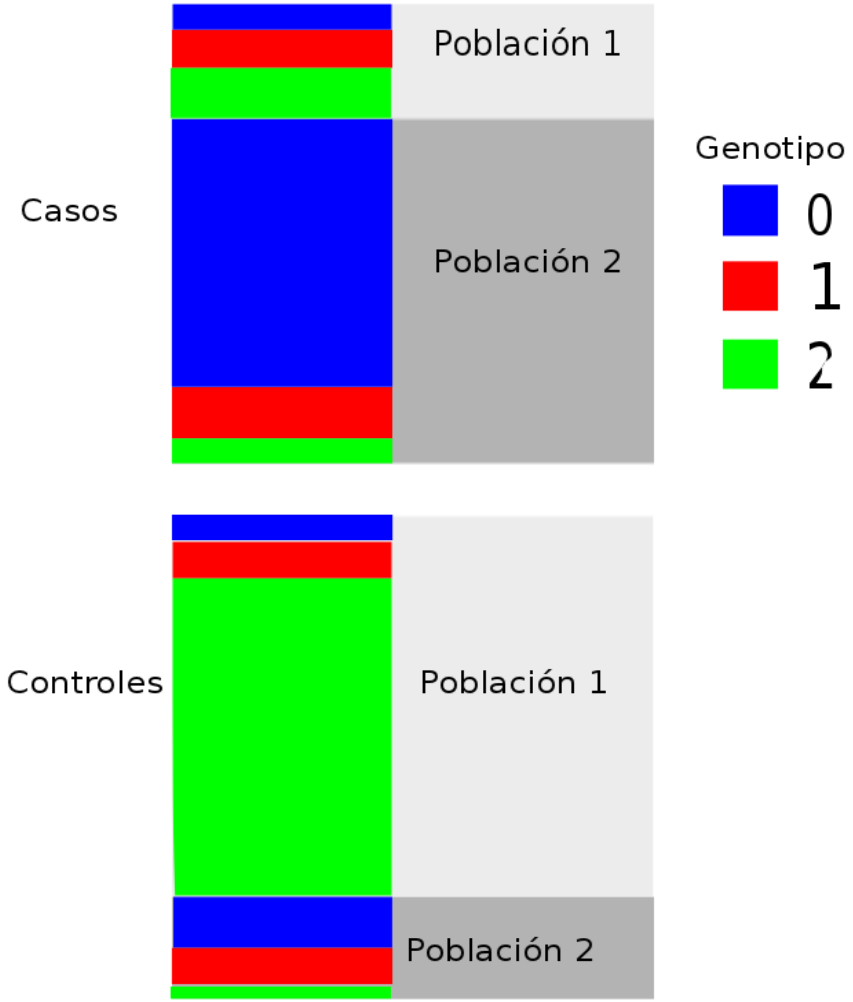


Tam et al Nat Rev Gen 2019

Potential causes for an association

- Causal variant
- Linkage disequilibrium
- Artifact due to poor quality control
- Population stratification

Population stratification



- Population stratification arises inclusion of individuals from different populations.
- We deal with it by adjusting for principal components in our association analysis or using mixed models.
- In the figure, genotype 0 is in a higher proportion in the cases than in the controls, but this is because it is more frequent in population 2 than in population 1. And there is a smaller number of individuals from population 2 in the controls.

Data visualization

- Q-Q plot
- Manhattan plot
- Locus Zoom plots

Q-Q plots

Quantile–quantile plot showing distribution of expected *P* values under a **null model** of no significance versus observed *P* values.

Expected $-\log_{10}$ transformed *P* values (*x* axis) for each association are plotted against observed values (*y* axis) to visualize the enrichment of association signal.

Deviation from the expectation under the null hypothesis (red line) indicates:

- the presence of either true causal effects (we will see it mostly at the end)
- insufficiently corrected population stratification (constant across all data).

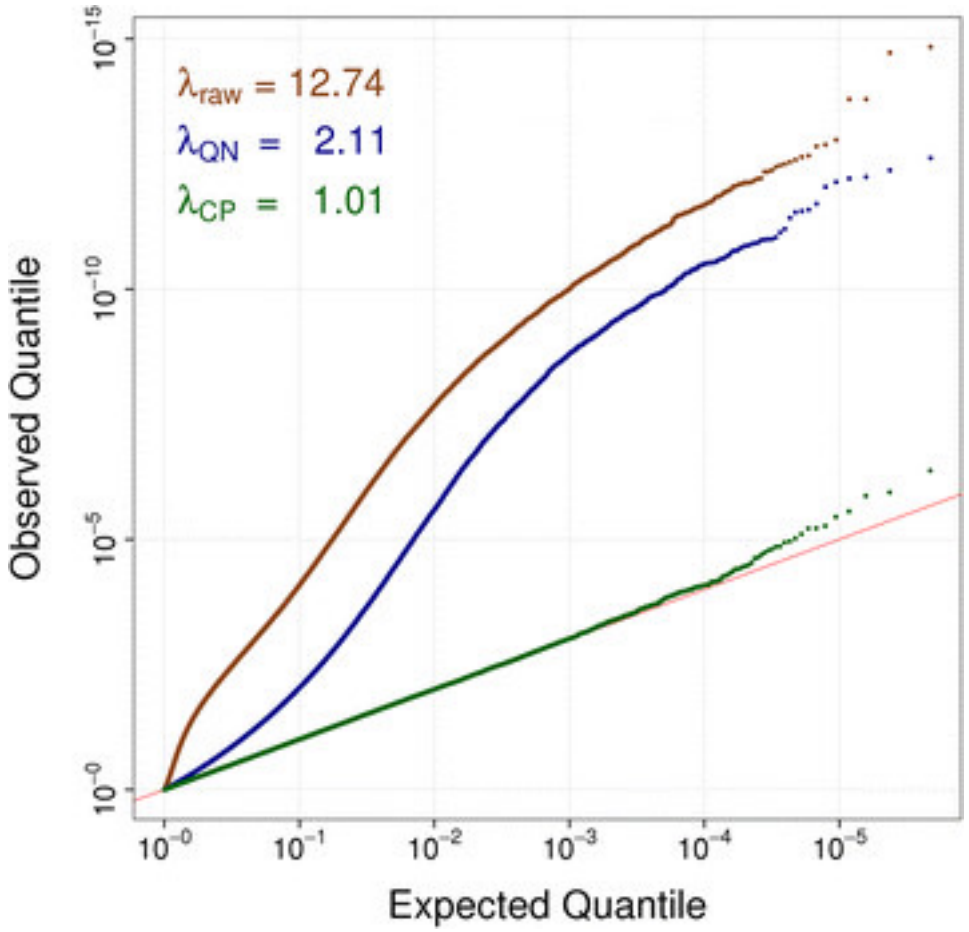


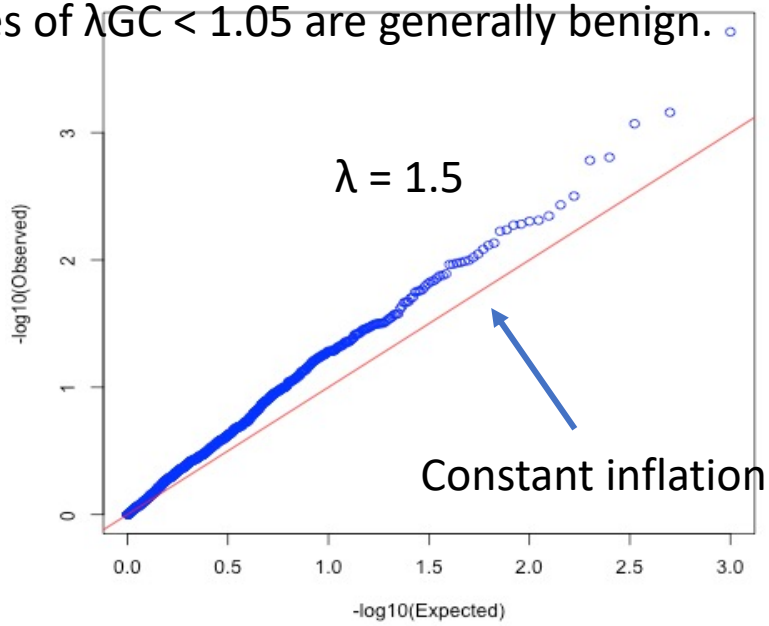
Imagen: Lehne et al 2015 Genome Biology

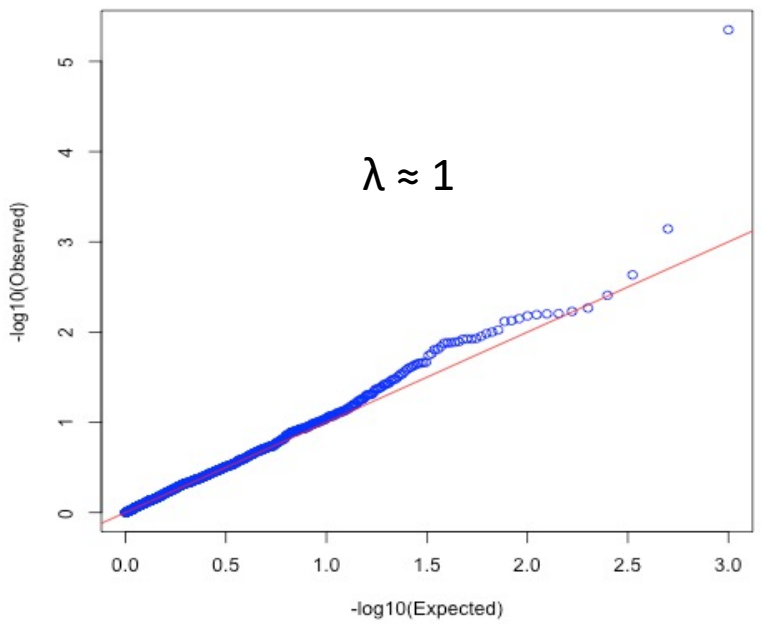
Genomic control

Objective: Correct inflation caused by population stratification and other confounding variables.
It is defined as

$$\lambda = \frac{\text{mediana}(\chi_1^2, \chi_2^2, \ldots, \chi_L^2)}{0.456}$$

Values of λ > 1 indicate the presence of population stratification and other confounding factors.
Values of λGC < 1.05 are generally benign.



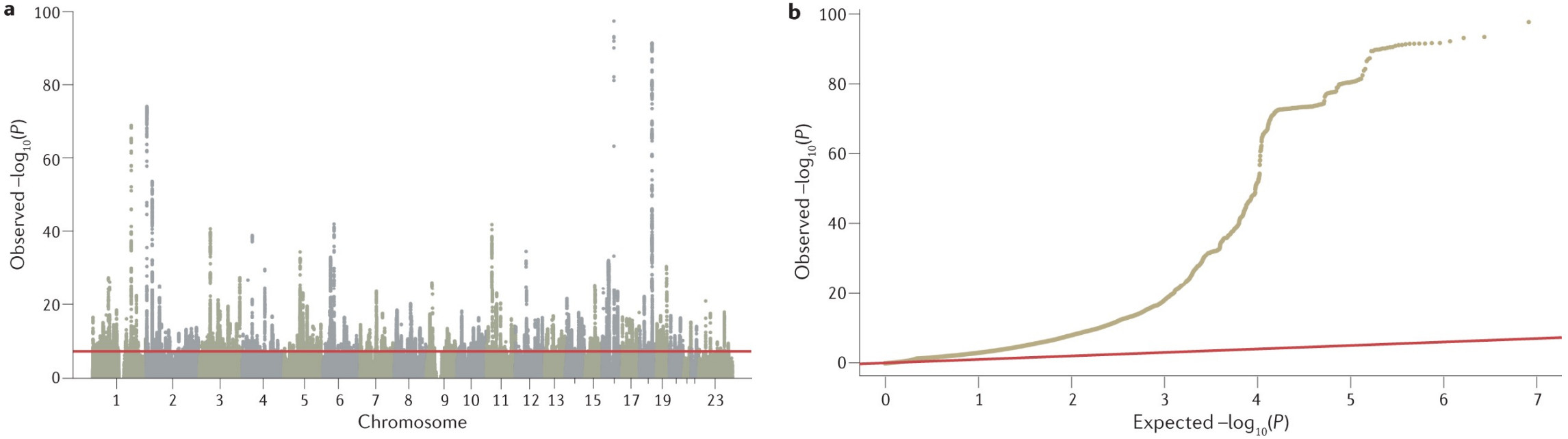


Devlin & Roeder, Biometrics 1999

What happens when we have a highly polygenic trait

BMI is extremely polygenic and the genome-wide association study (GWAS) was highly powered

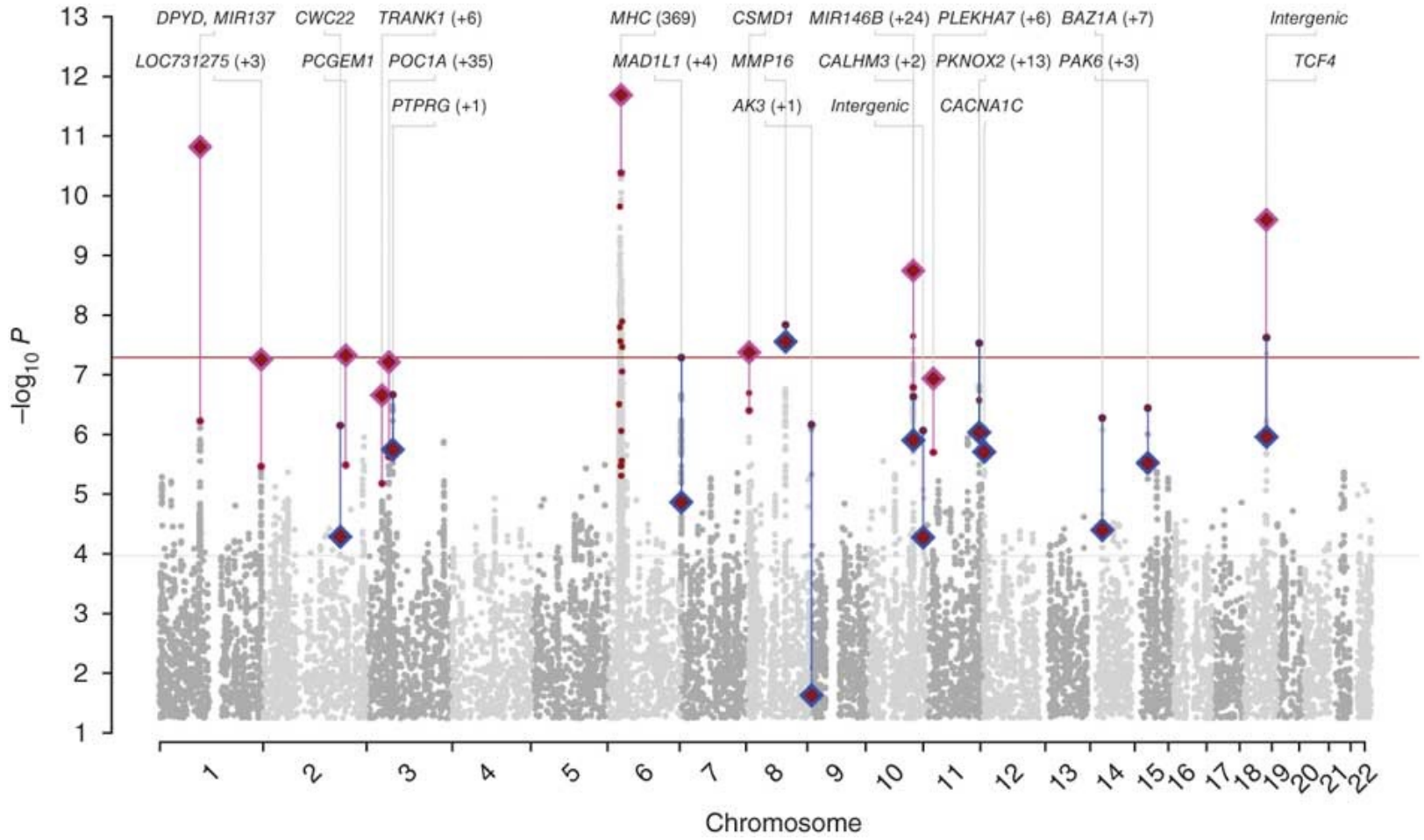
In this situation it because difficult to asses presence of bias vs polygenicity using genomic control.



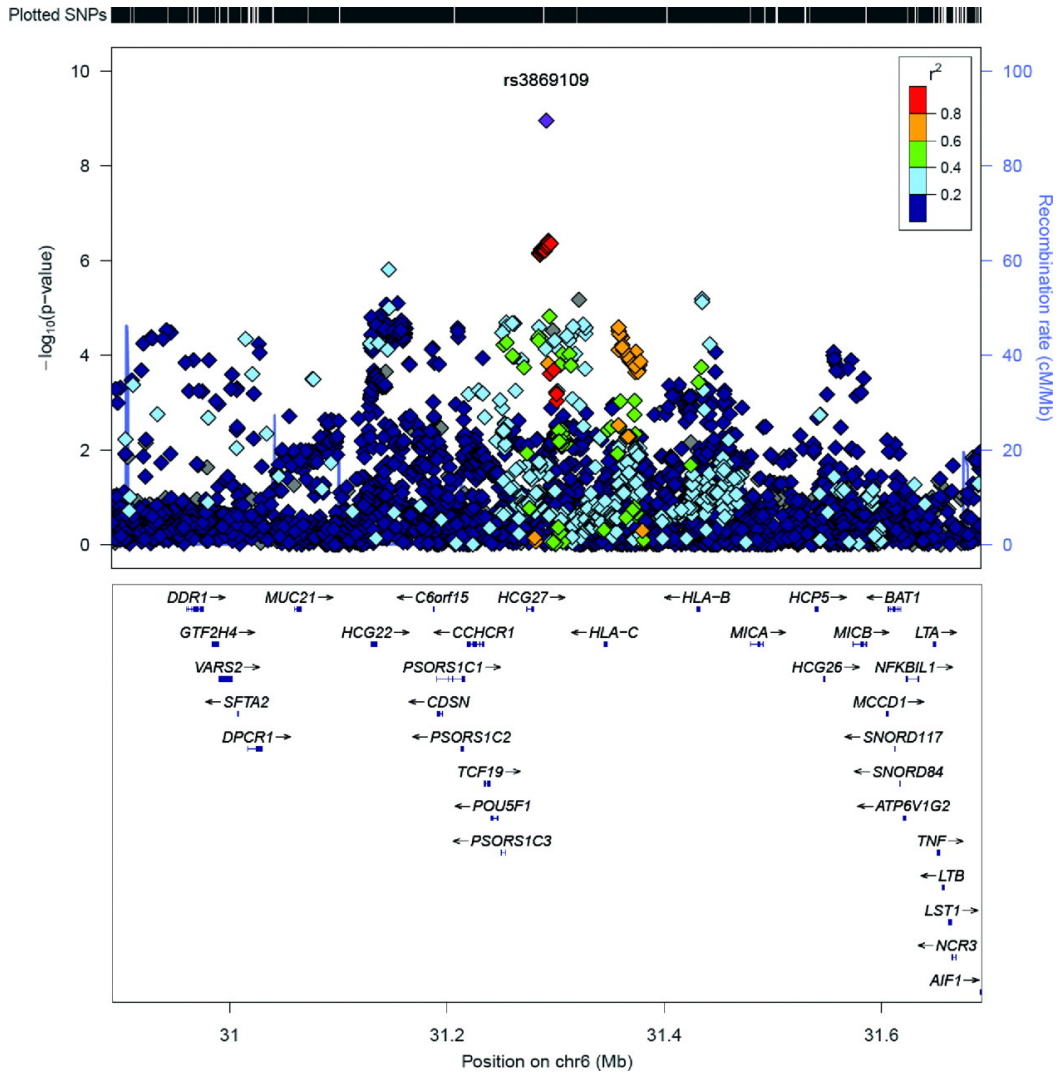
Solution: LD-score (LDSC) regression

Bulik-Sulivan et al Nat Genet 2015
Imagen: Uffelman et al, NatRev Methods Primers, 2021

Manhattan plots



Locus Zoom plots



Let’s keep playing with our data!