



# Genome-wide Association Study

3MR103 (2024)

Mark Jen-Hsiang Ou

PhD Student

Institutionen för medicinsk biokemi och mikrobiologi

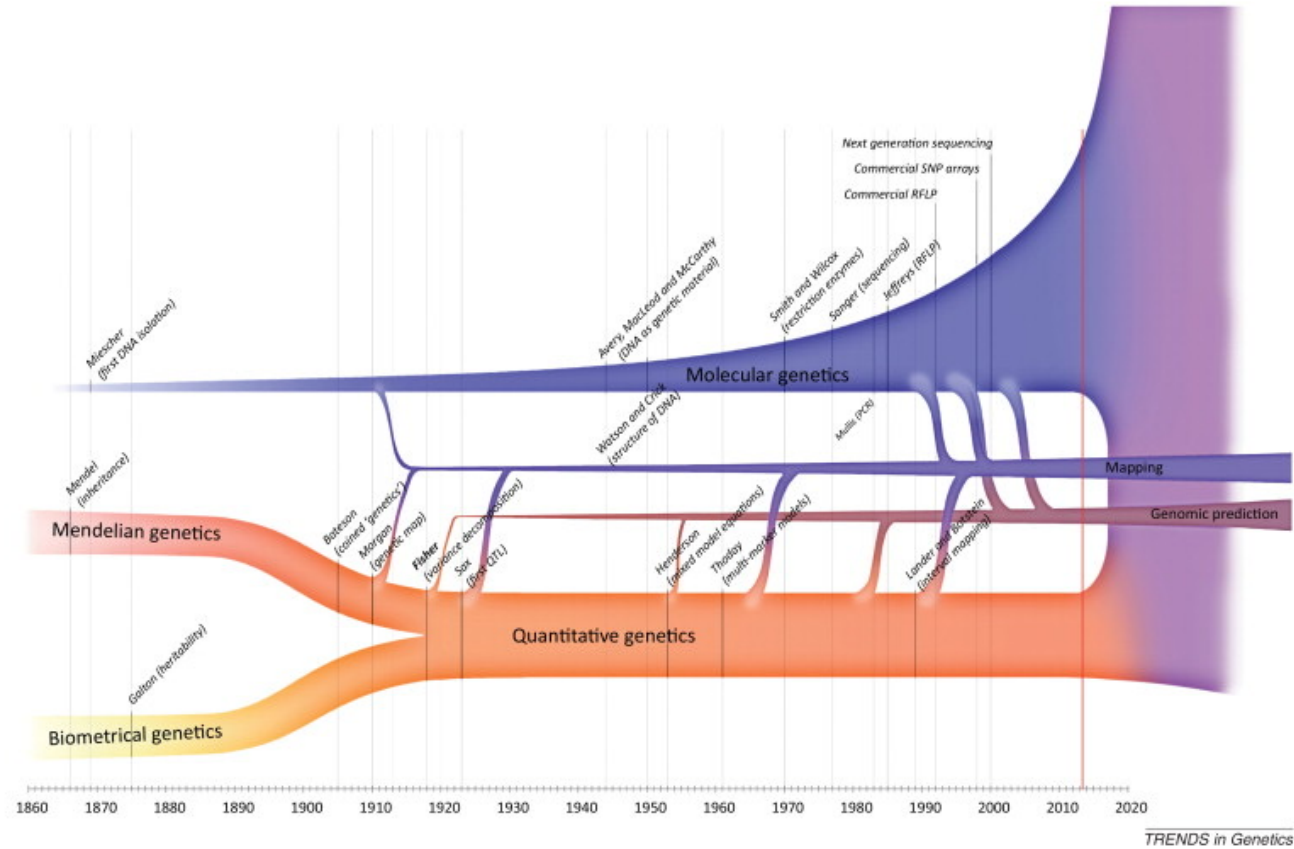
Uppsala Universitet



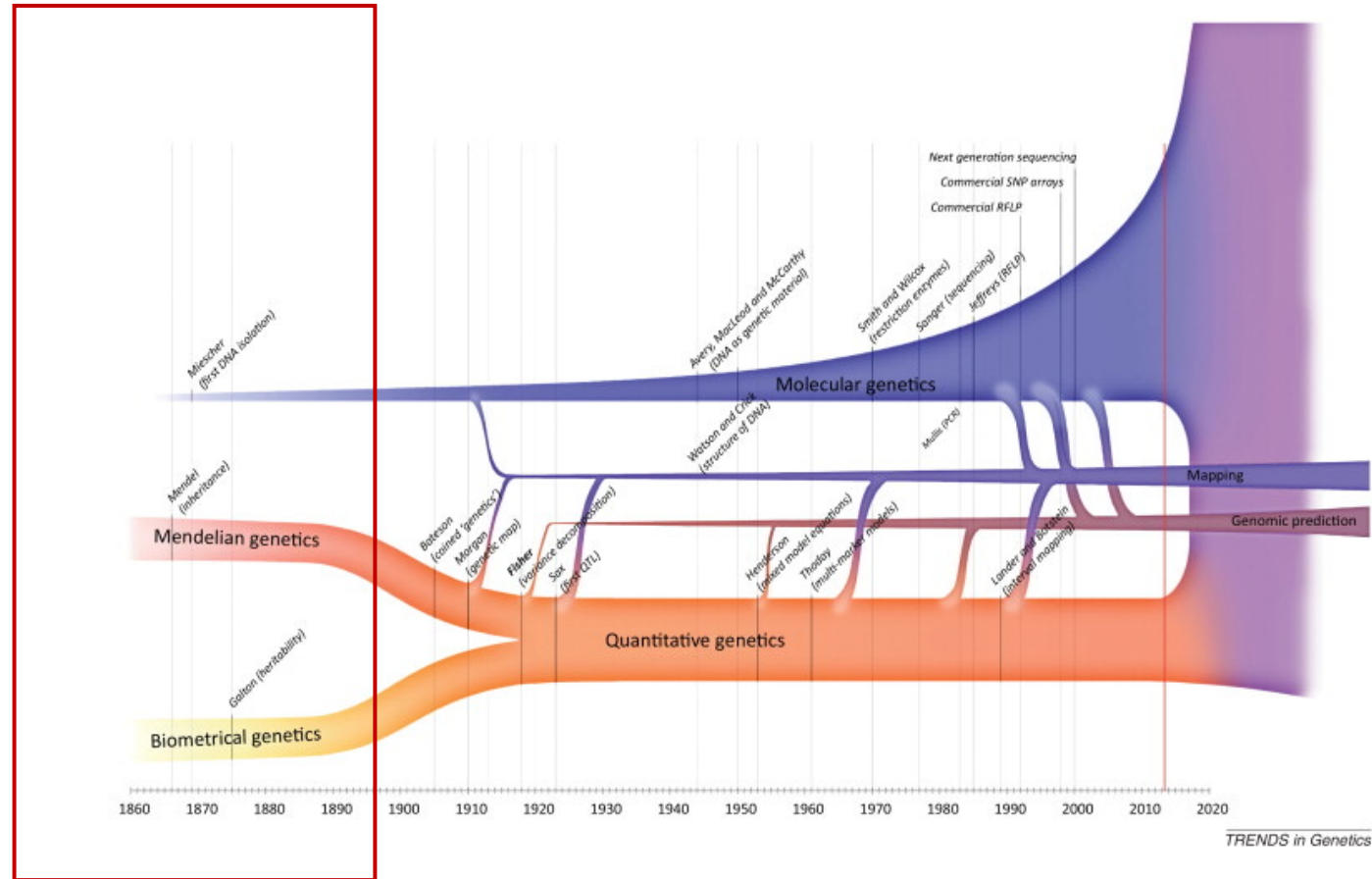
# How did it evolve?

Introduction to quantitative genetics

# Quantitative genetics. How did it evolve?



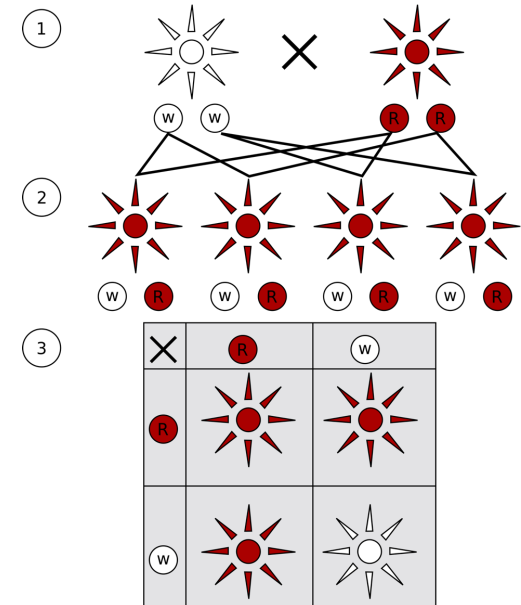
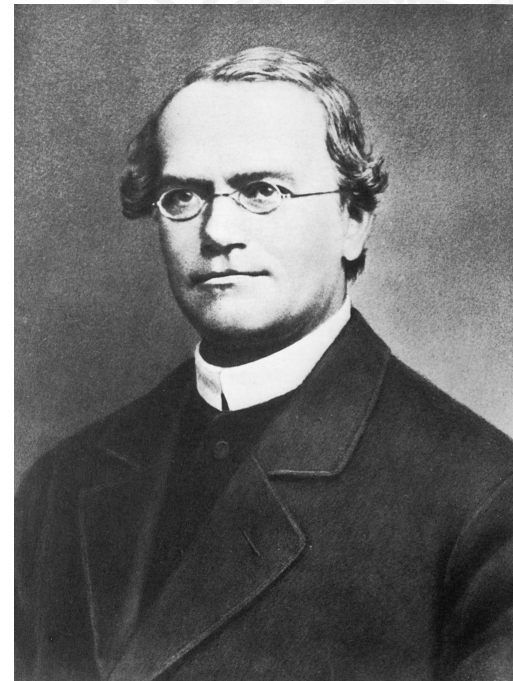
# Quantitative genetics. How did it evolve?



Early stage. Before the field is known as genetics.

# Gregor Mendel (1822-1884)

- Study variation in plants in his monastery's 2 hectares experimental garden.
- Studying seven traits that seemed to be inherited independently of other traits
  - seed shape, flower color, seed coat tint, pod shape, unripe pod color, flower location, and plant height
- Defined “recessive” and “dominant” traits based on their segregation in crosses.
- Traits can be predictably determined by “invisible factors”
  - Now know called gene
  - Genetic variation at DNA, protein, or phenotypic level can be found to follow Mendel’s Laws of Segregation

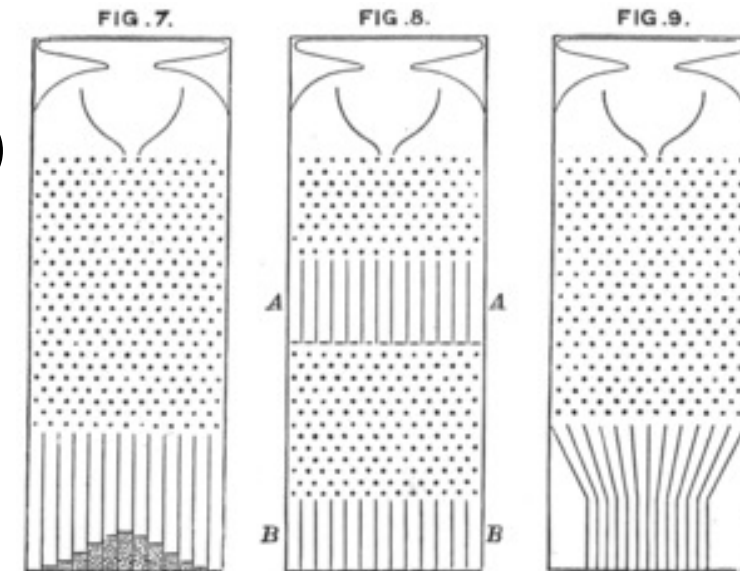
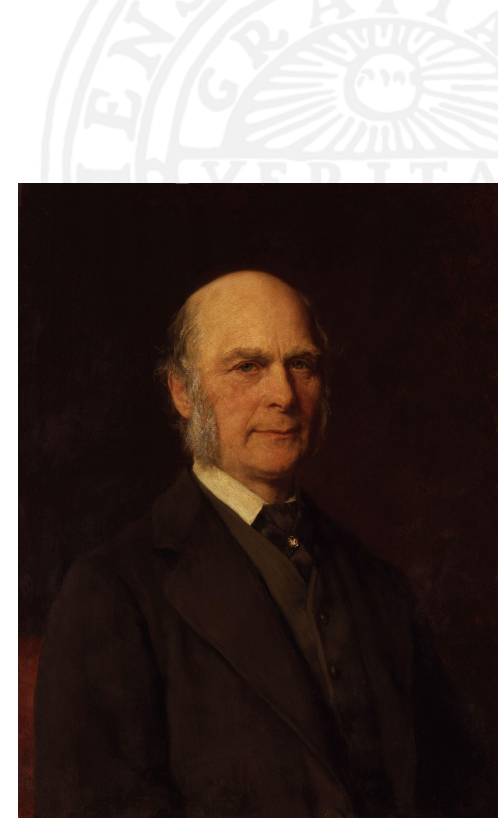


# Sir Francis Galton (1822-1911)

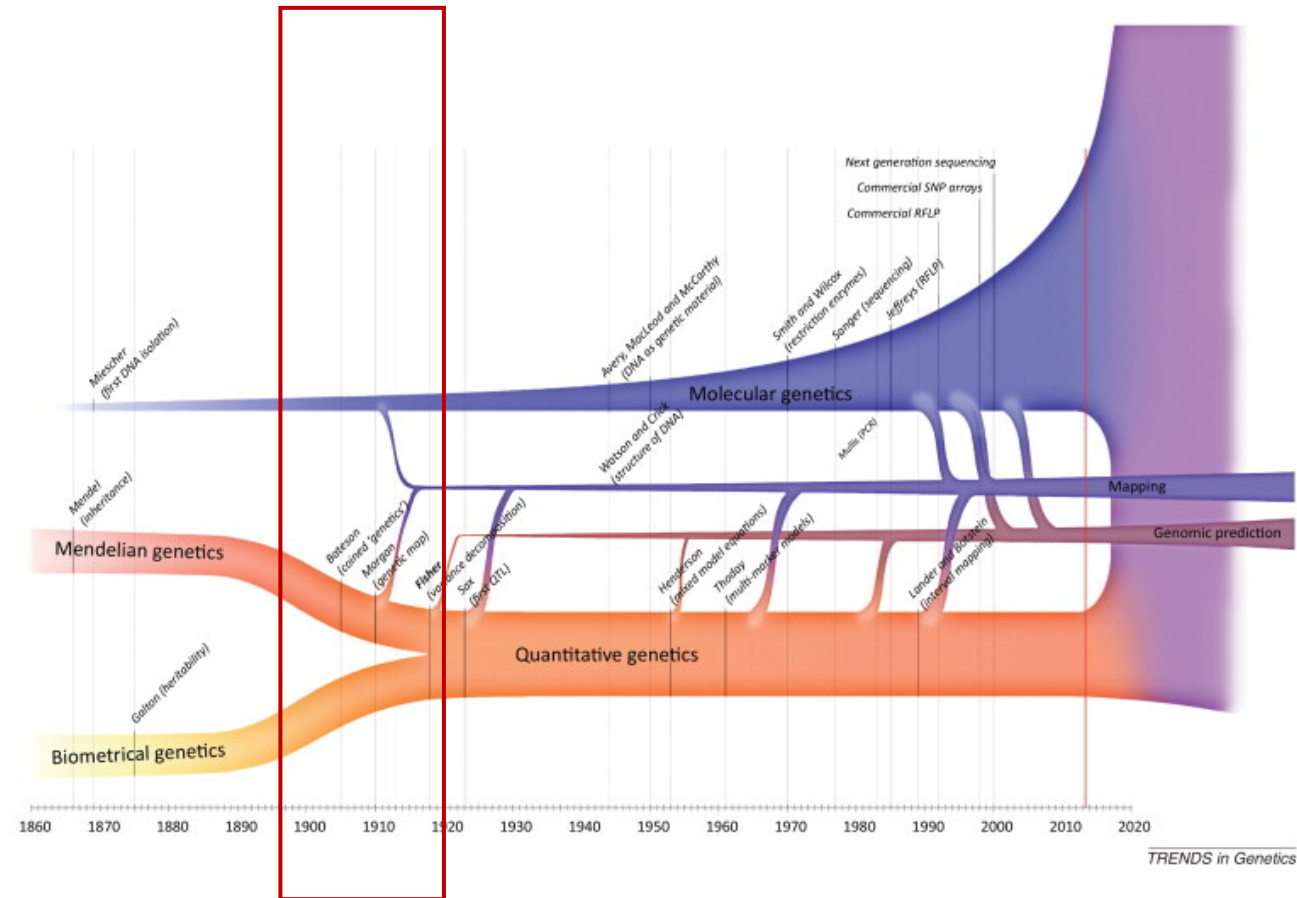
- Are traits hereditary?
  - Population studies with the hypothesis:  
“If traits are hereditary, relatives should be more similar to non-relatives.”
- Apply statistical methods to study human differences, intelligence inheritance, and biological data.
  - Pioneer of eugenics
  - His book Hereditary Genius (1869) was the first social scientific attempt to study genius and greatness.
  - Variance (SD) to qualify the normal variation
  - Experimental demonstration of normal distribution (bean machine)
  - Regression line and ‘ $r$ ’ to represent the regression coefficient  
=> parent-offspring correlation to estimate the heritability
  - Regression toward the mean

How to make a perfect cup of tea:

<https://galton.org/books/art-of-travel/galton-1855-art-travel-1st-ocr.pdf>



# Quantitative genetics. How did it evolve?



Quantitative traits <-> Mendelian traits

# William Bateson (1861-1926)

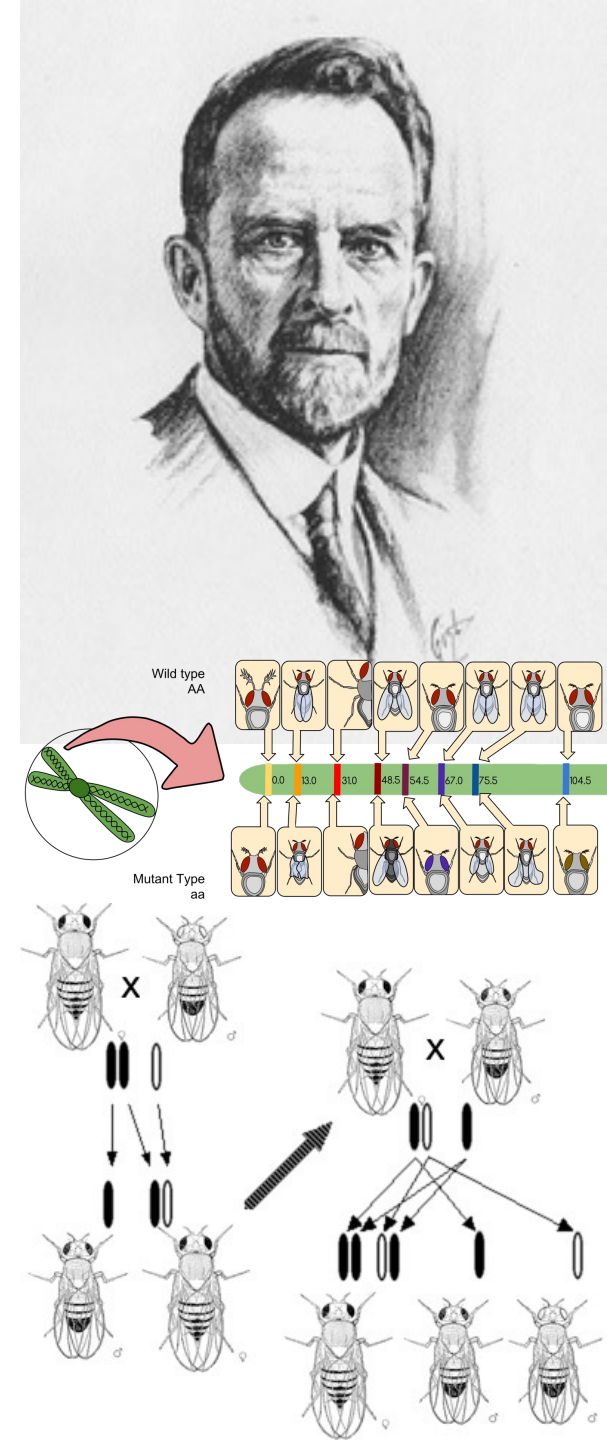
- Coined the term “genetics” to describe the study of heredity and “epistasis” to describe the genetic interactions
- Embraced both discontinuous and continuous traits
- Unaware of Mendel’s argument for making large crosses to study the inheritance of discontinuous traits. Later in his career replicated Mendel’s work
- The Mendelian antagonist of the biometrical school of thinking



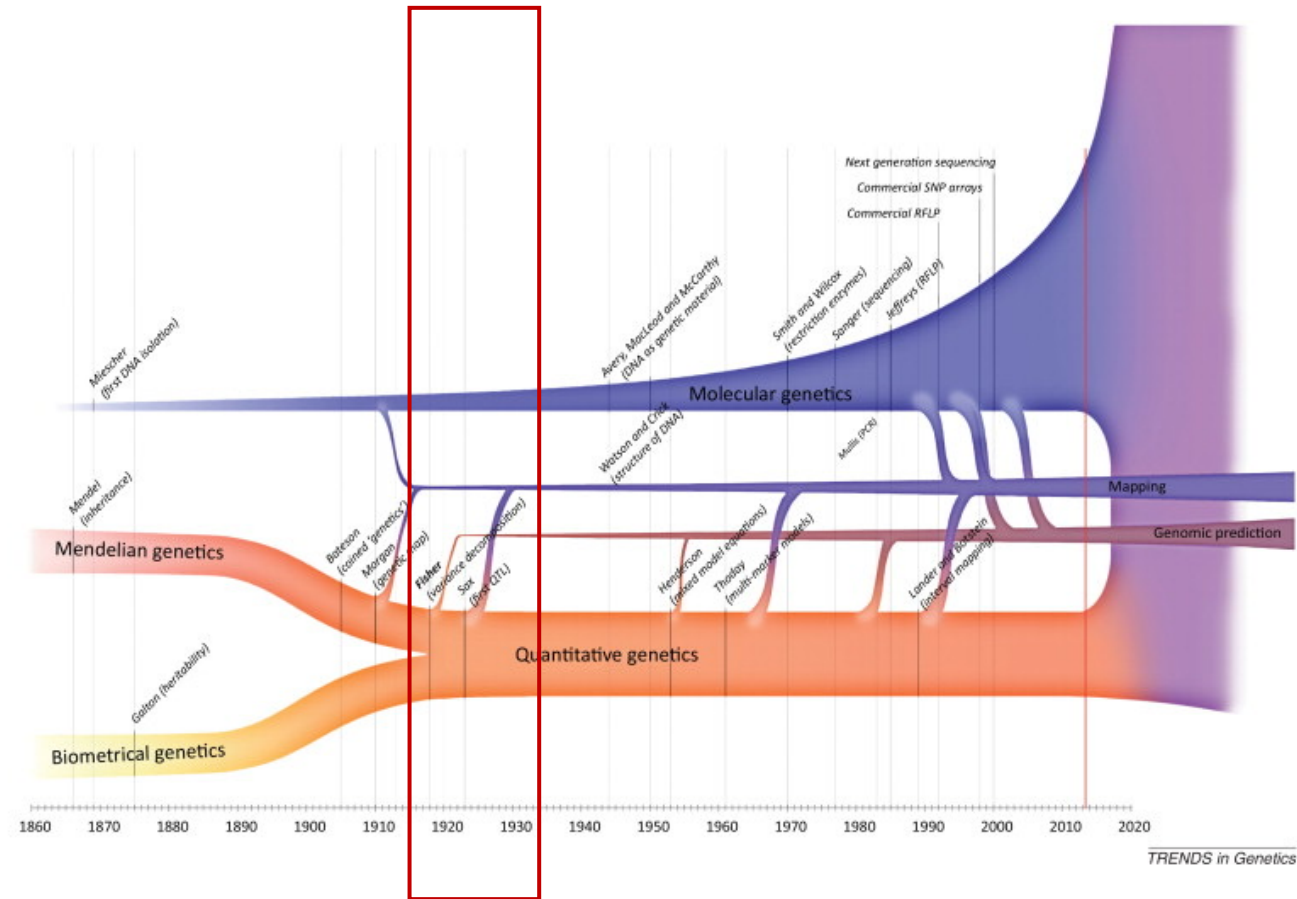


# Thomas Hunt Morgan (1866-1945)

- Large-scale experimental Mendelian genetics
  - Fruit fly (*D. Melanogaster*) to screen for mutants & the mechanical basis of heredity
  - Traits can be:
    - Sex-linked and by genes on sex chromosomes
    - Non-sex linked and by genes on other chromosomes
- Genes linked on chromosome
  - => Crossover frequencies indicate the distance separating them
  - => First genetic map in 1913 (morgan = unit of measurement)



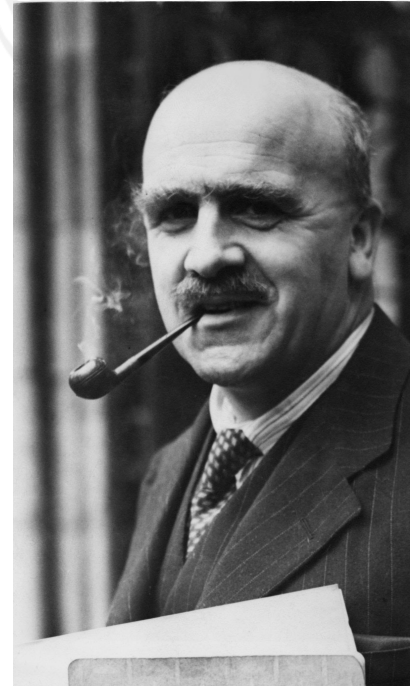
# Quantitative genetics. How did it evolve?





# JBS Haldane (1892-1964)

- Demonstration of genetic linkage in mammals (mice), chicken, human
- Develop statistical methods for human genetics
  - Using maximum likelihood for the estimation of human linkage maps
  - Linkage theory for polyploid
  - First estimates of mutation rate in humans
- Modern evolutionary synthesis
  - Natural selection is a central mechanism in evolution  
-> a mathematical consequence of Mendelian inheritance
  - Central mathematical theory in population genetics



# Ronald Aylmer Fisher (1890-1962)

- Statistician

- Analysis of variance (ANOVA)
- Fisher's z distribution ( $z = \frac{1}{2} \log F$ ), F-distribution
- Popularize the maximum likelihood
- 5% threshold for p-values

- Geneticist

- Model how continuous trait variation (Biometric) could result from many discrete genes (Mendelian)

Natural selection changes allele frequencies in the population

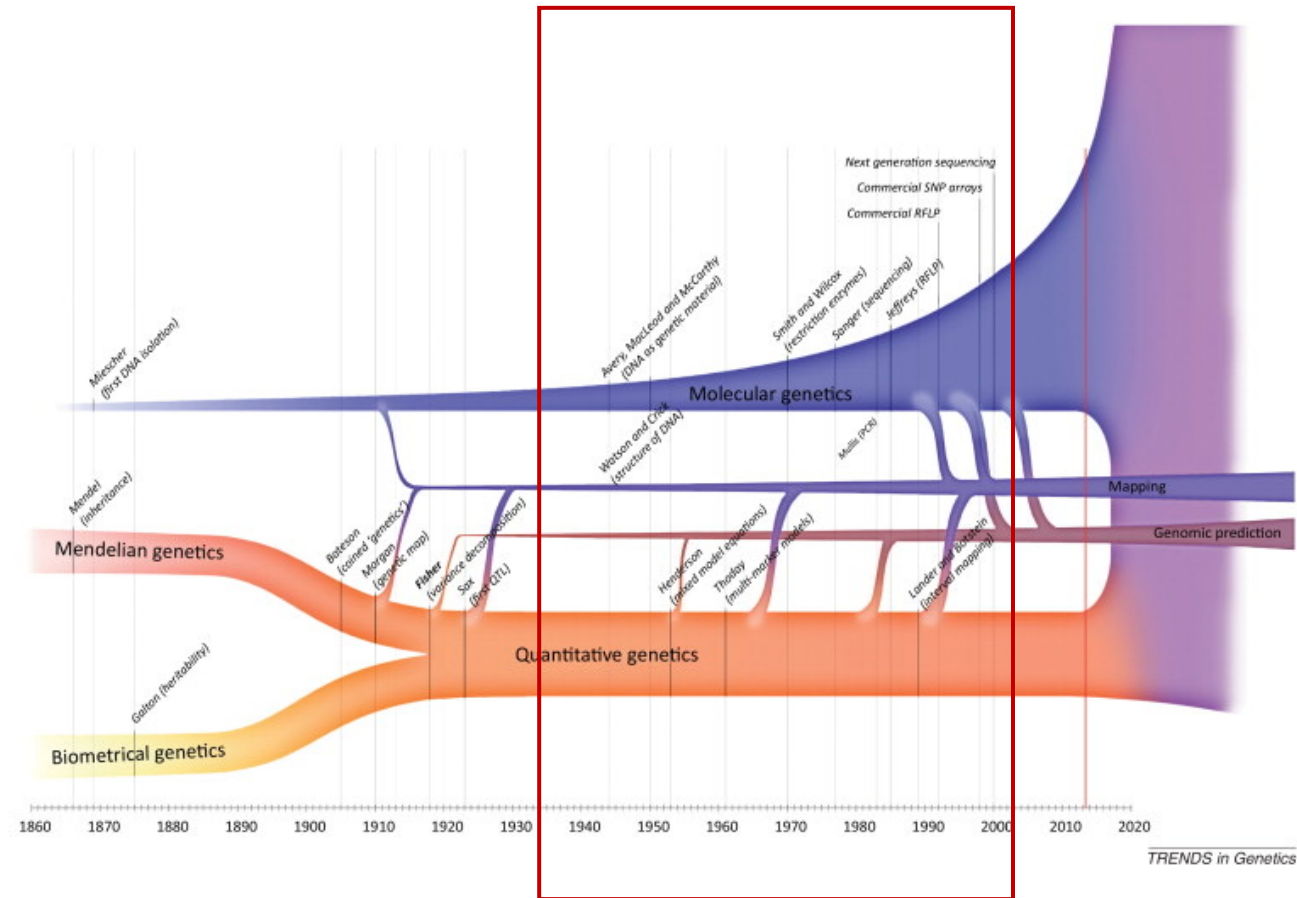
=> discontinuous Mendelian factors reconciled with gradual evolution



Known for

- [Fisher's exact test](#)
- [Fisher's inequality](#)
- [Fisher's principle](#)
- [Fisher's geometric model](#)
- [Fisher's Iris data set](#)
- [Fisher's linear discriminant](#)
- [Fisher's equation](#)
- [Fisher information](#)
- [Fisher's method](#)
- [Fisherian runaway](#)
- [Fisher's fundamental theorem of natural selection](#)
- [Fisher's noncentral hypergeometric distribution](#)
- [Fisher's z-distribution](#)
- [Fisher transformation](#)
- [Fisher consistency](#)
- [F-distribution](#)
- [F-test](#)
- [Fisher–Tippett distribution](#)
- [Fisher–Tippett–Gnedenko theorem](#)
- [Fisher–Yates shuffle](#)
- [Fisher–Race blood group system](#)
- [Behrens–Fisher problem](#)
- [Cornish–Fisher expansion](#)
- [von Mises–Fisher distribution](#)
- [family allowance](#)
- [Wright–Fisher model](#)
- [Ancillary statistic](#)
- [Fiducial inference](#)
- [Intraclass correlation](#)
- [Infinitesimal model](#)
- [Inverse probability](#)
- [Lady tasting tea](#)
- [Null hypothesis](#)
- [Maximum likelihood estimation](#)
- [Neutral theory of molecular evolution](#)
- [Particulate inheritance](#)
- [Random effects model](#)
- [Relative species abundance](#)
- [Reproductive value](#)
- [Sexy son hypothesis](#)
- [Sufficient statistic](#)
- [Analysis of variance](#)
- [Variance](#)

# Quantitative genetics. How did it evolve?





# Fundamental concept of quantitative genetics

Introduction to quantitative genetics



# The essence

- Focus on quantitative traits
- Statistics and phenotype centric
  - Variation in phenotype -> use statistic to quantify the genetic contribution to this trait variation
  - Framework provided via Modern Evolutionary Synthesis
- AIM
  - Decompose the genetic variance
  - Prediction of selection responses
- Mathematical central concept: heritability
  - Proportion of variation that is due to genetics
- Population based not individual!

# Multiple genes & environmental factors

$$\text{Binomial}(n, p) \xrightarrow{n \rightarrow \infty \text{ (CLT)}} \text{Normal}(np, npq)$$





Variation is everywhere





# The genetics of trait variation

- Many genes and environment factors -> continuous trait distribution
- Aim: Study contribution by genetics to this quantitative trait variation in populations using statistics
  - Decompose trait variation to genetic/environmental contributions
  - Predict selection responses in populations
    - $\Delta$  trait from  $\Delta$  allele frequency across many loci
    - Assume infinitesimal model: each gene minor effect and minor  $\Delta$  allele frequency during selection (remain constant or not)
- Key concept: additive genetic variance
  - Genetic variance transmittable from parents to offspring
  - Heritability: proportion of total trait variation that is inherited



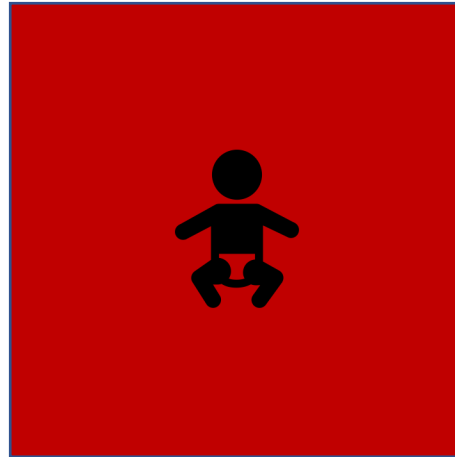
RA Fisher

# Decompose trait variation

Mostly environment



Mostly genes

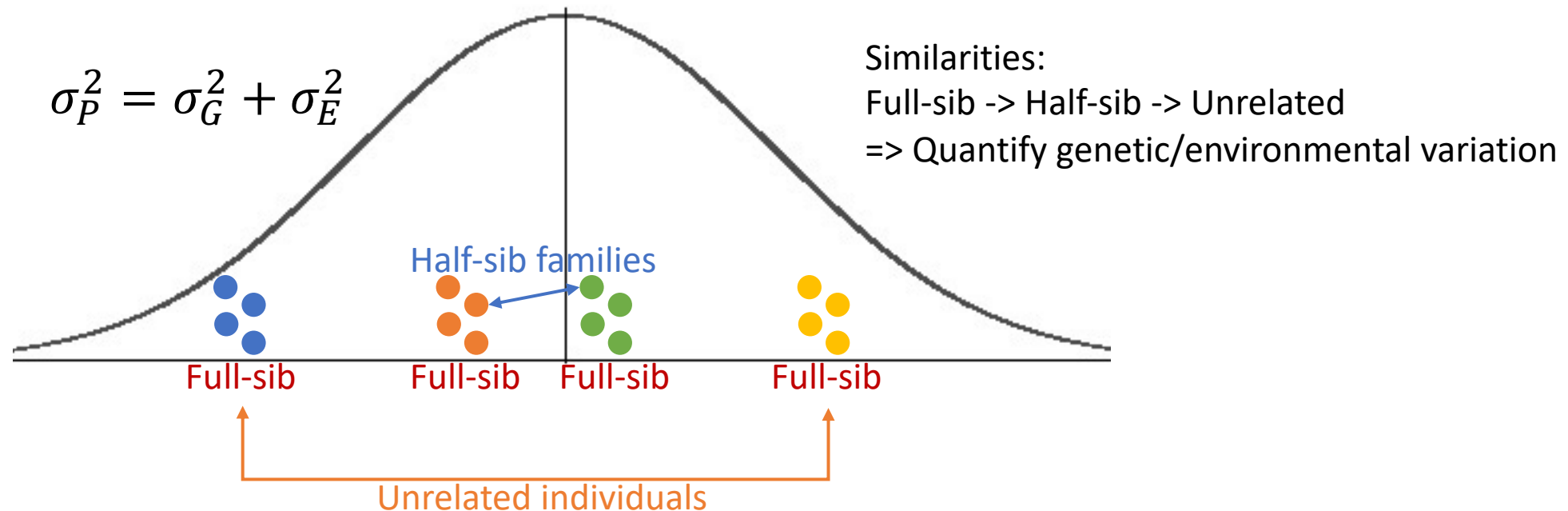


Gene + Environment



# Decompose trait variation in populations

- Basis: related individuals share genetic variants contributing to trait variation
- Statistical aim: identify how much more similar related individuals are (genetic variance,  $\sigma_G^2$ ) than unrelated ones (environmental variance,  $\sigma_E^2$ )



# Go beyond additive variation

Depending on how individuals are related, different components of the genetic variance can be estimated; additive ( $\sigma_A^2$ ), dominance ( $\sigma_D^2$ ) and interaction ( $\sigma_I^2$ ).

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

$$\sigma_P^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2$$



# Heritability

- A mathematically defined concept
- The broad-sense heritability:

The proportion of phenotypic variation due to genetics

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2}$$

- The narrow-sense heritability: (more often used)

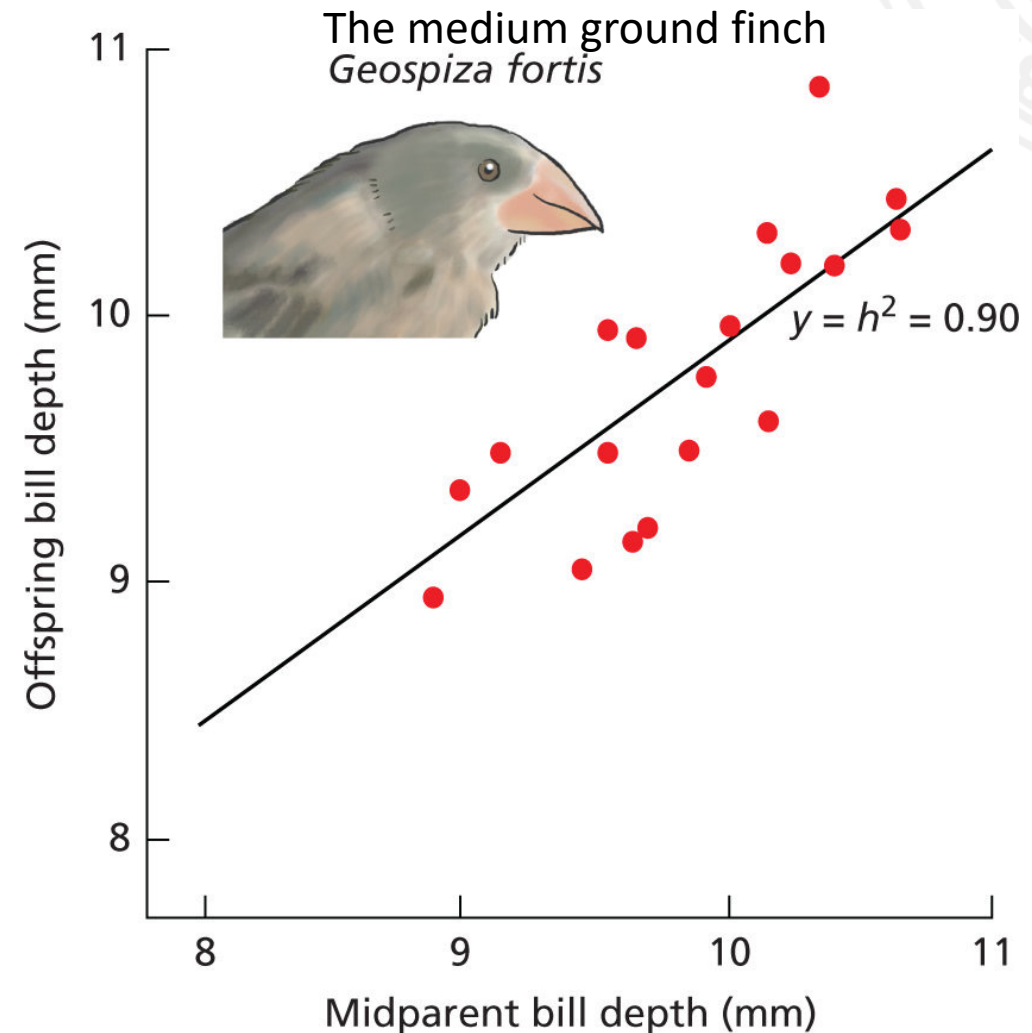
The fraction of phenotypic variance that can be attributed to variation in the additive effects of genes

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}$$

- Not a static property of a trait or a population, it a statistical estimate that is
  - Trait specific
  - Population specific
  - Dependent on the statistical approach used to estimate it
- The heritability does not show whether genetics is important or not

# Simply estimating the heritability: parent-offspring regression

- Galton's experiment, 1889
- Regress offspring phenotype on mid-parent value
- Slope = heritability
- “the resemblance between parent and offspring due to shared genes”

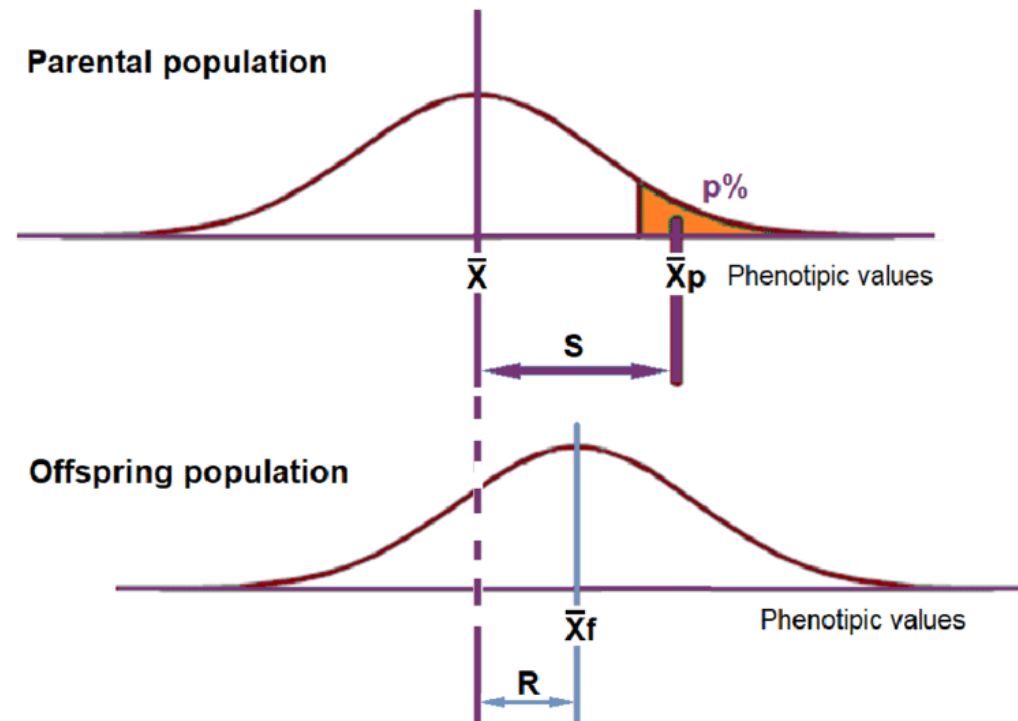


This figure is not done by Galton, just an example



# Selection response

- Heritability is excellent for within-population prediction
  - Benefits the process of breeding (predict selection response;  $R$ )
- It could also be another way to estimate the heritability



$$R = h^2 S$$

$S$ : Selection differential

$R$ : Selection response

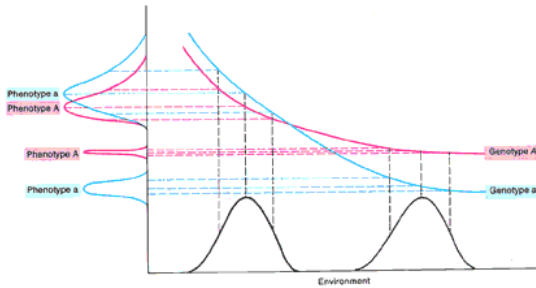




# *Molecular* quantitative genetics

Introduction to quantitative genetics

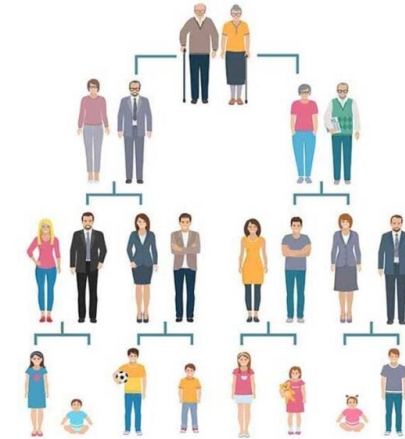
# Molecular quantitative genetics



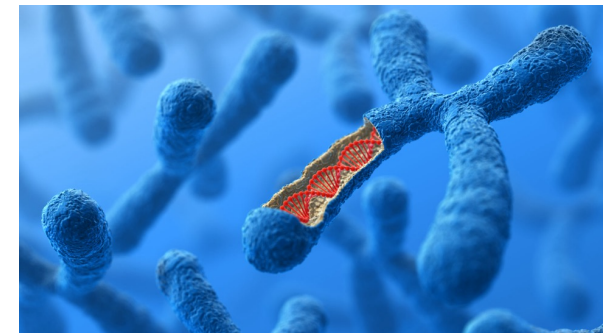
Statistical models



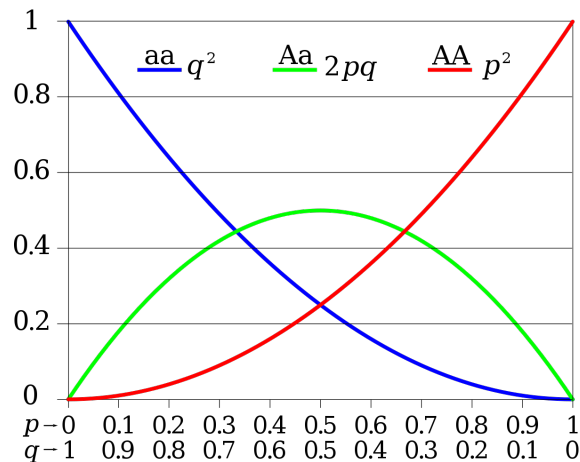
A genetic black box



Relationships



Molecular genetics data

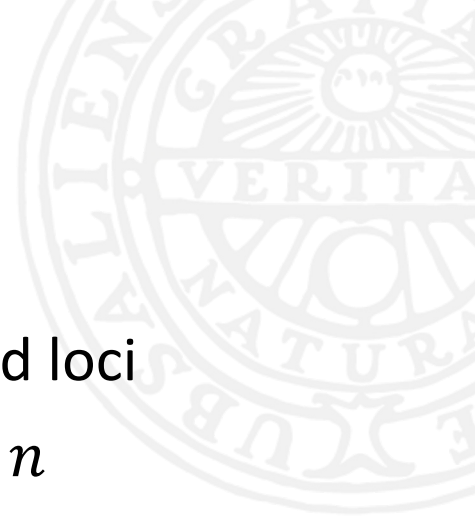


Population genetics models

# Quantitative genetics *data on genotypes at individual loci*

- Population genetics
  - Vital ingredient of the modern evolutionary synthesis
  - Focus on:
    - genetic variation within and between populations
    - contributions by genes and alleles to evolution
    - genetic forces contributing to evolution by changing allele frequencies
- *Molecular* quantitative genetics
  - Genotype data available on individual loci
  - Embrace population genetics theory on individual genes/alleles
  - Extend statistics modelling of phenotypic variation to defined loci





# Genetic variation from individual loci

- Basis: the relationship between individuals = allele-sharing at genotyped loci
- Statistical aim: estimate variance explained by allele-sharing at loci  $1 \cdots n$ 
  - Additive genetic variance:  $\sigma_{A1}^2, \sigma_{A2}^2, \dots, \sigma_{An}^2$
  - Population level  $\sigma_A^2$

$$\sigma_P^2 = \sigma_A^2 + \sigma_E^2$$

- Contribution from loci 1 to  $n$

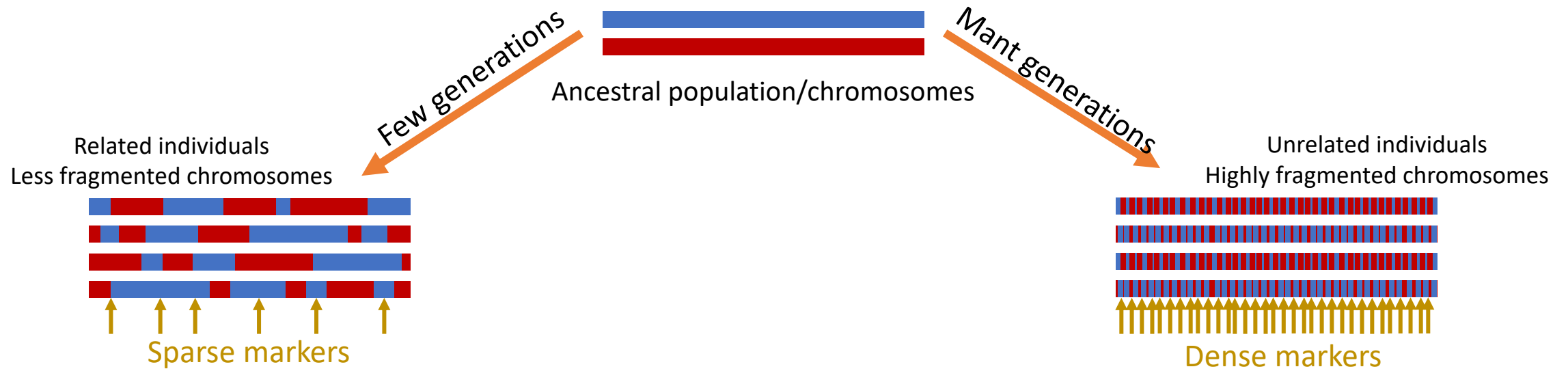
$$\sigma_P^2 = \sigma_{A1}^2 + \sigma_{A2}^2 + \dots + \sigma_{An}^2 + \sigma_E^2$$



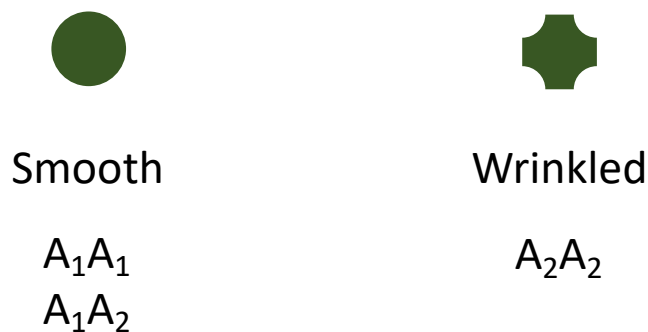
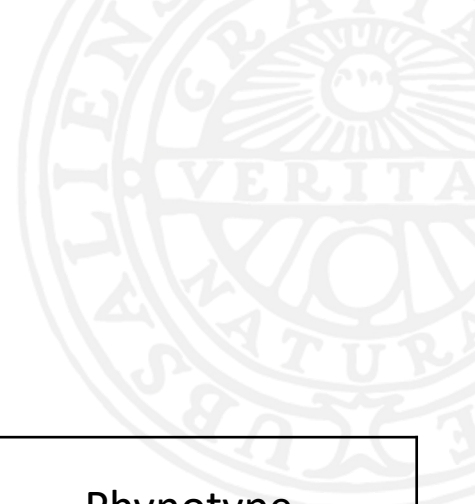
# Linkage & association mapping

What is linkage & association mapping?

- Identify loci explaining genetic variance in a quantitative trait
- Estimate allelic effects and amount of contributed variance
- Same quantitative genetics models, different populations



# Single-locus genotype-to-phenotype map



Quantitative genetics:  
 Set up a statistical model

- Genotypes  $A_1A_1$  and  $A_1A_2$  mapped to the smooth phenotype
- Genotype  $A_2A_2$  mapped to the wrinkled phenotype

Genotype		Phenotype
$A_1A_1$		0, $R=0$
$A_1A_2$		0, $R + a + d$ ( $d=-1/2$ )
$A_2A_2$		1, $R + 2a$ ( $a=1/2$ )

Two allele substitutions from  $A_1$  to  $A_2$   
 $\rightarrow \frac{1}{2}$  wrinkles for each allele substitution

In algebraic notation:

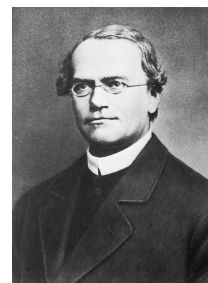
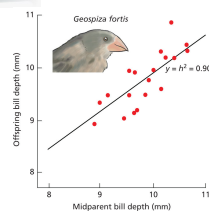
$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix} \times \begin{pmatrix} R \\ a \\ d \end{pmatrix}$$



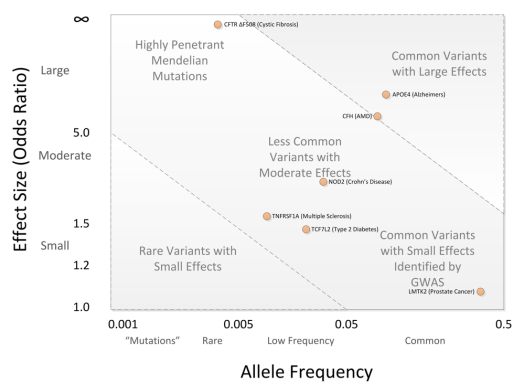
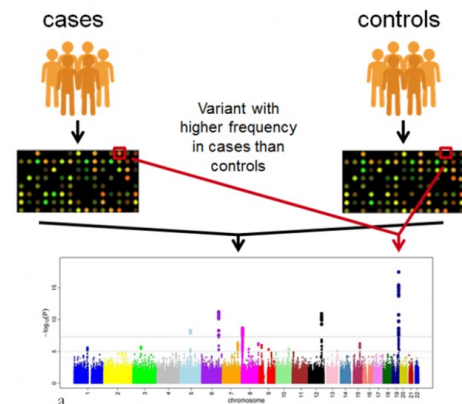
# What we are trying to do?



Diagram from Freeman & Herron, *Evolutionary Analysis*



$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix} \times \begin{pmatrix} R \\ a \\ d \end{pmatrix}$$





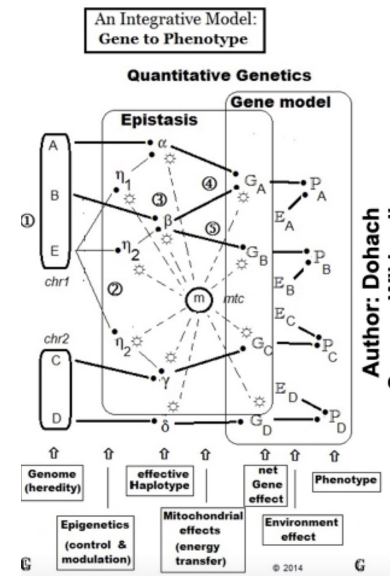
# Challenges when studying individual loci

- Quantitative genetics is a way to study genetics in the population  
→ Useful for predicting selection responses, breeding, evolution
- How about individuals?



Diagram from Freeman & Herron,  
*Evolutionary Analysis*

Healthy or not?







# Challenges when studying individual loci

- Analyses rely on statistical models providing population-level statistics, in particular, genetic effects and variances
- More reliable for Mendelian than Quantitative traits
- GWAS challenges:
  - Missing heritability
  - Failure of replication
  - Inter-population applicability
  - Statistical testing problems
  - ...
- Why so difficult?
  - Rare variants? Smaller effects than expected? Genetic interactions? ...?



# The problem of rare variants & small effects

- A locus contributes to trait variation as:

$$\sigma_A^2 = 2pq a^2$$

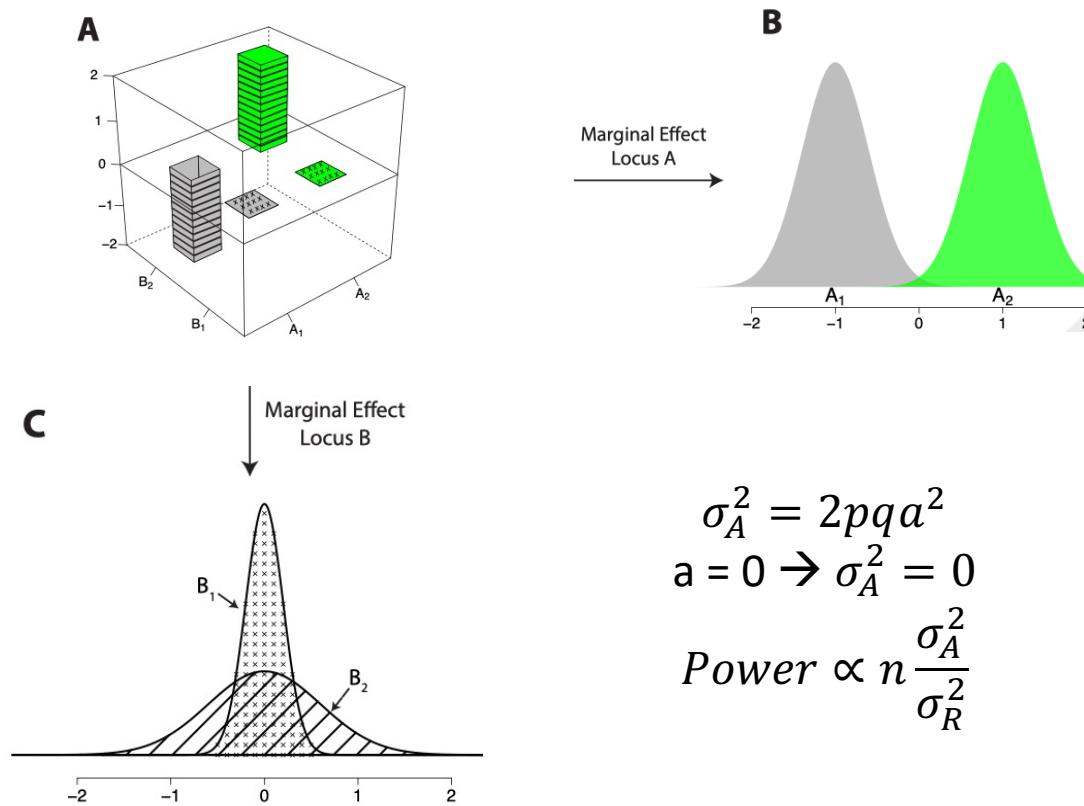
Allele frequencies

Additive genetic effect

$$Power \propto n \frac{\sigma_A^2}{\sigma_R^2}$$

- ➔ To detect rare variants, or variants with small effects, need large n  
(Motivated by this, sample sizes in many human studies today are >100k individuals)

# Genetic interactions (epistasis)



Epistasis can make loci undetectable even at very large n  
 Challenge to replicate and interpret effects of loci in different populations

# Missing heritability

- Heritability can be estimated using genomics data relationships by average genome sharing or by sharing of trait-associated SNPs

$$h_{SNP}^2 = \frac{\sigma_{A-SNP}^2}{\sigma_P^2} < h_{ped}^2 = \frac{\sigma_{A-ped}^2}{\sigma_P^2}$$

- GWAS variants accounts for little of the heritability
  - True for most diseases, behaviors, and other phenotypes
- Could, for example, be due to the factors mentioned above
  - Rare variants, small-effects or epistasis





# Quality Control

# Why quality control is important?

- We don't live in an ideal world...
- Large-scale experiments generate both true results and a proportion of false results
- Errors might come from any steps in the process
  - Sample selection → cryptic relatedness, population structure
  - Genotyping
  - ...

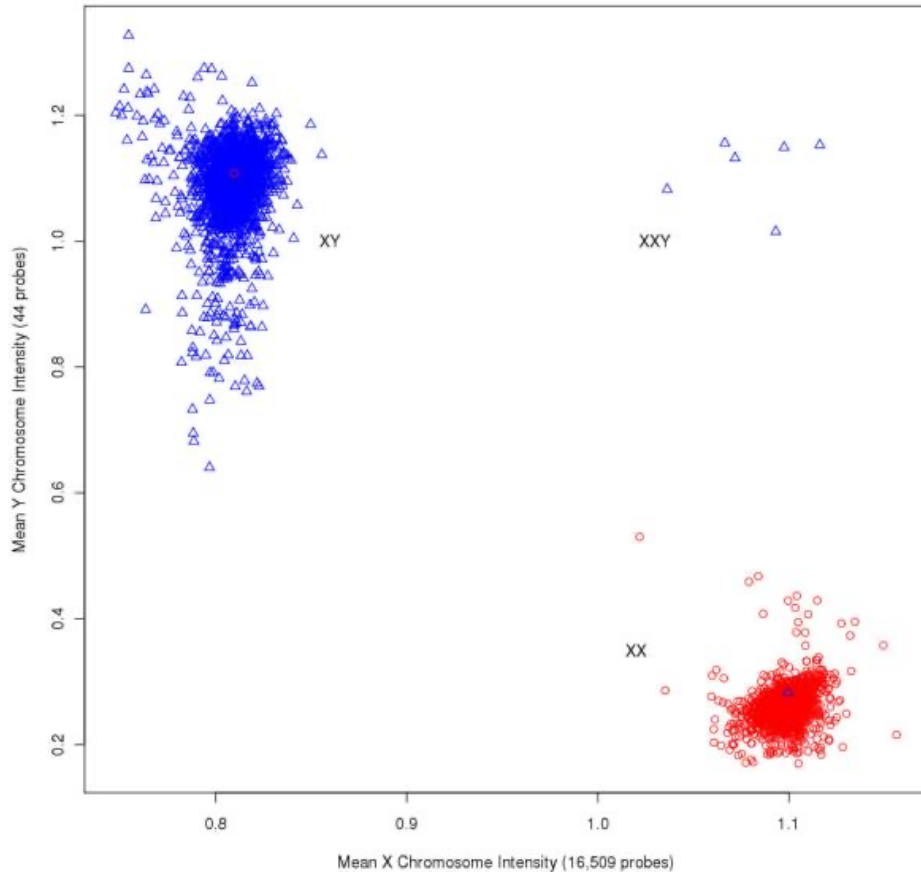




# An overview of QC steps

- Sample QC is aimed at the identification and removal of individuals with
  - sex discrepancy
  - low call rate
  - excess genome-wide heterozygosity and homozygosity
- Variant QC is aimed at identification and removal or refinement of variants with
  - low call rate
  - deviation from Hardy-Weinberg Equilibrium (HWE)
  - very low minor allele counts (MAC)

# Sex discrepancy

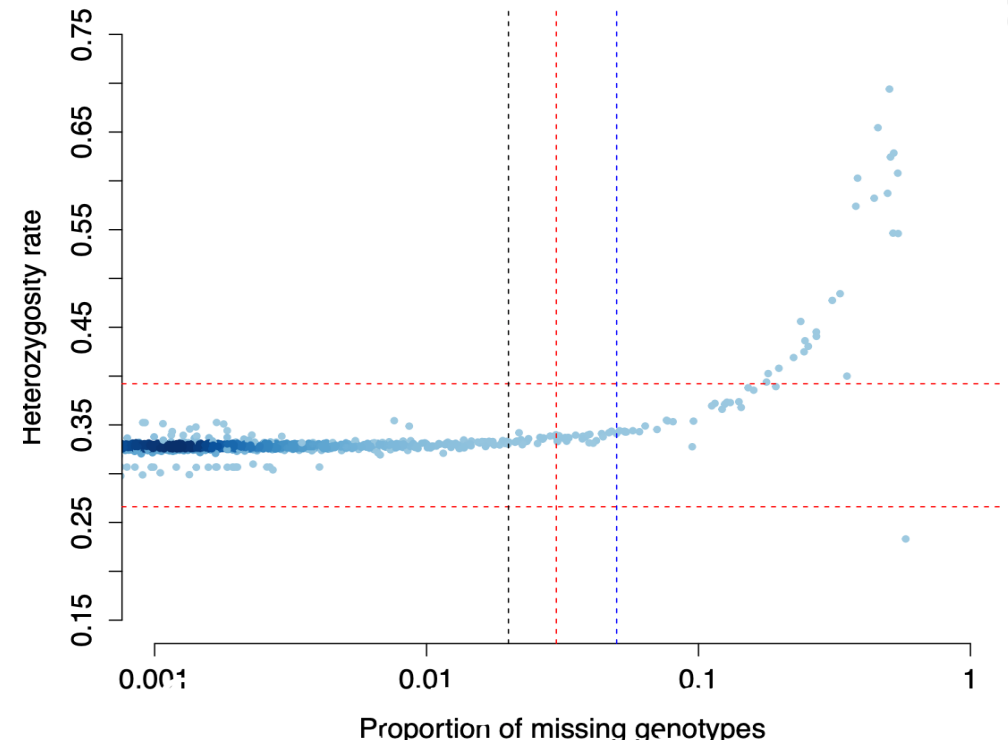


- --check-sex option in PLINK
- Report individuals for whom the sex recorded in the data does not match the predicted sex based on genetic data.
- Reveal sex chromosome anomalies
  - Turner syndrome (females having karyotype XO)
  - Klinefelter syndrome (male having karyotype XXY)
- The intensity plot
  - Females should have low Y intensity and high X
  - Males show similar levels of X and Y

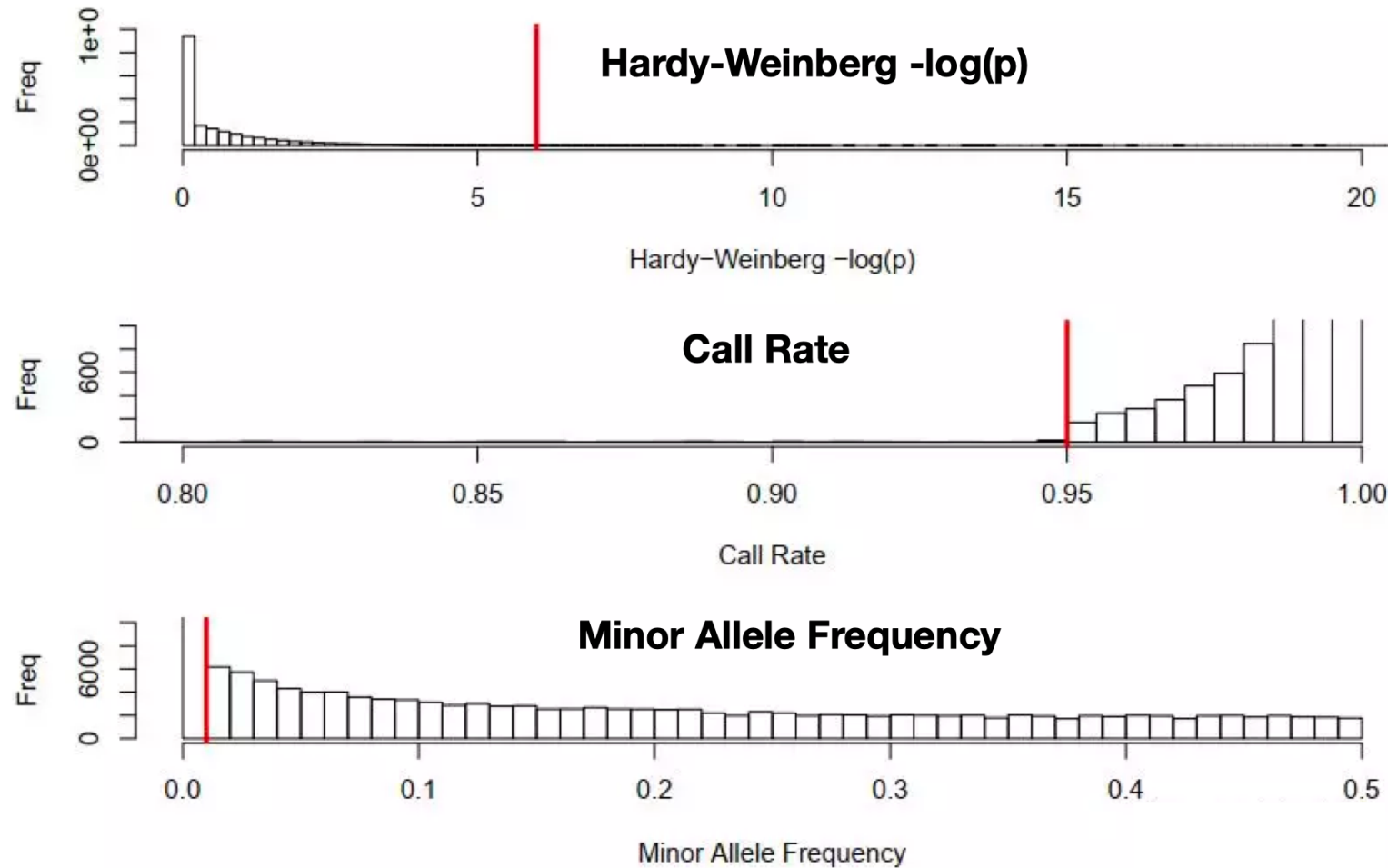


# Missingness and Heterozygosity

- Genotypic call rate
  - Per sample (individual) rate
  - The number of non-missing genotypes is divided by the number of genotyped markers.
- Heterozygosity Rate
  - Per sample (individual) rate
  - Excess heterozygosity: Possible sample contamination
  - Less than expected heterozygosity: Possibly inbreeding



# HWE, Genotype Call Rate, and MAF (variant level)

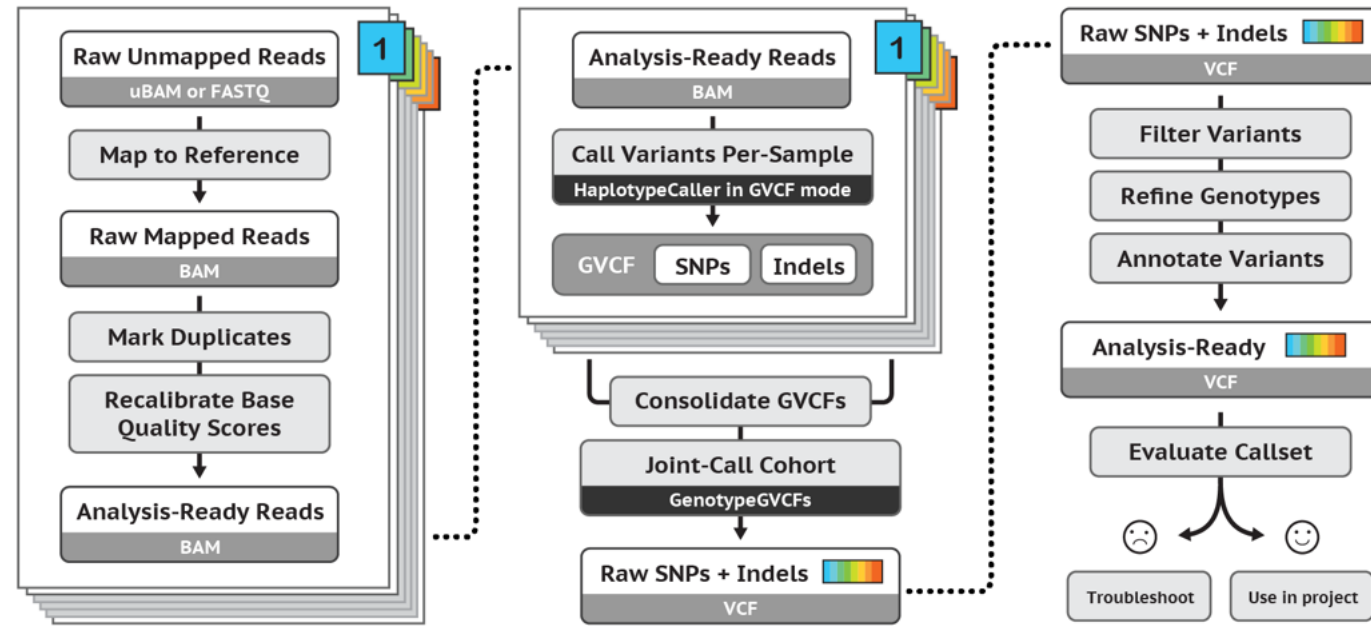


# Commonly used tools

- GATK: Genome Analysis Toolkit (<https://gatk.broadinstitute.org/hc/en-us>)
- VCFtools: <https://vcftools.github.io/index.html>
- R
- PLINK



# GATK



*Best Practices for SNP and Indel discovery in germline DNA  
- leveraging groundbreaking methods for combined power  
and scalability.*



# VCFtools

- Easy to use with a great manual: [https://vcftools.github.io/man\\_latest.html](https://vcftools.github.io/man_latest.html)

```
vcftools --vcf raw.vcf
--mac
--minQ
--minDP
--remove-indels / --keep-only-indels
--min-alleles 2 --max-alleles 2
--max-missing
--maf
```

# R packages

- plinkQC: <https://meyer-lab-cshl.github.io/plinkQC/>
- GenABEL: <https://github.com/GenABEL-Project/GenABEL>
- snpMatrix:  
<https://www.bioconductor.org/packages//2.7/bioc/html/snpMatrix.html>
- Write your own command?



# Plink



Option	Description
<code>--check-sex</code>	check for sex discrepancy
<code>--remove</code>	removes samples in the list
<code>--maf</code>	filters out SNPs with minor allele freq below threshold
<code>--hwe</code>	filters out SNPs with HWE exact test p-value below threshold
<code>--missing</code>	investigate missingness per individual and per SNP
<code>--geno</code>	filters SNPs with genotyping frequency below the threshold
<code>--mind</code>	exclude individuals with genotype rates below the threshold

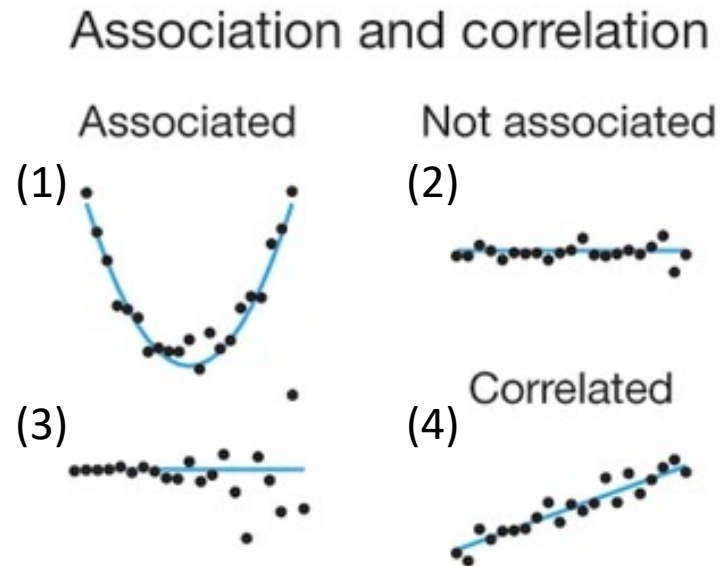


# GWAS

The discovery of **associations** between certain **variations** in the genetic code and the physical trait

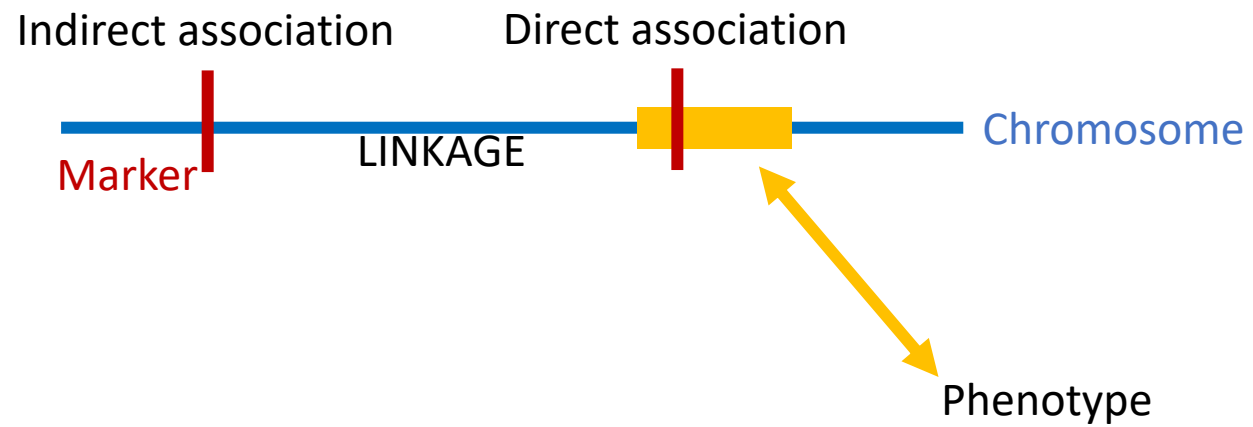


# Statistical association

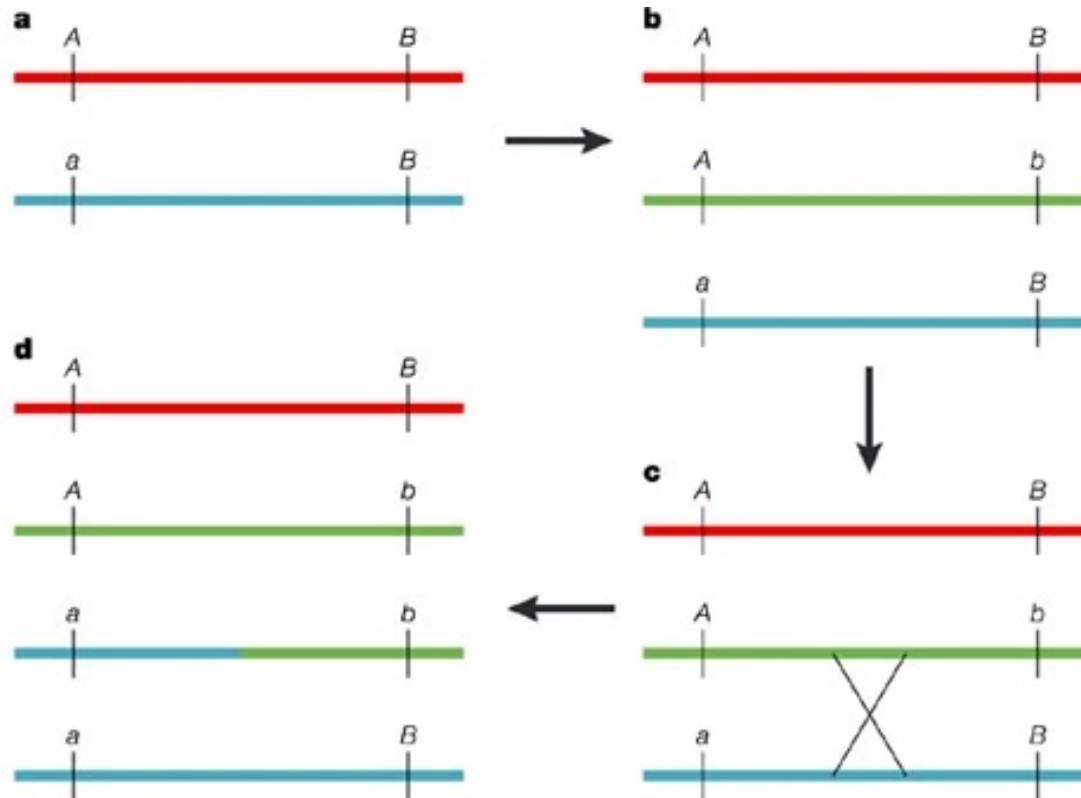


- Association is a very general relationship  
→ One variable provides information about another ( $\Delta x \rightarrow \Delta y$ )
- Correlation is more specific  
→ When displaying an increasing or decreasing trend  
( $x \uparrow \rightarrow y \uparrow$  or  $x \uparrow \rightarrow y \downarrow$ )

# Direct and indirect association

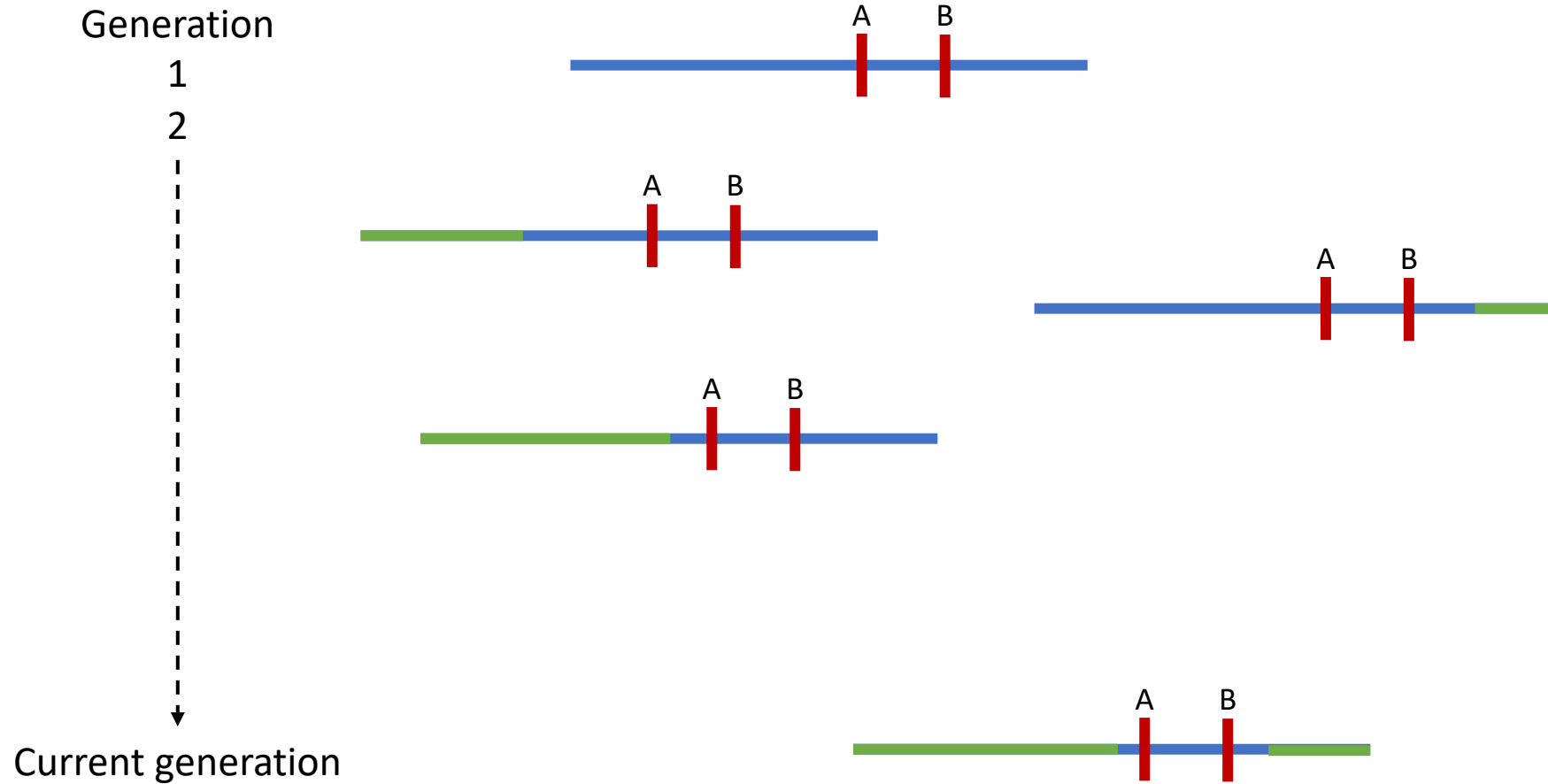


# Erosion of Linkage Disequilibrium (LD)



- There is a polymorphic locus with alleles A and a.
- A mutation occurs at a nearby locus, changing an allele B to b.
- Association between alleles at the two loci gradually be disrupted by recombination between the loci.
- Decline the LD among the markers in the population as the recombinant chromosome (a, b) increases in frequency.

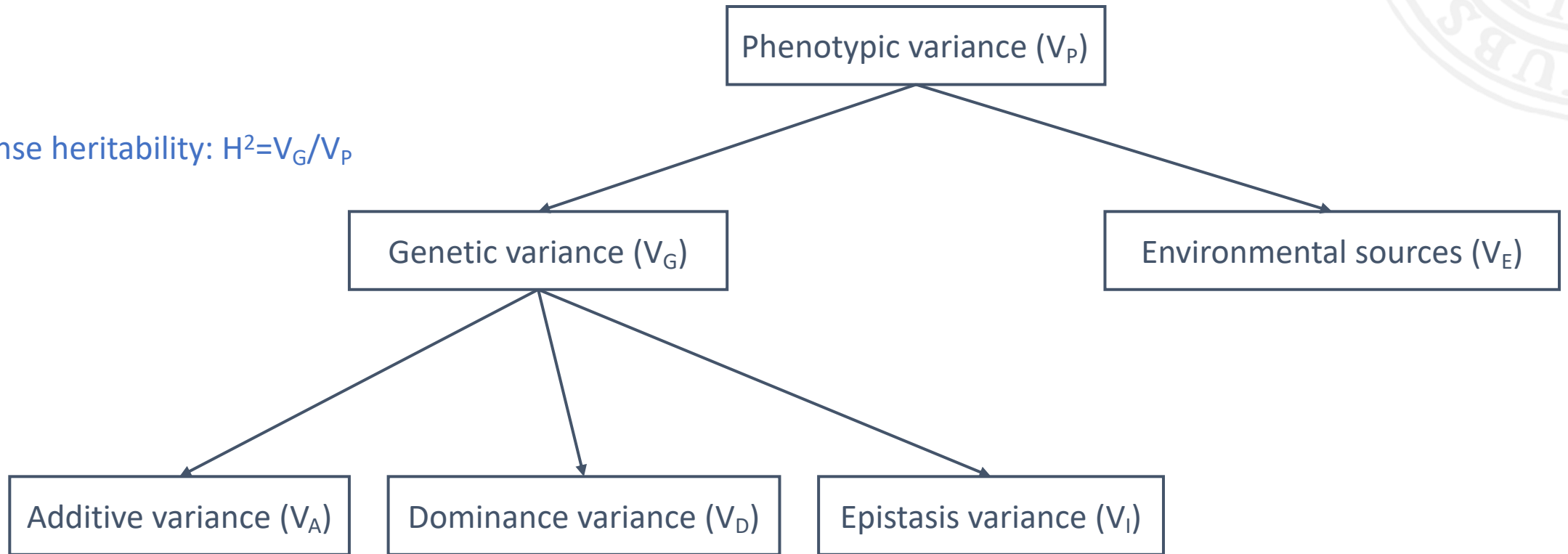
# Linkage Disequilibrium (LD)



# Variations

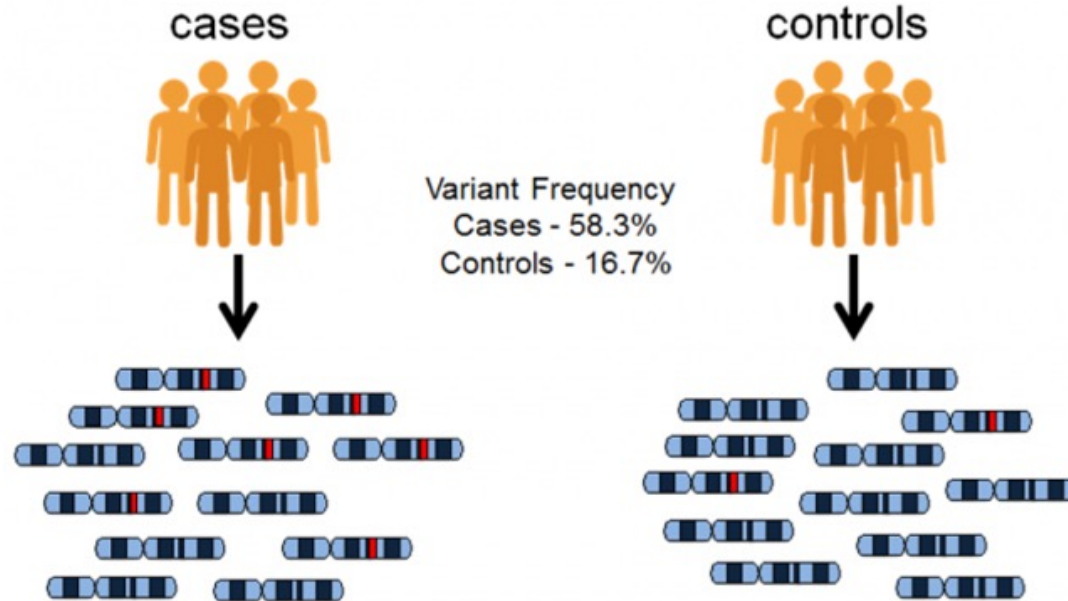


Broad-sense heritability:  $H^2 = V_G / V_P$



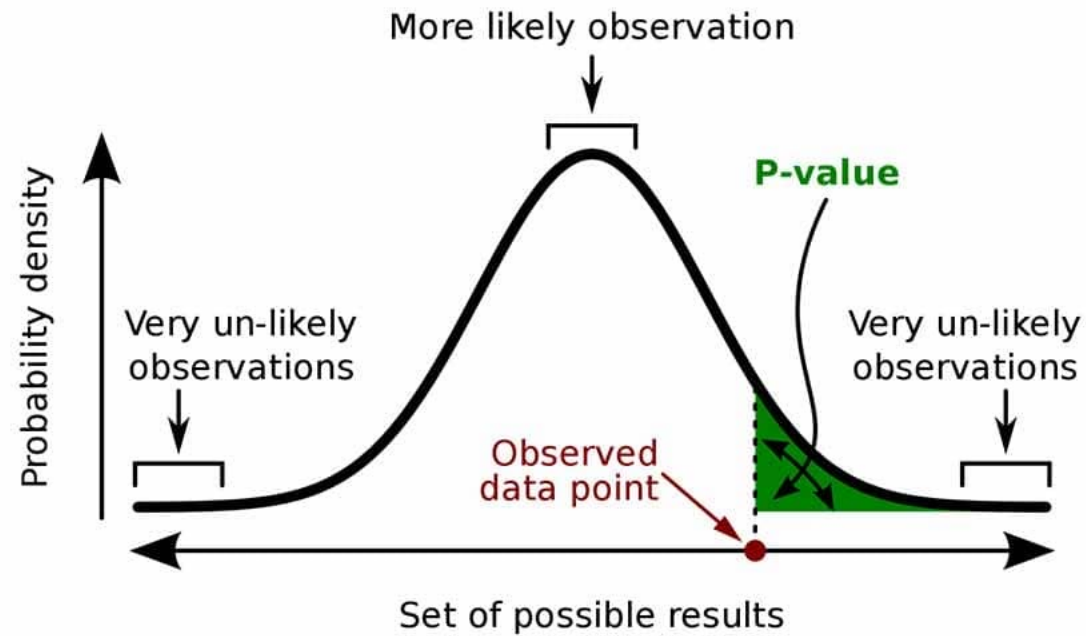
Narrow-sense heritability:  $h^2 = V_A / V_P$

# What are genome-wide association studies?



- Identify the association between genetic variance and phenotypic variance
- Cases have a higher frequency of carrying causal variants (and highly linked markers)

# p-value

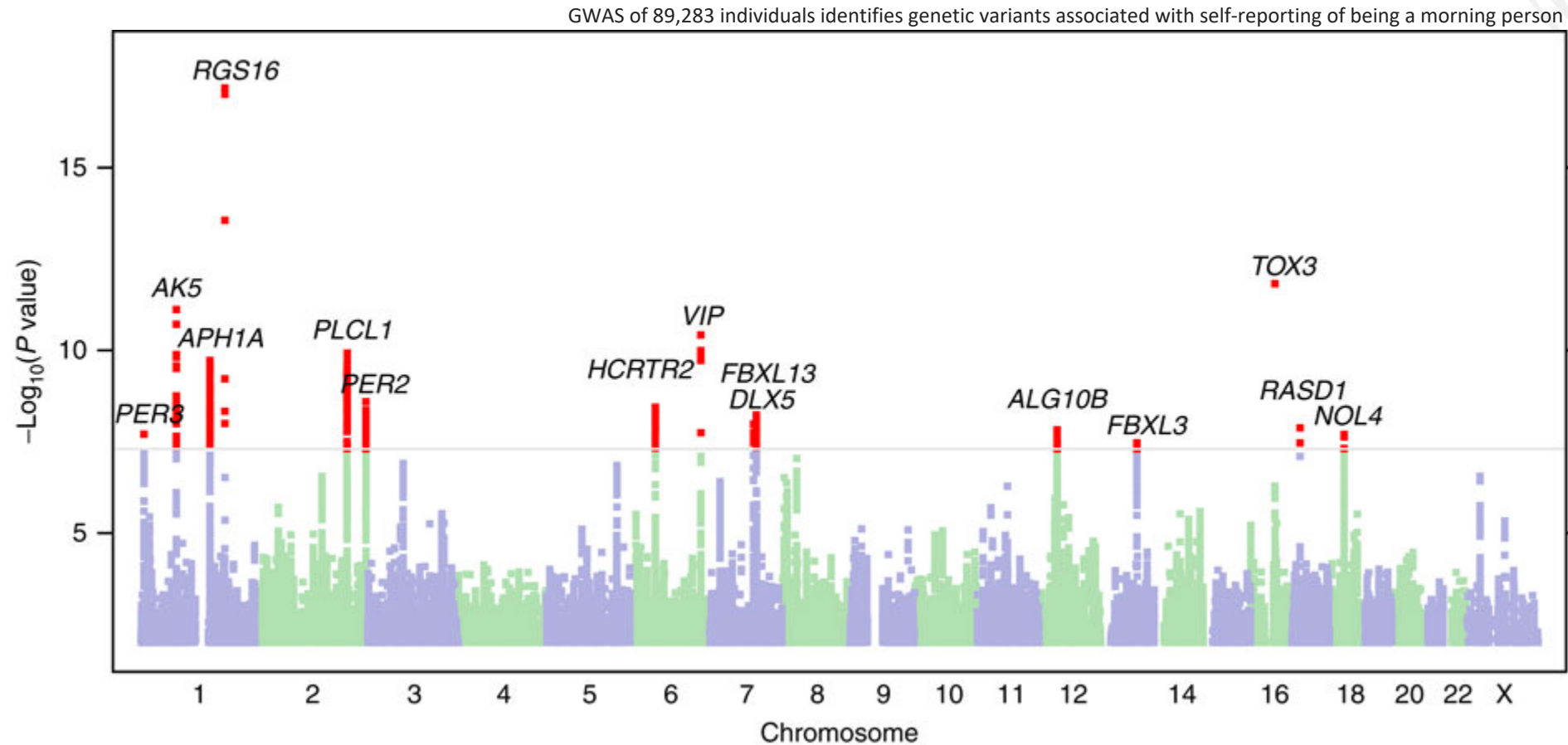


A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.



# Manhattan plot

Stronger  
Association

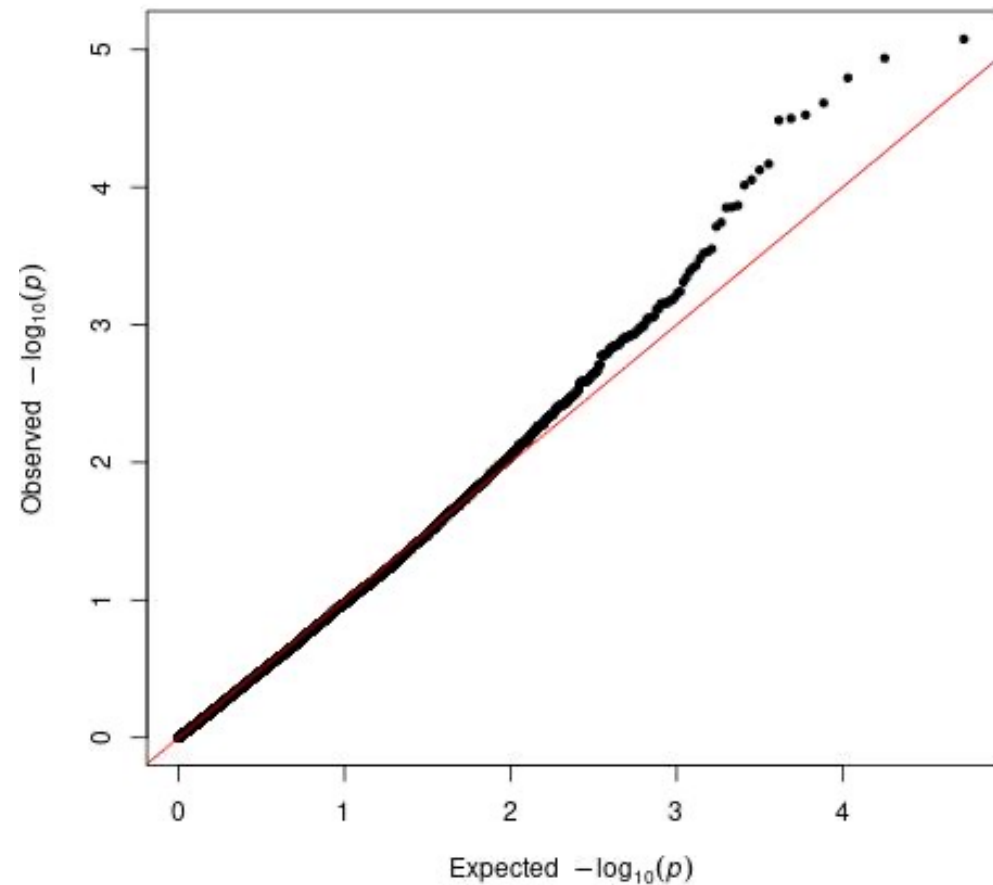
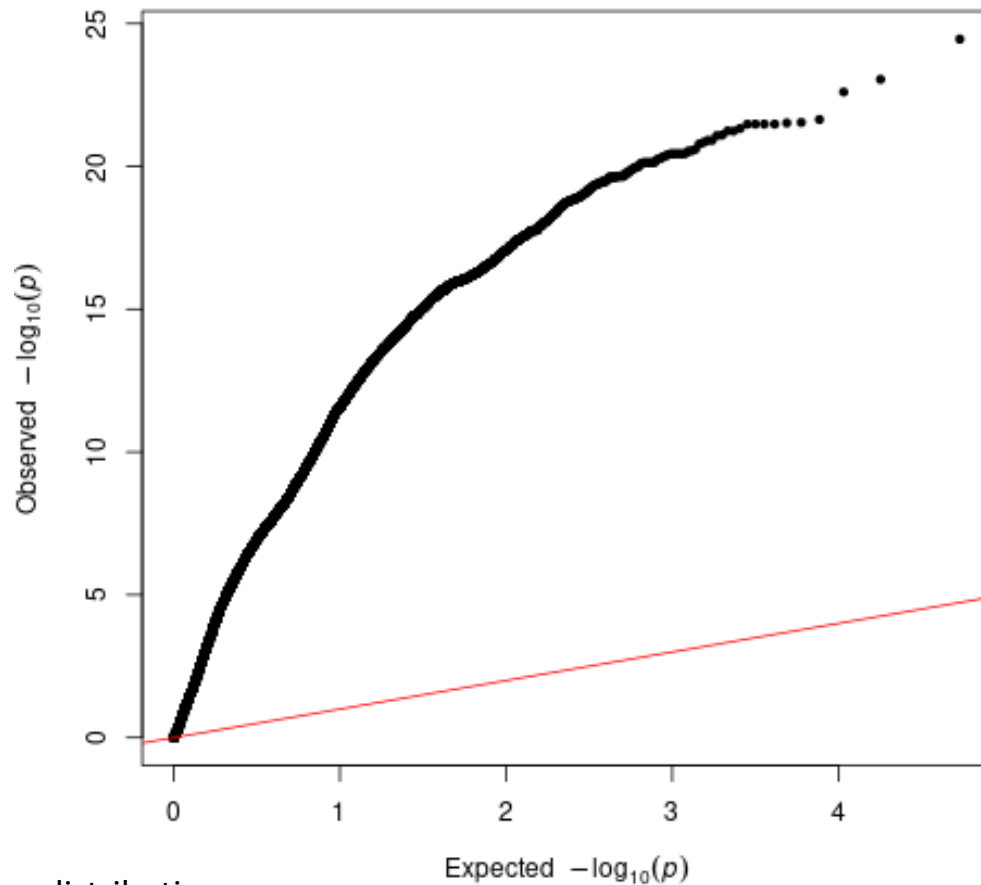




# Q-Q plot

comparing two probability distributions by plotting their quantiles against each other

		Truth	
		H0 is true (non-carrier)	H1 is true (carrier)
Decision	Fail to reject H0 (negative)	Correct decision (true negative)	Type II error (false negative)
	Reject H0 (positive)	Type I error (false positive)	Correct decision (true positive)





# Type of GWAS

## Populations used in GWAS

- Population-based
- Family-based
- Isolated population

## Phenotype measurement

- Qualitative (Usually binary; affected / not affected)
- Quantitative

## The complexity of genetic effect

- Single marker (one marker a time)
- Multi-marker (multivariate method)



# Association study example (I)

- Population-based GWAS
- Qualitative measurement (affected / not affected)
- Single marker

Observed	Case	Control	Total
AA	20	50	70
Aa	20	30	50
aa	60	20	80
Total	100	100	200

# Association study example (I)



Observed	Case	Control	Total
AA	20	50	70
Aa	20	30	50
aa	60	20	80
Total	100	100	200

Expected (No effect)	Case	Control	Total
AA	35	35	70
Aa	25	25	50
aa	40	40	80
Total	100	100	200



# Association study example (I)

The test statistic for the association study, in this case, is the  $\chi^2$  test.

## Chi-Square Test

- Association between two qualitative variables is statistically significant
- $H_0$ : There is no difference between the two variables  
 $H_1$ : There is a significant difference between the two variables
- The test statistic: determine whether the difference between the observed and expected values is statistically significant

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

# Association study example (I)

Observed	Case	Control	Total
AA	20	50	70
Aa	20	30	50
aa	60	20	80
Total	100	100	200

Expected	Case	Control	Total
AA	35	35	70
Aa	25	25	50
aa	40	40	80
Total	100	100	200

$$\chi^2 = \frac{(20 - 35)^2}{35} + \frac{(50 - 35)^2}{35} + \frac{(20 - 25)^2}{25} + \dots + \frac{(20 - 40)^2}{40} = 34.86$$

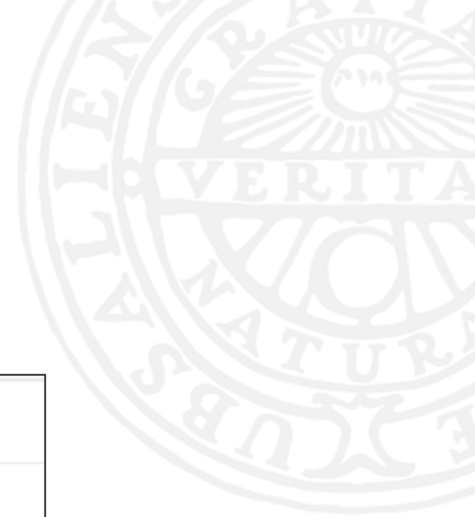
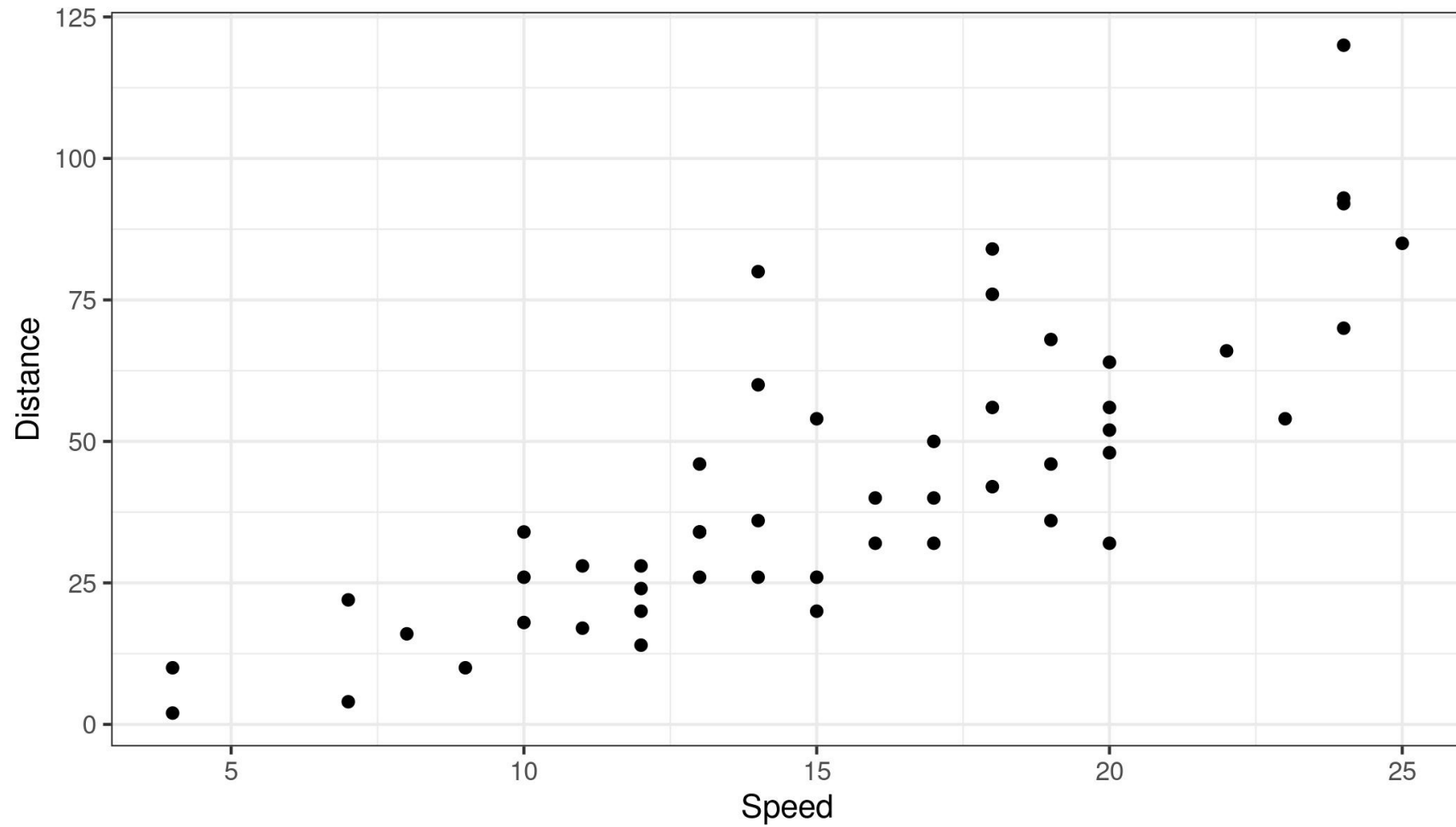
```
> dat = data.frame(case = c(20, 20, 60), control = c(50, 30, 20))
> dat
  case control
1   20     50
2   20     30
3   60     20
> chisq.test(dat)
```

Pearson's Chi-squared test

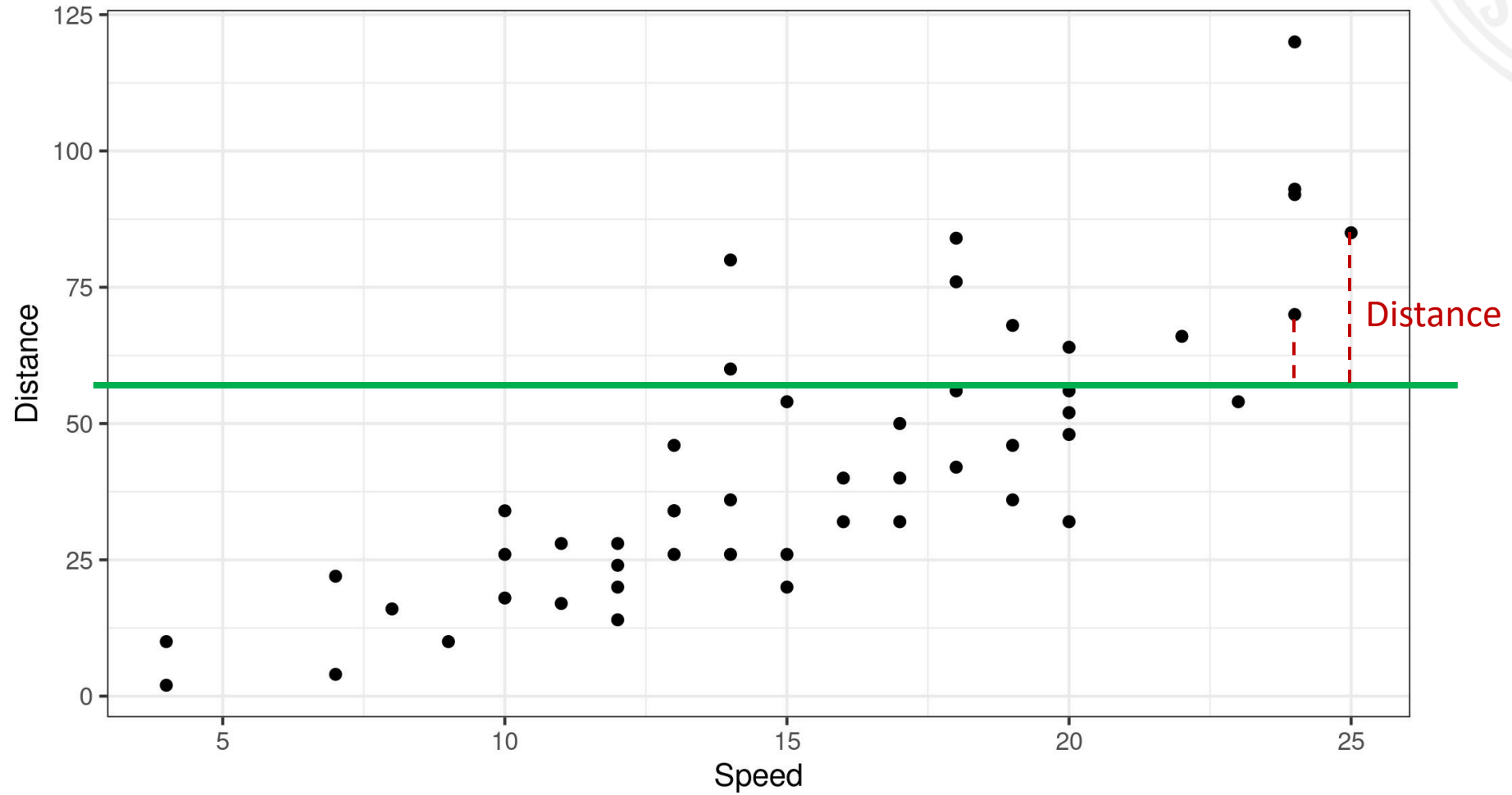
```
data: dat
X-squared = 34.857, df = 2, p-value = 2.697e-08
```

Reject  $H_0$ . There is a difference between case and control.

# Linear regression

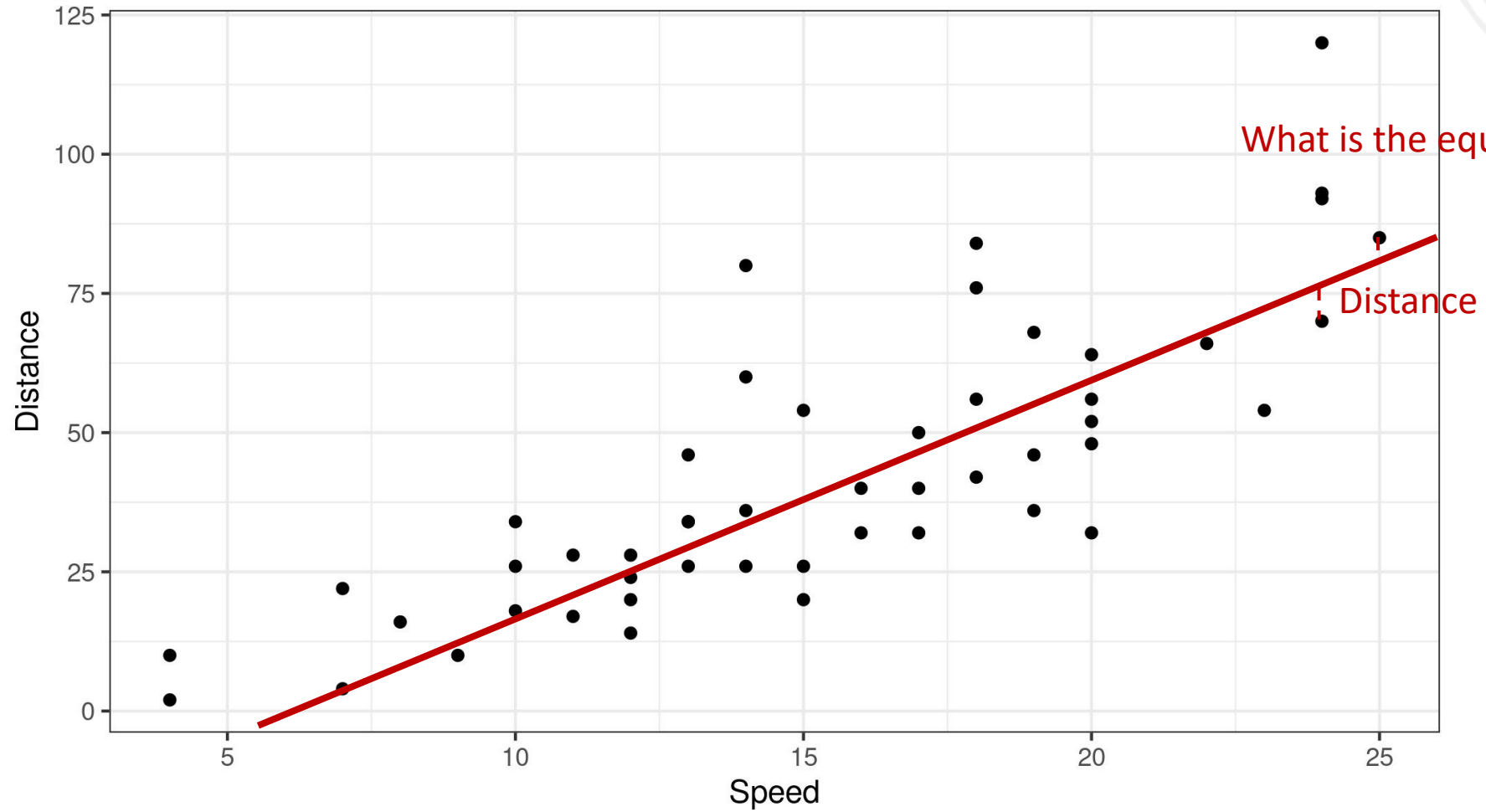
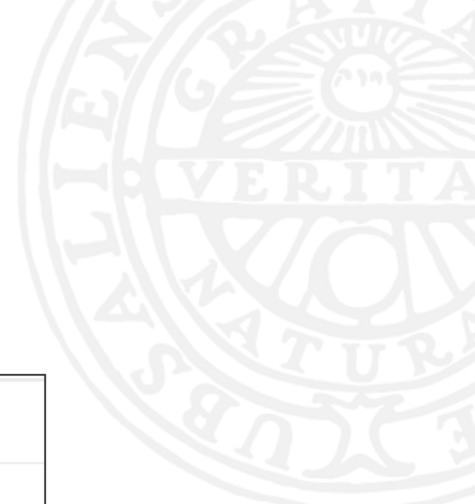


# Linear regression





# Linear regression





# Linear regression

$$y = \beta_0 + \beta_1 x + e$$

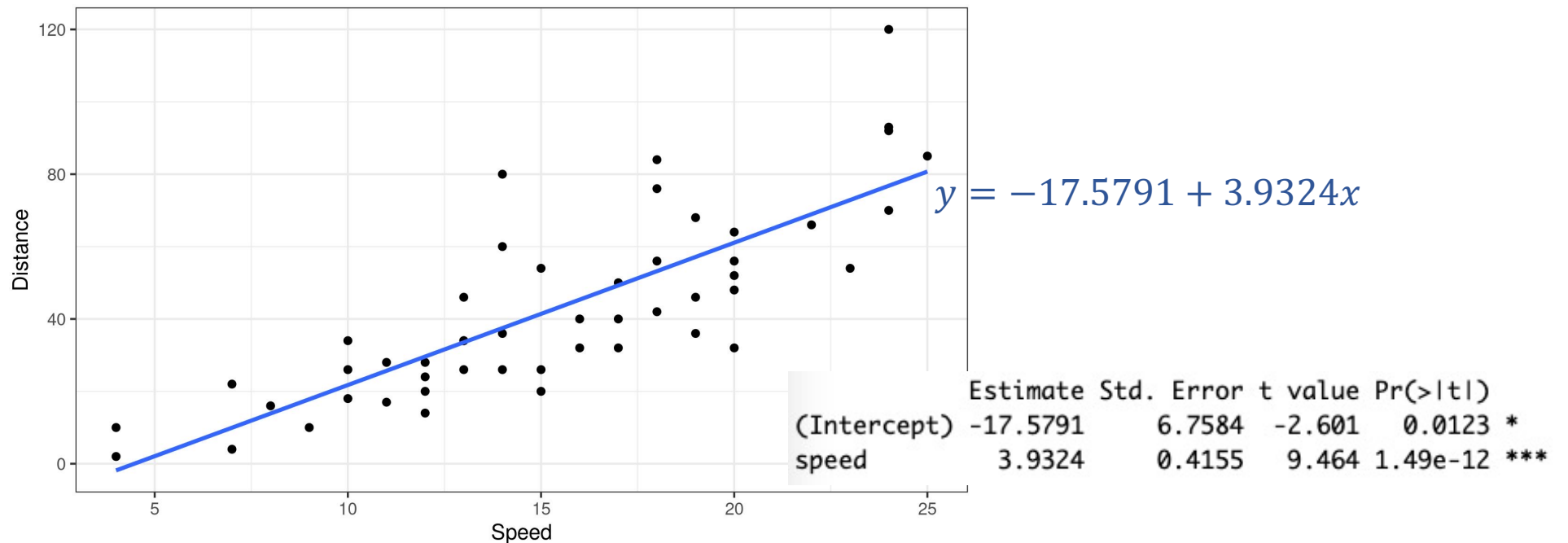
error

$\beta_0$ : y intercept

- $H_0: \beta_0 = 0$  vs.  $H_1: \beta_0 \neq 0$

$\beta_1$ : slope

- $H_0: \beta_1 = 0$  vs.  $H_1: \beta_1 \neq 0$





## Association study example (II)

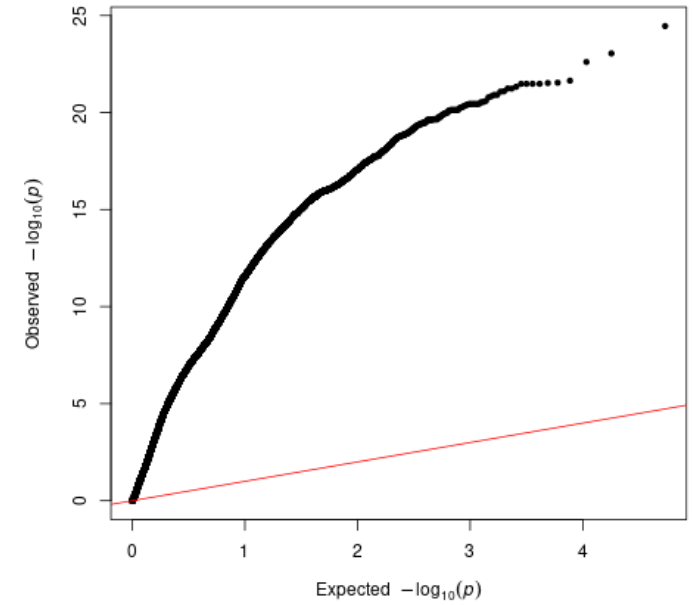
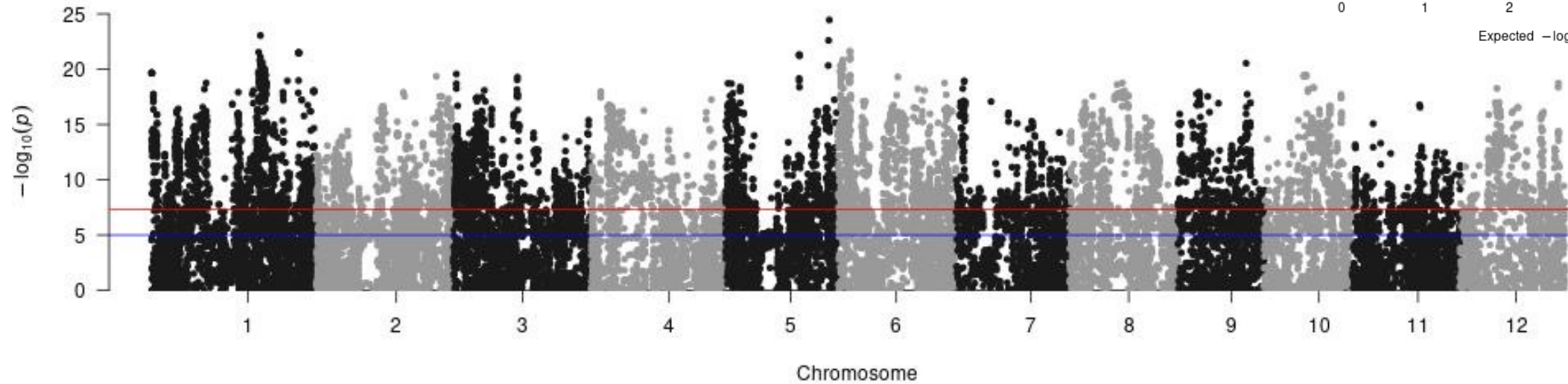
- Quantitative traits (such as blood glucose levels, BMI)
- Single marker

Use simple linear regression model for this study.

$$y = \beta_{0i} + \beta_{gi}x_{gi} + \epsilon_i$$

$i = 1, \dots$ , number of SNPs

# Association study example (II)



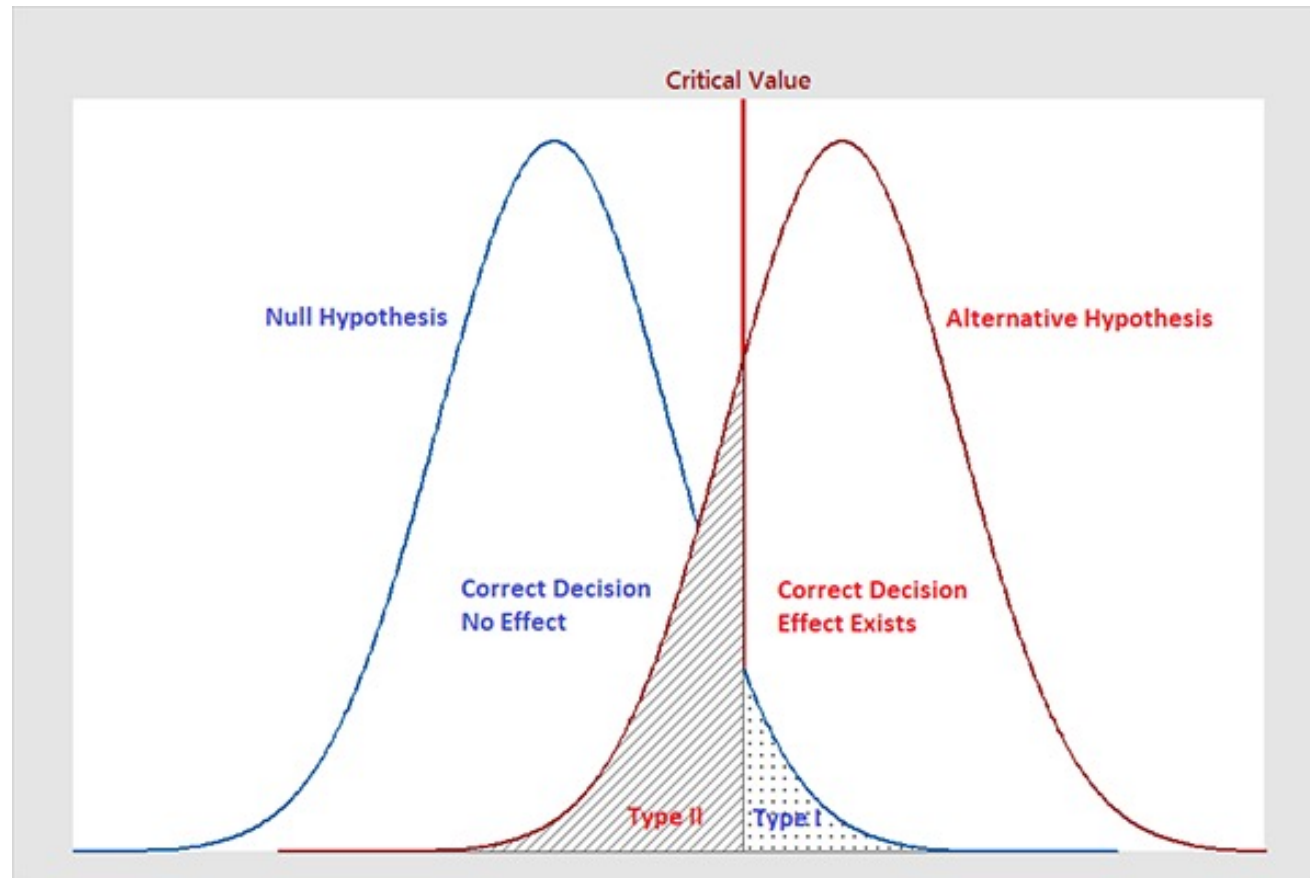


# Statistical Power

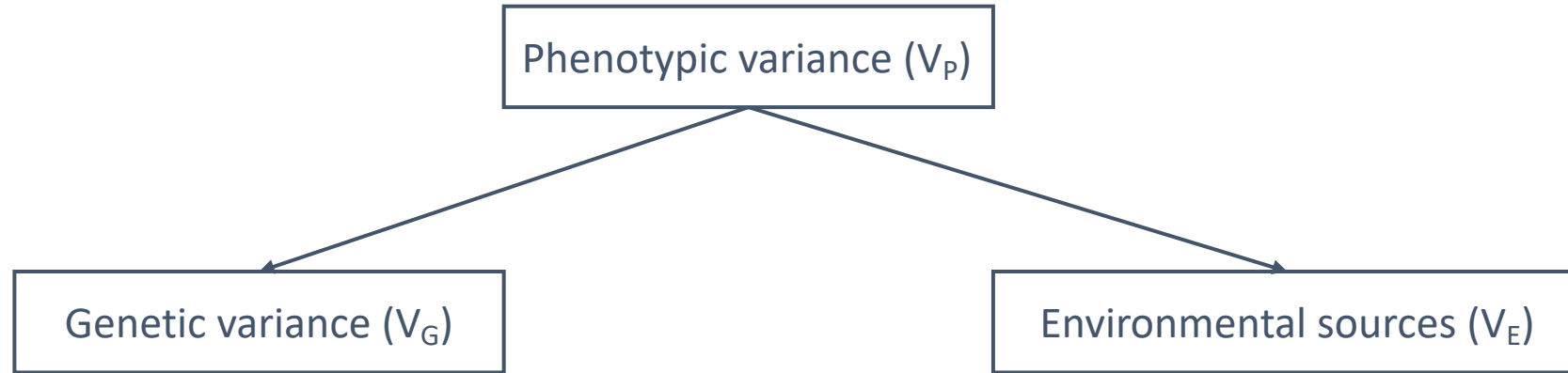
<https://www.nature.com/articles/s41437-019-0205-3>

# Statistical Power

Statistical power = The ability to correctly reject the null hypothesis



# Variance



$$Y = \mu + e_1 \rightarrow Y = \mu + G + e_2 \rightarrow Y = \mu + G + E + e_3 \rightarrow Y = \mu + G + P + E + e_4 \rightarrow Y = \mu + G + K + P + E + e_4$$

The diagram illustrates the decomposition of phenotypic variance into its components through a series of equations. Colored brackets and arrows indicate the addition of each component:

- Green:** Genetic variance ( $G$ ) is added to the mean ( $\mu$ ) in the second equation.
- Orange:** Environmental variance ( $E$ ) is added to the genetic component in the third equation.
- Red:** Phenotypic variance ( $P$ ) is added to the environmental component in the fourth equation.
- Black:** A fifth equation introduces a new component ( $K$ ), which is not explicitly defined in the diagram.



# Derivation (simple linear model)

The linear model

$$y = \mu + Z\gamma + e$$

Response variable  $y$ , Intercept  $\mu$ , regression coefficient  $\gamma$ , independent variable  $Z$ , residual error vector  $e \sim N(0, \sigma^2)$

$H_0: \gamma = 0$  vs.  $H_1: \gamma \neq 0$

Let  $\hat{\gamma}$  be the estimated regression coefficient, its variance can be shown as:

$$\text{var}(\hat{\gamma}) = \sigma_r^2 = \frac{\sigma^2}{\sum_{j=1}^n (Z_j - \bar{Z})^2} \approx \frac{\sigma^2}{n\sigma_Z^2}$$

The Wald chi-square test statistic:  $W = \frac{r^2}{\text{var}(r)} = n\sigma_Z^2 \frac{r^2}{\sigma^2}$

Wald test:  $W = \frac{(\hat{\theta} - \theta_0)^2}{\text{var}(\hat{\theta})}$



# Derivation (simple linear model)

$$W = \frac{r^2}{\text{var}(r)} = n\sigma_Z^2 \frac{r^2}{\sigma^2}$$

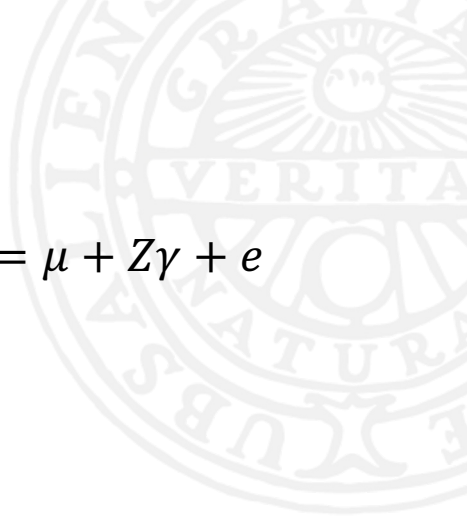
- When  $n$  is sufficiently large  $\rightarrow W \sim \chi^2(1)$  under the null model  
 $\rightarrow$  To simplify, use the central  $\chi^2(1)$  distribution as the null distribution
  - The critical value for the test:  $\chi_{1-\alpha}^2$
- If the null hypothesis is rejected, the Wald-test will follow a non-central  $\chi^2$  distribution with a non-central parameter

$$\delta = n\sigma_Z^2 \left(\frac{\gamma}{\sigma}\right)^2$$

If the independent variable is standardized prior to the analysis  $\rightarrow \sigma_Z^2 = 1$

$$\delta = n \left(\frac{\gamma}{\sigma}\right)^2$$

# Derivation (simple linear model)


$$y = \mu + Z\gamma + e$$

Have a look at the non-centrality parameter ( $\rightarrow$  statistical power)

$$\delta = n \left( \frac{\gamma}{\sigma} \right)^2$$

- Sample size:  $n$ 
  - Sample size  $\uparrow \rightarrow$  Power  $\uparrow$
- Effect size:  $\left( \frac{\gamma}{\sigma} \right)^2$ 
  - (Regression coefficient / Residual error)<sup>2</sup>
  - Effect size  $\uparrow \rightarrow$  Power  $\uparrow$

## Derivation (simple linear model)

Another way to represent the size of the QTL is by the heritability.

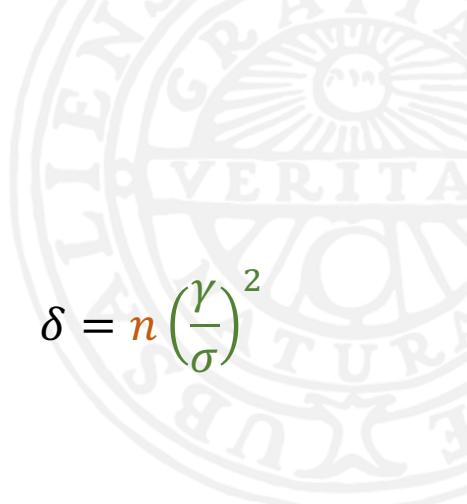
$$h_{QTL}^2 = \frac{V_G}{V_G + V_E} = \frac{\gamma^2}{\gamma^2 + \sigma^2} = \frac{\left(\frac{\gamma}{\sigma}\right)^2}{\left(\frac{\gamma}{\sigma}\right)^2 + 1}$$

Therefore,

$$\left(\frac{\gamma}{\sigma}\right)^2 = \frac{h_{QTL}^2}{1 - h_{QTL}^2}$$

And the non-central parameter

$$\delta = n \frac{h_{QTL}^2}{1 - h_{QTL}^2}$$


$$\delta = n \left(\frac{\gamma}{\sigma}\right)^2$$

# Derivation (simple linear model)

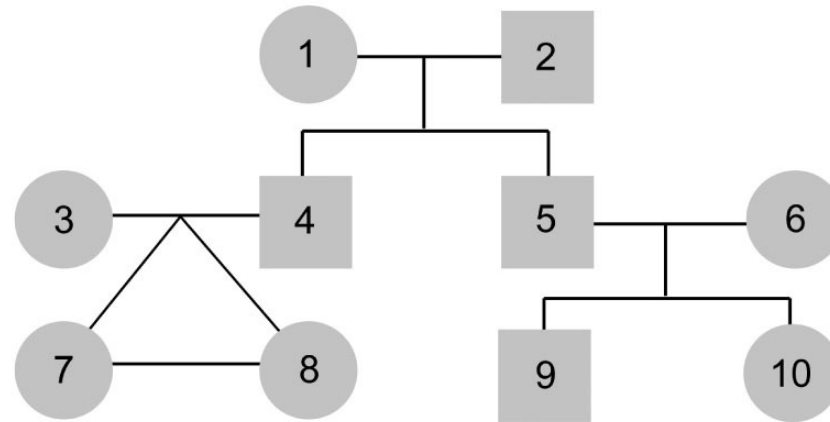
Have a look at the non-centrality parameter

$$\delta = n \frac{h_{QTL}^2}{1 - h_{QTL}^2}$$

- QTL size
  - QTL size  $\uparrow \rightarrow$  Power  $\uparrow$

# Kinship

- The probability that a random gene from subject A is identical with a gene at the same locus from subject B is the coefficient of kinship



Relationship	Example	$\phi$	$k_0$	$k_1$	$k_2$
Identical Twins	7-8	0.5000	0	0	1.00
Full-siblings	9-10	0.2500	0.25	0.50	0.25
Parent-child	1-5	0.2500	0	1.00	0
Grandparent-grandchild	1-10	0.1250	0.50	0.50	0
First cousins	8-9	0.0625	0.75	0.25	0



## A special kinship matrix

We simplify the kinship matrix so that a general trend can be found.

Assume that the kinship matrix has the following special structure

$$K = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

$\rho$  represents the correlation between any pair of individuals. Under this assumption, the eigenvalues are:

$$d_1 = n - (n - 1)(\rho - 1), d_2 = d_3 = \cdots = d_n = 1 - \rho$$

because  $\sum d_j = n$  (the sum of all eigenvalues of a correlation matrix equals to the sample size)



## A special kinship matrix

The effective sample size is defined as

$$n_0 = \sum_{j=1}^n (d_j \lambda + 1)^{-1} (\lambda + 1)$$

Effective sample size will equal the actual sample size if the polygenic variance were 0 ( $\lambda = 0$ ).

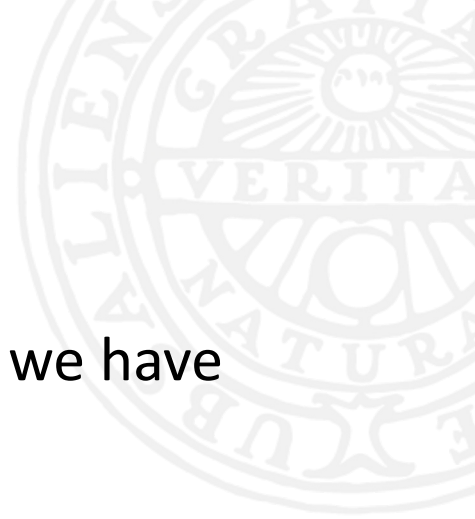
Calculate this by the eigenvalues we have

$$n_0 = (\lambda + 1) \left[ \frac{n - 1}{(1 - \rho)\lambda + 1} + \frac{1}{(1 + n\rho - \rho)\lambda + 1} \right]$$

## A special kinship matrix

Substituting the effective size into our non-centrality parameter formula, we have

$$\delta = (\lambda + 1) \left[ \frac{n - 1}{(1 - \rho)\lambda + 1} + \frac{1}{(1 + n\rho - \rho)\lambda + 1} \right] \frac{h_{QTL}^2}{1 - h_{QTL}^2}$$







# A special kinship matrix

Have a look at the non-centrality parameter under some special cases.

- If the polygenic variance approaches to nil,  $\lambda \rightarrow 0$ .

$$\lim_{\lambda \rightarrow 0} \delta = n \frac{h_{QTL}^2}{1 - h_{QTL}^2}$$

- If  $\rho \rightarrow 1$  and  $n$  is relatively large

$$\lim_{\rho \rightarrow 1} \delta = n(\lambda + 1) \frac{h_{QTL}^2}{1 - h_{QTL}^2}$$

If  $\rho \rightarrow 0$

$$\lim_{\rho \rightarrow 0} \delta = n \frac{h_{QTL}^2}{1 - h_{QTL}^2}$$

Adding the kinship matrix in GWAS boost the power by a factor  $(\lambda + 1)$

# Population structure

- Caused by population heterogeneity (or admixture) represented by multiple ethnic groups or subpopulations within the association population
- An example:
  - A locus is fixed to alleles that are unique to subpopulations
  - Subpopulations are strongly associated with the trait
  - ➔ The association observed may be caused by subpopulation (false positive)





# Population structure

Consider the Q + K mixed model for GWAS

$$y = \mu + Q\eta + Z\gamma + e$$

$Q$  is the design matrix for population structure (PCA or cluster analysis)

The non-centrality parameter for the Wald test is

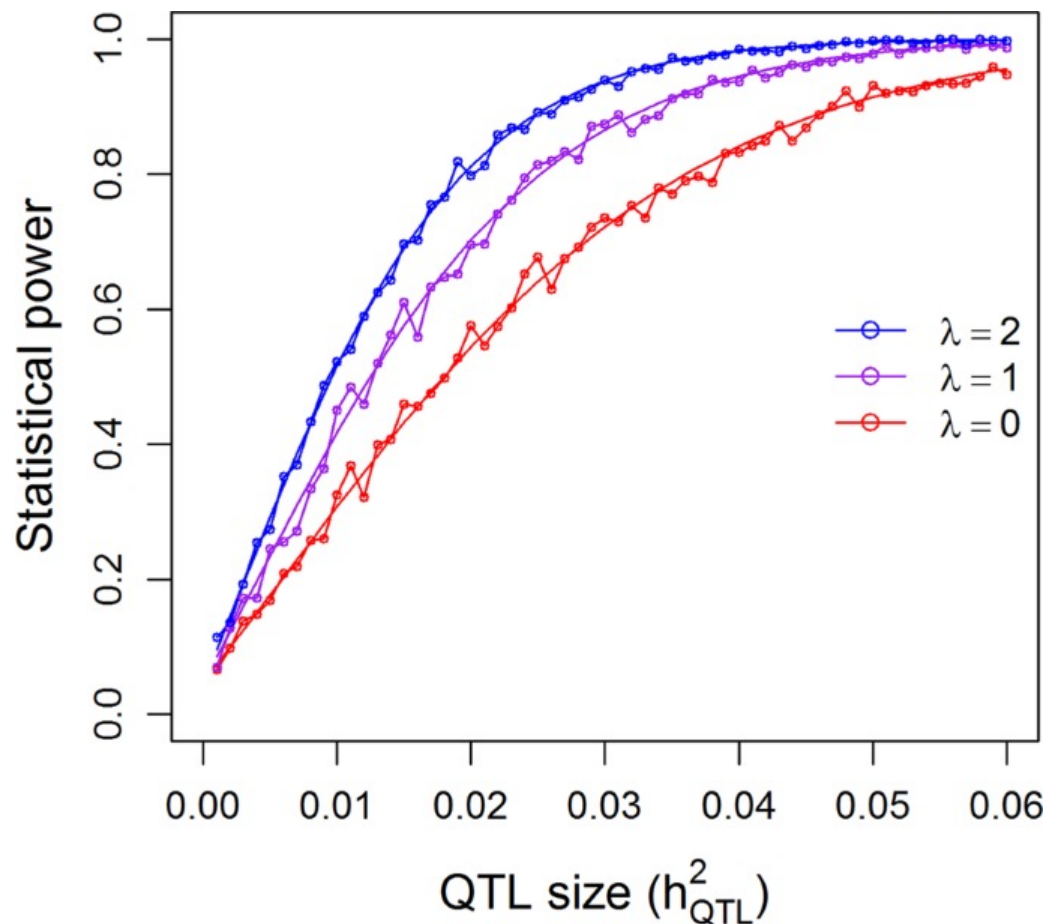
$$\delta = \frac{\gamma^2}{\text{var}(\gamma)} = n_0 \left( 1 - \sum_{i=1}^q r_{ZQ_i}^2 \right) \sigma_Z^2 \frac{h_{QTL}^2}{1 - h_{QTL}^2}$$

⇒ Population structure effects reduce the non-centrality parameter

⇒ Lower the power

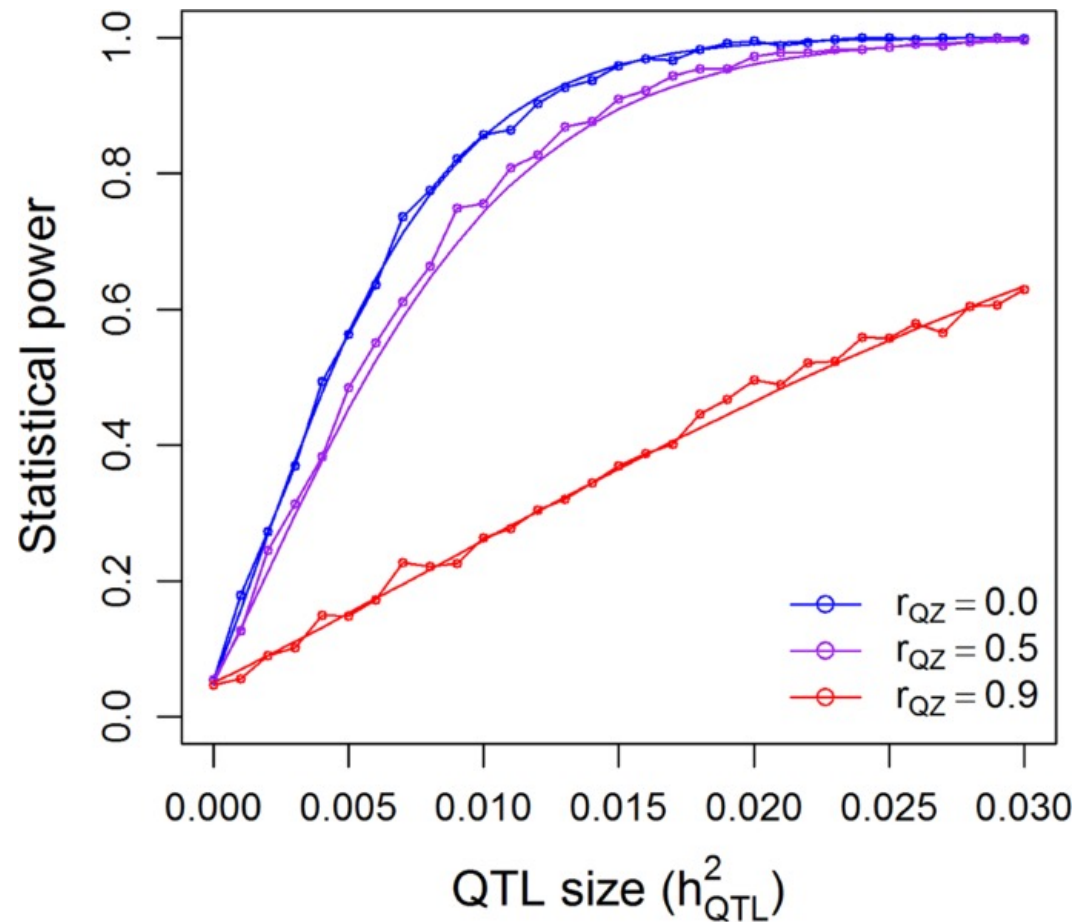
$r_{Z_i}^2$ : the squared correlation between matrix Q and Z

# Simulation results

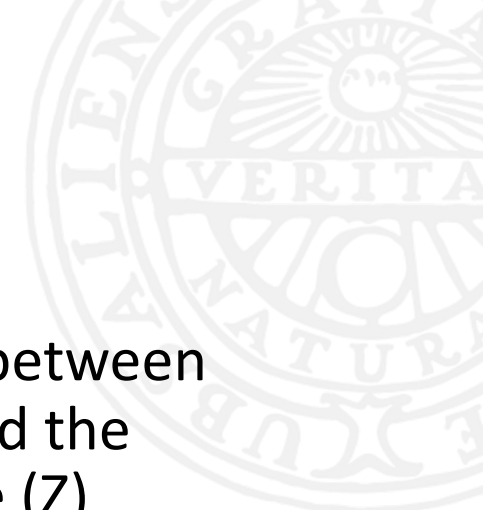


- Three levels of polygenic contribution are represented by the ratios of the polygenic variance to the residual variance.
- $h^2_{QTL} \uparrow \rightarrow \text{Power} \uparrow$

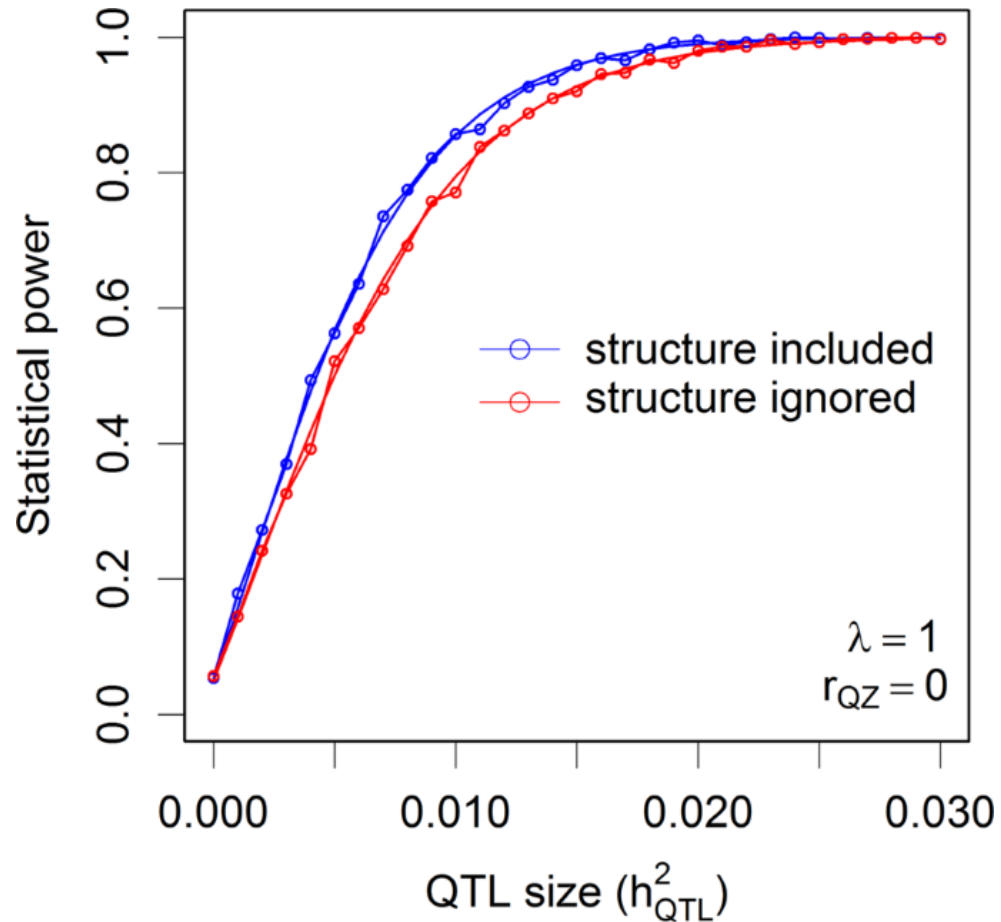
# Simulation results



- Three levels of correlation between population structure (Q) and the genotypic indicator variable (Z)



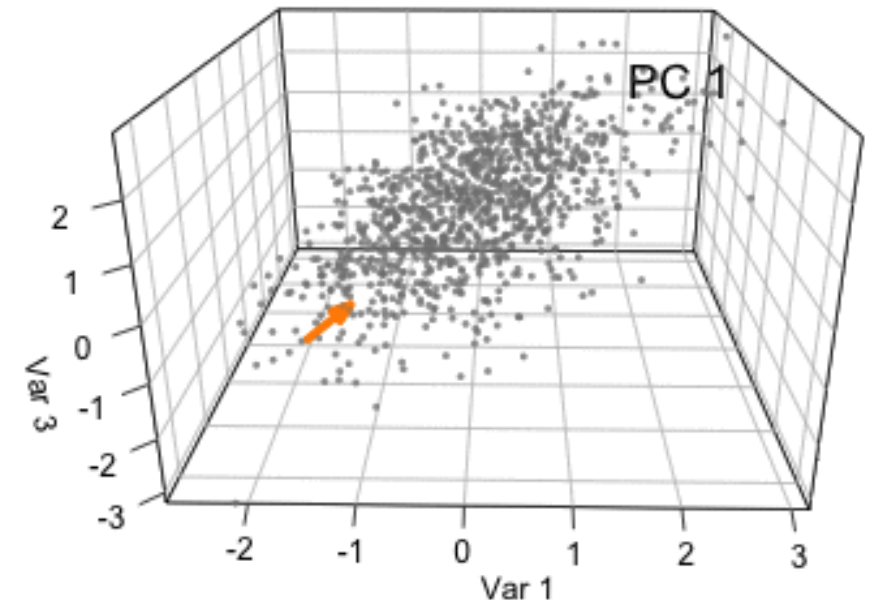
# Simulation results



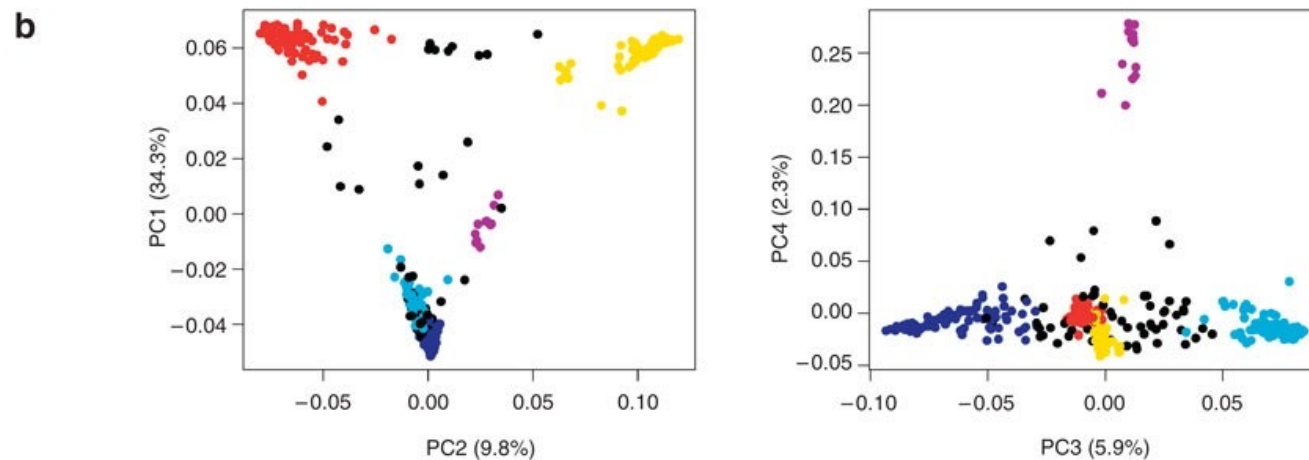
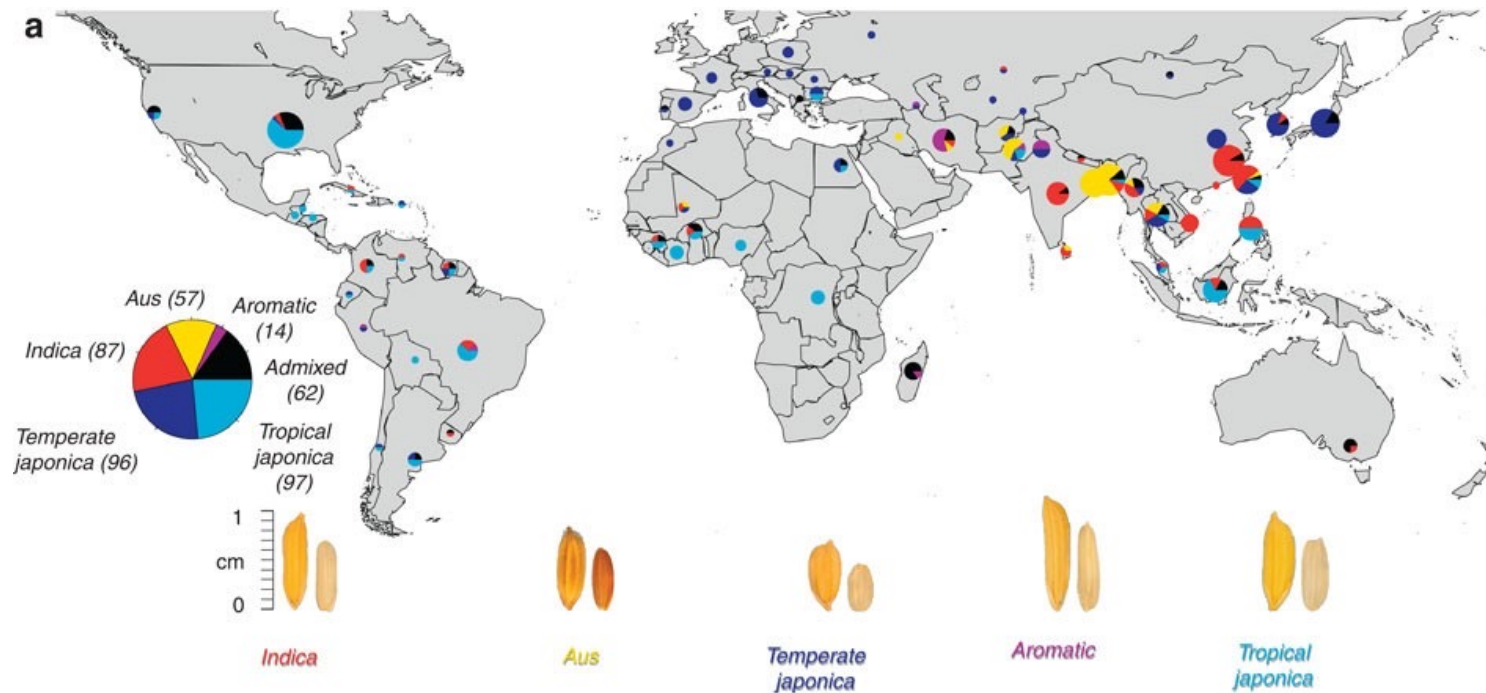
- The correlation between population structure ( $Q$ ) and the genotypic indicator variable ( $Z$ ) is 0.

# Principal component analysis (PCA)

PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance scalar projection of the data comes to lie on the first coordinate (aka the first principal component; PC1), the second greatest variance on the second coordinate (PC2), and so on.



# Rice 44k data





# Other factors affecting statistical power

- Quality of phenotype
- Quality of genotype
- Genders
- Seasons / Environments
- ...



# How could we do to increase statistical power

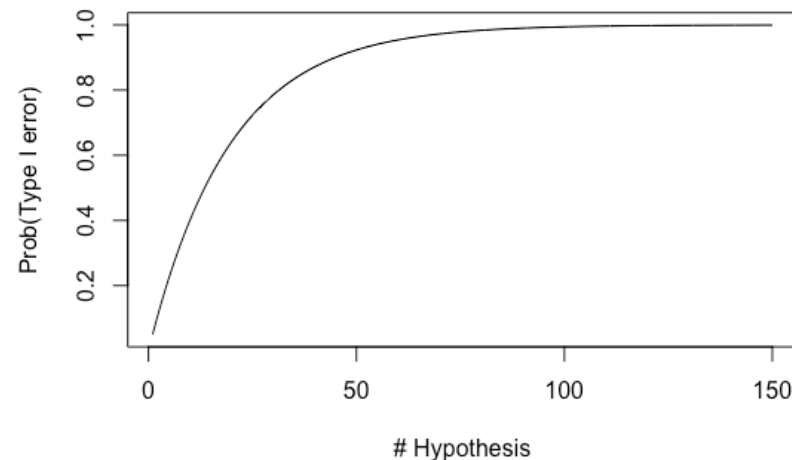
- Effect size: not much to do
- Increase rare alleles: a designed mating system can increase the frequency
- Sample size: increase sample size if it is affordable
- Population structure: a proper sampling population or better statistic model
- Experimental design





# Multiple testing problem

- We control the type I error by setting the significant level  $\alpha$   
→ The risk of making the type I error is  $\alpha$
- If we perform  $n$  hypothesis test and  $\alpha = 0.05$ .
  - The probability of not making any type I error is  $(1 - 0.05)^n$
  - The probability of making at least one type I error is  $1 - (1 - 0.05)^n$





# Multiple testing problem

## Two common solutions

- Bonferroni correction

For  $n$  tests and a desired significant level  $\alpha$ . The Bonferroni corrected threshold  $\alpha_B = \frac{\alpha}{n}$

- False discovery rate (FDR)

Use the expected number of false positives to set the significance level. For  $n$  tests and a desired significant level  $\alpha$ , the FDR is defined in terms or number of cases where the null hypothesis is rejected (R)

If  $p\text{-value}_i \leq \alpha \frac{i}{n} \rightarrow$  Significant