

# Genetics and Bioinformatics

**Kristel Van Steen, PhD<sup>2</sup>**

**Montefiore Institute - Systems and Modeling**

**GIGA - Bioinformatics**

**ULg**

**[kristel.vansteen@ulg.ac.be](mailto:kristel.vansteen@ulg.ac.be)**

## **Complicating factors in bioinformatics**

### **1 Trait heterogeneity in GWAs**

**Single traits association tests**

### **2 Confounding**

**3.a Epidemiology**

**3.b GWAs (population structure)**

### **3 Multiple testing**

**Locus heterogeneity**

### **4 Multiple studies**

**Meta-analysis**

## **5 When variants become rare – sparse data**

**Customizing GWAs for rare variants association analyses (future class)**

## **6 When effects become non-independent**

**Biological vs statistical epistasis (future class)**

# 1 Trait heterogeneity in GWAs

## The linear regression model

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon$$

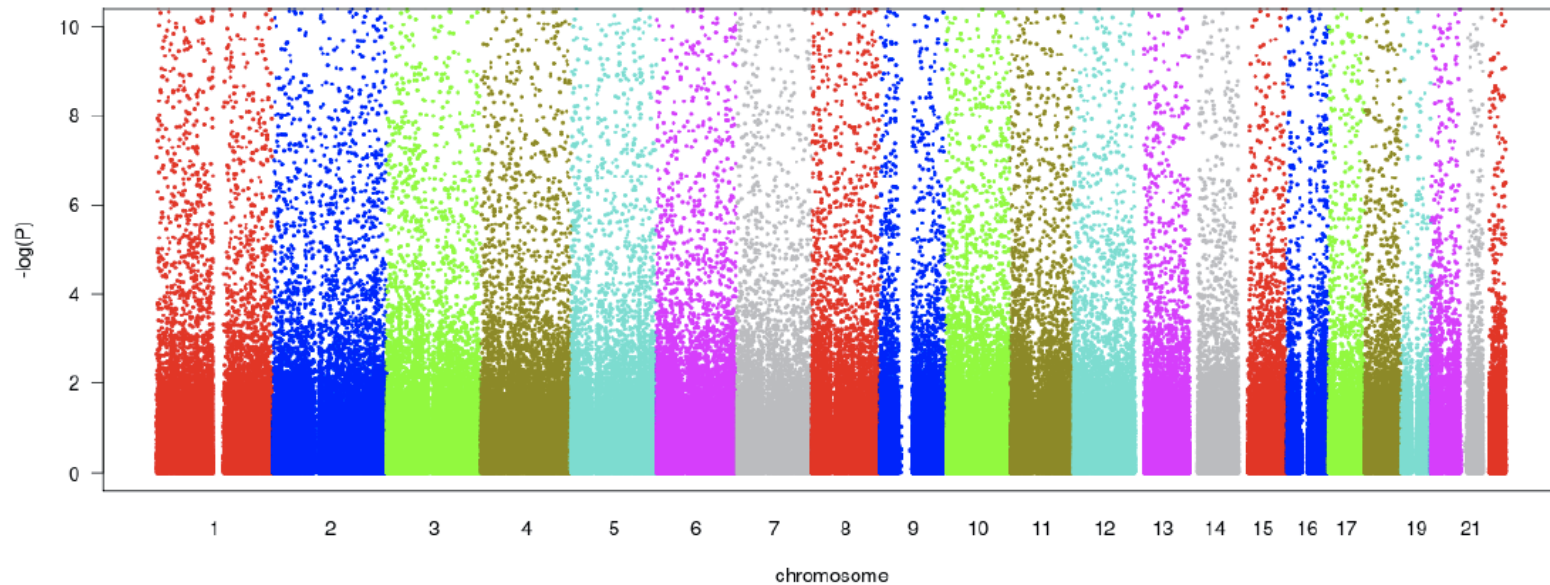
- $y$ : response variable.
- $x_1, \dots, x_k$ : regressor variables, independent variables.
- $\beta_0, \beta_1, \dots, \beta_k$ : regression coefficients.
- $\epsilon$ : model error.
  - ▶ Uncorrelated:  $\text{cov}(\epsilon_i, \epsilon_j) = 0, i \neq j$ .
  - ▶ Mean zero, Same variance:  $\text{var}(\epsilon_i) = \sigma^2$ . (homoscedasticity)
  - ▶ Normally distributed.

## **The linear regression model**

1. Input data need to be of high quality
2. The model needs to be appropriate (hence model assumptions need to be checked), before beta model parameters are estimated from the data at hand
3. The model needs to be appropriate before test results are derived/interpreted

# 1 High quality data

BEFORE QC → true signals are lost in false positive signals

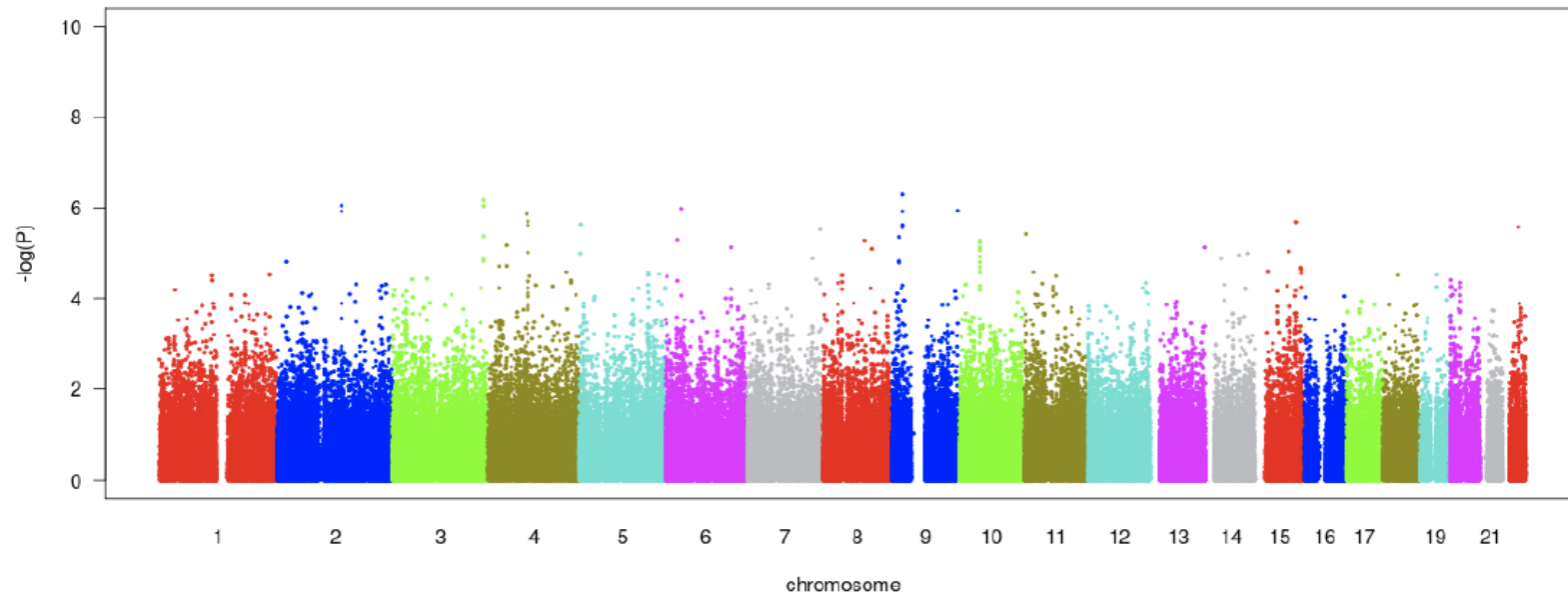


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

(Ziegler and Van Steen 2010)

## Why is quality control important?

**AFTER QC** → skyline of Manhattan (→ name of plot: Manhattan plot):



Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

SNPs passing standard quality control: 270,701

(Ziegler and Van Steen 2010)

## The Travemünde criteria

Level	Filter criterion	Standard value for filter
Sample level	Call fraction	$\geq 97\%$
	Cryptic relatedness	Study specific
	Ethnic origin	Study specific; visual inspection of principal components
	Heterozygosity	Mean $\pm$ 3 std.dev. over all samples
	Heterozygosity by gender	Mean $\pm$ 3 std.dev. within gender group
SNP level	MAF	$\geq 1\%$
	MiF	$\leq 2\%$ in any study group, e.g., in both cases and controls
	MiF by gender	$\leq 2\%$ in any gender
	HWE	$p < 10^{-4}$

(Ziegler 2009)



## The Travemünde criteria

Level	Filter criterion	Standard value for filter
SNP level	Difference between control groups	$p > 10^{-4}$ in trend test
	Gender differences among controls	$p > 10^{-4}$ in trend test
X-Chr SNPs	Missingness by gender	No standards available
	Proportion of male heterozygote calls	No standards available
	Absolute difference in call fractions for males and females	No standards available
	Gender-specific heterozygosity	No standard value available

(Ziegler 2009)

## 2 Appropriateness of the model

- There are 4 principal assumptions which justify the use of **linear regression** models for purposes of prediction:
  - **linearity** of the relationship between dependent and independent variables
  - independence of the errors (no serial correlation)
  - homoscedasticity (constant variance) of the errors
    - versus time (when time matters)
    - versus the predictions (or versus any independent variable)
  - normality of the error distribution. (<http://www.duke.edu/~rnau/testing.htm>)
- To check **model assumptions**: go to **quick-R** and regression diagnostics (<http://www.statmethods.net/stats/rdiagnostics.html>) : **role of qq plots!**

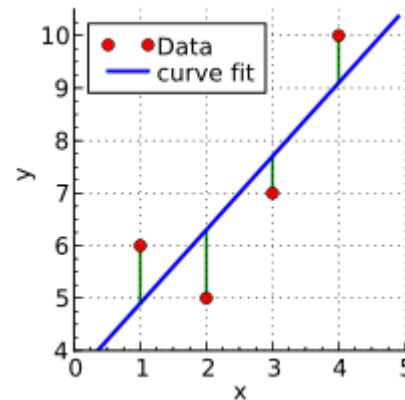
## Estimating model parameters

- A full model (continuous response, say “BMI”) may look like:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Fit the model by the **method of least squares** (this leads to estimations  $b$  for the beta parameters in the model)
- It will also lead to the **error sums of squares (SSE)**: the sum of the squared deviations of each observation  $Y$  around its estimated expected value
- The error sums of squares of the full model  $SSE(F)$ :

$$\begin{aligned} \sum [Y - b_0 - b_1 X_1 - b_2 X_2]^2 \\ = \sum (Y - \hat{Y})^2 \end{aligned}$$



## Estimating model parameters

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Use vector/matrix notations: e.g.,

$$b = (b_1, b_2)^T$$

- Least square estimation of the regression coefficients beta:

$$b = (X^T X)^{-1} X^T y$$

---

### 3 Association tests based on linear regression models

- For the full model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

we may consider the null hypothesis  $H_0$  of interest:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

- The model when  $H_0$  holds is called **the reduced or restricted model**. When  $\beta_1 = 0$ , then the regression model before reduces to

$$Y = \beta_0 + \beta_2 X_2 + \varepsilon$$

- Again we can fit this model with f.i. the least squares method and obtain an error sums of squares, now for the reduced model:  $SSE(R)$

## Association tests based on linear regression models

- By contrasting  $SSE(F)$  and  $SSE(R)$  a test statistic can be derived to test our null hypothesis

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} : \frac{SSE(F)}{df_F}$$

which follows an F distribution when  $H_0$  holds

- The decision rule (for a given alpha level of significance) is:
  - If  $F^* \leq F(1 - \alpha; df_R - df_F, df_F)$ , you cannot reject  $H_0$
  - If  $F^* > F(1 - \alpha; df_R - df_F, df_F)$ , conclude  $H_1$

## Association tests based on linear regression models

### Implications for GWAs:



(courtesy of Doug Brutlag 2010)

## Implication for GWAs:

$$Y = \beta_0 + \beta_1 SNP + \varepsilon$$

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$
- $df_F = n - 2$  (this links to df in variance estimation)
- $df_R = n - 1$  (this links to df in variance estimation)

It can be shown that for testing  $\beta_1 = 0$  versus  $\beta_1 \neq 0$

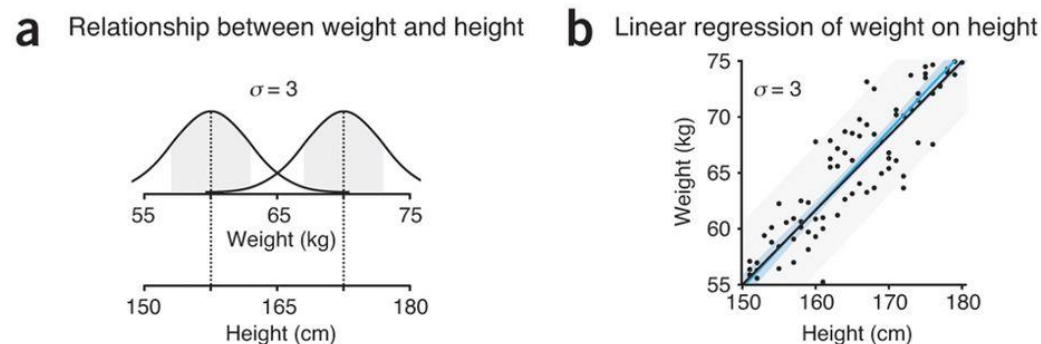
$$- F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} : \frac{SSE(F)}{df_F} = (t^*)^2$$

Note: the t-test is more flexible since it can be used for one-sided alternatives whereas the F-test cannot.

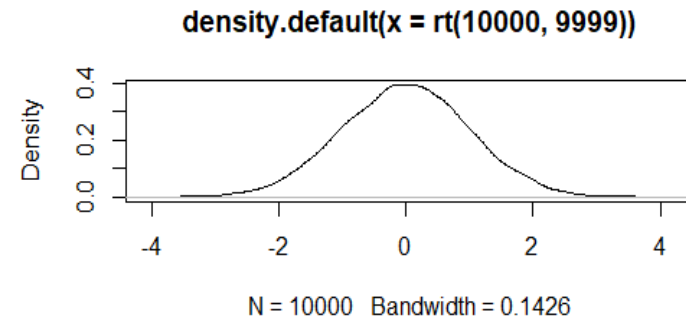
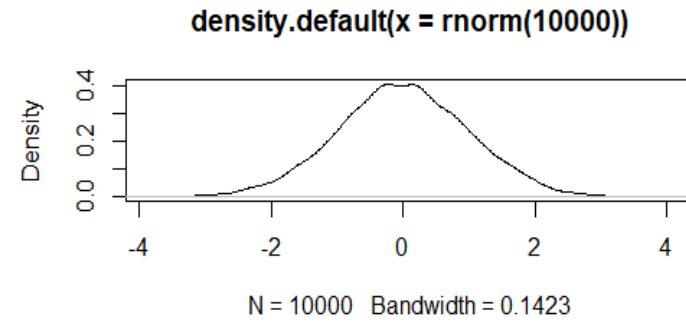
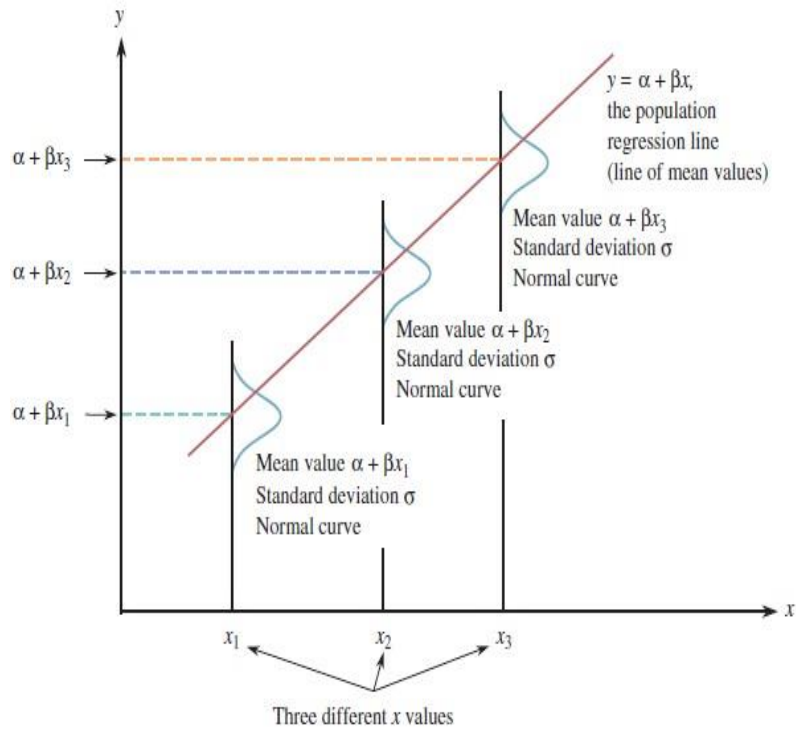


## Why to mention the relationship between the F and t (Z) statistics?

- **Analysis of variance (ANOVA)** is used to test for differences among more than two populations (“groups of samples”).  
It can be viewed as an extension of the t-test we used for testing two population means.
- In correlation, the two variables are treated as equals. In **linear regression**, one variable is considered independent (=predictor) variable (X) and the other the dependent (=outcome) variable Y



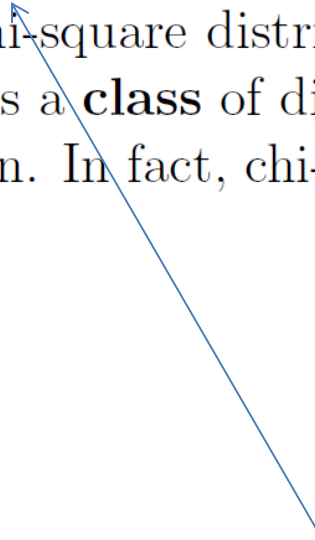
# Why to mention the relationship between the F and t statistics?



## Distributional relationships: F, t, chi-squared

$$Z_1, Z_2, \dots, Z_\kappa \text{ iid } N(0,1) \Rightarrow X^2 \equiv Z_1^2 + Z_2^2 + \dots + Z_\kappa^2 \sim \chi_\kappa^2.$$

Specifically, if  $\kappa = 1$ ,  $Z^2 \sim \chi_1^2$ . The density function of chi-square distribution will not be pursued here. We only note that: Chi-square is a **class** of distribution indexed by its *degree of freedom*, like the *t*-distribution. In fact, chi-square has a relation with *t*. We will show this later.

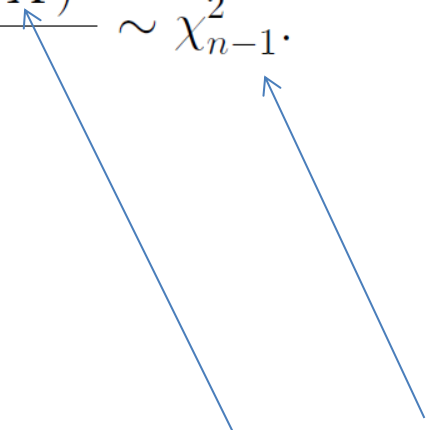


## Distributional relationships: F, t, chi-squared

If  $X_1, \dots, X_n$  iid  $N(\mu, \sigma^2)$ , then  $Z_j \equiv (X_j - \mu)/\sigma \sim N(0, 1), j = 1, \dots, n$ . We know, from a previous context, that  $\sum_1^n Z_j^2 \sim \chi_n^2$ , or equivalently,

$$\sum_{j=1}^n \left\{ \frac{X_j - \mu}{\sigma} \right\}^2 = \frac{\sum_1^n (X_j - \mu)^2}{\sigma^2} \sim \chi_n^2,$$

if  $\mu$  is *known*, or otherwise (if  $\mu$  is unknown)  $\mu$  needs to be estimated (by  $\bar{X}$ , say,) such that

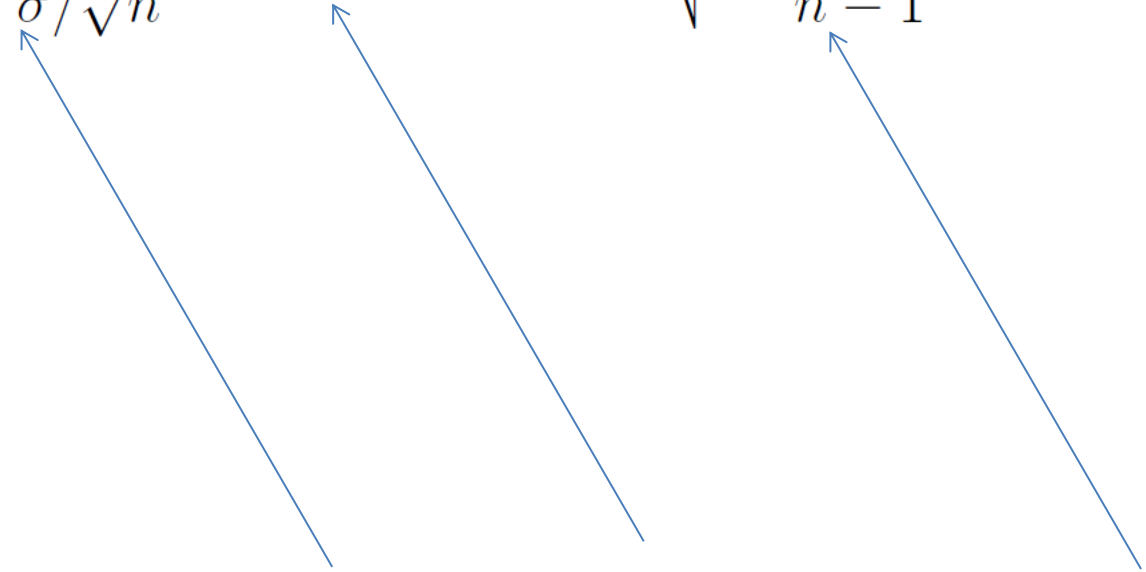
$$\frac{\sum_1^n (X_j - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2.$$


## Distributional relationships: F, t, chi-squared

If  $X_1, \dots, X_n$  iid  $N(\mu, \sigma^2)$ , then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

When  $\sigma$  is unknown,

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}, \text{ where } \hat{\sigma} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}.$$


Note that

$$\begin{aligned}\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} &= \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{1}{\frac{\hat{\sigma}}{\sigma}} \\ &= Z \cdot \frac{1}{\frac{\hat{\sigma}}{\sigma}} \\ &= \frac{Z}{\frac{\sqrt{\sum(X_i - \bar{X})^2}}{(n-1)\sigma^2}} \\ &= \frac{Z}{\sqrt{\frac{\chi_{n-1}^2}{n-1}}}.\end{aligned}$$

Combining (3) and (4) gives

$$t_{n-1} = \frac{Z}{\sqrt{\frac{\chi_{n-1}^2}{n-1}}},$$

or, in general,

$$t_{\kappa} = \frac{Z}{\sqrt{\frac{\chi_{\kappa}^2}{\kappa}}}.$$

## Distributional relationships: F, t, chi-squared

Can you see why our  $F^* = (t^*)^2$

$$F_{a,b} \equiv \frac{\chi_a^2/a}{\chi_b^2/b} \text{ (Sir R. A. Fisher).}$$

$$\begin{aligned} t_\nu &= \frac{Z}{\sqrt{\chi_\nu^2/\nu}} \\ &= \frac{\sqrt{\chi_1^2/1}}{\sqrt{\chi_\nu^2/\nu}} \\ &= \sqrt{F_{1,\nu}}. \end{aligned}$$

## Distributional relationships: F, t, chi-squared

$$F_{a,b} \equiv \frac{\chi_a^2/a}{\chi_b^2/b} \text{ (Sir R. A. Fisher).}$$

Can you see how the formula above relates to “taking a ratios of variances”?

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} : \frac{SSE(F)}{df_F}$$

$$\frac{SSE(R) - SSE(F)}{df_R - df_F} = SST / (df_R - df_F),$$

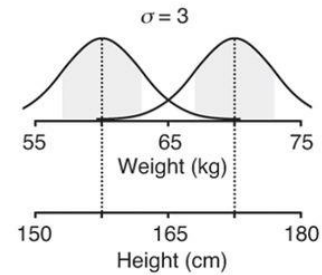
with SST referring to the **treatment sums of squares** (“**between**” source of variation in contrast to errors “**within**” source of variation)



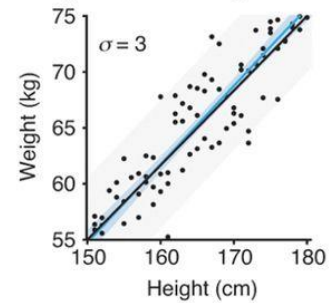
## Distributional relationships: F, t, chi-squared

Can you see why our  $F^* = \frac{b_1^2}{s^2(b_1)}$  ?

**a** Relationship between weight and height



**b** Linear regression of weight on height



## Logistic regression

*Variables:*

- Let  $Y$  be a binary response variable  
 $Y_i = 1$  if the trait is present in observation (person, unit, etc...)  $i$   
 $Y_i = 0$  if the trait is NOT present in observation  $i$
- $X = (X_1, X_2, \dots, X_k)$  be a set of explanatory variables which can be discrete, continuous, or a combination.  $x_i$  is the observed value of the explanatory variables for observation  $i$ . In this section of the notes, we focus on a single variable  $X$ .

*Model:*

$$\pi_i = Pr(Y_i = 1 | X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

Since

$$E[Y|X] = \text{Prob}(Y = 1|X) = \frac{\exp(\eta)}{(1+\exp(\eta))}$$

we have

$$\frac{\text{Prob}(Y = 1|X)}{1 - \text{Prob}(Y = 1|X)} = \exp(\eta)$$

and thus

$$\mathbf{g}(E[Y|X]) = \beta_0 + \beta_1 X = \log\left(\frac{\text{Prob}(Y = 1|X)}{1 - \text{Prob}(Y = 1|X)}\right) = \mathbf{\eta}$$

(g is called **the logit link function**)

## Appropriateness of the model

### *Assumptions:*

- The data  $Y_1, Y_2, \dots, Y_n$  are independently distributed, i.e., cases are independent.
- Distribution of  $Y_i$  is  $Bin(n_i, \pi_i)$ , i.e., binary logistic regression model assumes binomial distribution of the response. The dependent variable does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal,...)
- Does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the logit of the response and the explanatory variables;  $logit(\pi) = \beta_0 + \beta X$ .
- Independent (explanatory) variables can be even the power terms or some other nonlinear transformations of the original independent variables.
- The homogeneity of variance does NOT need to be satisfied. In fact, it is not even possible in many cases given the model structure.
- Errors need to be independent but NOT normally distributed.
- It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.
- Goodness-of-fit measures rely on sufficiently large samples, where a heuristic rule is that not more than 20% of the expected cells counts are less than 5.

(<https://onlinecourses.science.psu.edu/stat504>)

## Model fitting in logistic regression

- In **standard linear models** we estimate the parameters by minimizing the sum of the squared residuals
- This is equivalent to finding parameters that **maximize the likelihood**
- In a **logistic regression** we can also fit parameters by maximizing the likelihood

The *maximum likelihood estimator* (MLE) for  $(\beta_0, \beta_1)$  is obtained by finding  $(\hat{\beta}_0, \hat{\beta}_1)$  that maximizes:

$$L(\beta_0, \beta_1) = \prod_{i=1}^N \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} = \prod_{i=1}^N \frac{\exp\{y_i(\beta_0 + \beta_1 x_i)\}}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

In general, there are no closed-form solutions, so the ML estimates are obtained by using iterative algorithms such as *Newton-Raphson* (NR), or *Iteratively re-weighted least squares* (IRWLS). In Agresti (2013), see section 4.6.1 for GLMs, and for logistic regression

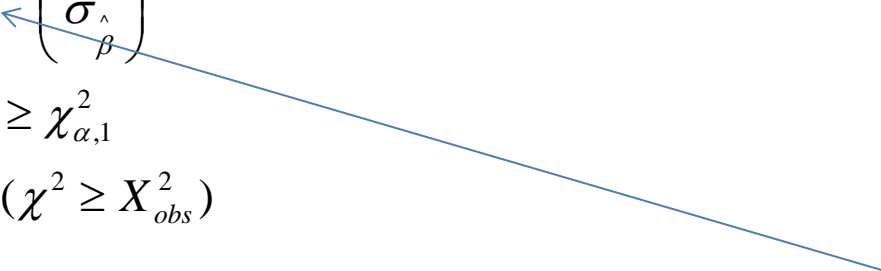
## Association tests based on logistic regression model

- Example:

$$\text{Logit}(P(Y = 1)|SNP) = \beta_0 + \beta_1 SNP$$

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

Large-sample “Wald test”:

$$T.S.: X_{obs}^2 = \left( \frac{\hat{\beta}}{\hat{\sigma}_{\beta}} \right)^2$$


$$R.R.: X_{obs}^2 \geq \chi_{\alpha,1}^2$$

$$P\text{-val} : P(\chi^2 \geq X_{obs}^2)$$

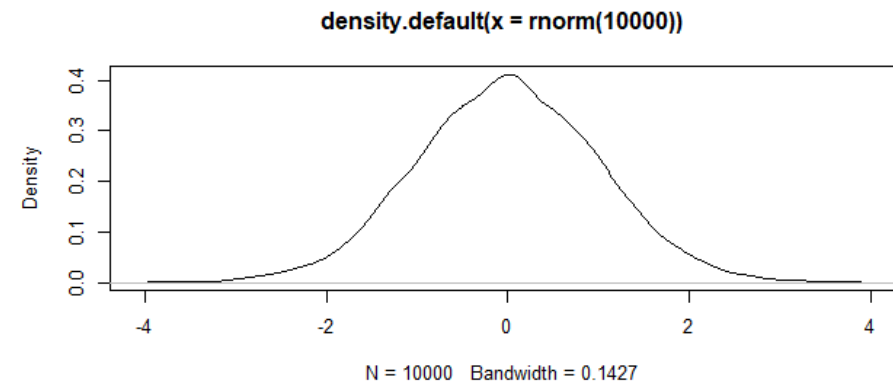
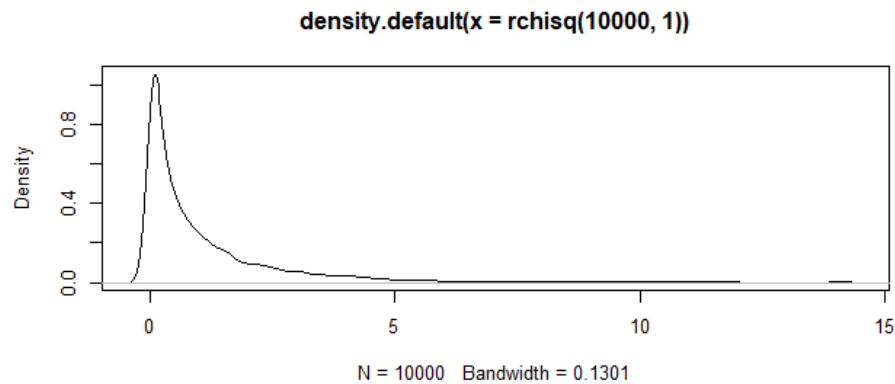
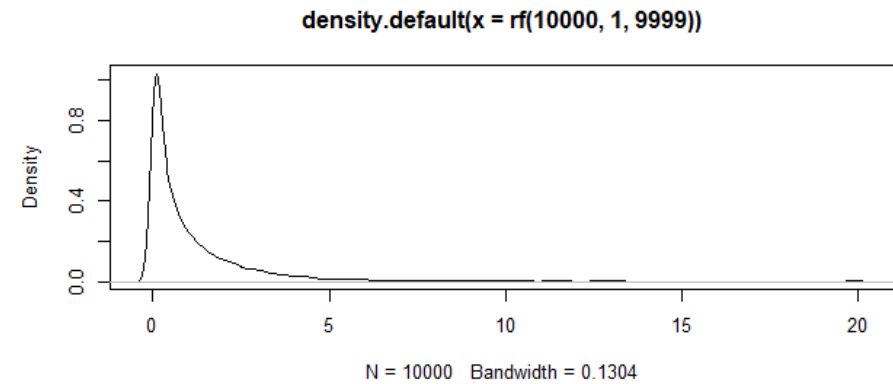
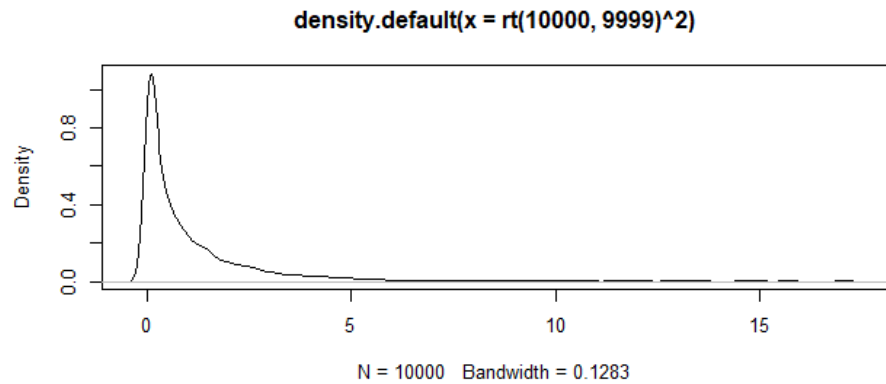
**Didn't we say that T.S. below was distributed like t-squared?**

$$T.S.: X_{obs}^2 = \left( \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \right)^2$$

$$R.R.: X_{obs}^2 \geq \chi_{\alpha,1}^2$$

$$P\text{-val}: P(\chi^2 \geq X_{obs}^2)$$

## The beauty of working with large sample statistics



```
> par(mfrow=c(2,2))  
> plot(density(rt(10000,9999)^2))  
> plot(density(rf(10000,1,9999)))
```

```
> plot(density(rchisq(10000,1)))  
> plot(density(rnorm(10000)))
```



## The Wald test statistic

- In the univariate case, the Wald test statistic is

$$\frac{(\hat{\theta} - \theta_0)^2}{\text{var}(\hat{\theta})}$$

which is **compared against a chi-squared distribution**.

- The Wald test statistic is almost but not exactly equal to the **square of the t-test statistic**, but they are asymptotically equivalent when  $n \rightarrow \infty$ .

$$\frac{b_1^2}{s^2(b_1)} = (t^*)^2$$

## The Wald test statistic

Note aside:

Alternatively, the difference can be compared to a normal distribution. In this case the test statistic is

$$\frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})}$$

where  $\text{se}(\hat{\theta})$  is the standard error of the maximum likelihood estimate (MLE). A reasonable estimate of the standard error for the MLE can be given by

$\frac{1}{\sqrt{I_n(MLE)}}$ , where  $I_n$  is the Fisher information of the parameter.

## Link between Wald and test of independence

- The **chi-square test of independence** is appropriate when the following conditions are met:
  - The sampling method is simple random sampling.
  - The variables under study are each categorical.
  - If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.
- There are four steps involved: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

## State the Hypotheses

- Suppose that Variable A has  $r$  levels, and Variable B has  $c$  levels. The null hypothesis states that knowing the level of Variable A does not help you predict the level of Variable B. That is, the variables are independent.

$H_0$ : Variable A and Variable B are independent.

$H_a$ : Variable A and Variable B are not independent.

- The alternative hypothesis is that knowing the level of Variable A **can** help you predict the level of Variable B.

**Note:** Support for the alternative hypothesis suggests that the variables are related; but the relationship is not necessarily causal, in the sense that one variable "causes" the other.

## Formulate an Analysis Plan

- The analysis plan describes how to use sample data to reject or not the null hypothesis. The plan specifies the following elements:
  - Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
  - Test method. Use the chi-square test for independence to determine whether there is a significant relationship between two categorical variables.
- Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the P-value associated with the test statistic.

- **Degrees of freedom.** The degrees of freedom (DF) is equal to:

$$DF = (r - 1) * (c - 1)$$

where r is the number of levels for one categorical variable, and c is the number of levels for the other categorical variable.

	<b>AA</b>	<b>Aa</b>	<b>aa</b>
<b>Cases</b>			
<b>Controls</b>			

Sum of entries =  
cases+controls

For example: r=2 (for a dichotomous Y) ; c=3 (for a SNP)

- **Expected frequencies.** The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute  $r * c$  expected frequencies, according to the following formula.

$$E_{r,c} = (n_r * n_c) / n$$

where  $E_{r,c}$  is the expected frequency count for level  $r$  of Variable A and level  $c$  of Variable B,  $n_r$  is the total number of observations at level  $r$  of Variable A,  $n_c$  is the total number of observations at level  $c$  of Variable B, and  $n$  is the total sample size.

	<b>AA</b>	<b>Aa</b>	<b>aa</b>
<b>Cases</b>	$E_{11}$	$E_{12}$	$E_{13}$
<b>Controls</b>	$E_{21}$	$E_{22}$	$E_{23}$

- **Test statistic.** The test statistic is a chi-square random variable ( $X^2$ ) defined by the following equation.

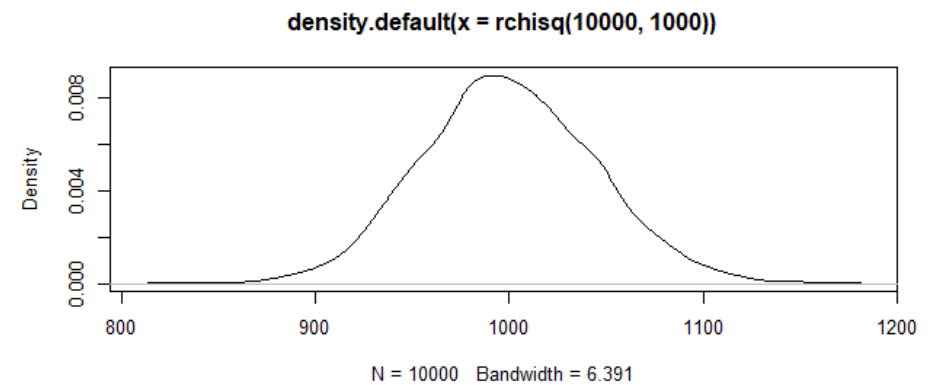
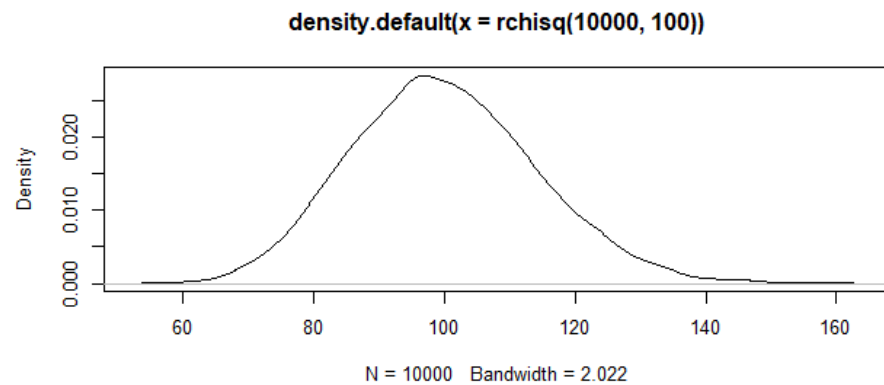
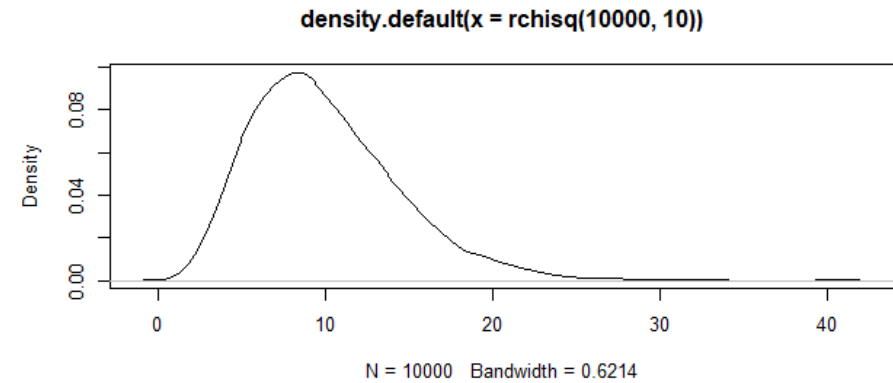
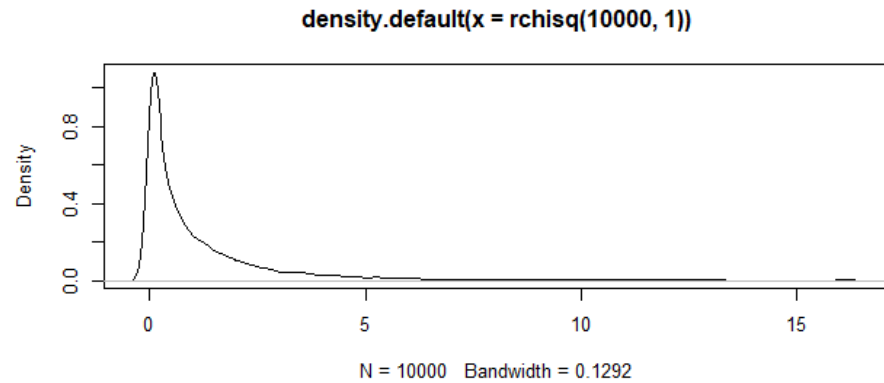
$$X^2 = \sum [ (O_{r,c} - E_{r,c})^2 / E_{r,c} ]$$

where  $O_{r,c}$  is the observed frequency count at level **r** of Variable A and level **c** of Variable B, and  $E_{r,c}$  is the expected frequency count at level **r** of Variable A and level **c** of Variable B.

- **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic, which can be proven to follow a chi-square distribution with degrees of freedom as derived before. The null hypothesis is rejected when the P-value is less than the pre-stated significance level (e.g., 0.05 or  $0.05/(\text{nr of SNPs to test})$ ).

(see <http://stattrek.com/chi-square-test> for a general example)





```
> par(mfrow=c(2,2))  
> plot(density(rchisq(10000,1)))  
> plot(density(rchisq(10000,10)))  
> plot(density(rchisq(10000,100)))  
> plot(density(rchisq(10000,1000)))
```

## Trait heterogeneity

- Trait heterogeneity exists when a trait has been defined with insufficient specificity such that it is actually two or more distinct traits
- Other forms of heterogeneity, complicating GWAs, exist as well:
  - In the case of **locus heterogeneity**, multiple predictor variables (i.e., multiple loci) are present, some of which may be unmeasured or unobserved and, therefore, unavailable for inclusion in the disease model.
  - Epistasis: **Gene-gene interactions** create a rugged model landscape for statistical analysis. There is clear and convincing evidence that gene-gene interactions, whether synergistic or antagonistic, are not only possible but probably omnipresent
- All of these forms may co-exist and severely hamper classical one-at-a-time SNP testing in GWAs.

	<b>Locus Heterogeneity</b>	<b>Trait Heterogeneity</b>	<b>Gene-Gene Interaction</b>
<b>Definition</b>	when two or more DNA variations in distinct genetic loci are independently associated with the same trait	when a trait, or disease, has been defined with insufficient specificity such that it is actually two or more distinct underlying traits	when two or more DNA variations interact either directly (DNA-DNA or DNA-mRNA interactions), to change transcription or translation levels, or indirectly by way of their protein products, to alter disease risk separate from their independent effects
<b>Diagram</b>			
<b>Example One</b>	<b>Retinitis Pigmentosa</b> (RP, OMIM# 268000) - genetic variations in at least fifteen genes have been associated with RP under an autosomal recessive model. Still more have been associated with RP under autosomal dominant and X-linked disease models <sup>2</sup> ( <a href="http://www.sph.uth.tmc.edu/RetNet">http://www.sph.uth.tmc.edu/RetNet</a> )	<b>Autosomal Dominant Cerebellar Ataxia</b> (ADCA, OMIM# 164500) - originally described as a single disease, three different clinical subtypes have been defined based on variable associated symptoms, <sup>6,7</sup> and different genetic loci have been associated with the different subtypes <sup>8</sup>	<b>Hirschsprung Disease</b> (OMIM# 142623) - variants in the RET (OMIM# 164761) and EDNRB (OMIM# 131244) genes have been shown to interact synergistically such that they increase disease risk far beyond the combined risk of the independent variants <sup>12</sup>
<b>Example Two</b>	<b>Tuberous Sclerosis</b> (TS, OMIM# 191100) - out of families informative for linkage analysis, half have mutations in the TSC1 gene (located at 9q34) and the other half have mutations in the TSC2 gene (located at 16p13) <sup>3,4,5</sup>	<b>Autism</b> (OMIM# 209850) - parents and other relatives of autistic individuals often exhibit one or two, but not all three, of the requisite autistic symptomatologies, suggesting autism may be the co-occurrence of three distinct traits. <sup>9</sup> Using subset analysis, some success has been achieved identifying genes associated with one of the three symptomatologies but not as strongly with the broader autistic phenotype <sup>10,11</sup>	<b>Creutzfeldt-Jakob Disease</b> (CJD, OMIM# 123400) and Fatal Familial Insomnia (OMIM# 176640.0010) - the Met129Val polymorphism and Asp178Asn mutation in the PRNP gene (OMIM# 176640) interact, such that when the val129 polymorphism is on the same chromosome as the asn178, the phenotype is fatal familial insomnia <sup>13-19</sup>

(Thornton-Wells et al. 2006)

## Trait heterogeneity

- Why is it a concern in GWAs?

It has been implicated as a **confounding factor** in traditional statistical genetics of complex human disease.

- How can you get a handle of trait heterogeneity?

In the absence of detailed phenotypic data collected consistently in combination with genetic data, unsupervised computational methodologies offer the potential for discovering underlying trait heterogeneity: “clustering” (see also Thornton-Wells et al. 2006)

## 2 Confounding

### 2.a Epidemiology ([www.dorak.info/epi](http://www.dorak.info/epi))

#### What is bias?

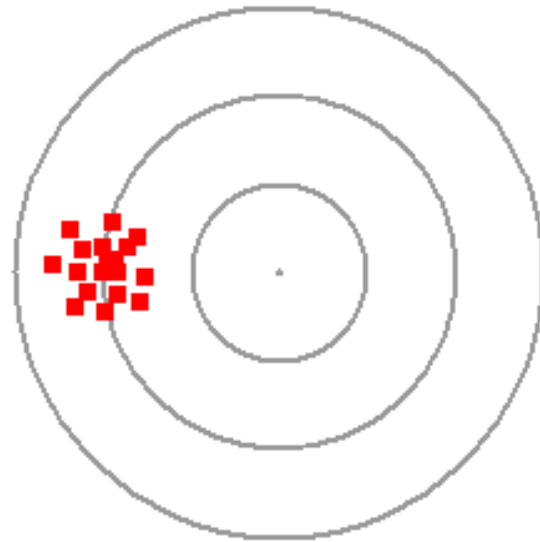
- Any trend in the collection, analysis, interpretation, publication or review of data that can lead to conclusions that are systematically different from the truth (Last, 2001)
- A process at any state of inference tending to produce results that depart systematically from the true values (Fletcher et al, 1988)
- Systematic error in design or conduct of a study (Szklo et al, 2000)

## What is bias?

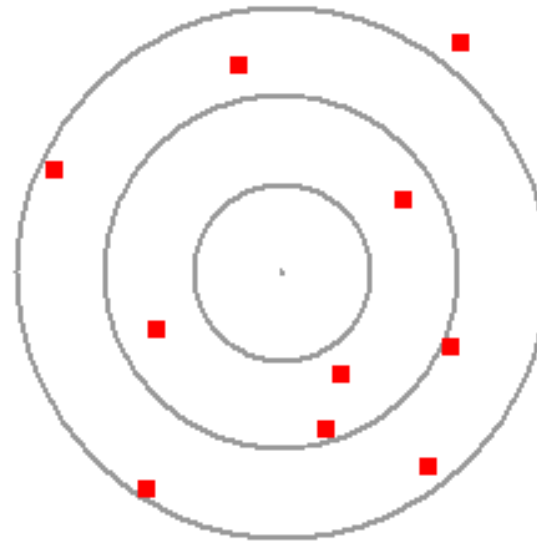
### Bias is systematic error

- Errors can be differential (systematic) or non-differential (random).  
The term “bias” should only be reserved for systematic errors
- **Random errors:** are statistical fluctuations (in either direction) in the measured data due to the precision limitations of the measurement device. Random errors usually result from the experimenter's inability to take the same measurement in exactly the same way to get exact the same number
- **Differential errors:** are reproducible inaccuracies that are consistently in the same direction. Systematic errors are often due to a problem which persists throughout the entire experiment.

## What is random error and what is systematic error?



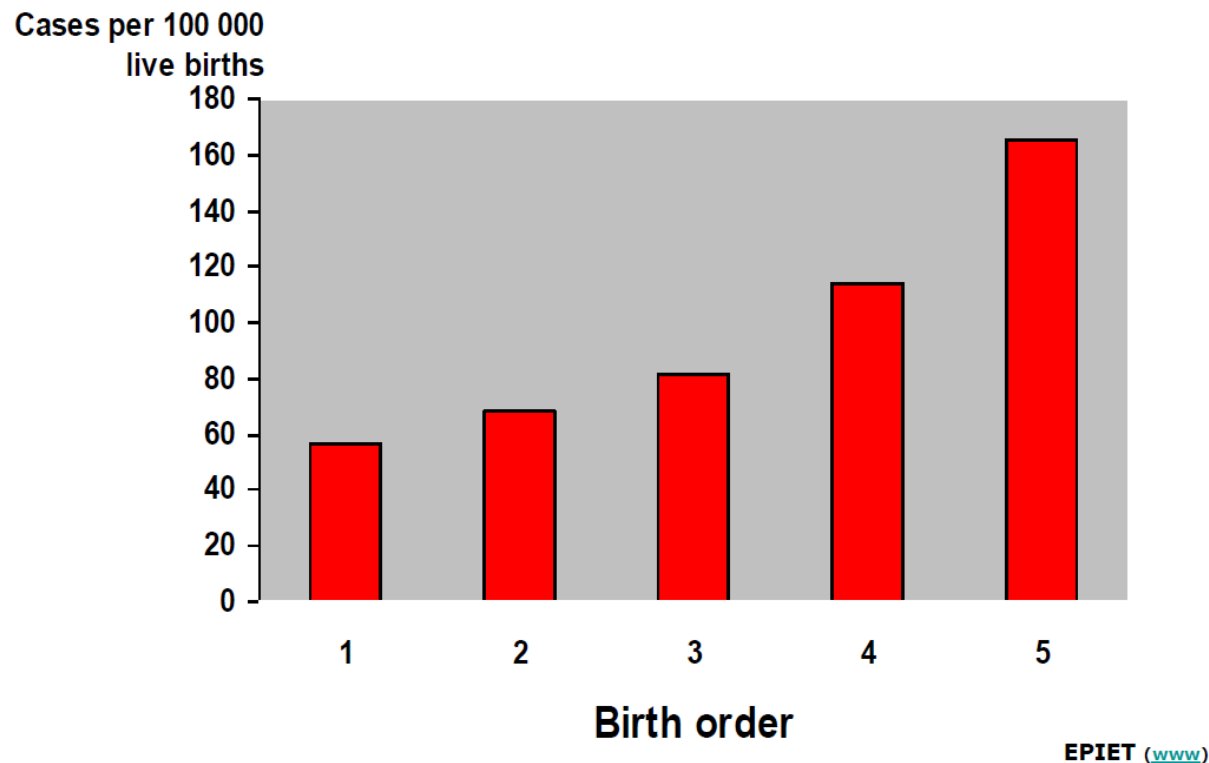
**Systematic Error**



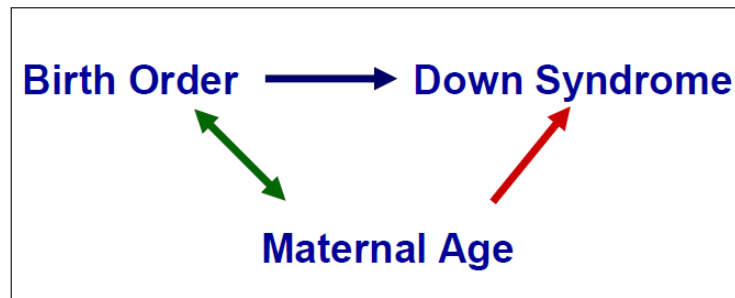
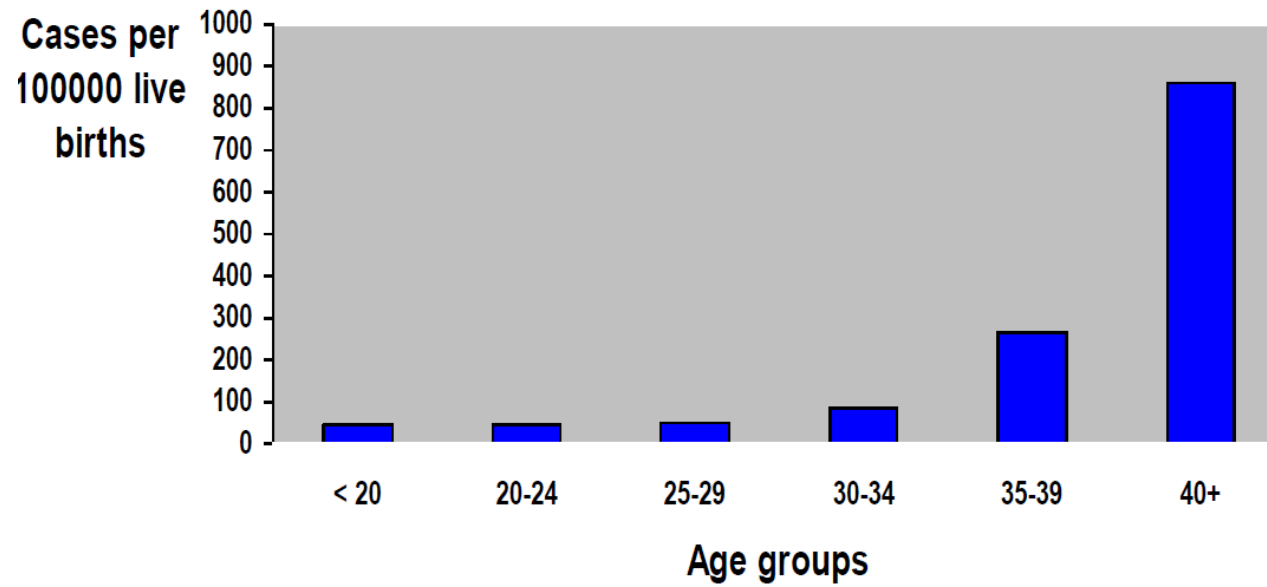
**Random Error**

## Confounding - by example

Apart from random error and systematic error (bias), another trouble maker in epidemiology is confounding





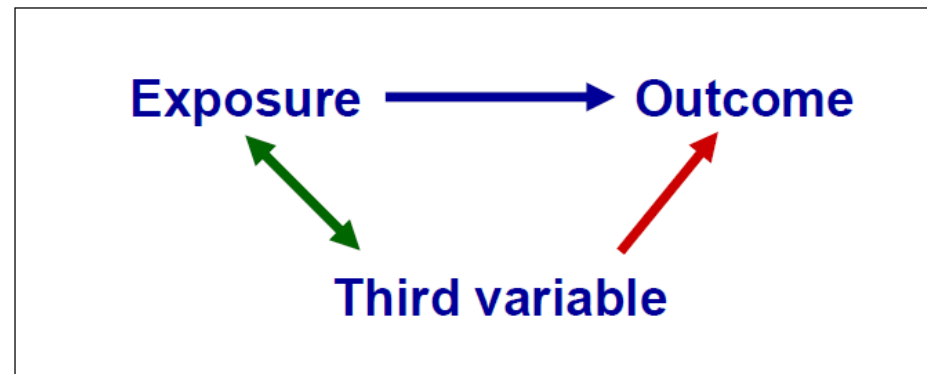


## Confounding - definition

- A third factor which is related to both exposure and outcome, and which accounts for some/all of the observed relationship between the two
- Confounder not a result of the exposure
  - e.g., association between child's birth rank (exposure) and Down syndrome (outcome); mother's age a confounder?
  - e.g., association between mother's age (exposure) and Down syndrome (outcome); birth rank a confounder?

## Confounding – definition

To be a confounding factor, two conditions must be met:



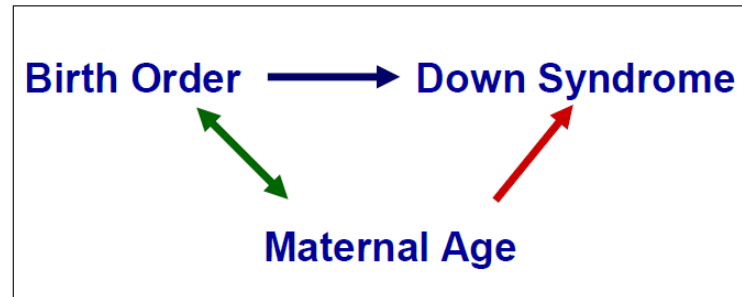
**Be associated with exposure**

- without being the consequence of exposure

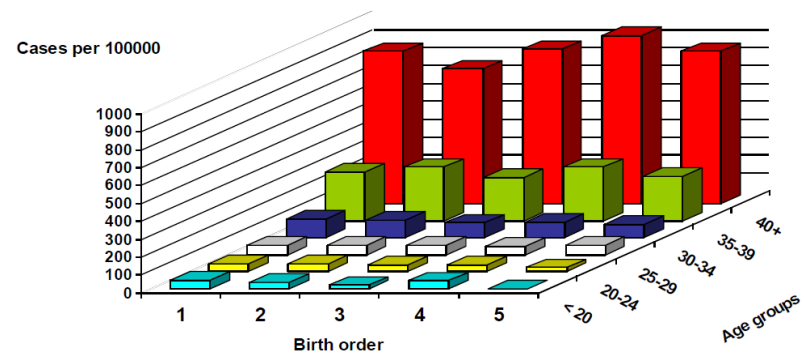
**Be associated with outcome**

- independently of exposure (not an intermediary)

# Confounding – definition



**Maternal age is correlated with birth order and a risk factor even if birth order is low**

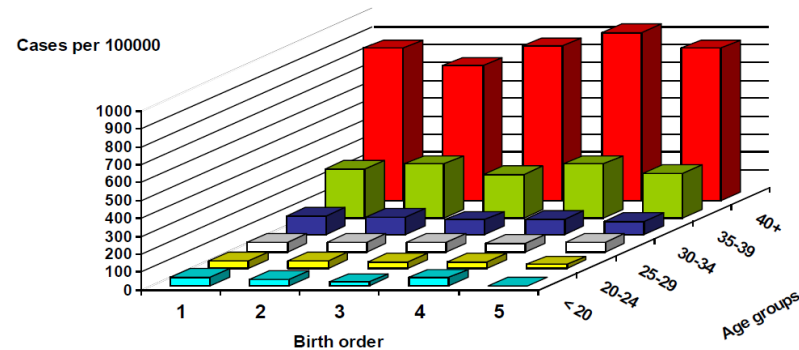


## Confounding – practical consequences

- Imagine you have repeated a positive finding of birth order association in Down syndrome. Would you be able to replicate it? If not why?
  - Spurious association? refers to false positive association result due to not having acknowledged the confounding factors in the analysis
  - What if a new sample only involved mothers below the age of 30?
- Two ways of handling/identifying confounders **during the analysis phase** (i.e., not the study design phase - randomization) are
  - performing a stratified analysis (e.g., subgroups of maternal age → power issues due to reduced sample size)
  - **adjustment by multivariable modelling**

## Confounding – practical consequences

- So, if analysis is repeated after stratification by age (during analysis), there will be no association with birth order.
- Sometimes confounding can be handled **during the design of the study:**



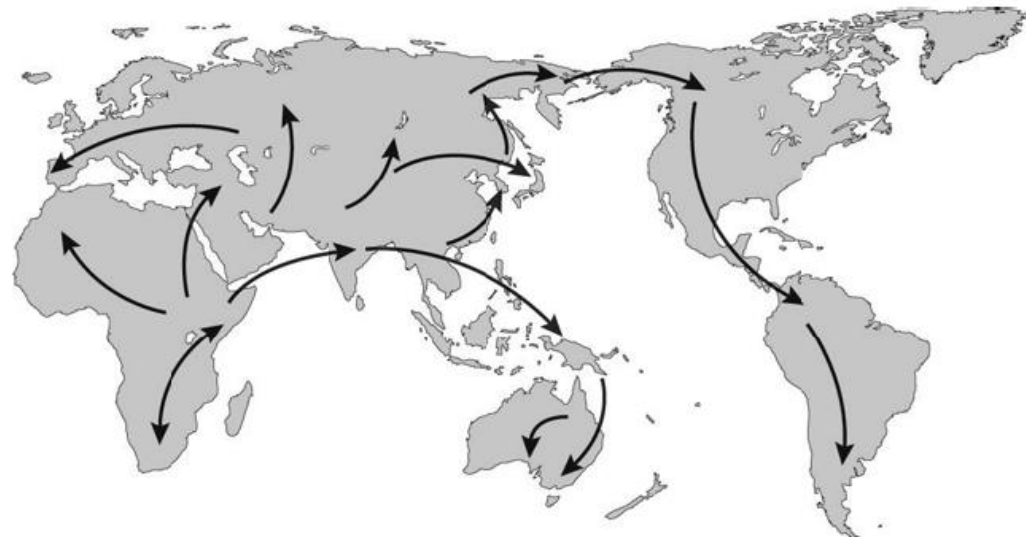
- If each case is matched with a same-age control, there will be no association.

## Confounding – practical consequences

- **Matching** is indeed another way of achieving what we want – not being hampered by confounding. It ensures equal representation of subjects with known confounders in study groups. It has to be coupled with matched analysis.
- In contrast, **randomisation** is an attempt to evenly distribute potential (unknown) confounders in study groups. It does not guarantee total control of confounding.

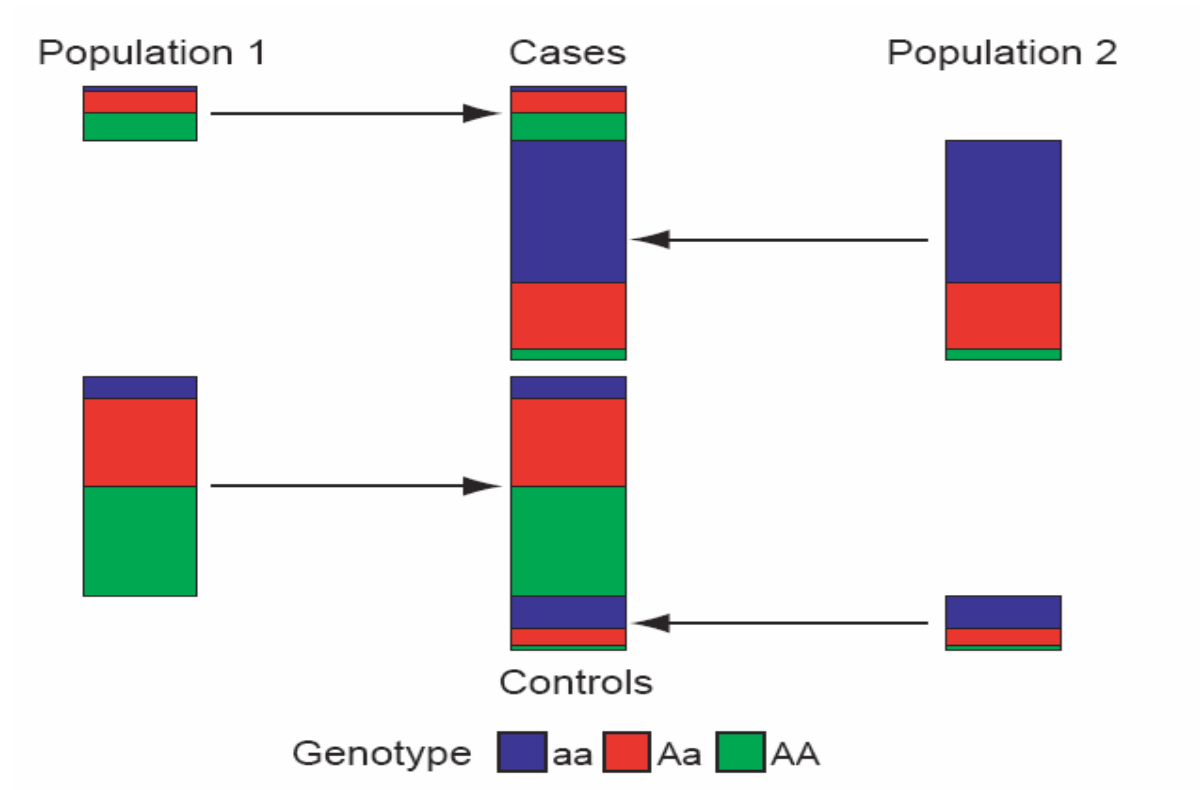
## Confounding – population structure

- Humans originally spread across the world many thousands years ago
- Migration (amongst others) led to genetic diversity between isolated groups

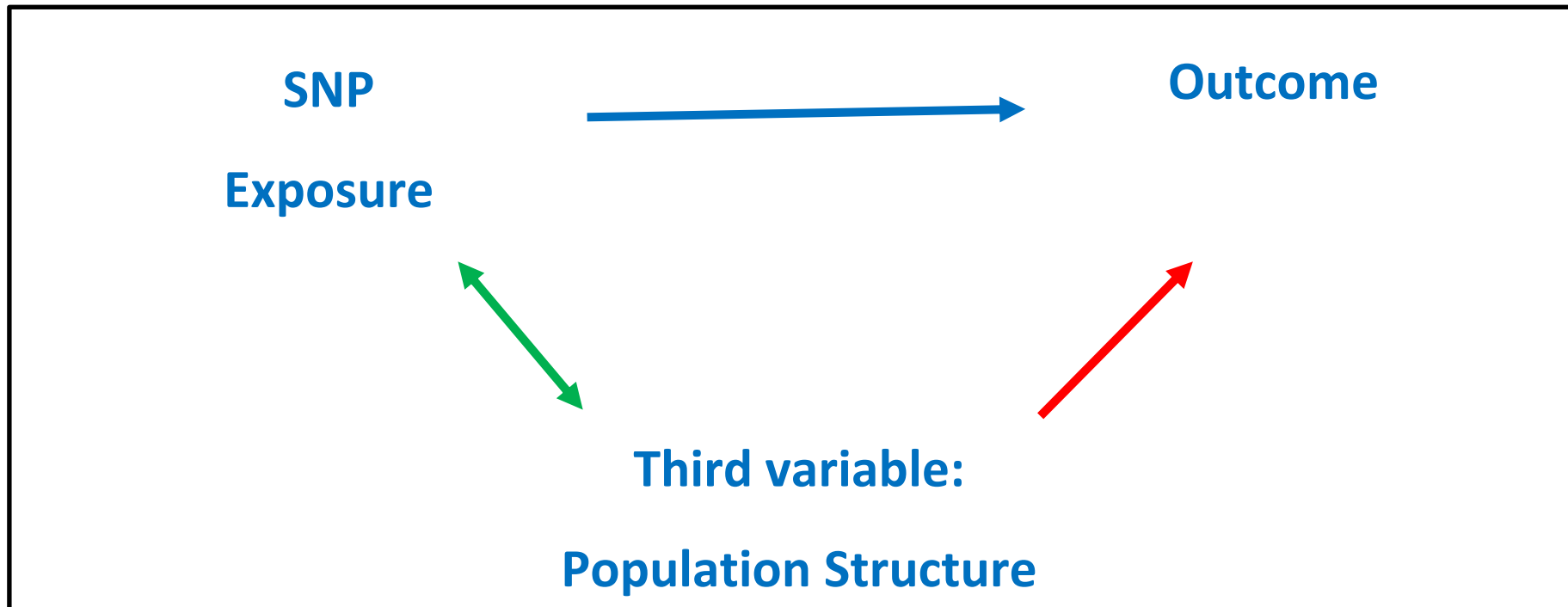




## Confounding – population structure (GWAs)



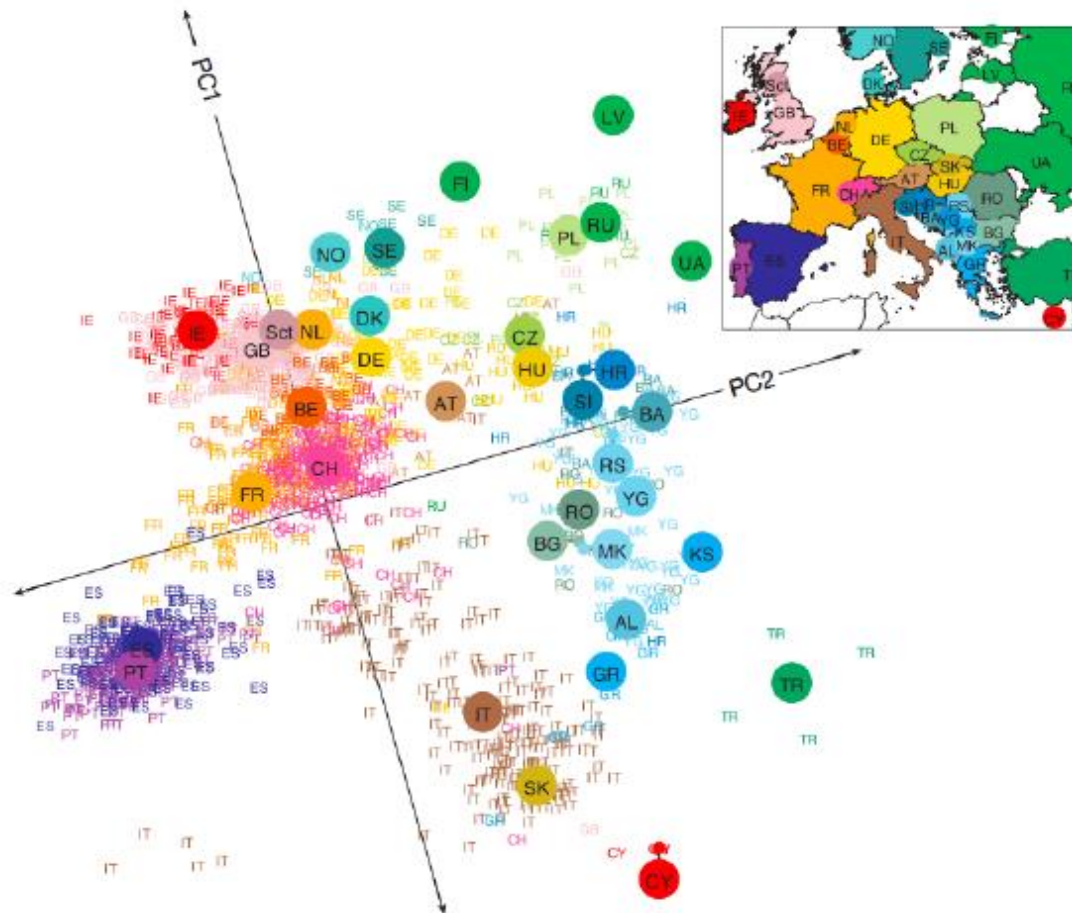
## Confounding – population structure (GWAs)



**Be associated with exposure (not a consequence of it)**

**Be associated with outcome (not an intermediary)**

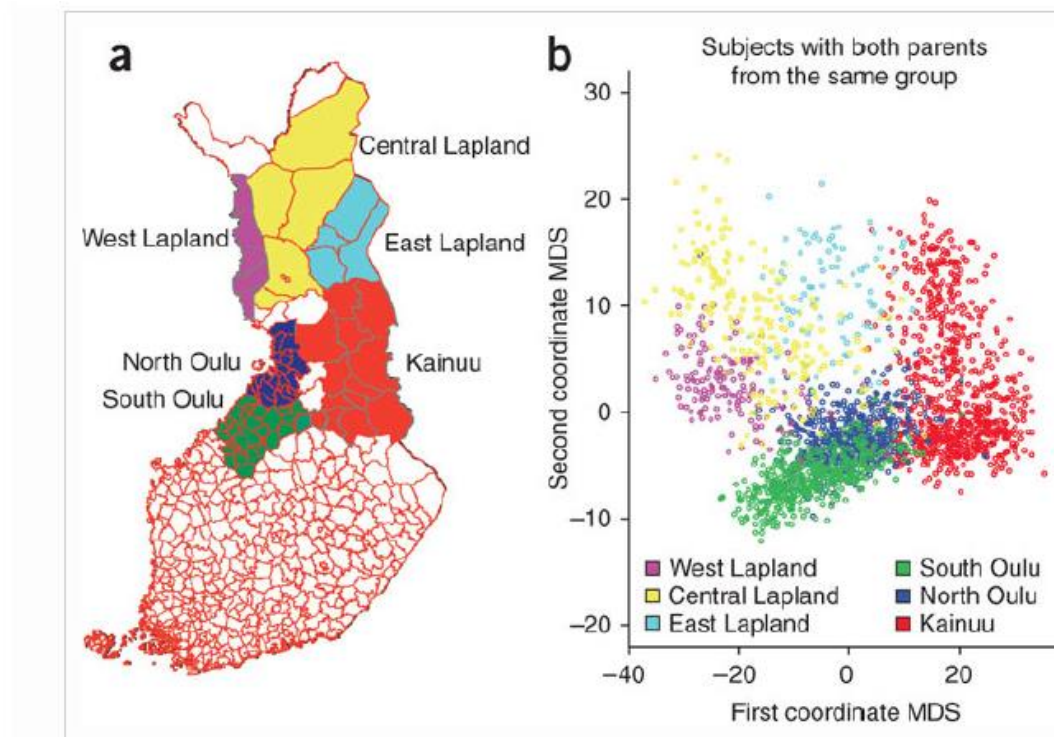
## Population structure – stratification in Europe



(Novembre et al 2008 – base don 200,000 SNPs)

## Population structure – stratification in Finland

- There can be population structure in all populations, even those that appear to be relatively “homogeneous”

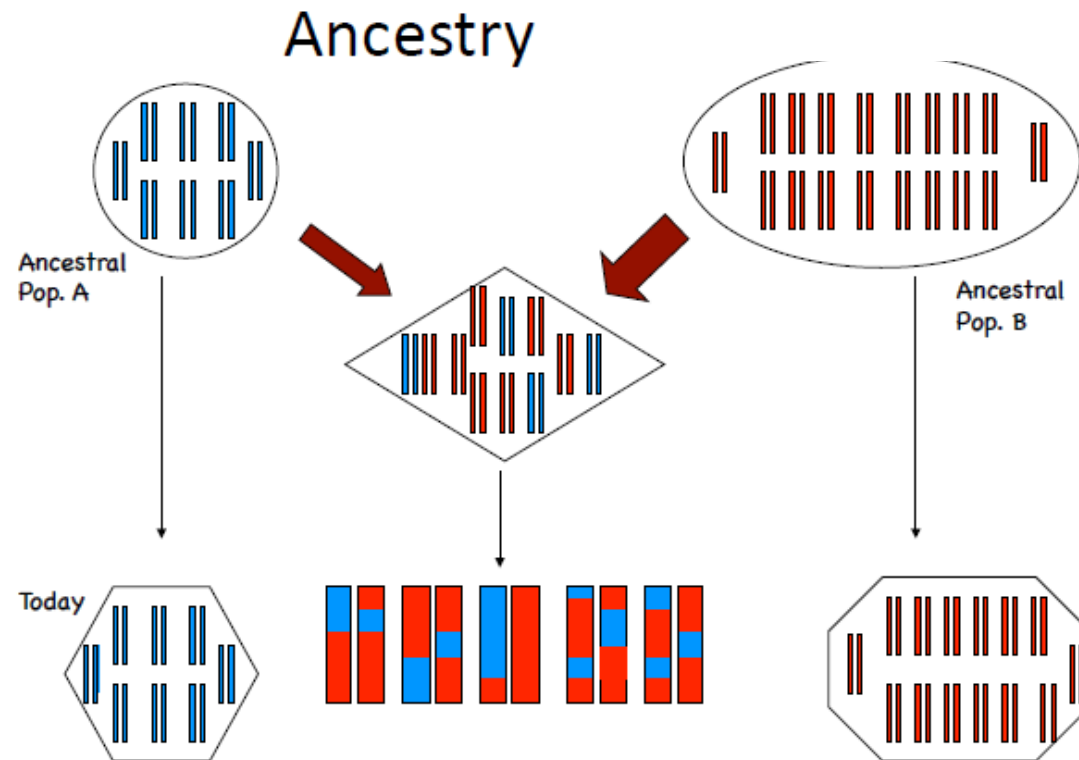


(Sabatti et al. 2009)

## Population structure – admixture

- ▶ Several recent and ongoing genetic studies have focused on **admixed populations**: populations characterized by ancestry derived from two or more ancestral populations that were reproductively isolated.
- ▶ Admixed populations have arisen in the past several hundred years as a consequence of historical events such as the transatlantic slave trade, the colonization of the Americas and other long-distance migrations.
- ▶ Examples of admixed populations include
  - ▶ African Americans and Hispanic Americans in the U.S
  - ▶ Latinos from throughout Latin America
  - ▶ Uyghur population of Central Asia
  - ▶ Cape Verdeans
  - ▶ South African "Coloured" population

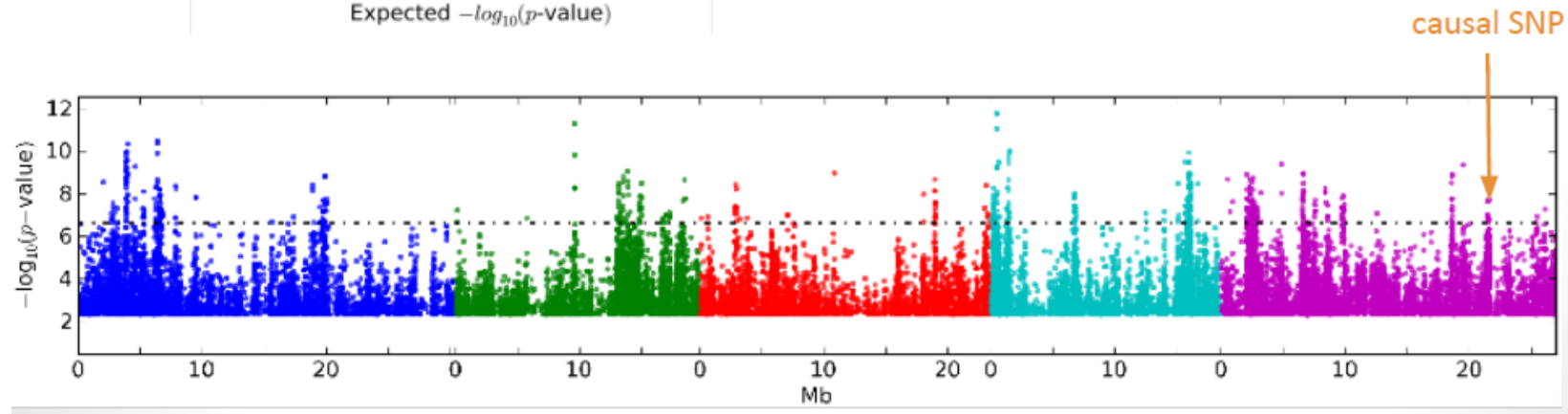
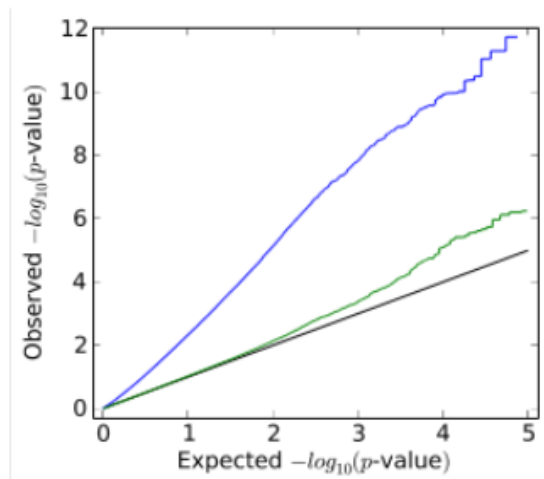
## Population structure – admixture



- ▶ The chromosomes of an admixed individual represent a mosaic of chromosomal blocks from the ancestral populations.

## Implications for GWAs due to shared genetic ancestry

- Inflated test statistics
- Too high false positive rates

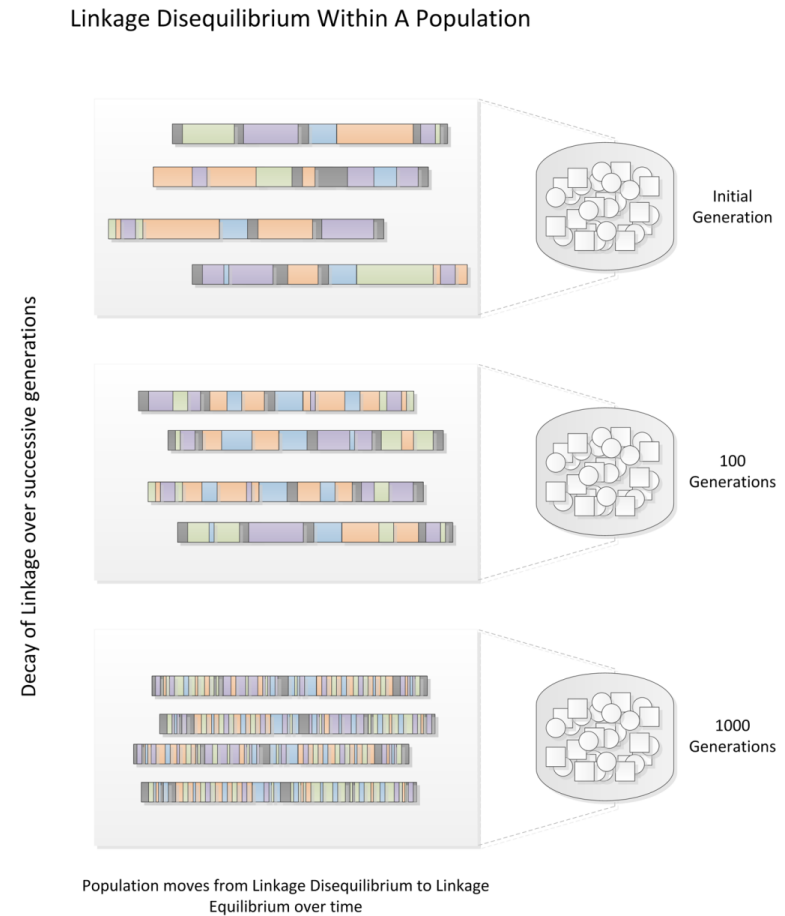
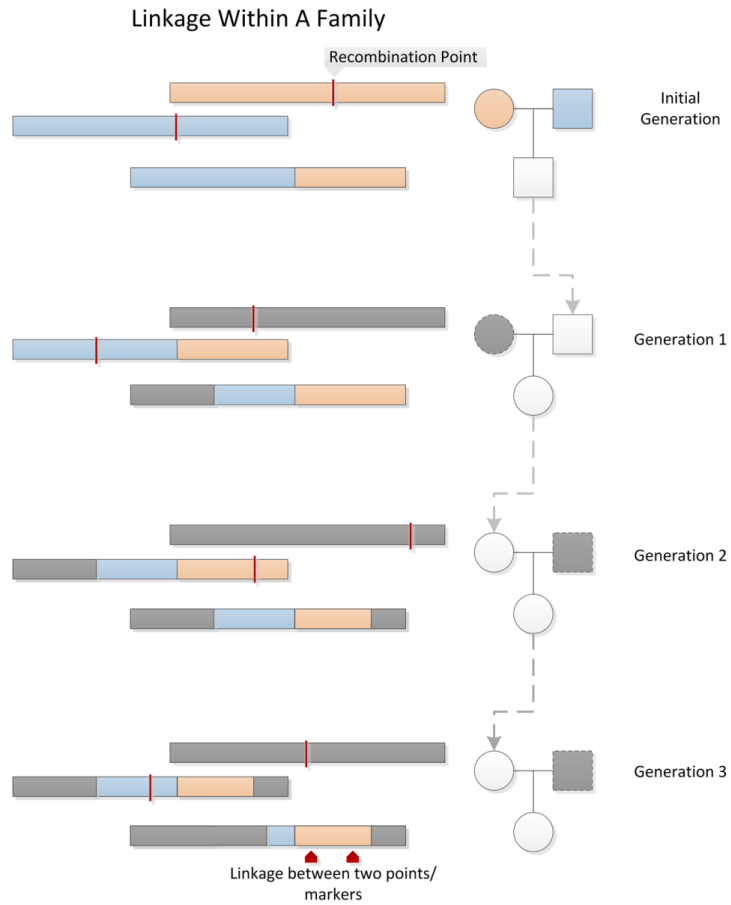


## Inference about population structure

- ▶ Inference on genetic ancestry differences among individuals from different populations, or **population structure**, has been motivated by a variety of applications:
  - ▶ population genetics
  - ▶ genetic association studies
  - ▶ personalized medicine
  - ▶ forensics
- ▶ Advancements in array-based genotyping technologies have largely facilitated the investigation of genetic diversity at remarkably high levels of detail
- ▶ A variety of methods have been proposed for the identification of genetic ancestry differences among individuals in a sample using high-density genome-screen data.



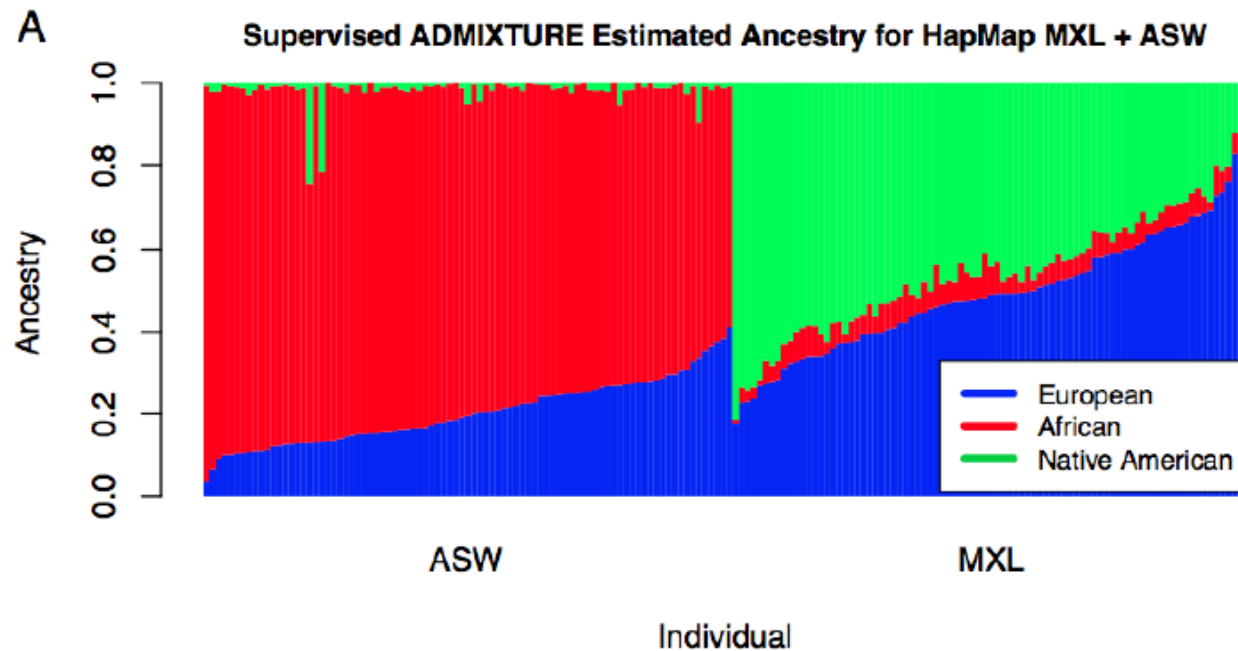
# Inference about admixture using ADMIXTURE



(Bush et al 2012)

# Inference about admixture using ADMIXTURE

## HapMap ASW and MXL Ancestry



Population	Estimated Ancestry Proportions (SD)		
	European	African	Native American
MXL	49.9% (14.8%)	6%(1.8%)	44.1% (14.8%)
ASW	20.5% (7.9%)	77.5% (8.4%)	1.9% (3.5%)

## Inference about admixture using ADMIXTURE

- ▶ Genome-screen data on 150,872 autosomal SNPs was used to estimate ancestry
- ▶ Estimated genome-wide ancestry proportions of every individual using the ADMIXTURE (Alexander et al., 2009) software
- ▶ A supervised analysis was conducted using genotype data from the following reference population samples for three "ancestral" populations
  - ▶ HapMap YRI for West African ancestry
  - ▶ HapMap CEU samples for northern and western European ancestry
  - ▶ HGDP Native American samples for Native American ancestry.

## Inference about population structure with PCA

- ▶ Principal Components Analysis (PCA) is the most widely used approach for identifying and adjusting for ancestry difference among sample individuals
- ▶ PCA applied to genotype data can be used to calculate **principal components** (PCs) that explain differences among the sample individuals in the genetic data
- ▶ The top PCs are viewed as continuous axes of variation that reflect genetic variation due to ancestry in the sample.
- ▶ Individuals with similar values for a particular top principal component will have similar ancestry for that axes.

## PCA in a nutshell

### Notation

- ▶  $\mathbf{x}$  is a vector of  $p$  random variables
- ▶  $\alpha_k$  is a vector of  $p$  constants
- ▶  $\alpha'_k \mathbf{x} = \sum_{j=1}^p \alpha_{kj} x_j$

### Procedural description

- ▶ Find linear function of  $\mathbf{x}$ ,  $\alpha'_1 \mathbf{x}$  with maximum variance.
- ▶ Next find another linear function of  $\mathbf{x}$ ,  $\alpha'_2 \mathbf{x}$ , uncorrelated with  $\alpha'_1 \mathbf{x}$  maximum variance.
- ▶ Iterate.

### Goal

It is hoped, in general, that most of the variation in  $\mathbf{x}$  will be accounted for by  $m$  PC's where  $m \ll p$ .

## Assumption and More Notation

- ▶  $\Sigma$  is the *known* covariance matrix for the random variable  $\mathbf{x}$
- ▶ Foreshadowing :  $\Sigma$  will be replaced with  $\mathbf{S}$ , the sample covariance matrix, when  $\Sigma$  is unknown.

## Shortcut to solution

- ▶ For  $k = 1, 2, \dots, p$  the  $k^{\text{th}}$  PC is given by  $z_k = \alpha_k' \mathbf{x}$  where  $\alpha_k$  is an eigenvector of  $\Sigma$  corresponding to its  $k^{\text{th}}$  largest eigenvalue  $\lambda_k$ .
- ▶ If  $\alpha_k$  is chosen to have unit length (i.e.  $\alpha_k' \alpha_k = 1$ ) then  $\text{Var}(z_k) = \lambda_k$

## First Step

- ▶ Find  $\alpha'_k \mathbf{x}$  that maximizes  $\text{Var}(\alpha'_k \mathbf{x}) = \alpha'_k \mathbf{\Sigma} \alpha_k$
- ▶ Without constraint we could pick a very big  $\alpha_k$ .
- ▶ Choose normalization constraint, namely  $\alpha'_k \alpha_k = 1$  (unit length vector).

## Constrained maximization - method of Lagrange multipliers

- ▶ To maximize  $\alpha'_k \mathbf{\Sigma} \alpha_k$  subject to  $\alpha'_k \alpha_k = 1$  we use the technique of Lagrange multipliers. We maximize the function

$$\alpha'_k \mathbf{\Sigma} \alpha_k - \lambda(\alpha'_k \alpha_k - 1)$$

w.r.t. to  $\alpha_k$  by differentiating w.r.t. to  $\alpha_k$ .

## Constrained maximization - method of Lagrange multipliers

- ▶ This results in

$$\begin{aligned}\frac{d}{d\alpha_k} (\alpha'_k \Sigma \alpha_k - \lambda_k (\alpha'_k \alpha_k - 1)) &= 0 \\ \Sigma \alpha_k - \lambda_k \alpha_k &= 0 \\ \Sigma \alpha_k &= \lambda_k \alpha_k\end{aligned}$$

- ▶ This should be recognizable as an eigenvector equation where  $\alpha_k$  is an eigenvector of  $\Sigma_b f$  and  $\lambda_k$  is the associated eigenvalue.
- ▶ Which eigenvector should we choose?



## Constrained maximization - method of Lagrange multipliers

- ▶ If we recognize that the quantity to be maximized

$$\boldsymbol{\alpha}'_k \boldsymbol{\Sigma} \boldsymbol{\alpha}_k = \boldsymbol{\alpha}'_k \lambda_k \boldsymbol{\alpha}_k = \lambda_k \boldsymbol{\alpha}'_k \boldsymbol{\alpha}_k = \lambda_k$$

then we should choose  $\lambda_k$  to be as big as possible. So, calling  $\lambda_1$  the largest eigenvalue of  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\alpha}_1$  the corresponding eigenvector then the solution to

$$\boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = \lambda_1 \boldsymbol{\alpha}_1$$

is the 1<sup>st</sup> principal component of  $\mathbf{x}$ .

- ▶ In general  $\boldsymbol{\alpha}_k$  will be the  $k^{\text{th}}$  PC of  $\mathbf{x}$  and  $\text{Var}(\boldsymbol{\alpha}'\mathbf{x}) = \lambda_k$

## Unsupervised learning with PCA – identifying genetic ancestry

- ▶ Suppose a genetic association study consists of a sample of  $N$  individuals
- ▶ Assume that genotype data is available at  $S$  SNPs in a genome-screen, where  $S$  can be very large (e.g. hundreds of thousands).
- ▶ For SNP  $s$  define  $\mathbf{G}_s = (G_1^s, \dots, G_n^s)^T$  is  $n \times 1$  vector of the genotypes, where  $G_i^s = 0, \frac{1}{2},$  or  $1$ , according to whether individual  $i$  has, respectively, 0, 1 or 2 copies of the reference allele at SNP  $s$ .
- ▶ We define  $\mathbf{Z}$  to be an  $N \times S$  standardized matrix with  $(i, s)$ -th entry

$$Z_{is} = \frac{G_i^s - \hat{p}_s}{\sqrt{\hat{p}_s(1 - \hat{p}_s)}}$$

and  $\hat{p}_s$  will typically be an allele frequency estimate for SNP  $s$

## Unsupervised learning with PCA – identifying genetic ancestry

- ▶ Can obtain a genetic relationship matrix (GRM)  $\hat{\Psi}$  where

$$\hat{\Psi} = \frac{1}{S} \mathbf{Z}\mathbf{Z}^T$$

The  $(i, j)$ -th entry of  $\hat{\Psi}$  is a measure of the average genetic similarity for individuals  $i$  and  $j$  in the sample.

$$\hat{\Psi}_{ij} = \frac{1}{S} \sum_{s=1}^S \frac{(G_i^s - \hat{\rho}_s)(G_j^s - \hat{\rho}_s)}{\hat{\rho}_s(1 - \hat{\rho}_s)}$$

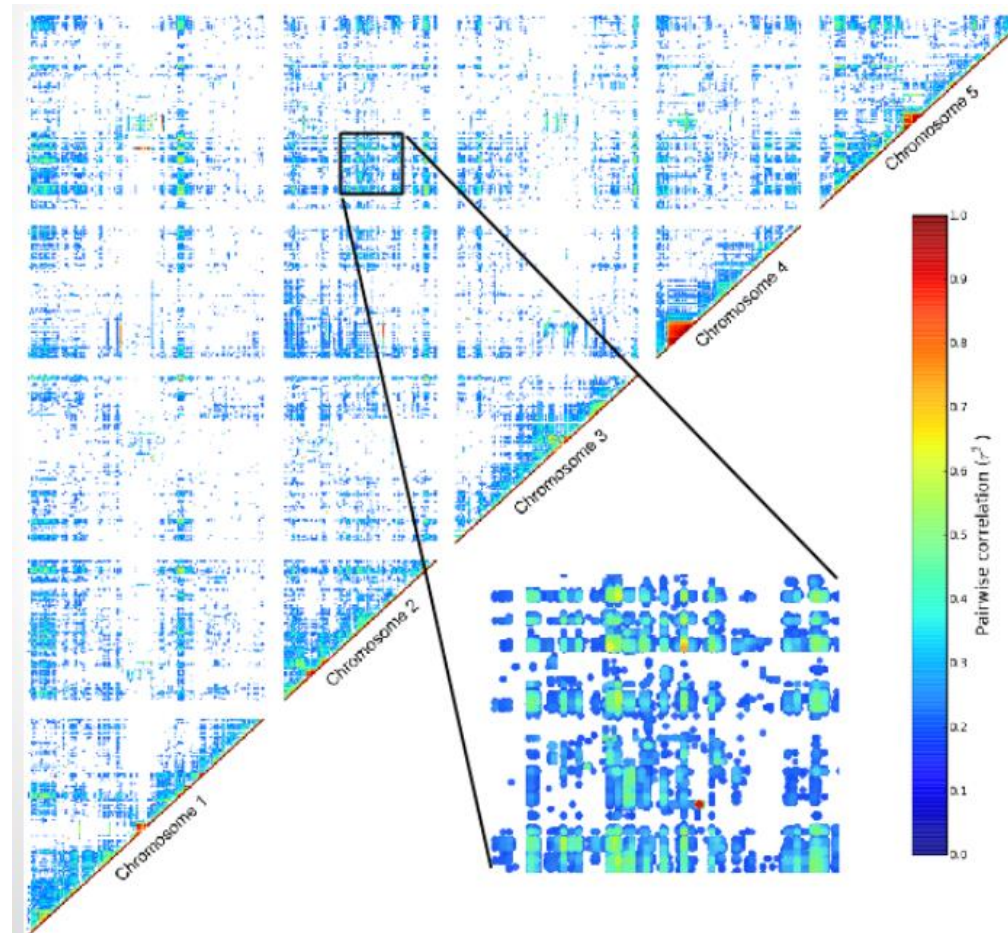
- ▶ Principal Components Analysis (PCA) is a dimension reduction that can be applied to GRM to identify ancestry differences among sample individuals
- ▶ PCA is performed by obtaining the eigendecomposition of the GRM  $\hat{\Psi}$ .



## Unsupervised learning with PCA – identifying genetic ancestry

- ▶ Orthogonal axes of variation, i.e. linear combinations of SNPs, that best explain the genotypic variability amongst the  $n$  sample individuals are identified.
- ▶ For the eigendecomposition we have  $\hat{\Psi} = \mathbf{VDV}^T$ , where  $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_n]$  is an  $n \times n$  matrix with orthogonal column vectors, and  $\mathbf{D}$  corresponding to a diagonal matrix of the length  $n$  eigenvalue vector  $\mathbf{\Lambda} = (\lambda_1, \lambda_2, \dots, \lambda_n)^T$
- ▶ The eigenvalues are in decreasing order,  $\lambda_1 > \lambda_2 > \dots > \lambda_n$ . The  $d^{\text{th}}$  principal component (eigenvector) corresponds to eigenvalue  $\lambda_d$ , where  $\lambda_d$  is proportional to the percentage of variability in the genome-screen data that is explained by  $\mathbf{V}_d$ .

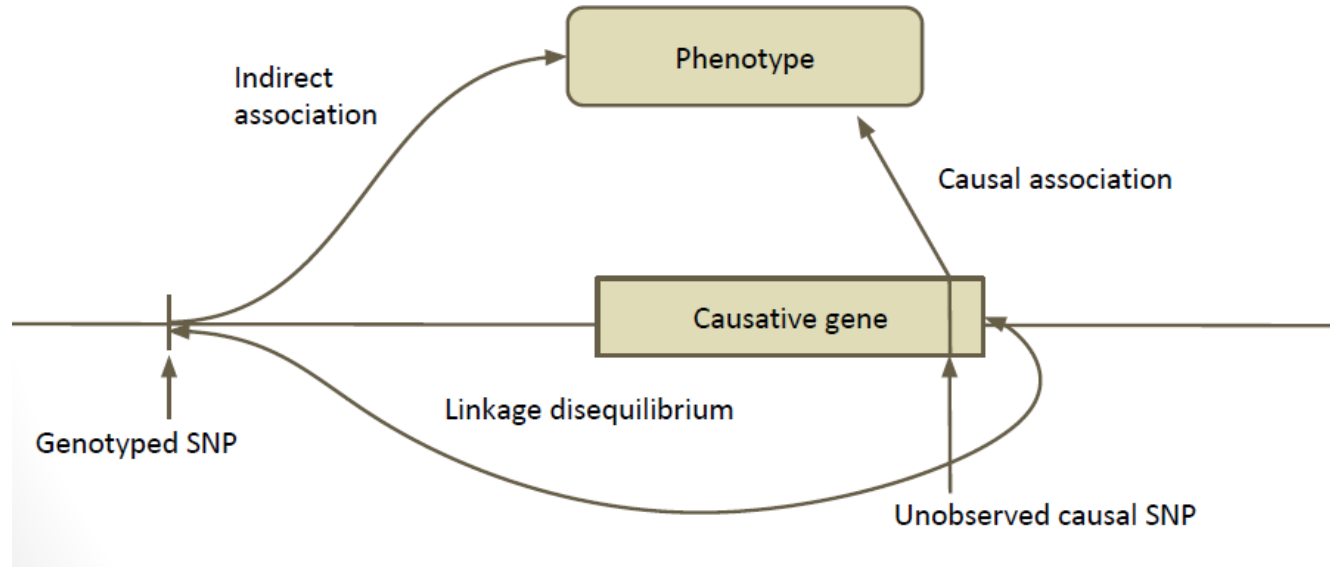
## Population structure may be reflected in long-range LD



Which SNPs to consider in the computation of the PCs? (“LD pruning”)

## LD: a nuisance or a merit?

- Since LD causes correlations between markers, in a given population we expect a lot of redundancy in the genotypes
- Neighboring markers will tend to be inherited together, causing linkage disequilibrium (LD) between the two markers → LD can help finding the causal loci!



## Confounding – practical consequences for GWAs

- Include a number (?) of **PCs** in the association models. When populations become highly complex, too many PCs may be needed. Top PCs are seen as continuous axes of variation that reflect genetic variation due to ancestry in the sample.
- **Genomic control:** Scale down the test statistic so that its median becomes the expected median. It is heavily used but may not entirely solve the problem (Devlin & Roeder 1999)
- **Mixed models:** these models model the genotype effect as a random term (with variation) in a so-called mixed model. This is done by explicitly describing the covariance structure between individuals (Yu et al. 2006; Kang et al 2018)

## Genomic control

- In Genomic Control (GC), a 1-df association test statistic is computed at each of the null SNPs, and a parameter  $\lambda$  is calculated as the empirical median divided by its expectation under the chi-squared 1-df distribution.
- Then the association test is applied at the candidate SNPs, and if  $\lambda > 1$  the test statistics are divided by  $\lambda$ .
  - Under  $H_0$  of no association p-values uniformly distributed
  - In case of population stratification: inflation of test statistics
  - $\hat{\lambda} = \frac{\text{median}(\chi_1^2, \chi_2^2, \dots, \chi_L^2)}{\text{median}(\mathcal{L}(\chi_1^2))} = \frac{\text{median}(\chi_1^2, \chi_2^2, \dots, \chi_L^2)}{0.456}$
  - $\chi_{GC}^2 = \chi^2 / \hat{\lambda}$



```
> median(rchisq(10,1))  
[1] 0.9641272  
  
> median(rchisq(100,1))  
[1] 0.5001173  
  
> median(rchisq(1000,1))  
[1] 0.4206546  
  
> median(rchisq(10000,1))  
[1] 0.4686072  
  
> median(rchisq(100000,1))  
[1] 0.455271  
  
> median(rchisq(1000000,1))  
[1] 0.4548966
```

## From linear regression to linear mixed models

A linear model generally refers to linear regression models in statistics.

$$y_j = \sum_{j=1}^P \beta_j x_{ij} + \epsilon_i \qquad Y = X' \beta + \epsilon$$

- **Y** typically consists of the phenotype values, or case-control status for **N** individuals.
- **X** is the **NxP** genotype matrix, consisting of **P** genetic variants (e.g. SNPs).
- $\hat{\beta}$  is a vector of **P** effects for the genetic variants.
- $\epsilon$  is still just known as the *noise* or *error* term.

## From linear regression to linear mixed models

$$Y = X\beta + u + \epsilon, \quad u \sim N(0, \sigma_g K), \quad \epsilon \sim N(0, \sigma_e I)$$

- Initially proposed in Association mapping by Yu et al. (2006)
- **Y** typically consists of the phenotype values, or case-control status for **N** individuals.
- **X** is the **N**×**P** genotype matrix, consisting of **P** genetic variants (e.g. SNPs).
- **u** is the random effect of the mixed model with  $\text{var}(u) = \sigma_g K$
- **K** is the **N** × **N** kinship matrix inferred from genotypes
- **β** is a vector of **P** effects for the genetic variants.
- **ε** is a **N** × **N** matrix of residual effects with  $\text{var}(\epsilon) = \sigma_e I$

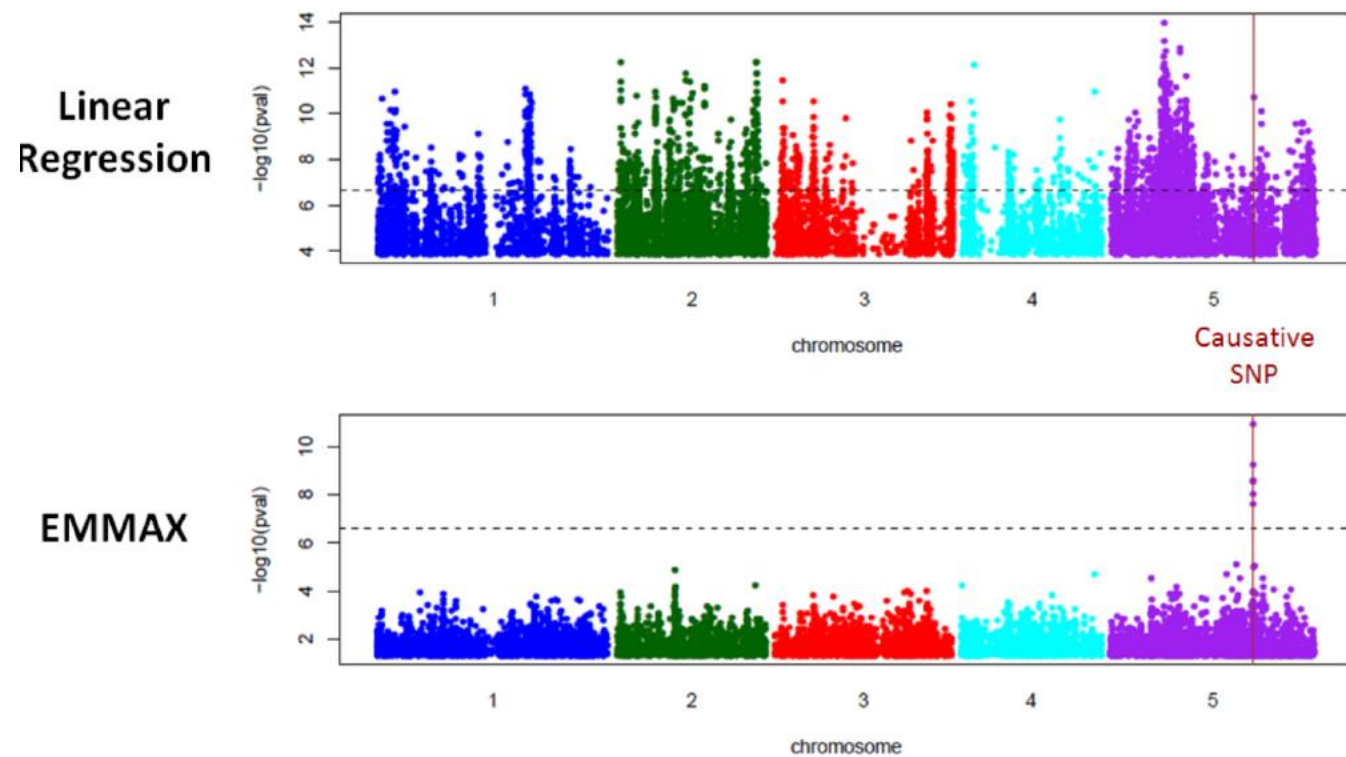
## From linear regression to linear mixed models

- The kinship measures the degree of relatedness, and is in general different from the covariance matrix.
- It is estimated using either pedigree (family relationships) data or (lately) using genotype data.
  - When estimating it from pedigree data, one normally assumes that the ancestral founders are “unrelated”.
  - They are sensitive to confounding by cryptic relatedness.
- Alternatively the kinship can be estimated from genotype data.
  - Genotype data may be incomplete.
  - Weights or scaling of genotypes can impact the kinship.

## From linear regression to linear mixed models - speed

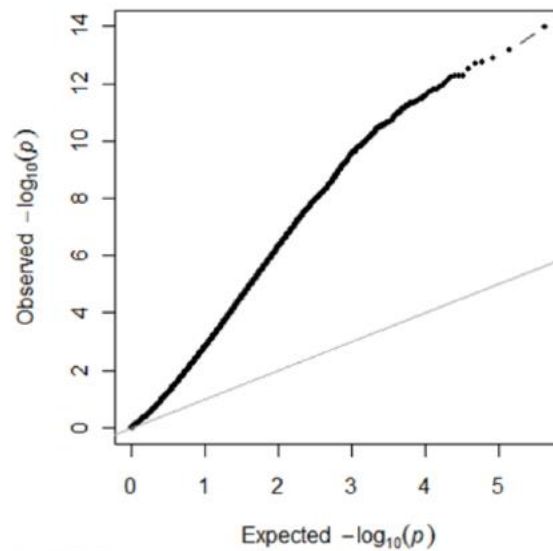
- Original implementation: EMMA (Kang *et al.*, 2008)
  - Problem:  $O(PN^3)$  → 1 GWAS in 1 day (500k individuals)
- Approximate methods  $O(PN^2)$ :
  - GRAMMAR (Aulchenko *et al.*, 2007) <http://www.genabel.org/packages/GenABEL>
  - P3D (Zhang *et al.*, 2010) <http://www.maizegenetics.net/#!tassel/c17q9>
  - EMMAX (Kang *et al.*, 2010) <http://genetics.cs.ucla.edu/emmax/>
- Exact methods:
  - FaST LMM (Lippert *et al.*, 2011) <http://mscompbio.codeplex.com/>
  - GEMMA (Zhou *et al.*, 2012) <http://www.xzlab.org/software.html>
- This is too slow for large samples (>20000 individuals), i.e. exactly the sample sizes where one might expect to see most gains.
  - BOLT-LMM (Loh *et al.*, 2015),  **$O(PN)$**  <https://data.broadinstitute.org/alkesgroup/BOLT-LMM/>.

# From linear regression to linear mixed models – adequate control for population structure

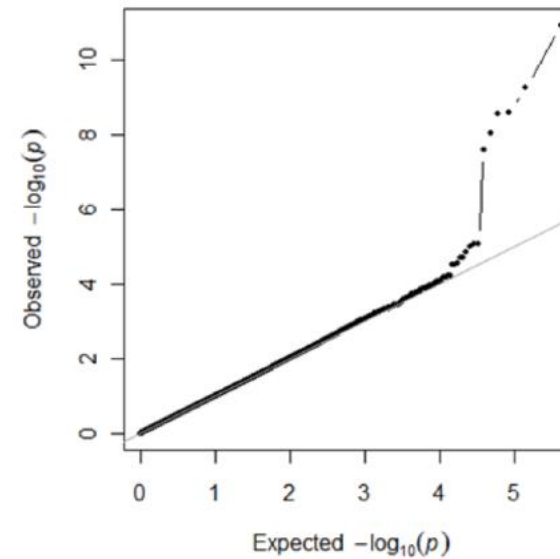


# From linear regression to linear mixed models – adequate control for population structure

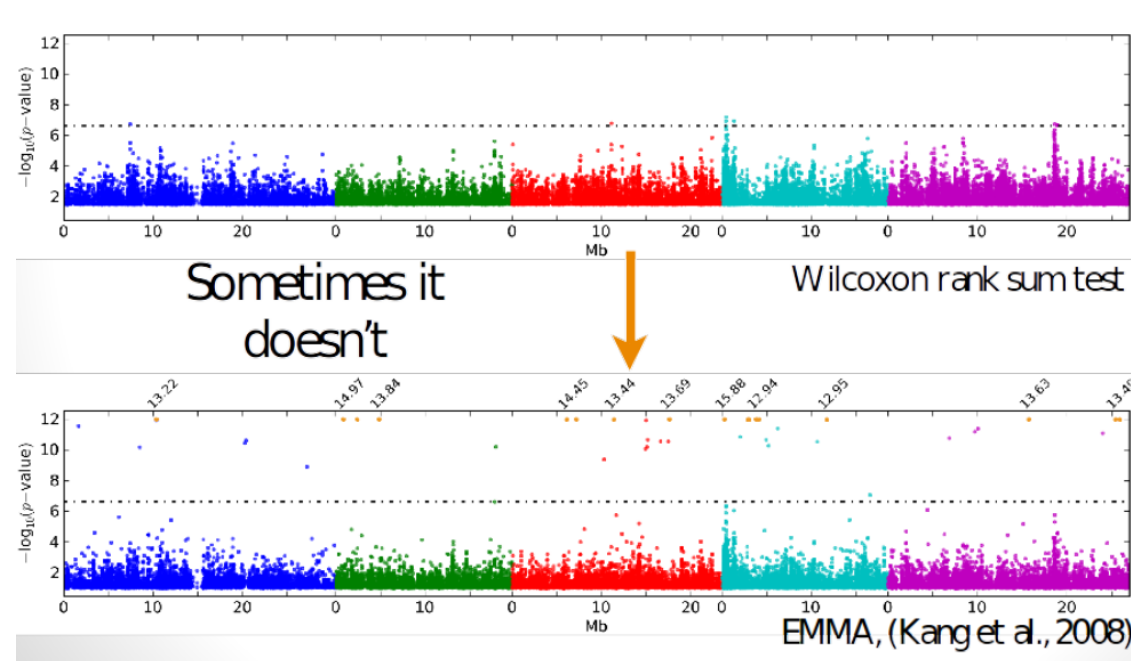
## Linear Regression



## EMMAX



## From linear regression to linear mixed models – Inadequate control for population structure



The Wilcoxon rank sum test (Mann-Whitney U) can be used to determine whether two independent samples were selected from populations having the same distribution. Unlike the  $t$ -test it does not require the assumption of normal distributions. It is nearly as efficient as the  $t$ -test on normal distributions.



## From linear regression to linear mixed models – advanced models

The mixed-model performs pretty well, but GWAS power remain limited and need to be improved:

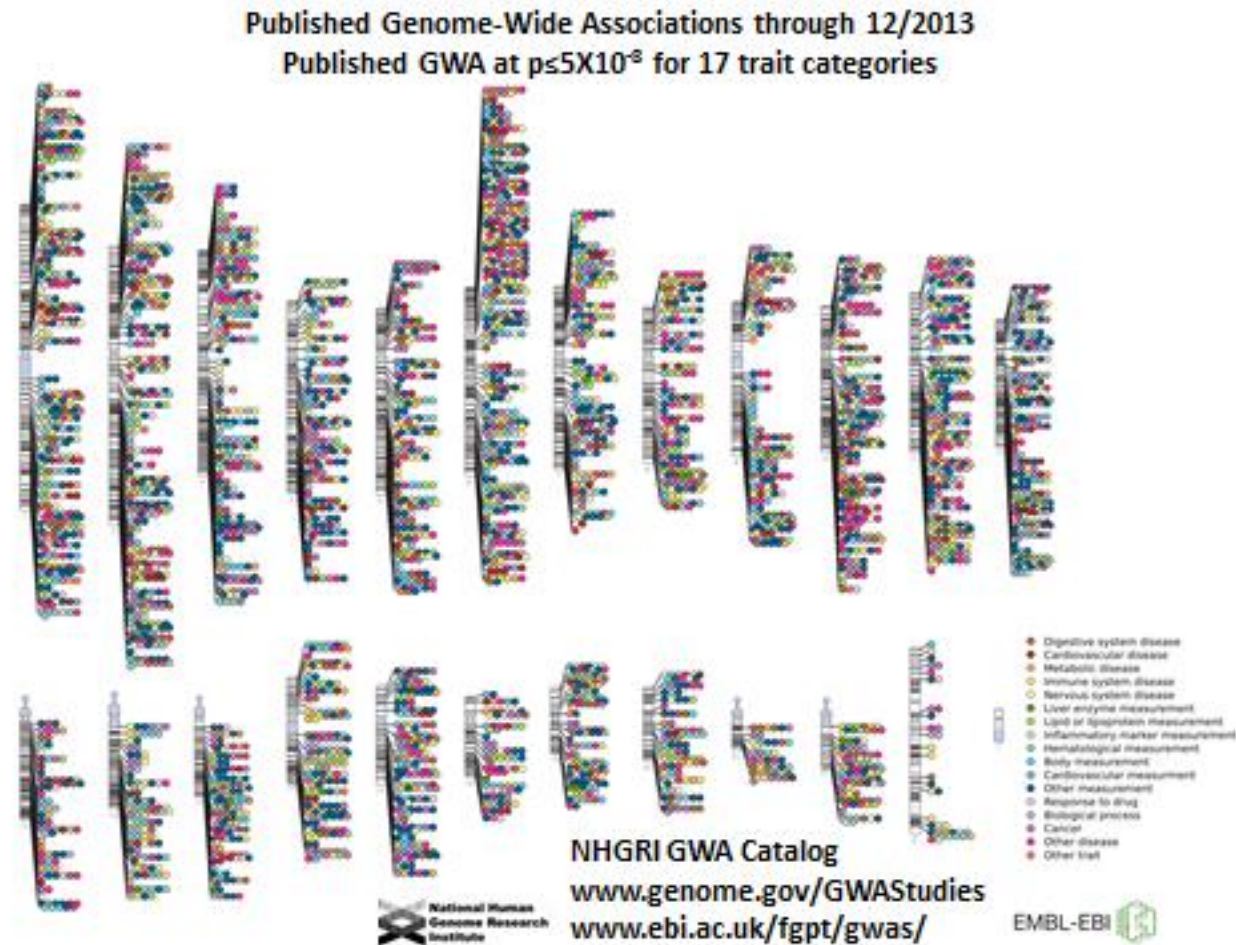
- **Multi Locus Mixed Model (MLMM, Segura *et al.*, 2012):**
  - Single SNP tests are wrong model for polygenic traits
  - Increase in power compared to single locus models
  - Detection of new associations in published datasets
  - Identification of particular cases of (synthetic associations) and/or allelic heterogeneity
- **Multi Trait Mixed Model (MTMM, Korte *et al.*, 2012):**
  - Traits are often correlated due to **pleiotropy** (shared genetics) or **linkage** between causative polymorphisms.
  - Combining correlated traits in a single model should thus increase detection power

## **Mixed models to account for population structure in GWAs – in conclusion**

- The underlying sources of confounding in GWASs are environmental and genetic.
- Population structure per se is not the problem, nor is relatedness: estimates of either can help us to reduce confounding, but to do this well, it is helpful to understand its true source.
- Mixed models that attempt to describe phenotypic covariance are a natural way to model this confounding. They have a solid mechanistic basis, and the variance components estimated are easily interpreted, allowing us to distinguish genetic from environmental components

### 3 Multiple testing

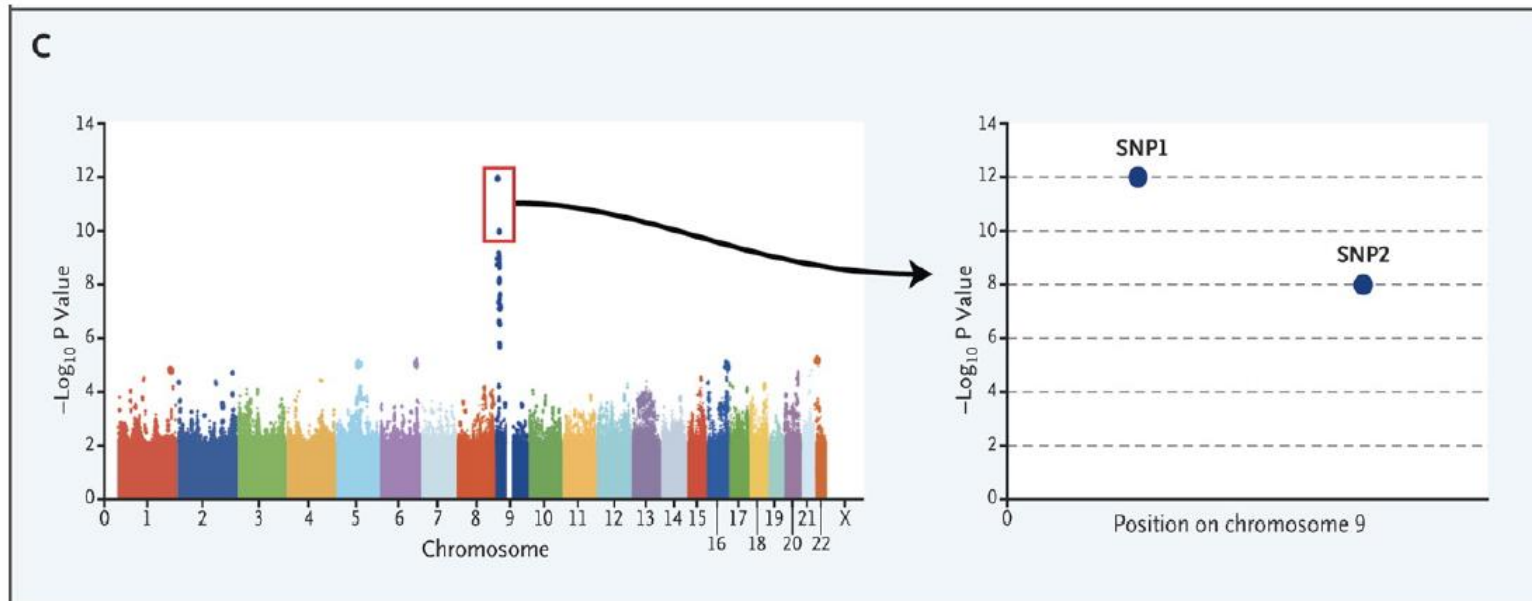
## Locus heterogeneity



## Multiple testing

- In GWAs a large number of marker tests are conducted, which leads to a multiple testing problem
- Using a 5% significance threshold, we would expect 5% of the markers that have true marker effects of 0 to be significant.
- Solutions include:
  - **Bonferroni correction:** By assuming our  $m$  available markers to be independent (is this true?) we can obtain a conservative bound on the probability of rejecting the null hypothesis for one or more markers:  $1 - P(T_1 \leq t, T_2 \leq t, \dots, T_m \leq t | H_0) \leq \alpha$ , with  $\alpha$ , a given significance threshold
  - **Permutation / rank based corrections**

## Multiple testing



- “The results implicate a locus on chromosome 9, marked by SNPs 1 and 2, which are adjacent to each other (graph at right), and other neighboring SNPs.”  
(Manolio 2010)

## Multiple testing

- Bonferroni correction for 500,000 markers:

$$p \leq \frac{0.05}{500000} = 10^{-7}$$

- Bonferroni correction for 1000,000 markers:

$$p \leq \frac{0.01}{1000000} = 10^{-8}$$

- Where does  $10^{-4}$  for HWE testing (Travemünde criteria) come from?

## 4 Multiple studies

### Meta-analysis

- Fisher's method combines extreme value probabilities from each test (p-values), into one test statistic, using the formula:

$$\sum_{i=1}^k -2\log(p_i)$$

- When all the null hypotheses are true, and the  $p_i$  (p-value for the  $i$ -th hypothesis test) are independent, the test statistic follows a chi-squared distribution with  $2k$  degrees of freedom, where  $k$  is the number of tests being combined.

## Meta-analysis

	METAL	GWAMA	MetABEL	PLINK	R packages
<b>Ability to process files from GWAS analysis tools; software used</b>	No	Yes; SNPTTEST, <u>PLINK</u>	Yes; ABEL	Yes; PLINK	No
<b>Fixed effects implemented?</b>	Yes	Yes	Yes	Yes	Yes
<b>Random effects implemented?</b>	No	Yes	No	No	Yes
<b>Heterogeneity metrics generated</b>	$Q, I^2$	$Q, I^2$	$Q, I^2$	$Q, I^2$	$Q, I^2$
<b>Graphical illustration of meta-analysis results</b>	No	Manhattan and QQ plots	Forest plots	No	Yes

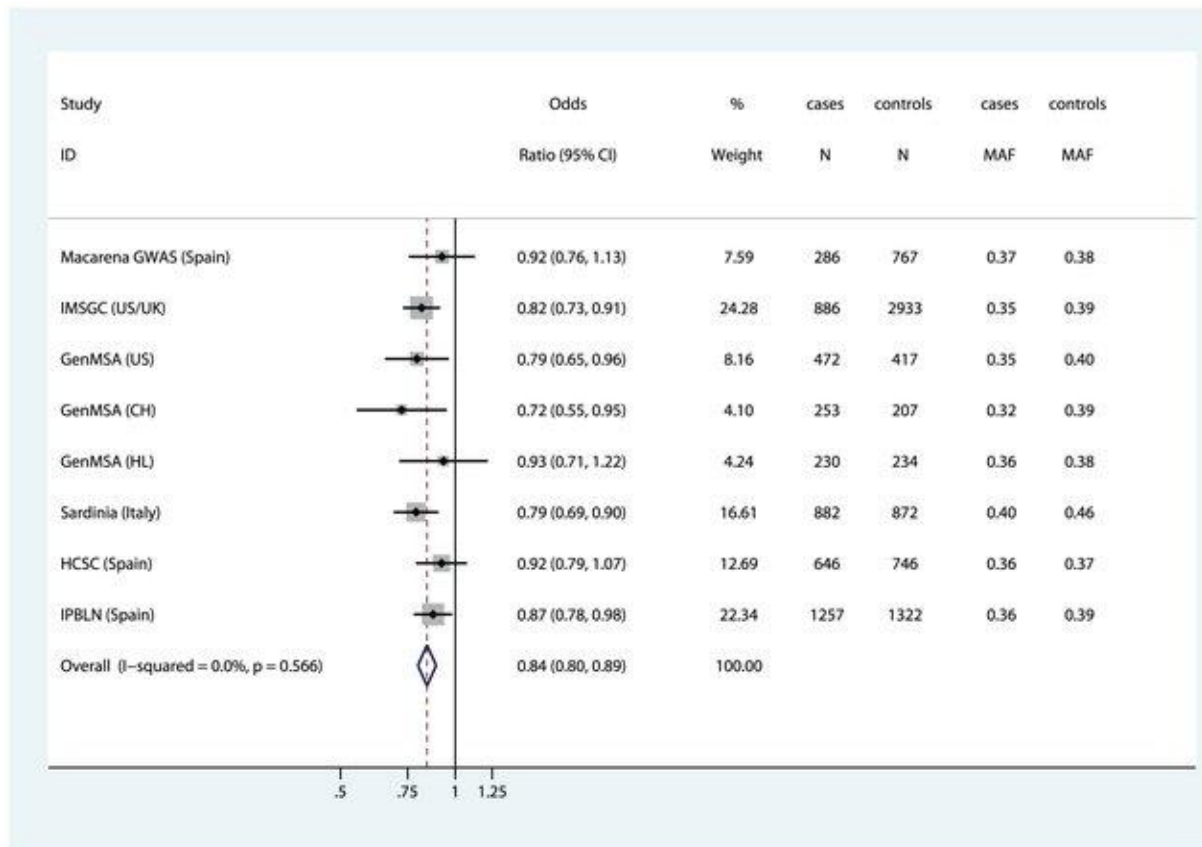
GWAS, genome-wide association study.

(Evangelou et al. 2013)



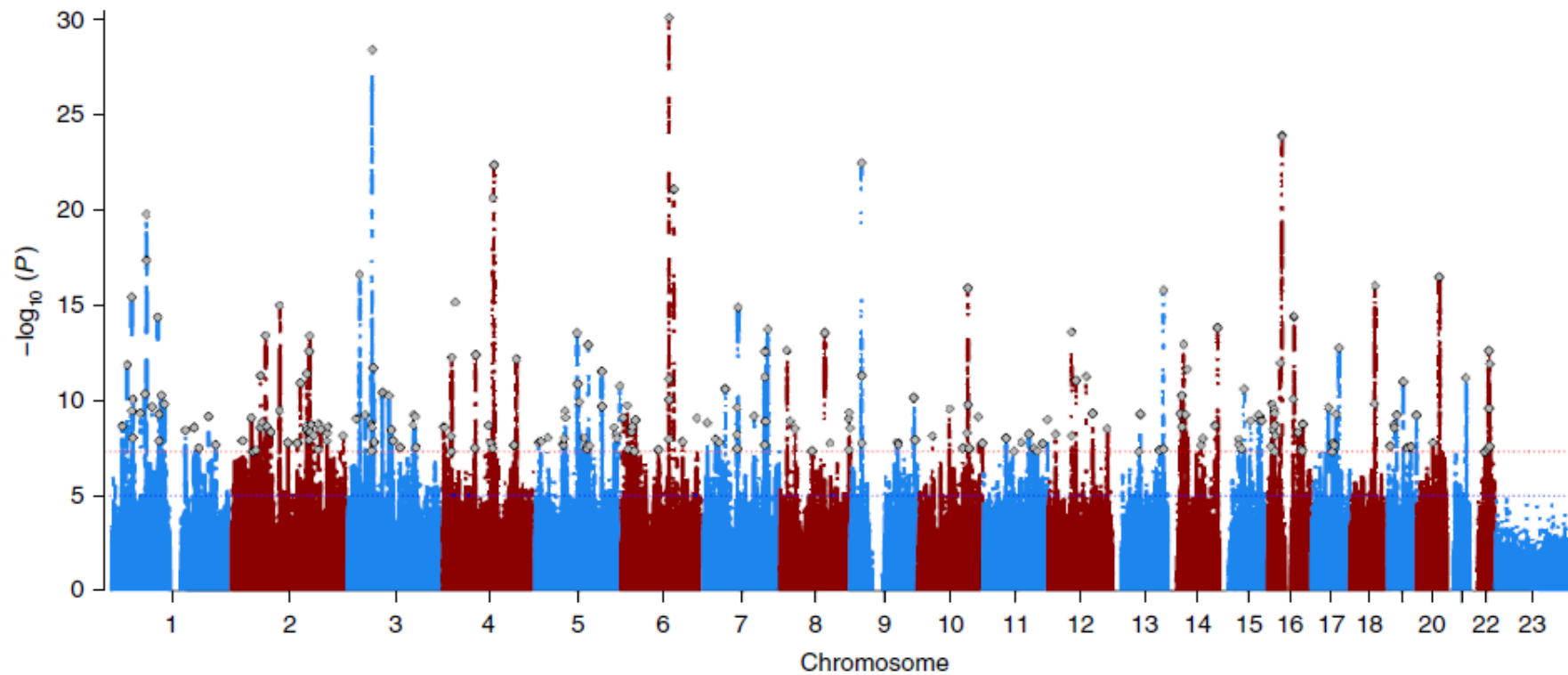
## Meta-analysis results presentation

- Forest plots (~ epidemiology)



## Meta-analysis results presentation

- Meta-analysis Manhattan plot (~ genetic epidemiology)



(Savage et al. 2018)

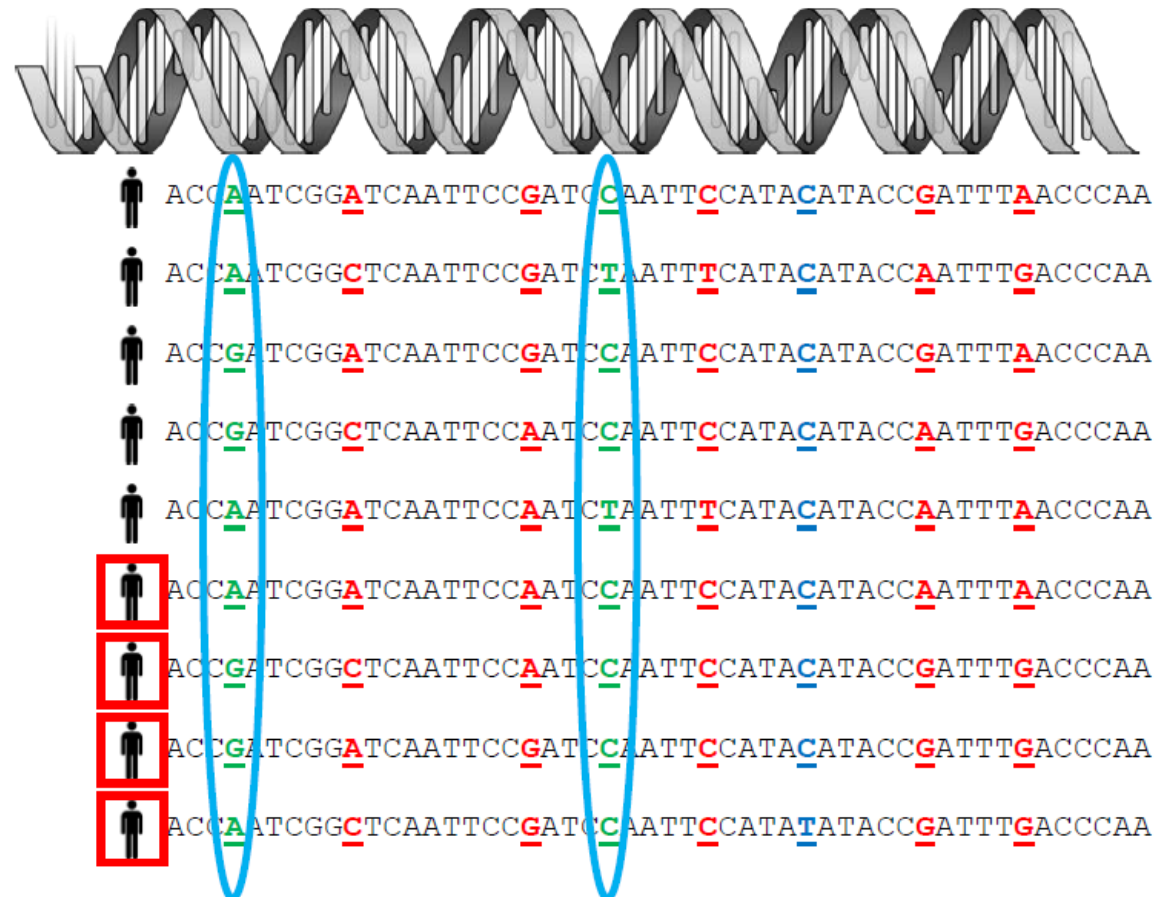
## 5 When variants become rare – sparse data

### Next generation sequencing (NGS)

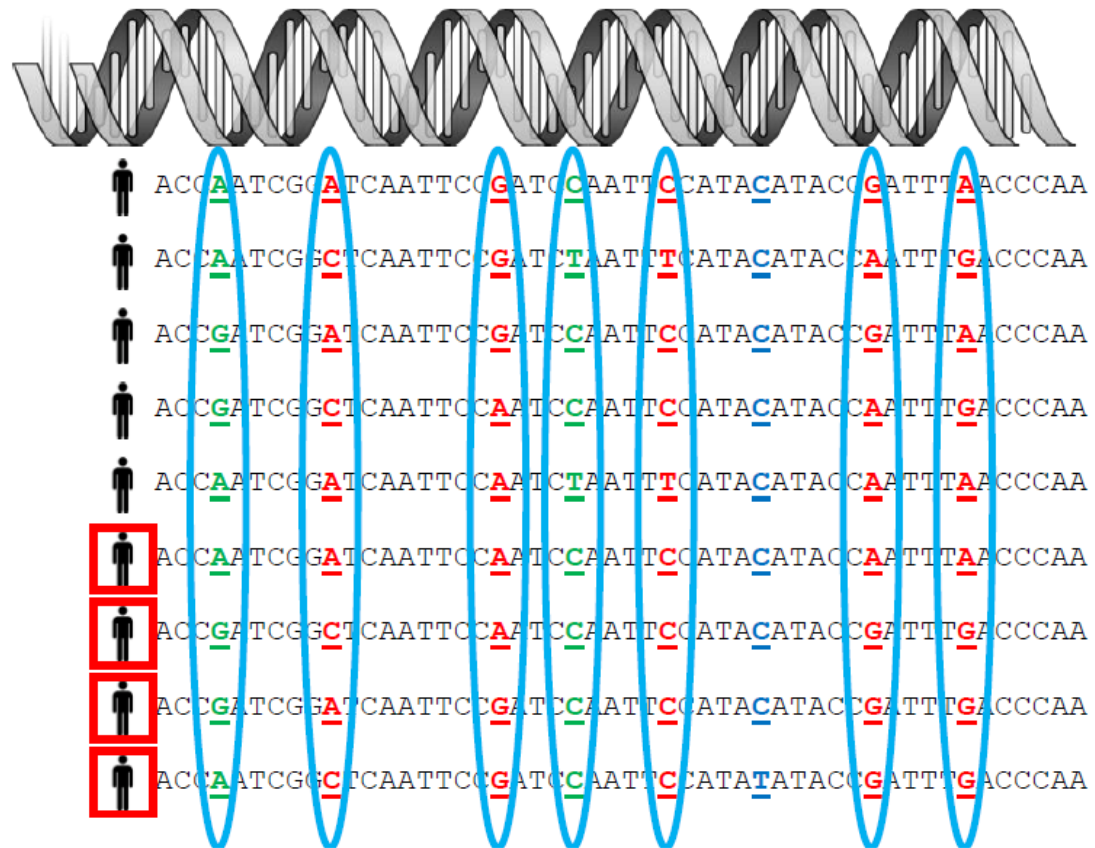
- Genotype all basepairs (bps) in a gene, the whole exomes, or the whole genome (recall: ~ 3 billion bps)
- Allow to identify all SNPs or other types of variants. No need to rely on LD to tag untyped causal SNPs



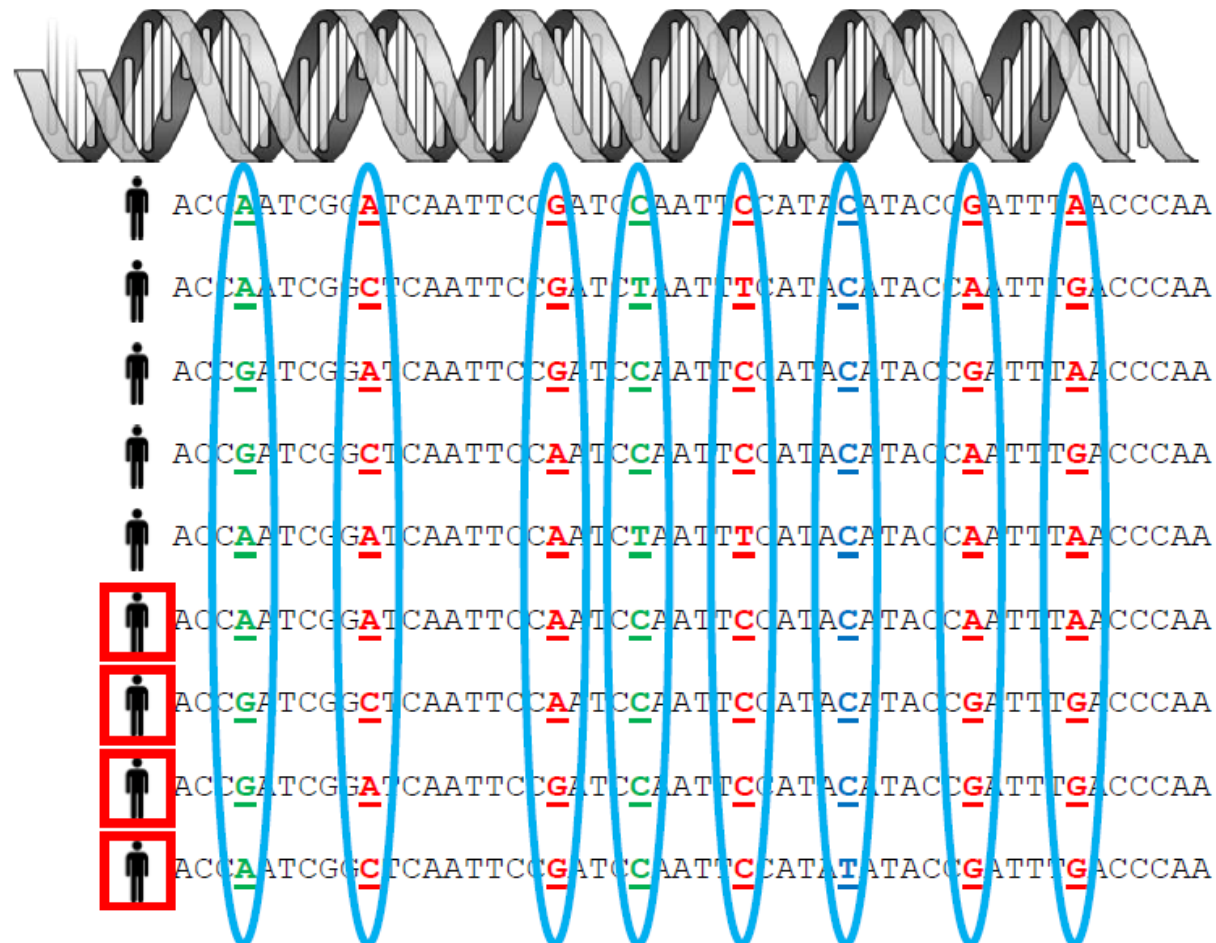
## GWAS in the early days



# GWAS nowadays (+ imputation)



## Sequencing based association

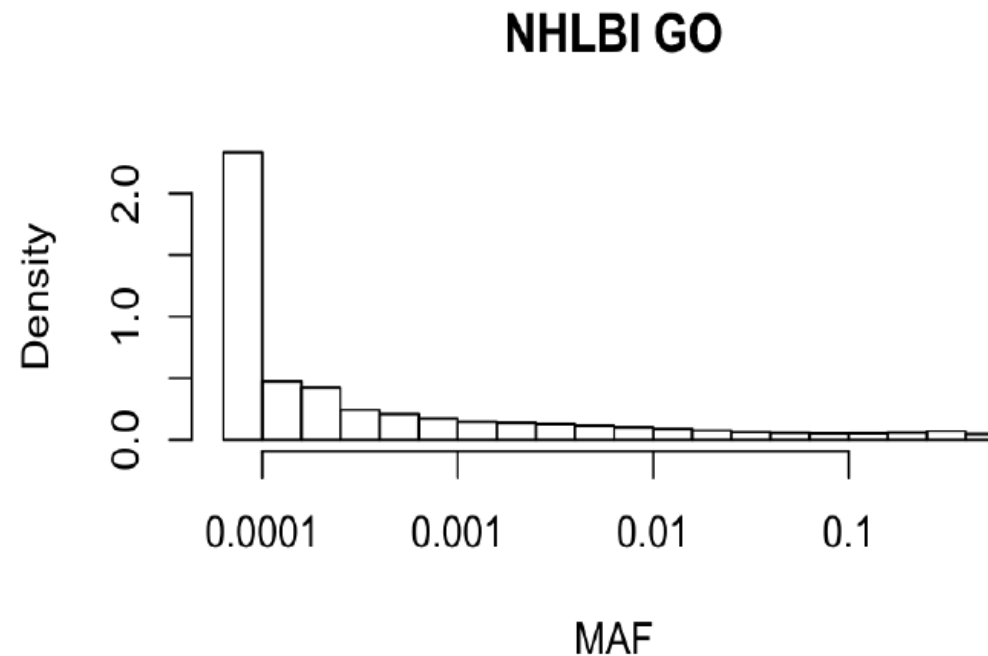


## Why studying rare variants?

- Common variants (SNPs)
  - $MAF > 0.01 \sim 0.05$
  - Often high correlation with adjacent SNPs (strong LD)
- Rare variants
  - $MAF \leq 0.01 \sim 0.05$
  - Relatively new mutations
  - Often weak correlation with other genetic variants

## Why studying rare variants?

- Most of human variants are rare!

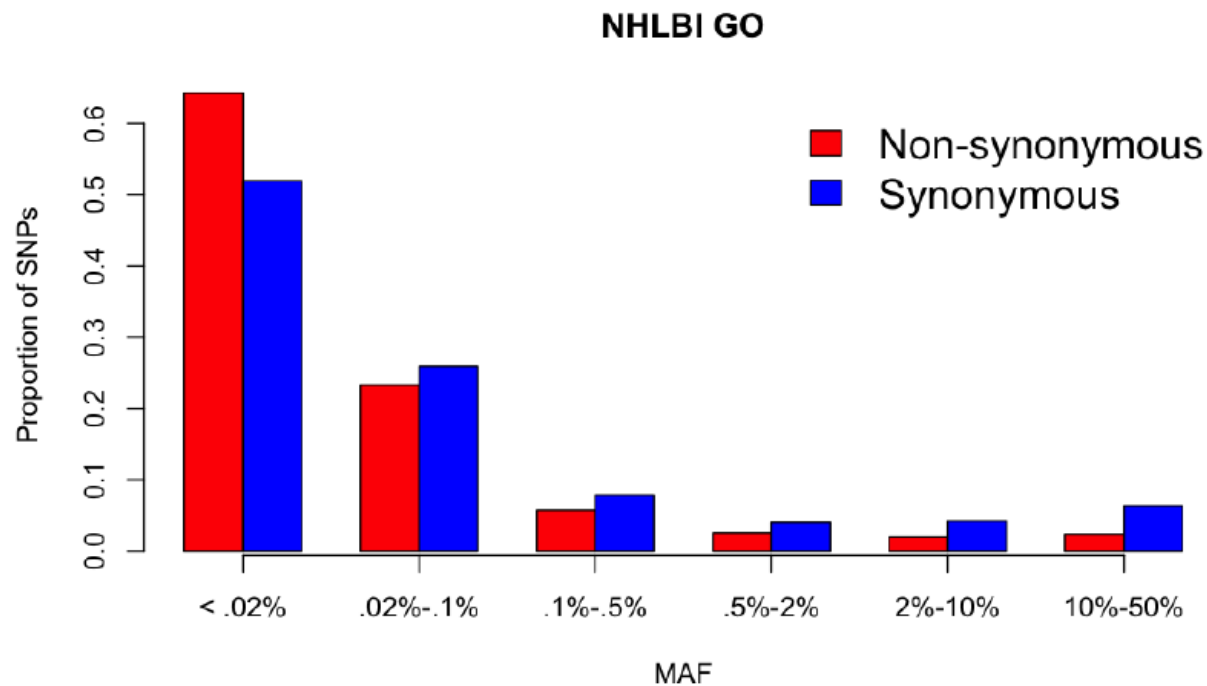


- But large samples are required (in terms of number of individuals) to observe rare variants!



## Why studying rare variants?

- Functional variants tend to be rare!



**Customizing GWAs for rare variants association analyses (future class)**

## 6 When effects become non-independent

### Confounding versus effect modification

- In an association study, if the strength of the association varies over different categories of a third variable, this is called effect modification. The third variable is changing the effect of the exposure.
- The effect modifier may be sex, age, an environmental exposure or a genetic effect.
- Effect modification is similar to interaction in statistics.
- There is no adjustment for effect modification. Once it is detected, stratified analysis can be used to obtain stratum-specific odds ratios.

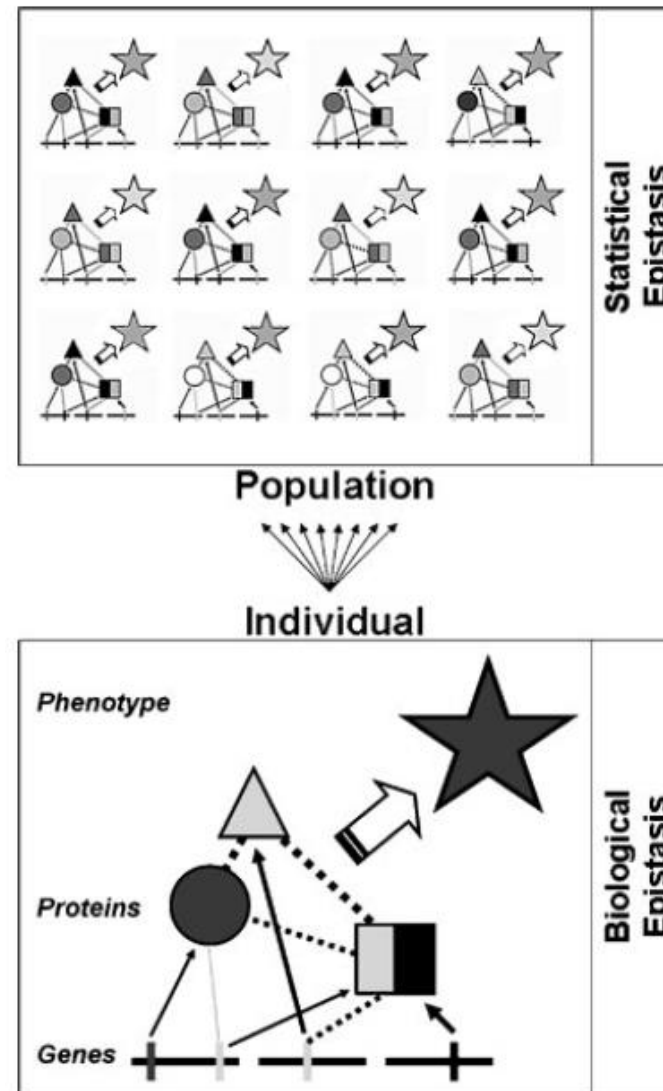
	<b>Locus Heterogeneity</b>	<b>Trait Heterogeneity</b>	<b>Gene-Gene Interaction</b>
<b>Definition</b>	when two or more DNA variations in distinct genetic loci are independently associated with the same trait	when a trait, or disease, has been defined with insufficient specificity such that it is actually two or more distinct underlying traits	when two or more DNA variations interact either directly (DNA-DNA or DNA-mRNA interactions), to change transcription or translation levels, or indirectly by way of their protein products, to alter disease risk separate from their independent effects
<b>Diagram</b>			
<b>Example One</b>	<b>Retinitis Pigmentosa</b> (RP, OMIM# 268000) - genetic variations in at least fifteen genes have been associated with RP under an autosomal recessive model. Still more have been associated with RP under autosomal dominant and X-linked disease models <sup>2</sup> ( <a href="http://www.sph.uth.tmc.edu/RetNet">http://www.sph.uth.tmc.edu/RetNet</a> )	<b>Autosomal Dominant Cerebellar Ataxia</b> (ADCA, OMIM# 164500) - originally described as a single disease, three different clinical subtypes have been defined based on variable associated symptoms, <sup>6,7</sup> and different genetic loci have been associated with the different subtypes <sup>8</sup>	<b>Hirschsprung Disease</b> (OMIM# 142623) - variants in the RET (OMIM# 164761) and EDNRB (OMIM# 131244) genes have been shown to interact synergistically such that they increase disease risk far beyond the combined risk of the independent variants <sup>12</sup>
<b>Example Two</b>	<b>Tuberous Sclerosis</b> (TS, OMIM# 191100) - out of families informative for linkage analysis, half have mutations in the TSC1 gene (located at 9q34) and the other half have mutations in the TSC2 gene (located at 16p13) <sup>3,4,5</sup>	<b>Autism</b> (OMIM# 209850) - parents and other relatives of autistic individuals often exhibit one or two, but not all three, of the requisite autistic symptomatologies, suggesting autism may be the co-occurrence of three distinct traits. <sup>9</sup> Using subset analysis, some success has been achieved identifying genes associated with one of the three symptomatologies but not as strongly with the broader autistic phenotype <sup>10,11</sup>	<b>Creutzfeldt-Jakob Disease</b> (CJD, OMIM# 123400) and Fatal Familial Insomnia (OMIM# 176640.0010) - the Met129Val polymorphism and Asp178Asn mutation in the PRNP gene (OMIM# 176640) interact, such that when the val129 polymorphism is on the same chromosome as the asn178, the phenotype is fatal familial insomnia <sup>13-19</sup>

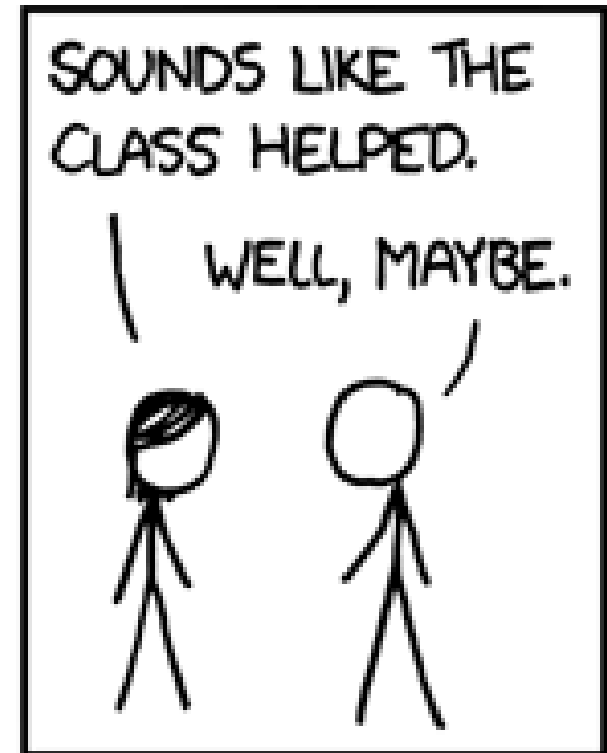
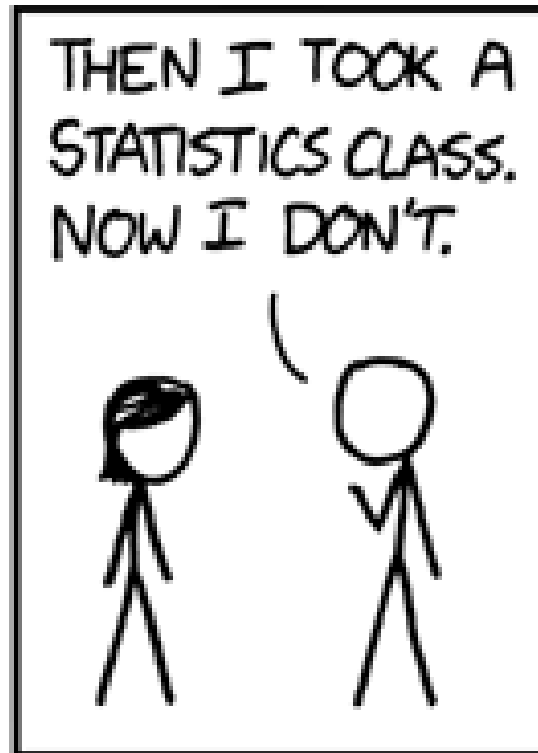
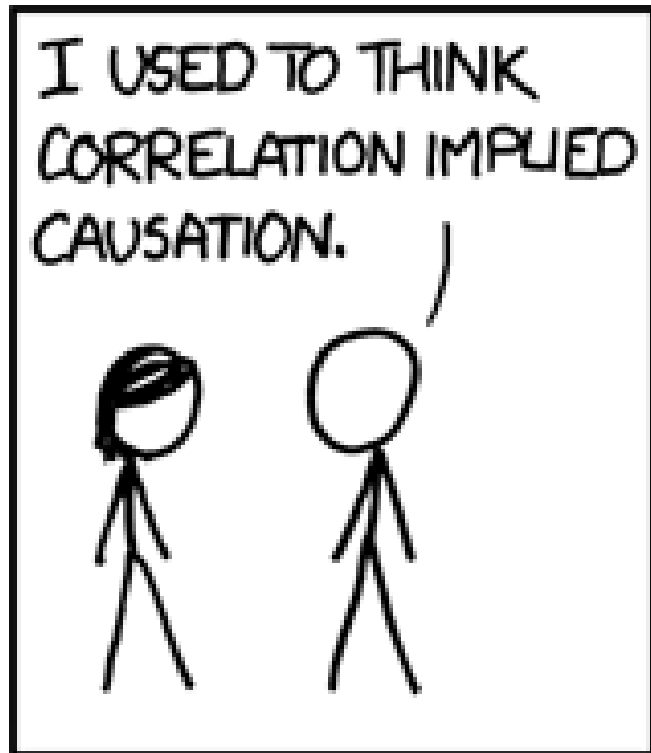
(Thornton-Wells et al. 2006)

## Biological vs statistical epistasis (future class)

(Moore et al. 2005)

**Figure 2.** The conceptual relationship between biological and statistical epistasis. Biological epistasis occurs at the level of the individual and involves DNA sequence variations (vertical bars), biomolecules (circle, square and triangle) and their physical interactions (dashed lines) giving rise to a phenotype (star) at a particular point in time and space (not shown). Statistical epistasis is a population phenomenon that is made possible by interindividual variability in genotypes, biomolecules and their physical interactions.





**Questions?**