Genetics and Bioinformatics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

Lecture 7: Beyond simple genome-wide Association studies

1 Capita selecta in GWAs

The role of regression analysis

Confounding: population stratification

2 When variants become rare

Impact

Remediation

3 When effects become non-independent

Impact and interpretation

Biological vs statistical epistasis

1 Capita selecta in GWAs

(slide Doug Brutlag 2010)

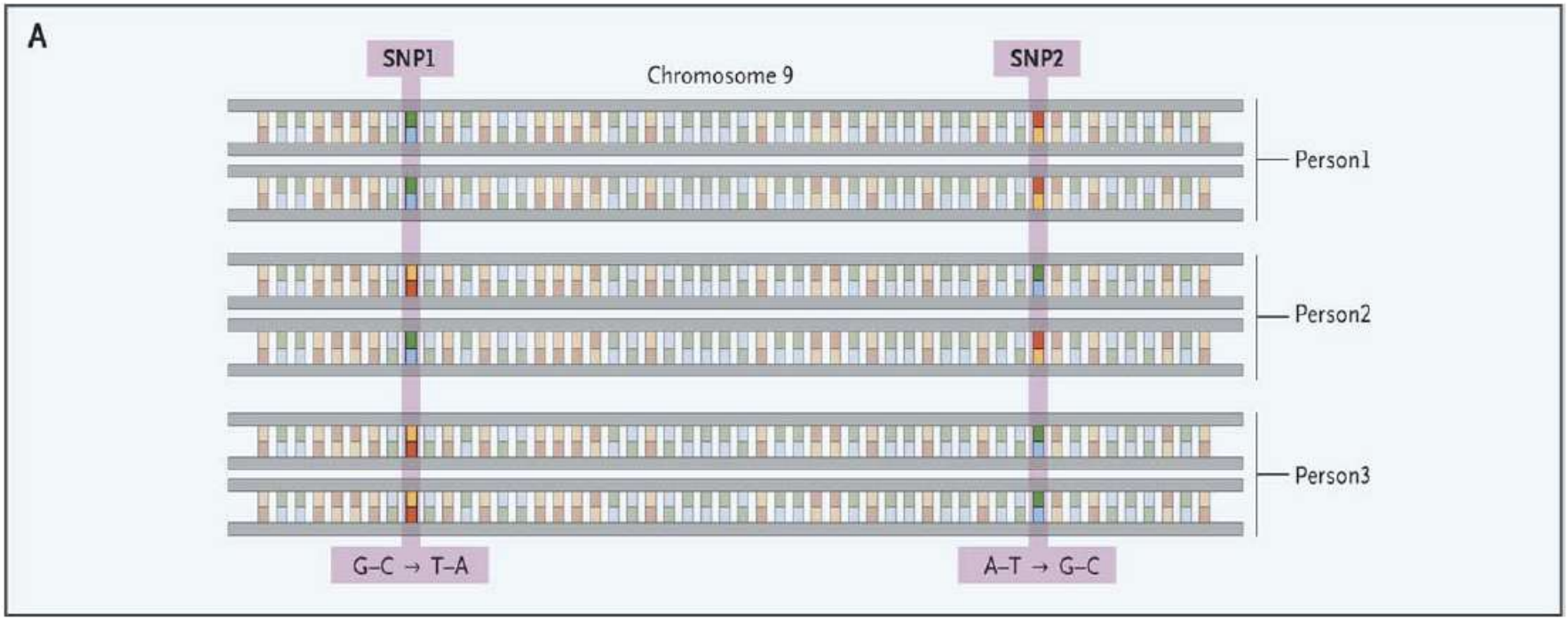
Definition (recap)

- A **genome-wide association study** is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular trait.

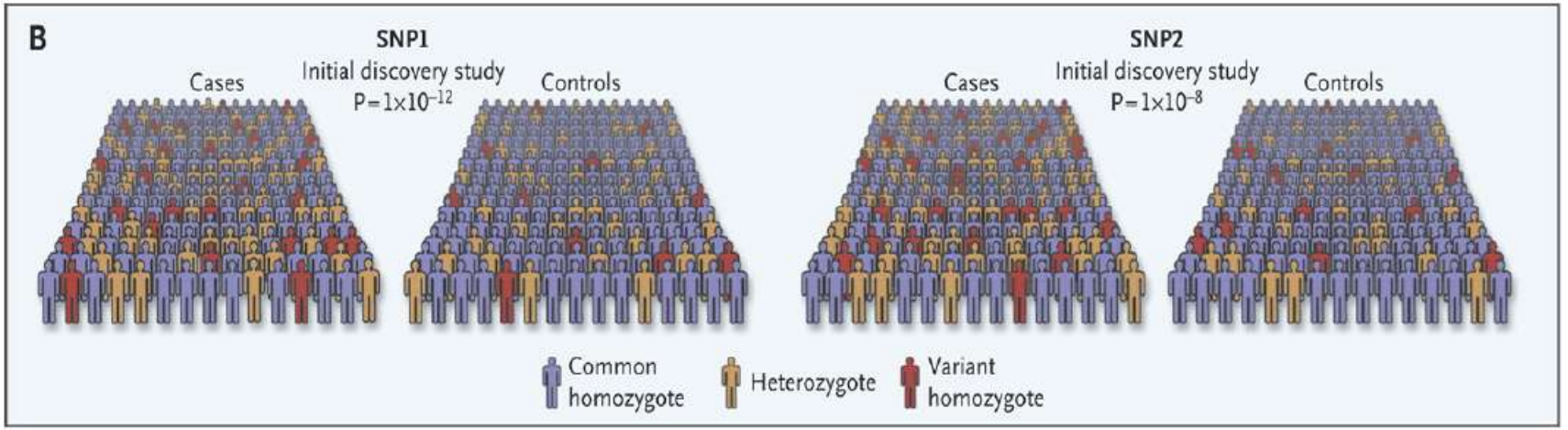
- A **trait** can be defined as a coded phenotype, a particular characteristic such as hair color, BMI, disease, gene expression intensity level, …

Genome-wide association studies in practice

The genome-wide association study is typically (but not solely!!!) based on a case–control design in which single-nucleotide polymorphisms (SNPs) across the human genome are genotyped ... (Panel A: small fragment)



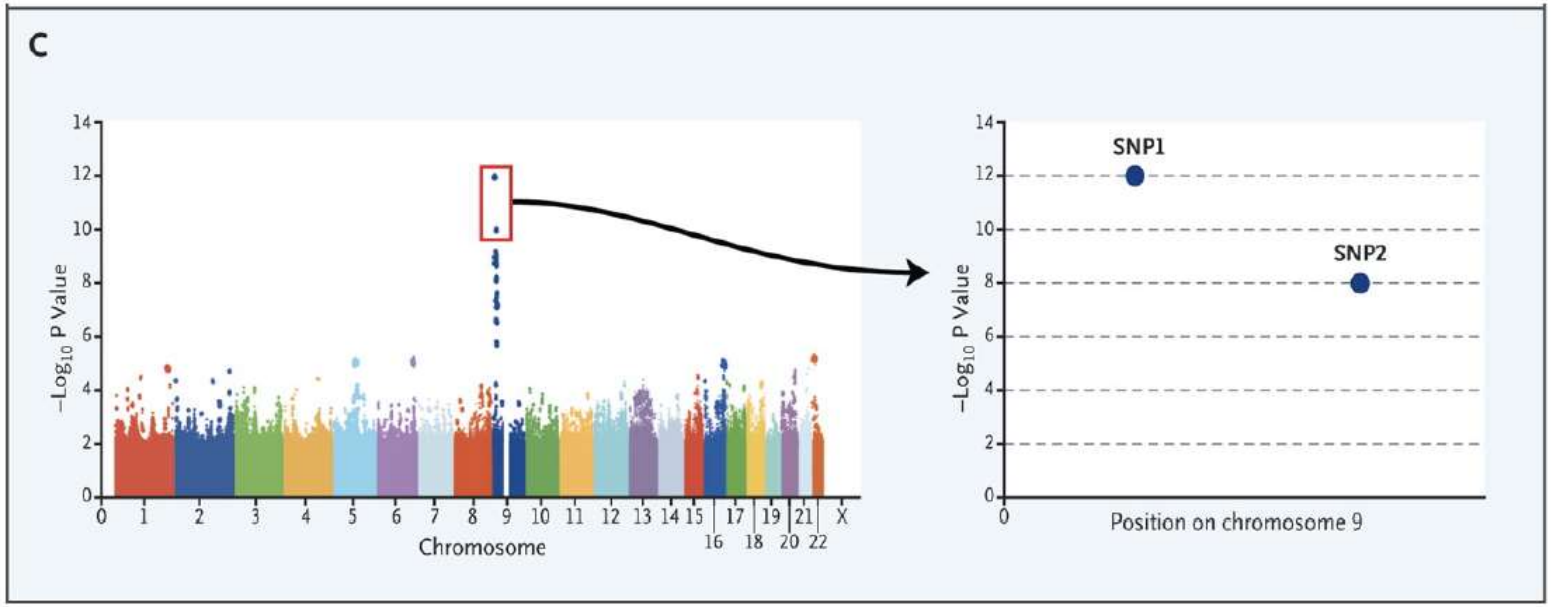
Genome-wide association studies in practice



- Panel B, the strength of association between each SNP and disease is calculated on the basis of the prevalence of each SNP in cases and controls. In this example, SNPs 1 and 2 on chromosome 9 are associated with disease, with P values of 10^{-12} and 10^{-8}, respectively

(Manolio 2010)

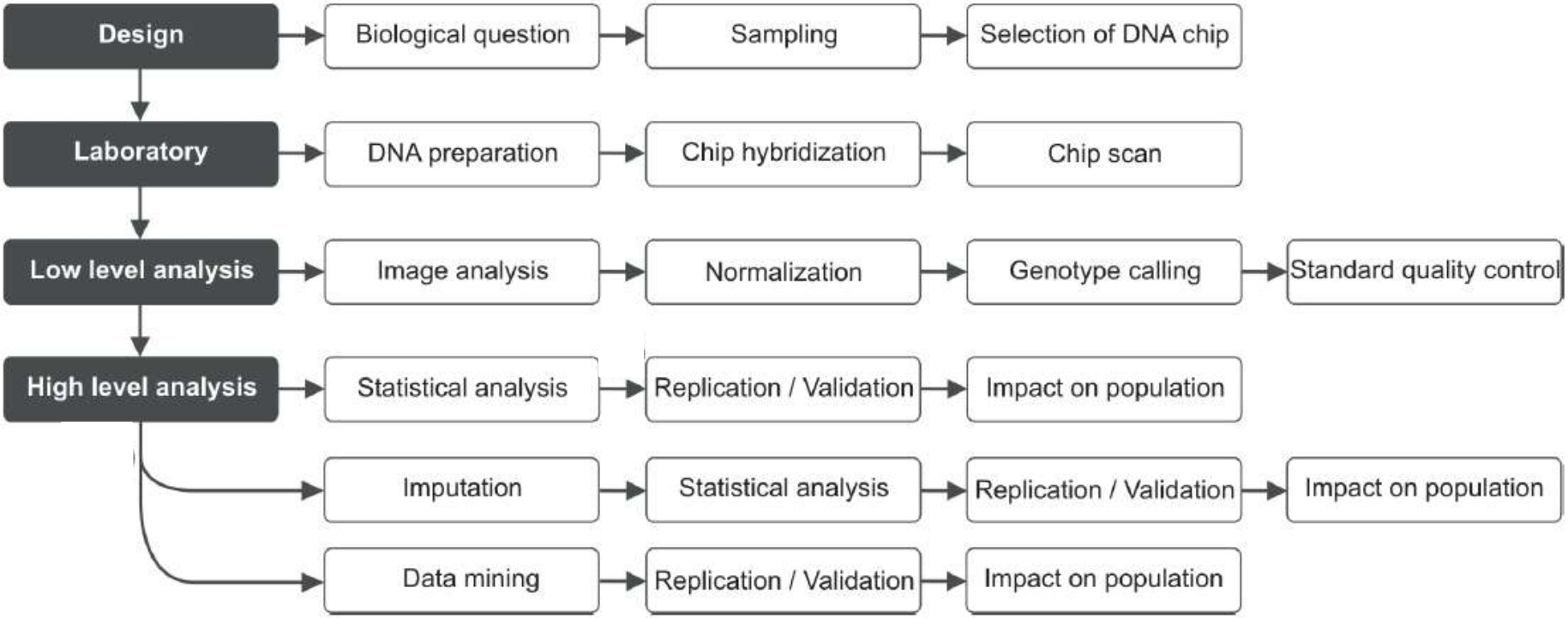
Genome-wide association studies in practice



- The plot in Panel C shows the P values for all genotyped SNPs that have survived a quality-control screen (each chromosome, a different color).
- 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)

Detailed flow of a genome-wide association study



(Ziegler 2009)

Standard file format for GWA studies

Standard data format: tped = transposed ped format file

FamID	PID	FID	MID	SEX	AFF	$SNP1_1$	$SNP1_2$	$SNP2_1$	$SNP2_2$
1	1	0	0	1	1	A	A	G	T
2	1	0	0	1	1	A	C	T	G
3	1	0	0	1	1	C	C	G	G
4	1	0	0	1	2	A	C	T	T
5	1	0	0	1	2	C	C	G	T
6	1	0	0	1	2	C	C	T	T

ped file

Chr	SNP name	Genetic distance	Chromosomal position
1	SNP1	0	123456
1	SNP2	0	123654

map file

Standard file format for GWA studies (continued)

Chr	SNP	Gen. dist.	Pos	PID 1		PID 2		PID 3		PID 4		PID 5		PID 6	
1	SNP1	0	123456	A	A	A	C	C	C	A	C	C	C	C	C
1	SNP2	0	123654	G	T	G	T	G	G	T	T	G	T	T	T

tped file

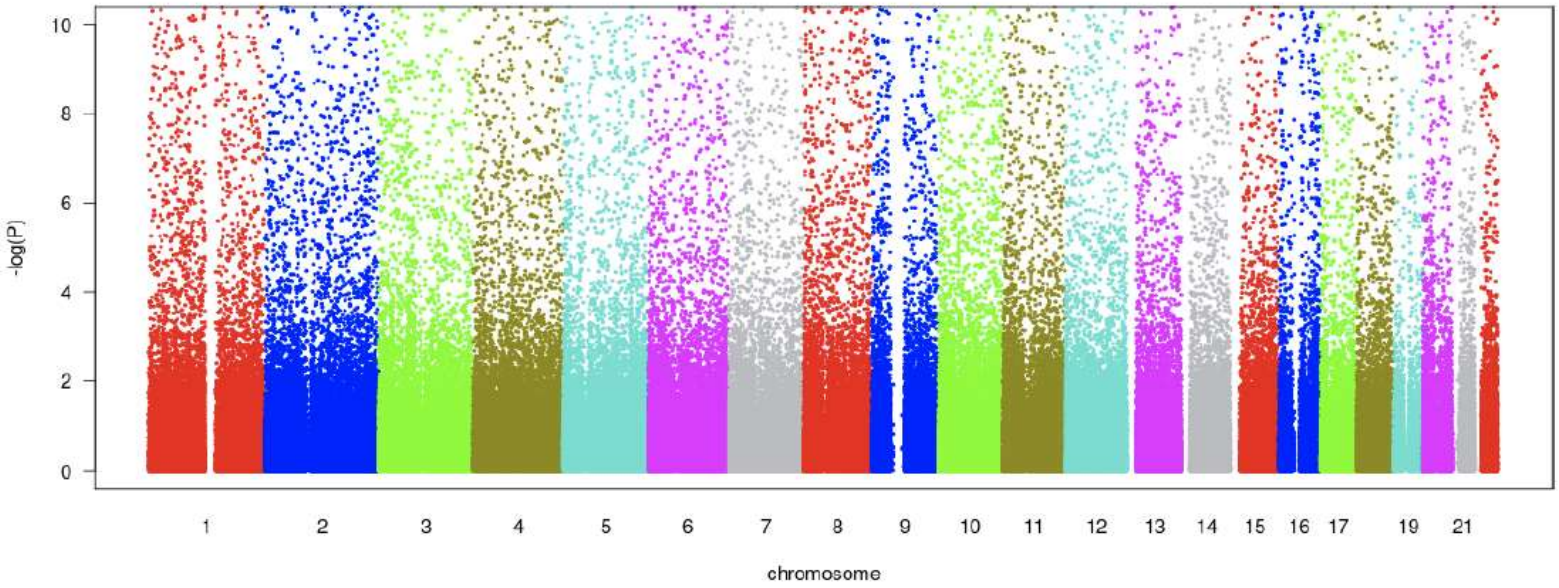
tfam file: First 6 columns of standard ped file

FamID	PID	FID	MID	SEX	AFF
1	1	0	0	1	1
2	1	0	0	1	1
3	1	0	0	1	1
4	1	0	0	1	2
5	1	0	0	1	2
6	1	0	0	1	2

tfam file

Why is quality control (QC) important?

BEFORE QC → true signals are lost in false positive signals

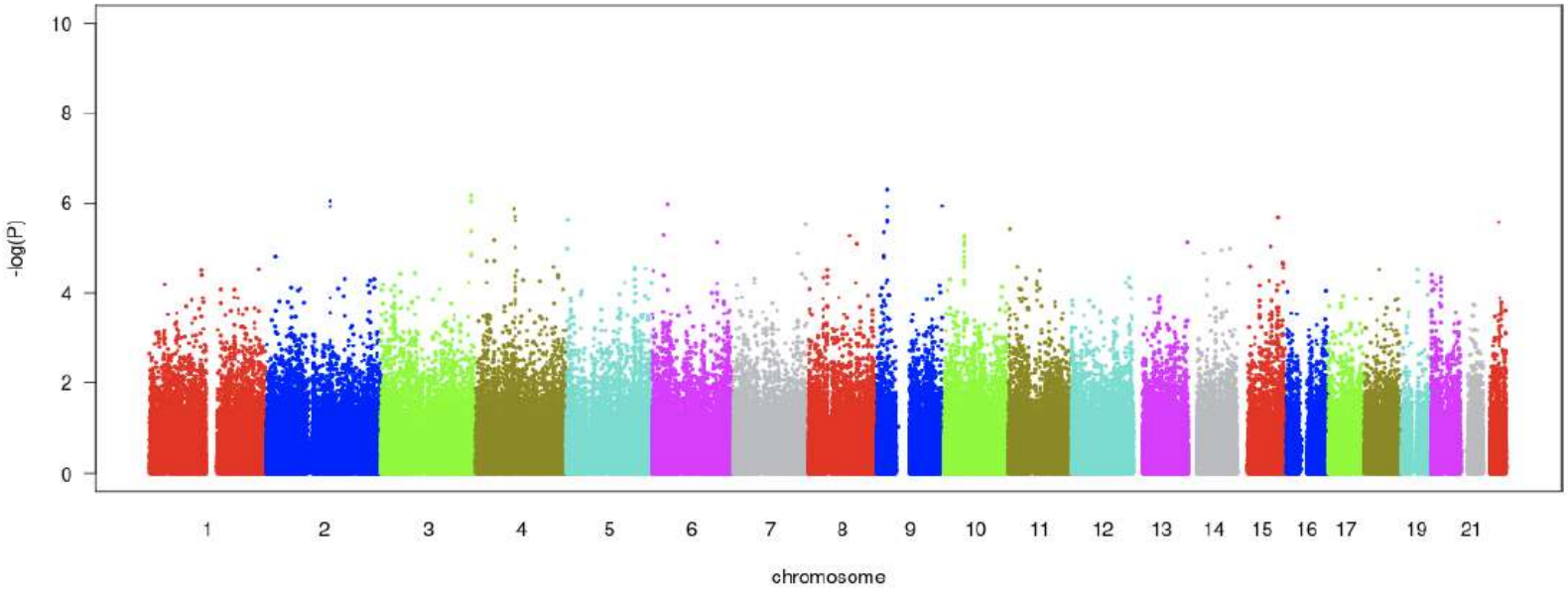


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

(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	≥ 97%
	Cryptic relatedness	Study specific
	Ethnic origin	Study specific; visual inspection of principal components
	Heterozygosity	Mean ± 3 std.dev. over all samples
	Heterozygosity by gender	Mean ± 3 std.dev. within gender group
SNP level	MAF	≥ 1%
	MiF	≤ 2% in any study group, e.g., in both cases and controls
	MiF by gender	≤ 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)

The role of regression analysis

- Galton used the following equation to explain the phenomenon that sons of tall fathers tend to be tall but not as tall as their fathers while sons of short fathers tend to be short but not as short as their fathers:

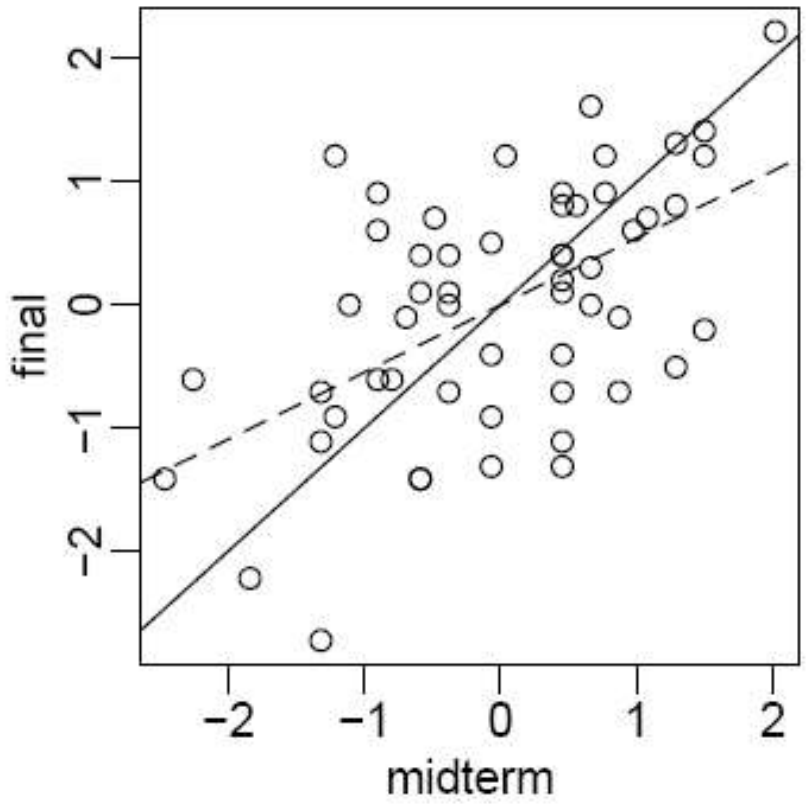
$$\frac{y - \bar{y}}{SD_y} = r\frac{(x - \bar{x})}{SD_x}.$$

This effect is called the regression effect.

- We can illustrate this effect with some data on scores from a course
  - When we scale each variable to have mean 0 and SD 1 so that we are not distracted by the relative difficulty of each exam and the total number of points possible.

The use of regression analysis

- **regression line** goes through (mean x, mean y)



(Faraway 2002)

The use of regression analysis

- **Regression analysis** is used for explaining or modeling the relationship between a single variable Y, called the response, output or dependent variable, and one or more predictor, input, independent or explanatory variables, X_1, …, X_p.
- When p=1 it is called simple regression but when p > 1 it is called multiple regression or sometimes multivariate regression.
- When there is more than one Y, then it is called multivariate multiple regression
- Regression analyses have several possible objectives including
  - Prediction of future observations.
  - Assessment of the effect of, or relationship between, explanatory variables on the response.
  - A general description of data structure

The linear regression model

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

- y: response variable.
- $x_1, \ldots, x_k$: regressor variables, independent variables.
- $\beta_0, \beta_1, \ldots, \beta_k$: regression coefficients.
- ϵ: 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.

Linear vs non-linear

Linear Models Examples:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$
$$y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + \epsilon$$
$$\log y = \beta_0 + \beta_1 \left(\frac{1}{x_1}\right) + \beta_2 \left(\frac{1}{x_2}\right) + \epsilon$$

Nonlinear Models Examples:

$$y = \beta_0 + \beta_1 x_1^{\gamma_1} + \beta_2 x_2^{\gamma_2} + \epsilon$$
$$y = \frac{\beta_0}{1 + e^{\beta_1 x_1}} + \epsilon$$

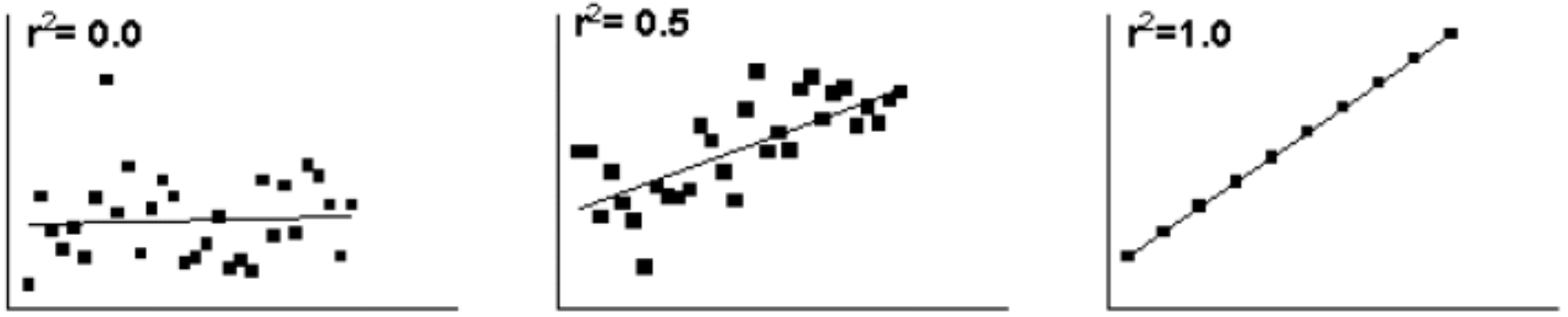
Regression inference

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

- Least square estimation of the regression coefficients. $b = (X^T X)^{-1} X^T y$.
- Variance estimation for σ^2: s^2.
- Coefficient of Determination. R^2.
- Partial F test or t-test for $H_0 : \beta_j = 0$.

Coefficient of determination = squared correlation coefficient

- The value r^2(also denoted as R^2) is a fraction between 0.0 and 1.0, and has no units. An r^2 value of 0.0 means that knowing X does not help you predict Y.
- There is no linear relationship between X and Y, and the best-fit line is a horizontal line going through the mean of all Y values. When
- r^2 equals 1.0, all points lie exactly on a straight line with no scatter. Knowing X lets you predict Y perfectly.



General linear test approach

- The full model:

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

- Fit the model by f.i. 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):

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

General linear test approach

- Next we consider a null hypothesis H_0 of interest:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

- The model when H0 holds is called **the reduced or restricted model.** When $\beta_1 = 0$, then the regression model 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)

- Question: which error sums of squares will be smaller? SSE(F) or SSE(R)

General linear test approach

- The logic now is to compare both SSEs. The actual test statistic is a function of SSE(R)-SSE(F):

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

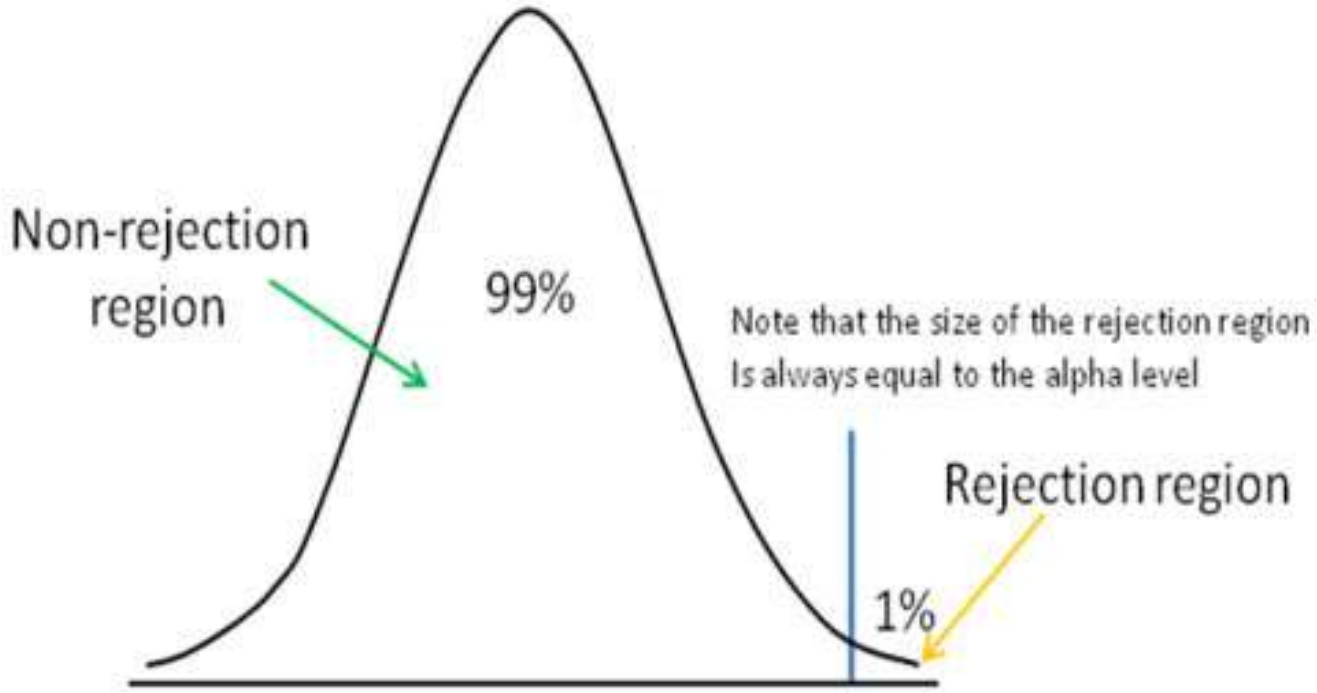
  which follows an F distribution when H_0 holds

- 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

Recall: alpha levels



(Partial) tests in GWAs

- **Example 1:**

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

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

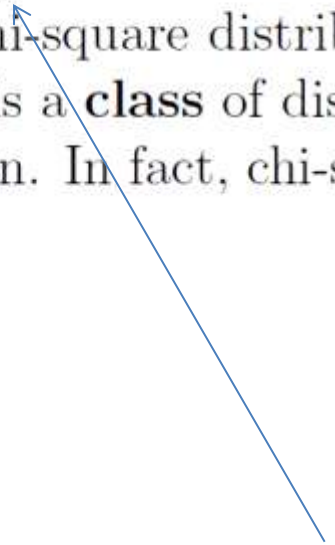
- **Example 2** (see more about this later)**:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 PC_1 + \beta_3 PC_2 + \varepsilon$$

  - $H_0: \beta_1 = 0$
  - $H_1: \beta_1 \neq 0$
  - $\mathrm{df_F} = \mathrm{n} - 4$
  - $\mathrm{df_R} = \mathrm{n} - 3$

Distributional relationships: F, t, chi-squared

$Z_1, Z_2, \ldots, Z_\kappa$ iid $\mathrm{N}(0,1) \Rightarrow X^2 \equiv Z_1^2 + Z_2^2 + \ldots + Z_\kappa^2 \sim \chi^2_\kappa$.
Specifically, if $\kappa = 1$, $Z^2 \sim \chi^2_1$. 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 $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$, then $Z_j \equiv (X_j - \mu)/\sigma \sim N(0,1), j = 1, \ldots, n$. We know, from a previous context, that $\Sigma_1^n Z_j^2 \sim \chi_n^2$, or equivalently,

$$\sum_{j=1}^{n} \{\frac{X_j - \mu}{\sigma}\}^2 = \frac{\Sigma_1^n (X_j - \mu)^2}{\sigma^2} \sim \chi_n^2,$$

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

$$\frac{\Sigma_1^n (X_j - \overline{X})^2}{\sigma^2} \sim \chi_{n-1}^2.$$

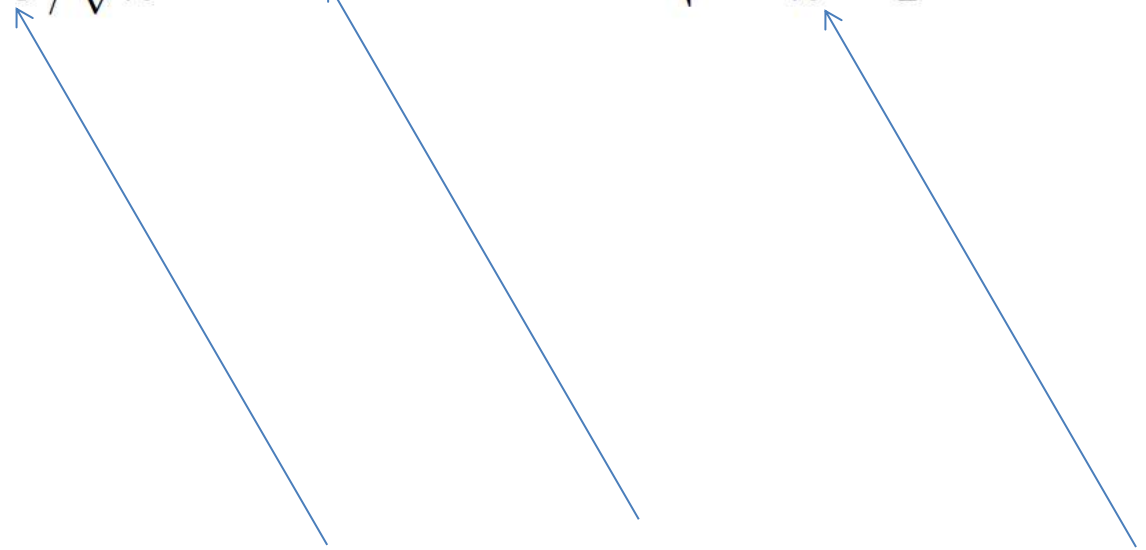
Distributional relationships: F, t, chi-squared

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

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

When σ is unknown,

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



Note that

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

Combining (3) and (4) gives

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

or, in general,

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

Distributional relationships: F, t, chi-squared

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

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

Implication to example 1:

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

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

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

- $F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} : \frac{SSE(F)}{df_F} = \frac{b_1^2}{s^2(b_1)} = (t^*)^2$

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

Regression analysis in R

- The basic syntax for doing regression in R is lm(Y~model) to fit linear models
- The R function glm() can be used to fit generalized linear models (i.e., when the response is not normally distributed).
- **Linear regression [and logistic regression]:** special type of regression models you can fit using lm() [and glm()] respectively.

Model assumptions for linear regression

- 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)

Use of `lm()` in genetics

For a continuous outcome,

```
lm(outcome ~ genetic.predictor, [...]  )
```

estimates the association between `outcome` and `predictor`

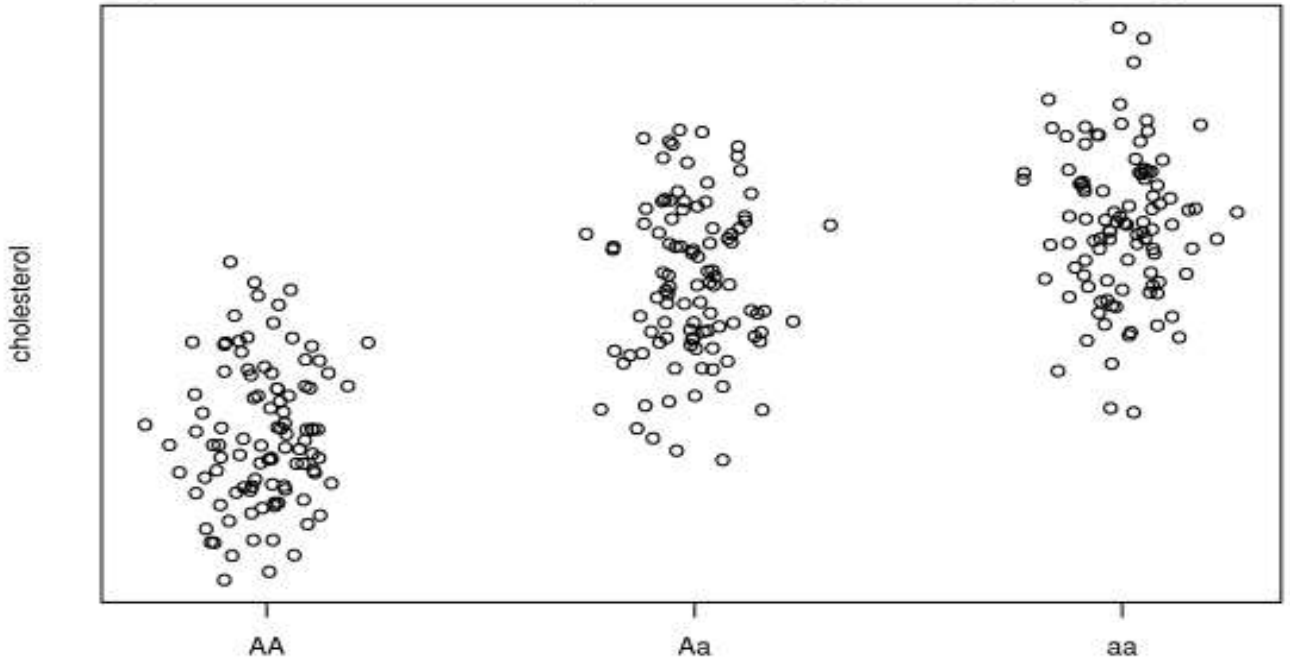
Model Description	`predictor`	Common name
Number of minor alleles	`(g=='Aa') + 2*(g=='aa')` or `as.numeric(g)`	Additive
Presence of minor allele	`(g=='Aa') | (g=='aa')`	Dominant
Homozygous for minor allele	`g=='aa'`	Recessive
Distinct effects for hetero/homozygous	`factor(g)`	2 parameter, or "2 df"

One SNP: different encodings imply different genetic models

	Coding scheme for statistical modeling/testing					
Indiv. genotype	**X1**	**X1**	**X2**	**X1**	**X1**	**X1**
	Additive coding	**Genotype coding (general mode of inheritance)**		**Dominant coding (for a)**	**Recessive coding (for a)**	**Advantage Heterozygous**
AA	**0**	**0**	**0**	**0**	**0**	**0**
Aa	**1**	**1**	**0**	**1**	**0**	**1**
aa	**2**	**0**	**1**	**1**	**1**	**0**

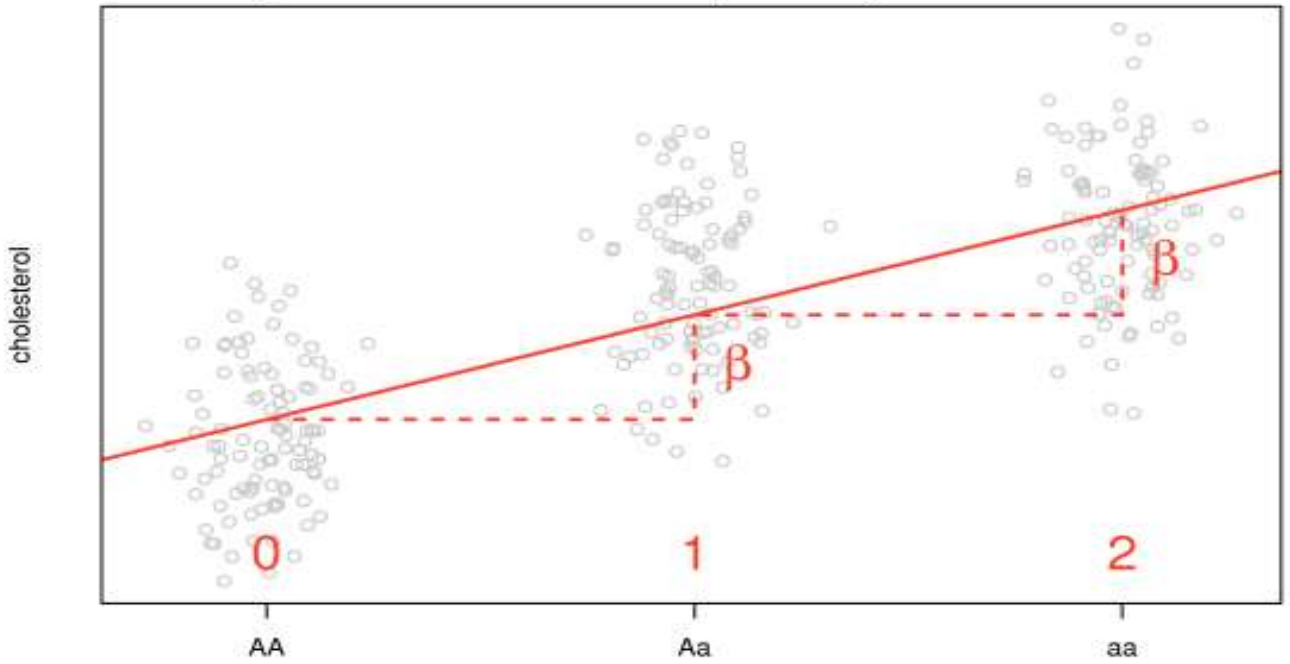
Use of `lm()` in genetics

Some data; cholesterol levels plotted by genotype (single SNP)



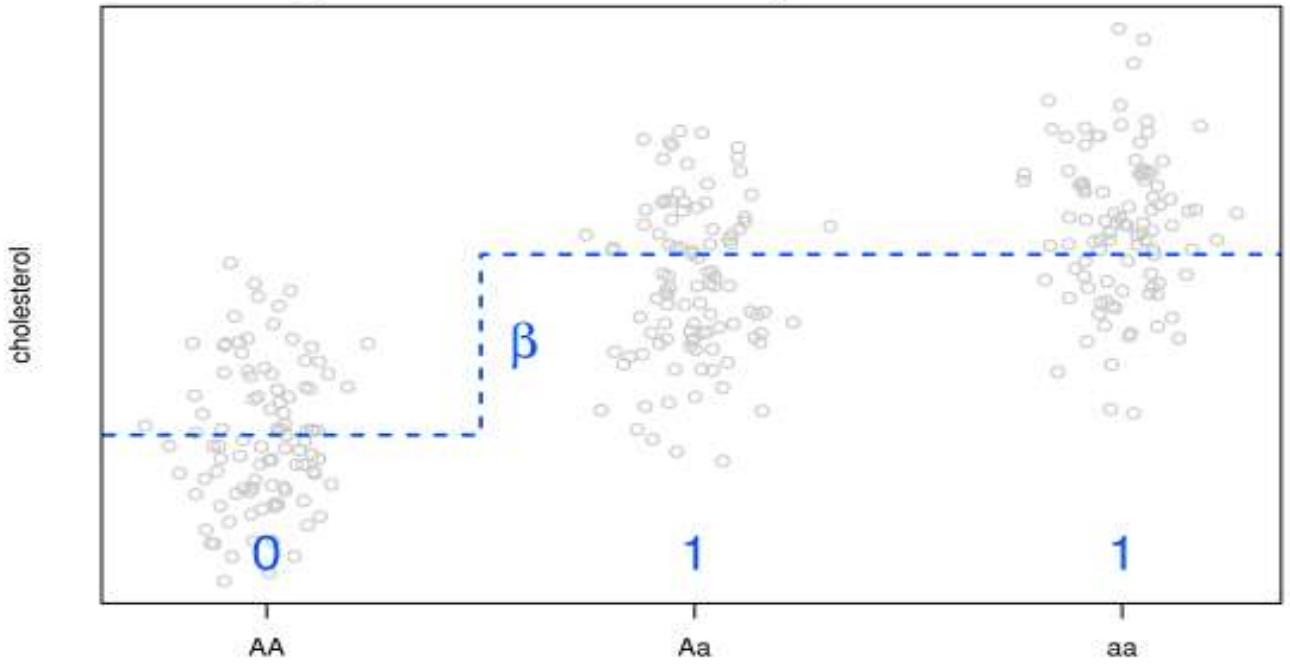
Use of `lm()` in genetics

Additive model (the most commonly used)



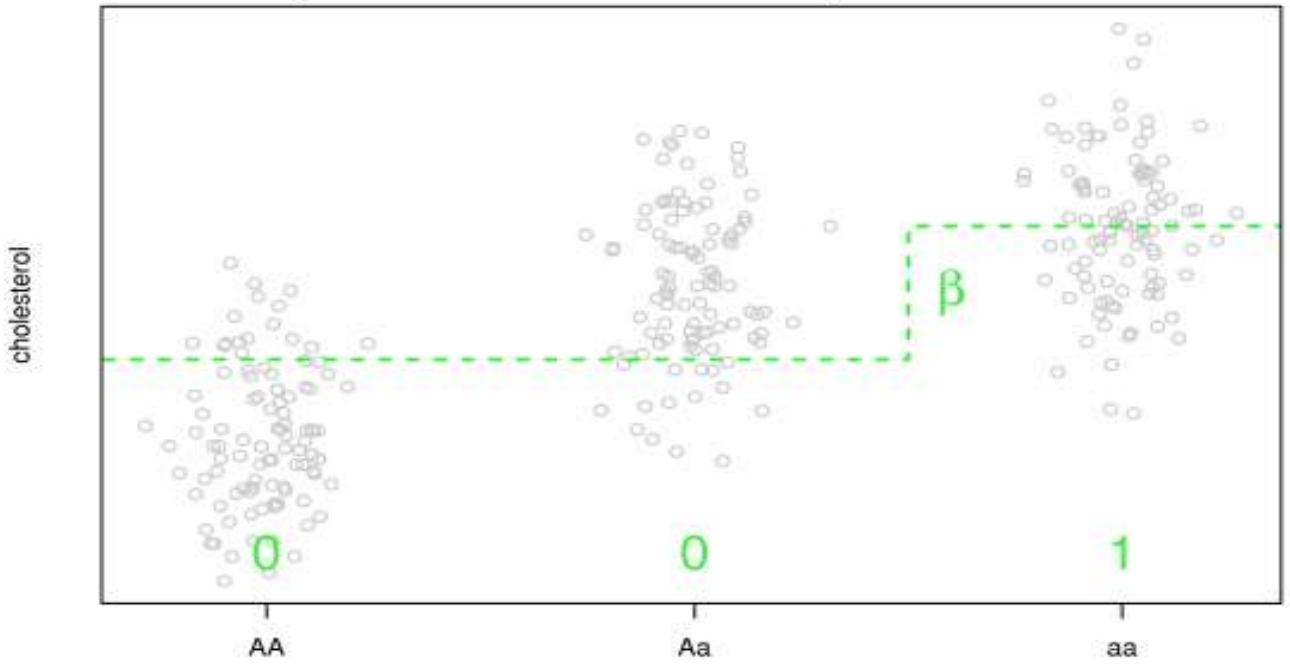
Use of `lm()` in genetics

Dominant model (best fit to this data)



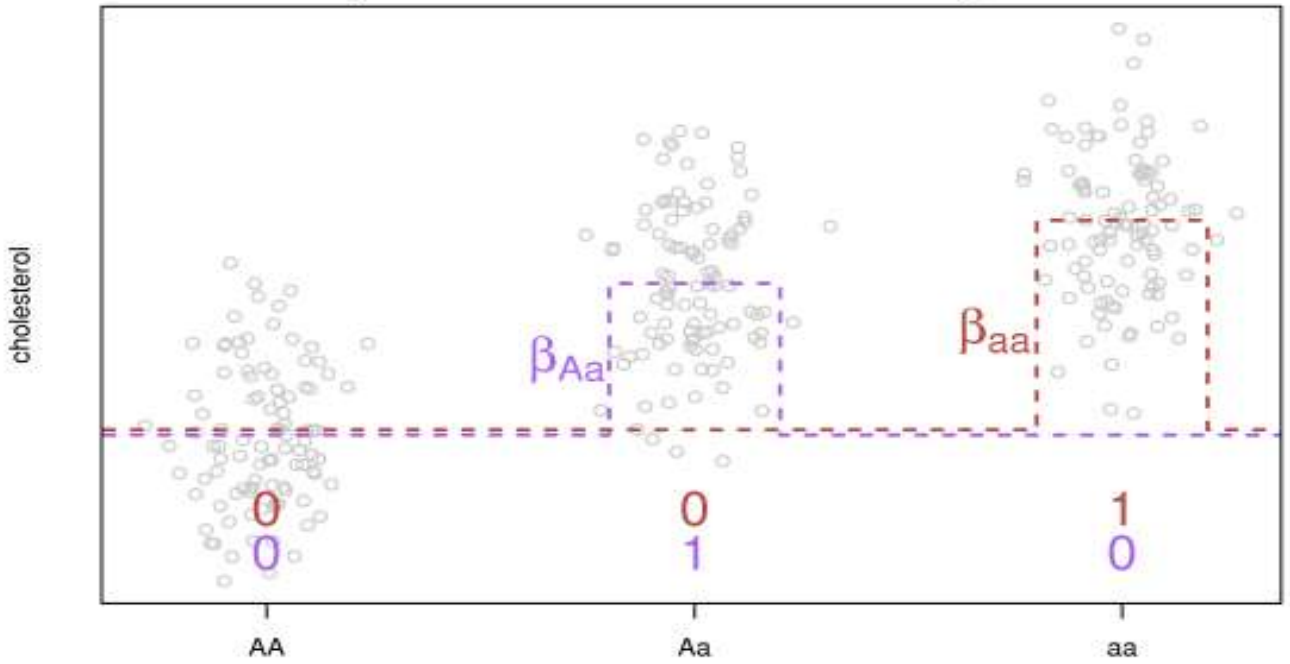
Use of `lm()` in genetics

Recessive model (least stable for rare aa)



Use of `lm()` in genetics

2 parameter model (robust but can be overkill)



Logistic regression (dichotomous traits; cases and controls)

In linear regression one equates

$$E[Y] = \beta_0 + \beta_1 X_1$$

In logistic regression one equates

$$E[Y] = P(Y = 1) = f(\beta_0 + \beta_1 X_1)$$

- y is binary: logistic regression.

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}}$$

- y is measured on an ordinal scale: ordinal logistic regression.
- y is measured on non-ordered scale: multinomial logistic regression.
- y is counts: Poisson or Negative Binomial regression.

Logistic regression (dichotomous traits; cases and controls)

$$E[Y] = P(Y = 1) = f(\beta_0 + \beta_1 X_1)$$

$$f^{-1}(E[Y]) = f^{-1}(P(Y = 1)) = (\beta_0 + \beta_1 X_1)$$

$$f^{-1}(E[Y]) = logit\big(P(Y = 1)\big) == \log\left(\frac{P(Y=1)}{1 - P(Y=1)}\right)$$

$$\log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \beta_1 X_1$$

Log(Odds|$X_{1\,==1}$) = $\beta_0 + \beta_1 1$

Log(Odds|$X_{1\,==0}$) = β_0

Log(OR) = β_1

Logistic regression (formal formulation)

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

- $X = (X_1, X_2, ..., X_k)$ 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:

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

Assumptions:

- The data Y_1, Y_2, ..., 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)

Parameter Estimation:

The *maximum likelihood estimator* (MLE) for (β_0, β_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, see sections 5.5.4-5.5.5.

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

Logistic regression test approach

- **Example 1:**

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

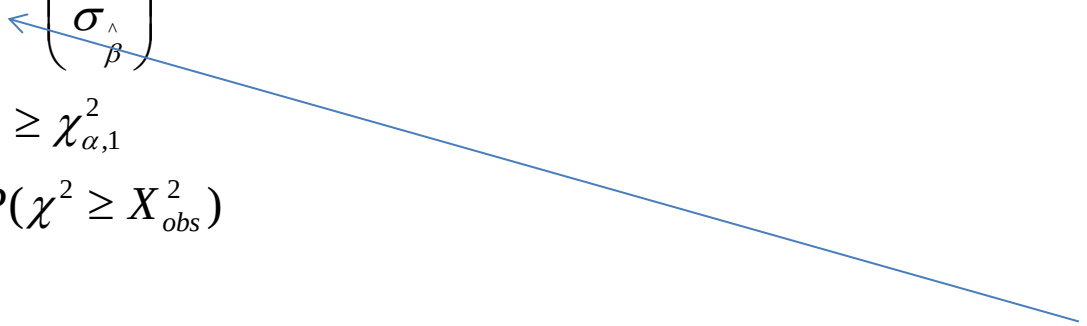
  - $H_0: \beta_1 = 0$
  - $H_1: \beta_1 \neq 0$

Large-sample "Wald test"**:**

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

$$R.R.: X^2_{obs} \geq \chi^2_{\alpha,1}$$

$$P-val: P(\chi^2 \geq X^2_{obs})$$



The Wald statistic

In the univariate case, the Wald statistic is

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

which is compared against a chi-squared distribution.

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

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

where $\operatorname{se}(\widehat{\theta})$ 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 accept or reject 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.

	AA	**Aa**	**aa**
Cases			
Controls			

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.

	AA	**Aa**	**aa**
Cases	E_{11}	E_{12}	E_{13}
Controls	E_{21}	E_{22}	E_{23}

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

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

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.

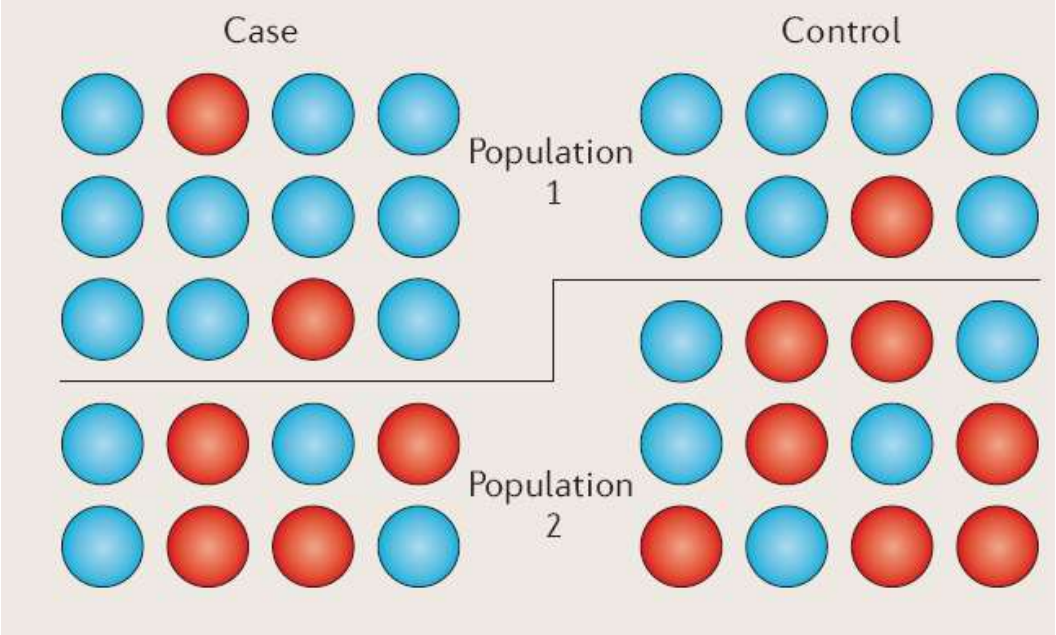
- **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic, which can be proven to follow a chi-square distribution with degrees of freedom as derived before. The null hypothesis is rejected when the P-value is less than the pre-stated significance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

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

Confounding: population stratification

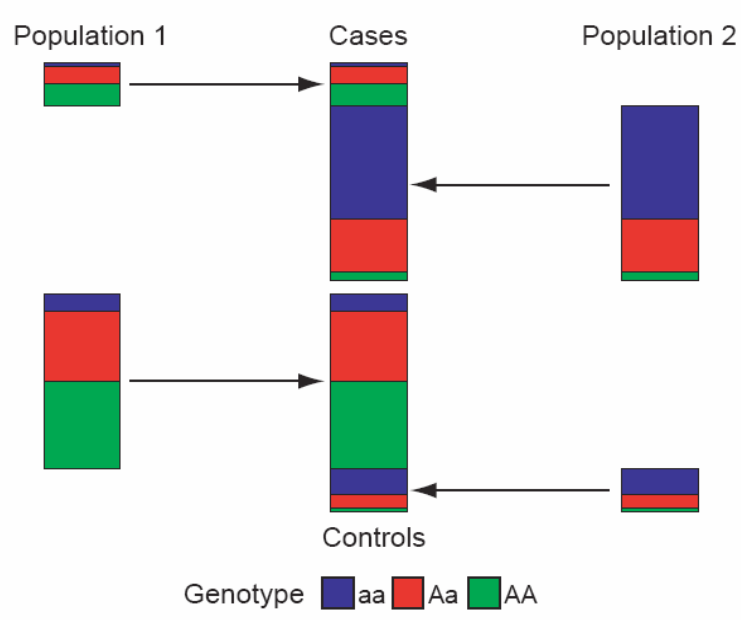
What is spurious association?

- **Spurious association** refers to false positive association results due to not having accounted for population substructure as a confounding factor in the analysis



What is spurious association?

- Typically, there are two characteristics present:
  - A difference in proportion of individual from two (or more) subpopulation in case and controls
  - Subpopulations have different allele frequencies at the locus.



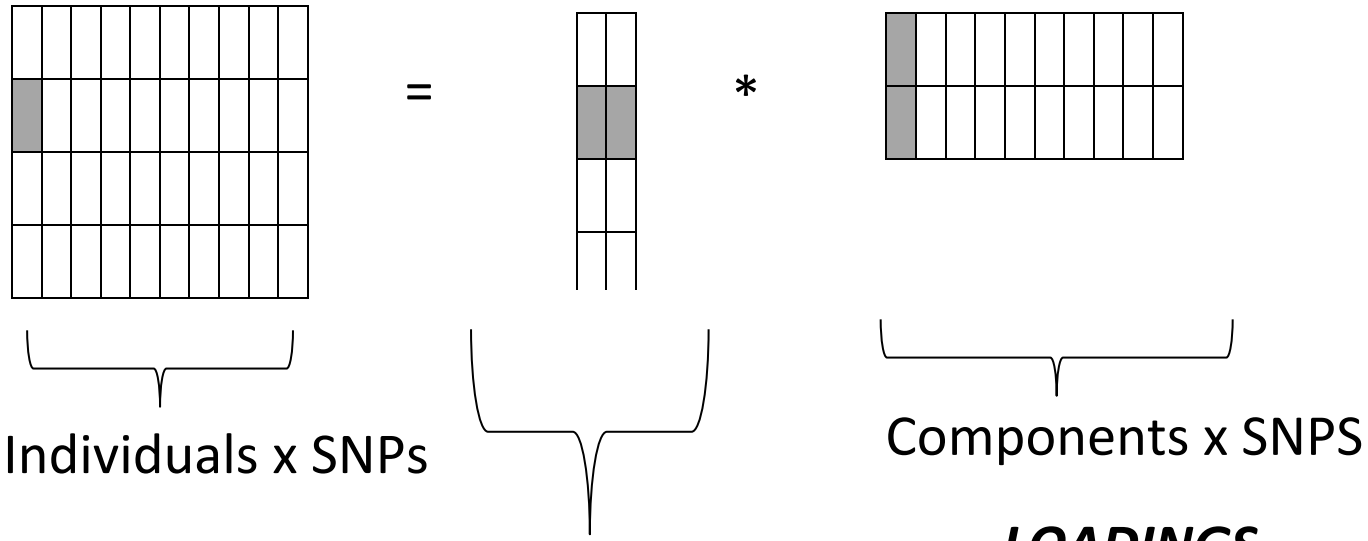
What are typical methods to deal with population stratification?

- Methods to deal with spurious associations generated by population structure generally require a number (at least >100) of widely spaced null SNPs that have been genotyped in cases and controls in addition to the candidate SNPs.

- These methods large group into:

  - **Principal components**
  - Structured association methods: “First look for structure (population clusters) and **second** perform an association **analysis** conditional on the cluster allocation”
  - **Genomic control methods**: “**First analyze** and second downplay association test results for over optimism” → see later

Principal components



- Principal components analysis (PCA) is one of a family of techniques for taking high-dimensional data, and using the dependencies between the variables to represent it in a more tractable, lower-dimensional form, without losing too much information.

- Data distribution : inputs in regression analysis

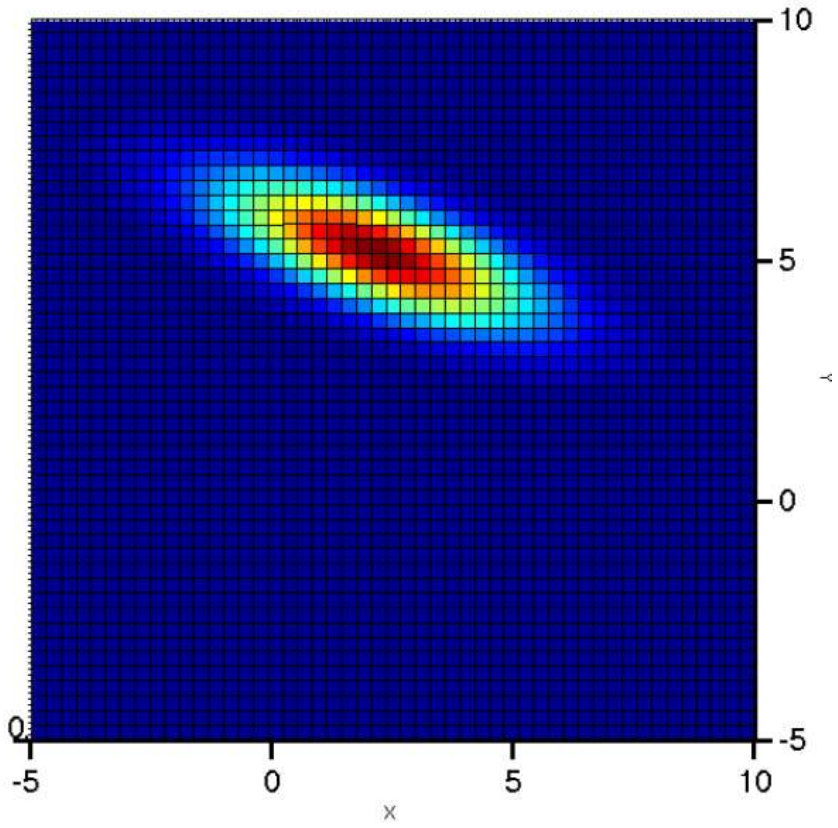


Figure: Gaussian PDF

- Uncorrelated projections of principal variation

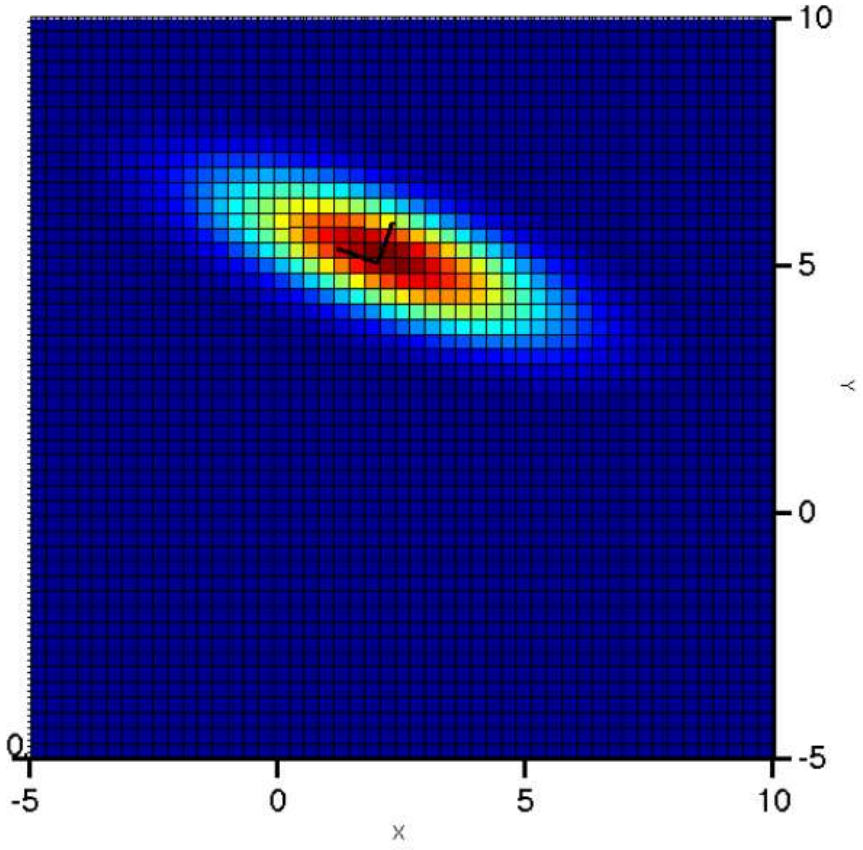


Figure: Gaussian PDF with PC eigenvectors

- PCA rotation

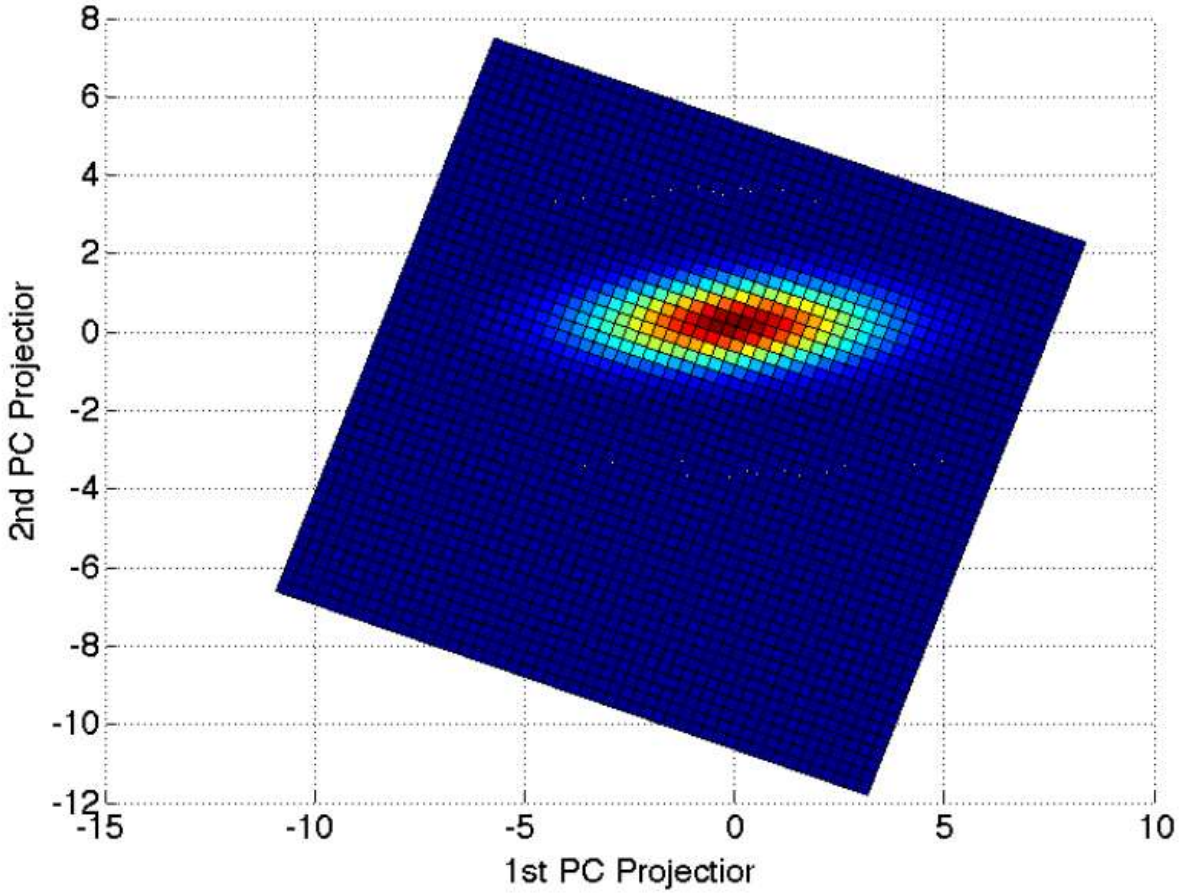


Figure: PCA Projected Gaussian PDF

PCA in a nutshell

Notation

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

Procedural description

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

Assumption and More Notation

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

Shortcut to solution

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

Derivation of PCA

First Step

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

Constrained maximization - method of Lagrange multipliers

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

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

w.r.t. to $\boldsymbol{\alpha}_k$ by differentiating w.r.t. to $\boldsymbol{\alpha}_k$.

Constrained maximization - method of Lagrange multipliers

- This results in

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

- This should be recognizable as an eigenvector equation where $\boldsymbol{\alpha}_k$ is an eigenvector of $\Sigma_b f$ and λ_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

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

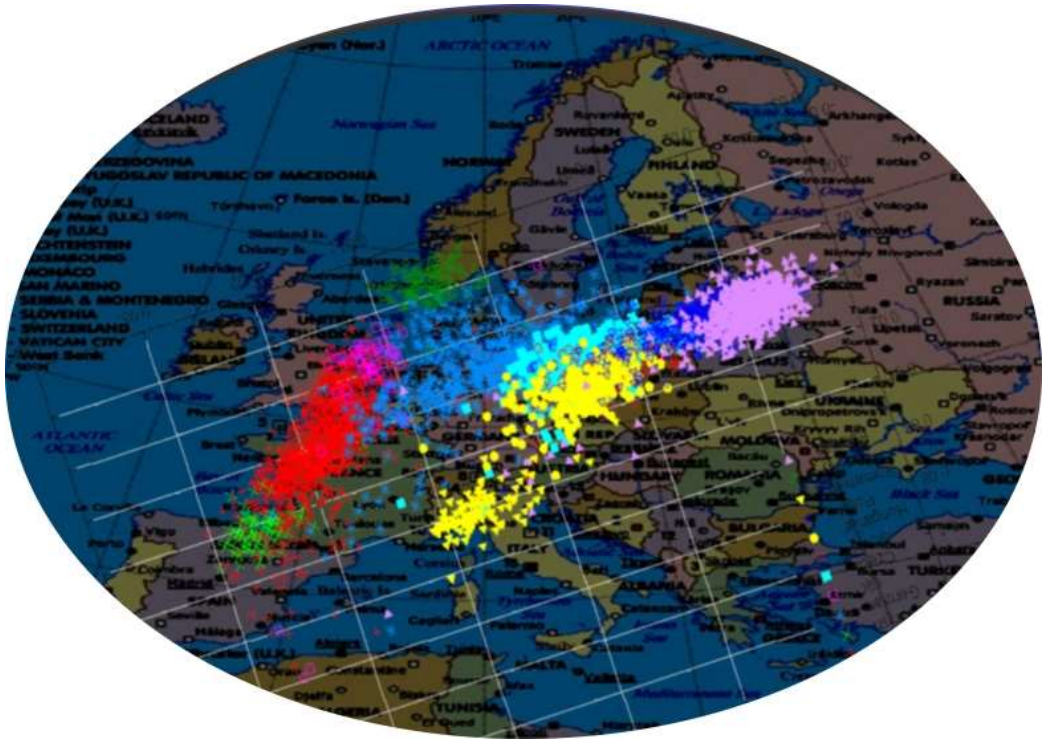
  then we should choose λ_k to be as big as possible. So, calling λ_1 the largest eigenvector of Σ and $\boldsymbol{\alpha}_1$ the corresponding eigenvector then the solution to

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

  is the 1^{st} principal component of $\mathbf{x}$.
- In general $\boldsymbol{\alpha}_k$ will be the k^{th} PC of $\mathbf{x}$ and $\text{Var}(\boldsymbol{\alpha}'\mathbf{x}) = \lambda_k$

Principal components in population genetics

- In European data, the first 2 principal components “nicely” reflect continuous axes of variation due to shared ancestry



Y-axis: PC2 (6% of variance); X-axis: PC1 (26% of variance)

Principal components in population genetics

- **Example 2:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 PC_1 + \beta_3 PC_2 + \varepsilon$$

  - $H_0: \beta_1 = 0$
  - $H_1: \beta_1 \neq 0$
  - $\mathrm{df_F} = \mathrm{n} - 4$
  - $\mathrm{df_R} = \mathrm{n} - 3$

Genomic control

- In Genomic Control (GC), a 1-df association test statistic is computed at each of the null SNPs, and a parameter λ 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 λ > 1 the test statistics are divided by λ.

  - Under H_0 of no association p-values uniformly distributed
  - In case of population stratification: inflation of test statistics
  - $\hat{\lambda} = \frac{\text{median}(\chi_1^2, \chi_2^2, \ldots, \chi_L^2)}{\text{median}(\mathcal{L}(\chi_1^2))} = \frac{\text{median}(\chi_1^2, \chi_2^2, \ldots, \chi_L^2)}{0.456}$
  - $\chi_{GC}^2 = \chi^2 / \hat{\lambda}$

```
> 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
```

2 When variants become rare

Impact

… on tests for association between trait and SNP

- Fill in the table below and perform a chi-squared test for independence between rows and columns

	AA	**Aa**	**aa**
Cases			
Controls			

Sum of entries = cases+controls

- How many observations do you expect to have two copies of a rare allele? Example: MAF for a = 0.001 → expected aa frequency is 0.001 x 0.001 or 1 out of 1 million

- **In a chi-squared test of independence setting**:
  When MAF <<< 0.05 then some cells above will be
  sparse and large-sample statistics (classic chi-squared tests of independence) will no longer be valid. This is the case when there are less than 5 observations in a cell

- **In a regression framework:**
  The minimum number of observations per independent variable should be 10, using a guideline provided by Hosmer and Lemeshow, authors of Applied Logistic Regression, one of the main resources for Logistic Regression

Remediation: rationale for burden tests

- Alpha level of 0.05, corrected by number of bp in the genome= $1.6*10^{-11}$
- One needs VERY LARGE samples sizes in order to be able to reach that level, even if you find “the variant”.
- So what to do in this situation?
- Do not test a single variant at a time, but pool variants: specification of a so-called “**region of interest**” (ROI)
- A region can be anything really:
  - Gene
  - Locus
  - Intra-genic area
  - Functional set

Key features of burden tests

- Collapse many variants into single risk score
- Several flavors exist:
  - In general they all combine rare variants into a genetic score
    Example: Combine minor allele counts into a single risk score (dominant genetic model)
  - Weighted or unweighted versions (f.i., to prioritize certain variant types, based on predictions about damaging effect)
- When high linkage disequilibrium (LD) [allelic non-independence] exists in the "region", combined counts may be artificially elevated
- Assume all rare variants in a set are causal and associated with a trait in the same direction
  - Counter-examples exist for different directionality (e.g. autoimmune GWAs)
  - Violations of this assumption leads to power loss

REVIEW

Rare-Variant Association Analysis: Study Designs and Statistical Tests

Seunggeung Lee,[1] Gonçalo R. Abecasis,[1] Michael Boehnke,[1] and Xihong Lin[2,*]

Despite the extensive discovery of trait- and disease-associated common variants, much of the genetic contribution to complex traits remains unexplained. Rare variants can explain additional disease risk or trait variability. An increasing number of studies are underway to identify trait- and disease-associated rare variants. In this review, we provide an overview of statistical issues in rare-variant association studies with a focus on study designs and statistical tests. We present the design and analysis pipeline of rare-variant studies and review cost-effective sequencing designs and genotyping platforms. We compare various gene- or region-based association tests, including burden tests, variance-component tests, and combined omnibus tests, in terms of their assumptions and performance. Also discussed are the related topics of meta-analysis, population-stratification adjustment, genotype imputation, follow-up studies, and heritability due to rare variants. We provide guidelines for analysis and discuss some of the challenges inherent in these studies and future research directions.

(Lee et al. 2014)

Other tests

- Variance-component tests (e.g., SKAT)
  - Variance tests outperform burden tests if many variants are non-causal
- Combined tests
  - These test the variance of genetic effects
  - Burden tests outperform variance tests if many variants are causal
  - Therefore, a test that combines both in different scenarios is useful.
  - SKAT-O is such a test: $Q = (1-p)Q_{SKAT} + pQ_{BURDEN}$
    - Can include covariates
    - Optimal p? Try several … (multiple testing)
- EC tests
  - These tests exponentially combine single variant score tests

(Lee et al. 2014)

PLOS | GENETICS

RESEARCH ARTICLE

The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease

Loukas Moutsianas[1☯*], Vineeta Agarwala[2,3☯], Christian Fuchsberger[4], Jason Flannick[3,5], Manuel A. Rivas[1], Kyle J. Gaulton[1], Patrick K. Albers[1], GoT2D Consortium[¶], Gil McVean[1], Michael Boehnke[4], David Altshuler[3,5,6,7], Mark I. McCarthy[1,8*]

1 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **2** Program in Biophysics, Harvard University, Cambridge, Massachusetts, United States of America, **3** Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America, **4** Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, United States of America, **5** Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, United States of America, **6** Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America, **7** Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **8** Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, United Kingdom

☯ These authors contributed equally to this work.
¶ A full list of GoT2D Consortium members and affiliations appears in S1 Text.
* moutsian@well.ox.ac.uk (LM); mark.mccarthy@drl.ox.ac.uk (MIM)

Citation: Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, et al. (2015) The

(Moutsianas et al. 2015)

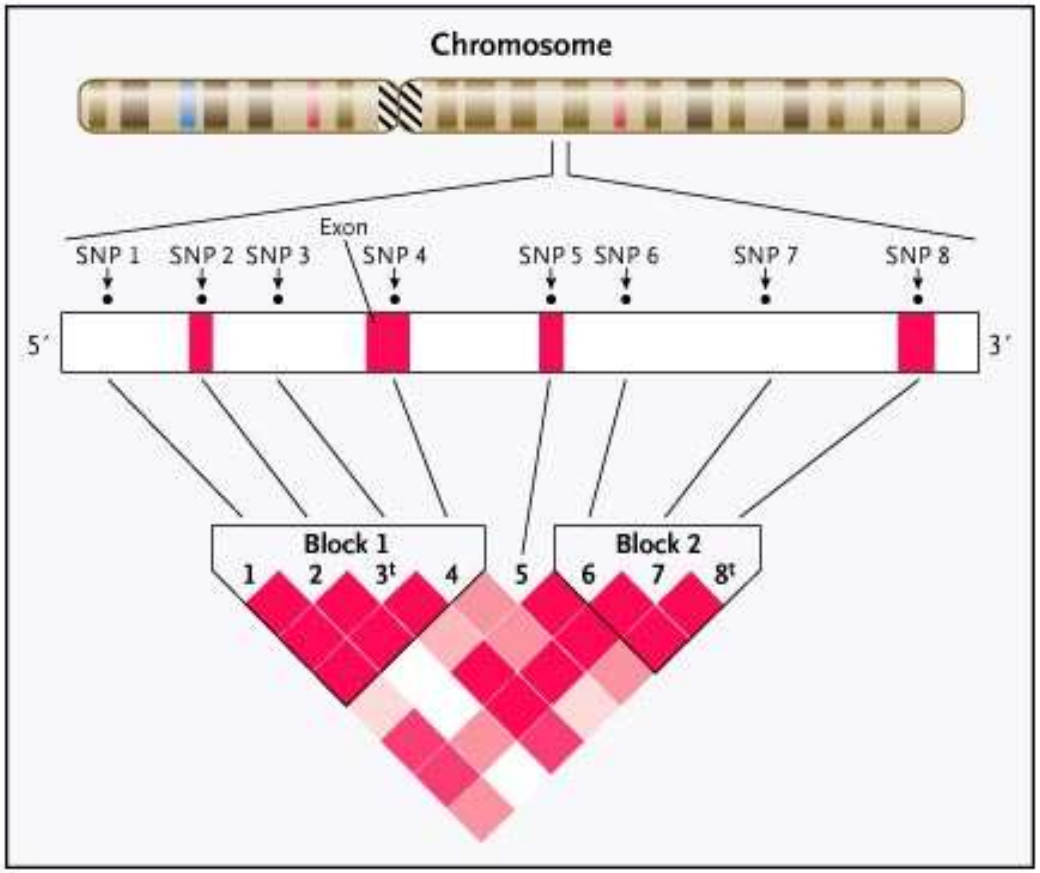
3 When effects become non-independent

Linkage disequilibrium (LD) between genetic markers

- **Linkage Disequilibrium** (LD) is a measure of co-segregation of alleles in a population – allelic association

> *Two alleles at different loci that occur together*
> *on the same chromosome (or gamete)*
> *more often*
> *than would be predicted by random chance.*

Mapping the "relationships" between SNPs (Christensen and Murray 2007)



(HaploView software)

Independence between SNPs

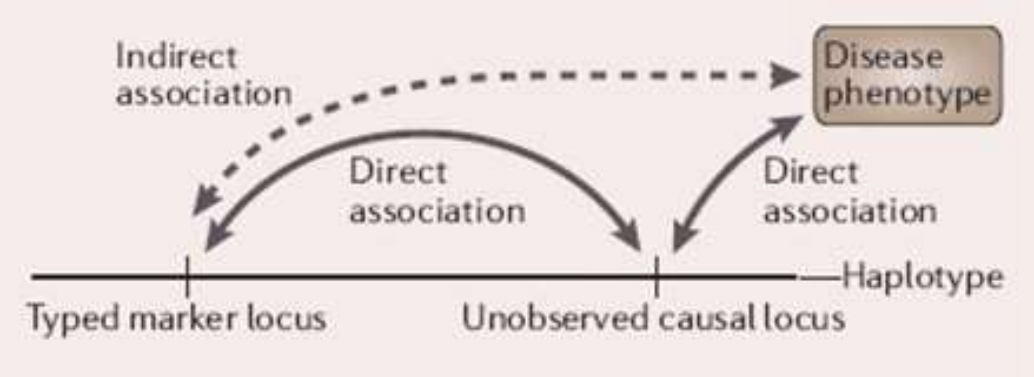
- The measure D is defined as the difference between the observed and expected (under the null hypothesis of independence) proportion of "haplotypes" bearing specific alleles at two loci: $p_{AB} - p_A p_B$

	A	a
B	p_{AB}	p_{aB}
b	p_{Ab}	p_{ab}

- Notice the link with a 2x2 table independence test … ("observed minus expected")
- Instead of testing all SNPs, use LD-block information to test "independent" SNPs or loci ... Then use the "dependency" structure again when interpreting results

Impact and interpretation

- LD is a very important concept for GWAs, since it gives the rational for performing genetic association studies



- Other measures of association than D exist: Because of its interpretation, the **measure r^2 (coefficient of determination**) is most often used for GWAs

Biological vs statistical epistasis

Definition of epistasis

- Our ability to detect epistasis depends on what we mean by epistasis

“compositional epistasis”

- The original definition (**driven by biology**) refers to distortions of Mendelian segregation ratios due to one gene masking the effects of another; a variant or allele at one locus prevents the variant at another locus from manifesting its effect (William Bateson 1861-1926).

- Example of phenotypes (e.g. hair colour) from different genotypes at 2 loci interacting epistatically under Bateson's (1909) definition (**compositional epistasis**):

Genotype at locus B/G	**gg**	**gG**	**GG**
bb	White	Grey	Grey
bB	Black	Grey	Grey
BB	Black	Grey	Grey

The effect at locus B is masked by that of locus G: locus G is epistatic to locus B.

(Cordell 2002)

Definition of epistasis

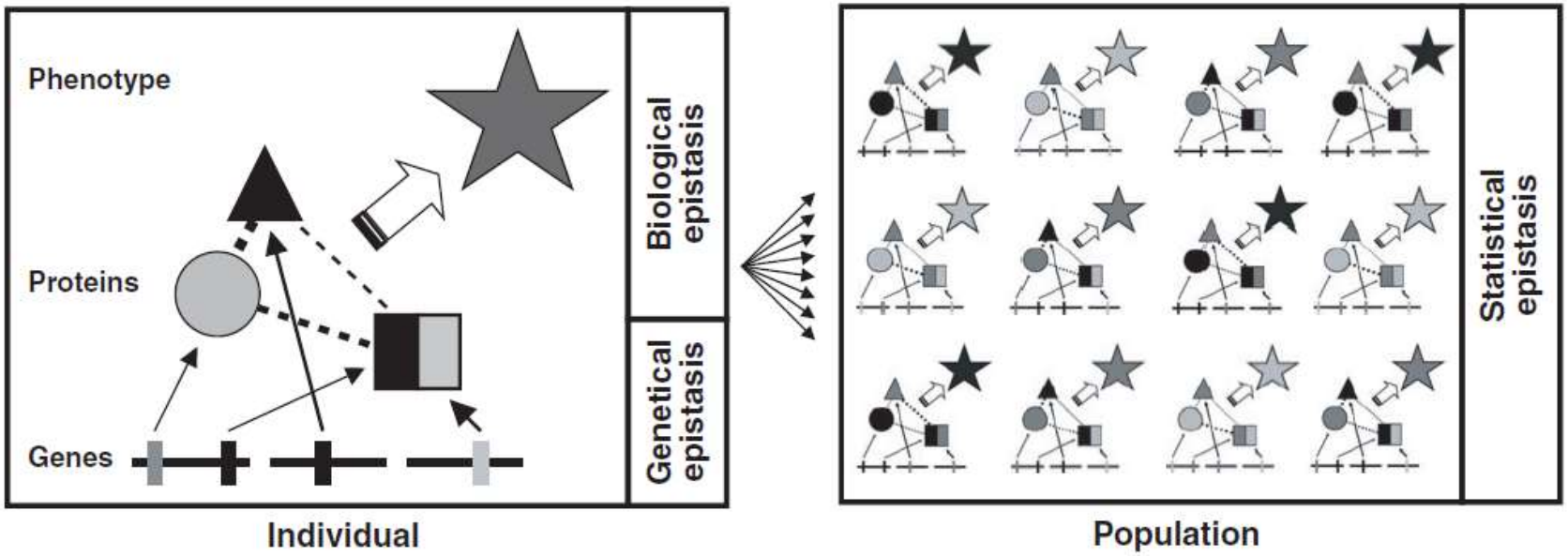
- Our ability to detect epistasis depends on what we mean by epistasis

“statistical epistasis”

- A later definition of epistasis (**driven by statistics**) is expressed in terms of deviations from a model of additive multiple effects.
- This might be on either a linear or logarithmic scale, which implies different definitions (Ronald Fisher 1890-1962).

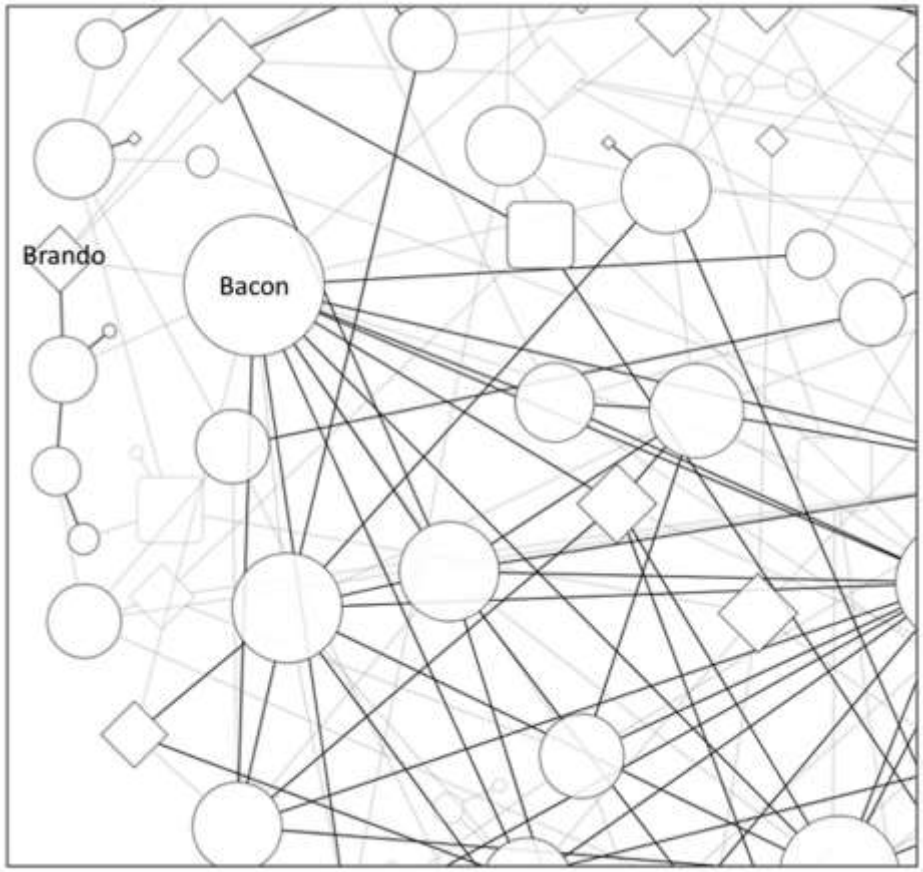
Genetic interactions:

… when two or more DNA variations interact either directly to change transcription or translation levels, or indirectly by way of their protein product, to alter disease risk separate from their independent effects …



(Moore 2005)

Why? Complementing insights from GWA studies



Edges **represent small gene–gene interactions between SNPs.** Gray nodes and edges **have weaker interactions.** Circle nodes **represent SNPs that do not have a significant main effect.** The diamond nodes **represent significant main effect association.** The size of the node **is proportional to the number of connections.**

(McKinney et al 2012)

How? Random Forests?

Winham *et al. BMC Bioinformatics* 2012, **13**:164
http://www.biomedcentral.com/1471-2105/13/164

BMC Bioinformatics

RESEARCH ARTICLE Open Access

SNP interaction detection with Random Forests in high-dimensional genetic data

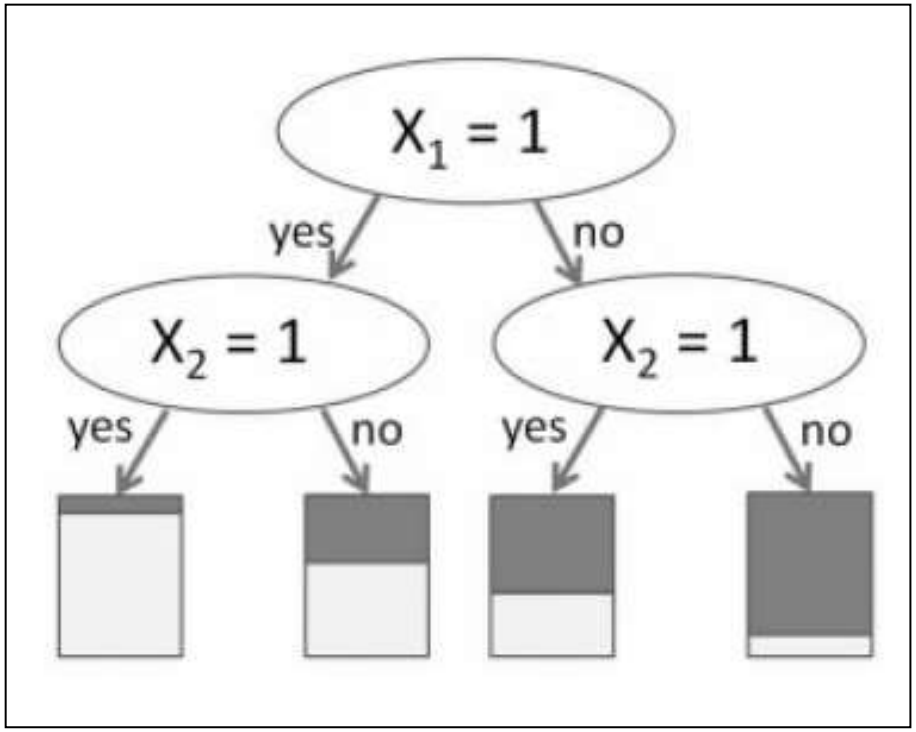
Stacey J Winham[1*], Colin L Colby[1], Robert R Freimuth[1], Xin Wang[1], Mariza de Andrade[1], Marianne Huebner[1,2] and Joanna M Biernacka[1,3*]

Abstract

Background: Identifying variants associated with complex human traits in high-dimensional data is a central goal of genome-wide association studies. However, complicated etiologies such as gene-gene interactions are ignored by the univariate analysis usually applied in these studies. Random Forests (RF) are a popular data-mining technique that can accommodate a large number of predictor variables and allow for complex models with interactions. RF

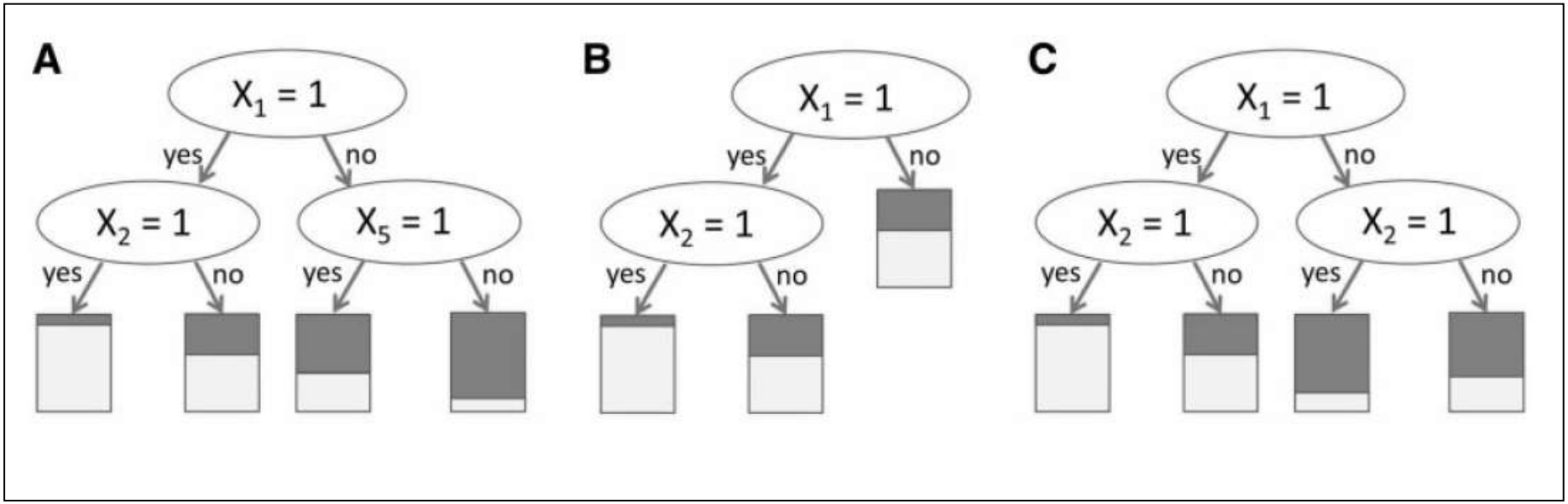
Be critical …

The split-based structure of classification and regression trees can advantageously take interaction effects into account. Let us consider the first two layers in a tree and how this tree might look when there are only two relevant binary predictor variables X_1 and X_2, with additional irrelevant predictor variables $X_3, \ldots, X_p$. If the root node is split by predictor variable X_1, the effect of X_2 may be different in the two child nodes, hence taking the potential interaction between X_1 and X_2 into account. If X_1 and X_2 have main effects only, one ideally expects X_2 to be selected in both child nodes with the same effect on the response, yielding the idealized picture

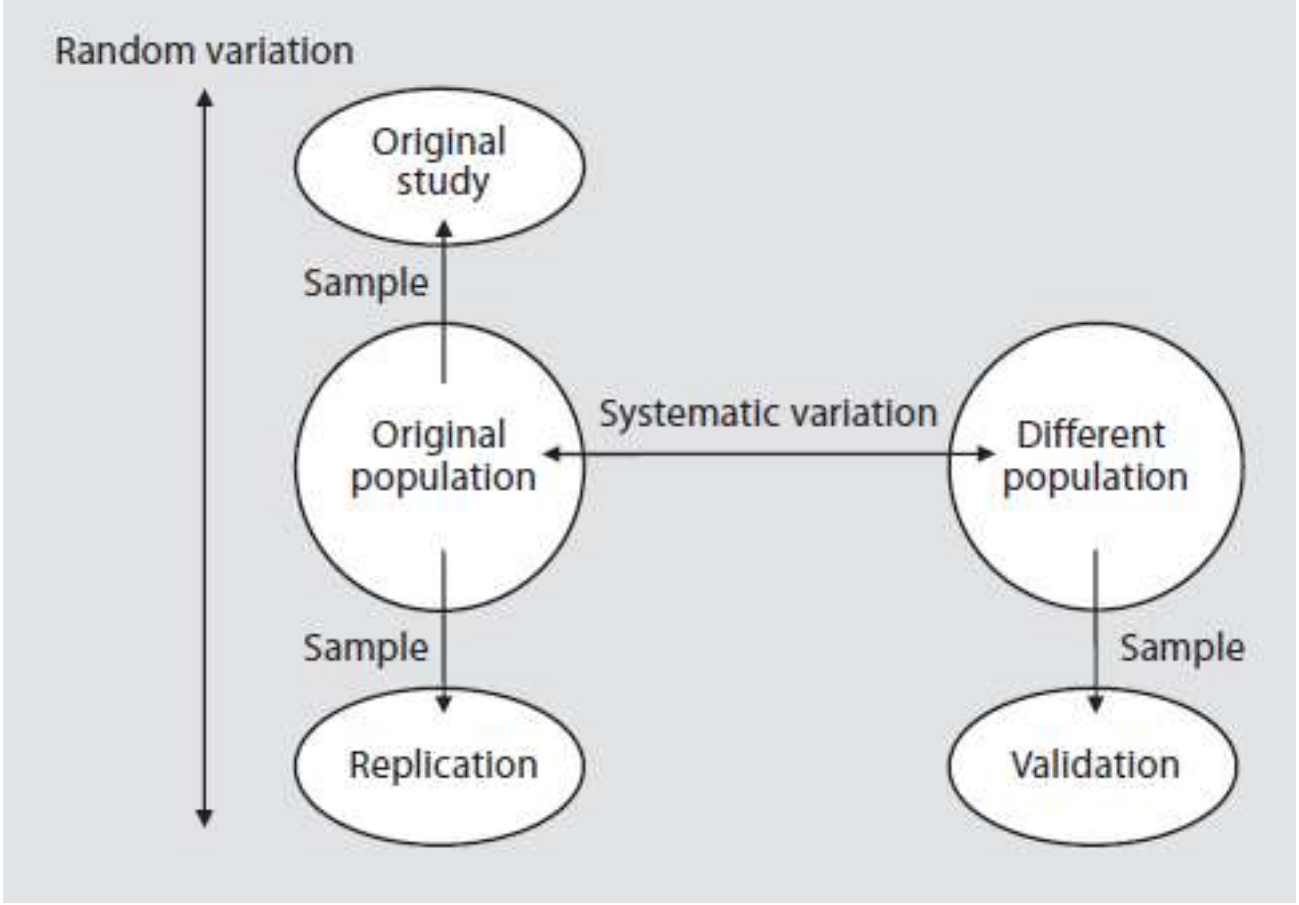


Be critical …

Everything else—selection of different predictor variables in the two child nodes, stopping on one side but not on the other, same predictor variable and same cutpoint on both sides but with different effects—indicates a potential interaction (Figures 2A, B and C as examples of these three situations) [23]. The problem is that, due to random variations in finite samples, it is extremely rare that the tree selects the same predictor variable with the same effect on both sides, except perhaps in the case of very large samples.



Replication / validation



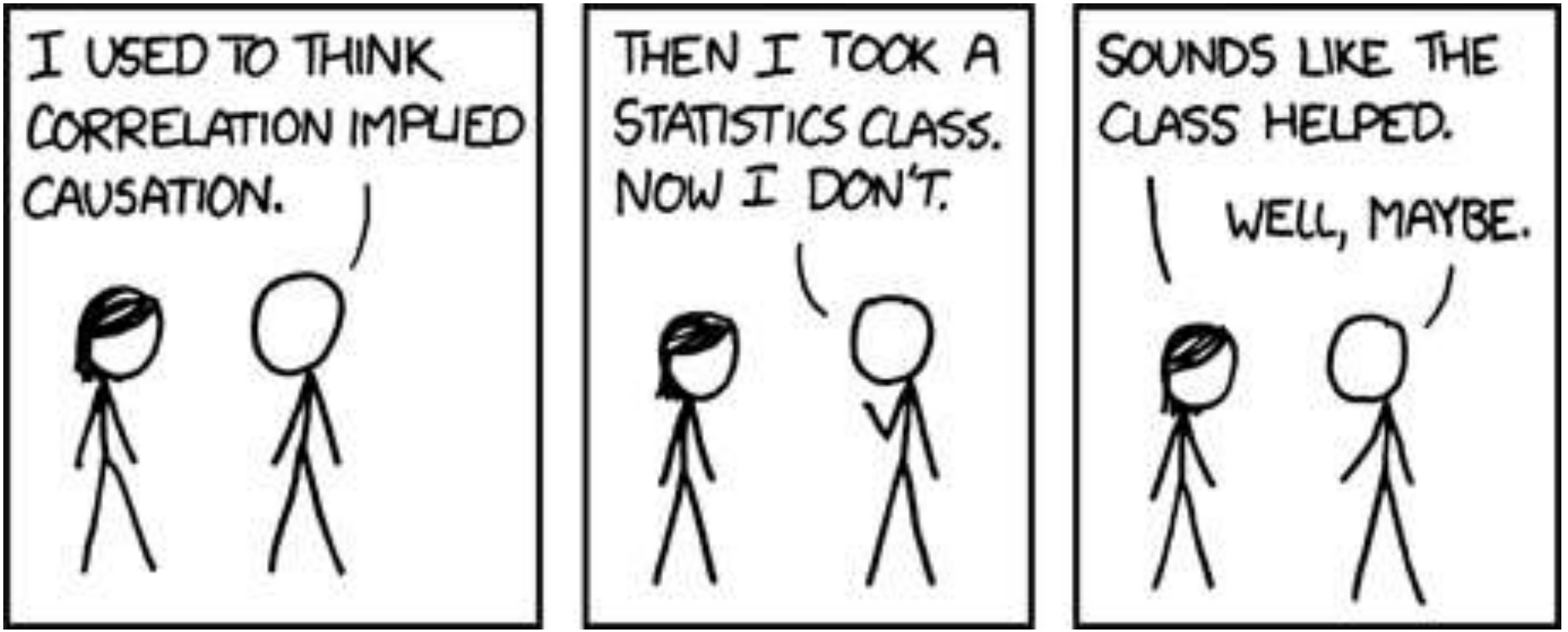
(Igl et al. 2009)

Guidelines for replication studies

- Replication studies should be of sufficient size to demonstrate the effect
- Replication studies should conducted in independent datasets
- Replication should involve the same phenotype
- Replication should be conducted in a similar population
- The same SNP should be tested
- The replicated signal should be in the same direction
- Joint analysis should lead to a lower *p*-value than the original report
- Well-designed negative studies are valuable

"Wishful thinking" for rare variant association or

large-scale interaction association studies?


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.