

Genetics and Bioinformatics

Kristel Van Steen, PhD²

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

Lecture 6: 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 and Remediation

DNA sequence analyses

3 When effects become non-independent

Impact and interpretation

Biological vs statistical epistasis

1 Capita selecta in GWAs



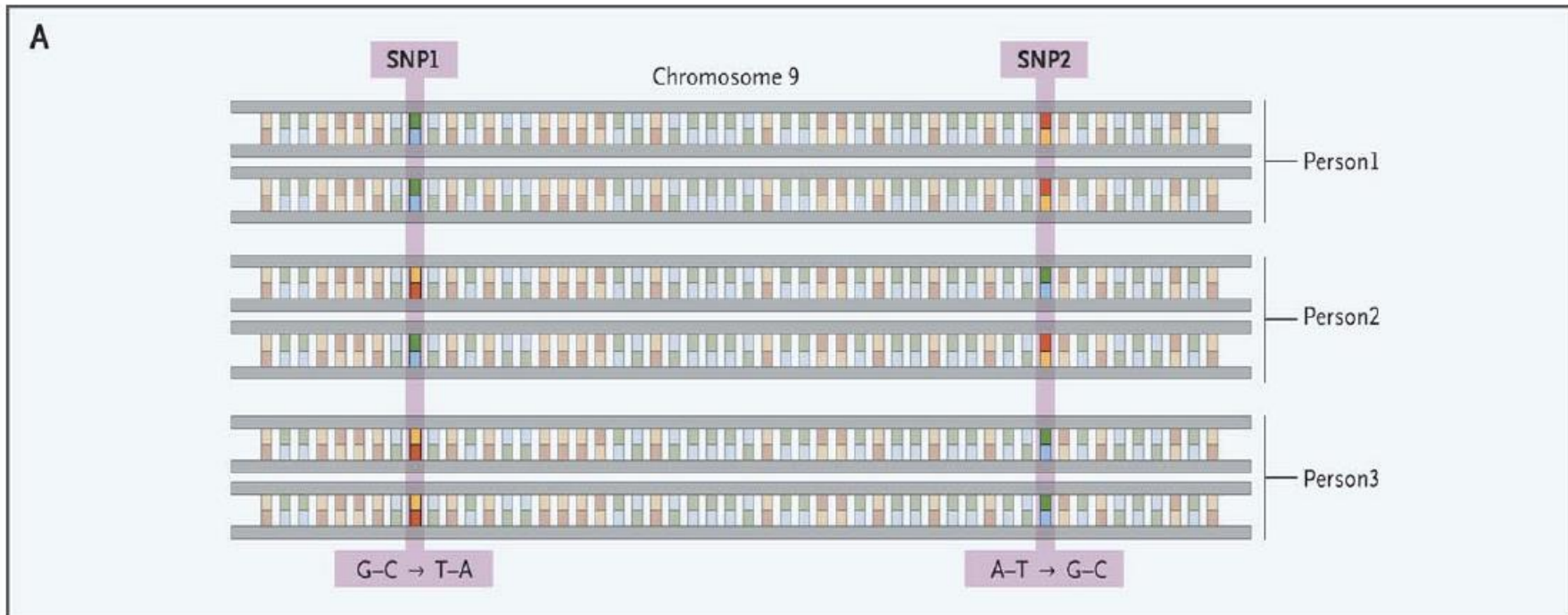
(slide Doug Brutlag 2010)

Definition (recap)

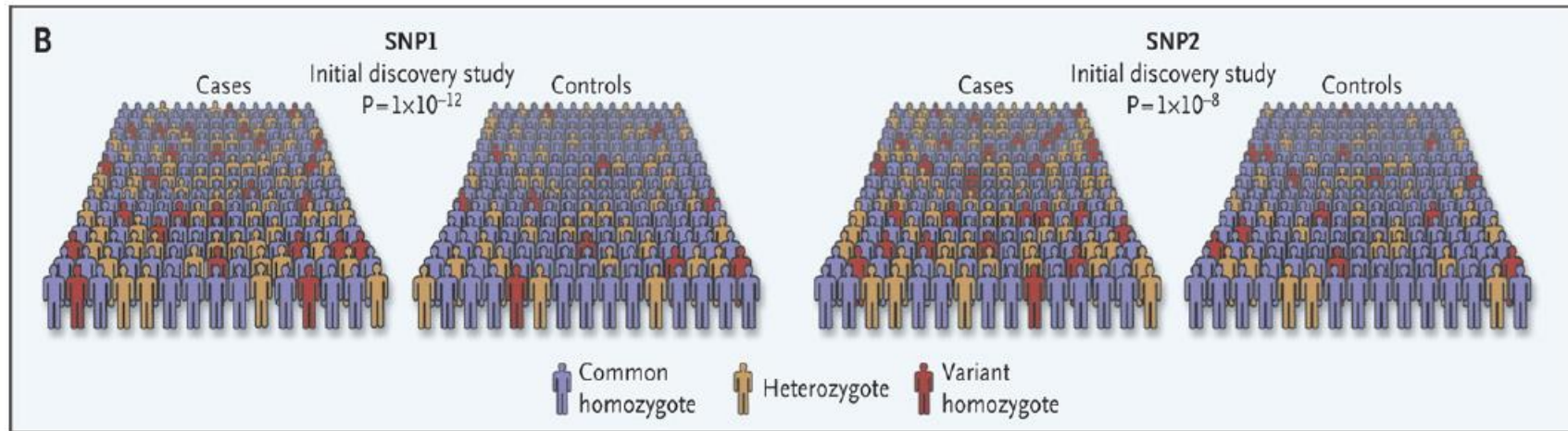
- A **genome-wide association study** is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular trait.
- A **trait** can be defined as a coded phenotype, a particular characteristic such as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



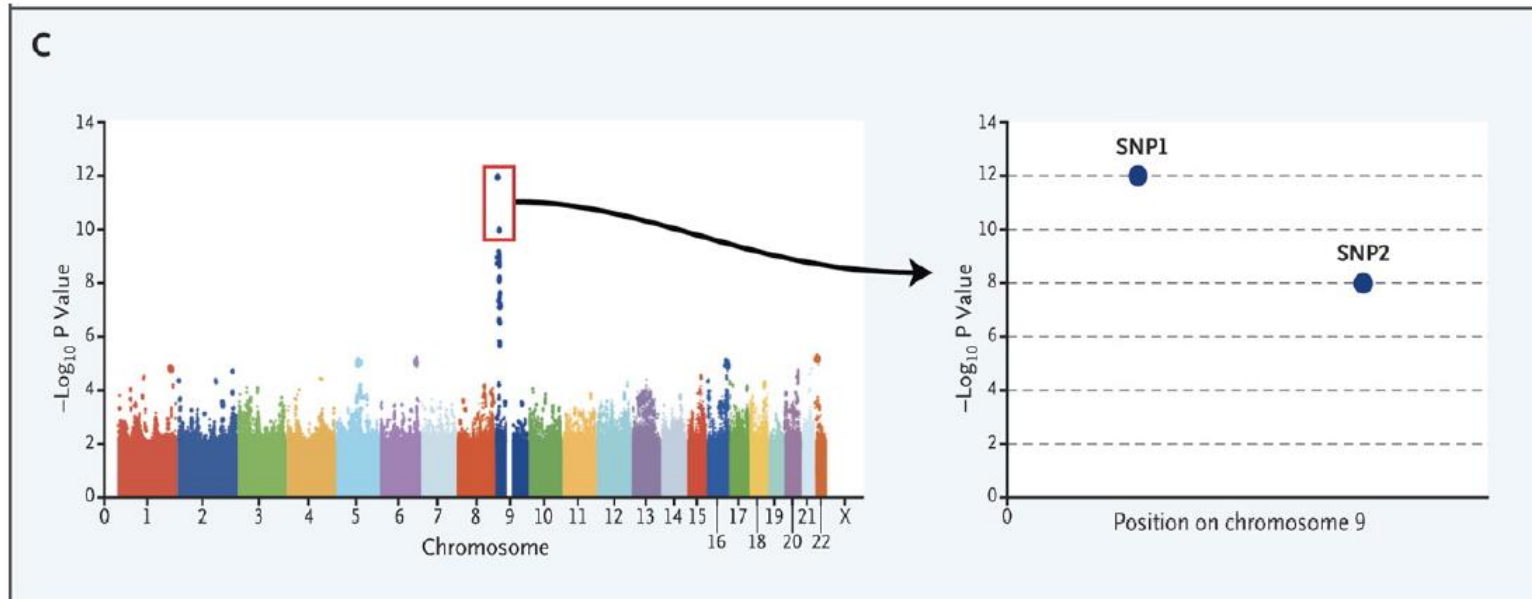
Genome-wide association studies in practice



- 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)

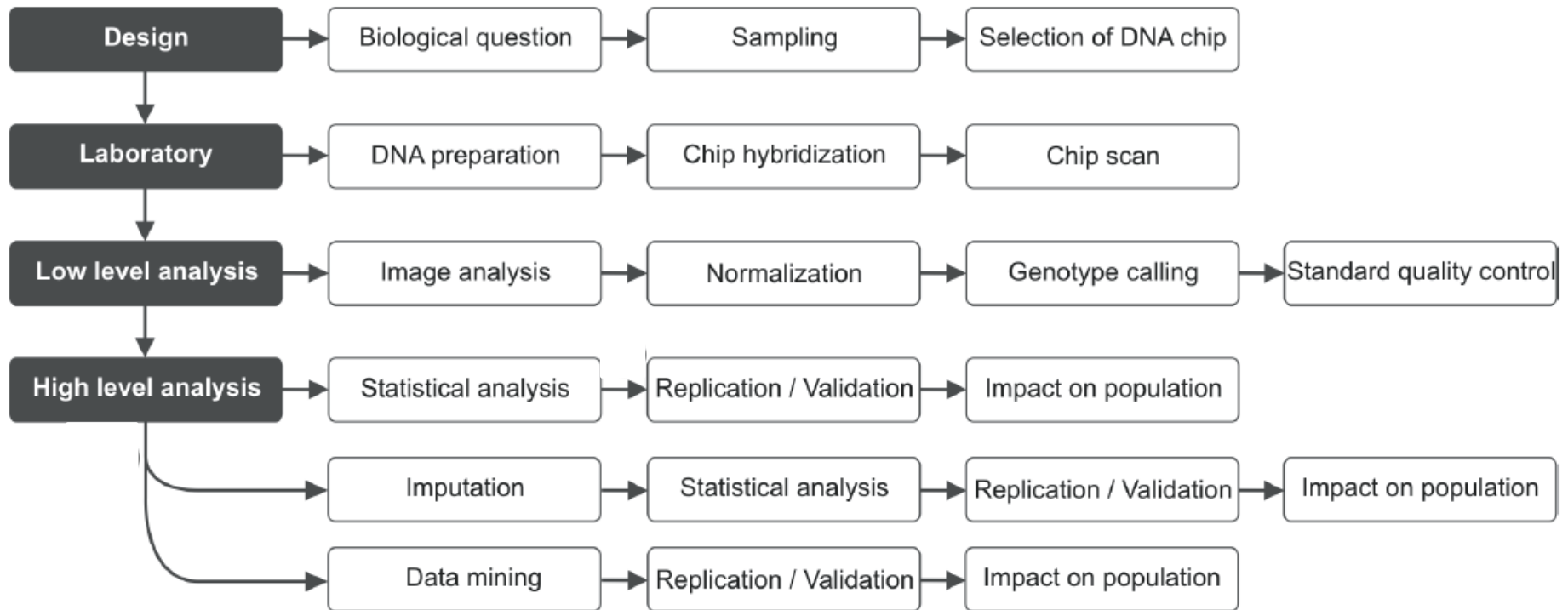
Genome-wide association studies in practice



- 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

tfam file: First 6 columns of standard ped file

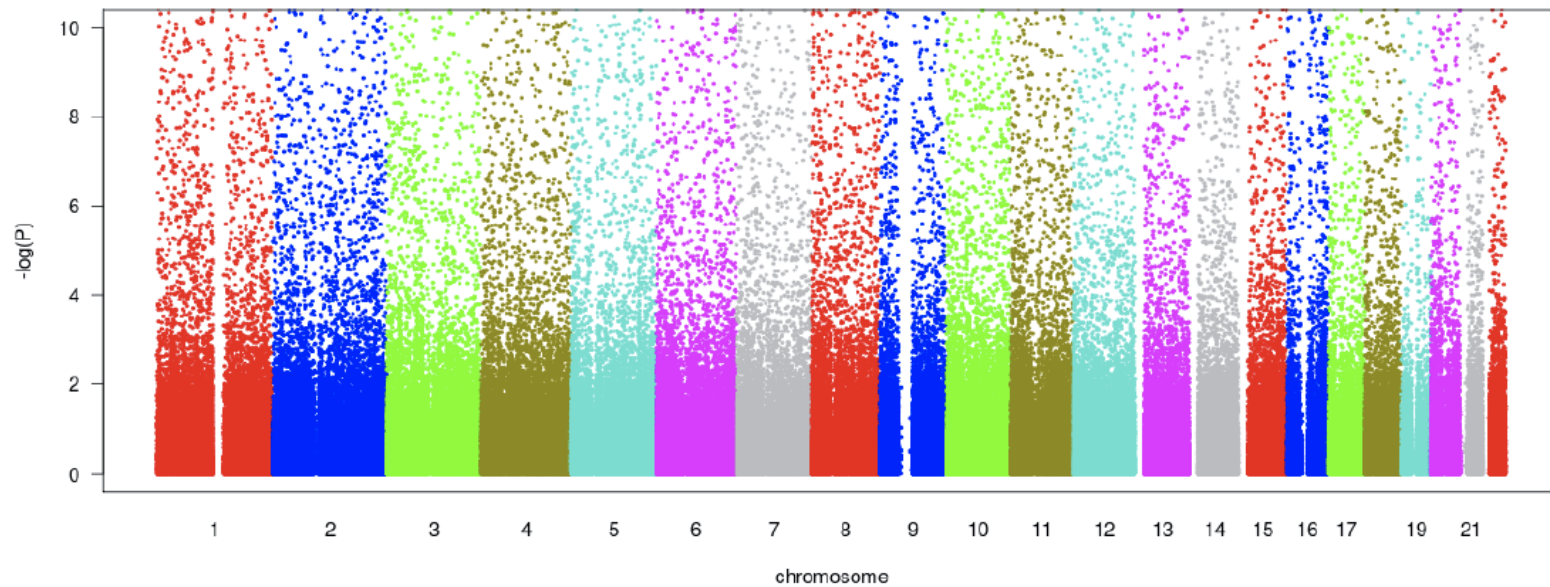
tped file

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

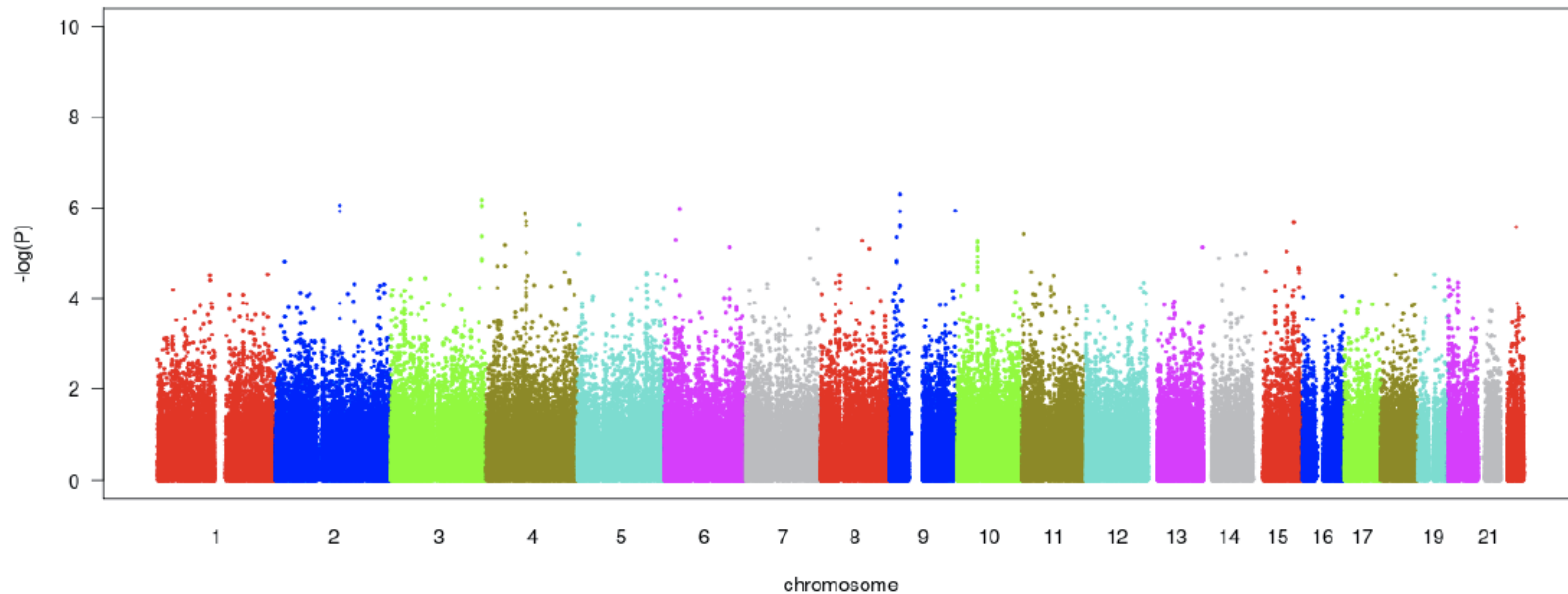


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

(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	$\geq 97\%$
	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)

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:

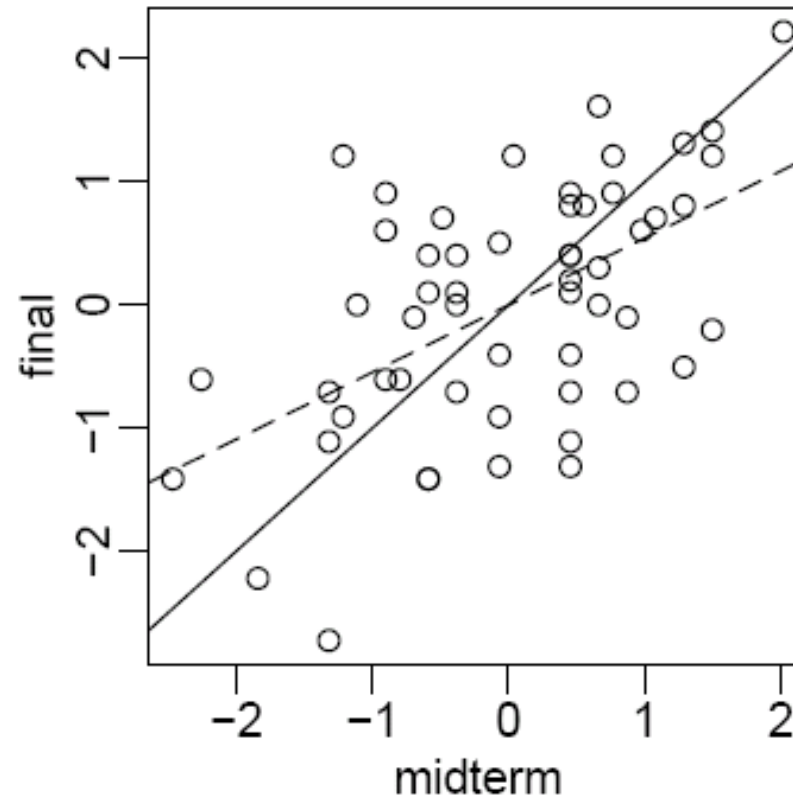
$$\frac{y - \bar{y}}{SD_y} = r \frac{(x - \bar{x})}{SD_x}.$$

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 (0,0)



(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, \dots, X_p .
- When $p=1$ it is called simple regression but when $p > 1$ it is called multiple regression (not recommended as term: multivariate regression).
- When there is more than one Y , then it is called multivariate (simple or 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

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

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

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

$$y = \frac{\beta_0}{1 + e^{\beta_1 x_1}} + \epsilon$$

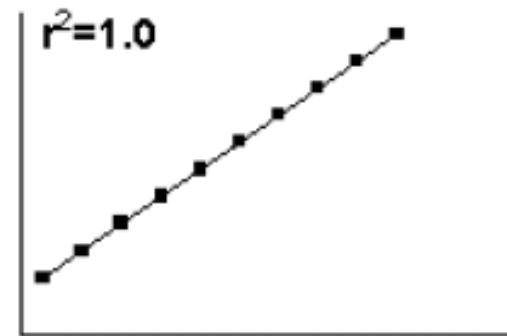
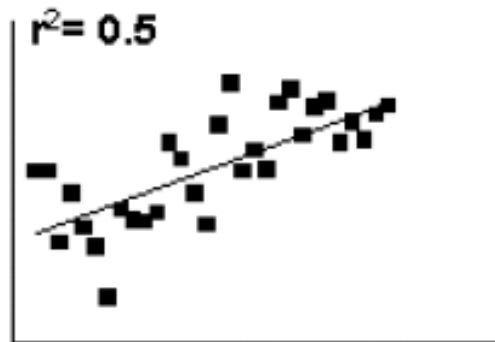
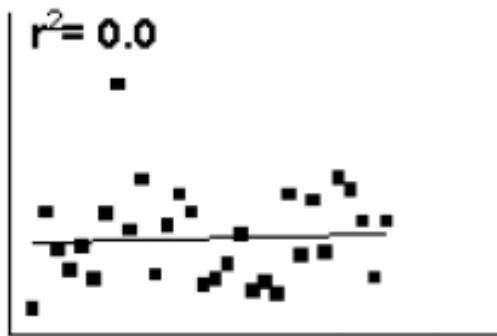
Regression inference

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

- 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. Most generally: $1 - \text{FUV}$ (Fraction of Unexplained Variation)
- 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 H_0 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

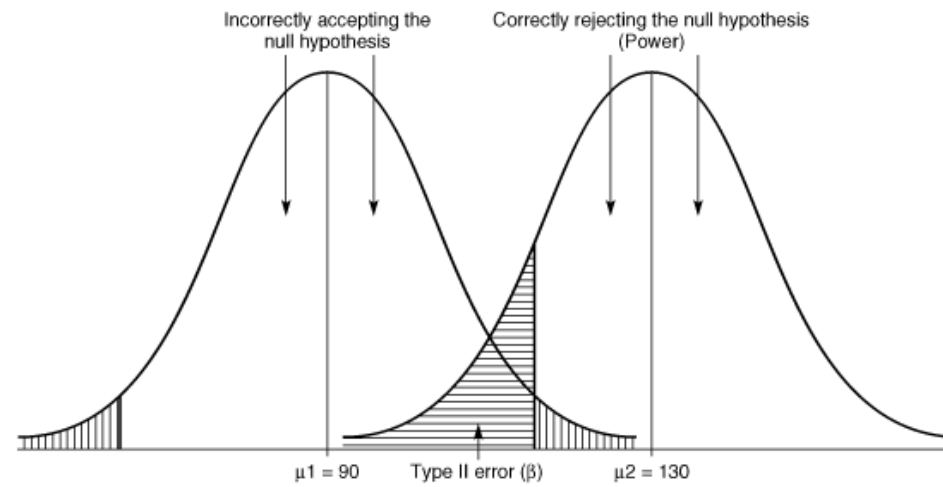
- The logic now is to compare both SSEs. The actual test statistic is a function of $SSE(R)$ - $SSE(F)$:

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

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:**

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

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

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

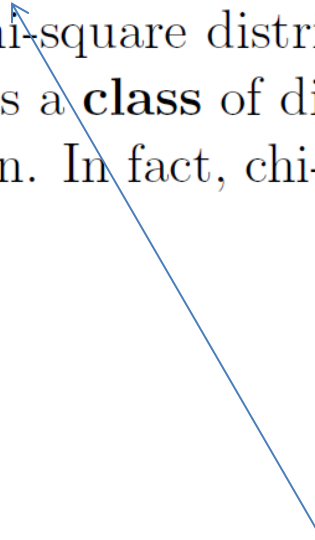
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 PC_1 + \beta_3 PC_2 + \varepsilon$$

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

Distributional relationships: F, t, chi-squared

$Z_1, Z_2, \dots, Z_\kappa$ 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.

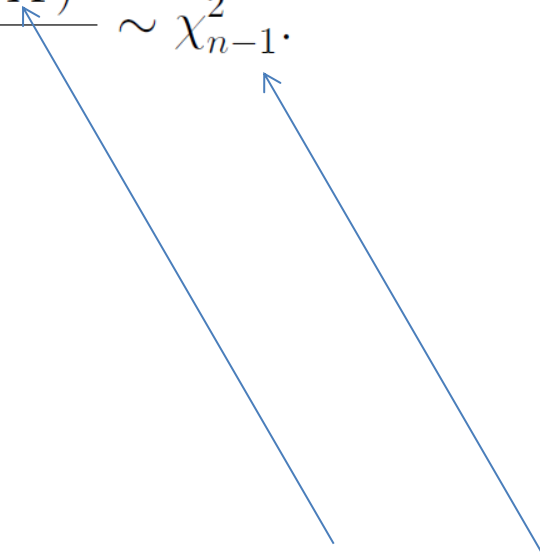


Distributional relationships: F, t, chi-squared

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

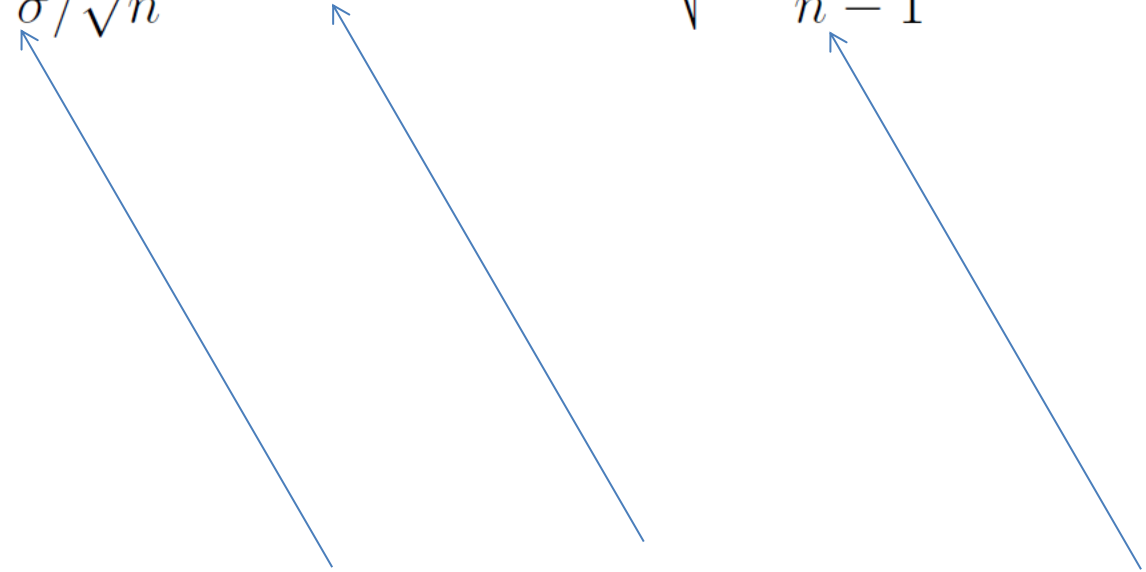
$$\frac{\sum_1^n (X_j - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2.$$


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 σ 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

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

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

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

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

Implication to example 1:

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

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$
- $df_F = n - 2$ (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 $\beta_1 \neq 0$

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

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

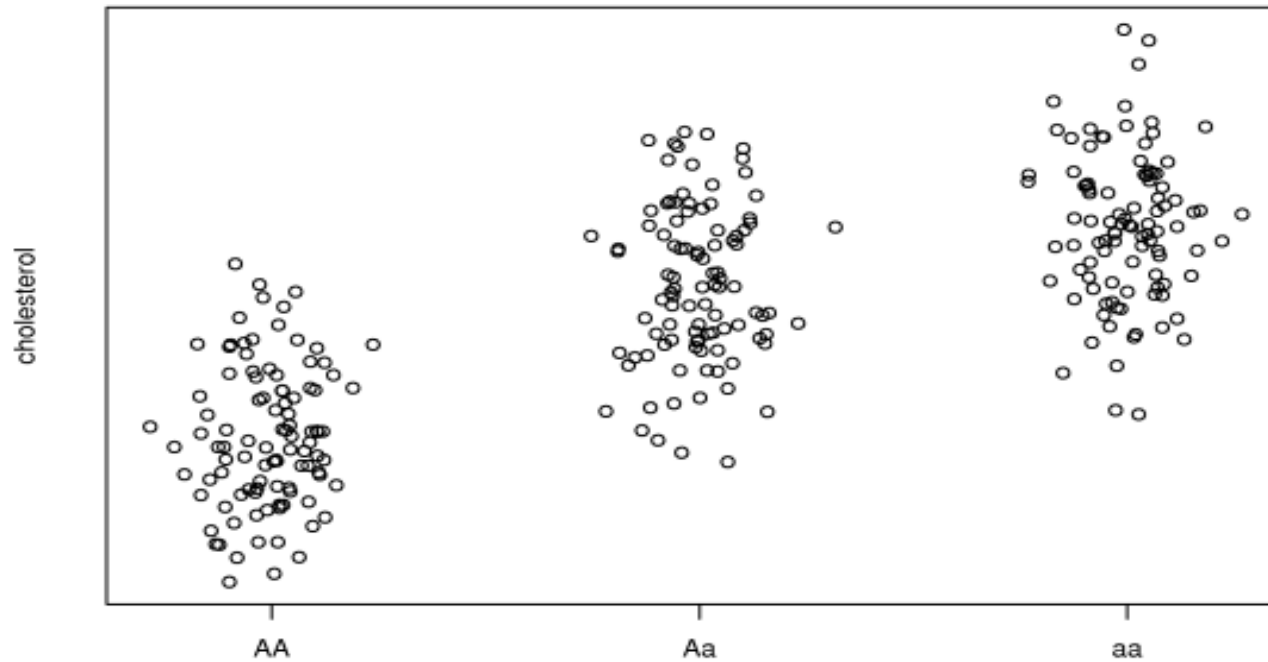
Model Description	predictor	Common name
Number of minor alleles	$(g=='Aa') + 2*(g=='aa')$ or <code>as.numeric(g)</code>	Additive
Presence of minor allele	$(g=='Aa') (g=='aa')$	Dominant
Homozygous for minor allele	$g=='aa'$	Recessive
Distinct effects for hetero/homozygous	<code>factor(g)</code>	2 parameter,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

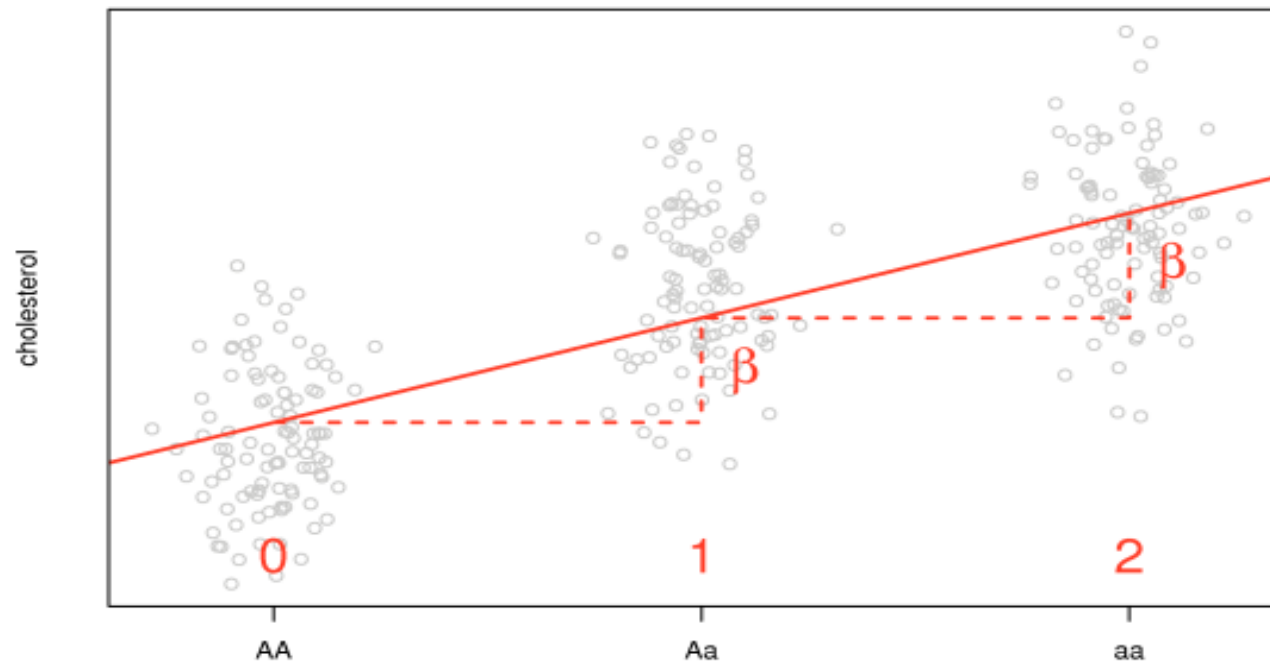
Use of `lm()` in genetics

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



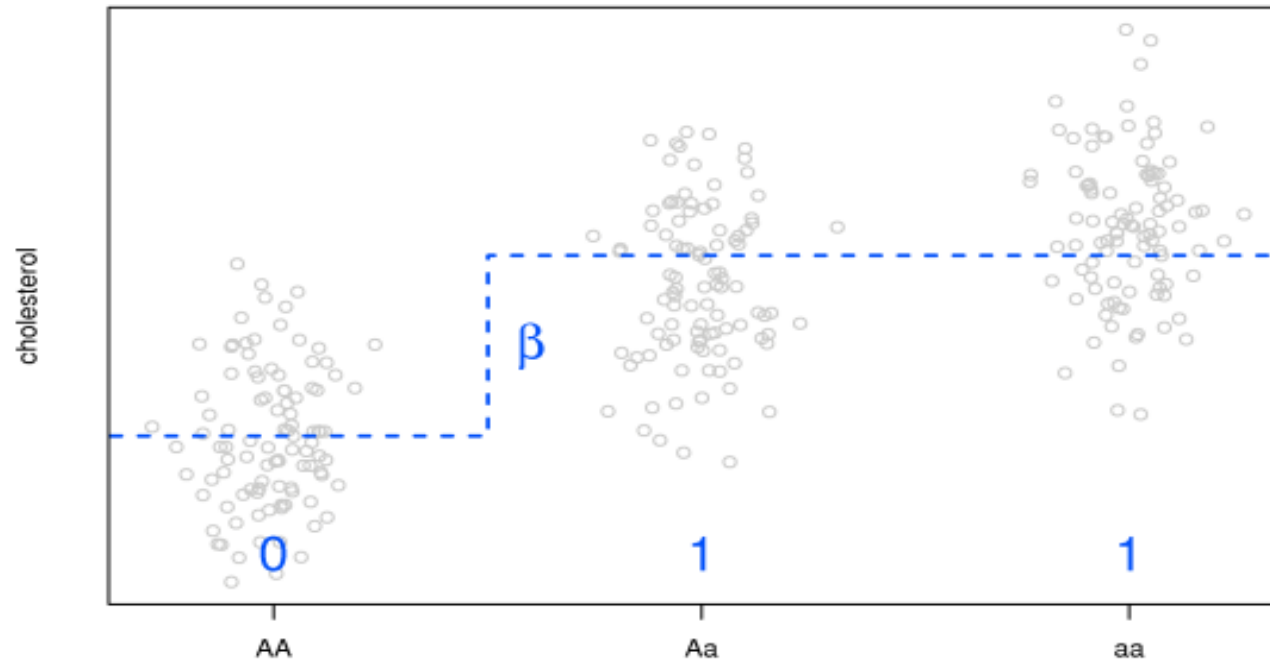
Use of `lm()` in genetics

Additive model (the most commonly used)



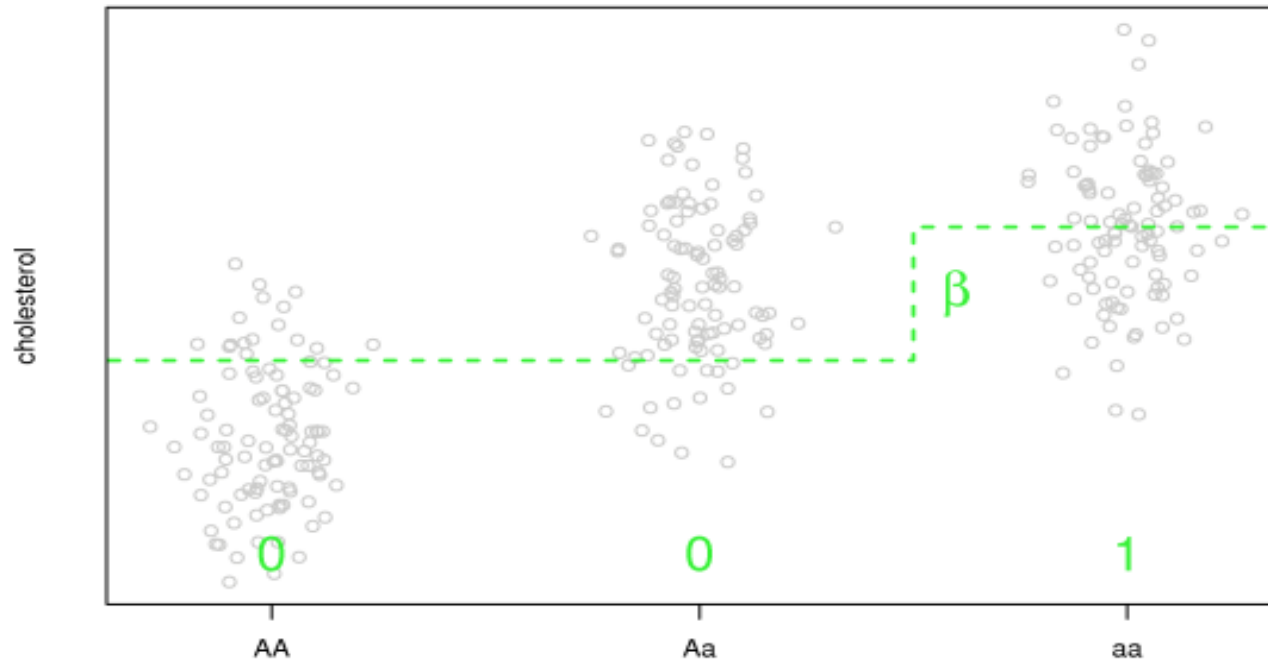
Use of `lm()` in genetics

Dominant model (best fit to this data)



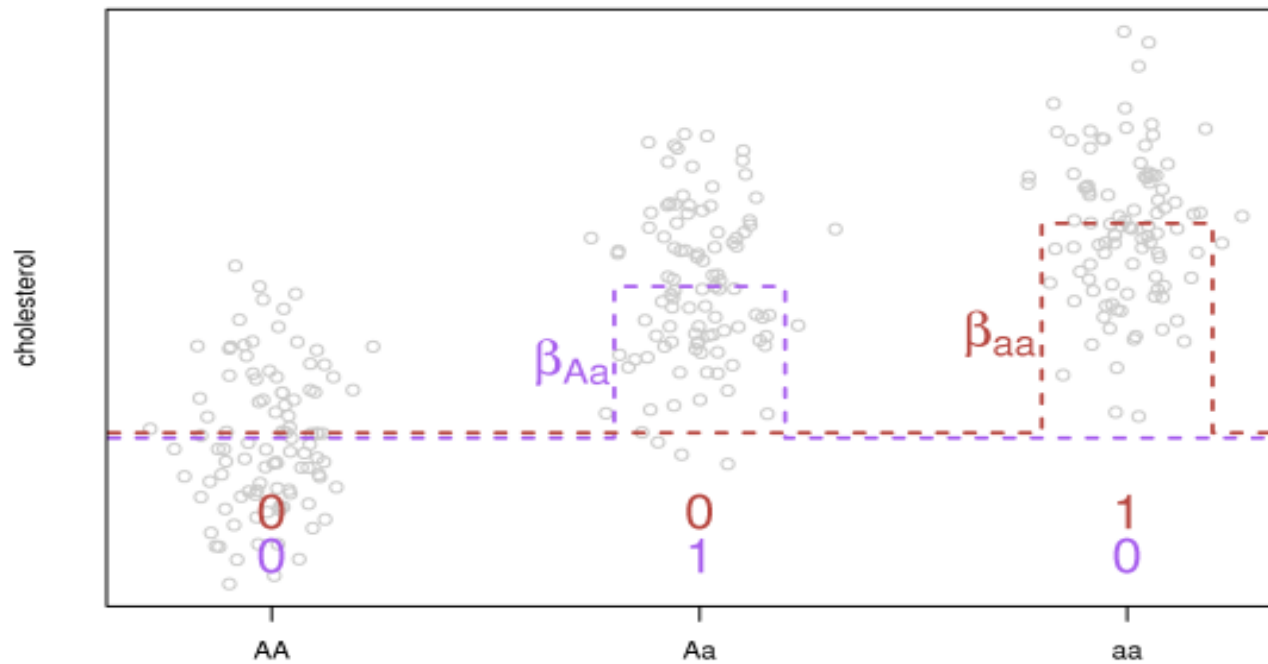
Use of `lm()` in genetics

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.

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}}$$

- 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)

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

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

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



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

$$\text{Log(Odds} | X_1 == 1) = \beta_0 + \beta_1 \cdot 1$$

$$\text{— Log(Odds} | X_1 == 0) = \beta_0$$

$$\text{Log(OR)} = \beta_1$$

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
- $X = (X_1, X_2, \dots, X_k)$ 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:

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

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>)

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

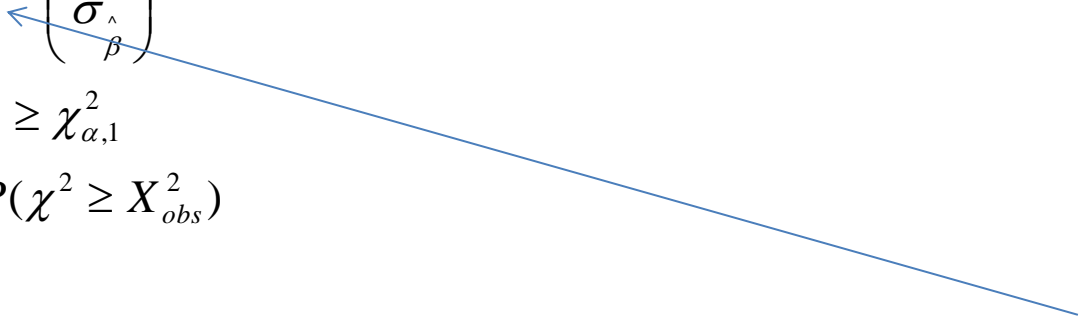
- **Example 1:**

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

- $H_0: \beta_1 = 0$

- $H_1: \beta_1 \neq 0$

Large-sample “Wald test”:

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


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

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

The Wald statistic

In the univariate case, the Wald statistic is

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

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

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

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 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 (χ^2) defined by the following equation.

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

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.

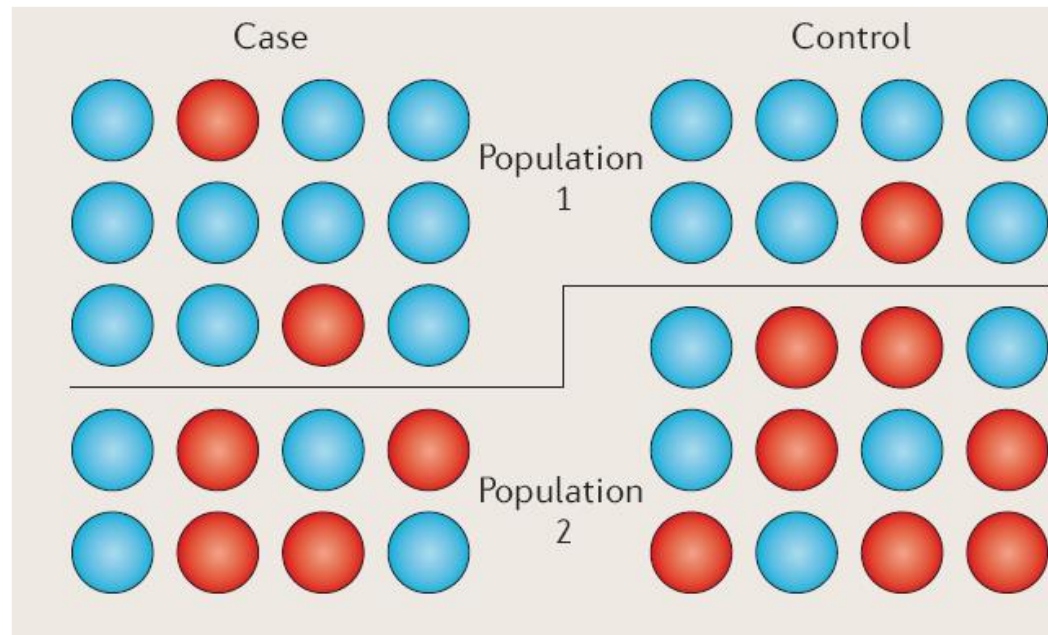
- **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic, which can be proven to follow a chi-square distribution with degrees of freedom as derived before. The null hypothesis is rejected when the P-value is less than the pre-stated significance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

Confounding: population stratification

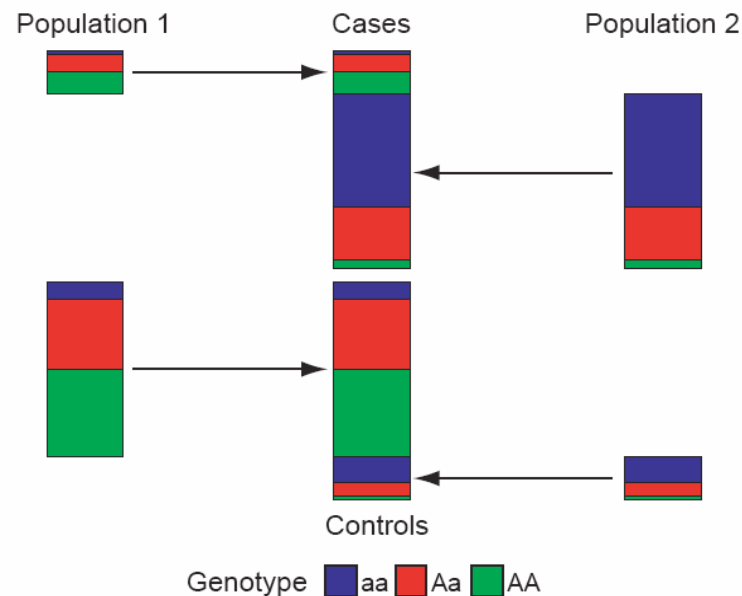
What is spurious association?

- **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?

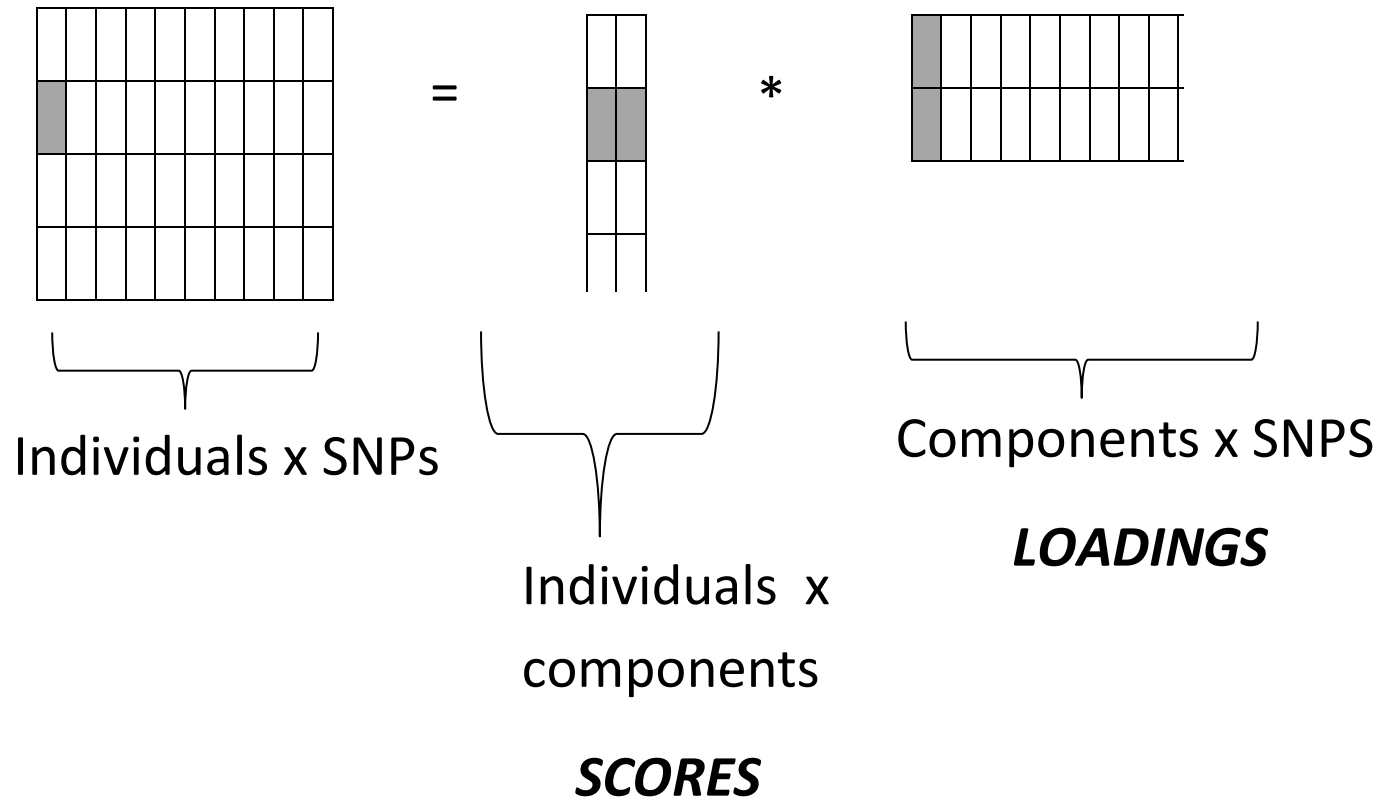
- Typically, there are two characteristics present:
 - A difference in proportion of individual from two (or more) subpopulation in case and controls
 - Subpopulations have different allele frequencies at the locus.



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”

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.

PCA in a nutshell

Notation

- ▶ \mathbf{x} is a vector of p random variables
- ▶ α_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

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

Shortcut to solution

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

Derivation of PCA

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 α_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 α_k by differentiating w.r.t. to α_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 α_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 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 1st 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 from Gaussian variates

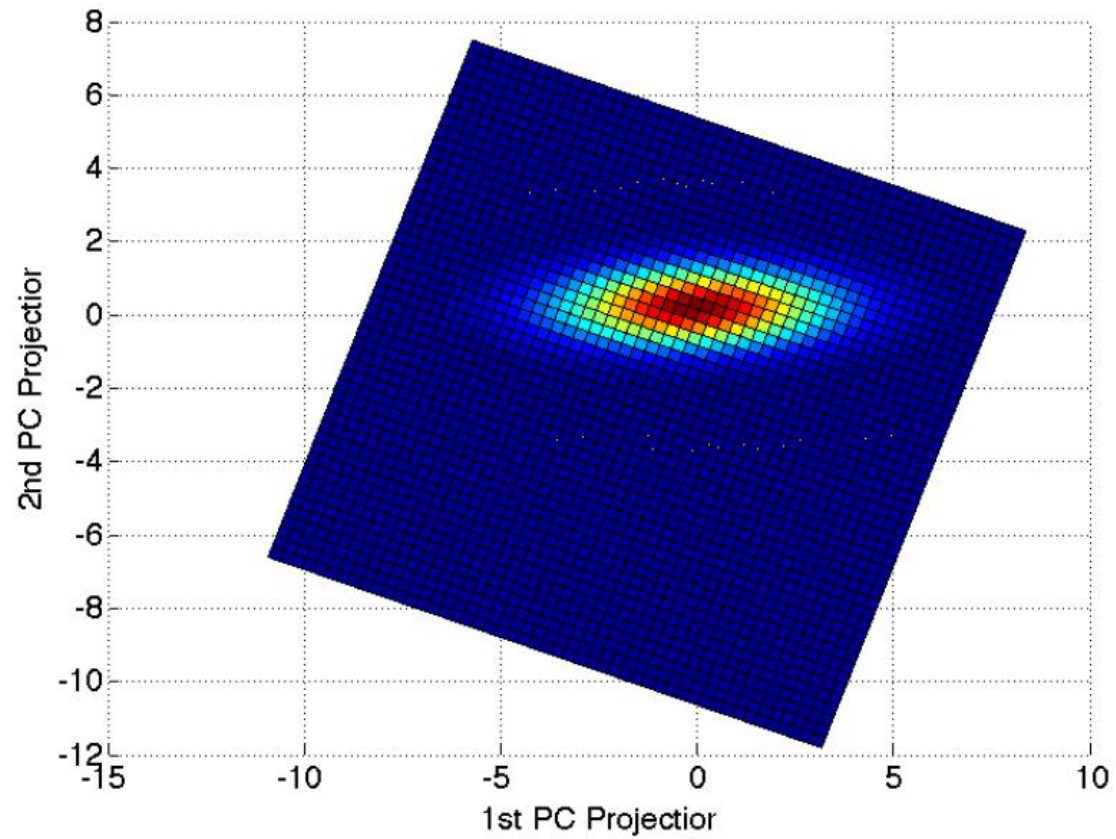
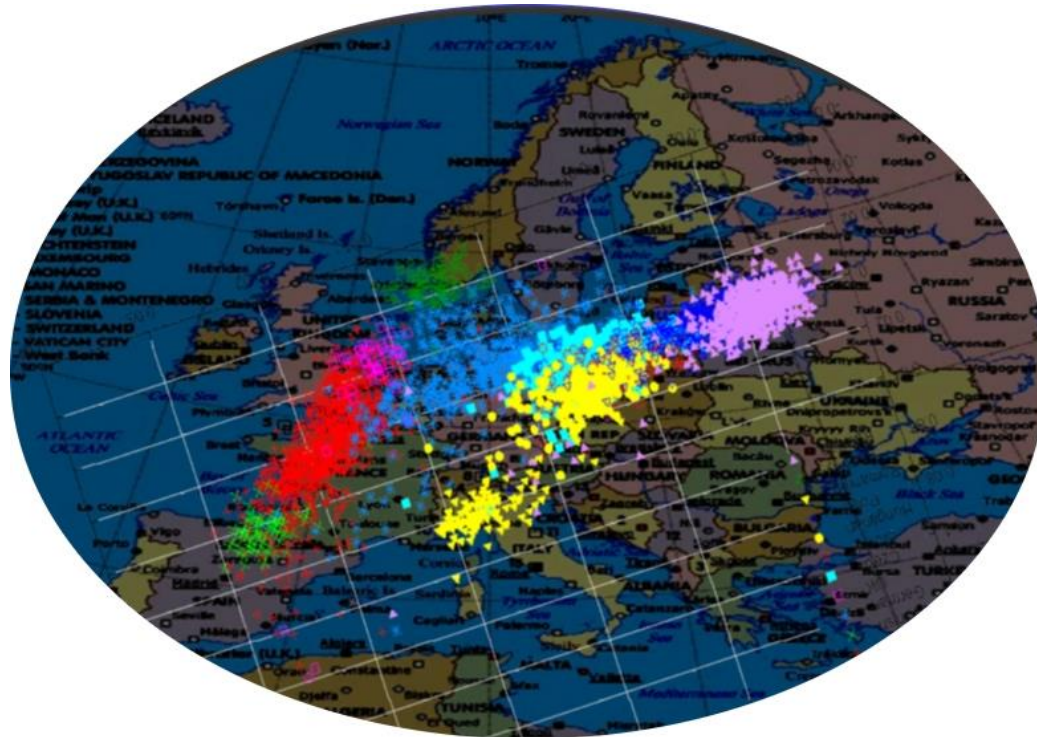


Figure: PCA Projected Gaussian PDF

Principal components in population genetics (variables are SNPs !)

- 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:**

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

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

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 $\lambda > 1$ 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
- $$\hat{\lambda} = \frac{\text{median}(\chi_1^2, \chi_2^2, \dots, \chi_L^2)}{\text{median}(\mathcal{L}(\chi_1^2))} = \frac{\text{median}(\chi_1^2, \chi_2^2, \dots, \chi_L^2)}{0.456}$$
- $$\chi_{GC}^2 = \chi^2 / \hat{\lambda}$$

```
> 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
```

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 $MAF \lll 0.05$ 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

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)
 - These test the variance of genetic effects
- Combined tests
 - Variance tests outperform burden tests if many variants are non-causal
 - 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)


RESEARCH ARTICLE

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


Loukas Moutsianas¹^{*}, Vineeta Agarwala^{2,3}[☉], Christian Fuchsberger⁴, Jason Flannick^{3,5}, Manuel A. Rivas¹, Kyle J. Gaulton¹, Patrick K. Albers¹, GoT2D Consortium[¶], Gil McVean¹, Michael Boehnke⁴, 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)

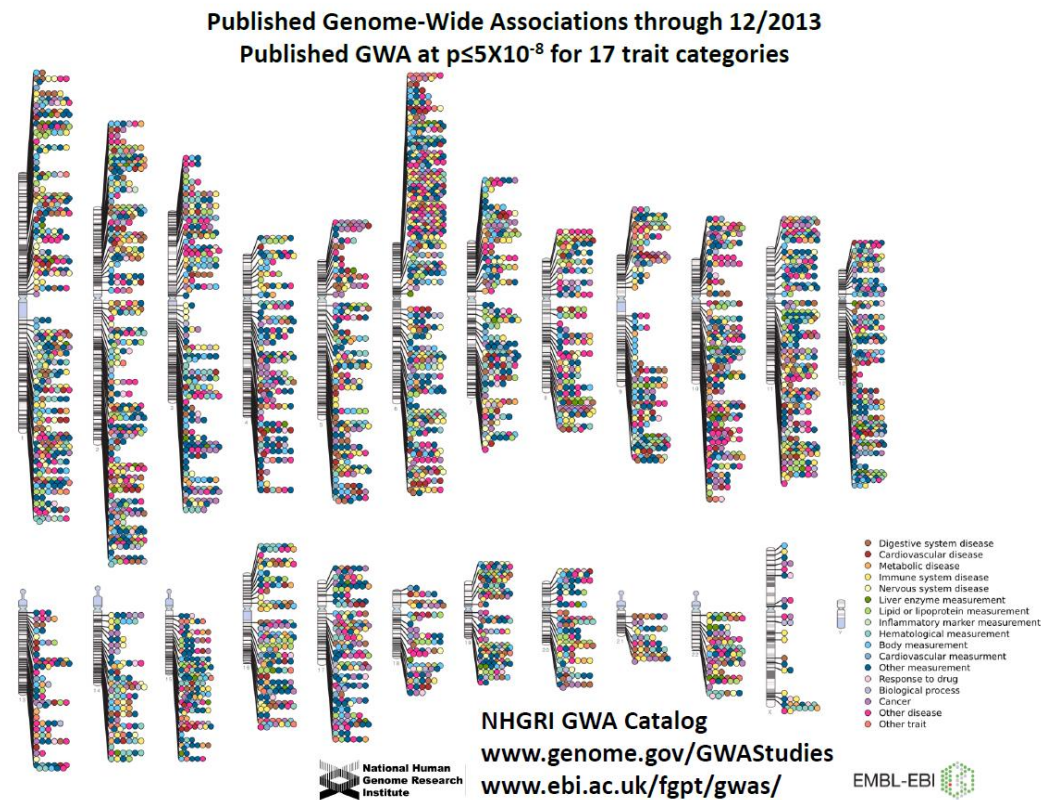
 OPEN ACCESS

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

(Moutsianas et al. 2015)

DNA sequence analyses

Motivation: classic GWAs are not sufficient to resolve “human genetic variation” ...



Sequencing projects

- In January 2008, when sequencing techniques became more advanced, more accurate, and less expensive, the **1000 Human Genome Project** was launched.

The main scope of this consortium is to sequence, ~1000 anonymous participants of different nationalities and concurrently compare these sequences to each other in order to better understand human genetic variation.

- Shortly after the 1000 Human Genome Project, the **1000 Plant Genome Project** (<http://www.onekp.com>) was launched, aiming to sequence and define the transcriptome of ~1000 plant species from different populations around the world.

Notably, out of the 370,000 green plants that are known today, only ~125,000 species have recorded gene entries in GenBank and many others still remain unclassified.

- While the 1000 Plant Genome Project was focused on comparing different plant species around the world, within the **1001 Genomes Project**, 1000 whole genomes of *A. Thaliana* plants across different places of the planet were sequenced.
- Similar to other consortiums, the **10,000 Genome Project** aims to create a collection of tissue and DNA specimens for 10,000 vertebrate species specifically designated for whole-genome sequencing.

Vertebrates have a series of nerves along the back which need support and protection. That need brings us to the backbones and notochords. Notochords were the first "backbones" serving as support structures.

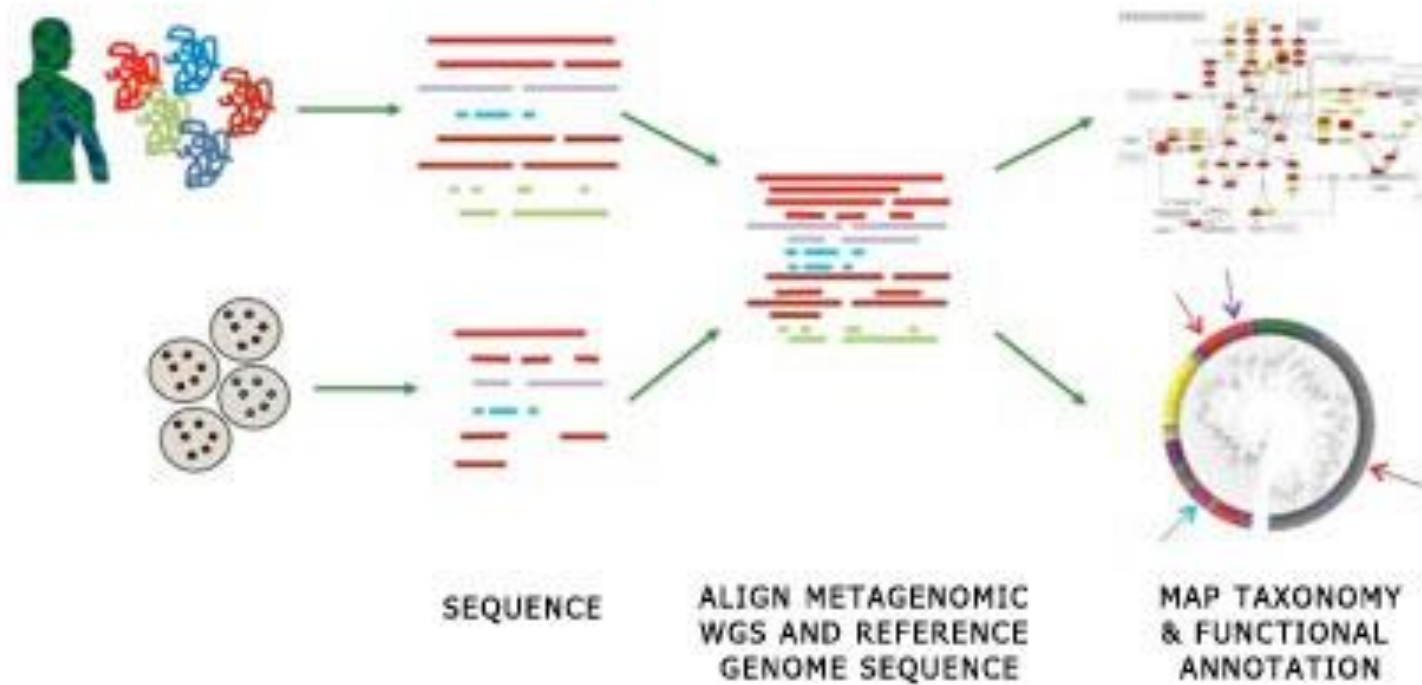
- The goal of the **1000 Fungal Genome Project** (<http://1000.fungalgenomes.org>) is to explore all areas of fungal biology.

- In human genetics, metagenome sequencing is becoming increasingly important, which lead to the **Human Microbiome Project** (<http://www.hmpdacc.org/>)
 - Metagenome sequencing is defined as an approach for the study of microbial populations in a sample representing a community by analysing the nucleotide sequence content.
 - The HMP plans to sequence 3000 genomes from both cultured and uncultured bacteria, plus several viral and small eukaryotic microbes isolated from human body sites.
 - This, in conjunction with reference genomes sequenced by HMP Demonstration Projects and other members of the International Human Microbiome Consortium (IHMC), will supplement the available selection of non-HMP funded human-associated reference genomes.

Metagenomics: the next hype

- Within the human body, it is estimated that there are 10x as many microbial cells as human cells.
- Our microbial partners carry out a number of metabolic reactions that are not encoded in the human genome and are necessary for human health (→ human genome = human genes + microbial genes).
- The majority of microbial species present in the human body have never been isolated, cultured or sequenced, typically due to the inability to reproduce necessary growth conditions in the lab (→ study microbial communities – metagenomics)
- In order to assign metagenomic sequence to taxonomic and functional groupings, and to differentiate the novel from the previously described, it is necessary to have a large pool of described genomes from the same environment (**reference genomes**).

Why Reference Sequences?



(<http://www.hmpdacc.org/>)

We have seen reference genomes before ... (the so-called “builds”)

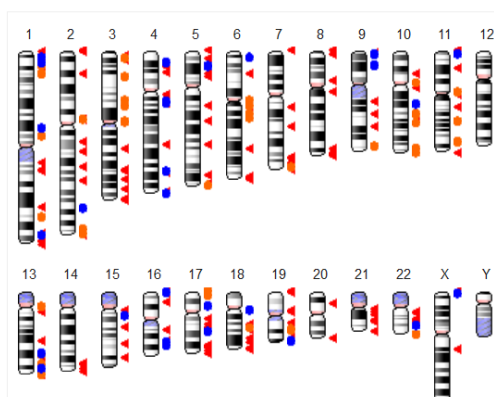
GRC Genome Reference Consortium

GRC Home
Data
Help
Report an Issue
Contact Us
Credits
Curators Only

Human Overview
Human Genome Issues
Human Assembly Data

Human Genome Overview

Information about the continuing improvement of the human genome



▲ Region containing alternate loci
● Region containing fix patches
● Region containing novel patches

Ideogram of the latest human assembly, GRCh38.p9

The GRC is working hard to provide the best possible reference assembly for human. We do this by both generating multiple representations ([alternate loci](#)) for regions that are too complex to be represented by a single path. Additionally, we are releasing regional fixes known as [patches](#). This allows users who are interested in a specific locus to get an improved representation without affecting users who need chromosome coordinate stability.

Download data:

- [GRCh38.p9 \(latest minor release\) FTP](#)
- [GRCh38 \(latest major release\) FTP](#)
- [Genomic regions under review FTP](#)
- [Current Tiling Path Files \(TPFs\)](#)

Transitioning to GRCh38? Try the [NCBI Remapping Service](#), which uses the same assembly-assembly alignments used by the GRC.

Next assembly update

The next assembly update (GRCh38.p10) will be a minor (patch) release in winter 2016.

GRC News

[New haplotype representations in the LRC_KIR region on chromosome 19q13.4](#) Nov 08, 2016

[Are you familiar with CYP2D6 and its importance in drug metabolism?](#) Oct 24, 2016

[see all](#)

Resolved Human Issues

[HG-2218](#) Oct 12, 2016

Evidence from MIsseq data for RP11-323F10, and RP11 WGS confirms the current assembly of AC017047.4 as

[HG-2416](#) Oct 12, 2016

This pathway has been uploaded to the Chr. 1 ALT_REF_LOCI_2 TPF.

[see all](#)

(<https://www.ncbi.nlm.nih.gov/grc/human>)

Which reference sequence?

Practical problems genomic reference sequence

- *for a human, a genomic reference sequence does not contain any useful information (a coding DNA reference sequence does)*
- a *gene can be very large* (over 2.0 Mb) - this makes nucleotide numbering based on a genomic reference sequence rather impractical (e.g. g.1567234_1567235insTG). Furthermore, genomic reference sequences based on GenBank NT_ files become increasingly long (e.g. the CFTR gene in [NT_007933.15](#), >77 Mb) and consequently lose their informativity. Downloading such large files is, even with good internet connections, time consuming and working with these files is rather difficult.
- when a genomic reference sequence is taken from a complete genome sequence, e.g. a bacterium or the human X-chromosome, the transcriptional orientation of the gene of interest may be on the *minus (-) strand*. This makes the description of sequence variants rather complicated, especially when the consequences on RNA and/or protein level need to be described; nucleotides on DNA and RNA level are complementary and numbering goes in different directions - a confusing situation that should be prevented.
- when different genes (partly) overlap, using the same or the minus (-) DNA strand, which reference sequence should one use to describe the variant and to which gene should the change be assigned? (see [Recommendations](#)).
- when the *gene sequence is incomplete* (especially when large introns are present) - a genomic sequence can not be used.
- genes may contain very large introns with many intronic (*length*) *variants* present in the population - it is thus very difficult to give **THE** genomic reference sequence (see [Genomic sequence changes regularly](#)).

Practical problems coding DNA reference sequence

- the exact *transcriptional start site* (cap-site) of a gene has often not been determined and/or its assignment is debated - the first nucleotide can thus not be assigned with certainty. The same might be true for the translation initiation site (ATG-codon).
- a gene may have *several transcripts*, using different promoters / 5'-first exons, alternatively spliced internal exons, different 3'-terminal exons and polyA-addition sites - **one** complete coding DNA reference sequence can thus not be generated (see [Alternatively spliced exons - nucleotide numbering](#)),
- the different transcripts may *encode different proteins* (isoforms) with, when different promoters are used, different N-terminal sequences and even using different reading frames in one or more exons. **One** complete protein reference sequence can thus not be assigned.
- when different genes (partly) overlap, using the same or the minus (-) DNA strand, which reference sequence should one use to describe the variant and to which gene should the change be assigned? (see [Recommendations](#)).

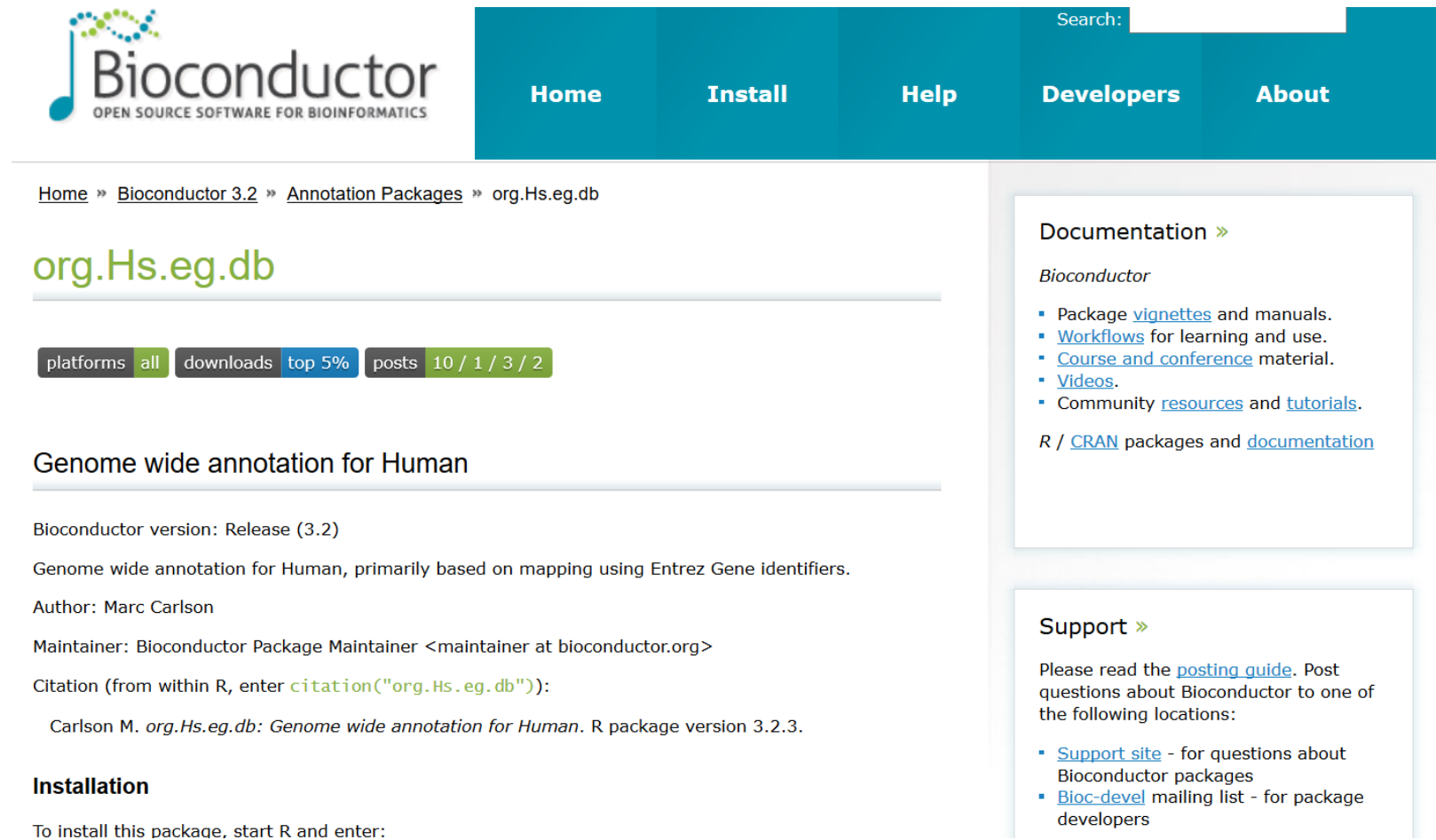
(<http://www.hgvs.org/mutnomen/refseq.html#standard>)

How is DNA sequencing used by scientists?

- A. In recent years, DNA sequencing technology has advanced many areas of science. For example, the field of **functional genomics** is concerned with
- figuring out what certain DNA sequences do, as well as
 - which pieces of DNA code for proteins and
 - which have important regulatory functions.
- B. An invaluable first step in making these determinations is **learning the nucleotide sequences** of the DNA segments under study.
- C. Another area of science that relies heavily on DNA sequencing is **comparative genomics**, in which researchers compare the genetic material of different organisms in order to learn about their evolutionary history and degree of relatedness.
- D. **Complex disease analysis**

A. Sequence annotation

(see practicals)



The screenshot shows the Bioconductor website interface. At the top left is the Bioconductor logo with the tagline "OPEN SOURCE SOFTWARE FOR BIOINFORMATICS". To the right is a teal navigation bar with links for "Home", "Install", "Help", "Developers", and "About", along with a search box. Below the navigation bar, the breadcrumb path is "Home » Bioconductor 3.2 » Annotation Packages » org.Hs.eg.db". The main title "org.Hs.eg.db" is displayed in green. Below the title are several colored buttons: "platforms", "all", "downloads", "top 5%", and "posts 10 / 1 / 3 / 2". The section "Genome wide annotation for Human" is highlighted with a horizontal line. Underneath, it states "Bioconductor version: Release (3.2)", "Genome wide annotation for Human, primarily based on mapping using Entrez Gene identifiers.", "Author: Marc Carlson", "Maintainer: Bioconductor Package Maintainer <maintainer at bioconductor.org>", and "Citation (from within R, enter `citation('org.Hs.eg.db')`): Carlsson M. *org.Hs.eg.db: Genome wide annotation for Human*. R package version 3.2.3." The "Installation" section begins with "To install this package, start R and enter:". On the right side, there are two sidebar boxes. The top one is titled "Documentation »" and lists resources for Bioconductor, including vignettes, workflows, course material, videos, and community resources. The bottom one is titled "Support »" and provides instructions on where to post questions, such as the support site and a mailing list.

Home » Bioconductor 3.2 » Annotation Packages » org.Hs.eg.db

org.Hs.eg.db

platforms all downloads top 5% posts 10 / 1 / 3 / 2

Genome wide annotation for Human

Bioconductor version: Release (3.2)

Genome wide annotation for Human, primarily based on mapping using Entrez Gene identifiers.

Author: Marc Carlson

Maintainer: Bioconductor Package Maintainer <maintainer at bioconductor.org>

Citation (from within R, enter `citation("org.Hs.eg.db")`):

Carlsson M. *org.Hs.eg.db: Genome wide annotation for Human*. R package version 3.2.3.

Installation

To install this package, start R and enter:

Documentation »

Bioconductor

- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about Bioconductor packages
- [Bioc-devel](#) mailing list - for package developers

B. Counting letters or words

- One of the most fundamental properties of a genome sequence is its GC content, the fraction of the sequence that consists of Gs and Cs, ie. the $\%(G+C)$.
- The GC content can be calculated as the percentage of the bases in the genome that are Gs or Cs. That is, $GC\ content = (\text{number of Gs} + \text{number of Cs}) * 100 / (\text{genome length})$. For example, if the genome is 100 bp, and 20 bases are Gs and 21 bases are Cs, then the GC content is $(20 + 21) * 100 / 100 = 41\%$.

Cell Reports
Article



Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition

Maayan Amit,^{1,4} Maya Donyo,^{1,4} Dror Hollander,^{1,4} Amir Goren,^{1,4} Eddo Kim,¹ Sahar Gelfman,¹ Galit Lev-Maor,¹ David Burstein,² Schraga Schwartz,³ Benny Postolsky,¹ Tal Pupko,² and Gil Ast^{1,*}

- The **CpG sites** or **CG sites** are regions of DNA where a cytosine nucleotide occurs next to a guanine nucleotide in the linear sequence of bases along its length. "CpG" is shorthand for "—C—phosphate—G—", that is, cytosine and guanine separated by only one phosphate. The "CpG" notation is used to distinguish this linear sequence from the CG base-pairing of cytosine and guanine.

(https://en.wikipedia.org/wiki/CpG_site)

```

CATTCCGCTTCTCTCCCGAGGTGGCGCGTGGGA
GGTGTTTTGTCTCGGGTCTGTAAGAATAGGCCAGG
CAGCTTCCCGCGGGATGCGCTCATCCCCTCTCGG
GGTTCGCTCCCACCGCGCGCGTTCGGCCGGTT
CCGCCTGCGAGATGTTTTCCGACCGACAAATGATTC
CACTCTCGCGCGCTCCCATGTTGATCCCAGCTCCT
CTGCGGGCGTCAGGACCCCTGGGCCCGCCCG
CTCCACTCAGTCAATCTTTGTCCCCTATAAGGCG
GATTATCGGGGTGGCTGGGGGCGGCTGATTCGGA
CGAATGCCCTTGGGGGTCACC CGGGAGGGAACTC
CGGGCTCCGGGCTTTGGCCAGCCCGCACCCCTGGT
TGAGCCGGGCCCGAGGGCCACCAGGGGGCGCTCG
ATGTTCTGCAGCCCCCGCAGCAGCCCCACTCC
CCGGCTCACCCCTACGATTGGCTGGCCCGCCCGAG
CTCTGTGCTGTGATTGGTACAGCCCGTGTCCGTC
GCGGGCGCGGGCGGATACGAGGTGACGCGCA
GAGGCCAGCTCGGGGCGGTGTCCCGCGCGCGG
GACTGCGGGCGGAGTTTCCGCGAGGGCCGAGCG
GGGCAGTGTGACGGCAGCGGTCCTGGGAGGCGC
CCGCGCGCGTCCGAGCAGCTCCCCTCCTCGCA
GCCTCACCGCGGCGTCCCGCCGCCCTGGCC
TCCCGCACTCGCGCACTCCTGTCCCGCCACCC
GCCACCTCCCACCTCGATGCGGTGC CGGCTGC
TGCGTGATGGGGCTGCGGAGCGGCGCCTGCGG
CTCGCGCGGCGCTGCTCGGCTGAGGTGCGT
CGGTGCCCGGCCCCCGCCGCCCGCGCGCGCG
GGCTCCTGTTGACC CGGTC CGCCCGT CGGTCTGC
AGCGCGGCTGAGGTAAGGCGCGGGGCTGGCCG
CGGTTGGCGCGCGGTCCGCGGGTTGGGGAGGG
GGCCGCTTCCCGCGGGGAGGAGCGGCCGGGCCG
GGTCCGGGCGGGTCTGAGGGGA
CTCTTAGTTTTGGGTGCATTTGTCTGGTCTTCCAAA
CTAGATTGAAAGCTCTGAAAAAAAAAACTATCTTGT
GTTTCTATCTGTTGAGCTCATAGTAGGTATCCAGGA
AGTAGTAGGGTTGACTGCATTGATTTGGGACTACAC
TGGGAGTTTTCTTCCCATCTCCCTTTAGTTTTCT
TTTTTCTTTCTTTCTTTCTTTTTTTTTTTTTTTTT
TTGAGATGTCTTTGCTCAGTCCCCCAGGCTGGA
GTGCAGTGGTGCGATCTTGGCTCACTGTAGCCTCC
ACCTCCCAGGTTCAAGCAATTCTACTGCCTTAGCCT
CCCGAGTAGCTGGGATTACAAGCACC CGCCACCAT
TCCTGGCTAATTTTTTTTTTTGTATTTTAGTTGAGA
CAGGGTTTACCATGTTGGTGATGCTGGTCTCAGA
CTCCTGGGGCCTAGCGATCCCCCTGCCTCAGCCT
CCCAGAGTGTAGGATTACAGGCATGAGCCACTGT
ACC CGGCCTCTCTCCAGTTTCCAGTTGGAATCCAA
GGGAAGTAAGTTAAGATAAAGTTACGATTTTGAAT
CTTTGGATTCAGAAGAATTTGTCACCTTTAACACCT
AGAGTTGAACTTCATACCTGGAGAGCCTTAACATT
AAGCCCTAGCCAGCCTCCAGCAAGTGGACATTGGT
CAGGTTTGGCAGGATTCTCCCTGAAGTGGACT
GAGAGCCACACCCTGGCCTGTACCATACCCATCC
CCTATCCTTAGTGAAGCAAACTCCTTTGTTCCCTT
CTCCTTCTCCTAGTGACAGGAAATATTGTGATCCTA
AGAATGAAAATAGCTTGTACCTCGTGGCCTCAG
GCCTCTTGACTTCAGGCGGTTCTGTTAATCAAGT
GACATCTTCCCGAGGCTCCCTGAATGTGGCAGATG
AAAGAGACTAGTTCAACCCTGACCTGAGGGGAAAG
CCTTTGTGAAGGGTCAGGAG

```

C. Comparing multiple sequences (evolutionary history)

- After collection of a set of related sequences, how can we compare them as a set?
- How should we line up the sequences so that the most similar portions are together?
- What do we do with sequences of different length?

```

                2430          2440          2450          2460          2470
HSA128 CACTTCCCCTAT---GCAGGTGTCCAACGGATGTGTGAGTAAAATTCTGGGCAGGTATTA
      ::  :::::  ::  ::::::::::::::::::::::::::::::::::::::::::::::::::::
pax6  CATTTCCCGAATTCTGCAGGTGTCCAACGGATGTGTGAGTAAAATTCTGGGCAGGTATTA
      540          550          560          570          580          590

                2480          2490          2500          2510          2520          2530
HSA128 CGAGACTGGCTCCATCAGACCCAGGGCAATCGGTGGTAGTAAACCGAGAGTAGCGACTCC
      ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
pax6  CGAGACTGGCTCCATCAGACCCAGGGCAATCGGTGGTAGTAAACCGAGAGTAGCGACTCC
      600          610          620          630          640          650

                2540          2550          2560          2570          2580          2590
HSA128 AGAAGTTGTAAGCAAAATAGCCCAGTATAAGCGGGAGTGCCCGTCCATCTTTGCTTGGGA

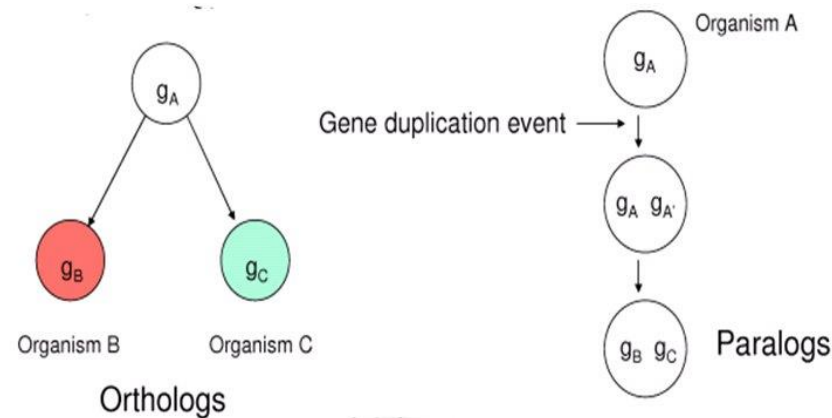
```


- Are sequences alike?
 - **Heterologs.** *{Heterologs differ in both origin and activity.}*
 - **Homologs.** *{Homologs have common origins but may or may not have common activity.}*
 - Genes that share an arbitrary threshold level of similarity determined by alignment of matching bases are termed **homologous**.
 - **Homology** is a qualitative term that describes a relationship between genes and is based upon the quantitative similarity.
 - **Similarity** is a quantitative term that defines the degree of sequence match between two compared sequences.
 - Homology implies that the compared sequences diverged in evolution from a common origin.

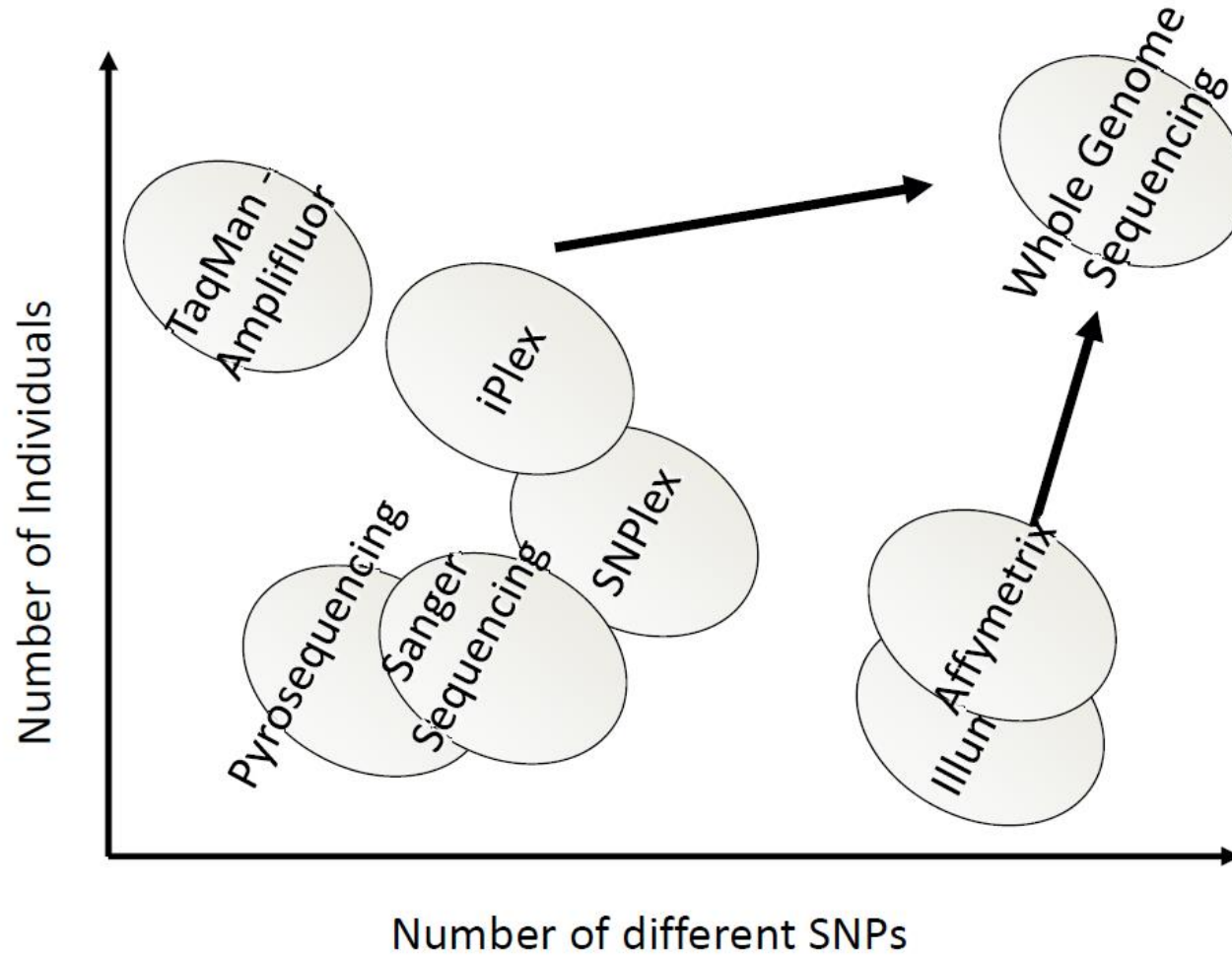
– **Analogs.** {*Analogs have common activity but not common origin.*}

- Genes or proteins that display the same activity but lack sufficient similarity to imply common origin are said to have **analogous** activity.
- The implication is that analogous proteins followed evolutionary pathways from different origins to converge upon the same activity.
- Analogs have homologous activity but heterologous origins.

– **Paralogs.** {*Paralogs are homologs produced by gene duplication.*}



D. Genomic variation for complex diseases

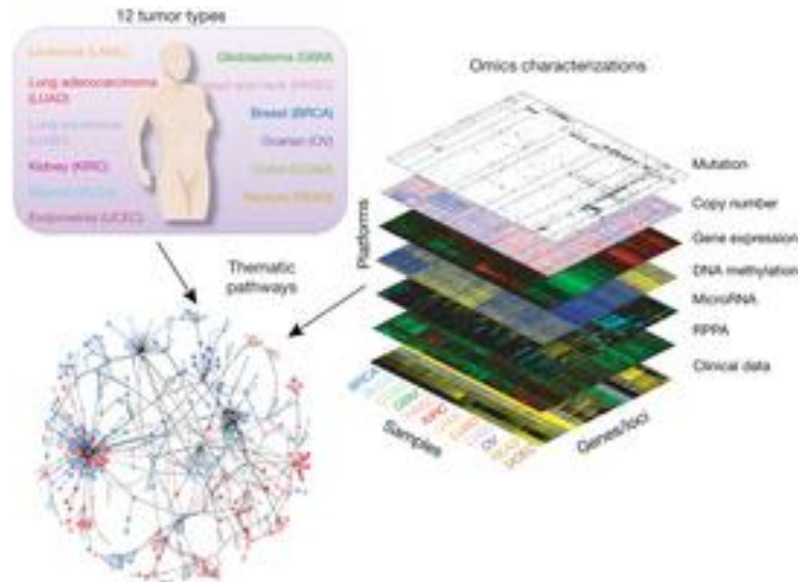


Genomic variation for complex diseases

- High throughput in numbers of individuals and variants matters
 - Only identical twins have the same DNA sequences
 - 2×10^7 bases in the human genome are variable
 - Average differences between two humans: 0.1% of their genome shared
 - Difference between human and chimpanzee is about 1% **DNA-wise!**

Genomic variation for complex diseases - continued

- Sequencing (DNA, RNA, ...) has already aided complex disease research by allowing scientists to catalogue certain genetic variations between individuals that may influence their susceptibility to different conditions or by identifying similar “patterns” between subgroups of patients (i.e., **molecular reclassification of patients → obtaining an accurate trait def**).



The TCGA Pan-Cancer project assembled data from thousands of patients with primary tumors occurring in different sites of the body, covering 12 tumor types. The idea of the TCGA PanCancer project was to integrate data set for comparing and contrasting multiple tumor types ... (Weinstein et al. 2013)

Genomic variation for complex diseases - continued

- In general, there are 3 common scenarios for human geneticists using NGS data to **understand complex diseases**:
 - Identification of causative genes in Mendelian disorders (germline mutations)
 - Identification of candidate genes in complex diseases for further functional studies (**complex diseases** are governed by multiple genes that are possibly interacting with each other and/or with environment)
 - Identification of constitutional mutations as well as driver and passenger genes in cancer (somatic mutations) (Pabinger et al 2013)

A **germline mutation** is one that was passed on to offspring because the egg or sperm cell was mutated.

A **somatic mutation** is a mutation of the somatic cells (all cells except sex cells) that cannot be passed on to offspring.

Whether A, B, C or D is the aim ...

the starting point
of any sequencing project
is the development of an appropriate study design,
which should start
with a well-defined question
(biological / medical / research / ...)

The application determines the analysis (software) tool

Suppose: You have been given a 5 KB piece of DNA sequence ...

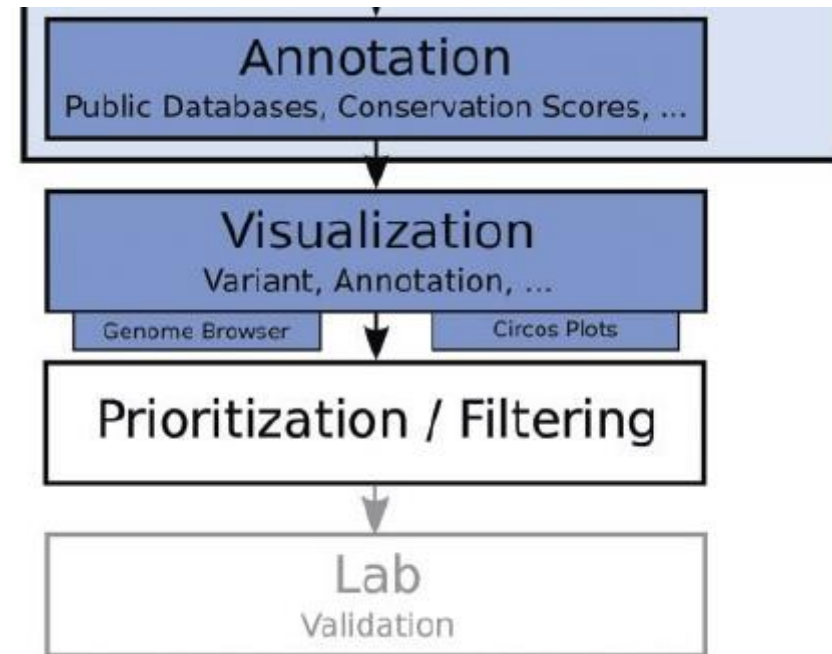
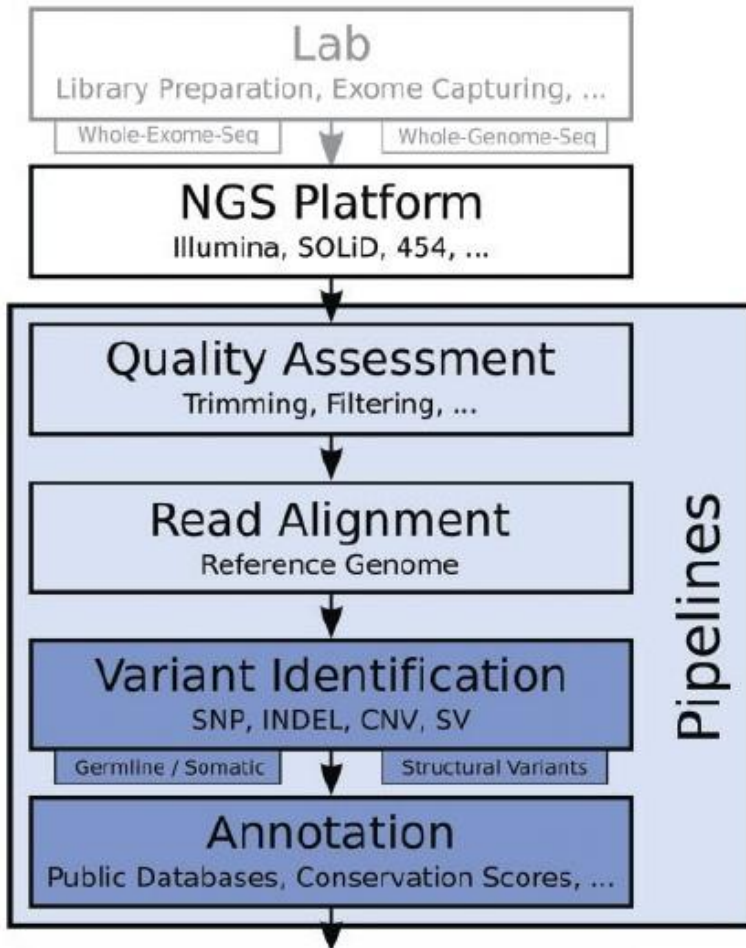
What to do next? ...

- GeneScan: find any exons in the DNA sequence and generate a predicted protein sequence
- ScanProsite: scan the protein sequence for domains/motifs/patterns found in the prosite database [**Motifs** are structural characteristics and **domains** are functional regions]
- BLASTP: run a BLASTP search against the Swissprot database find some of the best matches (hits) and copy each protein sequence into a word doc for the alignment
- MultAlin: conduct protein sequence alignments from the BLASTP search

The application determines the software tool

- The rule of thumb in the genomics community is that every dollar spent on sequencing hardware must be matched by a comparable investment in informatics (www.the-scientist.com/2011/3/1/60/1)
- There is a constant stream of new software
 - What is its quality?
 - How to install it?
 - How to get it working?

Common workflow for whole-exome and whole genome sequencing



(Pabinger et al. 2013)

R code for DNA seq analysis problems (at home)

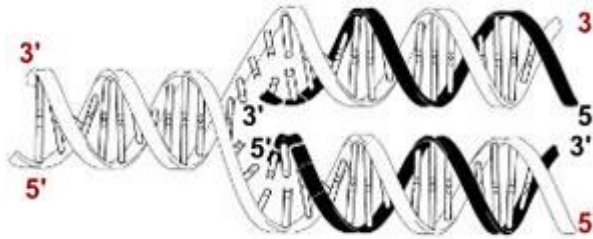
- R scripts illustrating relevant R packages for sequence pattern recognition and sequence-based analytics (see also practical session), includes:
 - DNA sequence statistics: <http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter1.html>
 - Querying sequence data bases: <http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter3.html>
 - Computational gene finding: <http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter7.html>

Introduction to investigating frequencies of occurrences of words

- Words are short strings of letters drawn from an alphabet
- In the case of DNA, the set of letters is A, C, T, G
- A word of length k is called a k -word or k -tuple
- Differences in word frequencies help to differentiate between different DNA sequence sources or regions
- Examples: 1-tuple: individual nucleotide; 2-tuple: dinucleotide; 3-tuple: codon
- The distributions of the nucleotides over the DNA sequences have been studied for many years → hidden correlations in the sequences

Biological words of length 1 – base composition

- There are constraints on base composition imposed by the genetic code
- The distribution of individual bases within a DNA molecule is not ordinarily uniform
 - There may be an excess of G over C on the leading strands



- This can be described by the “GC skew”, characterized by:
 - $(\#G - \#C) / (\#G + \#C)$
 - # = nr of
- What is the implication for AT skew on the lagging strand?

Biological words of length 1 – base composition

- GC or AT skew sign changes link to where DNA replication starts or finishes.
- Originally this asymmetric nucleotide composition was explained as different mechanism used in DNA replication between leading strand and lagging strand
- But recent research (2013) shows there is much more to it:

Research

GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination

Paul A. Ginno,^{1,3,4} Yoong Wearn Lim,^{1,3} Paul L. Lott,² Ian Korf,^{1,2}
and Frédéric Chédin^{1,2,5}

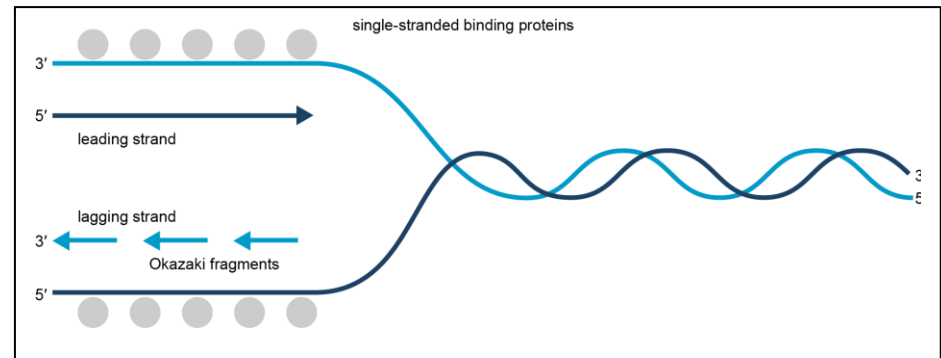
¹Department of Molecular and Cellular Biology, ²Genome Center, University of California, Davis, California 95616, USA

Strand asymmetry in the distribution of guanines and cytosines, measured by GC skew, predisposes DNA sequences toward R-loop formation upon transcription. Previous work revealed that GC skew and R-loop formation associate with a core set of unmethylated CpG island (CGI) promoters in the human genome. Here, we show that GC skew can distinguish four classes of promoters, including three types of CGI promoters, each associated with unique epigenetic and gene ontology signatures. In particular, we identify a strong and a weak class of CGI promoters and show that these loci

Biological words of length 1 - base composition

- DNA biosynthesis proceeds in the 5'- to 3'-direction. This makes it impossible for DNA polymerases to synthesize both strands simultaneously. A portion of the double helix must first unwind, and this is mediated by helicase enzymes.
- The leading strand is synthesized continuously but the opposite strand is copied in short bursts of about 1000 bases, as the lagging

strand template becomes available. The resulting short strands are called Okazaki fragments (after their discoverers, Reiji and Tsuneko Okazaki).



Probability distributions

Probability is the science of uncertainty

1. Rules → data: given the rules, describe the likelihoods of various events occurring
2. Probability is about prediction – looking forwards
3. Probability is mathematics

Statistics is the science of data

1. Rules \leftarrow data: given only the data, try to guess what the rules were. That is, some probability model controlled what data came out, and the best we can do is guess – or approximate – what that model was. We might guess wrong, we might refine our guess as we obtain / collect more data
2. Statistics is about looking backward
3. Statistics is an art. It uses mathematical methods but it is much more than maths alone
4. Once we make our best *statistical guess* about what the probability model is (what the rules are), based on looking backward, we can then use that probability model to predict the future \rightarrow the purpose of statistics is to make inference about unknown quantities from samples of data.

Statistics is the science of data

- Probability distributions are a fundamental concept in statistics.
- Before computing an interval or test based on a distributional assumption, we need to verify that the assumption is justified for the given data set.
- For this chapter, the distribution does not always need to be the best-fitting distribution for the data, but an adequate enough model so that the statistical technique yields valid conclusions.
- Simulation studies: one way to obtain empirical evidence for a probability model

Assumptions

- Simple rules specifying a probability model:
 - First base in sequence is either A, C, T or G with prob p_A, p_C, p_T, p_G
 - Suppose the first r bases have been generated, while generating the base at position $r+1$, no attention is paid to what has been generated before.
- Then we can actually generate A, C, T or G with the probabilities above
- Notation for the output of a random string of n bases may be: L_1, L_2, \dots, L_n
(L_i = base inserted at position i of the sequence)
- Whatever we would like to do with such strings, we will need to introduce the concept of a random variable

Probability distributions

- Suppose the “machine” we are using produces an output X that takes exactly 1 of the J possible values in a set $\chi = \{l_1, l_2, \dots, l_n\}$
 - In the DNA sequence $J=4$ and $\chi = \{A, C, T, G\}$
 - L is a discrete random variables (since its values are uncertain)
 - If p_j is the prob that the value (realization of the random variable L) l_j occurs, then
 - $p_1, \dots, p_J \geq 0$ and $p_1 + \dots + p_J = 1$
- The probability distribution (probability mass function) of L is given by the collection p_1, \dots, p_J
 - $P(L=l_j) = p_j, j=1, \dots, J$
- The probability that an event S occurs (subset of χ) is $P(L \in S) = \sum_{j:l_j \in S} (p_j)$

Probability distributions

- What is the probability distribution of the number of times a “given pattern” occurs in a random DNA sequence L_1, \dots, L_n ?

- New sequence X_1, \dots, X_n :

$$X_i=1 \text{ if } L_i=A \text{ and } X_i=0 \text{ else}$$

- The number of times N that A appears is the sum

$$N=X_1+\dots+X_n$$

- The prob distr of each of the X_i :

$$P(X_i=1) = P(L_i=A)=p_A$$

$$P(X_i=0) = P(L_i=C \text{ or } G \text{ or } T) = 1 - p_A$$

- **What is a “typical” value of N ?**

- Depends on how the individual X_i (for different i) are interrelated

Independence

- Discrete random variables X_1, \dots, X_n are said to be independent if for any subset of random variables and actual values, the joint distribution equals the product of the component distributions
- According to our simple model, the L_i are independent and hence

$$P(L_1=l_1, L_2=l_2, \dots, L_n=l_n) = P(L_1=l_1) P(L_2=l_2) \dots P(L_n=l_n)$$

Expected values and variances

- Mean and variance are two important properties of real-valued random variables and corresponding probability distributions.
- The “mean” of a discrete random variable X taking values x_1, x_2, \dots (denoted EX (or $E(X)$ or $E[X]$), where E stands for expectation, which is another term for mean) is defined as:

$$E(X) = \sum_i x_i P(X = x_i)$$

- $E(X_i) = 1 \times p_A + 0 \times (1 - p_A)$
 - If $Y = c X$, then $E(Y) = c E(X)$
 - $E(X_1 + \dots + X_n) = E(X_1) + \dots + E(X_n)$
- Because X_i are assumed to be independent and identically distributed (iid):

$$E(X_1 + \dots + X_n) = n E(X_1) = n p_A$$

Expected values and variances

- The idea is to use squared deviations of X from its center (expressed by the mean). Expanding the square and using the linearity properties of the mean, the $\text{Var}(X)$ can also be written as:

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

- If $Y=c X$ then $\text{Var}(Y) = c^2 \text{Var}(X)$
 - The variance of a sum of independent random variables is the sum of the individual variances
-
- For the random variables X_i :
 $\text{Var}(X_i) = [1^2 \times p_A + 0^2 \times (1 - p_A)] - p_A^2 = p_A(1 - p_A)$
 $\text{Var}(N) = n \text{Var}(X_1) = np_A(1 - p_A)$

Expected values and variances

- The expected value of a random variable X gives a measure of its location. Variance is another property of a probability distribution dealing with the spread or variability of a random variable around its mean.

$$\text{Var}(X) = E ([X - E(X)]^2)$$

- The positive square root of the variance of X is called its standard deviation $\text{sd}(X)$

The binomial distribution

- The binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial. These outcomes are appropriately labeled "success" and "failure". The binomial distribution is used to obtain the probability of observing x successes in a fixed number of trials, with the probability of success on a single trial denoted by p . The binomial distribution assumes that p is fixed for all trials.
- The formula for the binomial probability mass function is :

$$P(N = j) = \binom{n}{j} p^j (1 - p)^{n-j}, j = 0, 1, \dots, n$$

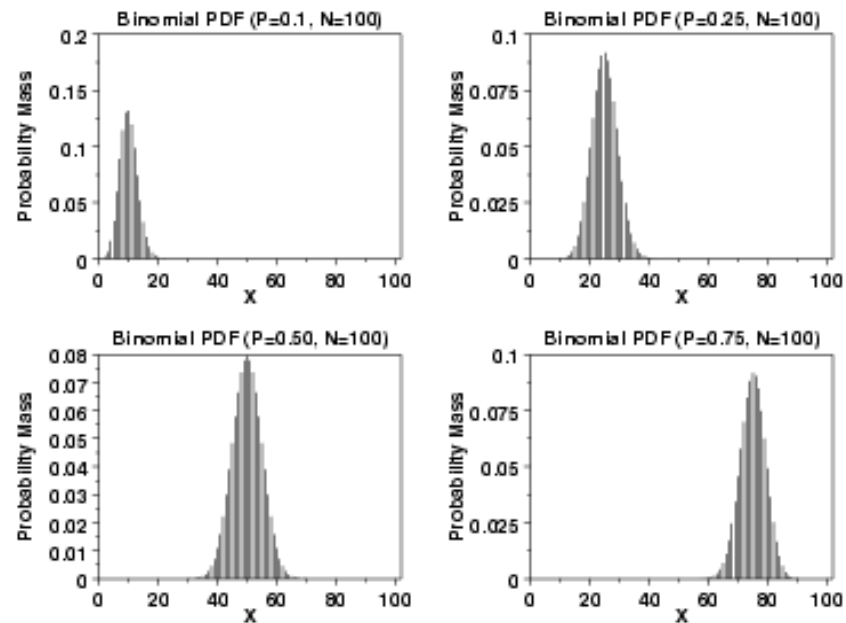
with the binomial coefficient $\binom{n}{j}$ determined by

$$\binom{n}{j} = \frac{n!}{j! (n - j)!}$$

and $j! = j(j-1)(j-2)\dots 3.2.1$, $0! = 1$

The binomial distribution

- The mean is np and the variance is $np(1-p)$
- The following is the plot of the binomial probability density function for four values of p and $n = 100$.



Simulating from probability distributions

- The idea is that we can study the properties of the distribution of N when we can get our computer to output numbers N_1, \dots, N_n having the same distribution as N

- We can use the sample mean to estimate the expected value $E(N)$:

$$\bar{N} = (N_1 + \dots + N_n)/n$$

- Similarly, we can use the sample variance to estimate the true variance of N :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (N_i - \bar{N})^2$$

Why do we use $(n-1)$ and not n in the denominator?

Simulating from probability distributions

- What is needed to produce such a string of observations?
 - Access to pseudo-random numbers: random variables that are uniformly distributed on (0,1): any number between 0 and 1 is a possible outcome and each is equally likely
- In practice, simulating an observation with the distribution of X_1 :
 - Take a uniform random number u
 - Set $X_1=1$ if $U \leq p \equiv p_A$ and 0 otherwise.
 - Why does this work? ... $P(X_1 = 1) = P(U \leq p_A) = p_A$
 - Repeating this procedure n times results in a sequence X_1, \dots, X_n from which N can be computed by adding the X 's

Simulating from probability distributions

- Simulate a sequence of bases L_1, \dots, L_n :
 - Divide the interval $(0,1)$ in 4 intervals with endpoints
 $p_A, p_A + p_C, p_A + p_C + p_G, 1$
 - If the simulated u lies in the leftmost interval, $L_1=A$
 - If u lies in the second interval, $L_1=C$; if in the third, $L_1=G$ and otherwise $L_1=T$
 - Repeating this procedure n times with different values for U results in a sequence L_1, \dots, L_n

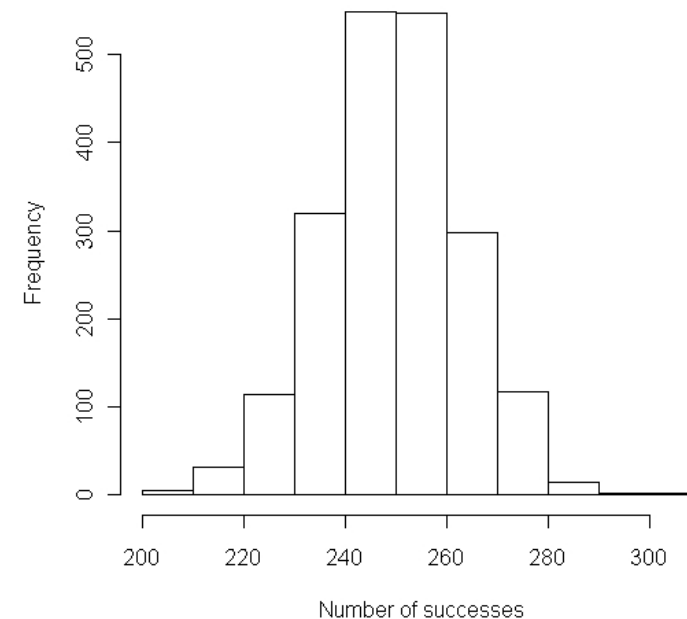
- Use the “sample” function in R:

```
pi <- c(0.25,0.75)
x<-c(1,0)
set.seed(2009)
sample(x,10,replace=TRUE,pi)
```

Simulating from probability distributions

- By looking through a given simulated sequence, we can count the number of times a particular pattern arises (for instance, the base A)
- By repeatedly generating sequences and analyzing each of them, we can get a feel for whether or not our particular pattern of interest is unusual

nr of successes in 1000 trials



```
x<- rbinom(2000,1000,0.25)
mean(x)
sd(x)^2
hist(x,xlab="Number of successes",main="")
```

R documentation

Binomial {stats}

R Documentation

The Binomial Distribution

Description

Density, distribution function, quantile function and random generation for the binomial distribution with parameters `size` and `prob`.

This is conventionally interpreted as the number of ‘successes’ in `size` trials.

Usage

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```

Arguments

<code>x, q</code>	vector of quantiles.
<code>p</code>	vector of probabilities.
<code>n</code>	number of observations. If <code>length(n) > 1</code> , the length is taken to be the number required.
<code>size</code>	number of trials (zero or more).

<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Binomial.html>

```
> rbinom(1,1000,0.25)
```

```
[1] 250
```


Simulating from probability distributions

- Using R code:

```
x<- rbinom(2000,1000,0.25)
mean(x)
sd(x)^2
hist(x,xlab="Number of successes",main="")
```

What is the number of observations (nr of “strings” generated)?

Simulating from probability distributions

- Using R code:

```
x<- rbinom(2000,1000,0.25)
mean(x)
sd(x)^2
hist(x,xlab="Number of successes",main="")
```

Number of observations = 2000

Number of trials = 1000

What is the number of observations?

Exercise

- Suppose we have a sequence of 1000bp and assume that every base occurs with equal probability. How likely are we to observe at least 300 A's in such a sequence?
 - Exact computation using a closed form of the relevant distribution
 - Approximate via simulation
 - Approximate using the Central Limit Theory

Solution 1: Exact computation via closed form of relevant distribution

- The formula for the binomial probability mass function is :

$$P(N = j) = \binom{n}{j} p^j (1 - p)^{n-j}, j = 0, 1, \dots, n$$

and therefore

$$\begin{aligned} P(N \geq 300) &= \sum_{j=300}^{1000} \binom{1000}{j} (1/4)^j (1 - 1/4)^{1000-j} \\ &= 0.00019359032194965841 \end{aligned}$$

	P: exactly 300 out of 1000	
Method 1. exact binomial calculation	0.00004566114740576488	
Method 2. approximation via normal	0.000038	
Method 3. approximation via Poisson	-----	
	P: 300 or fewer out of 1000	
Method 1. exact binomial calculation	0.9998520708293378	
Method 2. approximation via normal	0.999885	
Method 3. approximation via Poisson	-----	
	P: 300 or more out of 1000	
Method 1. exact binomial calculation	0.00019359032194965841	
Method 2. approximation via normal	0.000153	
Method 3. approximation via Poisson	-----	
For hypothesis testing	P: 300 or more out of 1000	
	One-Tail	Two-Tail
Method 1. exact binomial calculation	0.00019359032194965841	0.0003025705168772097
Method 2. approximation via normal	0.000153	0.000306
Method 3. approximation via Poisson	-----	-----

(<http://faculty.vassar.edu/lowry/binomialX.html>)

Solution 2: Approximate via simulation

- Using R code and simulations from the theoretical distribution, $P(N \geq 300)$ can be estimated as 0.000196 via

```
x<- rbinom(1000000,1000,0.25)
sum(x>=300)/1000000
```

- Note that the probability $P(N \geq 300)$ is estimated to be 0.0001479292 via

```
1-pbinom(300,size=1000,prob=0.25)
pbinom(300,size=1000,prob=0.25,lower.tail=FALSE)
```

Solution 3: Approximate via Central Limit Theory

- The central limit theorem offers a 3rd way to compute probabilities of a distribution
- It applies to sums or averages of iid random variables
- Assuming that X_1, \dots, X_n are iid random variables with mean μ and variance σ^2 , then we know that for the sample average

$$\bar{X}_n = \frac{1}{n} (X_1 + \dots + X_n),$$

$$E(\bar{X}_n) = \mu \text{ and } \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

- Hence,

$$E\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) = 0, \text{Var}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) = 1$$

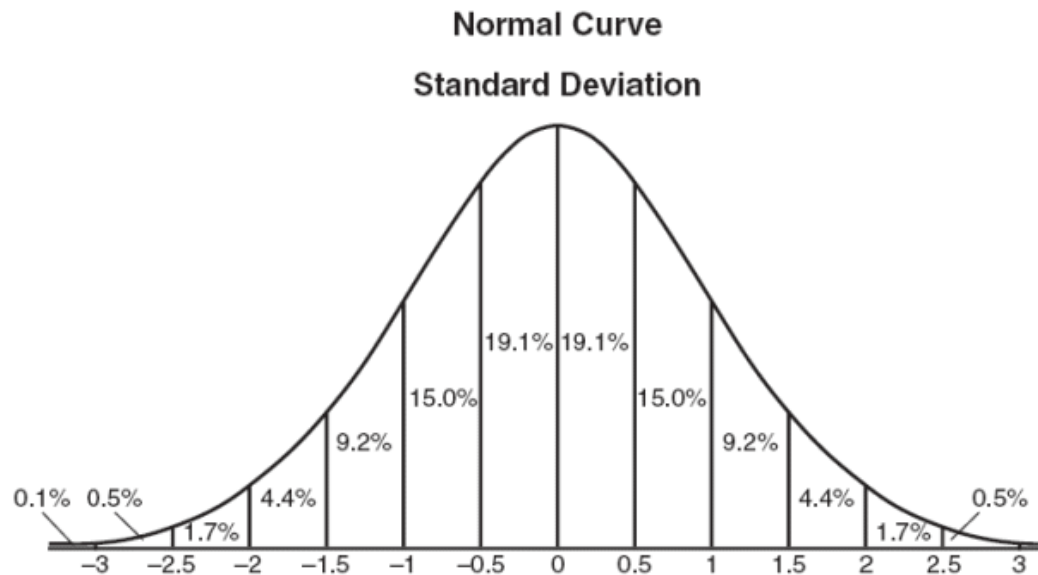
Approximate via Central Limit Theory

- The central limit theorem states that if the sample size n is large enough,

$$P\left(a \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq b\right) \approx \phi(b) - \phi(a),$$

with $\phi(\cdot)$ the standard normal distribution defined as

$$\phi(z) = P(Z \leq z) = \int_{-\infty}^z \phi(x) dx$$

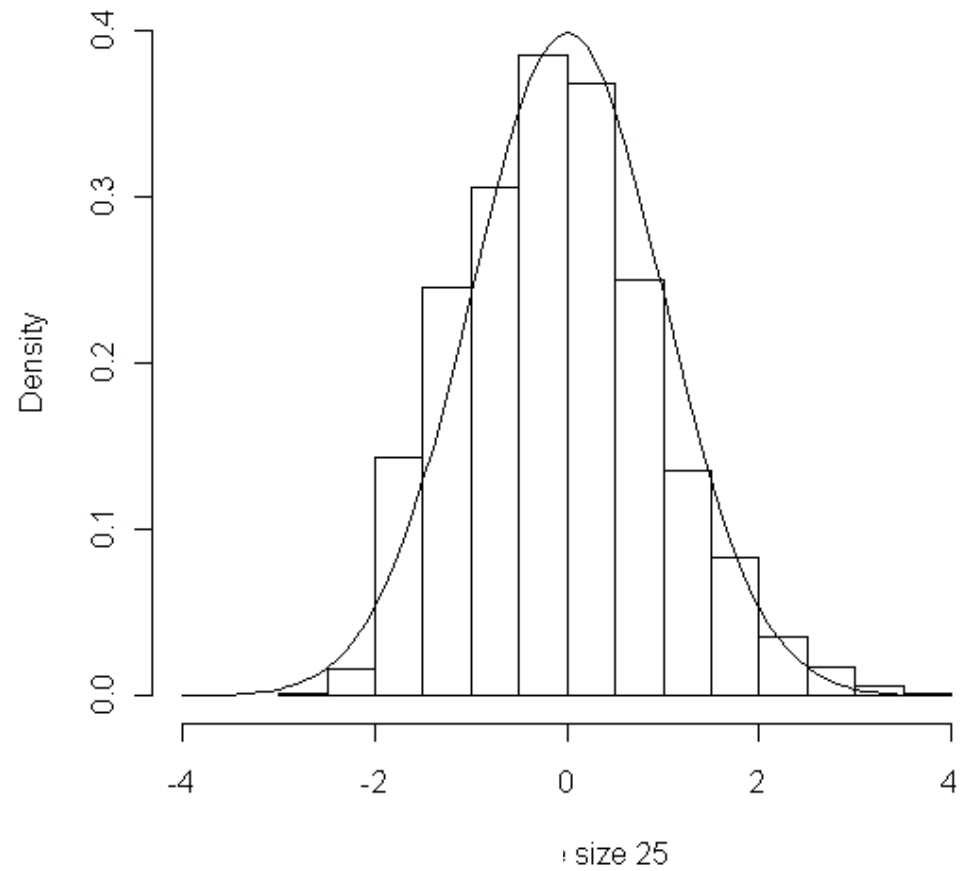


Approximate via Central Limit Theory

- The central limit theorem in action using R code:

```
bin25<-rbinom(1000,25,0.25)
av.bin25 <- 25*0.25
stdev.bin25 <- sqrt(25*0.25*0.75)
bin25<-(bin25-av.bin25)/stdev.bin25
hist(bin25,xlim=c(-4,4),ylim=c(0.0,0.4),prob=TRUE,xlab="Sample size
25",main="")
x<-seq(-4,4,0.1)
lines(x,dnorm(x))
```

Approximate via Central Limit Theory



Approximate via Central Limit Theory

- Estimating the quantity $P(N \geq 300)$ when N has a binomial distribution with parameters $n=1000$ and $p=0.25$,

$$E(N) = n\mu = 1000 \times 0.25 = 250,$$

$$sd(N) = \sqrt{n} \sigma = \sqrt{1000 \times \frac{1}{4} \times \frac{3}{4}} \approx 13.693$$

$$P(N \geq 300) = P\left(\frac{N - 250}{13.693} > \frac{300 - 250}{13.693}\right)$$

$$\approx P(Z > 3.651501) = 0.0001303560$$

- R code:

```
pnorm(3.651501,lower.tail=FALSE)
```

How do the estimates of $P(N \geq 300)$ compare?

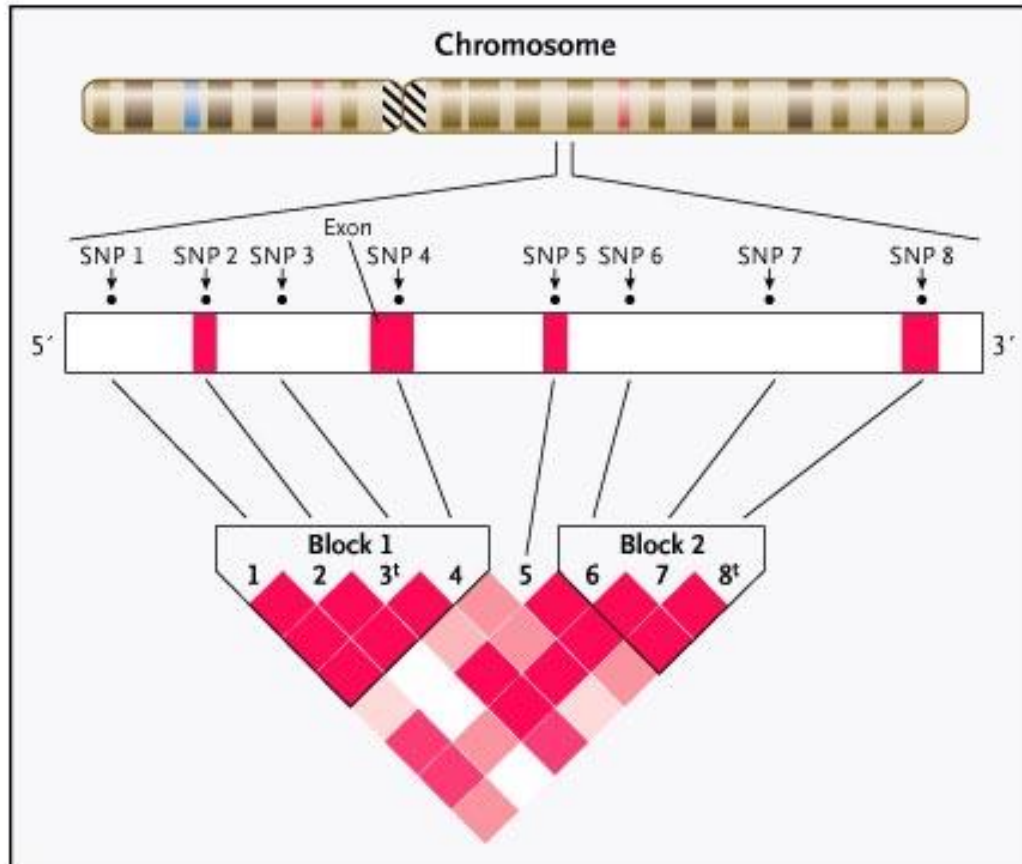
3 When effects become non-independent

Linkage disequilibrium (LD) between genetic markers

- **Linkage Disequilibrium (LD)** is a measure of co-segregation of alleles in a population. Strictly speaking, it therefore refers to linkage + 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

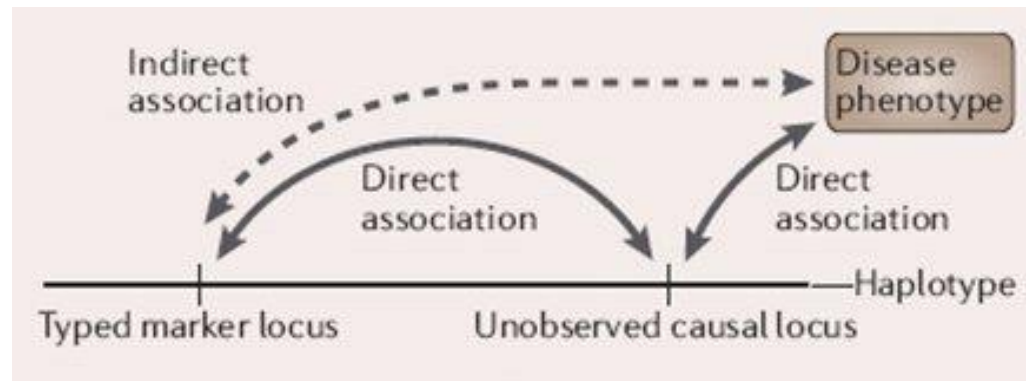
- The measure D is defined as the difference between the observed and expected (under the null hypothesis of independence) proportion of “haplotypes” bearing specific alleles at two loci: $p_{AB} - p_A p_B$

	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 rationale 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 color) 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

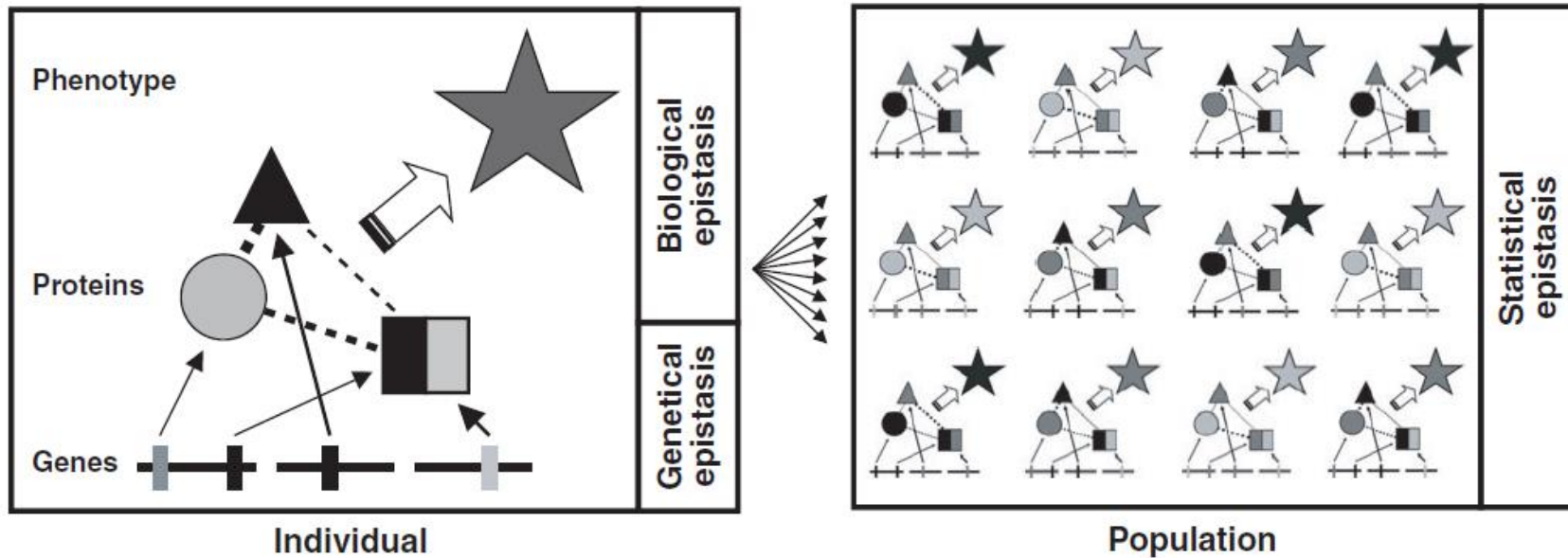
- Our ability to detect epistasis depends on what we mean by epistasis

“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).

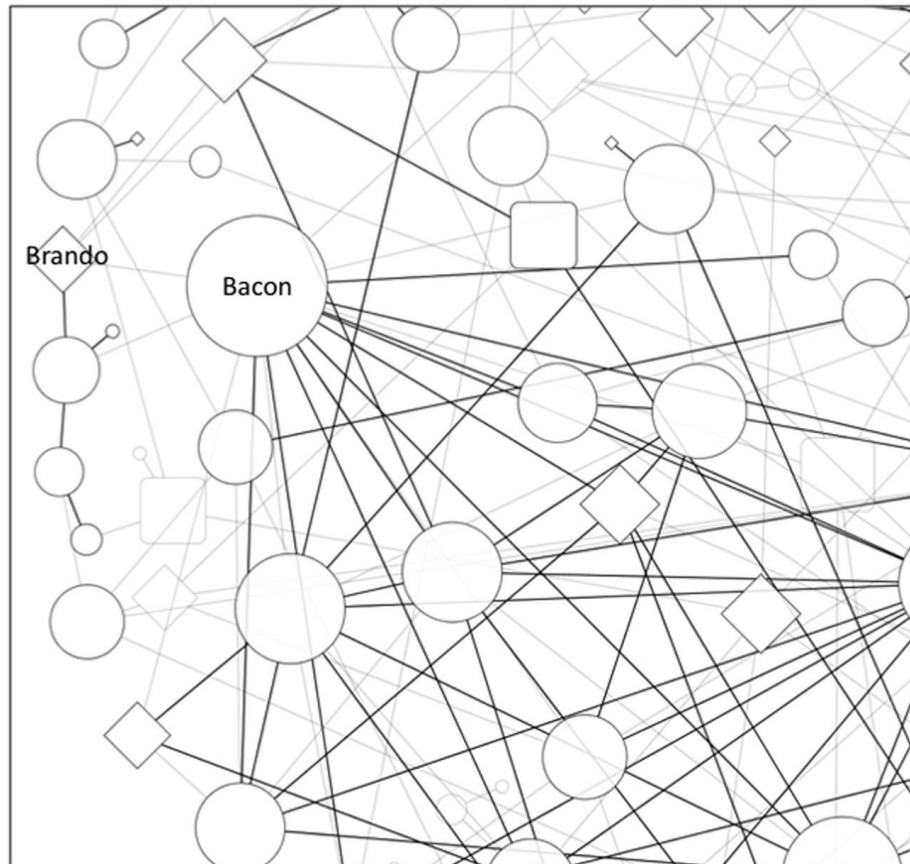
Genetic interactions:

... 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>



RESEARCH ARTICLE

Open Access

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

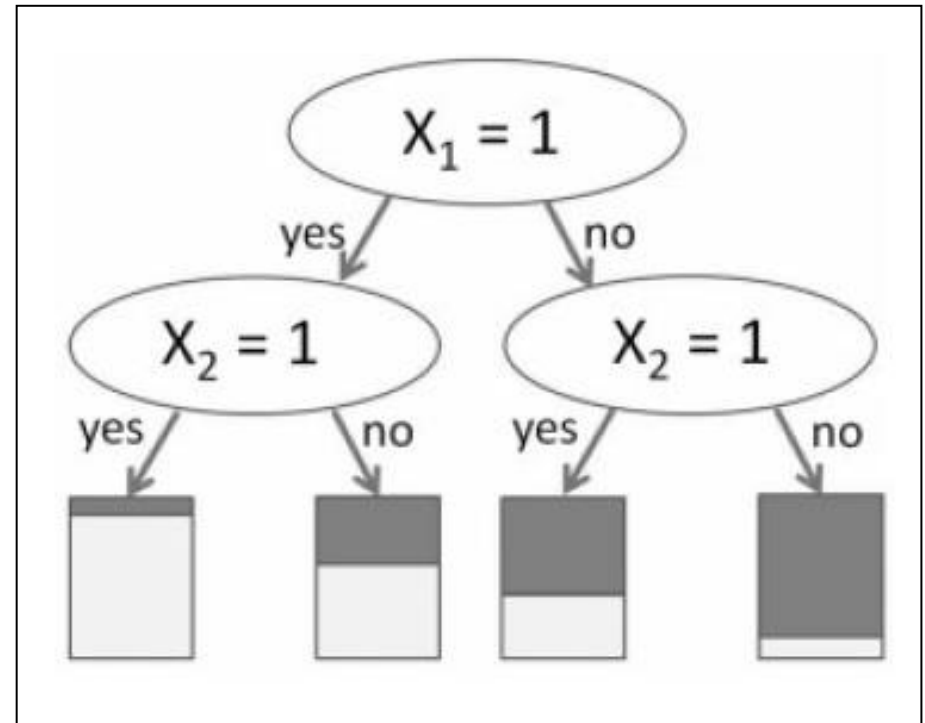
Stacey J Winham^{1*}, Colin L Colby¹, Robert R Freimuth¹, Xin Wang¹, Mariza de Andrade¹, Marianne Huebner^{1,2} and Joanna M Biernacka^{1,3*}

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

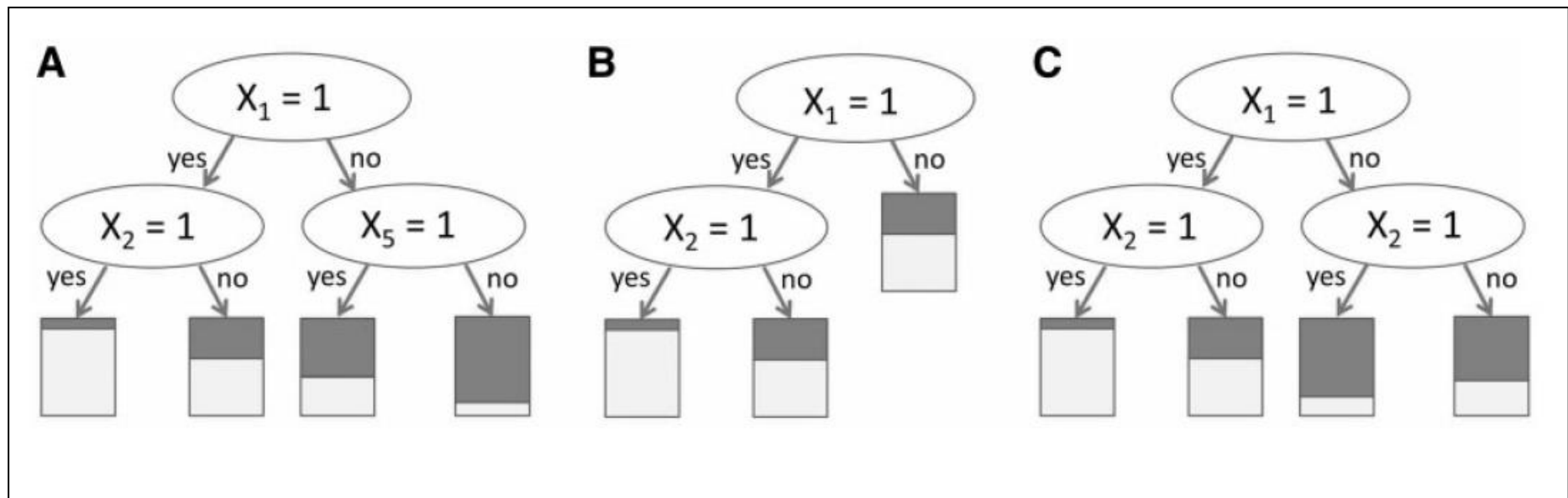
Be critical ...

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, \dots, 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

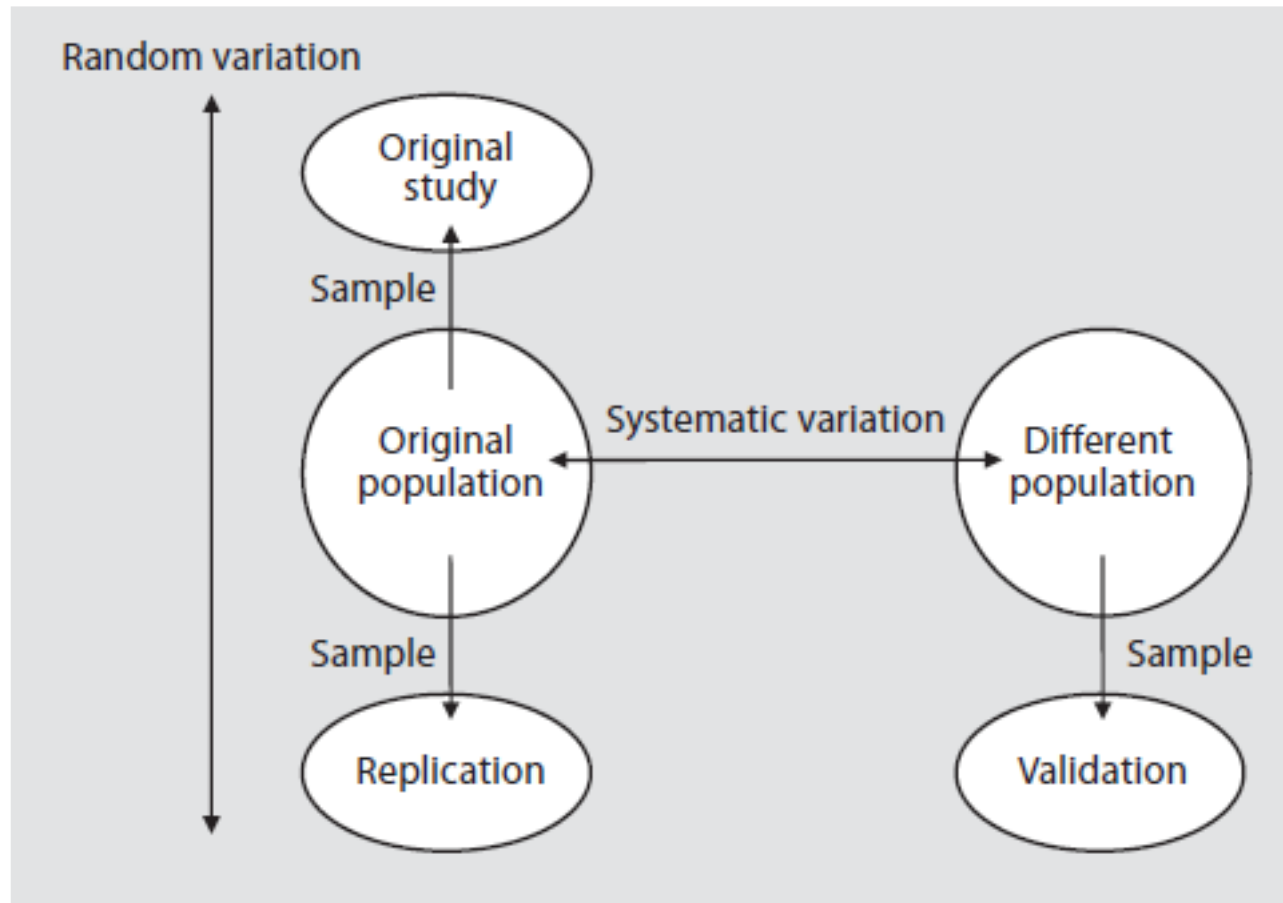


Be critical ...

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 be 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?

