Genetics and Bioinformatics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

From GWAS to Sequence Analyses

Part 1 When variants become rare

1. GWAS

2. Rare variants: promises and limitations

3. Frequency of sequence words: the stats perspective

Part 2 When effects become non-independent

Impact and interpretation

Biological vs statistical epistasis

(slide Doug Brutlag 2010)

1 GWAS

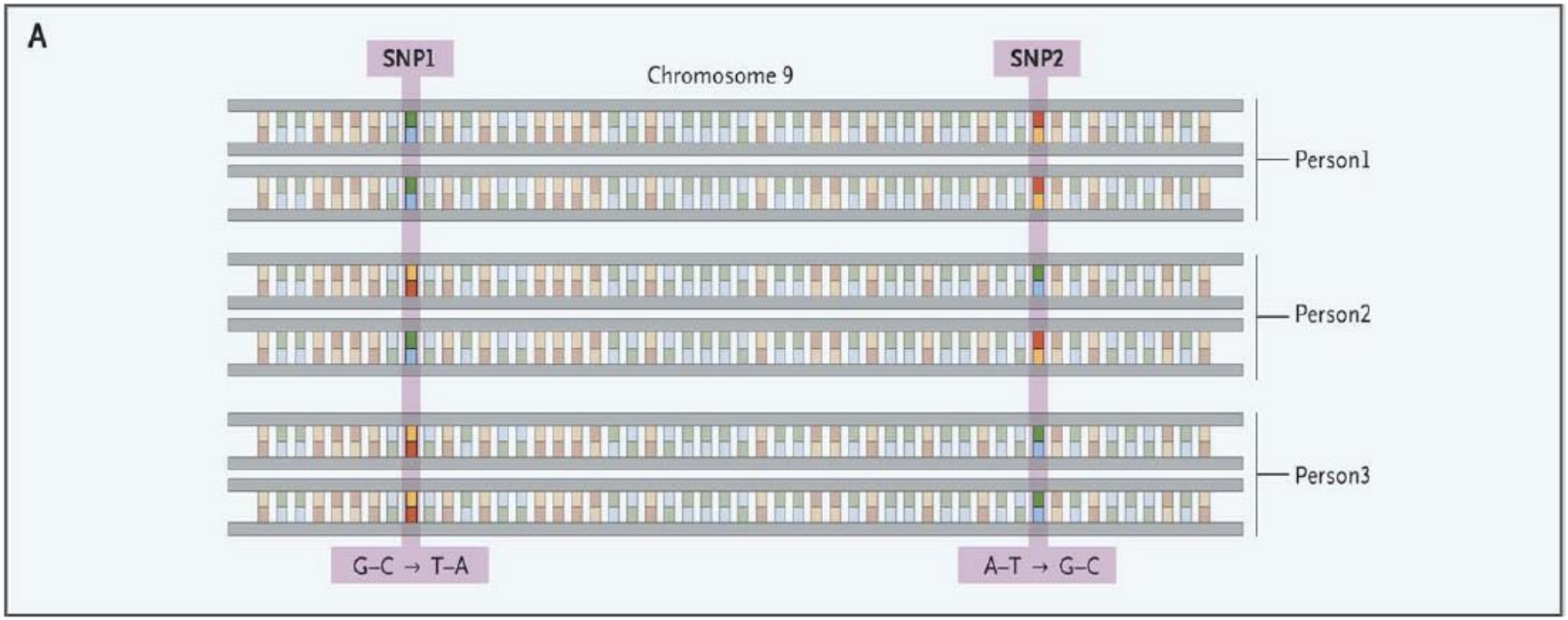
Definition (recap)

- 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.

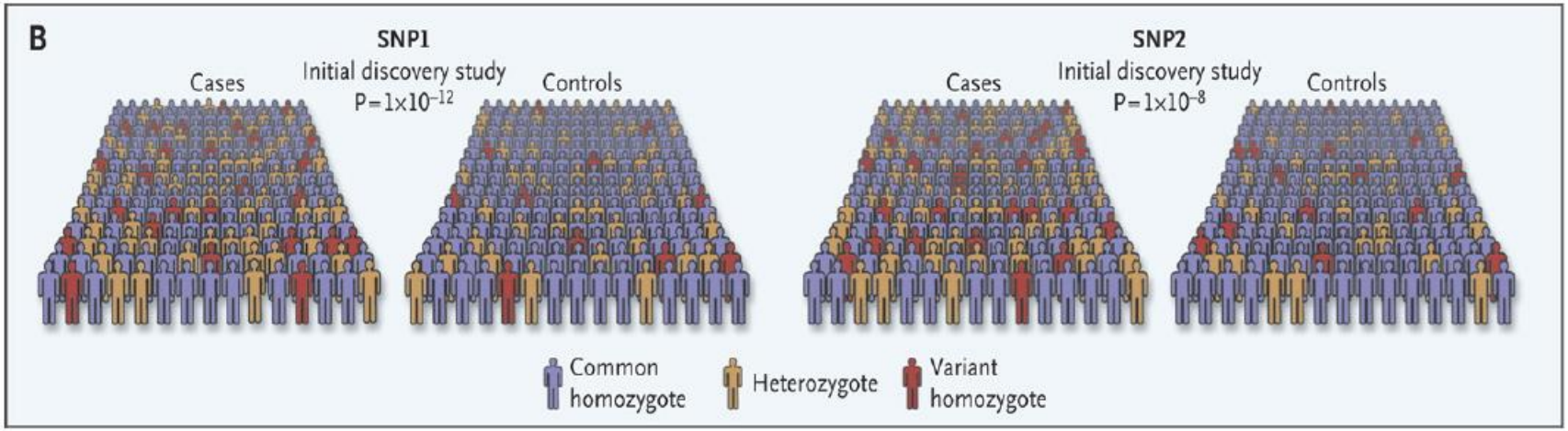
- A **trait** can be defined as a coded phenotype, a particular characteristic such as hair color, BMI, disease, gene expression intensity level, …

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)



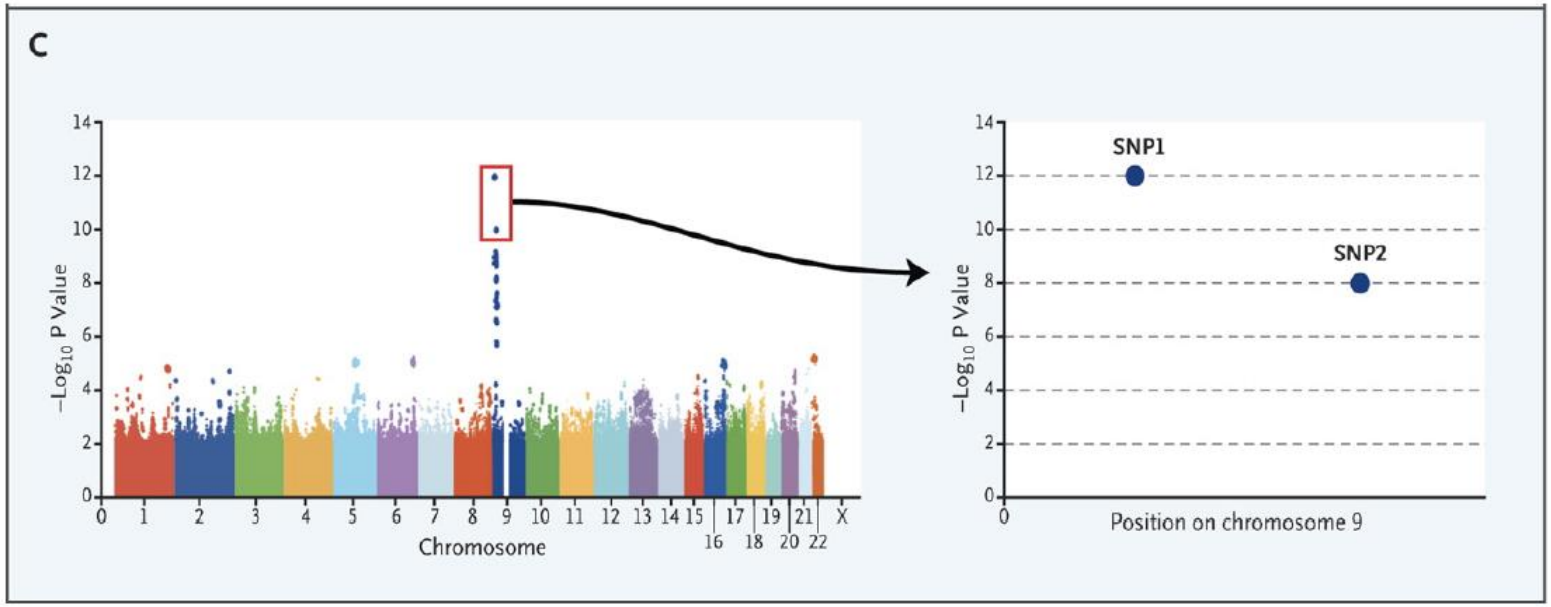
Genome-wide association studies in practice



- 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)

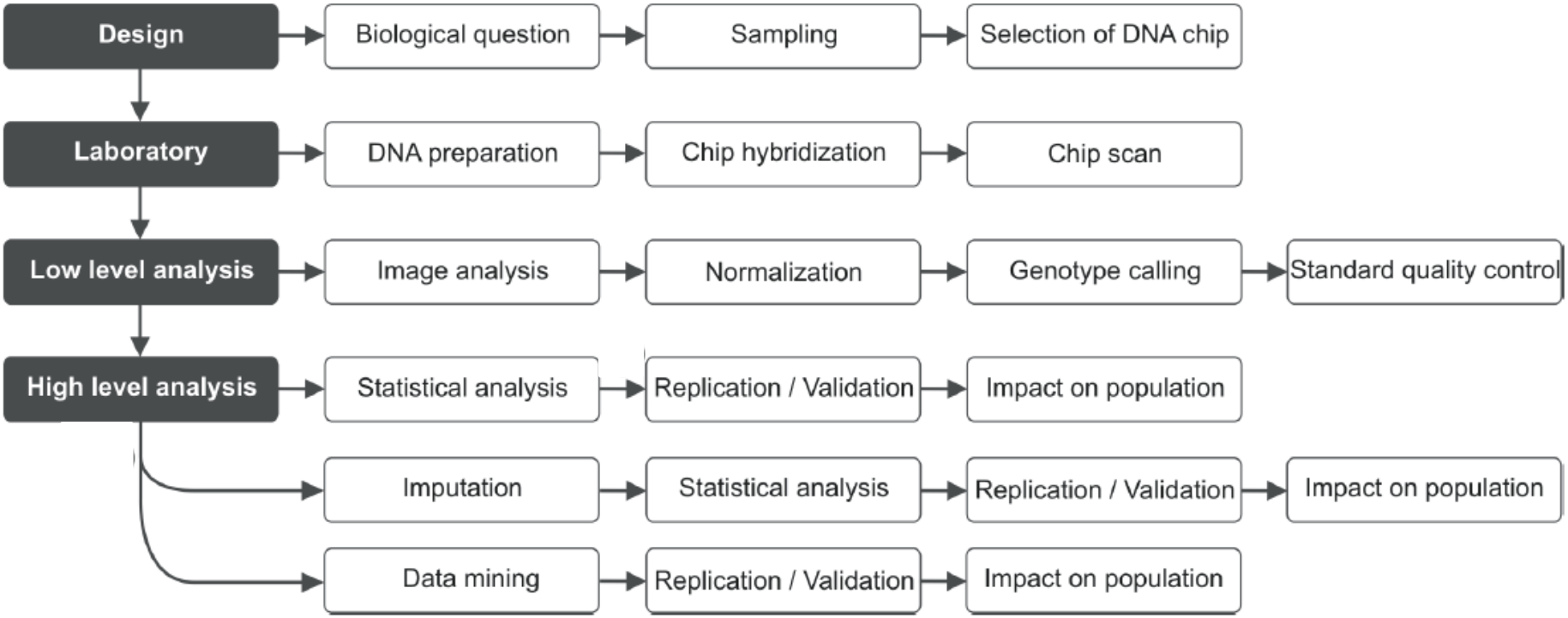
Genome-wide association studies in practice



- 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).

(Manolio 2010)

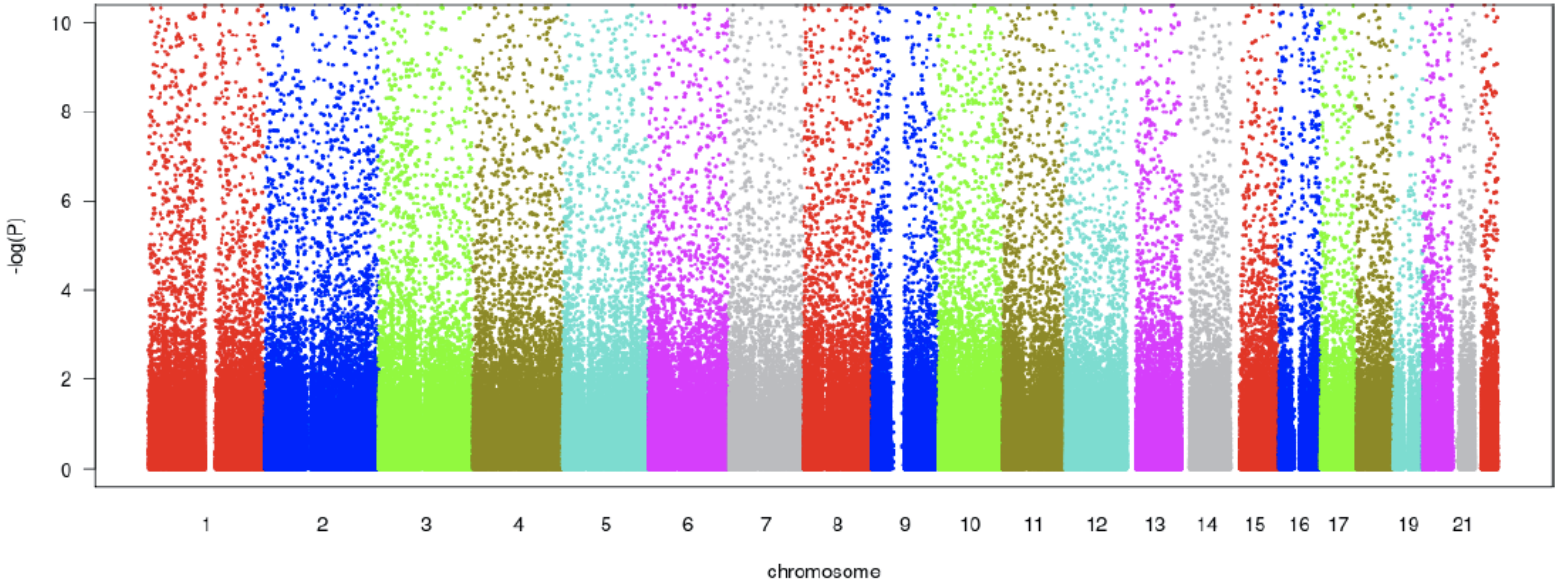
Detailed flow of a genome-wide association study



(Ziegler 2009)

Why is quality control (QC) important?

BEFORE QC → true signals are lost in false positive signals

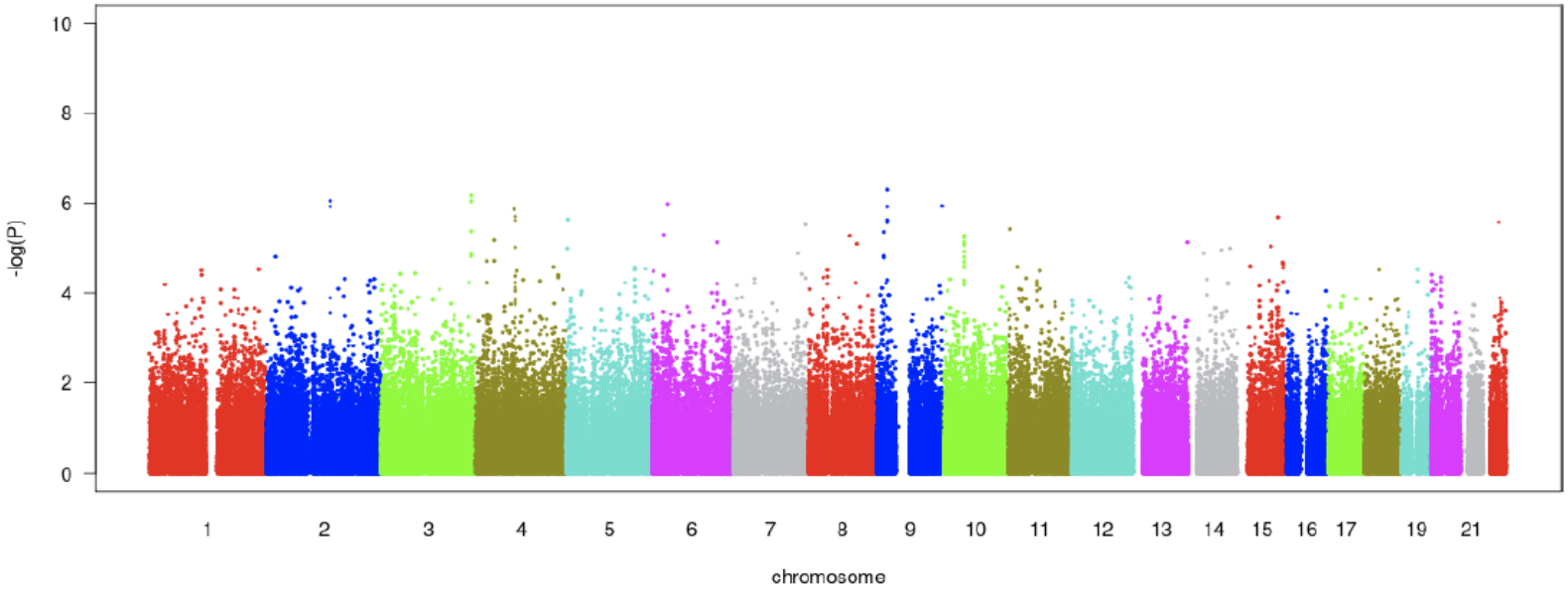


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

(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

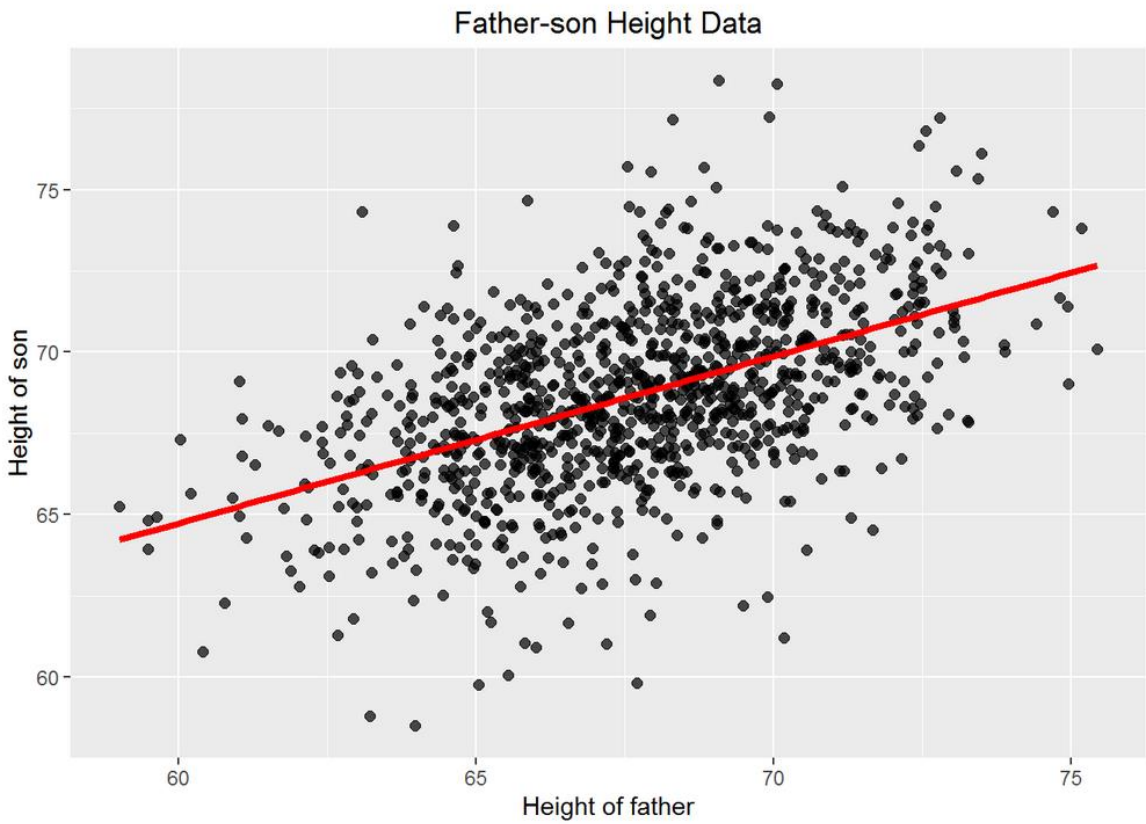
- Galton used the following equation to explain the phenomenon that sons of tall fathers tend to be tall but not as tall as their fathers while sons of short fathers tend to be short but not as short as their fathers:

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

This effect is called the **regression effect**.

The use of regression analysis

- **regression line** goes through (mean Y, mean X)



(https://rstudio-pubs-static.s3.amazonaws.com/204984_dd2112475db84af2a03260c4a4f830ac.html)

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, $X_1, \ldots, X_p$.
- 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.
  $\underline{b = (X^T X)^{-1} X^T y.}$
- Variance estimation for σ^2 (see later)
- Coefficient of Determination. R^2.
- Partial F test or t-test for $H_0 : \beta_j = 0$.

What is R-squared?

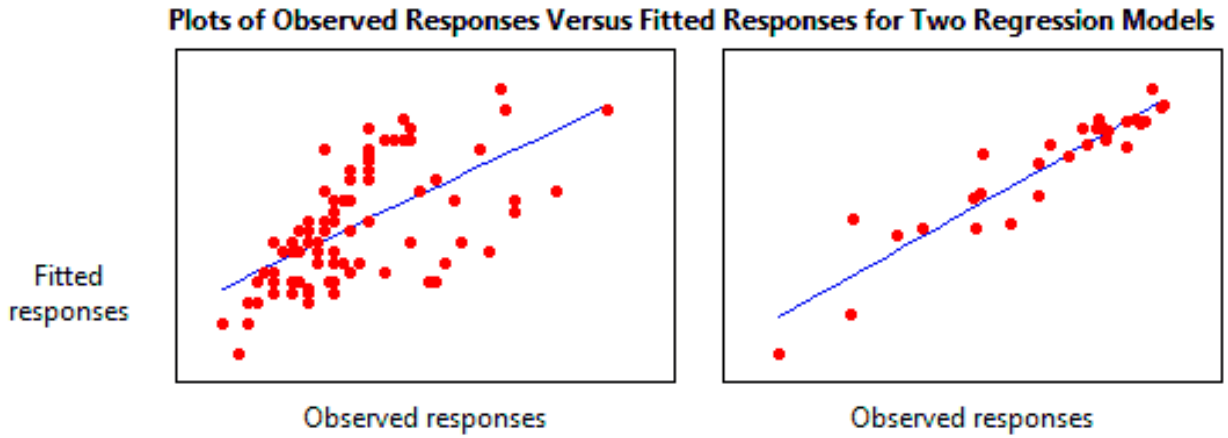
- R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the **coefficient of determination, or the coefficient of multiple determination for multiple regression**.
- The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model:

R-squared = Explained variation / Total variation

- R-squared is always between 0 and 100%:
  - 0% indicates that the model explains none of the variability of the response data around its mean.
  - 100% indicates that the model explains all the variability of the response data around its mean.

Graphical representation of R-squared

- Plotting fitted values by observed values graphically illustrates different R-squared values for regression models.

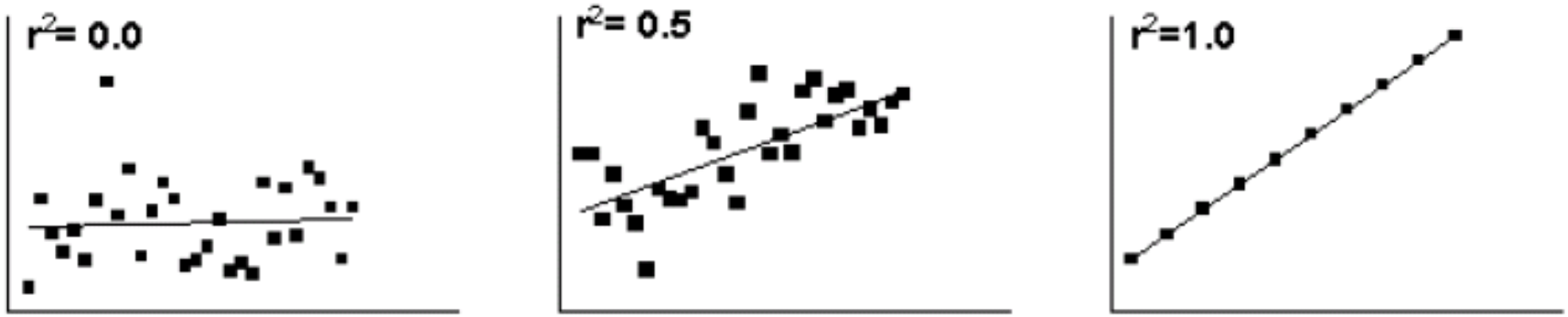


- The regression model on the left accounts for 38.0% of the variance while the one on the right accounts for 87.4%. The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line.

Coefficient of determination ~ squared correlation coefficient r^2

- 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):

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

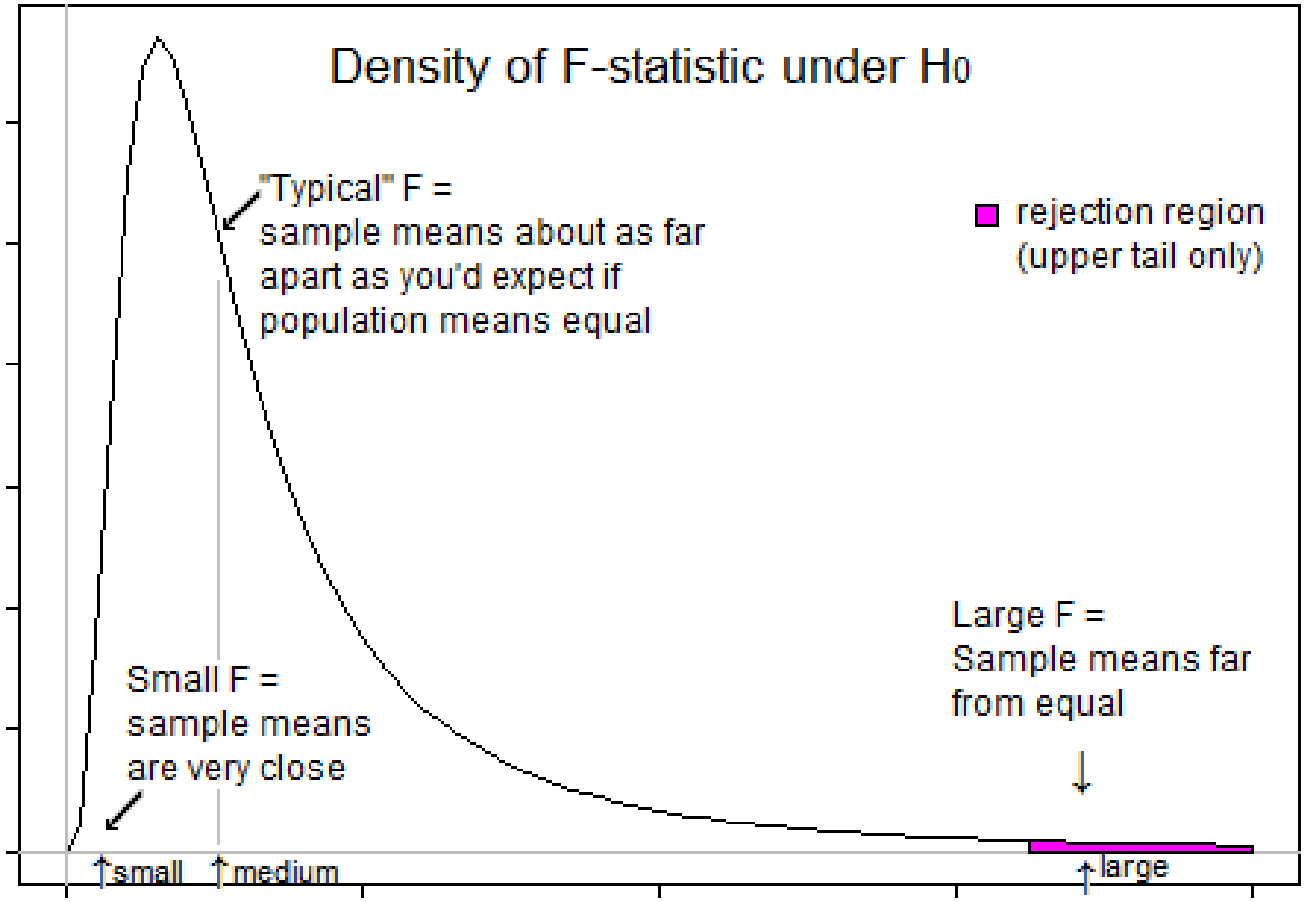
  which follows an F distribution when H_0 holds

- 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: rejection and non-rejection regions



Tests in GWAS using the regression framework

- **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 $\beta_1 \neq 0$

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

Why is the t-test more flexible?

Tests in GWAS using the regression framework

- **Example 2:**

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

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

How many dfs would the corresponding F-test have?

How many dfs would a corresponding $t(^2)$ test have?

Regression analysis in R

- Main functions
  - The basic syntax for doing regression in R is **lm()** 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)
- General syntax rules in R model fitting are given on the next slide.

Regression analysis in R

Syntax	Model	Comments
Y ~ A	$Y = \beta_o + \beta_1 A$	Straight-line with an implicit y-intercept
Y ~ -1 + A	$Y = \beta_1 A$	Straight-line with no y-intercept; that is, a fit forced through (0,0)
Y ~ A + I(A^2)	$Y = \beta_o + \beta_1 A + \beta_2 A^2$	Polynomial model; note that the identity function **I( )** allows terms in the model to include normal mathematical symbols.
Y ~ A + B	$Y = \beta_o + \beta_1 A + \beta_2 B$	A first-order model in A and B without interaction terms.
Y ~ A:B	$Y = \beta_o + \beta_1 AB$	A model containing only first-order interactions between A and B.
Y ~ A*B	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 AB$	A full first-order model with a term; an equivalent code is Y ~ A + B + A:B.
Y ~ (A + B + C)^2	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 AC$	A model including all first-order effects and interactions up to the nth order, where n is given by **( )^n**. An equivalent code in this case is Y ~ A*B*C – A:B:C.

Coding matters

	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**

Coding matters

Use of `lm()` in genetics

For a continuous outcome,

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

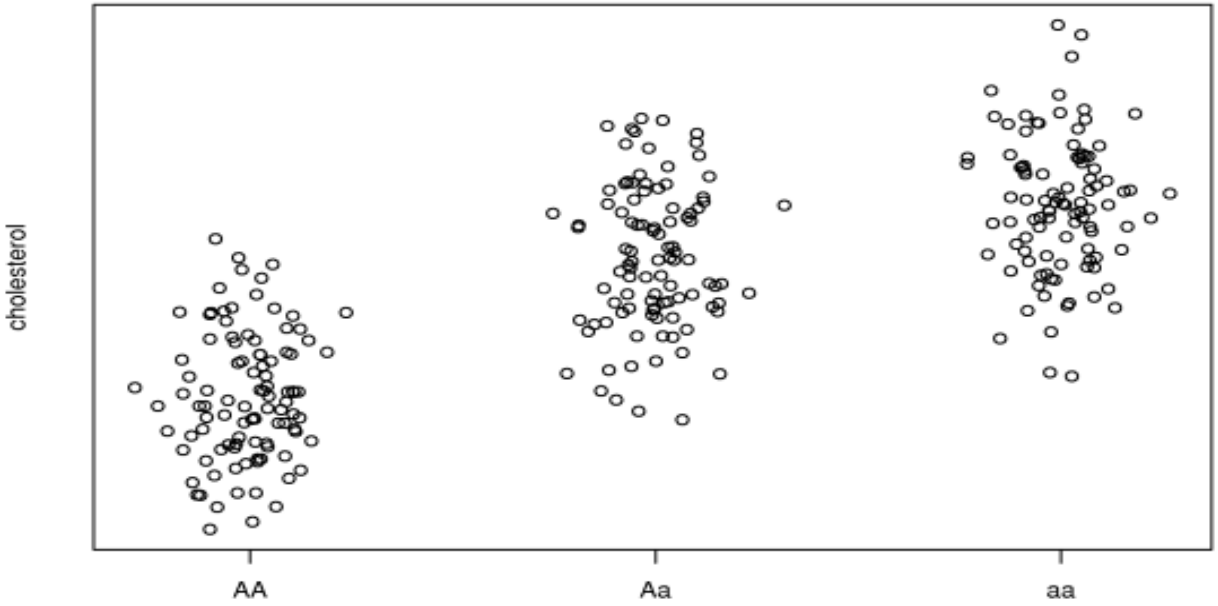
and predictor

- `weights` – for advanced analyses

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"

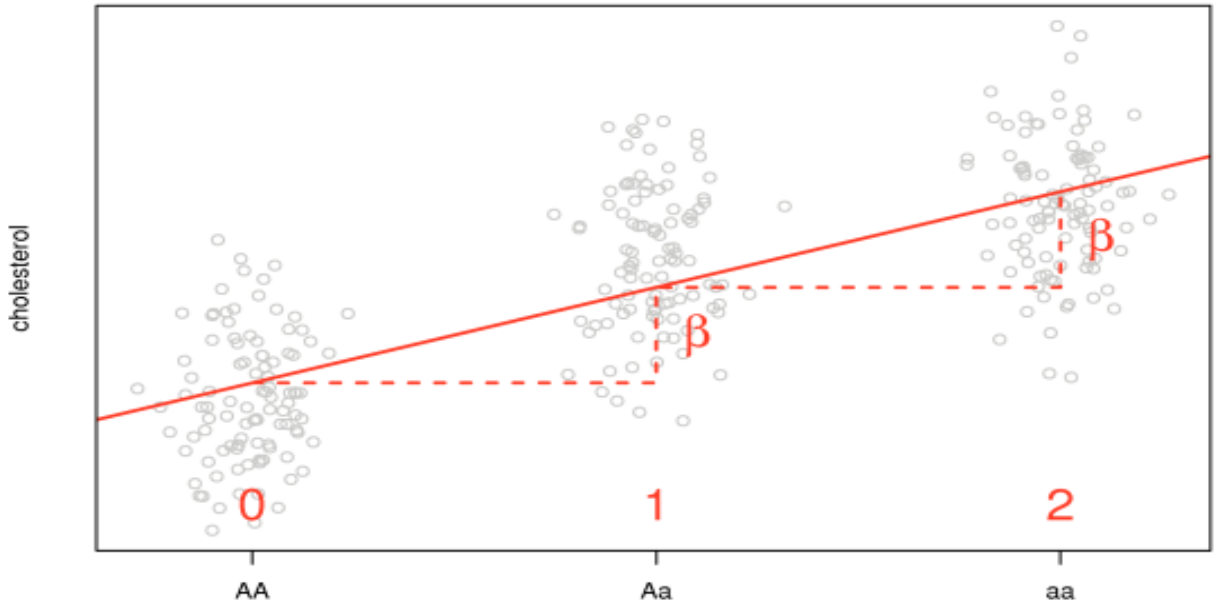
Use of `lm()` in genetics

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



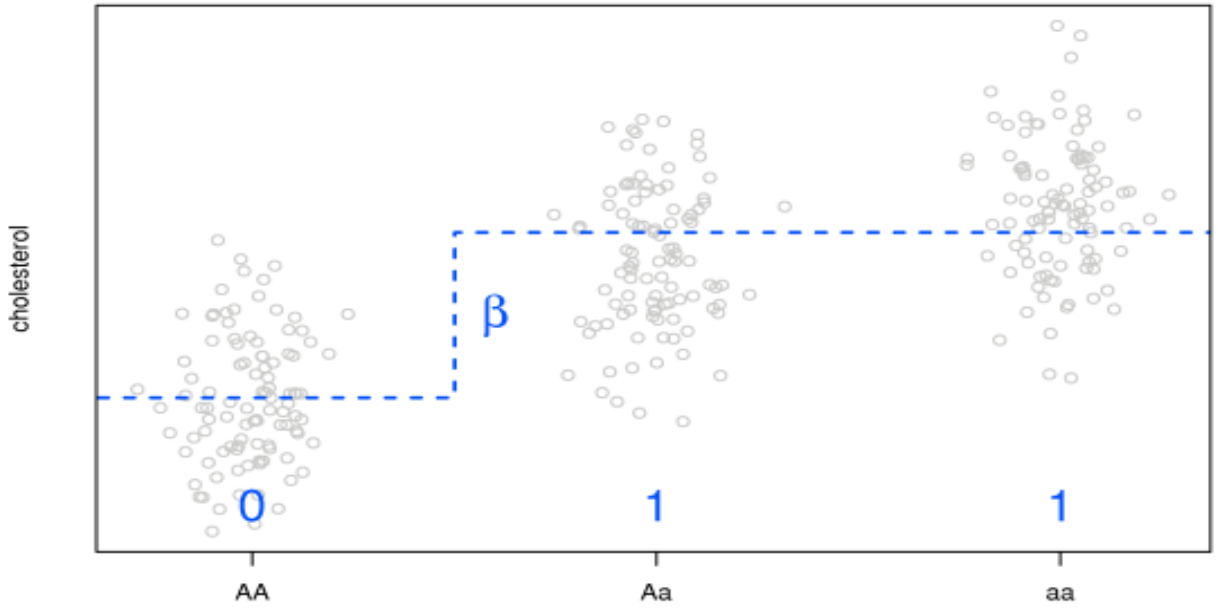
Use of `lm()` in genetics

Additive model (the most commonly used)



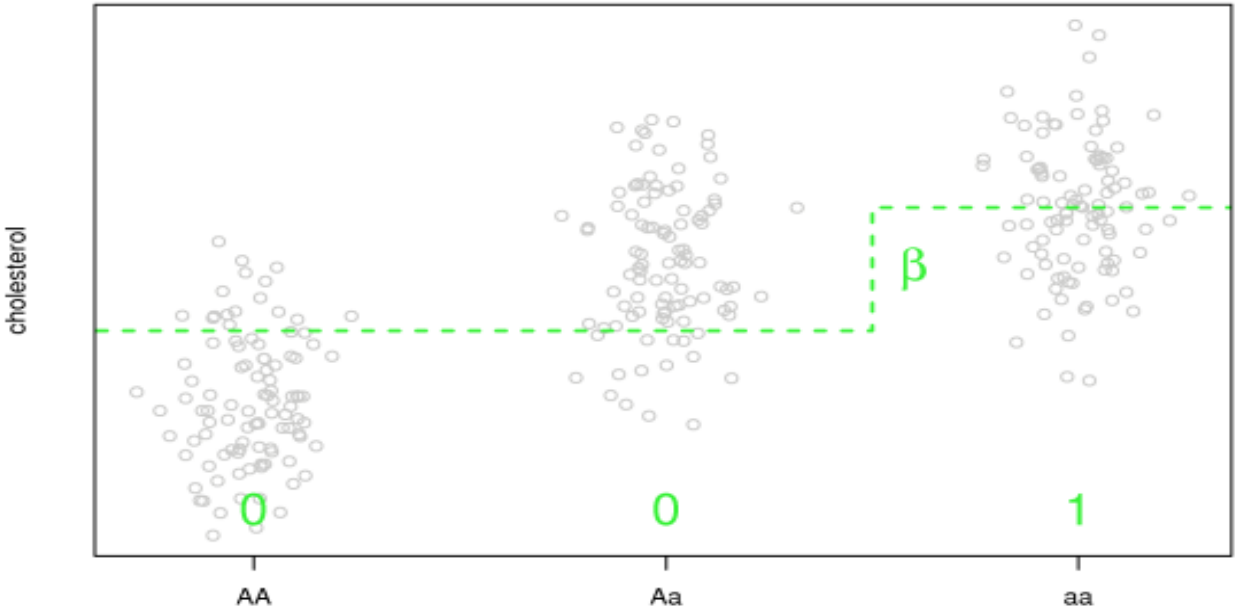
Use of `lm()` in genetics

Dominant model (best fit to this data)



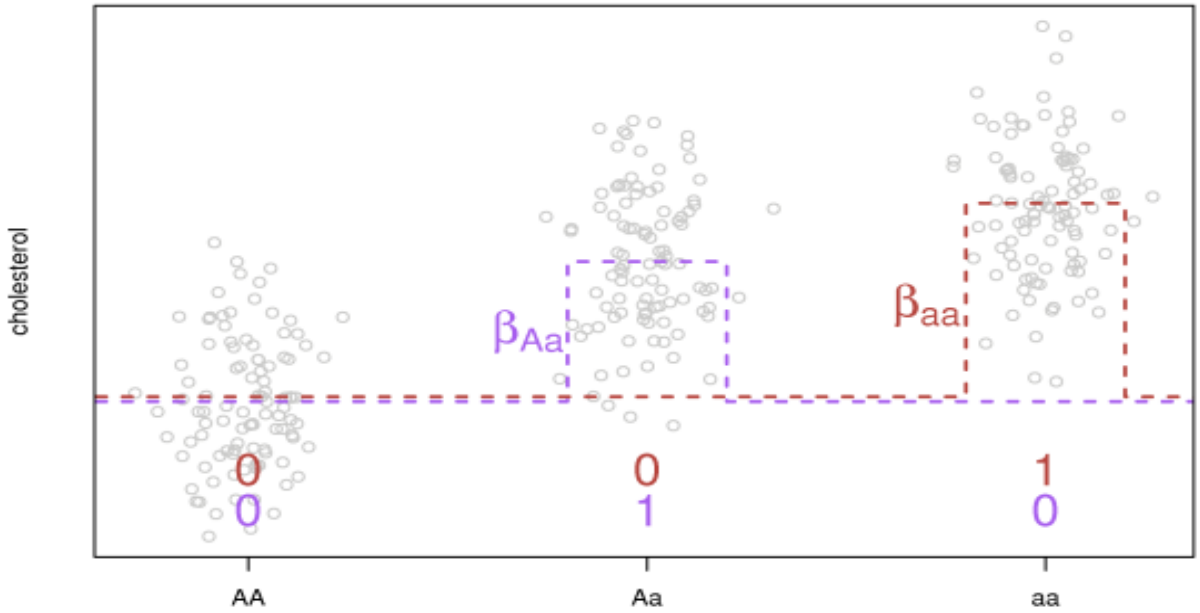
Use of `lm()` in genetics

Recessive model (least stable for rare aa)



Use of `lm()` in genetics

2 parameter model (robust but can be overkill)



lm(): Estimates, Intervals, p-values

lm() produces **point estimates** for your model;

```
> n.minor <- (g=="Aa") + 2*(g=="aa")
> my.lm <- lm( cholesterol ~ n.minor )
> my.lm
Call:
lm(formula = cholesterol ~ n.minor)
Coefficients:
(Intercept)      n.minor
     0.2104       0.9507
```

– also available via my.lm$coefficients.

The coefficients in the output tell you the **additive increase** in outcome associated with a **one-unit** difference in the genetic predictor.

The coefficient for n.minor is in units of cholesterol

lm(): Estimates, Intervals, p-values

You will also want **confidence intervals**;

```
> confint.default(my.lm)
                 2.5 %    97.5 %
(Intercept) 0.08391672 0.3368275
n.minor     0.85279147 1.0486953
```

Remember to **round these numbers** to an appropriate number of significant figures! (2 or 3 is usually enough)

We are **seldom** interested in the `Intercept`

lm(): Estimates, Intervals, p-values

Two-sided **p-values** are also available;

```
> summary(my.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.21037    0.06426   3.274  0.00119 **
n.minor      0.95074    0.04977  19.101  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this data, we have **strong evidence** of an **additive effect** of the minor allele on cholesterol

`summary(my.lm)` gives **many** other details – ignore for now

Confidence intervals are just `Estimate` $\pm$ 2$\times$`Std.Error`

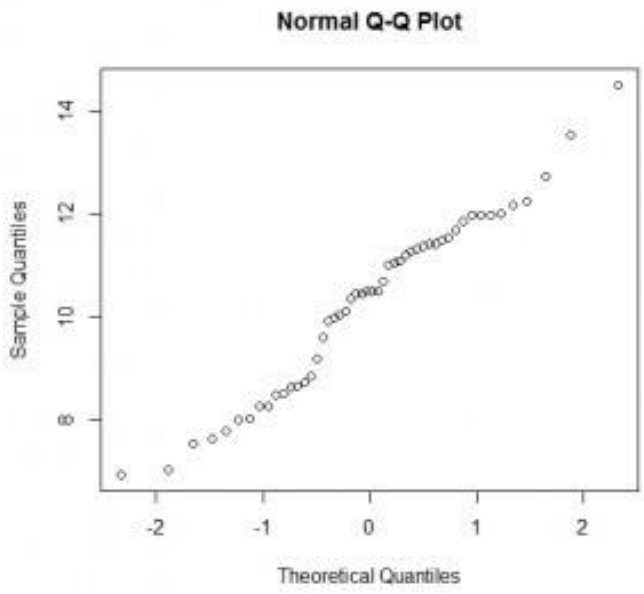
Model diagnostics are model-dependent …

- 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)

QQ plots for model diagnostics

- A Q-Q plot is a scatterplot created by plotting **two sets of quantiles** against one another.
- If both sets of quantiles come from the same distribution, we should see the points forming a line that's roughly straight.
- Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



QQ plots for model diagnostics

- Quantiles are points in your data below which a certain proportion of your data fall.

  What is the 0.5 quantile for normally distributed data?

- Here we generate a random sample of size 200 from a normal distribution and find the quantiles for 0.01 to 0.99 using the quantile function:

```
quantile(rnorm(200),probs = seq(0.01,0.99,0.01))
```

- Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution.
  The number of quantiles is selected to match the size of your sample data.

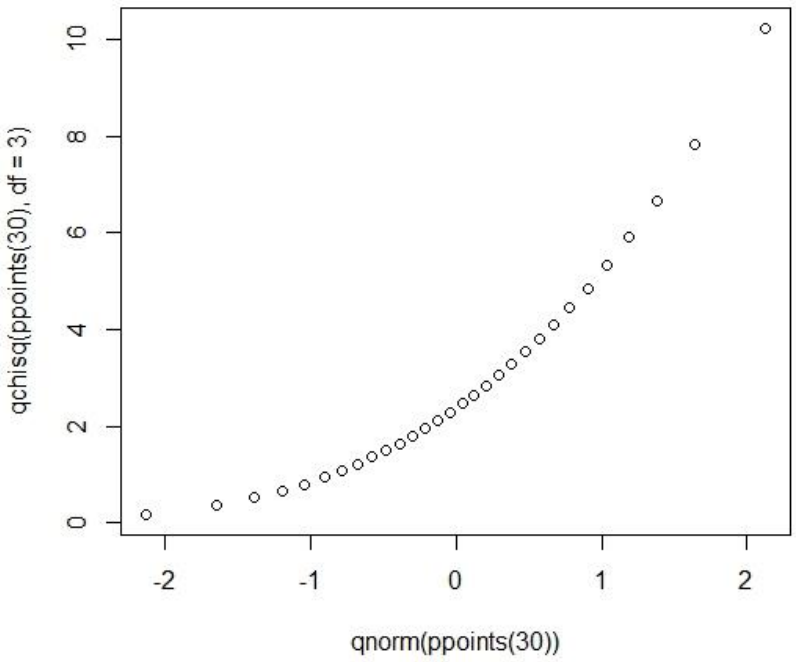
  The quantile function in R offers 9 different quantile algorithms!

  See help(quantile)

Examples of QQ plots: no straight line

- QQ plot of a distribution that's skewed right; a Chi-square distribution with 3 degrees of freedom against a Normal distribution

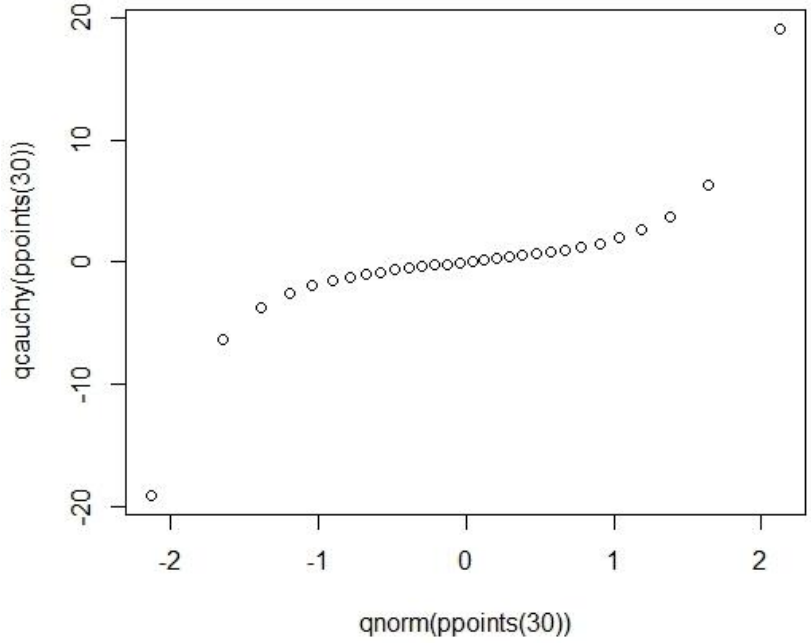
qqplot(qnorm(ppoints(30)), qchisq(ppoints(30),df=3))



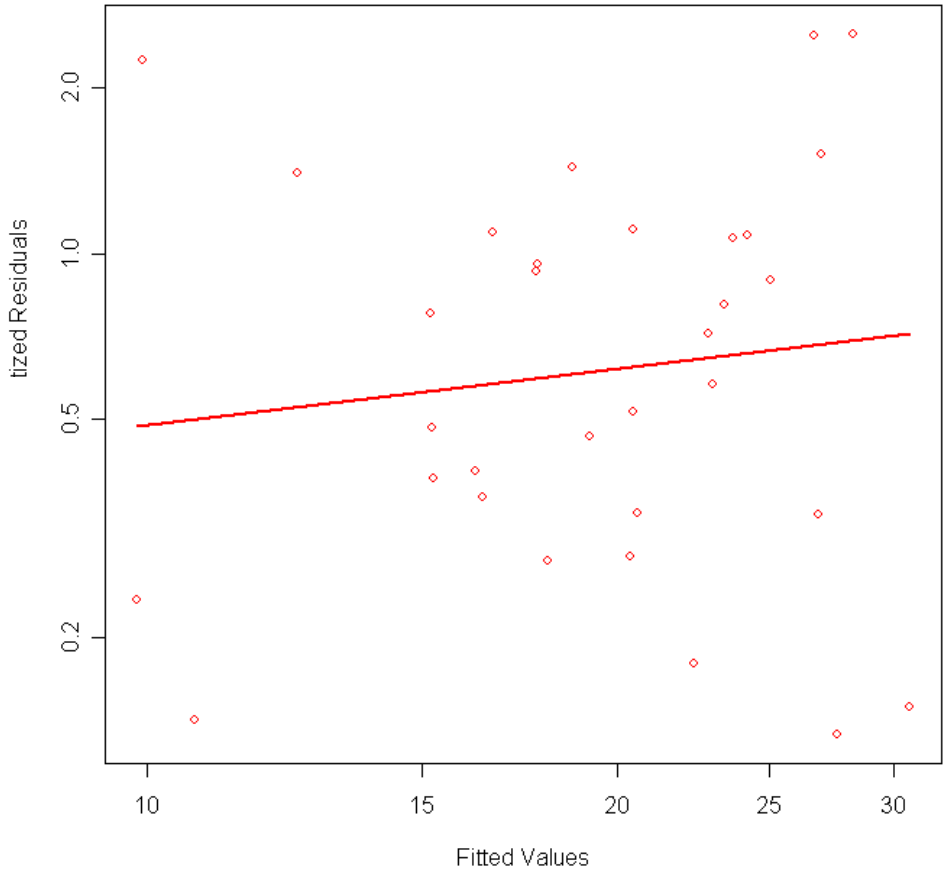
Examples of QQ plots: no straight line

- QQ plot of a distribution with heavy tails (vs Normal)

qqplot(qnorm(ppoints(30)), qcauchy(ppoints(30)))



Residual plots for model diagnostics



Logistic regression (dichotomous traits; cases and controls)

In linear regression one equates

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

In logistic regression one equates

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

- 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; conditional expectations)

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

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

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

$$\Downarrow$$

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

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

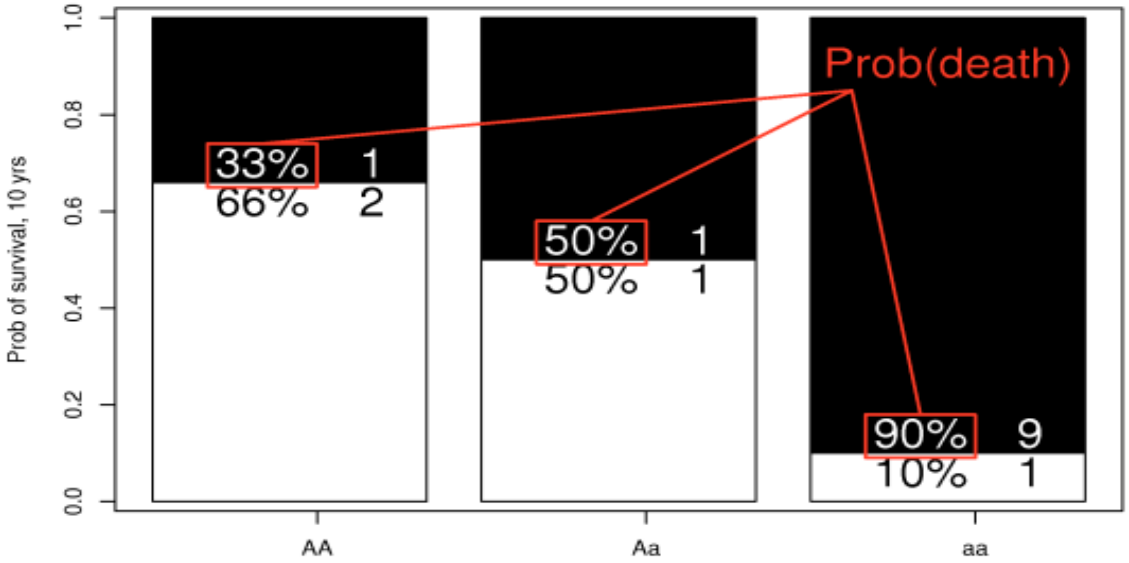
Log(Odds|$X_{1\,==0}$) = β_0

Log(OR) = β_1

Coding matters

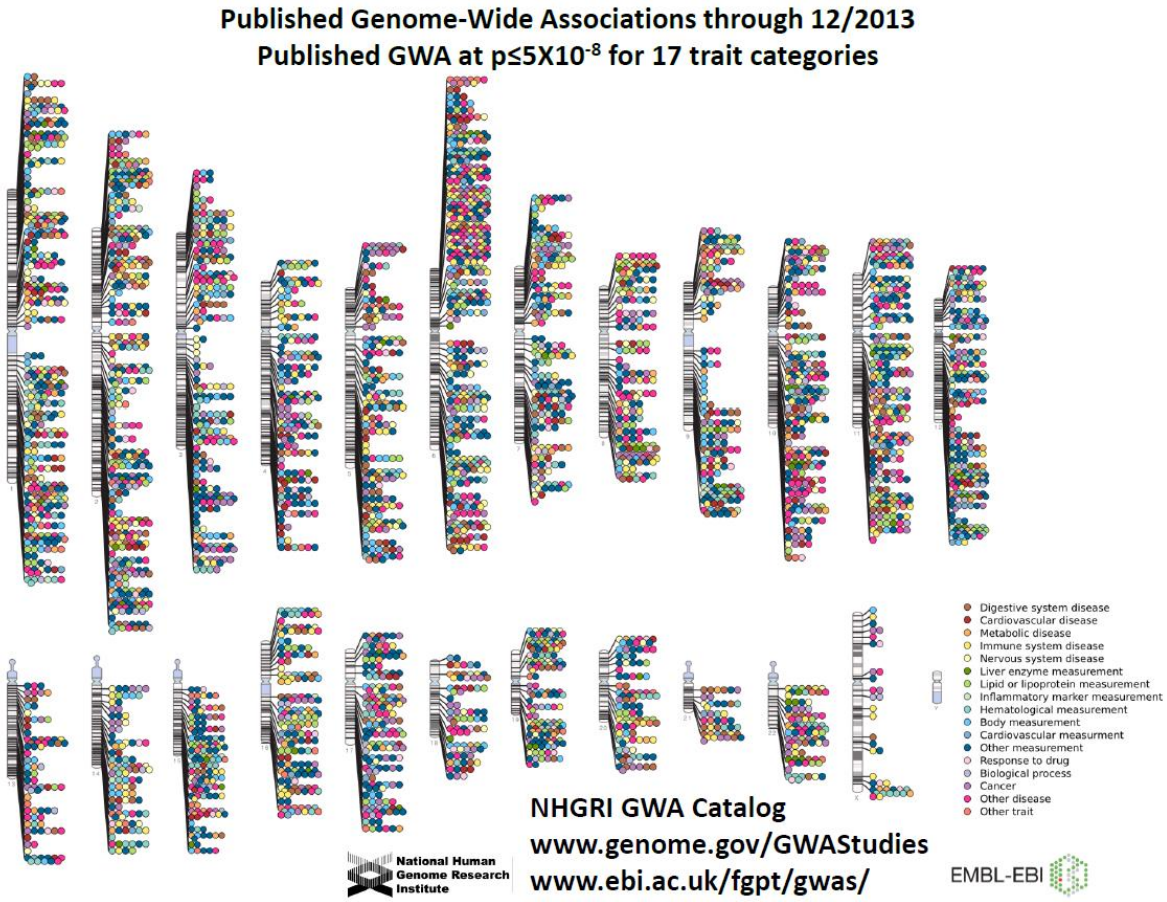
Use of `glm()` in genetics

Odds are a [gambling-friendly] measure of chance;



2 When the need emerges to look at rare variants

DNA sequence analyses: motivation



Sequencing projects

- Few years later, as sequencing techniques became more advanced, more accurate, and less expensive, the **1000 Human Genome Project** was launched (January 2008).

  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.

- The **International HapMap Project** (short for “haplotype map”) aims to identify common genetic variations among people, making use of data from six different countries.
- 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.

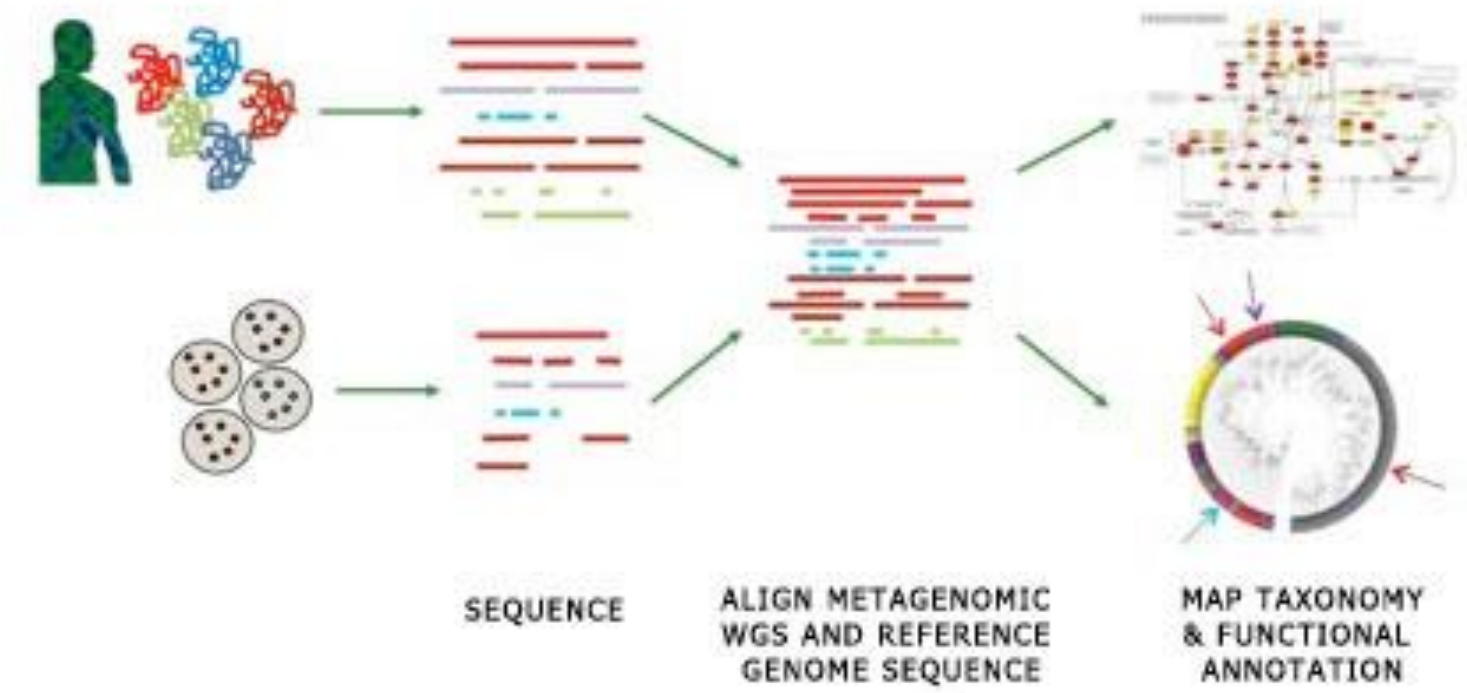
- The goal of the **1000 Fungal Genome Project** (http://1000.fungalgenomes.org) is to explore all areas of fungal biology.

- 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.

Why do we need reference sequences?

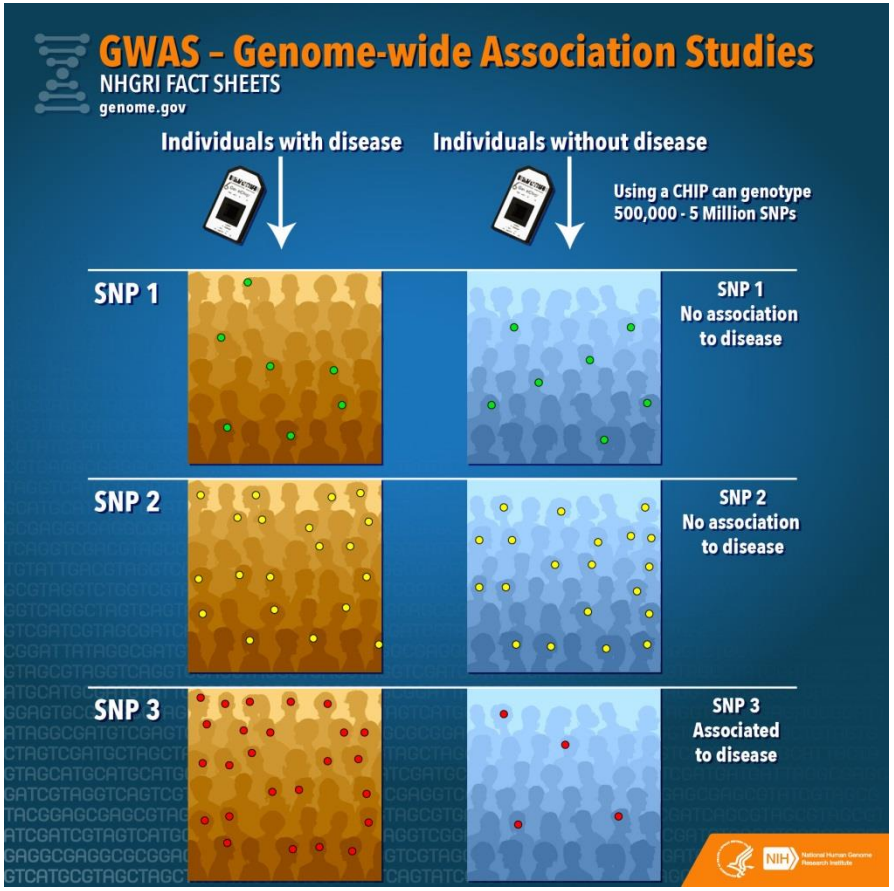
- Within the human body, it is estimated that there are 10x as many microbial cells as human cells.
- Our microbial partners carry out a number of metabolic reactions that are not encoded in the human genome and are necessary for human health (→ human genome = human genes + microbial genes).
- The majority 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/)

Why reference sequences?



(https://www.genome.gov/images/content/gwas_infographic.jpg)

Which reference sequence?

www.hgvs.org/mutnomen/refseq.html

Search Most Visited MyULg lpvs Montefiore http://www.ulg.ac.be/

HGVS HUMAN GENOME VARIATION SOCIETY

A reference sequence - discussions and FAQs

Last modified September 11, 2015

Since references to WWW-sites are not yet acknowledged as citations, please mention den Dunnen JT and Antonarakis SE (2000). Hum.Mutat. 15:7-12 when referring to these pages.

Contents

(http://www.hgvs.org/mutnomen/refseq.html)

Which reference sequence?

Practical problems genomic reference sequence

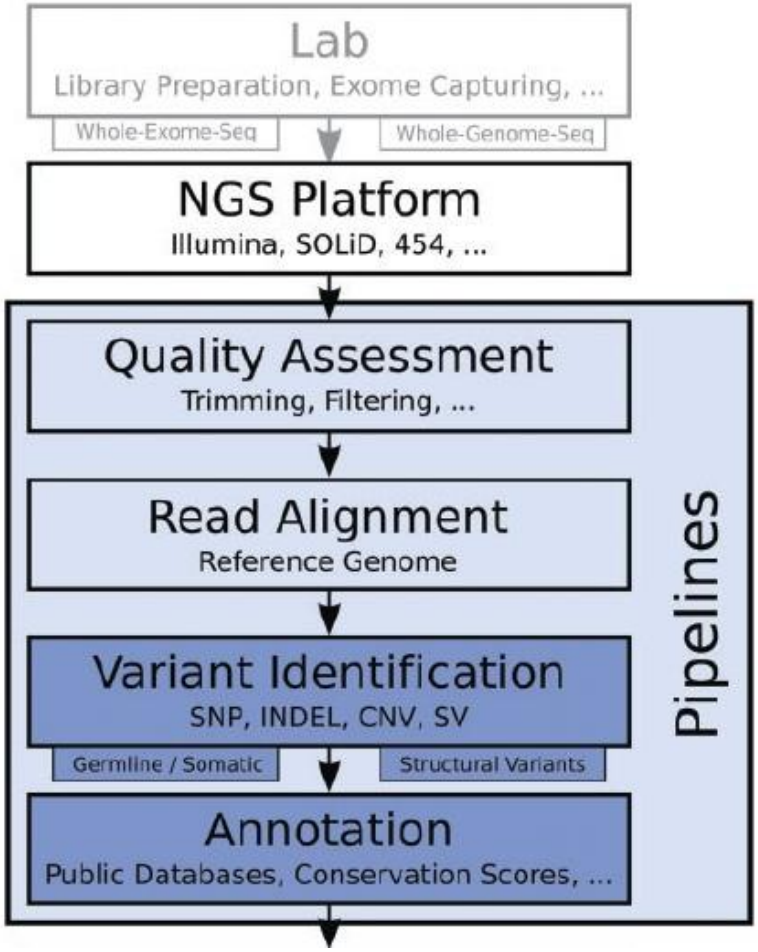
- a *gene can be very large* (over 2.0 Mb) - this makes nucleotide numbering based on a genomic reference sequence rather impractical (e.g. g.1567234_1567235insTG). Furthermore, genomic reference sequences based on GenBank NT_ files become increasingly long (e.g. the CFTR gene in NT_007933.15, >77 Mb) and consequently loose 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*).

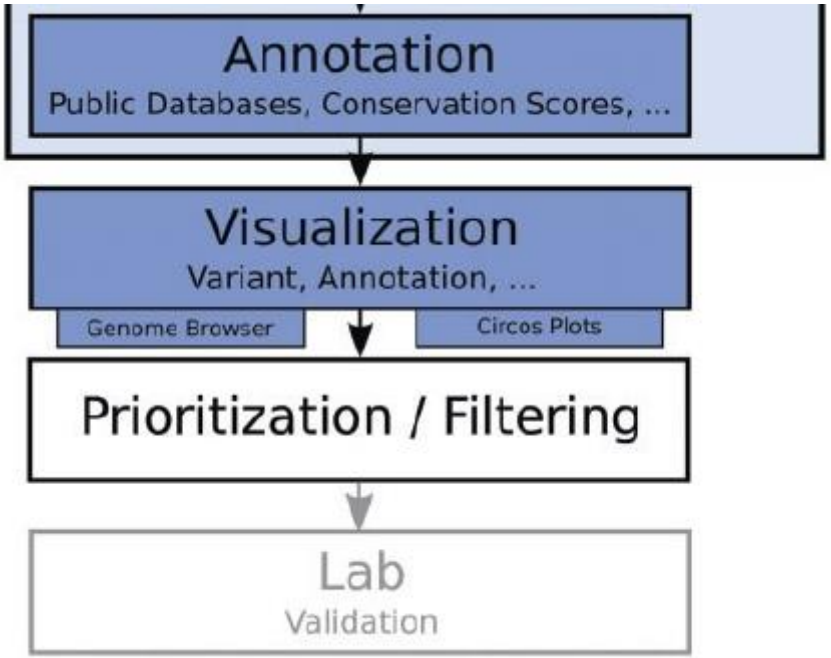
Practical problems coding DNA reference sequence

- 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)

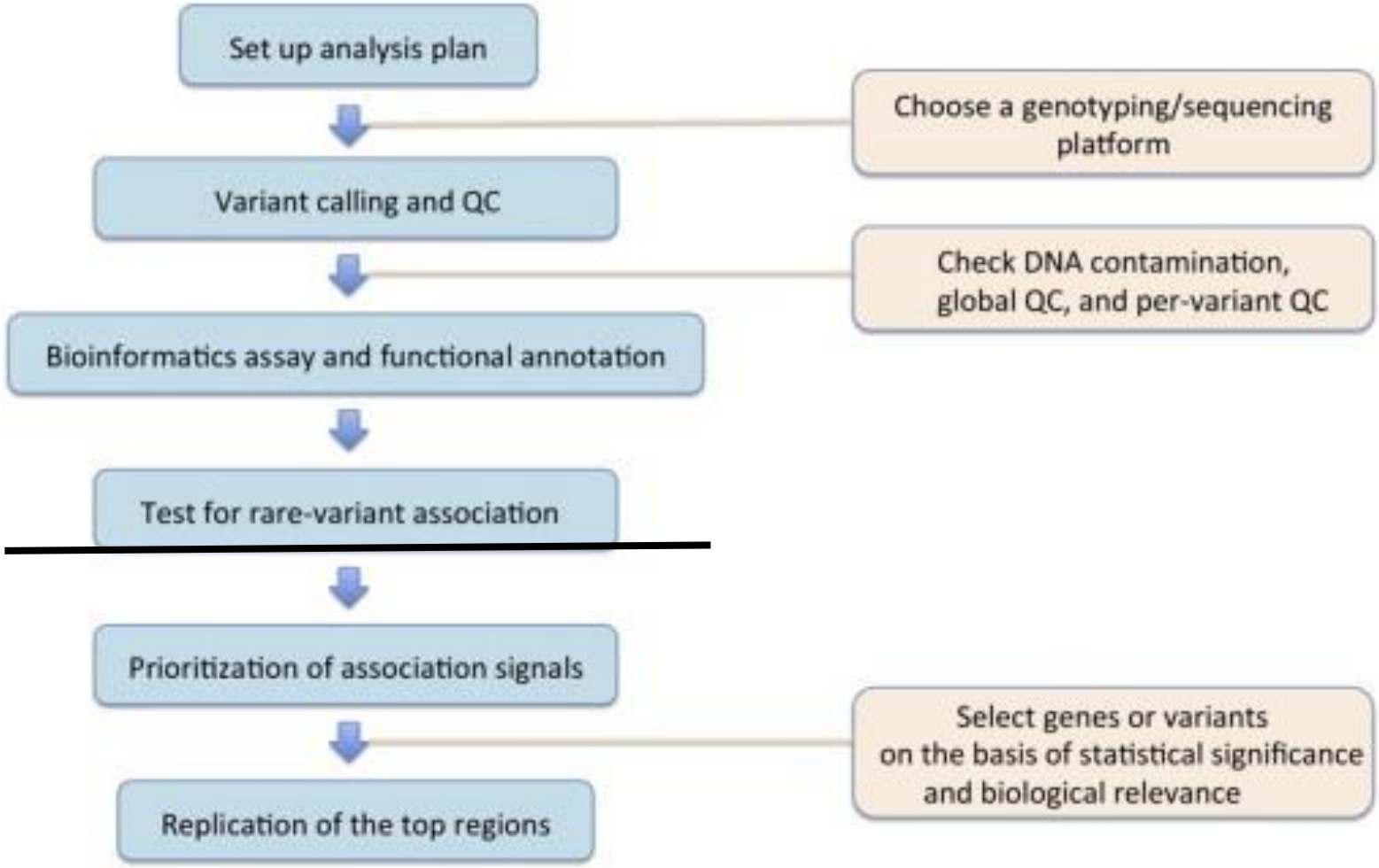
Common workflow for whole-exome and whole genome sequencing





(Pabinger et al. 2013)

Work flow genome-wide association study with sequence data



(Lee et al. 2014)

Impact of rare variants arising from sequence data on inference

- A variant – genetic association test implies filling in the table below and performing 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** (comparing two variables in a contingency table to see if they are related):
  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

  $X^2 = \sum_{all\ cells\ i} \frac{(O_i - E_i)^2}{E_!}$ (contrasting Observed minus Expected)

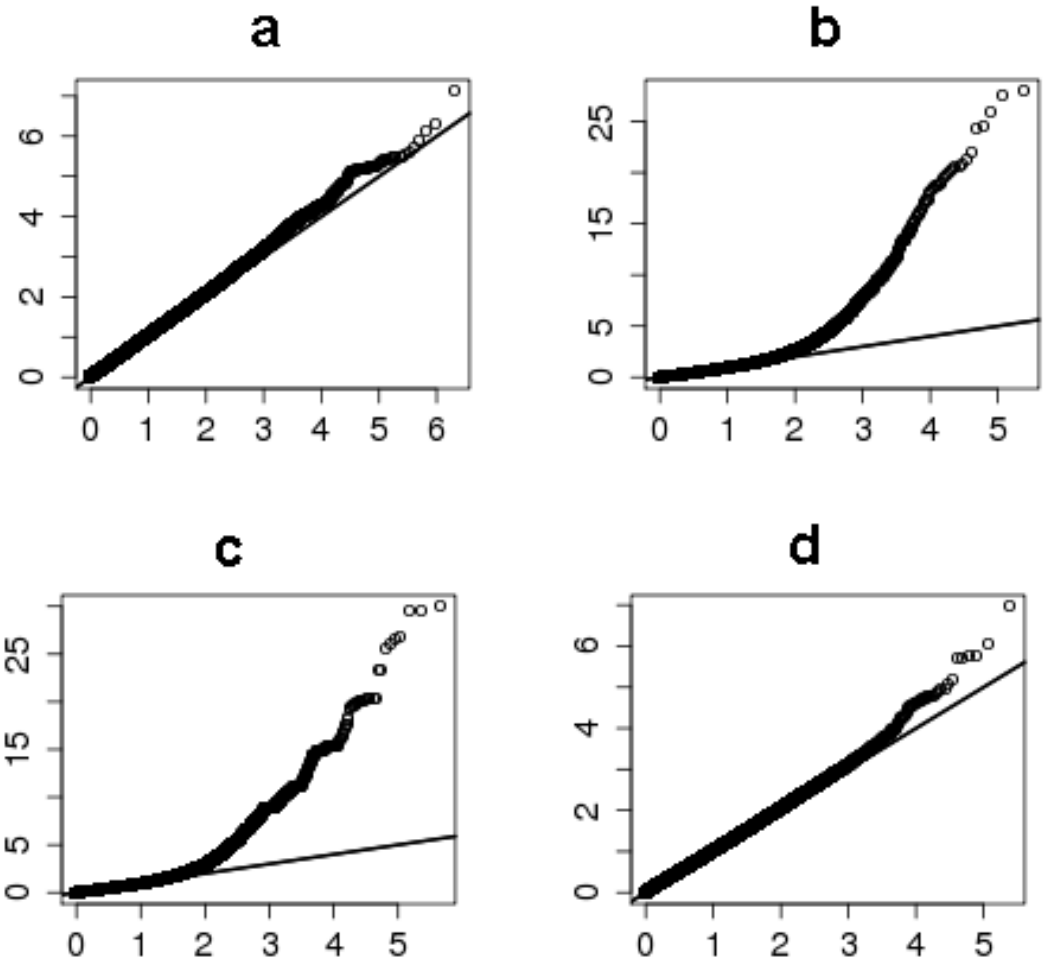
- **In a regression framework:**
  The minimum number of observations per independent variable should be 10, using a guideline provided by Hosmer and Lemeshow (Applied Logistic Regression, one of the main resources for Logistic Regression)

Increased false positive rates

Q-Q plots from GWAS data, unpublished

N=~2500

MAF>0.03



N=~2500

MAF<0.03

N=~2500

MAF<0.03

Permuted

N=50000

MAF<0.03

Bootstrapped

Remediation: do not look at a single variant at a time, but collapse

- Rationale for aggregation 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”.
- Remedy = aggregate / pool variants
  - Requires specification of a so-called “region of interest” (ROI)
  - A ROI 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)

Some problems with burden tests

- Problem 1: When high linkage disequilibrium (LD) [allelic non-independence] exists in the “region”, combined counts may be artificially elevated
- Problem 2: Assumes that 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

	Description	Methods	Advantage	Disadvantage	Software Packages[a]
Burden tests	collapse rare variants into genetic scores	ARIEL test,[50] CAST,[51] CMC method,[52] MZ test,[53] WSS[54]	are powerful when a large proportion of variants are causal and effects are in the same direction	lose power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	EPACTS, GRANVIL, PLINK/SEQ, Rvtests, SCORE-Seq, SKAT, VAT
Adaptive burden tests	use data-adaptive weights or thresholds	aSum,[55] Step-up,[56] EREC test,[57] VT,[58] KBAC method,[59] RBT[60]	are more robust than burden tests using fixed weights or thresholds; some tests can improve result interpretation	are often computationally intensive; VT requires the same assumptions as burden tests	EPACTS, KBAC, PLINK/SEQ, Rvtests, SCORE-Seq, VAT
Variance-component tests	test variance of genetic effects	SKAT,[61] SSU test,[62] C-alpha test[63]	are powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	are less powerful than burden tests when most variants are causal and effects are in the same direction	EPACTS, PLINK/SEQ, SCORE-Seq, SKAT, VAT

(Lee et al. 2014)

Other tests

Combined tests	combine burden and variance-component tests	SKAT-O,[64] Fisher method,[65] MiST[66]	are more robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants	can be slightly less powerful than burden or variance-component tests if their assumptions are largely held; some methods (e.g., the Fisher method) are computationally intensive	EPACTS, PLINK/SEQ, MiST, SKAT
EC test	exponentially combines score statistics	EC test[67]	is powerful when a very small proportion of variants are causal	is computationally intensive; is less powerful when a moderate or large proportion of variants are causal	no software is available yet

Abbreviations are as follows: ARIEL, accumulation of rare variants integrated and extended locus-specific; aSum, data-adaptive sum test; CAST, cohort allelic sums test; CMC, combined multivariate and collapsing; EC, exponential combination; EPACTS, efficient and parallelizable association container toolbox; EREC, estimated regression coefficient; GRANVIL, gene- or region-based analysis of variants of intermediate and low frequency; KBAC, kernel-based adaptive cluster; MiST, mixed-effects score test for continuous outcomes; MZ, Morris and Zeggini; RBT, replication-based test; Rvtests, rare-variant tests; SKAT, sequence kernel association test; SSU, sum of squared score; VAT, variant association tools; VT, variable threshold; and WSS, weighted-sum statistic.
[a]More information is given in Table 3.

(Lee et al. 2014)

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 15 September 2014
doi: 10.3389/fgene.2014.00323

A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required

Carmen Dering[1], Inke R. König[1], Laura B. Ramsey[2], Mary V. Relling[2], Wenjian Yang[2] and Andreas Ziegler[1,3,4]*

[1] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Lübeck, Germany
[2] Pharmaceutical Department, St. Jude Children's Research Hospital, Memphis, TN, USA
[3] Zentrum für Klinische Studien, Universität zu Lübeck, Lübeck, Germany
[4] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

Edited by:
Daniel C. Koboldt, Washington University in St. Louis, USA

Reviewed by:
Michelle Lacey, Tulane University, USA
Jian Li, Tulane University, USA

***Correspondence:**
Andreas Ziegler, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany
e-mail: ziegler@imbs.uni-luebeck.de

The advent of next generation sequencing (NGS) technologies enabled the investigation of the rare variant-common disease hypothesis in unrelated individuals, even on the genome-wide level. Analysis of this hypothesis requires tailored statistical methods as single marker tests fail on rare variants. An entire class of statistical methods collapses rare variants from a genomic region of interest (ROI), thereby aggregating rare variants. In an extensive simulation study using data from the Genetic Analysis Workshop 17 we compared the performance of 15 collapsing methods by means of a variety of pre-defined ROIs regarding minor allele frequency thresholds and functionality. Findings of the simulation study were additionally confirmed by a real data set investigating the association between methotrexate clearance and the *SLCO1B1* gene in patients with acute lymphoblastic leukemia. Our analyses showed substantially inflated type I error levels for many of the proposed collapsing methods. Only four approaches yielded valid type I errors in all considered scenarios. None of the statistical tests was able to detect true associations over a substantial proportion of replicates in the simulated data. Detailed annotation of functionality of variants is crucial to detect true associations. These findings were confirmed in the analysis of the real data. Recent theoretical work showed that large power is achieved in gene-based analyses only if large sample sizes are available and a substantial proportion of causing rare variants is present in the gene-based analysis. Many of the investigated statistical approaches use permutation requiring high computational cost. There is a clear need for valid, powerful and fast to calculate test statistics for studies investigating rare variants.

(Dering et al. 2014)

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 15 September 2014
doi: 10.3389/fgene.2014.00323

A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required

Carmen Dering[1], Inke R. König[1], Laura B. Ramsey[2], Mary V. Relling[2], Wenjian Yang[2] and Andreas Ziegler[1,3,4]*

[1] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Lübeck, G
[2] Pharmaceutical Department, St. Jude Children's Research Hospital, Memphis, TN, USA
[3] Zentrum für Klinische Studien, Universität zu Lübeck, Lübeck, Germany
[4] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South

Edited by:
Daniel C. Koboldt, Washington University in St. Louis, USA

Reviewed by:
Michelle Lacey, Tulane University, USA
Jian Li, Tulane University, USA

***Correspondence:**
Andreas Ziegler, I
Medizinis
Sta
Univ
Schle
Lübeck,
23562 L
e-mail: zi imbs.uni-luebeck.de

Collapsing tests typically do not perform well

The advent of next generati enabled the investigation of the rare variant-co related individuals, even on the genome-wide I equires tailored statistical methods as single m entire class of statistical methods collapses ra of interest (ROI), thereby aggregating rare variants. udy using data from the Genetic Analysis Workshop 17 rformance of 15 collapsing methods by means of a variety of s regarding minor allele frequency thresholds and functionality. Findings ulation study were additionally confirmed by a real data set investigating the ociation between methotrexate clearance and the *SLCO1B1* gene in patients with acute lymphoblastic leukemia. Our analyses showed substantially inflated type I error levels for many of the proposed collapsing methods. Only four approaches yielded valid type I errors in all considered scenarios. None of the statistical tests was able to detect true associations over a substantial proportion of replicates in the simulated data. Detailed annotation of functionality of variants is crucial to detect true associations. These findings were confirmed in the analysis of the real data. Recent theoretical work showed that large power is achieved in gene-based analyses only if large sample sizes are available and a substantial proportion of causing rare variants is present in the gene-based analysis. Many of the investigated statistical approaches use permutation requiring high computational cost. There is a clear need for valid, powerful and fast to calculate test statistics for studies investigating rare variants.

(Dering et al. 2014)

For what else are human DNA sequences 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)

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

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")`):

Carlson 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:

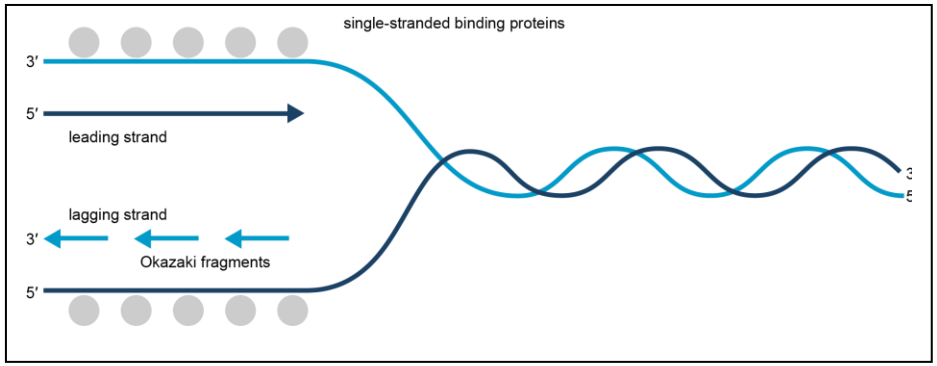
- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

B. Counting letters or words

- 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)

Recall: DNA biosynthesis

- 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).



C. Comparing multiple sequences

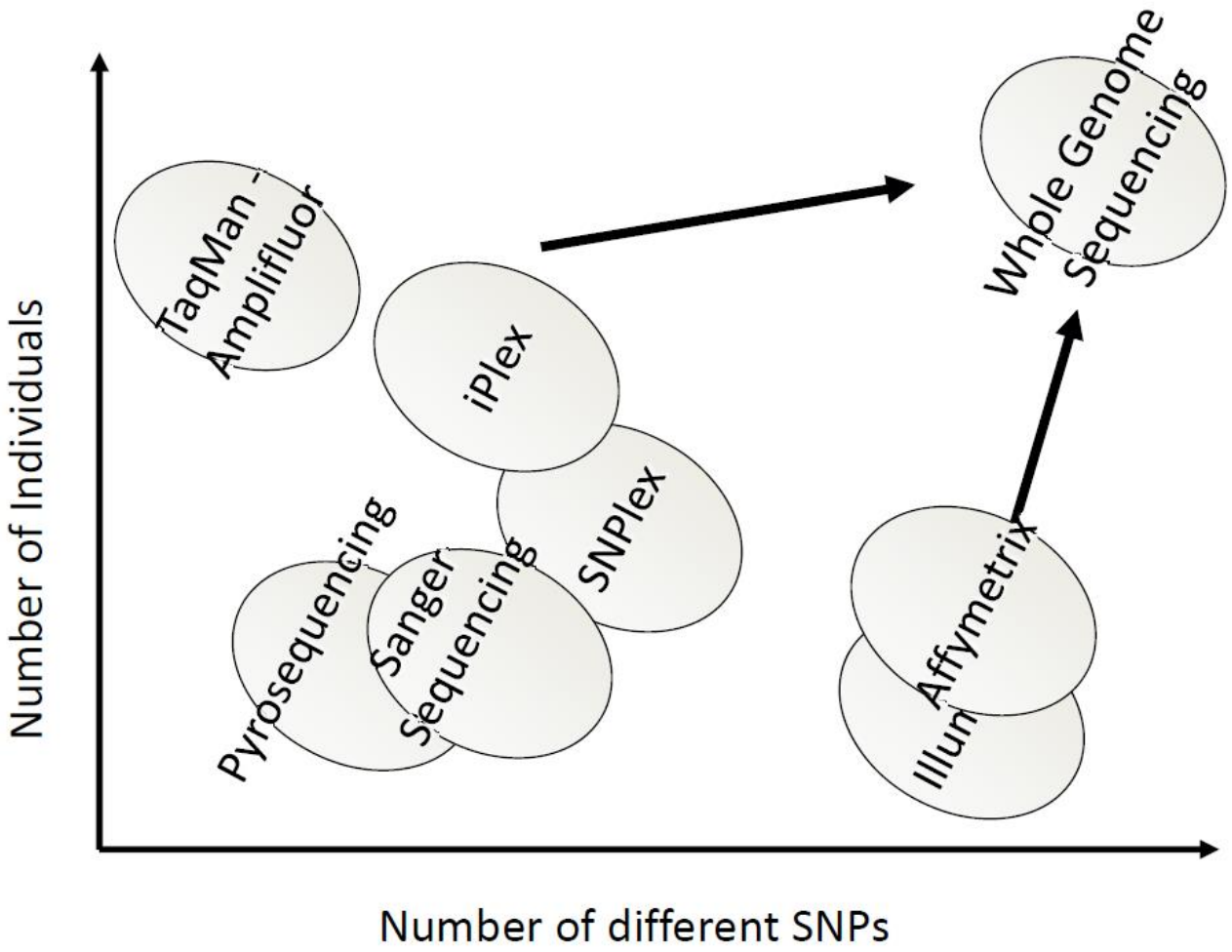
- After collection of a set of related sequences, how can we compare them as a set?
- How should we line up the sequences so that the most similar portions are together?
- What do we do with sequences of different length?

```
            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
```

D. Genomic variation for complex diseases



3 Investigating frequencies of occurrences of words

Introduction

- 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 (e.g., CpGs)

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 ← 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. 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
3. Statistics is an art. It uses mathematical methods but it is much more than maths alone
4. 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, …, 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 $\chi = \{A, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$
  - 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
    - $p_1, \dots, p_J \geq O$ 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 $L_1, \ldots, L_n$?
  - New sequence $X_1, \ldots, X_n$:

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

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

$$N=X_1+\ldots+X_n$$

  - 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 $X_1, \ldots, X_n$ 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

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

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 $x_1, x_2, \ldots$ (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 $E(Y) = c\,E(X)$
  - $E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n)$
- Because X_i are assumed to be independent and identically distributed (iid):

$$E(X_1 + \ldots + X_n) = n\,E(X_1) = n\,p_A$$

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 Var(X) can also be written as:

$$Var(X) = E(X^2) - [E(X)]^2]$$

  - If Y=c X then Var (Y) = c^2 Var (X)
  - The variance of a sum of independent random variables is the sum of the individual variances

- For the random variables X_i:
  Var (X_i) = $[1^2 \times p_A + 0^2 \times' (1 - p_A)] - p_A^2$ = $p_A(1 - p_A)$
  Var (N) = n 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.

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

  - The positive square root of the variance of X is called its standard deviation 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 :

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

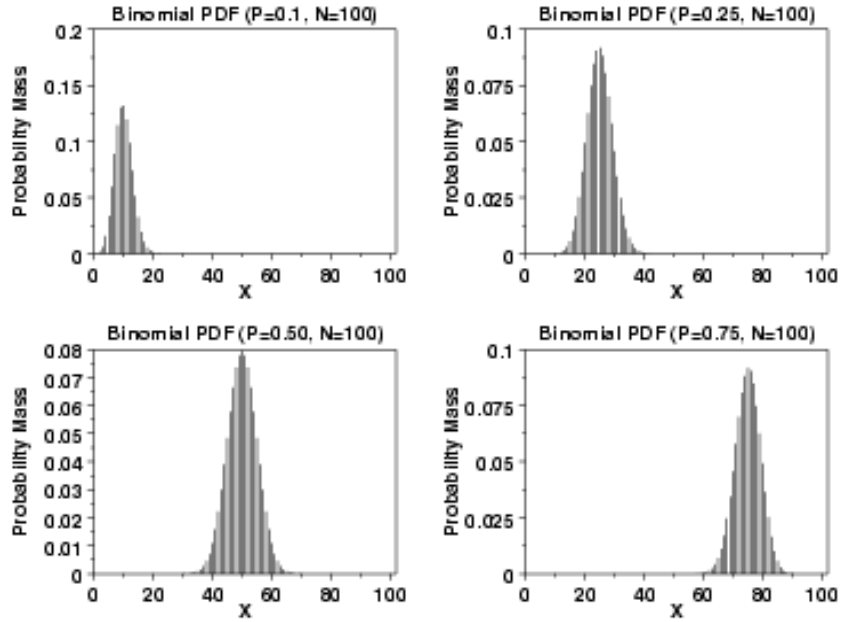
with the binomial coefficient $\binom{n}{j}$ determined by

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

and *j!=j(j-1)(j-2)…3.2.1, 0!=1*

The binomial distribution

- The mean is n*p* and the variance is n*p(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 $N_1, \ldots, N_n$ having the same distribution as N
  - We can use the sample mean to estimate the expected value E(N):

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

  - 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, …, X_n from which N can be computed by adding the X's

Simulating from probability distributions

- Simulate a sequence of bases $L_1, \ldots, L_n$:
  - Divide the interval (0,1) in 4 intervals with endpoints

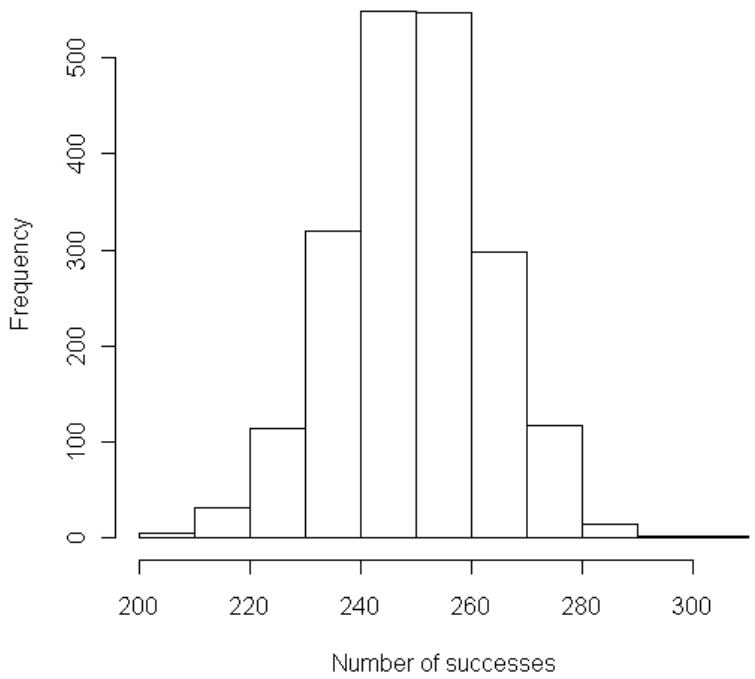
$$0, p_A, p_A + p_C, p_A + p_C + p_G, 1$$

  - 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 $L_1, \ldots, L_n$
- 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




```
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

`x, q`
: vector of quantiles.

`p`
: vector of probabilities.

`n`
: number of observations. If `length(n) > 1`, the length is taken to be the number required.

`size`
: 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 → you got lucky!!!!

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?

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?

Number of sequences = 2000

Number of trials = 1000

Back to our original question

- 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

Exact computation via closed form of relevant distribution

- The formula for the binomial probability mass function is :

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

and therefore

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

$$= 0.00019359032194965841$$

- 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)
```

	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)

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
```

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 $X_1, \ldots, X_n$ are iid random variables with mean μ and variance σ^2, then we know that for the sample average

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

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

- Hence,

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

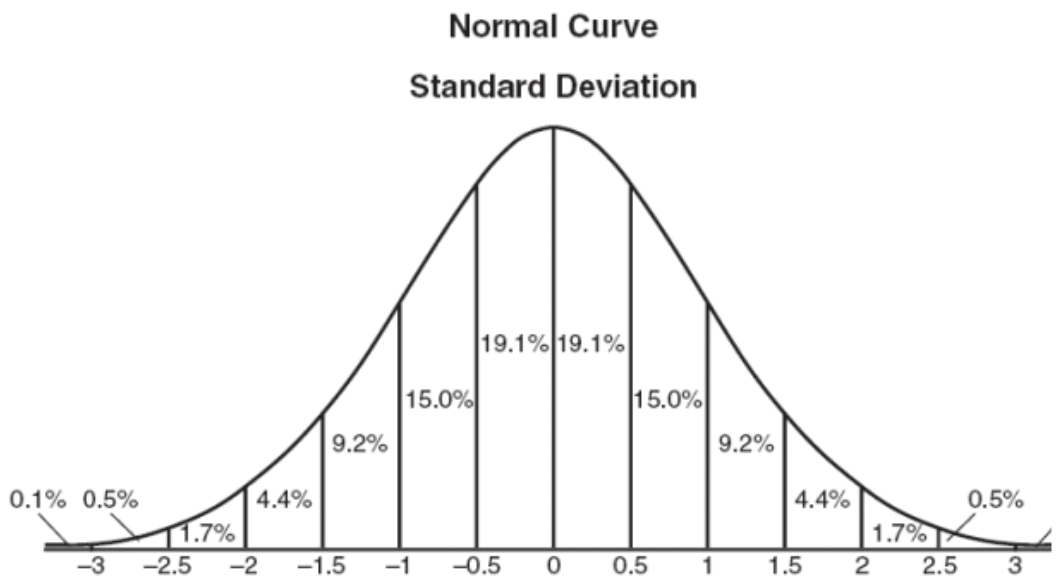
Approximate via Central Limit Theory

- 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 $\phi(.)$ the standard normal distribution defined as

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



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)
```

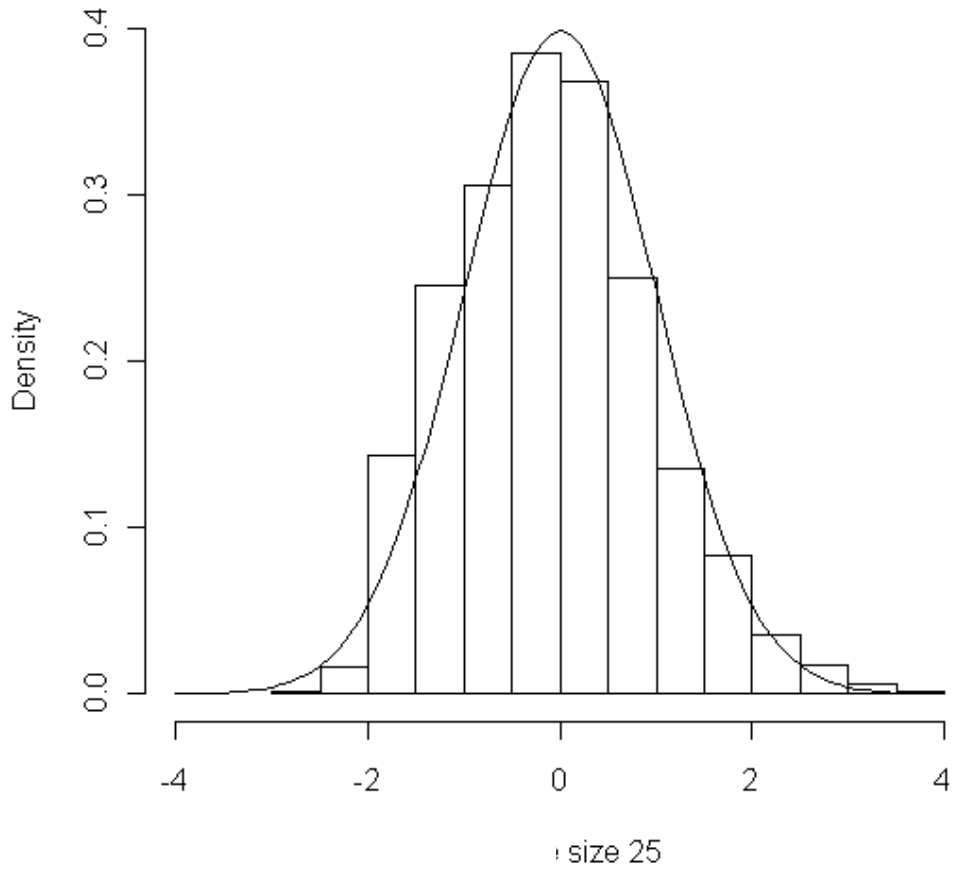
How do the estimates of $P(N \geq 300)$ compare?

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



Supporting doc to this class (complementing course slides)

√

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.

AJHG 2014; 95, 5-23

Questions?