

Model comparison and hypothesis testing

**Prof. Manuel A. Rivas (rivaslab.stanford.edu
(rivaslab.stanford.edu))**

Topics in Biomedical Data Science : Large-scale inference

Lecture number 2

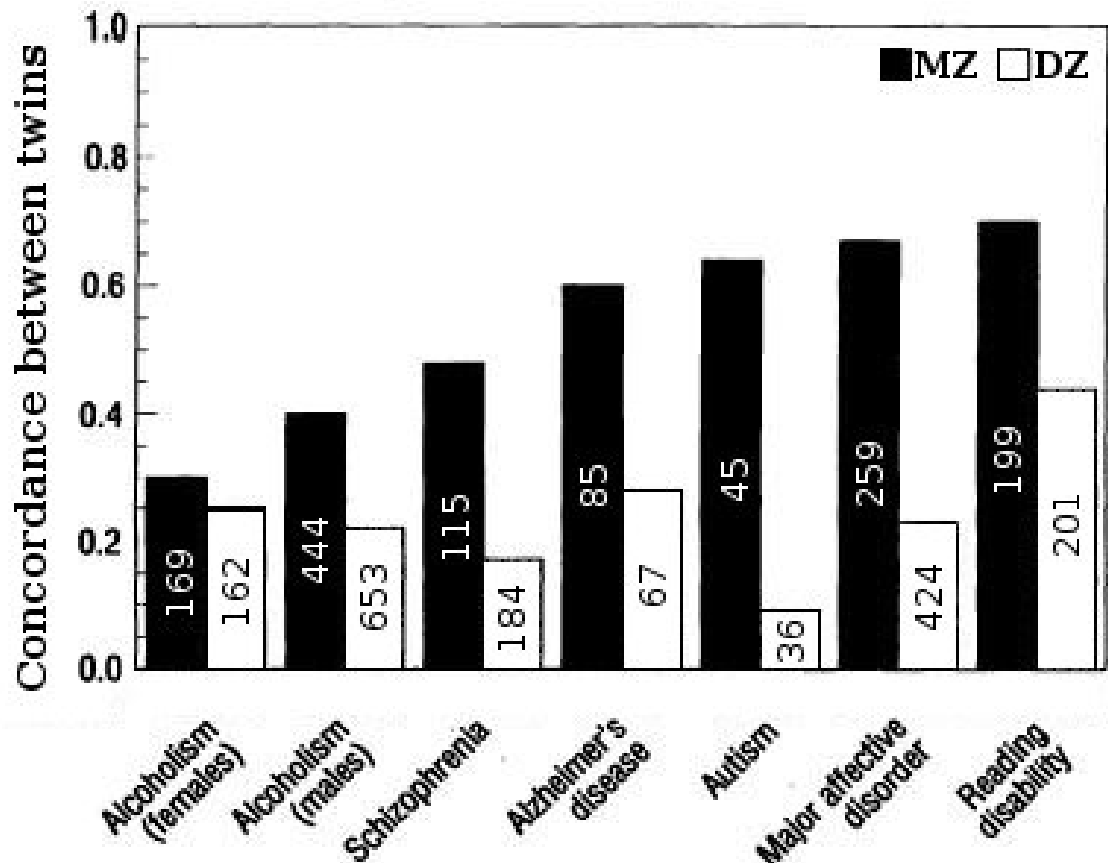
04/6/2017

Stanford University

Genome-wide association studies (motivating example)

Does genetics affect trait? Is the trait heritable?

NB: We will revisit variance components in Lecture #3.



Source: Wikipedia *Twin concordances for seven psychological traits (sample size shown inside bars), with DZ being fraternal and MZ being identical twins.*

Question we are trying to ask is: Do more close relatives have more similar phenotypes (on average)? Twin studies compare monozygotic twins with dizygotic twins (but environment may confound)

Motivation

1. What is a genetic association?
2. Why is this important?

Human Genome Primer

- 22 autosome pairs + 2 sex chromosome
- 3.3 billion base pairs with alphabet {A, C, G, T}
- Approximately 1% encodes proteins

Single nucleotide polymorphisms

On average, 1:300 positions has (common) variation in population, called "SNP"

Genomes in population:

... A G T G ... (96%)

... A T T G ... (4%)

SNP/alleles: G/T, minor allele frequency (MAF) = 4%

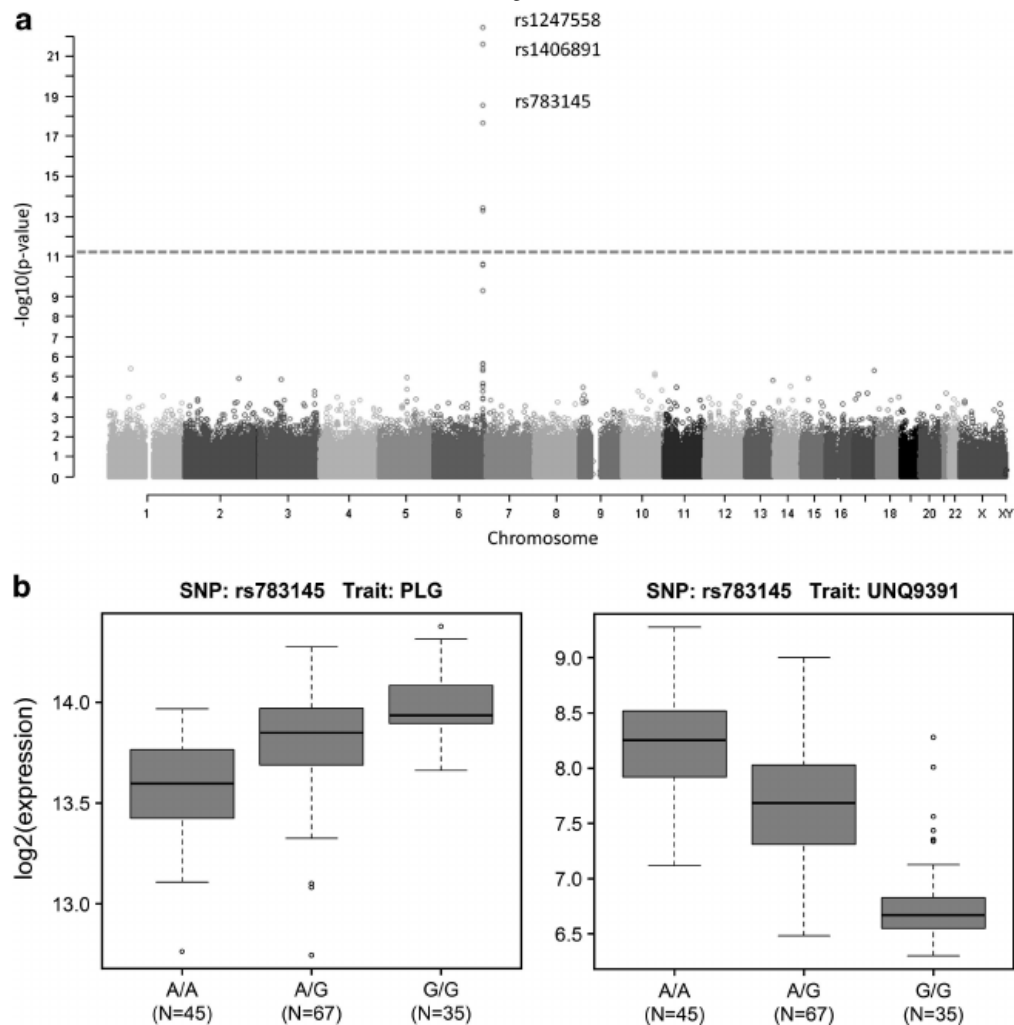
Individuals in a population

For individuals in a population

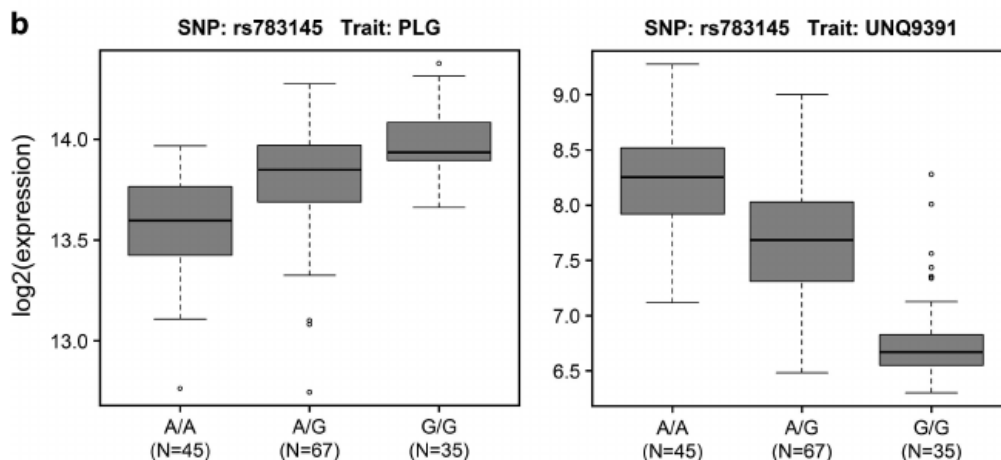
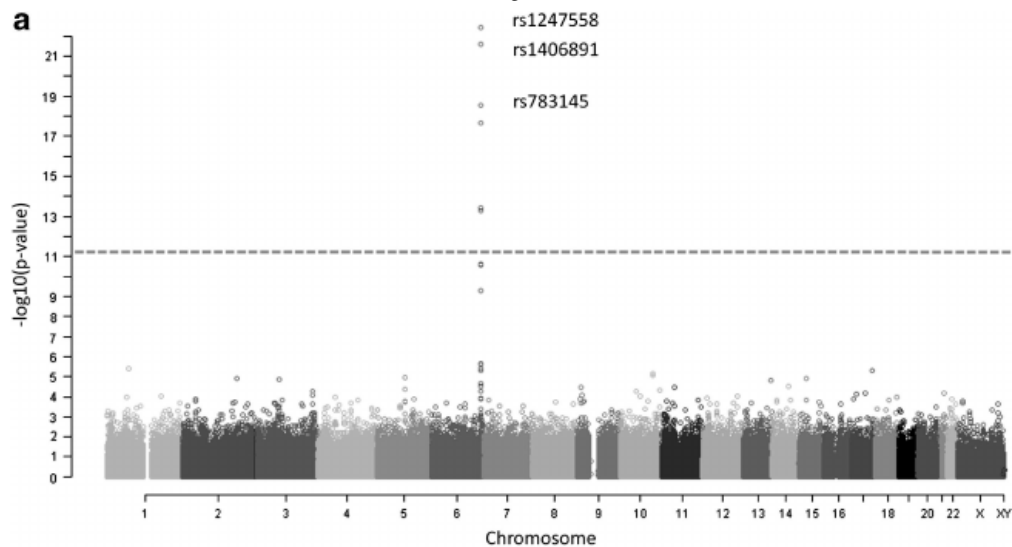
GG = 92.1% (p^2)

GT = 7.7% ($2 \times p \times q$, where p = major allele frequency, and q = minor allele frequency)

TT = 0.2% (q^2)



Genomics of ADME gene expression: Mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver, Schröder et al. 2013



Boxplot shows : (1) medians (thick lines), (2) interquartile range (boxes), (3) 1.5 x interquartile range (dotted segments), and (4) outliers (points).

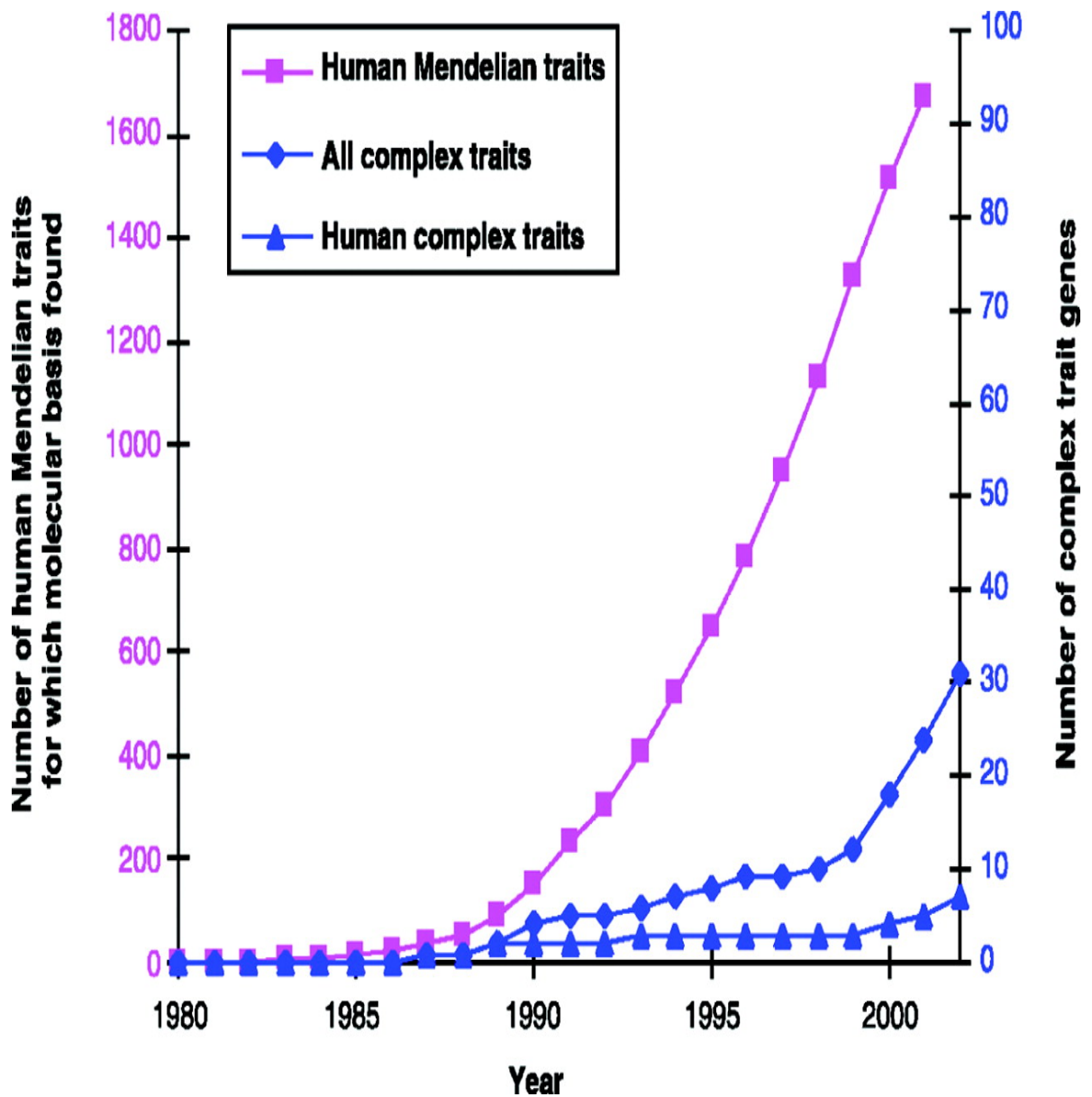
Carriers of A allele have lower *PLG* gene expression value.

Why are genetic associations important?

- Hint of biology behind the diseases and traits
- Hint of targets for therapeutics

Large effect variants were found for

many Mendelian traits but not for complex traits during 1985-2000



Glazier, Nadeau and Aitman, Science 2002

Generating effective therapeutic hypotheses

PCSK9 and LDL-C

2003: *PCSK9* and LDL-C

Mutations in *PCSK9* cause autosomal dominant hypercholesterolemia

Marianne Abifadel^{1,2}, Mathilde Varret¹, Jean-Pierre Rabès^{1,3},
Delphine Allard¹, Khadija Ouguerram⁴, Martine Devillers¹,
Corinne Cruaud⁵, Suzanne Benjannet⁶, Louise Wickham⁶,
Danièle Erlich¹, Aurélie Derré¹, Ludovic Villéger¹, Michel Farnier⁷,
Isabel Beucler⁸, Eric Bruckert⁹, Jean Chambaz¹⁰, Bernard Chanu¹¹,
Jean-Michel Lecerf¹², Gerald Luc¹², Philippe Moulin¹³,
Jean Weissenbach⁵, Annick Prat⁶, Michel Krempf⁶,
Claudine Junien^{1,3}, Nabil G Seidah⁶ & Catherine Boileau^{1,3}

- Hypercholesterolemia (high LDL-C) is a risk factor for heart disease
- 2003 Nature Genetics

2

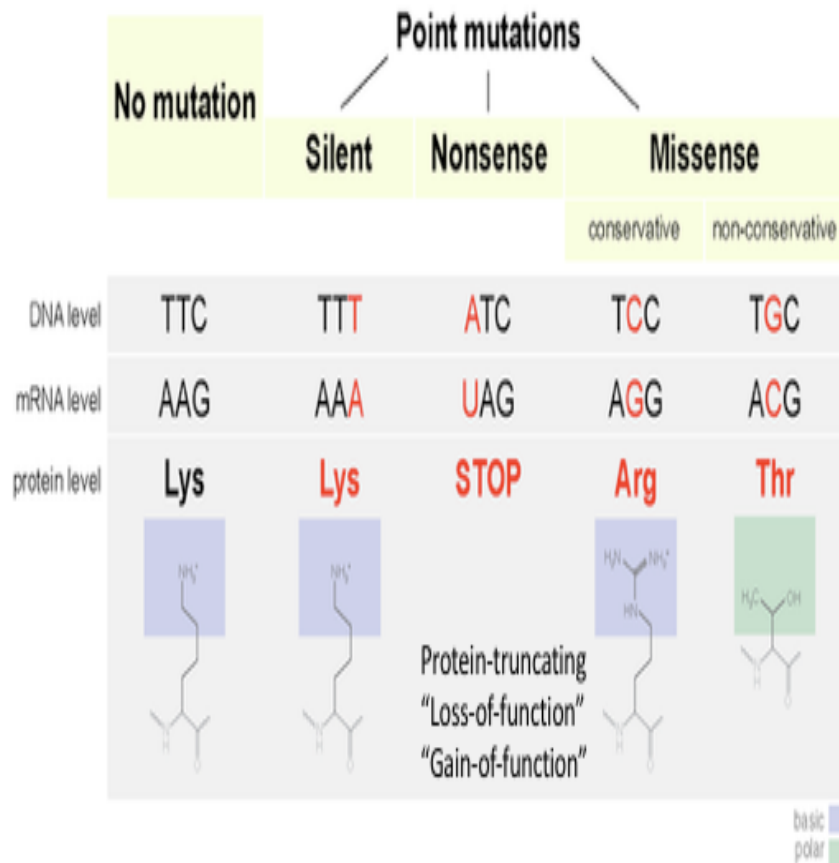
Source: *Matti Pirinen*

Generating effective therapeutic hypotheses

Protein-coding genes

Mutations in protein-coding sequence

In genes, a group of three DNA bases codes for one amino acid



Rosalind

3

Generating effective therapeutic hypotheses

Clinical trials - lower LDL levels

PCSK9 inhibition with evolocumab (AMG 145) in heterozygous familial hypercholesterolaemia (RUTHERFORD-2): a randomised, double-blind, placebo-controlled trial

Prof Frederick J Raal, PhD, Prof Evan A Stein, PhD, Robert Dufour, MD, Traci Turner, MD, Fernando Civeira, MD, Prof

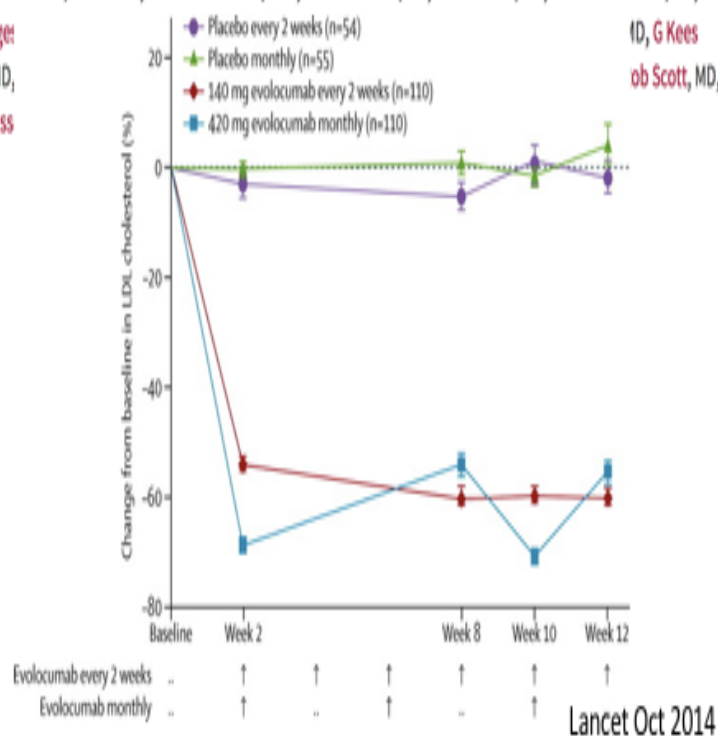
Lesley Burger

Hovingh, MD,

Scott M Wass

Dr G Kees

Rob Scott, MD,



6

Generating effective therapeutic hypotheses


Clinical trials - disease endpoints (March 2017)

ORIGINAL ARTICLE

Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease

Marc S. Sabatine, M.D., M.P.H., Robert P. Giugliano, M.D., Anthony C. Keech, M.D., Narimon Honarpour, Stephen D. Wiviott, M.D., Sabina A. Murphy, M.P.H., Julia F. Kuder, M.A., Hui Wang, Ph.D., Thomas Wasserman, M.D., Peter S. Sever, Ph.D., F.R.C.P., and Torje R. Pedersen, M.D., for the FOURIER Study Investigators*

March 17, 2017 | DOI: 10.1056/NEJMoa1615664

 Comments open through March 24, 2017

Share

[Abstract](#)

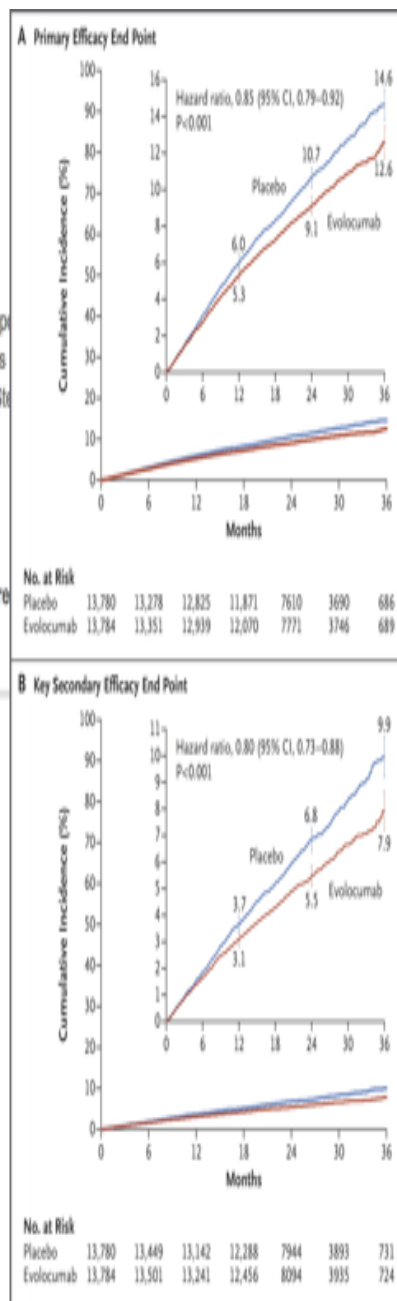
[Article](#)

[References](#)

[Citing Articles \(4\)](#)

[Comments \(15\)](#)

[Metrics](#)



Genome-wide association studies (~2006 onwards)

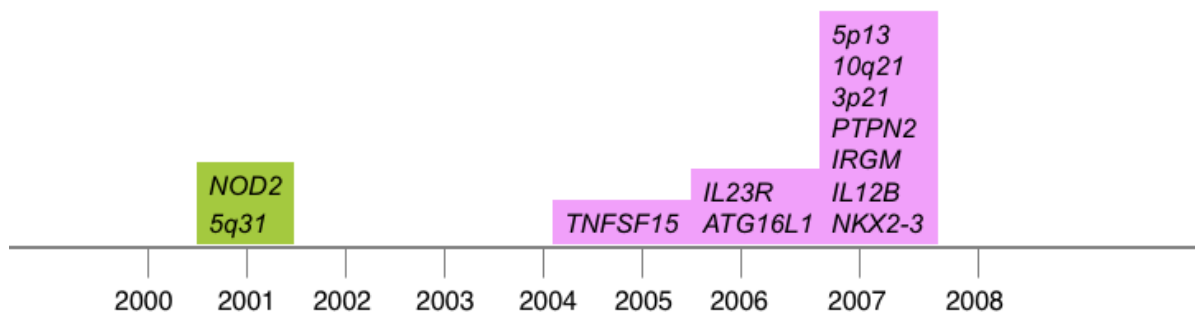
- Idea : To look for associations in a detailed map of common variation across the genome
 - Common disease common variant hypothesis
- Facilitated by
 - Technologies

- Collaboration (consortia)

Progress of GWAS over the past decade

Case study 1: Inflammatory bowel disease

Individual GWA studies – Crohn's



Source: Mark Daly

Today, over 200 loci mapped

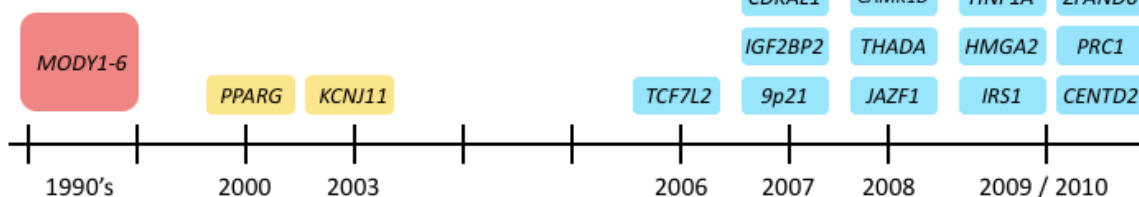
Progress of GWAS over the past decade

Case study 1: type 2 diabetes

T2D genetics through 2011

54 new regions containing genes influencing T2D

GWAS of Related Traits
GWAS of Type 2 diabetes
Candidate Gene Studies
Linkage studies of Mendelian subtypes

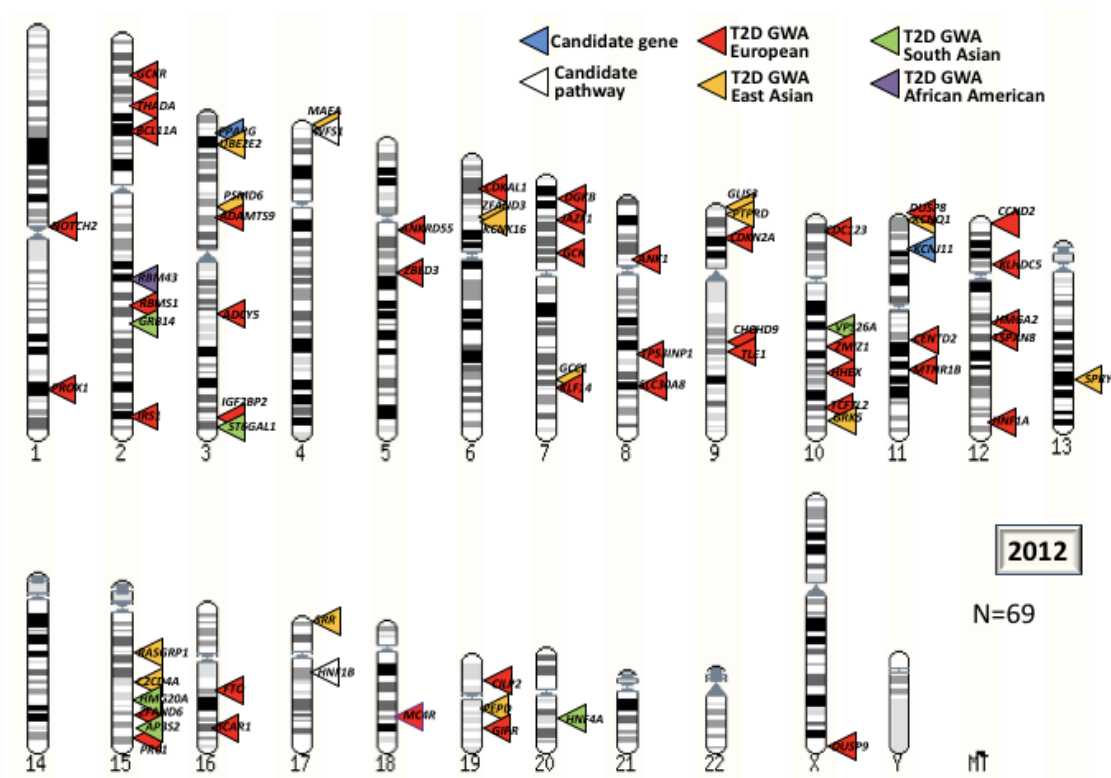


For purposes of presentation, loci are named according to a nearby gene of interest. In only a few cases is the causal gene yet proven.

Source: Mark Daly

Progress of GWAS over the past decade

Case study 1: type 2 diabetes



Established T2D susceptibility loci

Today, over 60 loci mapped to type 2 diabetes.

Source: Anubha Mahajan, Andrew Morris, and Mark McCarthy

Early divide in Genome-wide association studies

How to assess for association?

Most straightforward: compare proportion of each SNP allele in cases and controls

rs11209026	Allele A	Allele G
Cases	22	976
Controls	68	932

Chi-sq = 24.5, p=7.3 x 10⁻⁷

Simplest tests (single marker regression, chi-square) rule the day - association results requiring arcane statistics/complex multi-marker models are often less reliable

How to compute a chi-squared test

Widely used test is Pearson's chi-squared test

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^n \frac{(O_i/N - p_i)^2}{p_i}$$

where

χ^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution.

O_i = the number of observations of type i .

N = total number of observations

$E_i = Np_i$ = the expected (theoretical) frequency of type i , asserted by the null hypothesis that the fraction of type i in the population is p_i

n = the number of cells in the table.

Question: Degrees of freedom for a case-control association table? Hint: 2×2 table.

P-values

The usual approach to assess evidence for a *population association* between genetic variants and a phenotype of interest is to compute a *p-value* for the null hypothesis (H_0) of no association.

Limitation of p-values - From a *p-value* alone it is difficult to quantify how confident one should be that a given SNP is truly associated with a phenotype.

Setup in a Bayesian framework

Statistical model

The term **statistical model** is a complete description of a random process by which observed data are generated.

Example: Bernoulli process

For example, you may have a bernoulli process:

- A finite or infinite sequence of binary random variables. (think of a series of coin flips).

Example: Normal random variables

Another process is: x_1, x_2, \dots being independent standard normal random variables.

NB: We will look at Bayesian setups for both along with their frequentist analogs.

Let's return to a human genetics example

Let S be the number of case samples. Let R be the number of control samples. Let $N = S + R$

Let y be the number of allelic observations for a particular variant in cases.

Let n be the number of allelic observations for a particular variant in total samples (Hence, number of observations in controls is $n - y$).

Assume that $n \ll \min(S, R)$ we assume that the alleles among cases and controls will be distributed according to a binomial distribution with n trials and success probability θ .

Bayes Factor

The Bayes factor (BF), which quantifies the support for a model over another (regardless of whether these models are correct) is obtained by comparing the **marginal likelihoods** for two models:

Question: Why is it referred to as the marginal likelihood?

Answer: It is referred to as the marginal likelihood because the parameters have been marginalized/"integrated out".

Sometimes it is referred to as $g(y)$, $g(\text{DATA})$, or $f_m(\text{DATA})$.

Bayes Factor

The Bayes factor (BF), which quantifies the support for a model over another (regardless of whether these models are correct) is obtained by comparing the marginal likelihoods for two models:

Bayes Factor

$$BF = \frac{\int \binom{n}{y} \theta^y (1-\theta)^{n-y} g(\theta) d\theta}{\int \binom{n}{y} \theta^y (1-\theta)^{n-y} f(\theta) d\theta},$$

where $f(\theta)$ is the prior density for θ under the null model and $g(\theta)$ is the prior density for θ under the alternative model.

Null model

Under the null model, the prior density for θ is a point mass at $\theta = \frac{S}{N}$ because that is the probability we expect to observe an allele in cases, which is dependent on the case to control ratio. If some bias was expected, maybe as a result of population stratification or familial sharing then a different prior density would be appropriate that reflects the uncertainty regarding the probability of observing an allele in cases.

The marginal likelihood for y is given by

$$\int P(y|\theta) P(\theta) d\theta = \binom{n}{y} \left(\frac{S}{N}\right)^y \left(1 - \frac{S}{N}\right)^{n-y}.$$

Exercise: Please *derive* marginal likelihood for y .

Propose an alternative model

Alternative model

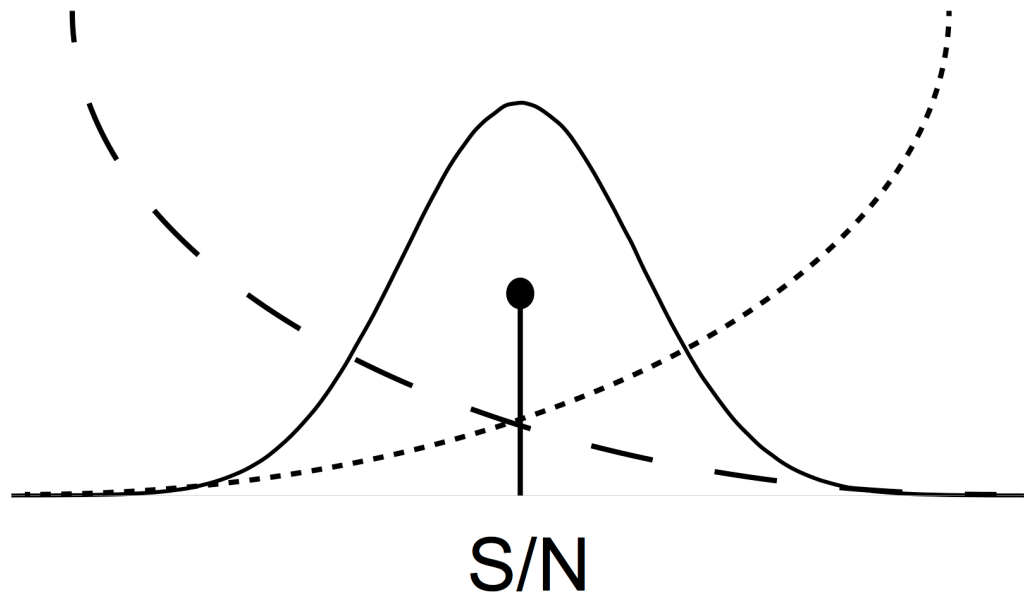
Under the alternative model, to allow protective or risk allele in a gene we assumed that the prior on θ is a mixture of beta distributions:

$$\theta \sim \frac{1}{2} \text{Beta}(\alpha_1, \beta_1) + \frac{1}{2} \text{Beta}(\alpha_2, \beta_2).$$

Then, the marginal likelihood for y is given by

$$\int P(y|\theta) P(\theta) d\theta = \frac{1}{2} \binom{n}{y} \left[\frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1) \Gamma(\beta_1)} \frac{\Gamma(y + \alpha_1) \Gamma(n - y + \beta_1)}{\Gamma(n + \alpha_1 + \beta_1)} + \frac{\Gamma(\alpha_2 + \beta_2)}{\Gamma(\alpha_2) \Gamma(\beta_2)} \frac{\Gamma(y + \alpha_2) \Gamma(n - y + \beta_2)}{\Gamma(n + \alpha_2 + \beta_2)} \right]$$

Exercise: Please derive the Bayes Factor.



The density of the null distribution (solid curve), and a point mass reflecting the prior density of the null model with value at $\theta = S/N$.

An alternative take on quantitative traits

Assume that among N individuals studied, n individuals carried one of the non-reference alleles for a variant of interest.

Typically $n \ll N$.

Let Y_1, \dots, Y_N be the standardised quantitative trait values of the individuals and we assumed that the trait values Y_1, \dots, Y_n correspond to the carriers of the non-reference allele and the values Y_{n+1}, \dots, Y_N correspond to the non-carriers.

We assumed that standardised trait values across the whole sample follow a standard normal distribution, which can be achieved by applying quantile normalisation.

Null model

Under the null model, the gene does not affect the trait and the trait values of the non-reference allele carriers and the non-carriers follow the standard normal distribution:

$$\begin{aligned}\text{NULL} : Y_i &\sim \mathcal{N}(0, 1^2), \text{ for } i = 1, 2, \dots, n \\ Y_j &\sim \mathcal{N}(0, 1^2), \text{ for } j = n + 1, n + 2, \dots, N.\end{aligned}$$

The statistical challenge is to look for strong evidence against the null model. If the variant under consideration affect trait values this will cause a deviation from normality for Y_1, \dots, Y_n .

The Bayesian approach requires specification of the alternative hypothesis.

An alternative model

We assume that the effect of the variant is to shift the mean of the distribution of trait values, so that the trait values of the carriers follow a normal distribution with mean μ and standard deviation s whereas the trait values for the remaining individuals follow a standard normal distribution:

$$\begin{aligned}\text{alternative model} : Y_i &\sim \mathcal{N}(\mu, s^2), \text{ for } i = 1, 2, \dots, n \\ Y_j &\sim \mathcal{N}(0, 1^2), \text{ for } j = n + 1, n + 2, \dots,\end{aligned}$$

Here, we fix the value of s , to $s = 1$, but more general approaches could also allow a change under the alternative hypothesis in the variance of trait values, or potentially in their distribution.

Question: Since it will not be known whether the variant will increase or decrease trait values (as in the case of case-control

data), how will you specify the prior?

The distribution of the trait mean μ is specified under the alternative hypothesis.

Since it will not be known in advance whether variant will increase or decrease trait values we used a 50:50 mixture of two normal distributions as a prior for μ :

$$\mu \sim \frac{1}{2} \mathcal{N}(-a, t^2) + \frac{1}{2} \mathcal{N}(a, t^2),$$

Question: What are a and t called?

We let the hyperparameters be $a = 1.5$ and $t^2 = 0.5$. With these values 95% of the prior mass for μ lies in the set $(-2.89, -0.12) \cup (0.12, 2.89)$ following the signal the method is tailored for, i.e. trait values of individuals carrying the variant in genes contributing to trait variation strongly deviate from normality.

What is a hyperparameter?

A hyperparameter is a parameter of a prior distribution; the term is used to distinguish them from parameters of the model for the underlying system under analysis.

Example of a coin flip

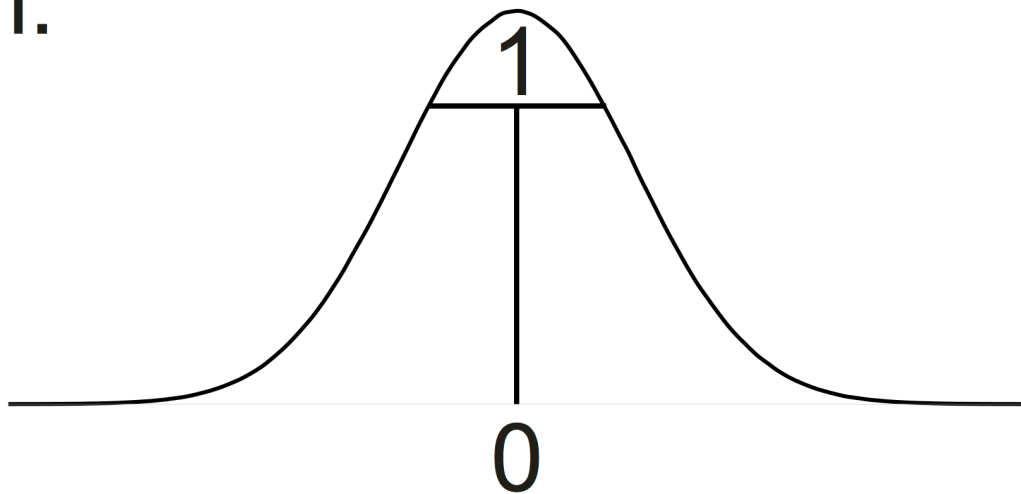
If we are using a beta distribution to model the distribution of the parameter p (sometimes referred to as θ) of a Bernoulli distribution,

then:

p or θ is a parameter of the underlying system (Bernoulli distribution), and α and β are parameters of the prior distribution (beta distribution), hence hyperparameters.

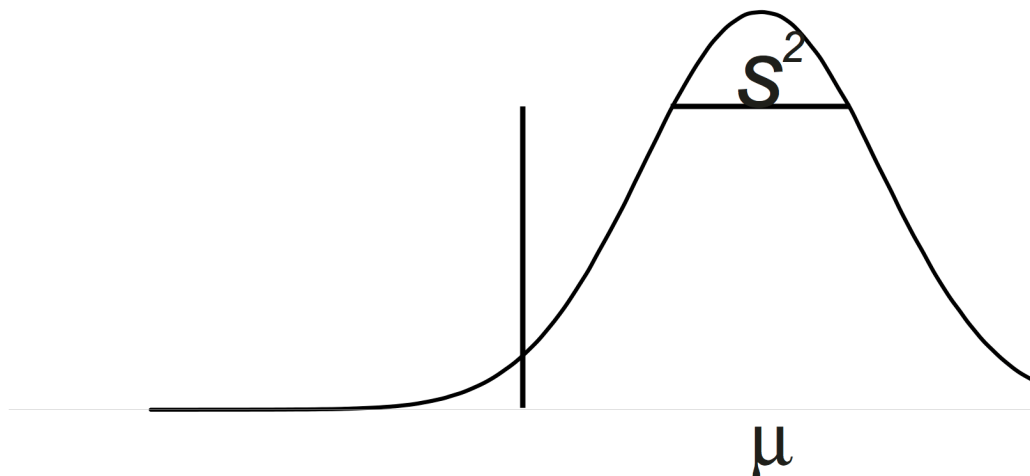
Distribution of the trait values under the null model

i.



Distribution of the trait values under the alternative model

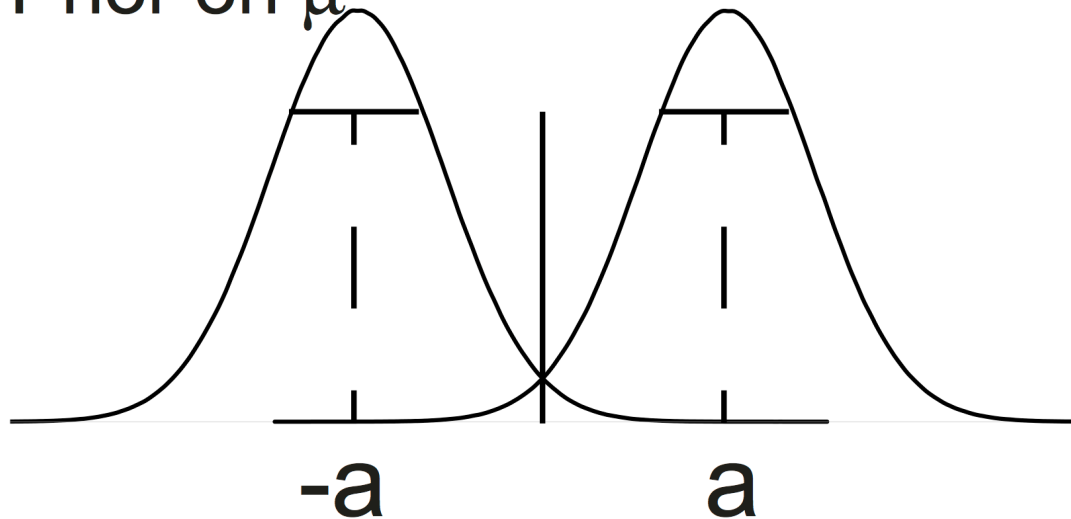
ii.



50:50 mixture of two normal distributions as prior for $\mu \sim 1/2\mathcal{N}(-a, t^2) + 1/2\mathcal{N}(+a, t^2)$

iii.

Prior on μ



Exercise: What would you do in the frequentist setting?

Significance levels in genome-wide studies

Classical multiple testing theory in statistics is concerned with the problem of "multiple tests" of a single "global" null hypothesis. This, we would argue, is a problem far removed from that which faces us in genome-wide association studies, where we face the problem of testing "multiple hypotheses" (for a particular disease, one hypothesis for each SNP, or region of correlated SNPs, in the genome) and we thus do

not subscribe to the view that one should correct significance levels for the number of tests performed to obtain "genome-wide significance levels".

WTCCC, 2007

Nonetheless, our aim is to keep the false positive rate within acceptable bounds and this still leads to the view that very low P values are needed for strong evidence of association. But the factor determining the threshold is not the number of tests performed, but the a priori probability that there is likely to be a true association at any specified location in the genome. Of course, we cannot know this prior probability from objective evidence, but we can perhaps estimate an order of magnitude.

WTCCC, 2007

There are two linked questions. The first concerns the choice of an appropriate "threshold" for reporting possible associations as likely to be genuine. Here the mathematics is quite straightforward if we make the simplifying assumption that we have the same power to

detect all true associations. Then we have *posterior odds* for true association = Prior odds x Power/Significance threshold. That is, for a given significance threshold, the probability of a true association depends on the prior odds and, crucially, the power. A plausible estimate for the prior odds of true association at any specified locus might be of the order of 100,000:1 against, for example, on the basis of 1,000,000 "independent" regions of the genome and an expectation of 10 detectable genes involved in the condition. (Other plausible estimates might vary from this by an order of magnitude or so in either direction.)

Then, assuming a power of 0.5 and a significance threshold of 5×10^{-7} , the posterior odds in favour of a "hit" being a true association would be 10:1. However, if we relax this significance threshold by a factor of ten, or alternatively if the power were lower by a factor of 10, the posterior odds that a "hit" is a true association would also be reduced by a factor of ten. This simple mathematical analysis is little affected by allowing for the fact that true associations come in various sizes with varying power to detect them; the above formula is simply modified by interpreting "power" as the mean power.

WTCCC, 2007

Posterior Odds

After the association data are available, a related but different question is whether a particular positive finding is likely to be a true one.

For that calculation,
posterior odds (PO) = prior odds \times Bayes Factor

For instance, if the prior odds of a finding were 1 : 100,000 then a Bayes Factor of 1,000,000; that is $\log_{10}(\text{Bayes Factor})$; corresponds to a posterior odds of 10 to 1.

Posterior probability of association (PPA)

From the expression of posterior odds (PO) we can obtain **PPA**:
$$\text{PPA} = \text{PO} / (1 + \text{PO})$$

Let's review

What concepts have we learned about?

Uncertainty about the model

So far we have assumed that we can formulate the statistical problem in terms of a single model.

Question: What are the key components of that model?

- A likelihood: $f(x|\theta)$
- A prior distribution: $f(\theta)$

Model averaging

It is possible to obtain $f(\psi|\text{DATA})$, the posterior distribution of a parameter ψ over all models, as

$$f(\psi|\text{DATA}) = \sum_{m=1}^M f(m|\text{DATA}) f_m(\psi|\text{DATA})$$

Connection between frequentist and Bayesian approaches

Bayes factors and likelihood ratio tests

Likelihood ratios are computed at the Maximum Likelihood estimate (MLE) of the data for two models, usually at

- The null model
- an alternative model

Let's turn to an R shiny app for intuition

Source: https://github.com/EtzAlex/shiny_likelihooods (https://github.com/EtzAlex/shiny_likelihooods), an app originally created by Fabian Dablander for this blog post: <http://alexanderetz.com/2015/04/15/understanding-bayes-a->

[look-at-the-likelihood/](http://alexanderetz.com/2015/04/15/understanding-bayes-a-look-at-the-likelihood/)

[\(http://alexanderetz.com/2015/04/15/understanding-bayes-a-look-at-the-likelihood/\)](http://alexanderetz.com/2015/04/15/understanding-bayes-a-look-at-the-likelihood/)

Alternative approaches to model comparison

We will return to this in the mixture models section where we discuss **information criteria**.

Instructions for students to download and run the shiny app

1. mkdir shiny
2. cd shiny
3. git clone https://github.com/EtzAlex/shiny_likelihoods
(https://github.com/EtzAlex/shiny_likelihoods)
4. R
5. source('shiny_likelihoods.R')
6. runApp('shiny_likelihoods.R')

We strongly recommend to install anaconda

1. <https://www.continuum.io/downloads>
(<https://www.continuum.io/downloads>)
2. Install conda for R <https://www.continuum.io/conda-for-r> (<https://www.continuum.io/conda-for-r>)

Please do e-mail the instructor if you have additional questions.