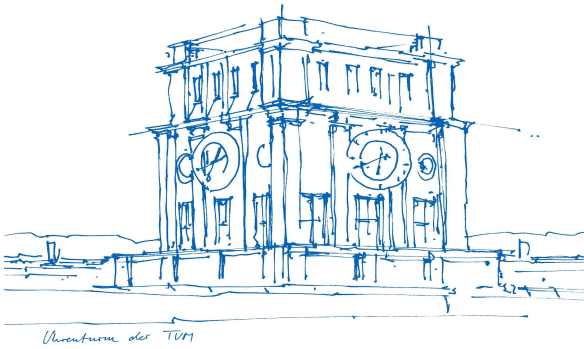
Statistical primer for systems genetics

Julien Gagneur

Prof. for Computational Molecular Medicine
Technical University of Munich

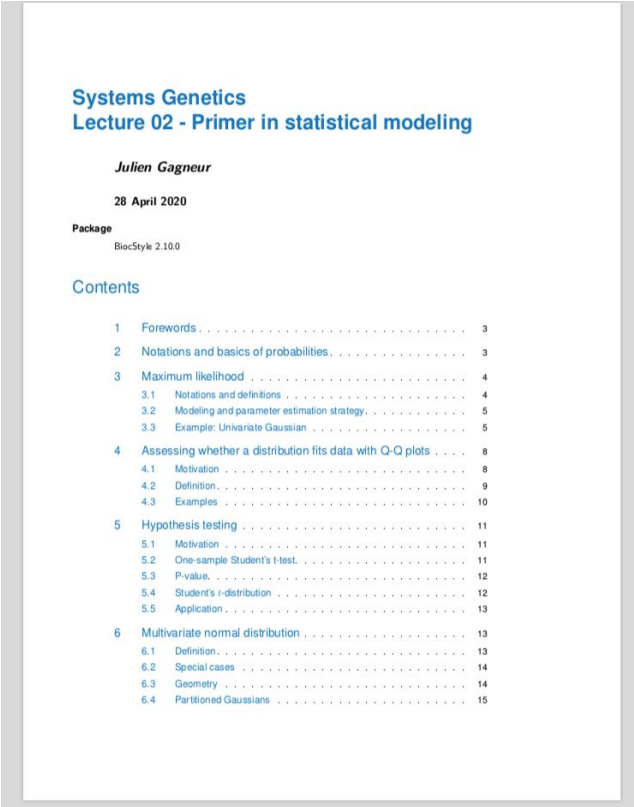
www.gagneurlab.in.tum.de

To understand the genetic basis of gene regulation and its implication in diseases



Motivation

- Lecture 02 provides a primer in statistical modeling covering topics necessary for the module
- This presentation gives an overview of the primer and how it connects to systems genetics modeling questions



Systems Genetics
Lecture 02 - Primer in statistical modeling

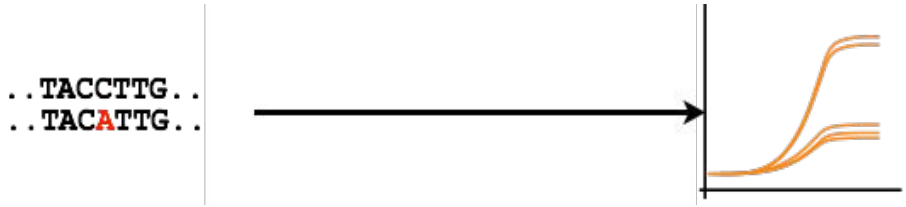
Julien Gagneur

28 April 2020

Package
BiocStyle 2.10.0

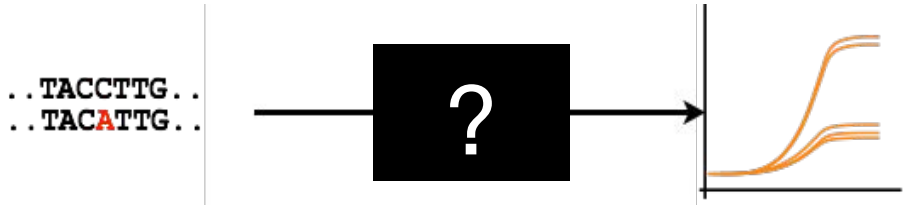
Contents

Goals of Systems Genetics



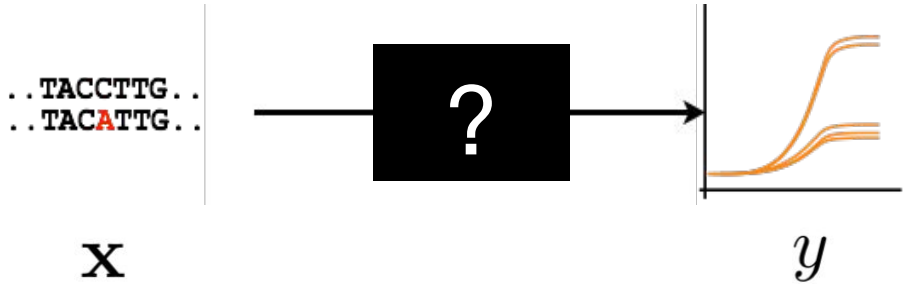
- Which genetic variants cause phenotypic variations?
- How strong are the effects?
- Through which molecular mechanisms?

Challenges of Systems Genetics



- Black box / complex biology

Modeling approach: data generative models



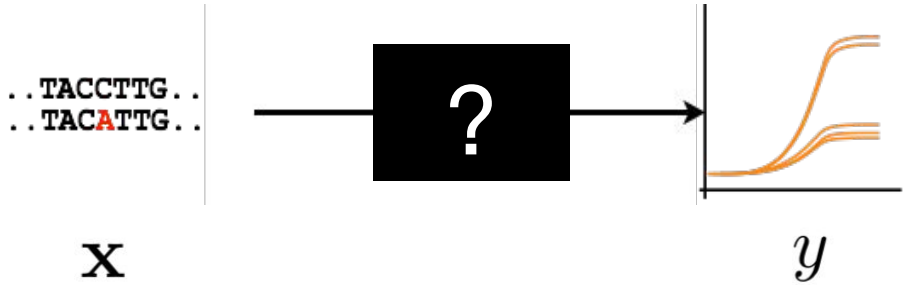
- Black box / complex biology
- Simple, high-level abstractions, often via linear models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- **Data generative models** that describe the distribution of the phenotype y in function of genetic and other explanatory variables (x) and parameters (beta)

$$p(y|\mathbf{x}, \boldsymbol{\beta})$$

Parameter estimation



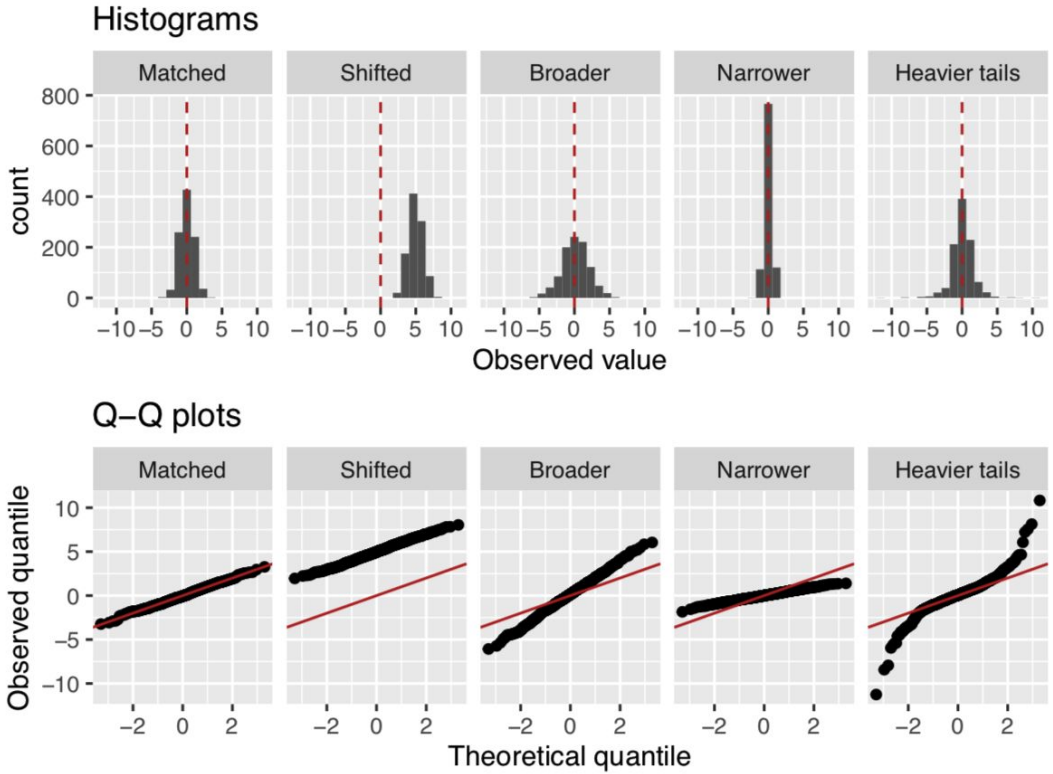
$$p(y|\mathbf{x}, \boldsymbol{\beta})$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- Parameter estimation (a.k.a model inference) aims at fitting a model to data
- **Maximum likelihood estimation** estimates the parameters that maximizes $p(y|\mathbf{x}, \boldsymbol{\beta})$
- Allows answering “How strong are the effects?”

Does the model distribution fits the data?

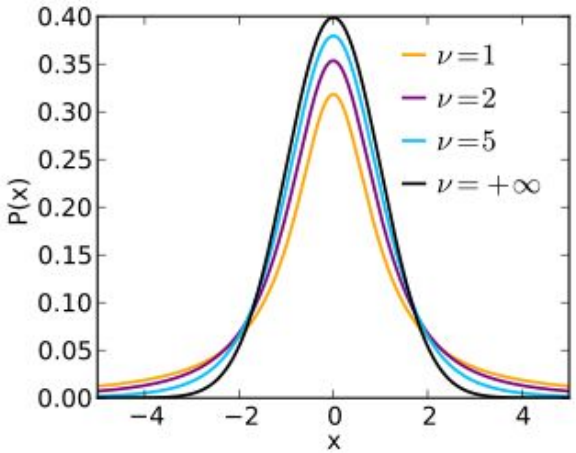
- Data generative models model the distribution of the data. Are they reasonable?
- **Quantile-quantile plots** allow comparing empirical data distribution compared to fitted distribution



Hypothesis testing

- Which genetic variants cause phenotypic variations?
- Inferring causality is hard. There will be lectures about it.
- For now: which effects are statistically significant? In other words, which effects would still be seen would the experiment be reproduced?
- **Hypothesis testing** allows answering this question.

$$P = p(t \geq t_{\text{obs}} | H_0)$$

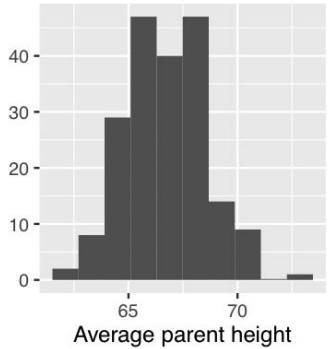


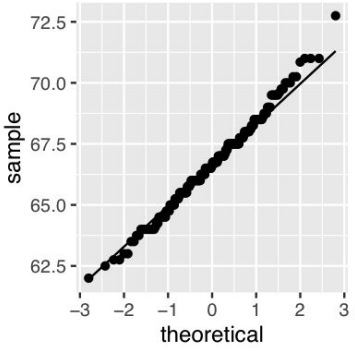
Multivariate Normal Distribution

- Normal distribution often arises among phenotypic traits

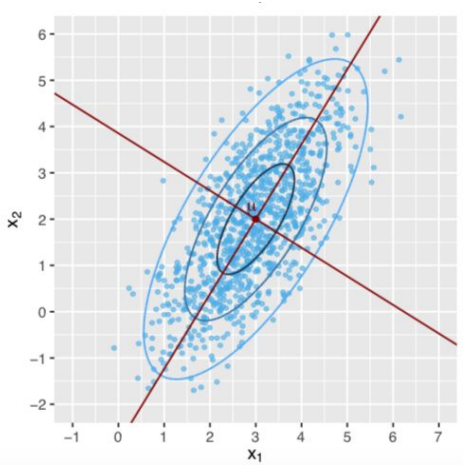
- Multivariate Normal (MVN) describes joint distributions
  - Multiple traits
  - Multi-omics

- The script describes properties of the MVN





Data from Galton, 1886

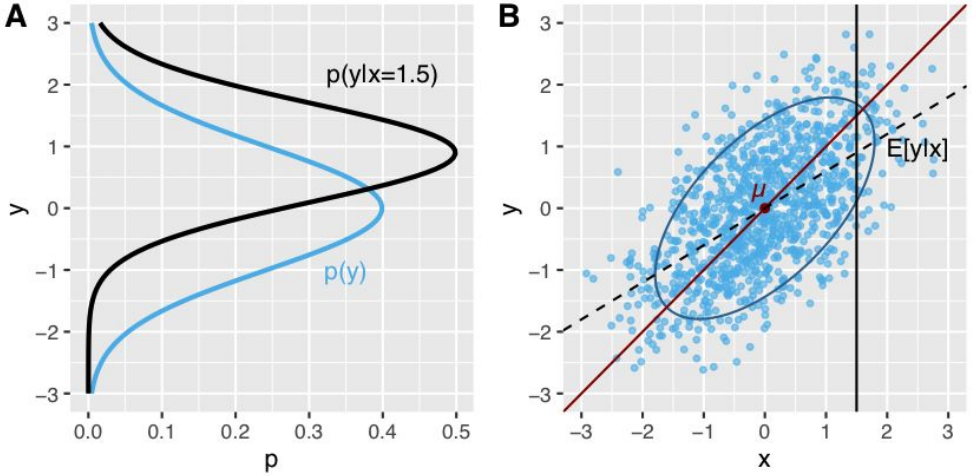


Conditioning and regression

- The **conditional distribution** is the distribution of one variable given (or conditioned on) the value of another one.

  E.g: How does cholesterol levels distribute among individuals with a specific BMI?

- Fundamentally related to the concept of **regression**, which aims at predicting a particular trait y given some explanatory variables x (genotype, sex, etc.)

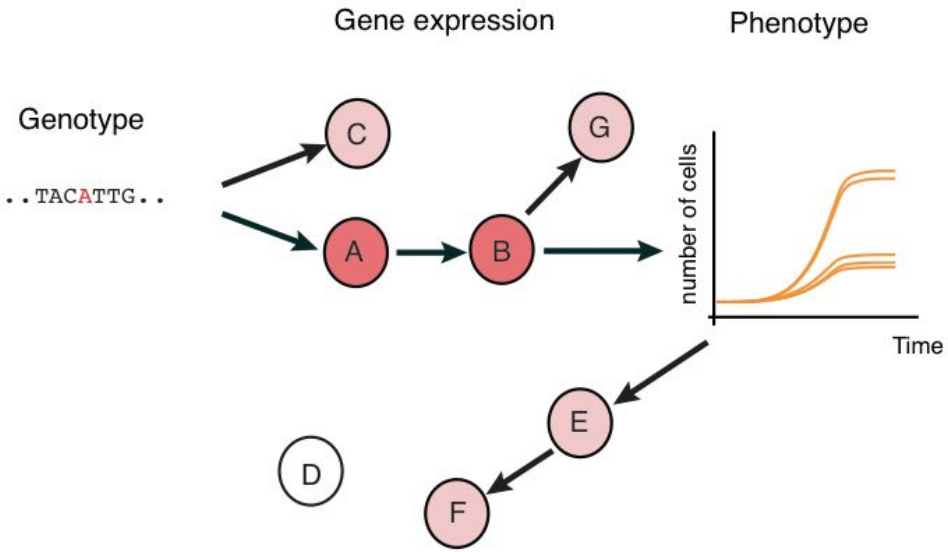


Linear systems

- Gaussian variables whose means are linear combinations of other Gaussian variables

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right)$$

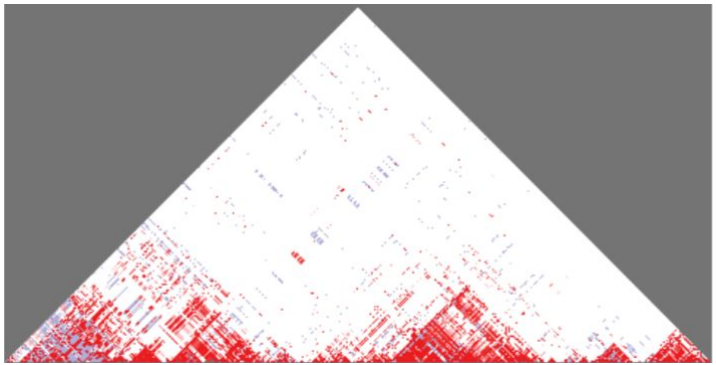
- E.g. in Bayesian networks that are used to model abundances of molecules, such as RNAs, proteins or metabolites



Estimating effects for millions of variants is hard

- Millions of genetic variants in human genome.
- Not enough data to estimate the entire model
- Strong genetic correlation in the population (e.g. ethnicity)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_{1,000,000} x_{1,000,000} + \epsilon$$



Genotype correlation matrix

Chromosome

Estimating a single controlling for all others?

- Millions of genetic variants in human genome.
- Not enough data to estimate the entire model
- Strong genetic correlation in the population (e.g. ethnicity)
- Can we estimate effect of one variant controlling for all millions of other variants?

$$y = \beta_0 + \beta_1 x_1 + \boxed{\beta_2 x_2 + ...\beta_{1,000,000} x_{1,000,000}} + \epsilon$$

Marginalization with linear systems

- Millions of genetic variants in human genome.
- Not enough data to estimate the entire model
- Strong genetic correlation in the population (e.g. ethnicity)
- Can we estimate effect of one variant controlling for all millions of other variants?
- By assuming MVN distribution (mixed effect models) on other parameters and marginalizing out.

$$y = \beta_0 + \beta_1 x_1 + \boxed{\beta_2 x_2 + ...\beta_{1,000,000} x_{1,000,000}} + \epsilon$$

$$p(y|\mathbf{x}, \beta_1, \beta_2, ..., \beta_{10^6})$$

$$p(\beta_2, ..., \beta_{10^6}) = \mathcal{N}(\beta_2, ..., \beta_{10^6}|\boldsymbol{\mu}, \Sigma)$$

$$p(y|\mathbf{x}, \beta_1) = \int p(y|\mathbf{x}, \beta_1 x_1, \beta_2 x_2, ..., \beta_{10^6} x_{10^6}) d\beta_2 ... d\beta_{10^6}$$

Conclusion

- Genetics and statistics are old friends (Galton, Fisher, Pearson,...), driven by the need of detecting relationships in nature without having access to mechanistic details

- The primer gives you the necessary concepts and results for the module

- The exam does not involve complicated proofs

- The MVN formulae will be provided