University of Zurich[UZH]

Institute of Molecular Life Sciences

STA 426: Statistical Analysis of High-Throughput Genomic and Transcriptomic Data

- Learning outcomes
- Administrative: course structure and organization, presentations
- Course materials: via github
- Intro to: {Unix, Bioconductor, Molecular Biology}

Mark D. Robinson, Statistical Genomics, IMLS

Today's structure

9.00-9.45: Ice Breakers, Surveys

10.00-10.45: Course structure, evaluations, Introduction to Molecular Biology (Hubert)

11.00-11.45: Troubleshooting computing/logins; Introduction to Bioconductor exercise

Survey 1: A bit of background on you

movo.ch

Token:

NA HY LY QY

University of Zurich[UZH]

Institute of Molecular Life Sciences

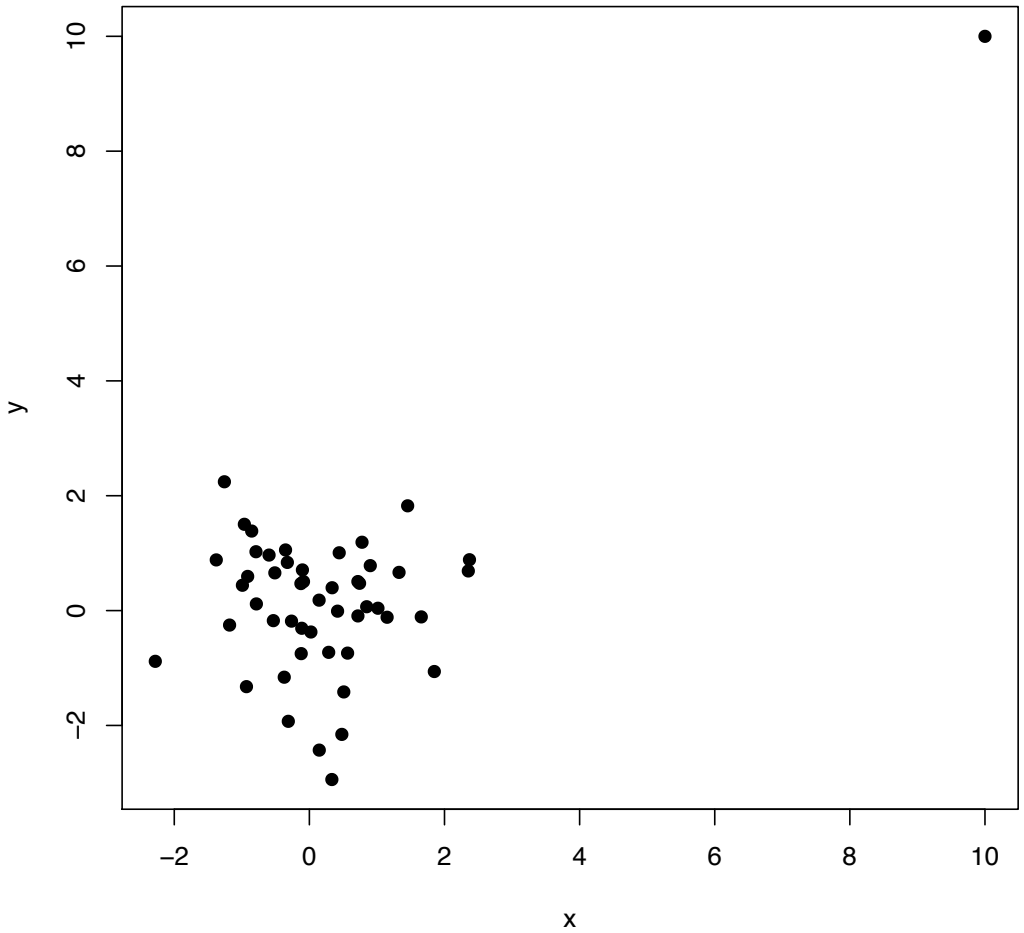
Survey 2: Statistical Insight

movo.ch

Token:

CI PY ZO SA

Question 1



Question 3

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Question 5

1 $$\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

2 $$\sum^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

3 $$\frac{(\overline{x}_1 - \overline{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

Rough structure of Monday mornings

We will run {9,10,11}.00-{9,10,11}.45

- Lecture/journal club presentation (9.00-whenever)
- Remaining time: in the computer lab (Y11-J-05) doing exercises/project

M.Sc. thesis projects

If you are:

- in a M.Sc. programme (ETHZ or UZH)
- have a solid background in mathematics / statistics
- have an interest in research in this field ("statistical bioinformatics")
- looking for a thesis project

→ Discuss a project in my lab

Critical skills needed by statisticians (Jeffrey Leek's words):

With all the excitement going on around statistics, there is also increasing diversity. It is increasingly hard to define "statistician" since the definition ranges from very mathematical to very applied. An obvious question is: what are the most critical skills needed by statisticians?

So just for fun, I made up my list of the top 5 most critical skills for a statistician by my own definition. They are by necessity very general (I only gave myself 5).

1. **The ability to manipulate/organize/work with data on computers** - whether it is with excel, R, SAS, or Stata, to be a statistician you have to be able to work with data.
2. **A knowledge of exploratory data analysis** - how to make plots, how to discover patterns with visualizations, how to explore assumptions
3. **Scientific/contextual knowledge** - at least enough to be able to abstract and formulate problems. This is what separates statisticians from mathematicians.
4. **Skills to distinguish true from false patterns** - whether with p-values, posterior probabilities, meaningful summary statistics, cross-validation or any other means.
5. **The ability to communicate results to people without math skills** - a key component of being a statistician is knowing how to explain math/plots/analyses.

Learning outcomes (in my words)

- Understand the fundamental "scientific process" in the field of Statistical Bioinformatics
- Be equipped with the skills/tools to preprocess genomic data (Unix, Bioconductor, mapping, etc.) and ensure reproducible research (markdown)
- Have a general knowledge of (some) **types** of data and **biological applications** encountered with high throughput genomic data
- Have the general knowledge of the range of statistical methods that get used with microarray and sequencing data
- Gain the ability to apply statistical methods/knowledge/software to a collaborative biological project
- Gain the ability to critical assess the statistical bioinformatics literature
- Write a coherent summary of a bioinformatics problem and it's solution in statistical terms

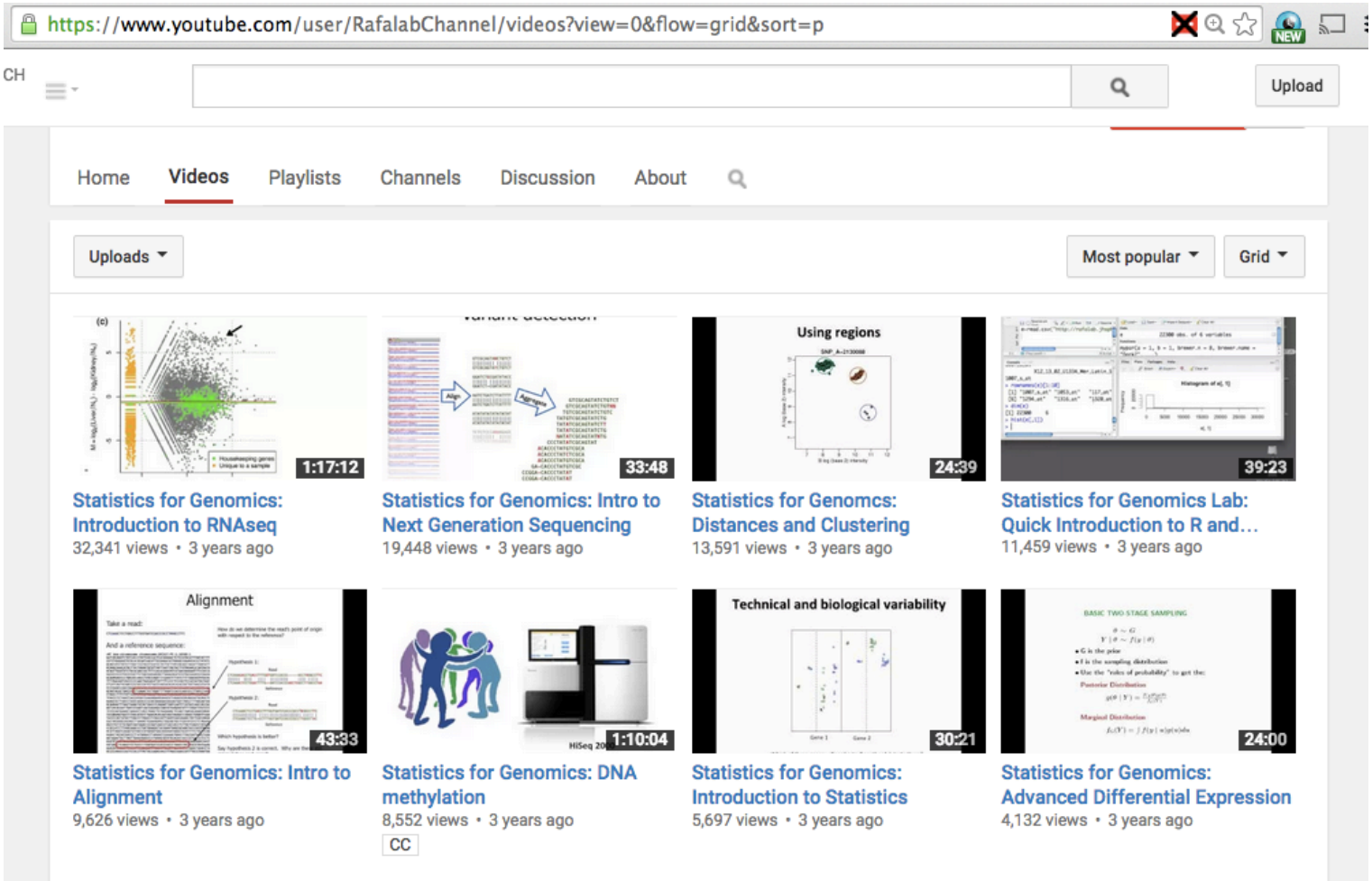
Course evaluation

1. Journal club presentation	20%
2. Project	50%
3. Exercises	30%
4. Technology day (participation)	0% or -10%

The semester-long course structure (subject to change)

Date	Lecturer	Content	Paper Title 1 (with link to PubMed)	Paper Presenter 1 (first and last name)	Paper Title 2 (with link to PubMed)	Paper Presenter 2 (first name and last name)
Mo 14.09.2015	Mark; Hubert	administrative structure Linux basics Bioconductor basics Molecular Biology basics: genome; genes; transcription; DNA binding; DNA modification; histone modifications				
Mo 21.09.2015	Hubert	exploratory data analysis: clustering, PCA, ... error types: FP, FN, power error rates: FPR, FDR, FWER				
Mo 30.09.2015	Hubert; Mark	technologies: RNA, DNA variants, de novo, meth, chip-seq				
Mo 05.10.2015	Mark	limma				
Mo 12.10.2015	Mark	beyond limma				
Mo 19.10.2015	Mark	NGS intro; intro to mapping				
Mo 26.10.2015	Hubert	more on mapping				
Mo 02.11.2015	Hubert	RNA-seq quantification				
Mo 09.11.2015	Mark	differential counts				
Mo 16.11.2015	Hubert	more on differential counts				
Mo 23.11.2015	Mark	isoform switching				
Mo 30.11.2015	Mark	epigenomics, DNA methylation				
Mo 07.12.2015	Mark	ChIP; GeneSet Analysis				
Mo 14.12.2015	Hubert	Classification				

Trial with flipped classroom using Statistics for Genomics MOOCs



Expectations: journal club presentation

- 20 minutes (+5 minutes discussion)
- MUST be a paper about a **statistical method in genomics** paper + MUST be approved by Mark/Hubert
- Should describe the biological context
- Should describe the (new) model used
- Should describe comparisons to existing methods
- Should not be one of the papers discussed in detail in lectures: limma, edgeR, DEXSeq, etc.

Expectations: project

- ~10-15 page report, with R code in line (e.g. **knitR**)
- Describe the biological setting, statistical analysis, exploratory analysis with publication-quality graphics embedded
- Three possibilities:
  - Comparison of statistical methods (simulation/independent reference data + metrics)
  - Reproduce an analysis from a paper from the raw data
  - (NEW in 2015!) Real collaborative project with FGCZ or a local laboratory
- Be strategic: work on something related to your interests!

Soft technical skills needed (developed) in this course …

- Use unix-like operating system to run command-line programs
- Options are:
  - Use your own Linux/MacOSX computer; N.B.: you may be able to do everything from Windows (e.g., cygwin), but we will not help with this
  - Use the Macs in Y11-J-05
- R: from the command line or R studio; know how to get help; how to make plots in R, pipe them to a file
- knitr/Rmarkdown
- Bioconductor – www.bioconductor.org

Main resource to make this work – Jenny Bryan's course at UBC:
1. http://stat545-ubc.github.io/bit004_stat545-use-of-github.html
2. We will work it out together (new to us as well!)

(NEW for 2015!) All submissions occur via github

Homework for today (part 1):

1. Acquaint yourself with the idea of github [1]
2. Create a github account at github.com
3. Make sure you know to check in / check out files (git clone ..) from the command line or from an app [2]
4. Create a repository and a README.md (learn a bit of markdown [3]) in a public repository and add some text
   - Include an image
   - Include a web link

[1] https://gist.github.com/andrewpmiller/9668225
[2] https://confluence.atlassian.com/stash/basic-git-commands-278071958.html
[3] http://markdowntutorial.com/

Rmarkdown / knitR for executable documents / reproducibility

Homework for today (part 2):

1. Acquaint yourself with **knitR** PDF/HTML Rmarkdown documents [1], perhaps both in R studio and from command prompt
2. Create an HTML/PDF document that samples 100 values from a log-normal distribution (say, mu=1, sigma=.25); create a histogram of the distribution and the distribution on the log scale; report the mean and variance of the sample in line in the text.
   - Do not just dump the R code and plots in the HTML/PDF document; add some text and headings to give a full explanation (i.e., the document should be self-explanatory)