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:

DE ZA NY GY

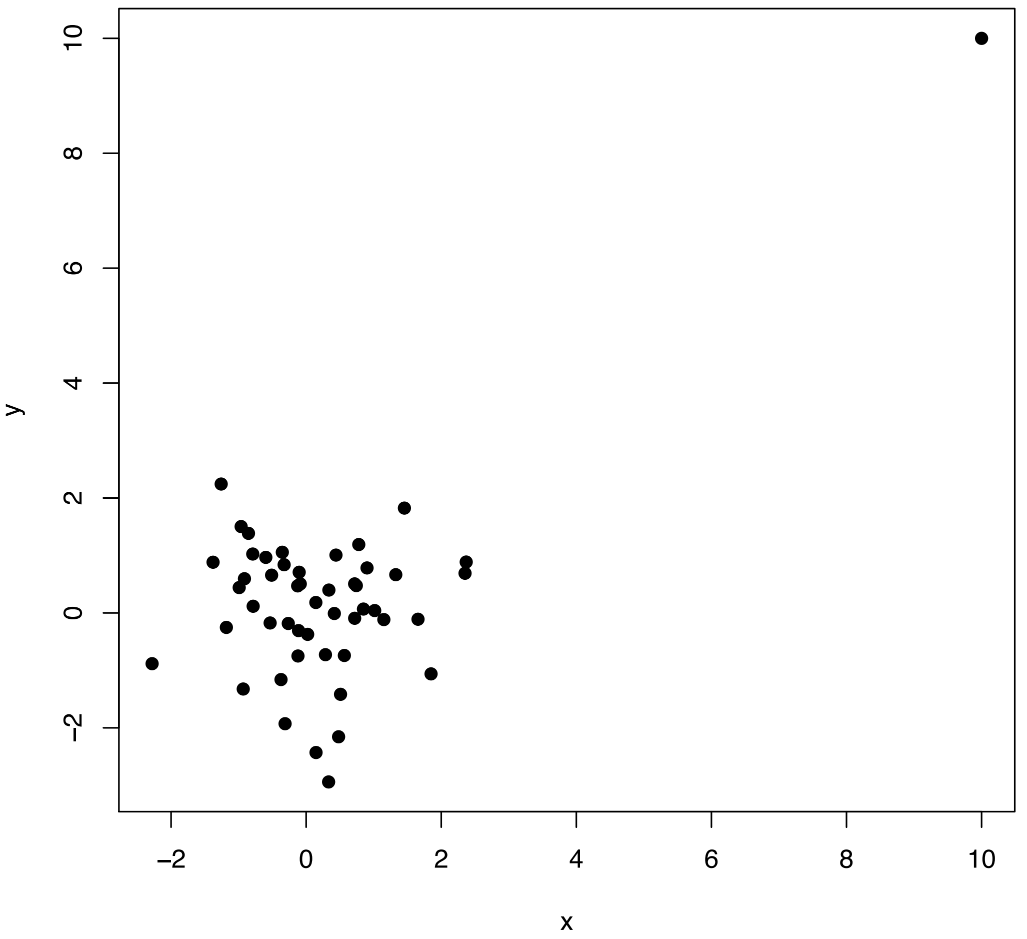
Survey 2: Statistical Insight

movo.ch

Token:

GO CI DO RA

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 X.00-X.45; X in {9,10,11}

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

M.Sc. thesis projects

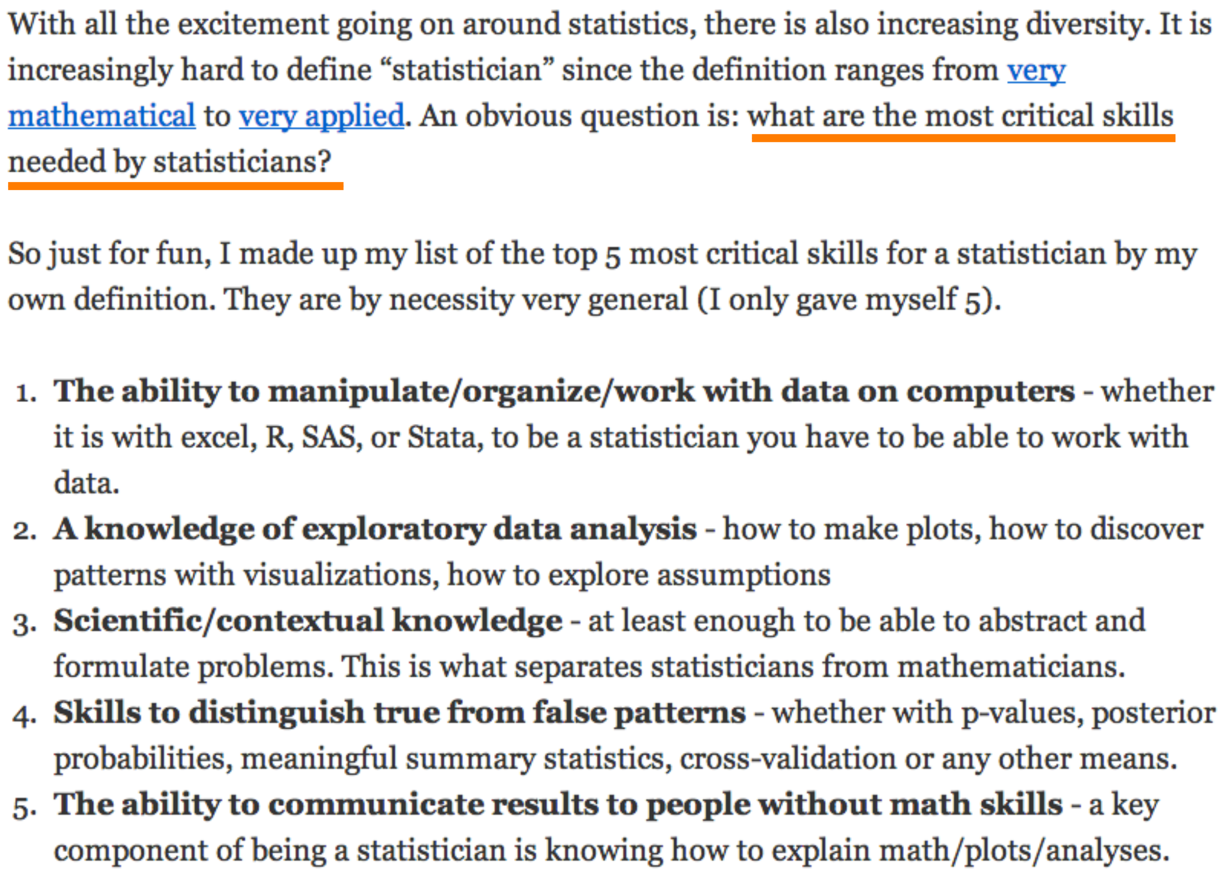
If you are:

- 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

http://simplystatistics.org/2013/07/03/repost-the-5-most-critical-statistical-concepts/

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.

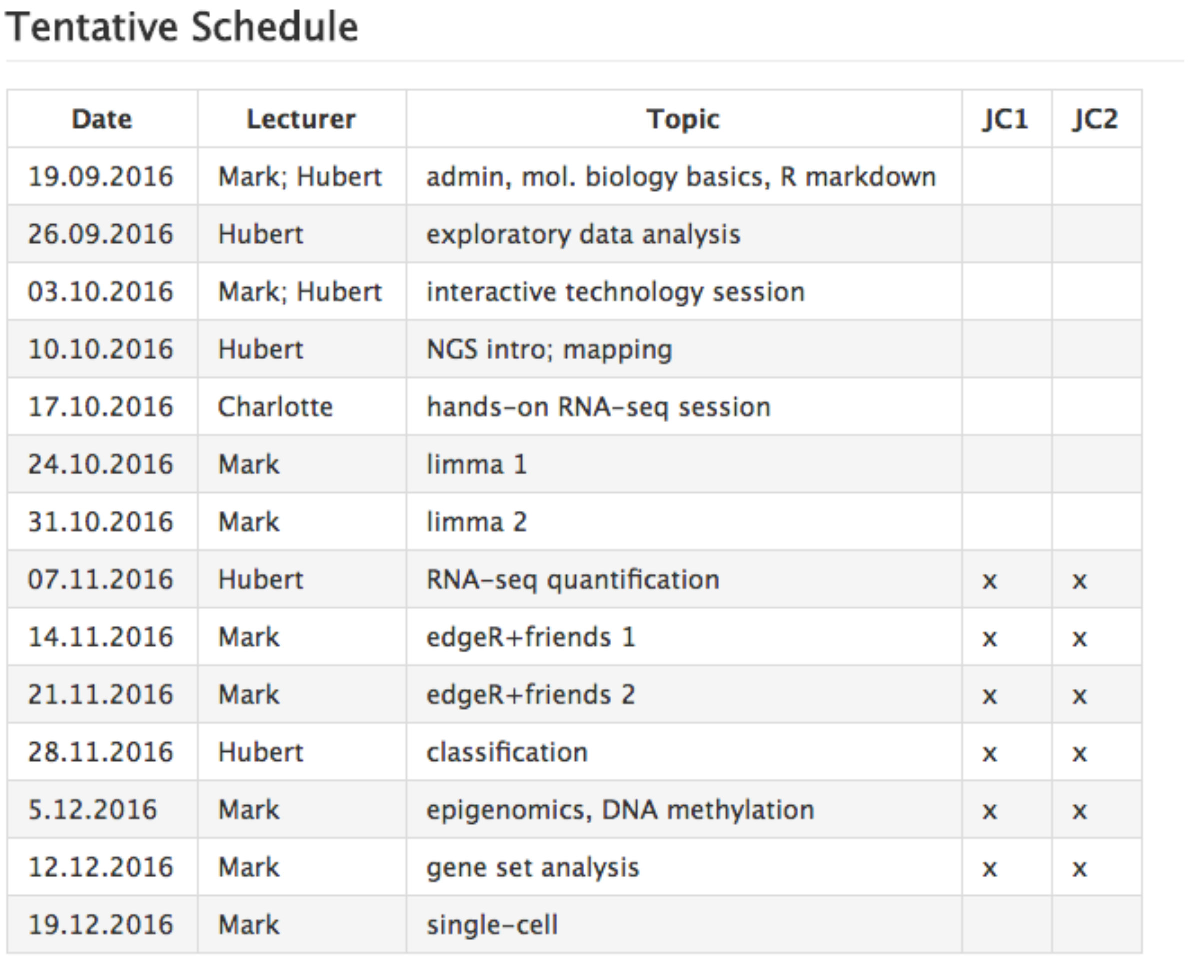
Learning outcomes (in my words)

- 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 (R/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)



Tentative Schedule

Date	Lecturer	Topic	JC1	JC2
19.09.2016	Mark; Hubert	admin, mol. biology basics, R markdown		
26.09.2016	Hubert	exploratory data analysis		
03.10.2016	Mark; Hubert	interactive technology session		
10.10.2016	Hubert	NGS intro; mapping		
17.10.2016	Charlotte	hands-on RNA-seq session		
24.10.2016	Mark	limma 1		
31.10.2016	Mark	limma 2		
07.11.2016	Hubert	RNA-seq quantification	x	x
14.11.2016	Mark	edgeR+friends 1	x	x
21.11.2016	Mark	edgeR+friends 2	x	x
28.11.2016	Hubert	classification	x	x
5.12.2016	Mark	epigenomics, DNA methylation	x	x
12.12.2016	Mark	gene set analysis	x	x
19.12.2016	Mark	single-cell		

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
  - 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)
  - 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:
http://stat545-ubc.github.io/bit004_stat545-use-of-github.html

All submissions occur via github

Main repo for course: https://github.com/sta426hs2016/material

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 new public repository, add a README.md (learn a bit of markdown [3]) and add some content
   - Include an image
   - Include a web link
   - add an issue to the materials repo to let me know that you've done it (https://github.com/sta426hs2016/material/issues)
   - (you can delete the repo after, if you want)

[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 and make it into a readable story (i.e., the document should be self-explanatory)

A Slack channel (trial for 2016)

https://sta426hs2016.slack.com/messages/general/

(Or, you can use the Slack app)