STA 426 First Lecture

- survey
- course structure
- Some thoughts on Molecular Biology (Hubert's slides given by Mark)
- Exercise 1

Mark D. Robinson, Statistical Bioinformatics

Today's structure

9.00-9.45: Survey + Course Structure (Mark)

10.00-10.45: Introduction to Molecular Biology (Hubert)

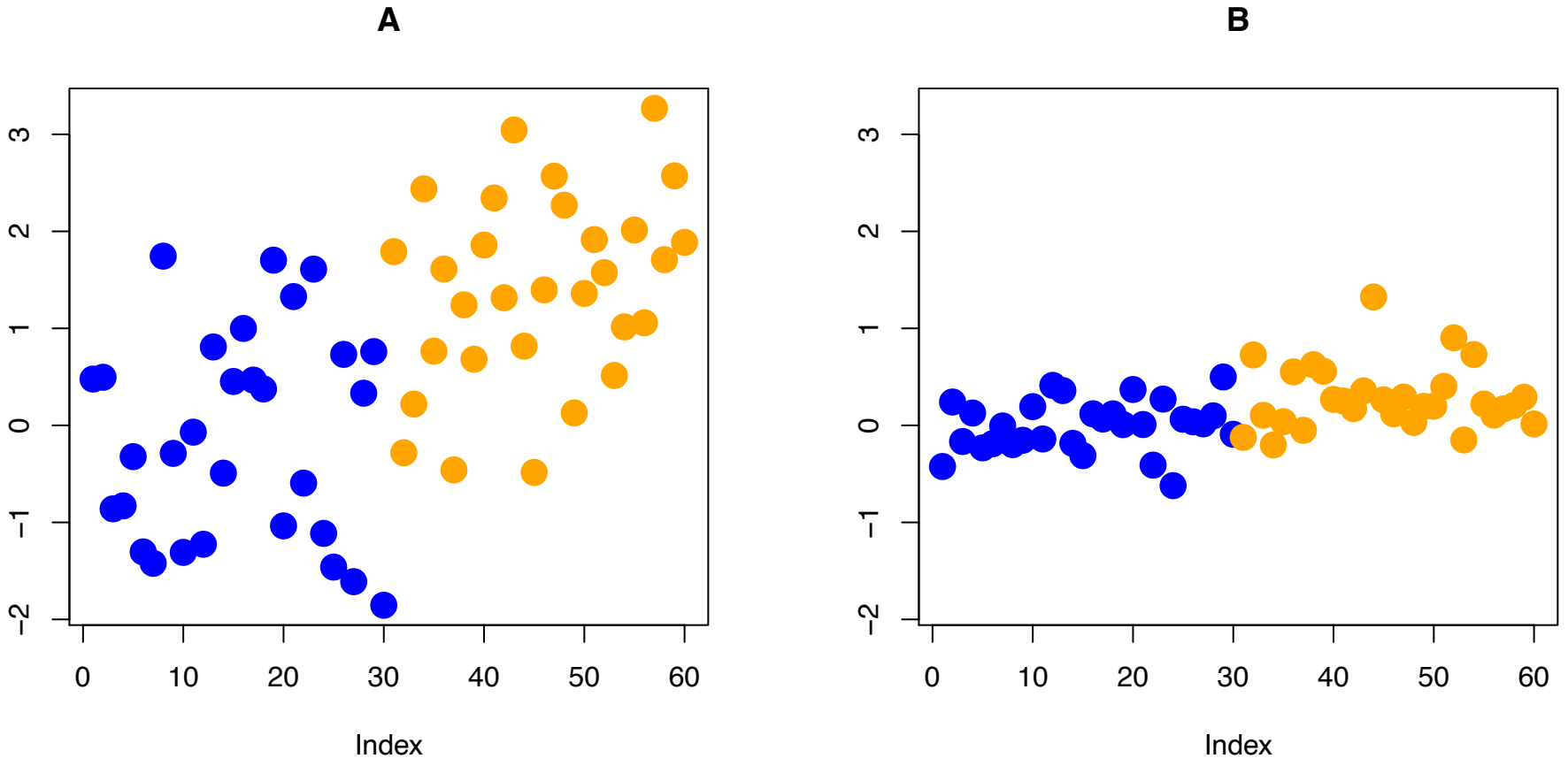
11.00-11.45: Quarto + Exercise 1

https://pwa.klicker.uzh.ch/join/marobi

Survey: Statistical Insight

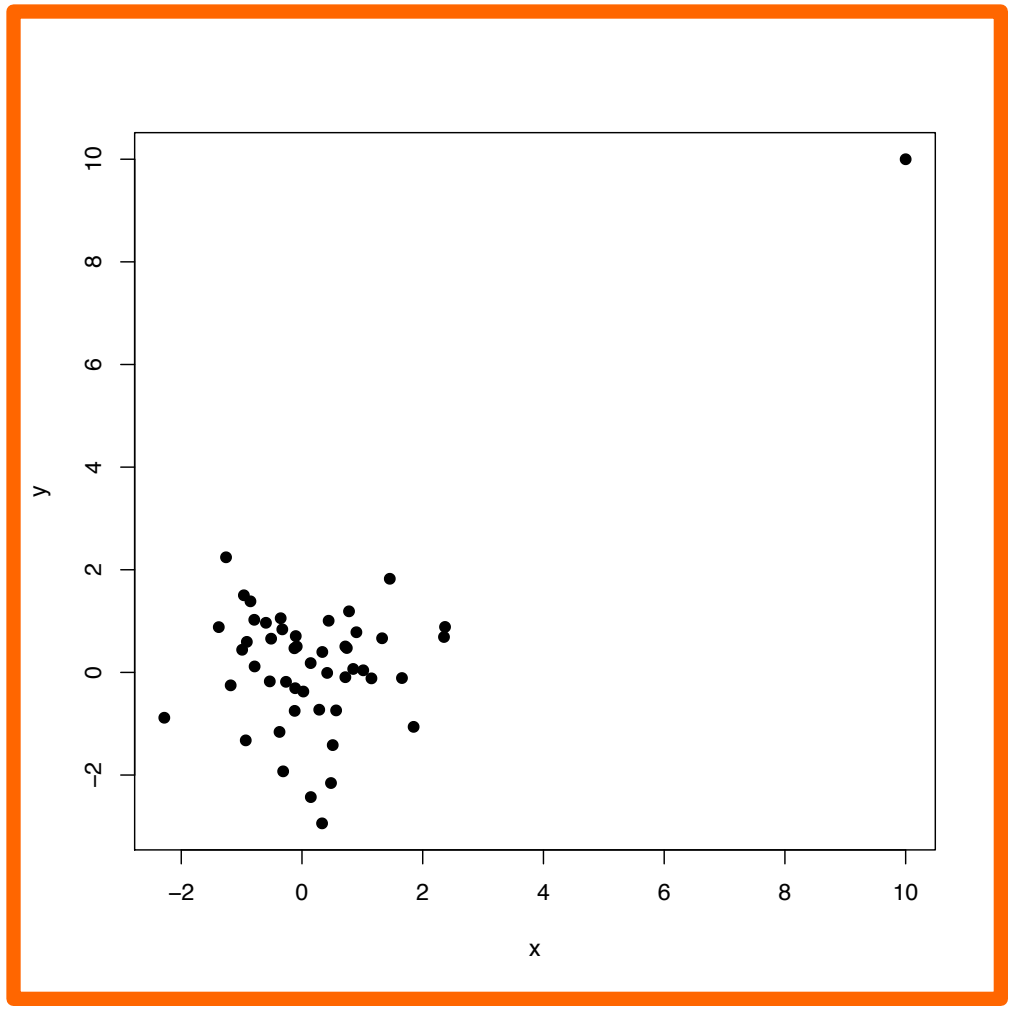
https://pwa.klicker.uzh.ch/join/marobi

Question 1: Which plot highlights more (statistical) evidence for a change in the population means (between orange and blue)?



https://pwa.klicker.uzh.ch/join/marobi

Question 2: In your view, what best describes the associations shown in the plot of 'x' and 'y' ?



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

https://pwa.klicker.uzh.ch/join/marobi

Question 3: Of these equations, which one resembles the standard two sample t-test ?

1

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

2

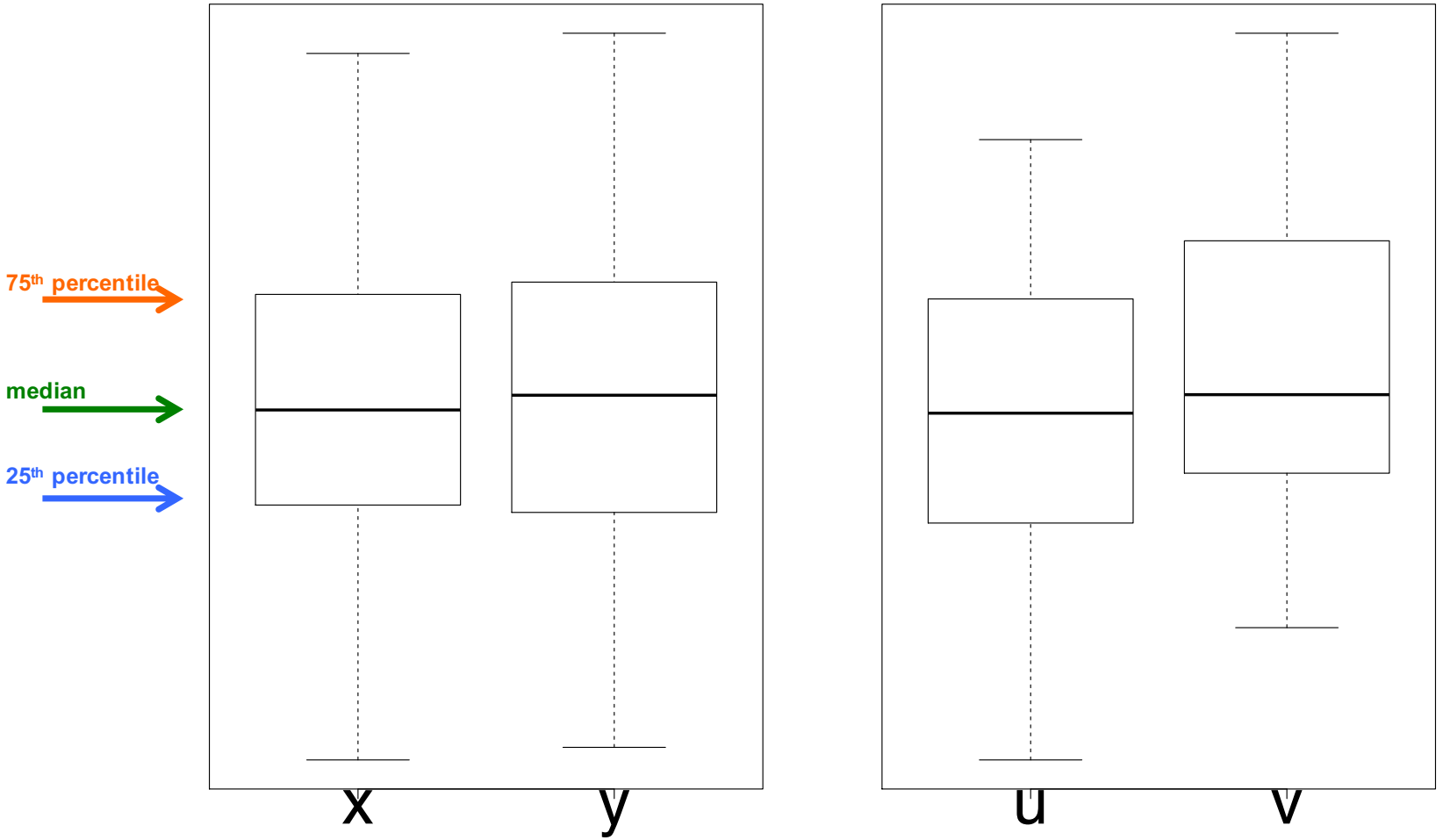
$$\sum^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

3

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

https://pwa.klicker.uzh.ch/join/marobi

Question 4: Given these boxplots, which of two underlying distributions are more similar?



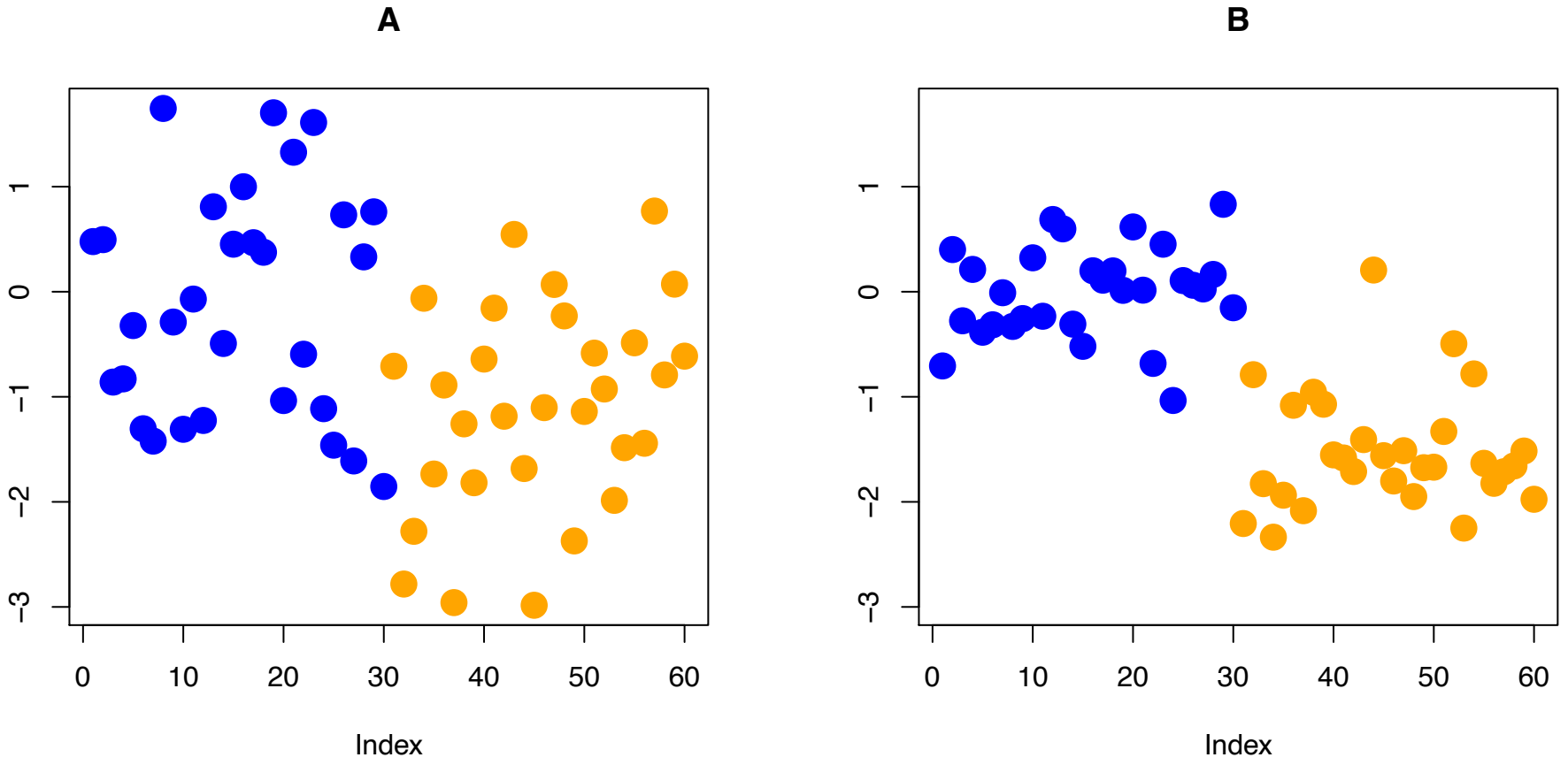
https://pwa.klicker.uzh.ch/join/marobi

Question 6: Given this design matrix, describe the experimental design.

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

https://pwa.klicker.uzh.ch/join/marobi

Question 8: Which plot highlights more (statistical) evidence for a change in the population means (between orange and blue)?



Course communication

- Video situation: we may have recordings, but no live stream
- Slack: all communication on Slack (note: all invitations were sent to UZH email addresses)
- Except for exceptional circumstances, **no emails please**; communicate on Slack
- Slack policy: unless really private, ask questions in a public channel (please note: *good questions get good answers*); use threads when relevant; good manners/behaviours are expected

Course evaluation

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

Rough structure of lecture/exercise time

Monday mornings: we will run X.00-X.45; X in {9,10,11}

- Lectures and Exercises
- Lecture/journal club presentation (9.00-whenever)
- Remaining time: free (can be used to work on exercises; we are available for questions)

M.Sc. thesis projects

If you are:

- in a M.Sc. programme (ETHZ or UZH)
- have a solid background/experience in mathematics / statistics / computation
- 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.

STA 426 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 / quarto)
- 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

Use your AI tool of choice

- Goal for this year's course: let's chart out where ChatGPT (or similar) works well and where it doesn't
- In particular, in Exercise 2 (technologies); document where/how you use it
- ChatGPT and variants often hallucinate and given non-functioning code, but can be useful to get an overall structure and it's becoming better all the time
- At the moment, I use https://www.perplexity.ai/ (free version) and you might find it useful too. Please share useful tools with your colleagues.

The semester-long course structure (subject to change)

Date	Lecturer	Topic	Exercise	JC1	JC2
16.09.2024	Mark	admin; mol. bio. basics	quarto; git(hub)		
23.09.2024	Mark	interactive technology/statistics session	group exercise: technology pull request		
30.09.2024	Mark	limma + friends	linear model simulation + design matrices		
07.10.2024	Hubert	NGS intro; exploratory data analysis	EDA in R		
14.10.2024	Hubert	mapping	Rsubread		
21.10.2024	Hubert	RNA-seq quantification	RSEM	X	X
28.10.2024	Mark	edgeR+friends 1	basic edgeR/voom	X	X
04.11.2024	Mark	edgeR+friends 2	advanced edgeR/voom	X	X
11.11.2024	Mark	hands-on session #1: RNA-seq	FASTQC/Salmon/etc.	X	X
18.11.2024	Hubert	single-cell 1: preprocessing, dim. reduction, clustering	clustering	X	X
25.11.2024	tba	hands-on session #2: cytometry	cytof null comparison	X	X
02.12.2024	Mark	single-cell 2: clustering, marker gene DE	marker gene DE	X	X
09.12.2024	tba	hands-on session #3: single-cell RNA-seq (cell type definition, differential state)	full scRNA-seq pipeline	X	X
16.12.2024	Mark	spatial omics	spatial statistics	X	X

Hands-on sessions

Date	Lecturer	Topic	Exercise	JC1	JC2
16.09.2024	Mark	admin; mol. bio. basics	quarto; git(hub)		
23.09.2024	Mark	interactive technology/statistics session	group exercise: technology pull request		
30.09.2024	Mark	limma + friends	linear model simulation + design matrices		
07.10.2024	Hubert	NGS intro; exploratory data analysis	EDA in R		
14.10.2024	Hubert	mapping	Rsubread		
21.10.2024	Hubert	RNA-seq quantification	RSEM	X	X
28.10.2024	Mark	edgeR+friends 1	basic edgeR/voom	X	X
04.11.2024	Mark	edgeR+friends 2	advanced edgeR/voom	X	X
11.11.2024	Mark	hands-on session #1: RNA-seq	FASTQC/Salmon/etc.	X	X
18.11.2024	Hubert	single-cell 1: preprocessing, dim. reduction, clustering	clustering	X	X
25.11.2024	tba	hands-on session #2: cytometry	cytof null comparison	X	X
02.12.2024	Mark	single-cell 2: clustering, marker gene DE	marker gene DE	X	X
09.12.2024	tba	hands-on session #3: single-cell RNA-seq (cell type definition, differential state)	full scRNA-seq pipeline	X	X
16.12.2024	Mark	spatial omics	spatial statistics	X	X

Expectations: **journal club** presentation

- 20-25 minutes (+5 minutes discussion)
- MUST:
  - ➡ be a paper about a **statistical** method in bioinformatics
  - ➡ be approved by Mark/Hubert
- Should:
  - ➡ describe the biological context and/or data collected
  - ➡ describe the (new) model used
  - ➡ describe comparisons to existing methods
- Should not:
  - ➡ be one of the papers discussed in detail in lectures: limma, edgeR, DEXSeq, etc.

- part of the JC grade: giving feedback to fellow students (online feedback forms)

Expectations: **project**

- ~10-15 page report, with R code in line (e.g. **quarto**)
- Describe the biological setting, statistical analysis, exploratory analysis with publication-quality graphics embedded
- Three possibilities:
  - Comparison of statistical methods (simulation / 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!
- Typically due at end of first working week of January

Expectations: **exercises**

- There will be an exercise **every** week
- Across 14 weeks, the best 9 exercises are counted towards the 30%

Soft technical skills needed (developed) in this course …

- **Data Science!**
- Use unix-like operating system to run command-line programs
- Options:
  - use your own computer (if Windows, use cygwin / gitbash / etc.)
  - use cloud: renkulab.io
- R: from the command line or RStudio / Posit (https://posit.co/); getting help; creating workflows; how to make publication-quality graphics (ggplot2); knitr / Rmarkdown / quarto
- Bioconductor – www.bioconductor.org
- (Python)
- git/github
- bioconda / Docker

Hubert's slides given by me.

Demos:

- Slack
- git/github
- quarto.org

Quick intro to Git/Github (version control)

```
git clone
git pull
git status
git branch
git commit
git add
git checkout
git push
```

https://www.simplilearn.com/tutorials/git-tutorial/git-commands

Exercise 1
Part a: GitHub
Part b: quarto

Note: all exercise submissions occur via GitHub-classroom

Exercise 1 Part A:

1. If you have note already, install R 4.4.1 (https://www.r-project.org/), RStudio (https://posit.co/download/rstudio-desktop/), git (https://github.com/git-guides/install-git), quarto (https://quarto.org/docs/get-started/).
2. If you have not already, create an account at https://github.com/; share your GitHub username with Mark via https://forms.gle/BrYozKiuvKVbwziy9
3. Acquaint yourself with git / github (gitlab) [1] (recommendation: use command line; but there are apps too); make sure you can check in (`push`) to a personal repository and check out (`pull/clone`) files from a repository.
4. Create your (private) Exercise 1 repository using GitHub-classroom: https://classroom.github.com/a/_yPLR4vK. Add a `README.md` file to this repository and put your name and matriculation number in the file.
5. Add an Issue to the 'material' repo [3] with a link to your repo.

[1] https://confluence.atlassian.com/stash/basic-git-commands-278071958.html
[2] https://quarto.org/docs/get-started/hello/rstudio.html
[3] https://github.com/sta426hs2024/material

Quarto for executable documents / reproducibility

Exercise 1 Part B:

1. Test your R knowledge here: https://forms.gle/wueUwbQt2eG8rP9t7 (only 9 questions)
2. Acquaint yourself with quarto for building executable documents [1].
3. Using quarto and R, create an executable HTML document with R code that solves Roger Peng's Coursera selfquiz: https://www.biostat.jhsph.edu/~rpeng/coursera/selfquiz/quiz.html
4. Add both the QMD and HTML files to the repo you made in Part A.

[1] https://quarto.org/