STA 426 First Lecture
- ice breakers
- survey
- course structure
- Molecular Biology lecture (Hubert)
- R survey + computing + Exercise 1

Mark D. Robinson, Statistical Bioinformatics

Today's structure

9.00-9.45 (Y27-H-46): Ice Breakers + Survey + Course Structure (Mark)

10.00-10.45 (Y27-H-46): Introduction to Molecular Biology (Hubert)

11.00-11.45 (Y01-F-50): Computing + R quiz + Rmarkdown exercise
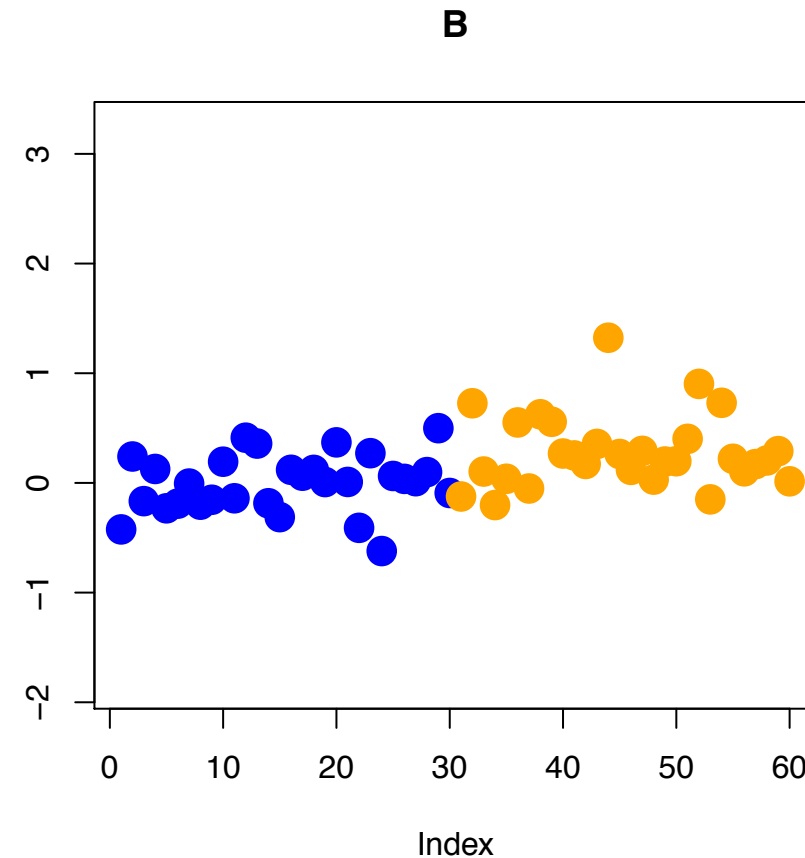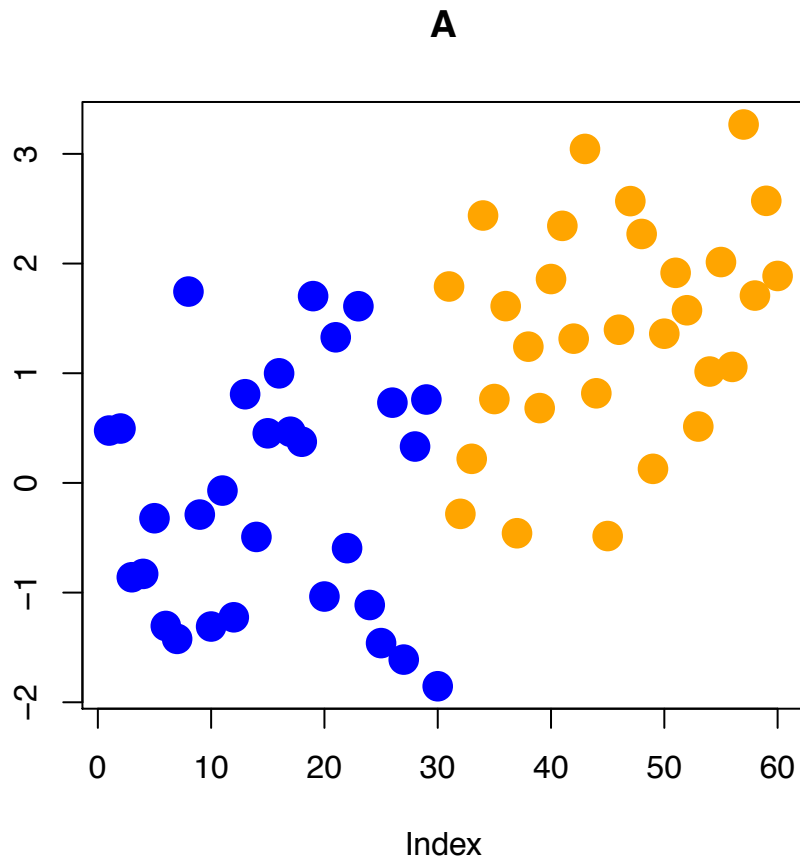
Ice breakers

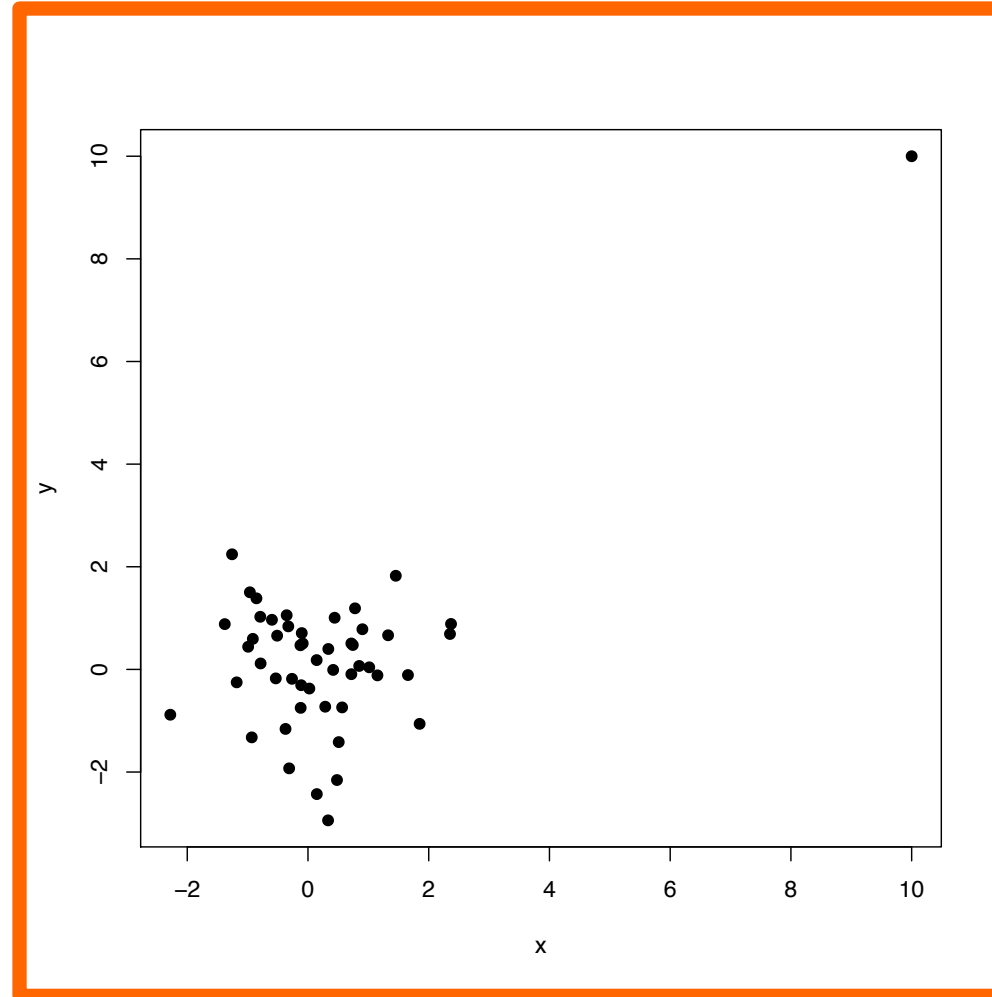# Survey: Statistical Insight

movo.ch

Token: TA XE LO QO

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

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

Question 4: Given this design matrix, describe the experimental design.

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$
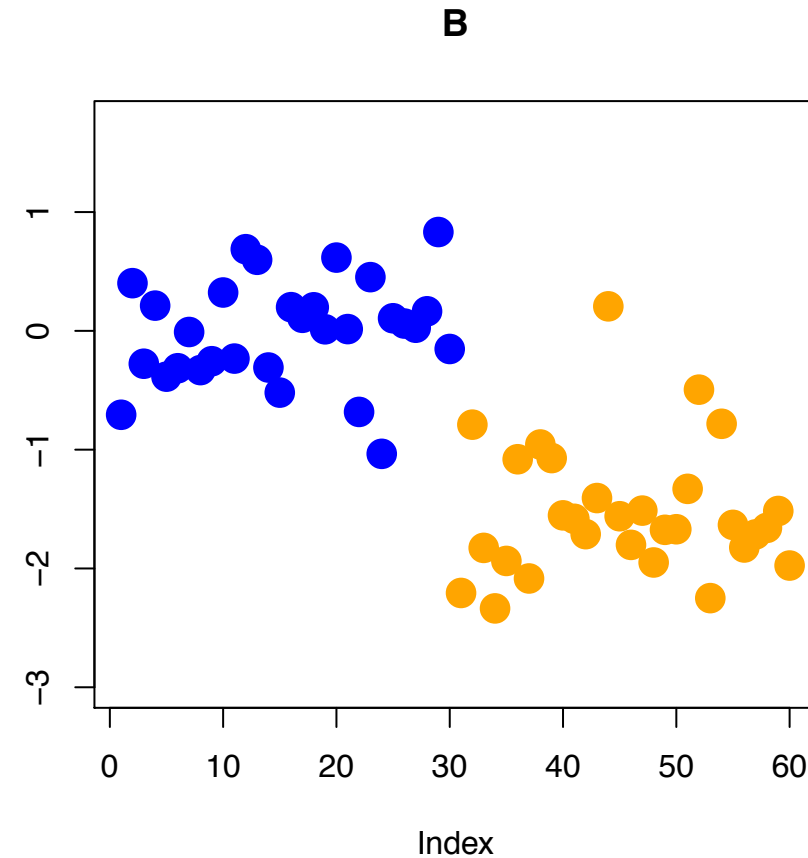
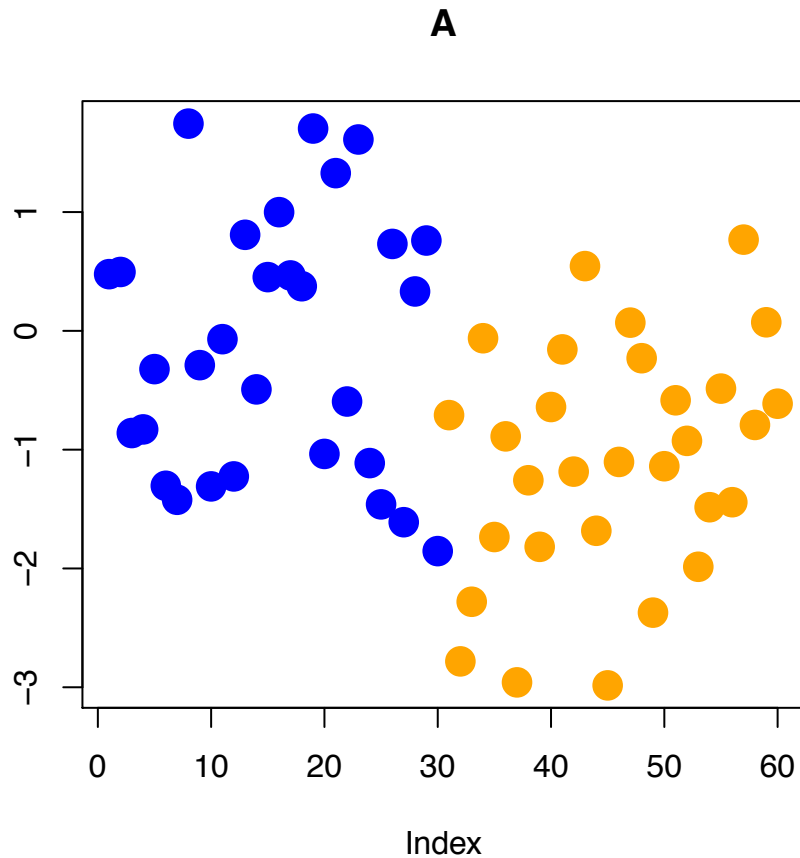Question 6: Of these equations, which one resembles the standard two sample t-test ?

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

2 $\quad \displaystyle\sum_{}^{k} \dfrac{(\text{observed} - \text{expected})^2}{\text{expected}}$

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

Question 7: Which plot highlights more (statistical) evidence for a change in the population means (between orange and blue)?

Question 8: Given these boxplots, which of two underlying distributions are more similar?

75th percentile →

median →

25th percentile →

x   y   u   v
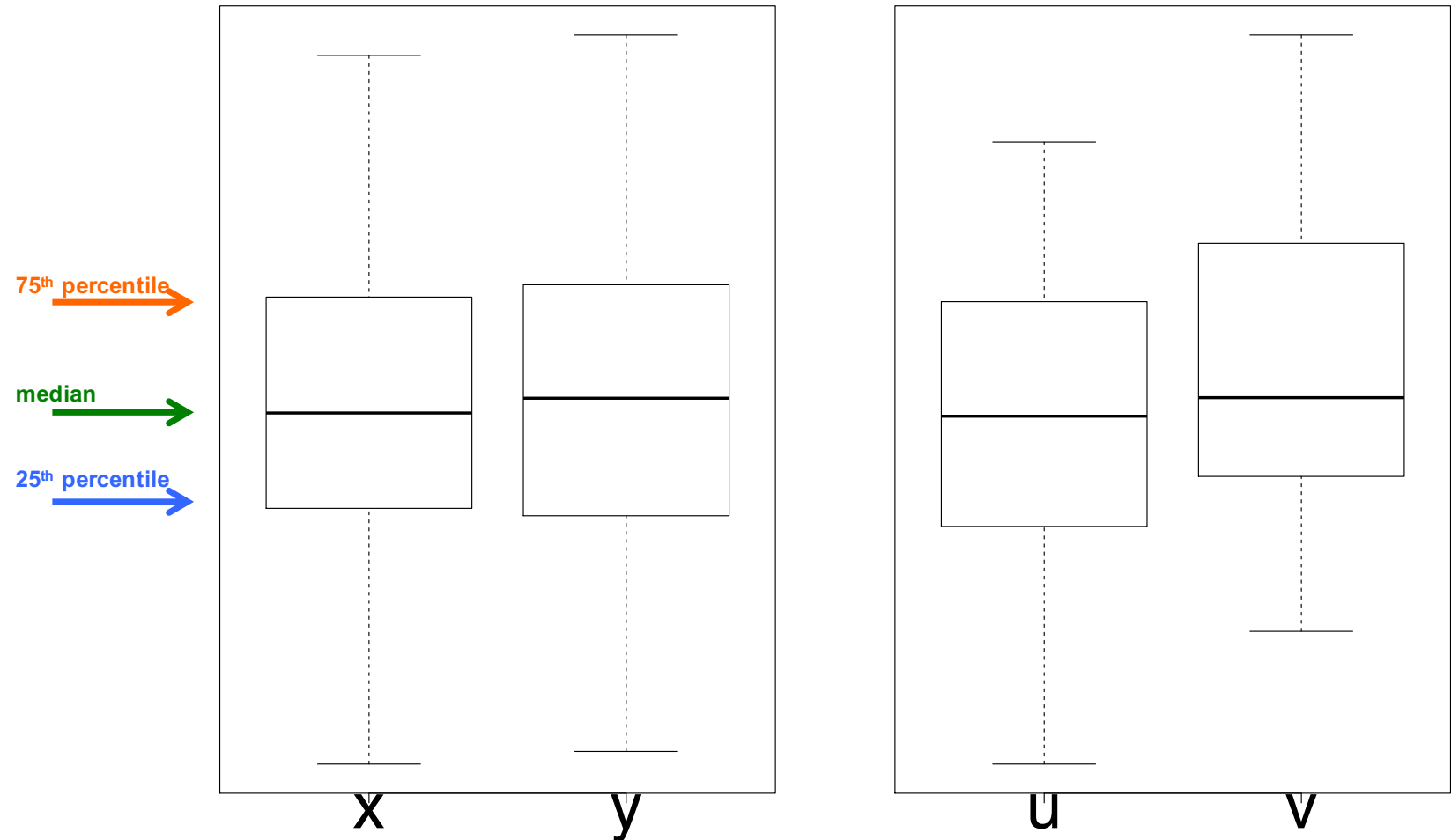
Rough structure of lecture/exercise time

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

- Lectures: Y27-H-46

- Exercises: Y01-F-50

- Lecture/journal club presentation (9.00-whenever)

- Remaining time: in the computer lab (Y01-F-50) doing exercises/project

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

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.

13

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)

Date	Lecturer	Topic	Exercise	JC1	JC2
16.09.2019	Mark + Hubert	admin; mol. bio. basics	R markdown; git(hub)		
23.09.2019	Mark	interactive technology/statistics session	group exercise: technology pull request		
30.09.2019	Hubert	NGS intro; exploratory data analysis	EDA in R		
07.10.2019	Hubert	mapping	Rsubread		
14.10.2019	Mark	limma + friends	linear model simulation + design matrices		
21.10.2019	Hubert	RNA-seq quantification	RSEM		
28.10.2019	Kathi	hands-on session #1: RNA-seq	FASTQC/Salmon/etc.	X	X
04.11.2019	Mark	edgeR+friends 1	basic edgeR/voom		
11.11.2019	Mark	edgeR+friends 2	GLM/DEXSeq		
18.11.2019	Hubert	single-cell 1: preprocessing, dim. reduction, clustering	scRNA exercise 1		
25.11.2019	YYY	hands-on session #2: cytometry	cytof null comparison	X	X
02.12.2019	Mark	single-cell 2: cell type definition, differential state	scRNA exercise 2		
09.12.2019	Pierre-Luc	hands-on session #3: single-cell RNA-seq	full scRNA-seq pipeline	X	X
16.12.2019	Mark	loose ends: HMM, EM, robustness	segmentation, peak finding		

Expectations: **journal club** presentation

- 20-25 minutes (+5 minutes discussion)
- MUST:
  - ➡ be a paper about a **statistical** method in genomics
  - ➡ 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.
- (since 2017) feedback forms from fellow students

Expectations: **project**

- ~10-15 page report, with R code in line (e.g. **knitR** / **Rmarkdown**)
- 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, cygwin)
  - use renkulab.io
  - (use Macs in 01-F-50)
- R: from the command line or R studio; getting help; creating workflows; how to make publication-quality graphics (ggplot2); knitr/Rmarkdown
- Bioconductor – www.bioconductor.org
- git/github
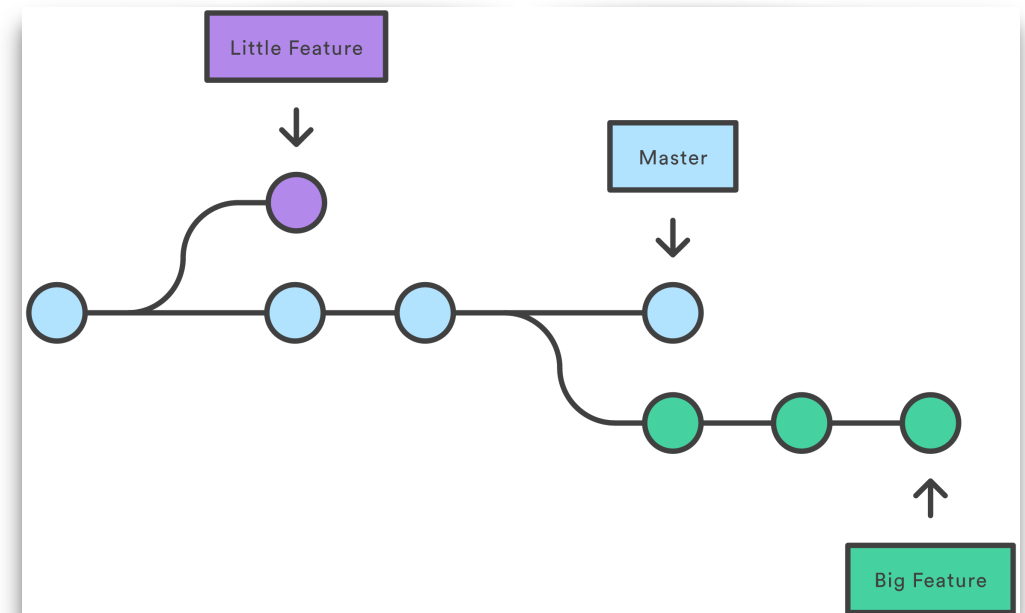- bioconda/Docker (cloud computing)

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

Quick intro to Git/Github (version control)

Branching

https://blog.seibert-media.net/blog/2015/07/31/git-mit-branches-arbeiten-git-branch/
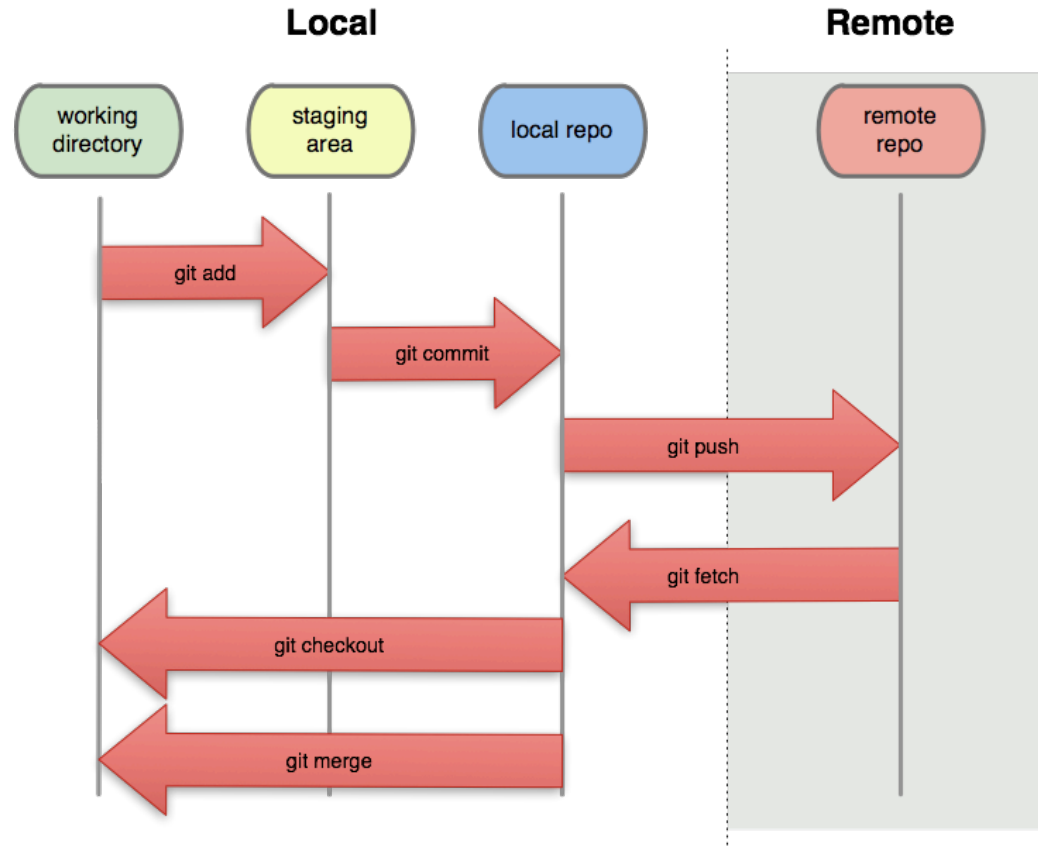
Surveys

1. R
2. Computing

Exercise 1
Part a: GitHub
Part b: knitR/Rmarkdown

All homework submissions occur via github

Week 1 Exercise (part a):

1. Create an account at github.com; give userid to Mark via Slack

2. Acquaint yourself with git/github [1]; make sure you can check in / out files from command line or app [2]

3. Create a new public repository, add a README.md (using markdown [3]) and add some content; include an image; include a web link, etc.

4. Add an Issue to the 'material' repo [4] with a link to your repo (you can delete the repo after I've closed the issue, 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/
[4] https://github.com/sta426hs2019/material

Rmarkdown / knitR for executable documents / reproducibility

Week 1 Exercise (part b):

1. Acquaint yourself with knitR PDF/HTML Rmarkdown documents [1].

2. Create an HTML 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. (In general, do not just dump the R code and plots in the HTML document; break your exercises into sections with headings, add some text and make it readable.

[1] https://www.r-bloggers.com/r-markdown-and-knitr-tutorial-part-1/