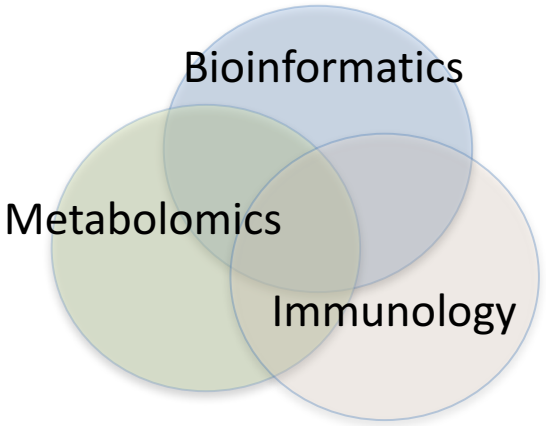
A bioinformatics primer to data science



Shuzhao Li, Ph.D
Assistant Professor
Department of Medicine
Emory University
August 27, 2019

shuzhao.li@gmail.com shuzhao-li@GitHub

Computing environment and setup

- Desktops, laptops, clusters and the cloud
- Operating systems for bioinformatics
- UNIX / Linux
- GUI vs the command line
- Compute languages

  Python, R, C, Java, Javascript

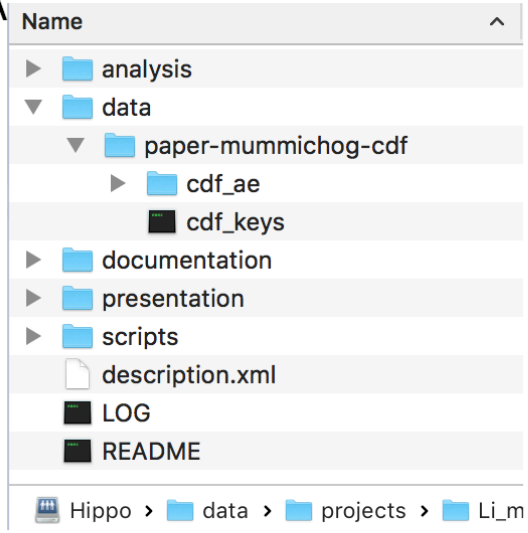
Data management

- Do not stockpile code, data, and spreadsheets into a single folder. Create meaningfully sub folders up front to contain highly related information even if you do not yet have a significant amount of information to manage.
- Use meaningful file names, okay using long names with date stamp, version numbers.
- Place data on a volume that is backed up or replicated to an off-premise site
- Use notebook tools
  Jupyter Notebook (https://jupyter.org/),
  knitr (https://yihui.name/knitr)

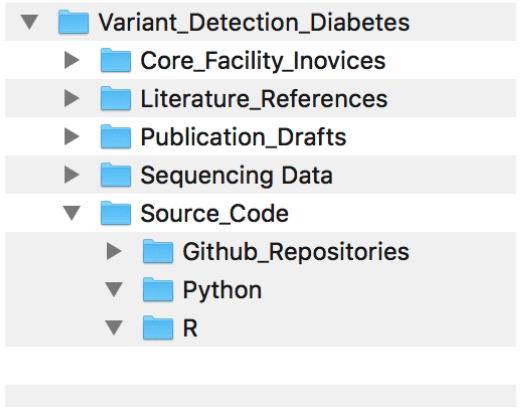
Slide modified from Steve Pittard

File tree examples

A



B



C

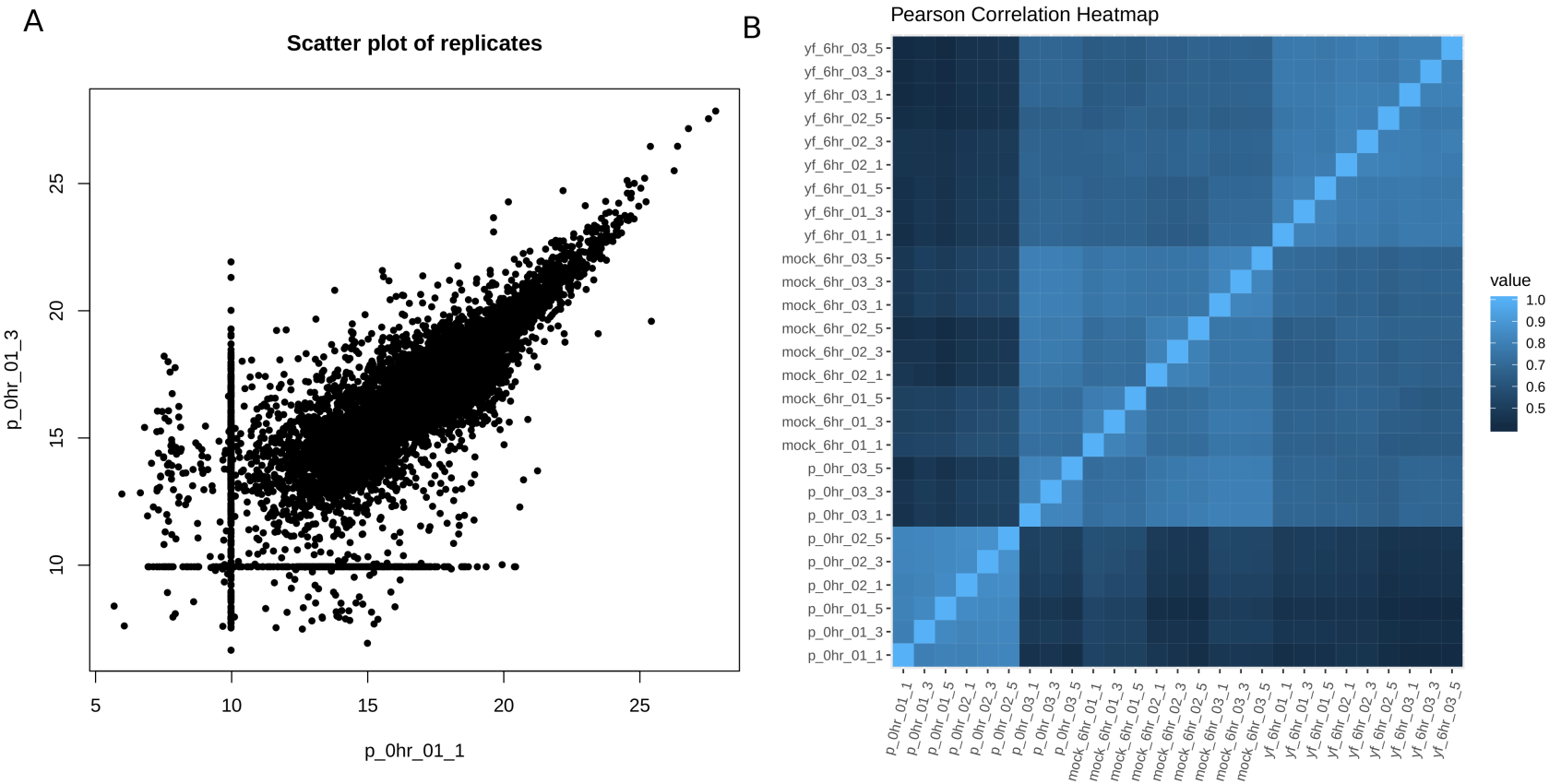
```
➜ manuscript_expometabo_pls_201808 tree -L 2
.
├── Bibliography__.docx
├── ExpoMetabo_Shuzhao_20180910_v1.docx
├── Fig1
│   ├── communities_expo_Fig1.pdf
│   ├── communities_expo_Fig1.pptx
│   └── communities_expo_Fig1_PFAS.pdf
├── Fig2
│   └── Fig2-E_M_pls_network_20180823.tiff
├── Fig3
│   ├── Fig3-20180825.pdf
│   ├── Fig3-20180825.pptx
│   ├── Fig3-20180908.pdf
│   ├── Fig3a.png
│   └── edit.mcg_pathwayanalysis_M_4.xlsx
├── Fig4
│   ├── Bubble-plot-LOG.txt
│   ├── Fig4-20180825.pdf
│   ├── Fig4-20180825.pptx
│   ├── Fig4.png
│   ├── Mclus.order.txt
│   ├── good-bubbles-0.01-v1.pdf
│   ├── good-bubbles-0.05.pdf
│   ├── good-bubbles-v2.pdf
│   ├── good-bubbles-v3.pdf
│   ├── mcg_pathwayanalysis_M_1.tsv
│   ├── mcg_pathwayanalysis_M_12.tsv
│   ├── mcg_pathwayanalysis_M_13.tsv
│   ├── mcg_pathwayanalysis_M_2.tsv
│   ├── mcg_pathwayanalysis_M_25.tsv
│   ├── mcg_pathwayanalysis_M_28.tsv
│   ├── mcg_pathwayanalysis_M_30.tsv
│   ├── mcg_pathwayanalysis_M_31.tsv
│   ├── path.order.txt
│   └── pathway.0.01.data.txt
├── Fig5
│   ├── Cyto-Fig5.pdf
│   ├── Fig5-20180909.pdf
│   ├── Fig5-20180909.pptx
│   ├── Fig5-20181030.pdf
│   ├── Fig5-20181030.pptx
│   ├── Screen Shot 2018-09-09 at 4.49.27 PM.png
│   └── attri_GSA_score.txt
├── Suppl_Table_1__.xlsx
├── Suppl_Table_3.xlsx
└── Suppl_fig1
    ├── Suppl_Fig_1_metabo_clustering_HCL.png
    ├── draw_heatmap_metabo_clusters.py
    └── sh.txt
```

Panel B contributed by Steve Pittard

Common bioinformatics tasks

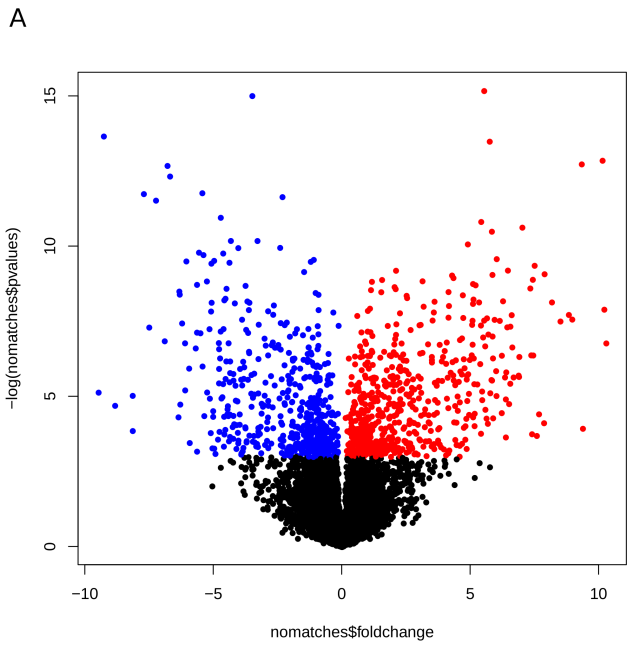
- Performing quality control (QC) and quality assurance (QA)
- Data wrangling, transformation, scaling and normalization
- Statistical analyses
- Visualization
- Online help can be found via websites such as Stack Overflow (https://stackoverflow.com/) and BioStars (https://www.biostars.org).

Example of quality control

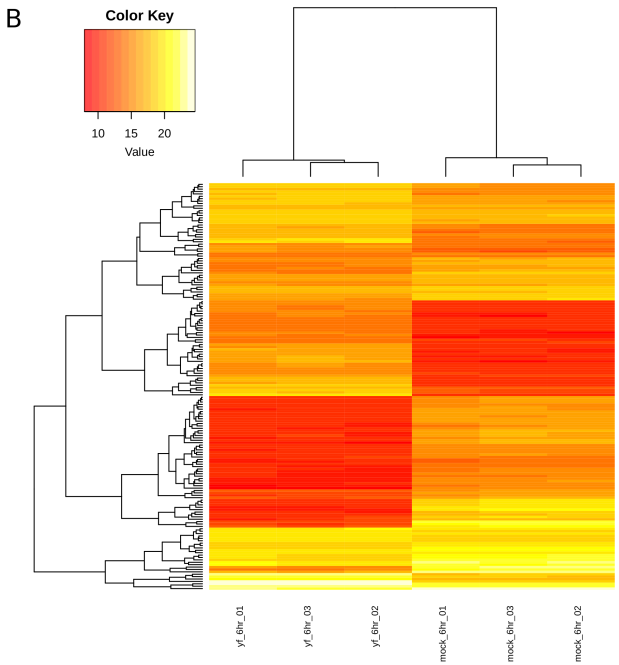


Example of data visualization

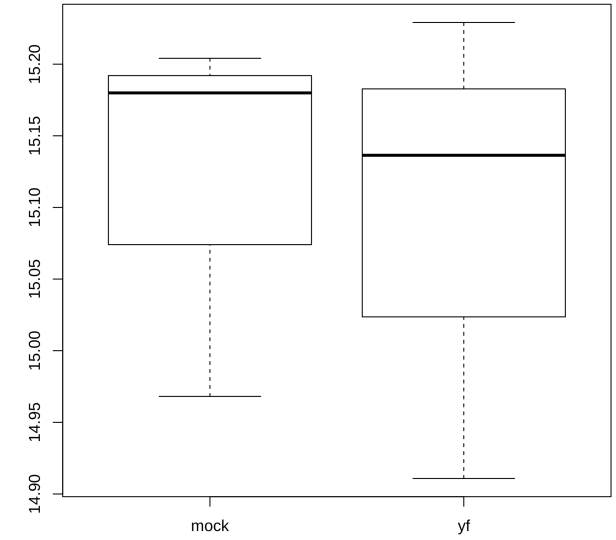
Volcano plot



Heat map



Box plot



Basics of scripting and Code editors

- Learn to use a code editor, not MS Word. A code editor works on plain text and gives you total transparency.
- The default text editor on MS Windows is Notepad, on Mac OS is TextEdit. A few free code editors I've used include Atom, VS Code and Geany.
- Beginners should avoid IDEs (Integrated Development Environment, e.g. Eclipse)
- Rstudio, Matlab, IPython are not exactly programming IDEs, more interactive environment for data analysis.
- Command line is empowering. A good tutorial on UNIX Shell is http://swcarpentry.github.io/shell-novice/
- Understand absolute and relative paths on computers

Jupyter Notebook

- By keeping code and result in a web browser, it keeps research record, facilitates collaboration, and makes good tutorials. The Jupyter notebook can run kernels on different computer languages. Quick guide: https://jupyter.readthedocs.io/en/latest/content-quickstart.html.

- Option 1. Using Anaconda, a software distribution for Python/R data science https://www.anaconda.com/distribution/ After installing, you can use the Jupyter notebook within.

- Option 2 (not recommended for novices). Using Docker container. https://docs.docker.com/get-started/ Docker is not easiest on MS Windows, but it works. After Docker is working, modify this command line to run a Jupyter notebook with a local work directory: `docker run -v /home/shuzhao/project_1_megaID:/home/jovyan/p1 -p 8888:8888 jupyter/scipy-notebook`

  The directory after "-v" is my working directory, and you should change to yours. That's where the input data file should be, and you will see the mounted directory "p1" via the notebook.