Towards reproducible research

Roland Krause

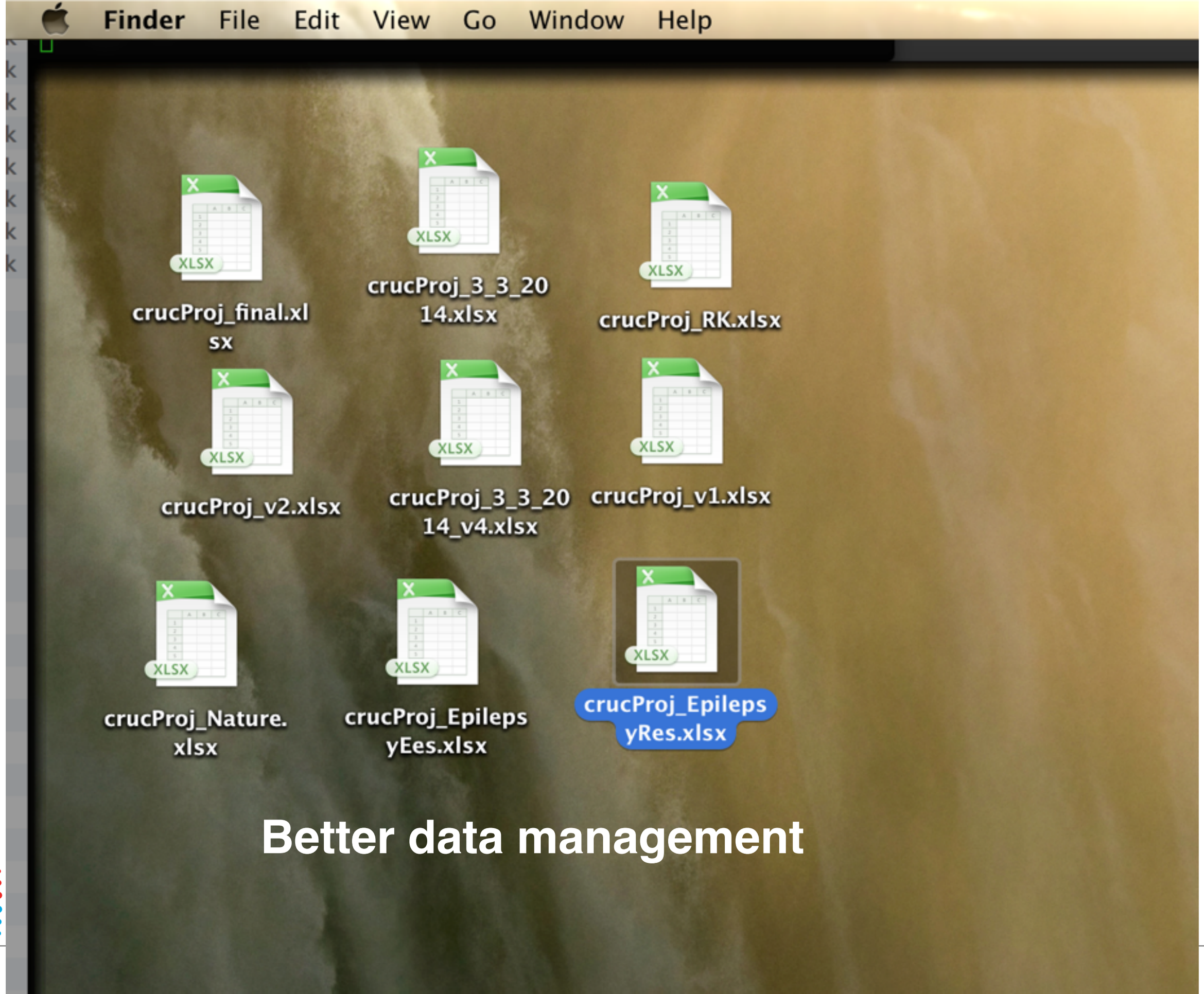
Luxembourg Centre for Systems Biomedicine

Content

- Basic data management
- R and the shell
- Tools for reproducible research
  - RStudio and markdown
  - Software versioning, make etc.

Learning objectives

- Understand large scale bioinformatics
- Know principles of *tidy data*
- Connecting the Shell with R and R with the Shell
- Running R within RStudio
- Writing basic markdown documents


Finder File Edit View Go Window Help
crucProj_final.xlsx
crucProj_3_3_2014.xlsx
crucProj_RK.xlsx
crucProj_v2.xlsx
crucProj_3_3_2014_v4.xlsx
crucProj_v1.xlsx
crucProj_Nature.xlsx
crucProj_EpilepsyEes.xlsx
crucProj_EpilepsyRes.xlsx
Better data management

Bad data

1. Patient names
2. Identical column names
3. Inconsistent variables
4. Non-English columns names
5. Color coding
6. Inconsistent dates (ISO8601)
7. “Disease”
8. Multiple columns for one item
9. Redundant information
10. Repeated rows
11. Comment field
12. Uncoded syndromes
13. Unnecessary information (Birthdate, comments)
14. Name of the table

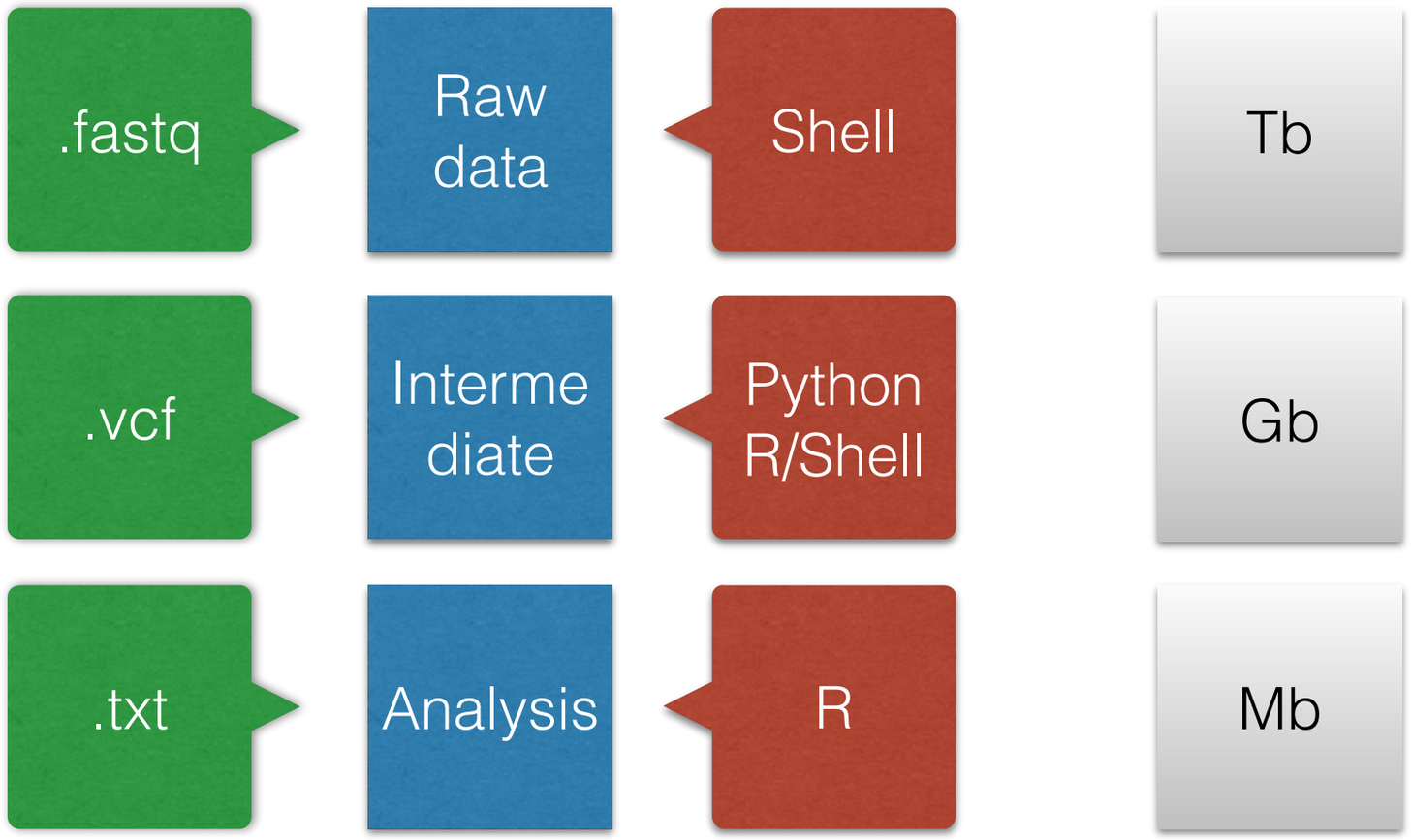
Tidy data

- One variable per column
- One observation per row
- Tables hold elements of only one kind

Even tidier data

- Column names are easy to use and informative
- Row names are easy to use and informative
- Obvious mistakes in the data have been removed
- Variable values are internally consistent
- Appropriate transformed variables have been added

Workflow in NGS



Why is computational research ***not*** reproducible?

- Copy & paste
- Manual text entry
- Data from downloaded from “some website”
- Code version not tractable

How to improve

- Document before you write code
- Everything is a text file
- No files edited manually or adhoc
- No manual selection
- Share data and code

General structure

Separate

- *data collection - “raw data”*
  - Experimental data - omics and small scale
  - Electronic data collection, clinical support
  - Web-scraped resources, data downloads
- *processed*
  Computed results
- *analysis*
  R scripts, Markdown documents, Manuscripts

How bad is Excel?

- All your analysis will be required to be made available
  - Reinhart-Rogoff “Growth in a time of debt”
  - Potti et al., “Genomic signatures to guide the use of chemotherapeutics”
  - CLCN2 variants in IGE
- Excel manipulations are inherently non-tractable
- It’s the data and data structure, not the tool

The shell in R

Most shell commands have an equivalent in R

- **getwd()** - pwd
- **setwd()** - cd
- **list.files()** - ls
  ls() - list elements in workspace!

The shell is directly accessible via

system(command)

system(“ls”)

Calling R on the shell

Rscript file.R

More direct matches

- **head(df), tail(df)**
  Yield the elements of the data structures, not files!
- **grep(vec)**
- **cat()** - Try comparing to print()
- **seq(), unique()**

Data frames

- **cbind(), rbind()**
- **merge()** - Merge two data frames by a common column
- **names(df), dim(df), nrow(df), ncol(df)**
- **df[row, col]** - Subsetting

Conditional

- if (cond) expression
- ifelse(test, yes, no)

for loops in R

- for (i in seq(0,3,1)) print(i)

- for (i in seq(0,3,1)) {
  print(i)
  }

- General advice: don’t use for loops on data frames!

Alternatives to loops

- Many functions are vectorised and accept multiple input
  - paste("Hello", “World!")
  - paste(c(“a”, “b”, “c”), c(1,2,3))
- The zoo of apply functions
  - lapply(list, function)

The value of NA

```
> ifelse(T, 1 ,2)

[1] 1

> ifelse(F, 1 ,2)

[1] 2

> ifelse(NA, 1 ,2)

[1] NA
```

Resources

- http://goo.gl/TPX7GI
- R tutorial - http://www.cyclismo.org/tutorial/R/hwI.html

Literate programming

- An article is a stream of text and code
- Analysis code is divided into text and code “chunks”
- Presentaiton codeformats results (tables, figures, etc.)
- Article text explains what is going on
- Literate programs are weaved to produce human-readable documents and tangled to produce machine-readable documents

What is markdown?

- A simple formatting structure

What you can do in knitr

- Manuals
- Short/medium-length technical documents
- Tutorials
- Reports (esp.if generated periodically)
- Home work
- Data preprocessing documents/summaries

Markdown exercise

- Create the template for your paper
  - Introduction, Methods, Results, Discussion
    - Insert the compartment data
    - Visualize the compartment data in a histogram or barplot
      - How often is a particular validity code being used?
    - Use the shell or R to preprocess the data