Introduction to reproducible research with Sweave and ORG-MODE

Yurii Aulchenko
Consultant and Independent Researcher
yurii [dot] aulchenko [at] gmail [dot] com

Reproducible research

- Embeds executable code in research reports and publications, with the aim of allowing readers to re-run the analyses described

My view

- Precise way of documenting analysis
- Allows me to return to my previous analysis and
  - Reproduce / verify it
  - Build more analyses on the top
  - Find my previous analyses / results
- Allows running same analysis on different data set (“pipelining”)

Literate programming

- Enhances traditional software development by embedding code in explanatory essays and encourages treating the act of development as one of communication with future maintainers

My view

- Write the code / pipeline in a way that I can understand what is going on years after (or that someone can understand the code without me explaining it)
- This allows easy maintenance and code re-use

Problem

Prepare data

Import Data

Perform QC

Analyze association

Generate report

Primitive way to do RR

```
drwxr-xr-x  17 yuryaulchenko  staff     578 May  8 23:01 .
drwxr-xr-x  12 yuryaulchenko  staff     408 May  9 15:06 ..
-rw-r--r--   1 yuryaulchenko  staff     390 May  8 22:49 .pversion
-rwxr-xr-x   1 yuryaulchenko  staff     305 May  8 22:42 00_prepare_data.sh
-rw-r--r--   1 yuryaulchenko  staff     178 May  8 21:47 01_import2R.R
-rw-r--r--   1 yuryaulchenko  staff     206 May  8 22:46 02_doQC.R
-rw-r--r--   1 yuryaulchenko  staff     110 May  8 22:47 03_analysis.R
-rw-r--r--   1 yuryaulchenko  staff    7581 May  8 22:50 Rplots.pdf
-rw-r--r--   1 yuryaulchenko  staff   14639 May  8 22:50 dta.RData
-rw-r--r--   1 yuryaulchenko  staff   13213 May  8 22:50 dtaQCed.RData
-rw-r--r--   1 yuryaulchenko  staff    2003 May  8 22:49 phenotypes.txt
-rw-r--r--   1 yuryaulchenko  staff    2470 May  8 22:49 phenotypes_goodEOL.csv
-rw-r--r--   1 yuryaulchenko  staff    1844 May  8 22:49 plink.log
-rw-r--r--   1 yuryaulchenko  staff   59220 May  8 22:50 plink.raw
-rw-r--r--   1 yuryaulchenko  staff    1824 May  8 22:49 plink.tfam
-rw-r--r--   1 yuryaulchenko  staff  251312 May  8 22:49 plink.tped
-rw-r--r--   1 yuryaulchenko  staff     105 May  8 22:41 reformat.pl
```

R Sweave

- A module allowing documenting R analyses

- Our example:
  - Import files to R
  - Perform QC
  - Run association analysis

Summary R Sweave

- **Great tool** for pedagogic; for RR in R (and may be some shell commands...)
- Lack of instant feedback (-RR): with larger analysis it may be very annoying that you need to re-run large chunks
- Not so good to combine multiple tools and scripts and pieces of own code

ORG-MODE

- OrgMode paper
  - http://www.jstatsoft.org/v46/i03
- OrgMode web-site
  - http://orgmode.org

Languages supported

Language	Identifier	Language	Identifier
Asymptote	asymptote	Awk	awk
Emacs Calc	calc	C	C
C++	C++	Clojure	clojure
CSS	css	ditaa	ditaa
Graphviz	dot	Emacs Lisp	emacs-lisp
gnuplot	gnuplot	Haskell	haskell
Java	java		
Javascript	js	LaTeX	latex
Ledger	ledger	Lisp	lisp
Lilypond	lilypond	MATLAB	matlab
Mscgen	mscgen	Objective Caml	ocaml
Octave	octave	Org mode	org
Oz	oz	Perl	perl
Plantuml	plantuml	Python	python
R	R	Ruby	ruby
Sass	sass	Scheme	scheme
GNU Screen	screen	shell	sh
SQL	sql	SQLite	sqlite

Languages supported

Language	Identifier	Language	Identifier
Asymptote	asymptote	Awk	awk
Emacs Calc	calc	C	C
C++	C++	Clojure	clojure
CSS	css	ditaa	ditaa
Graphviz	dot	Emacs Lisp	emacs-lisp
gnuplot	gnuplot	Haskell	haskell
Java	java		
Javascript	js	LaTeX	latex
Ledger	ledger	Lisp	lisp
Lilypond	lilypond	MATLAB	matlab
Mscgen	mscgen	Objective Caml	ocaml
Octave	octave	Org mode	org
Oz	oz	Perl	perl
Plantuml	plantuml	Python	python
R	R	Ruby	ruby
Sass	sass	Scheme	scheme
GNU Screen	screen	shell	sh
SQL	sql	SQLite	sqlite

Org-Mode

- Platform for RR & **Literate Programming**
- Supports >30 languages
- Instant feedback
- Flexible controls
- Reports in many formats - PDF, HTML, ...