Steps toward reproducible research

Karl Broman

Biostatistics & Medical Informatics
Univ. Wisconsin–Madison

kbroman.org
github.com/kbroman
@kwbroman
Slides: bit.ly/TAGC2016

Karl -- this is very interesting, however you used an old version of the data (n=143 rather than n=226).

I'm really sorry you did all that work on the incomplete dataset.

Bruce

The results in Table 1 don’t seem to correspond to those in Figure 2.

In what order do I run these scripts?

Where did we get this data file?

Why did I omit those samples?

How did I make that figure?

Reproducible

vs.

Replicable

Reproducible

vs.

Correct

Steps toward reproducible research

kbroman.org/steps2rr

1. Organize your data & code

File organization and naming
are powerful weapons against chaos.

– Jenny Bryan

1. Organize your data & code

Your closest collaborator is you six months ago,
but you don't reply to emails.

(paraphrasing Mark Holder)

1. Organize your data & code

```
RawData/                Notes/
DerivedData/            Refs/

Python/                 ReadMe.txt
R/                      ToDo.txt
Ruby/                   Makefile
```

Chaos

```
AimeeNullSims/     Deuterium/             Ping/
AimeeResults/      ExtractData4Gary/      Ping2/
AnnotationFiles/   FromAimee/             Ping3/
Brian/             GoldStandard/          Ping4/
Chr6_extrageno/    HumanGWAS/             Play/
Chr6_segdis/       Insulin/               Prdm9/
ChrisPlaisier/     Int2_for_Mark/         RBM_PlasmaUrine_2012-03-08/
Code4Aimee/        Islet_2011-05/         Slco1a6/
CompAnnot/         MappingProbes/         StudyLineupMethods/
CondScans/         MultiProbes/           kidney_chr6.R
D2O_2012-02-14/    NewMap/                pck2_sucla2.R
D2O_cellcycle/     Notes/                 penalties.txt
D2Ocorr/           NullSims/              transeQTL4Lude/
Data4Aimee/        NullSims_2009-09-10/
Data4Tram/         PepIns_2012-02-09/
```

2. Everything with a script

If you do something once,
you'll do it 1000 times.

3. Automate the process (GNU Make)

```
R/analysis.html: R/analysis.Rmd Data/cleandata.csv
    cd R;R -e "rmarkdown::render('analysis.Rmd')"

Data/cleandata.csv: R/prepData.R RawData/rawdata.csv
    cd R;R CMD BATCH prepData.R

RawData/rawdata.csv: Python/xls2csv.py RawData/rawdata.xls
    Python/xls2csv.py RawData/rawdata.xls > RawData/rawdata.csv
```

3. Automate the process (GNU Make)

```
R/analysis.html: R/analysis.Rmd Data/cleandata.csv
    cd R;R -e "rmarkdown::render('analysis.Rmd')"

Data/cleandata.csv: R/prepData.R RawData/rawdata.csv
    cd R;R CMD BATCH prepData.R

RawData/rawdata.csv: Python/xls2csv.py RawData/rawdata.xls
    Python/xls2csv.py RawData/rawdata.xls > RawData/rawdata.csv
```

3. Automate the process (GNU Make)

```
R/analysis.html: R/analysis.Rmd Data/cleandata.csv
    cd R;R -e "rmarkdown::render('analysis.Rmd')"

Data/cleandata.csv: R/prepData.R RawData/rawdata.csv
    cd R;R CMD BATCH prepData.R

RawData/rawdata.csv: Python/xls2csv.py RawData/rawdata.xls
    Python/xls2csv.py RawData/rawdata.xls > RawData/rawdata.csv
```

3. Automate the process (GNU Make)

```
R/analysis.html: R/analysis.Rmd Data/cleandata.csv
    cd R;R -e "rmarkdown::render('analysis.Rmd')"

Data/cleandata.csv: R/prepData.R RawData/rawdata.csv
    cd R;R CMD BATCH prepData.R

RawData/rawdata.csv: Python/xls2csv.py RawData/rawdata.xls
    Python/xls2csv.py RawData/rawdata.xls > RawData/rawdata.csv
```

4. Turn scripts into reproducible reports

Gough project diagnostics

Karl Broman, 3 March 2014

Combine genotypes and phenotypes

I've combined the initial genotypes (using the re-clustered genotypes for plates 14-16) with the well-behaved portion of the re-run genotypes. I'm focusing on 36813 markers that are informative (though, as we'll see, there are still a lot of badly behaved and basically non-informative markers that need to be removed). I've combined data on replicate samples, to give one set of genotype calls for each sample.

There are 1497 genotyped mice and 1464 phenotyped mice. All of the mice in the phenotype data have genotypes, but there are 33 genotyped mice with no phenotypes, including 3 Gough mice and 30 F2 progeny.

4. Turn scripts into reproducible reports

Gough project diagnostics

Karl Broman, 3 March 2014

Combine genotypes and phenotypes

I've combined the initial genotypes (using the re-clustered genotypes for plates 14-16) with
the well-
informat
informat
give one

There ar
data hav
mice and

```
I've combined the initial genotypes (using the re-clustered genotypes
for plates 14-16) with the well-behaved portion of the re-run
genotypes. I'm focusing on `r totmar(g)` markers that are informative
(though, as we'll see, there are still a lot of badly behaved and
basically non-informative markers that need to be removed).
I've combined data on replicate samples, to give one set of genotype
calls for each sample.

There are `r nind(g)` genotyped mice and `r nrow(phe)` phenotyped
```

5. Turn repeated code into functions

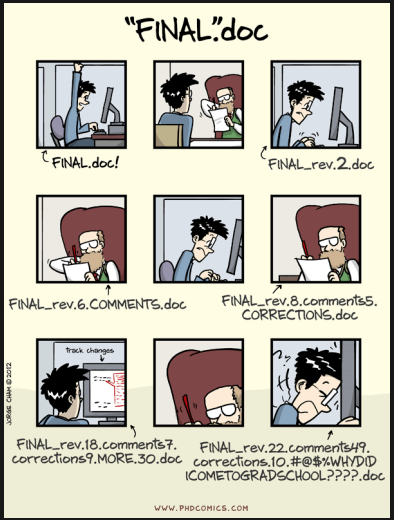
```
# Python
def read_genotypes (filename):
    "Read matrix of genotype data"
```

```
# R
plot_genotypes <-
function(genotypes, ...)
{
}
```

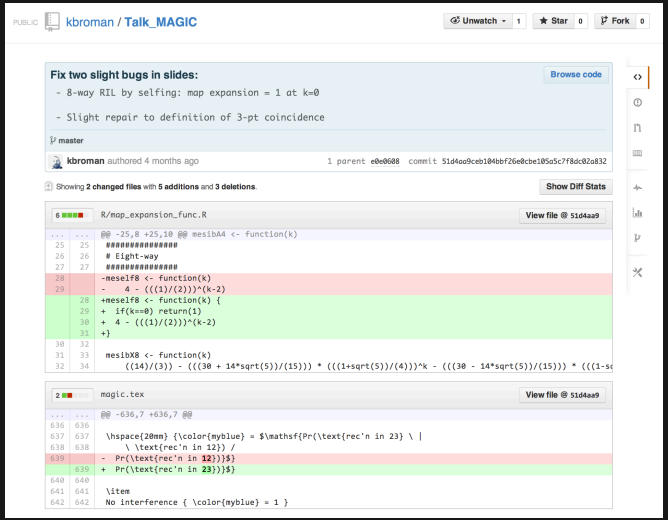
6. Create a package/module

Don’t repeat yourself

7. Use version control (git/GitHub)



7. Use version control (git/GitHub)



8. License your software

Pick a license, any license

– Jeff Atwood

The most important tool is the mindset,
when starting, that the end product
will be reproducible.

– Keith Baggerly

Slides: bit.ly/TAGC2016

kbroman.org

github.com/kbroman

@kwbroman