Reproducible Research

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman
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?
"Your script is now giving an error."
"The attached is similar to the code we used."
Reproducible
Reproducible vs. Replicable
Reproducible

vs.

Correct
Levels of quality

- Are the tables and figures reproducible from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear why it was done?
  
  (e.g., how were parameter settings chosen?)

- Can the code be used for other data?
- Can you extend the code to do other things?
Steps toward reproducible research

kbroman.org/steps2rr
1. Everything with a script

If you do something once, you'll do it 1000 times.
2. Organize your data & code

File organization and naming are powerful weapons against chaos.

– Jenny Bryan
2. Organize your data & code

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

(paraphrasing Mark Holder)
2. Organize your data & code

<table>
<thead>
<tr>
<th>RawData/</th>
<th>Notes/</th>
</tr>
</thead>
<tbody>
<tr>
<td>DerivedData/</td>
<td>Refs/</td>
</tr>
<tr>
<td>Python/</td>
<td>ReadMe.txt</td>
</tr>
<tr>
<td>R/</td>
<td>ToDo.txt</td>
</tr>
<tr>
<td>Ruby/</td>
<td>Makefile</td>
</tr>
</tbody>
</table>
3. Automate the process (GNU Make)

- `cd R; R -e "rmarkdown::render('analysis.Rmd')"
- `cd R; R CMD BATCH prepData.R`
- `Python/xls2csv.py RawData/rawdata.xls > RawData/rawdata.csv`
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

25 I've combined the initial genotypes (using the re-clustered genotypes
26 for plates 14-16) with the well-behaved portion of the re-run
27 genotypes. I'm focusing on `r totmar(g)` markers that are informative
28 (though, as we'll see, there are still a lot of badly behaved and
29 basically non-informative markers that need to be removed).
30 I've combined data on replicate samples, to give one set of genotype
31 calls for each sample.
32
33 There are `r nind(g)` genotyped mice and `r nrow(phe)` phenotyped
34 mice. All of the mice in the phenotype data have genotypes, but there
35 are `r sum(is.na(match(gid, pid)))` genotyped mice with no phenotypes,
36 including `r sum(g$pheno$gen[which(is.na(match(gid, pid)))]==0)`
37 Gough mice and `r sum(g$pheno$gen[which(is.na(match(gid, pid)))]==2)`
38 F2 progeny.
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)

Talk for MAGIC Workshop in Cambridge, UK

These are slides for a talk I will give at the Workshop on MAGIC-type populations in Cambridge, UK, on 12 June 2013.

The PDF is here.

To the extent possible under law, Karl Broman has waived all copyright and related or neighboring rights to "MAGIC design and other topics". This work is published from: United States.
7. Use version control (git/GitHub)

Greatly simplify the public domain stuff in the ReadMe

- **kbroman** authored 15 days ago
- **latest commit** f177ef192

- **Figs**
  - Add crazy table from preCC paper
  - 4 months ago

- **Perl**
  - Add lines_of_code_by_version.csv to repository
  - 4 months ago

- **R**
  - Another fix regarding map expansion in 8-way RIL by selfing at k=0
  - 4 months ago

- **.gitignore**
  - Add lines_of_code_by_version.csv to repository
  - 4 months ago

- **Makefile**
  - Revise Readme to link to version for web
  - 4 months ago

- **ReadMe.md**
  - Greatly simplify the public domain stuff in the ReadMe
  - 15 days ago

- **magic.tex**
  - Fix two slight bugs in slides:
  - 4 months ago
7. Use version control (git/GitHub)
7. Use version control (git/GitHub)
8. License your software

Pick a license, any license

– Jeff Atwood
Other considerations

- **Testing**
  are you getting the right answers?

- **Software versions**
  will your stuff work when dependencies change?

- **Large-scale computations**
  computation time + dependence on cluster environment

- **Collaborations**
  coordinating who does what and where things live
Summary

1. Everything with a script
2. Organize your data & code
3. Automate the process (GNU Make)
4. Turn scripts into reproducible reports
5. Turn repeated code into functions
6. Create a package/module
7. Use version control (git/GitHub)
8. Pick a license, any license

kbroman.org

github.com/kbroman

@kwkbroman