

# Steps toward reproducible research

Karl Broman

Biostatistics & Medical Informatics  
Univ. Wisconsin–Madison

[kbroman.org](http://kbroman.org)

[github.com/kbroman](https://github.com/kbroman)

@kbroman

Slides: [bit.ly/jax2017-05](https://bit.ly/jax2017-05)



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

vs.

Replicable

Reproducible

vs.

Correct

# Steps toward reproducible research

[kbroman.org/steps2rr](http://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
D20_2012-02-14/    NewMap/             pck2_sucla2.R
D20_cellcycle/     Notes/              penalties.txt
D20corr/           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

Comb

I've comb  
the well-  
informat  
informat  
give one

There are  
data have  
mice and

```
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

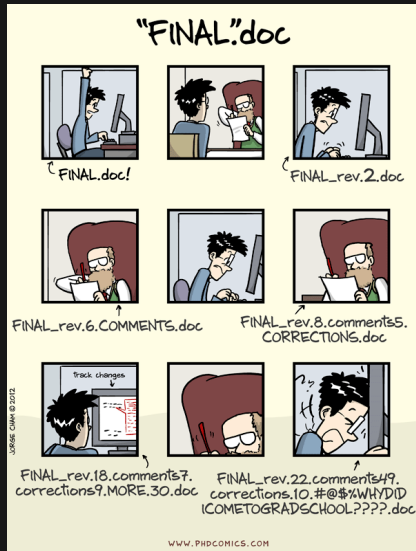
```
# Python
def read_genotypes (filename):
    "Read matrix of genotype data"
```

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

## 6. Create a package/module

Don't repeat yourself

# 7. Use version control (git/GitHub)



# No “final” in file names

Deprecated/	hypo_prcomp.RData
ReadMe.txt	islet_int1_final.RData
adipose_int1_final.RData	islet_int2_final.RData
adipose_int2_final.RData	islet_mlratio_final.RData
adipose_mlratio_final.RData	islet_mlratio_nqrank_final.RData
adipose_mlratio_nqrank_final.RData	islet_prcomp.RData
adipose_prcomp.RData	kidney_int1_final.RData
aligned_genome_with_pmap.RData	kidney_int2_final.RData
batches_final.RData	kidney_mlratio_final.RData
batches_raw_final.RData	kidney_mlratio_nqrank_final.RData
cpl_final.RData	kidney_prcomp.RData
d2o_final.RData	lipomics_final_rev2.RData
gastroc_int1_final.RData	liverTG_final.RData
gastroc_int2_final.RData	liver_int1_final.RData
gastroc_mlratio_final.RData	liver_int2_final.RData
gastroc_mlratio_nqrank_final.RData	liver_mlratio_final.RData
gastroc_prcomp.RData	liver_mlratio_nqrank_final.RData
hypo_int1_final.RData	liver_prcomp.RData
hypo_int2_final.RData	mirna_final.RData
hypo_mlratio_final.RData	necropsy_final_rev2.RData
hypo_mlratio_final_old.RData	plasmaurine_final_rev.RData
hypo_mlratio_nqrank_final.RData	pmark.RData
hypo_mlratio_nqrank_final_old.RData	rbm_final.RData
hypo_omit.RData	

# No “final” in file names

Deprecated/	hypo_prcomp.RData
ReadMe.txt	islet_int1_final.RData
adipose_int1_final.RData	islet_int2_final.RData
adipose_int2_final.RData	islet_mlratio_final.RData
adipose_mlratio_final.RData	islet_mlratio_nqrank_final.RData
adipose_mlratio_nqrank_final.RData	islet_prcomp.RData
adipose_prcomp.RData	kidney_int1_final.RData
aligned_genome_with_pmap.RData	kidney_int2_final.RData
batches_final.RData	kidney_mlratio_final.RData
batches_raw_final.RData	kidney_mlratio_nqrank_final.RData
cpl_final.RData	kidney_prcomp.RData
d2o_final.RData	lipomics_final_rev2.RData
gastroc_int1_final.RData	liverTG_final.RData
gastroc_int2_final.RData	liver_int1_final.RData
gastroc_mlratio_final.RData	liver_int2_final.RData
gastroc_mlratio_nqrank_final.RData	liver_mlratio_final.RData
gastroc_prcomp.RData	liver_mlratio_nqrank_final.RData
hypo_int1_final.RData	liver_prcomp.RData
hypo_int2_final.RData	mirna_final.RData
hypo_mlratio_final.RData	necropsy_final_rev2.RData
hypo_mlratio_final_old.RData	plasmaurine_final_rev.RData
hypo_mlratio_nqrank_final.RData	pmark.RData
hypo_mlratio_nqrank_final_old.RData	rbm_final.RData
hypo_omit.RData	

# 7. Use version control (git/GitHub)

The screenshot shows a GitHub repository page for 'kbroman / Talk\_MAGIC'. At the top, it indicates 'PUBLIC' and 'kbroman / Talk\_MAGIC'. There are buttons for 'Unwatch' (1), 'Star' (0), and 'Fork' (0). The repository name is 'Talk\_MAGIC' with a link to the repository. Below this, it says 'Talk for MAGIC workshop in Cambridge, UK, 12 June 2013 — Edit'. There are statistics for '97 commits', '1 branch', '0 releases', and '1 contributor'. A green bar indicates the current branch is 'master' and the repository is 'Talk\_MAGIC'. A commit message is shown: 'Greatly simplify the public domain stuff in the ReadMe', authored by 'kbroman' 15 days ago, with the latest commit hash 'f1777ef192'. A list of files is shown with their commit dates: 'Figs' (4 months ago), 'Perf' (4 months ago), 'R' (4 months ago), '.g@gnore' (4 months ago), 'Makefile' (4 months ago), 'ReadMe.md' (15 days ago), and 'magic.tex' (4 months ago). Below the file list, the 'ReadMe.md' content is displayed, featuring the title 'Talk for MAGIC Workshop in Cambridge, UK'. The text describes the slides for a talk at the 'Workshop on MAGIC-type populations' in Cambridge, UK, on 12 June 2013. It includes a link for the PDF and a copyright notice stating that Karl Broman has waived all copyright and related or neighboring rights to 'MAGIC design and other topics'. A 'PUBLIC DOMAIN' logo is also present. On the right side, there are navigation options: 'Code', 'Issues' (0), 'Pull Requests' (0), 'Wiki', 'Pulse', 'Graphs', 'Network', and 'Settings'. At the bottom right, there are options to 'Clone in Desktop' and 'Download ZIP', along with the HTTPS clone URL: 'https://github.com/kbroman/Talk\_MAGIC.git'.

# 7. Use version control (git/GitHub)

The screenshot shows the GitHub interface for the repository 'kbroman / Talk\_MAGIC'. At the top, it indicates the repository is 'PUBLIC'. There are buttons for 'Unwatch' (1), 'Star' (0), and 'Fork' (0). The repository description is 'Talk for MAGIC workshop in Cambridge, UK, 12 June 2013 — Edit'. Below the description, there are statistics: 97 commits, 1 branch, 0 releases, and 1 contributor. On the right side, there are links for 'Code', 'Issues', and 'Pull requests'.

## Greatly simplify the public domain stuff in the ReadMe



**kbroman** authored 15 days ago

latest commit [f1777ef192](#)

<a href="#">Figs</a>	Add crazy table from preCC paper	4 months ago
<a href="#">Perl</a>	Add lines_of_code_by_version.csv to repository	4 months ago
<a href="#">R</a>	Another fix regarding map expansion in 8-way RIL by selfing at k=0	4 months ago
<a href="#">.gitignore</a>	Add lines_of_code_by_version.csv to repository	4 months ago
<a href="#">Makefile</a>	Revise Readme to link to version for web	4 months ago
<a href="#">ReadMe.md</a>	Greatly simplify the public domain stuff in the ReadMe	15 days ago
<a href="#">magic.tex</a>	Fix two slight bugs in slides:	4 months ago

rights to "MAGIC design and other topics". This work is published from: United States.



# 7. Use version control (git/GitHub)

The screenshot shows a GitHub repository page for 'kbroman / Talk\_MAGIC'. At the top, it indicates the repository is 'PUBLIC' and shows navigation options: 'Unwatch 1', 'Star 0', and 'Fork 0'. Below this, the current branch is 'master' and the view is 'Talk\_MAGIC / Commits'. The commits are listed in chronological order, grouped by date. The most recent commit is from September 27, 2013, followed by a commit from June 17, 2013, and a series of commits from June 10, 2013. The commit from June 17, 2013, is highlighted in yellow and contains two bullet points: '- 8-way RIL by selfing: map expansion = 1 at k=0' and '- Slight repair to definition of 3-pt coincidence'. Each commit entry includes a commit hash, a 'Browse code' link, and the author's name and time since commit.

PUBLIC kbroman / Talk\_MAGIC Unwatch 1 Star 0 Fork 0

branch: master - Talk\_MAGIC / Commits

Sep 27, 2013

- Greatly simplify the public domain stuff in the ReadMe  
kbroman authored 15 days ago f1777ef192 - Browse code
- Fix url in ReadMe.md file  
kbroman authored 15 days ago a6515823f9 - Browse code

Jun 17, 2013

- Another fix regarding map expansion in 8-way RIL by selfing at k=0  
kbroman authored 4 months ago 208a482f2c - Browse code
- Fix two slight bugs in slides: `||||`  
- 8-way RIL by selfing: map expansion = 1 at k=0  
- Slight repair to definition of 3-pt coincidence  
kbroman authored 4 months ago 51d4aa9ceb - Browse code

Jun 10, 2013

- Change one page number  
kbroman authored 4 months ago e0e0688615 - Browse code
- Add missing paren  
kbroman authored 4 months ago f4975dee6e - Browse code
- who's -> who is  
kbroman authored 4 months ago 886f20f098 - Browse code
- rublish -> bad  
kbroman authored 4 months ago e6fbf2f647 - Browse code
- Add link to R/qtl page  
kbroman authored 4 months ago 4edf3e8676 - Browse code
- Revise slide re analysis issues  
kbroman authored 4 months ago 14ebb1eeb5 - Browse code
- italicize 'de novo'  
kbroman authored 4 months ago 45dda4b4c7 - Browse code
- replace plain right arrow with fat arrow  
kbroman authored 4 months ago 8bbe385d6c - Browse code

# 7. Use version control (git/GitHub)

PUBLIC kbroman / Talk\_MAGIC Unwatch 1 Star 0 Fork 0

**Fix two slight bugs in slides:** [Browse code](#)

- 8-way RIL by selfing: map expansion = 1 at k=0
- Slight repair to definition of 3-pt coincidence

master

kbroman authored 4 months ago 1 parent e0e8608 commit 51d4aa9ceb104bbf26e8cbe185a5c7f8dc02a832

Showing 2 changed files with 5 additions and 3 deletions. [Show Diff Stats](#)

**6** R/map\_expansion\_func.R [View file @ 51d4aa9](#)


```
... .. @@ -25,8 +25,10 @@ mesibA4 <- function(k)
25 25 #####
26 26 # Eight-way
27 27 #####
28 -mesif8 <- function(k)
29 - 4 - ((1)/(2))^(k-2)
+mesif8 <- function(k) {
+ if(k==0) return(1)
+ 4 - ((1)/(2))^(k-2)
+ }
30 32
31 33 mesibX8 <- function(k)
32 34 ((14)/(3)) - ((30 + 14*sqrt(5))/(15)) * (((1+sqrt(5))/(4)))^k - ((30 - 14*sqrt(5))/(15)) * (((1-sq
```

**2** magic.tex [View file @ 51d4aa9](#)

```
... .. @@ -636,7 +636,7 @@
636 636
637 637 \hspace{20mm} {\color{myblue} = $\mathsf{Pr}(\text{rec'n in 23}) \setminus
638 638 \ \text{rec'n in 12}) /
639 - Pr(\text{rec'n in 12})}$
639 + Pr(\text{rec'n in 23})}$
640 640
641 641 \item
642 642 No interference { \color{myblue} = 1 }
```

## 7. Use version control (git/GitHub)

```
27 27 #####
28 28 -meself8 <- function(k)
29 29 -   4 - (((1)/(2)))^(k-2)
30 28 +meself8 <- function(k) {
31 29 +   if(k==0) return(1)
32 30 +   4 - (((1)/(2)))^(k-2)
33 31 +}
34 30 mesibX8 <- function(k)
35 31   ((14)/(3)) - (((30 + 14*sqrt(5))/(15)))
36 32
```

2  magic.tex

```
... ... @@ -636,7 +636,7 @@
636 636
637 637 \hspace{20mm} {\color{myblue} = $\mathsf{Pr(}
638 638 \ \ \text{rec'n in 12})} /
639 639 - Pr(\text{rec'n in 12})}$}
640 639 + Pr(\text{rec'n in 23})}$}
641 640
```

## 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

- ▶ **Distribution**

where and how to distribute data and code?

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

– Keith Baggerly

# Summary

1. Organize your data & code
2. Everything with a script
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

Slides: [bit.ly/jax2017-05](https://bit.ly/jax2017-05)



[kbroman.org](https://kbroman.org)

[github.com/kbroman](https://github.com/kbroman)

@kwbroman