Steps toward reproducible research

Karl Broman

Biostatistics & Medical Informatics
Univ. Wisconsin–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.

Where did we get this data file?

Why did I omit those samples?

Which image goes with which experiment?

How did I make that figure?

In what order do I run these scripts?

“Your script is now giving an error.”

“The attached is similar to the code we used.”

Reproducible

vs.

Replicable

Reproducible

vs.

Correct

kbroman.org/steps2rr

A little bit reproducible
is better than not reproducible.

A little bit open
is better than not open.

Strive to make each project
a bit better organized than the last.

Organize your project

File organization and naming
are powerful weapons against chaos.

– Jenny Bryan

Organize your project

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

(paraphrasing Mark Holder)

Organize your project

Have sympathy for your future self.

Organize your project

```
RawData/                    Notes/
DerivedData/                Refs/

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

Analysis/
Figures/
```

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/
```

Choose good names for things

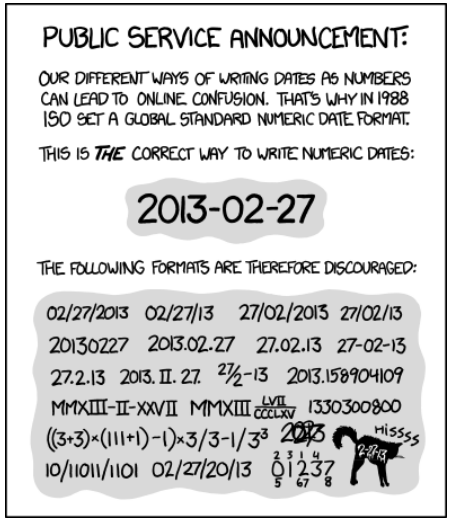
```
betw_tissue_corr.R      expr_scatterplot_allprobes.R  gve_similarity_alltissues.R
coatcolor_lod.R         expr_scatterplots_dup.R       gve_similarity.R
colors.R                expr_scatterplots_mix.R       gve_supp.R
cover_fig.R             expr_scatterplots_swap.R      insulin_lod.R
eqtl_counts_10.R        expr_swaps.R                  local_eqtl_locations.R
eqtl_counts.R           func.R                        my_plot_map.R
eve_hist.R              genotype_plates.R             my_plot_scanone.R
eve_scheme.R            gve_hist.R                    sex_vs_X.R
eve_similarity.R        gve_new.R                     xchr_fig.R
eve_similarity_supp.R   gve.R                         xist_and_y.R
expr_corr_dup.R         gve_scheme.R
expr_corr_mix.R         gve_similarity_2ndbest.R
```

Choose good names for things

```
fig1.png          fig5.png
fig10.png         fig6.png
fig2.png          fig7.png
fig3.png          fig8.png
fig4.png          fig9.png
```

Choose good names for things

- Machine readable
  - No spaces
  - No special characters except _ and -
- Human readable
  - Explain the contents
- Consistent
  - Name similar files in a similar way
- Make use of computer's sorting
  - pad numbers with 0's (e.g., `01`, `02`, ...)
  - start with general grouping, then more specific
  - dates like `2019-05-14`

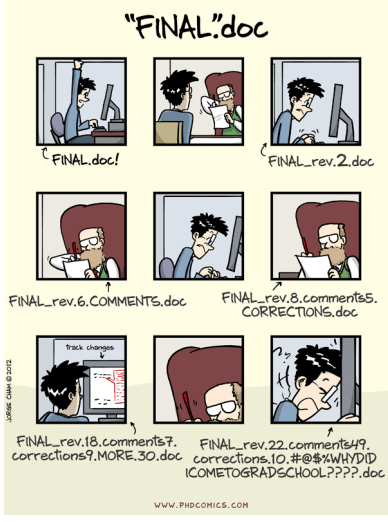


xkcd.com/1179

Choose good names for things

```
0_vcf2db.R
1_prep_geno.R
2_prep_pheno_clin.R
2_prep_pheno_otu.R
3_prep_covar.R
4_prep_analysis_pheno_clin.R
4_prep_analysis_pheno_otu.R
5_scans.R
6_grab_peaks.R
7_find_nearby_peaks.R
```

No “final” in file names



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

Choose good names for things

```
batches_raw_v1.rds                  geneexpr_mlratio_gastroc_v2.rds
batches_v1.rds                      geneexpr_mlratio_hypo_v1.rds
clinical_cpl_v2.rds                 geneexpr_mlratio_hypo_v2.rds
clinical_d2o_v2.rds                 geneexpr_mlratio_islet_v2.rds
clinical_lipomics_v4.rds            geneexpr_mlratio_kidney_v2.rds
clinical_liverTG_v2.rds             geneexpr_mlratio_liver_v2.rds
clinical_mirna_v2.rds               geneexpr_mlratio_nqrank_adipose_v2.rds
clinical_necropsy_v4.rds            geneexpr_mlratio_nqrank_gastroc_v2.rds
clinical_plasmaurine_v3.rds         geneexpr_mlratio_nqrank_hypo_v1.rds
clinical_rbm_v2.rds                 geneexpr_mlratio_nqrank_hypo_v2.rds
Deprecated/                         geneexpr_mlratio_nqrank_islet_v2.rds
geneexpr_int1_adipose_v2.rds        geneexpr_mlratio_nqrank_kidney_v2.rds
geneexpr_int1_gastroc_v2.rds        geneexpr_mlratio_nqrank_liver_v2.rds
geneexpr_int1_hypo_v2.rds           geneexpr_omit_hypo.rds
geneexpr_int1_islet_v2.rds          geneexpr_prcomp_adipose_v2.rds
geneexpr_int1_kidney_v2.rds         geneexpr_prcomp_gastroc_v2.rds
geneexpr_int1_liver_v2.rds          geneexpr_prcomp_hypo_v2.rds
geneexpr_int2_adipose_v2.rds        geneexpr_prcomp_islet_v2.rds
geneexpr_int2_gastroc_v2.rds        geneexpr_prcomp_kidney_v2.rds
geneexpr_int2_hypo_v2.rds           geneexpr_prcomp_liver_v2.rds
geneexpr_int2_islet_v2.rds          geno_aligned_w_pmap.rds
geneexpr_int2_kidney_v2.rds         geno_pmark.rds
geneexpr_int2_liver_v2.rds          ReadMe.txt
geneexpr_mlratio_adipose_v2.rds
```

Document your work

- What is all of this stuff?
- What was your analysis process?

→ `ReadMe` files

Organizing data in spreadsheets

	A	B	C	D	E	F	G
1	1MIN						
2			Normal			Mutant	
3	B6	146.6	138.6	155.6	166	179.3	186.9
4	BTBR	245.7	240	243.1	177.8	171.6	188.1
5							
6	5MIN						
7			Normal			Mutant	
8	B6	333.6	353.6	408.8	450.6	474.4	423.8
9	BTBR	514.4	610.6	597.9	412.1	447.4	446.5

Organizing data in spreadsheets

	A	B	C	D
1	ttt_min	strain	mutation	response
2	1	B6	normal	146.6
3	1	B6	normal	138.6
4	1	B6	normal	155.6
5	1	B6	mutant	166
6	1	B6	mutant	179.3
7	1	B6	mutant	186.9
8	1	BTBR	normal	245.7
9	1	BTBR	normal	240
10	1	BTBR	normal	243.1
11	1	BTBR	mutant	177.8
12	1	BTBR	mutant	171.6
13	1	BTBR	mutant	188.1
14	5	B6	normal	333.6
15	5	B6	normal	353.6

Organizing data in spreadsheets

- Make it a rectangle
- Individual measurements as rows; variables as columns
- Single header row
- One item per cell
- No empty cells
- No calculations in the raw data
- No highlighting or coloring as data

Broman and Woo (2018) Am Stat 72:2-10
doi: 10.1080/00031305.2017.1375989

“What the heck is ‘FAD_NAD SI 8.3_3.3G’?”

Metadata

- Create a data dictionary
  - Explain each column
  - Include different versions of the variable names (compact vs descriptive)
  - Units
  - Allowable values
- The metadata are data
  - Make it a rectangle

Data dictionary

	A	B	C	D
1	name	plot_name	group	description
2	mouse	Mouse	demographic	Animal identifier
3	sex	Sex	demographic	Male (M) or Female (F)
4	sac_date	Date of sac	demographic	Date mouse was sacrificed
5	partial_inflation	Partial inflation	clinical	Indicates if mouse showed partial pancreatic inflation
6	coat_color	Coat color	demographic	Coat color, by visual inspection
7	crumblers	Crumblers	clinical	Indicates if mouse stored food in their bedding
8	diet_days	Days on diet	clinical	Number of days on high−fat diet

Everything with a script

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

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.

Reproducible reports

Gough project diagnostics

Karl

Co

I've
the
infor
infor
give

Ther
data
mice

```
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
mice. All of the mice in the phenotype data have genotypes, but there
are `r sum(is.na(match(gid, pid)))` genotyped mice with no phenotypes,
including `r sum(g$pheno$gen[which(is.na(match(gid, pid)))]==0)`
Gough mice and `r sum(g$pheno$gen[which(is.na(match(gid, pid)))]==2)`
F2 progeny.
```

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
```

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
```

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
```

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
```

Fancier example

```
FIG_DIR = Figs

mypaper.pdf: mypaper.tex ${FIG_DIR}/fig1.pdf ${FIG_DIR}/fig2.pdf
    pdflatex mypaper

# One line for both figures
${FIG_DIR}/%.pdf: R/%.R
    cd R;R CMD BATCH $(<F)

# Use "make clean" to remove the PDFs
clean:
    rm *.pdf Figs/*.pdf
```

How do you use make?

- If you name your make file `Makefile`, then just go into the directory containing that file and type `make`
- If you name your make file `something.else`, then type
  `make -f something.else`
- Actually, the commands above will build the first target listed in the make file. So I'll often include something like the following.
  `all: target1 target2 target3`

  Then typing `make all` (or just `make`, if `all` is listed first in the file) will build all of those things.
- To be build a specific target, type `make target`. For example, `make Figs/fig1.pdf`

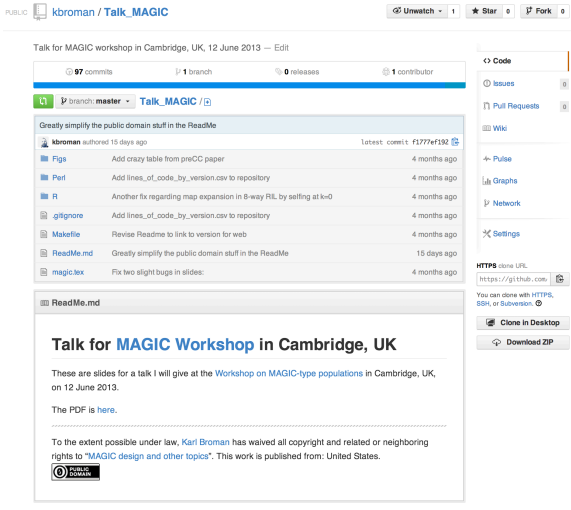
Write modular code

- Modular code is easier to understand, maintain, and reuse.
- Turn repeated code into functions
- Combine useful functions into a package or module

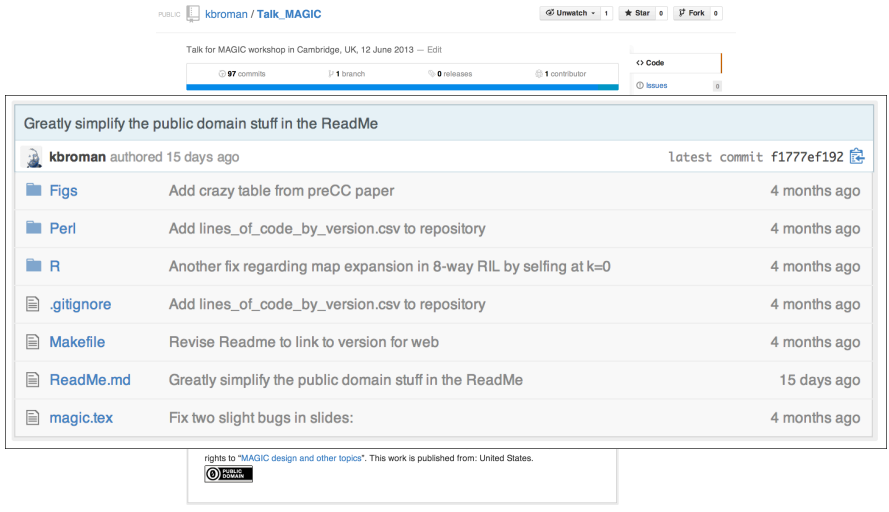
Keeping track of versions

- Google drive / Dropbox / Box
- Version numbers in file names
- Formal version control (e.g., git/GitHub)
  - Browse changes
  - Try new things without fear of breaking what works
  - Jump to the state of the project at any time point
  - Merge simultaneous changes from multiple people

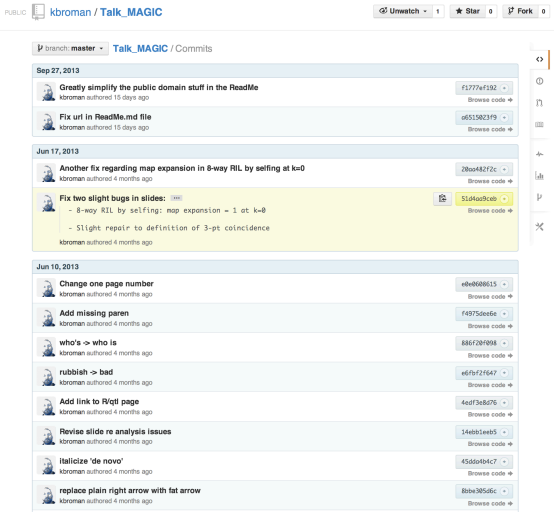
Version control (git/GitHub)



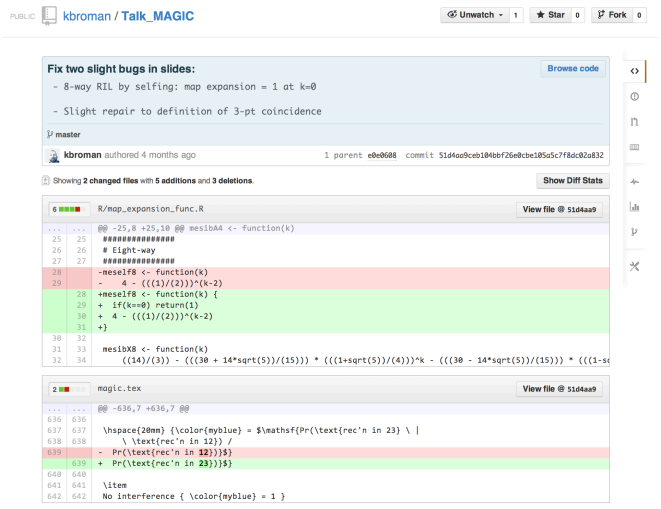
Version control (git/GitHub)



Version control (git/GitHub)



Version control (git/GitHub)



Version control (git/GitHub)

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

2 magic.tex

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

Backups

- Multiple places, including off-site
- Automatic

License your software

Pick a license, any license

– Jeff Atwood

Share your stuff

- Code
  - GitHub / BitBucket
  - Zenodo (archival, with DOIs)
- Data
  - Domain-specific repository (e.g., dbGAP)
  - General repository (e.g., github, figshare, zenodo, datadryad)
  - Institutional repository

Summary

1. Organize your project
2. Choose good names for things
3. Document what's what
4. Organize data as a rectangle
5. Metadata is data
6. Everything with a script
7. Even better: reproducible reports
8. Automate the process (GNU Make)
9. Write modular code (functions and packages)
10. Use version control (git/GitHub)
11. License your software
12. Share your data and code