

# 1: Intro to Reproducible Research

`bit.ly/SISBID3`

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?

# Basic principles

- ▶ Project encapsulated in one directory
- ▶ Organize and document
- ▶ Keep track of the **provenance** of all data files
- ▶ Everything via code
- ▶ Use a version control system
- ▶ Keep track of versions of dependencies
- ▶ Everything automated

# Why do we care?

- ▶ Avoid embarrassment
- ▶ More likely correct
- ▶ Save time, in the long run
- ▶ Greater potential for extensions; higher impact

# Try to avoid

- ▶ Open a file to extract as CSV
- ▶ Open a data file to do even a slight edit
- ▶ Paste results into the text of a manuscript
- ▶ Copy-paste-edit tables
- ▶ Copy-paste-adjust figures



# Problem: Variations across data files

- ▶ Different files (or parts of files!) may have different formats.
- ▶ Variables (or factor levels) may have different names in different files.
- ▶ The names of files may inconsistent.
- ▶ It's tempting to hand-edit the files. Don't!
- ▶ Create another meta-data file that explains what's what.

# Basic tools

- ▶ File organization and naming
- ▶ RMarkdown
- ▶ R packages
- ▶ Version control with git/GitHub
- ▶ Automation with Make

File organization and naming  
are powerful weapons against chaos.

– Jenny Bryan

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

(paraphrasing [Mark Holder](#))

# Organizing your stuff

```
Code/d3examples/  
  /Others/  
  /PyBroman/  
  /Rbroman/  
  /Rqtl/  
  /Rqtlcharts/  
Docs/Talks/  
  /Meetings/  
  /Others/  
  /Papers/  
  /Resume/  
  /Reviews/  
  /Travel/  
Play/  
Projects/AlanAttie/  
  /BruceTempel/  
  /Hassold_QTL/  
  /Hassold_Age/  
  /Payseur_Gough/  
  /PhyloQTL/  
  /Tar/
```

# Organizing your projects

```
Projects/Hassold_QTL/
```

```
  Data/
```

```
  Notes/
```

```
  R/
```

```
  R/Figs/
```

```
  R/Cache/
```

```
  Rawdata/
```

```
  Refs/
```

```
  Makefile
```

```
  Readme.txt
```

```
  Python/convertGeno.py
```

```
  Python/convertPheno.py
```

```
  Python/combineData.py
```

```
  R/prepData.R
```

```
  R/analysis.R
```

```
  R/diagnostics.Rmd
```

```
  R/ql_analysis.Rmd
```

# Organizing a paper

```
Docs/Papers/PhyloQTL/
```

```
    Analysis/
```

```
    Data/
```

```
    Figs/
```

```
    Notes/
```

```
    R/
```

```
    SuppFigs/
```

```
    ReadMe.txt
```

```
    Makefile
```

```
    phyloqtl.tex
```

```
    phyloqtl.bib
```

```
    Submitted/
```

```
    Reviews/
```

```
    Revised/
```

```
    Final/
```

```
    Proofs/
```

# Organizing a talk

```
Docs/Talks/SampleMixups/
```

```
  Figs/
```

```
  R/
```

```
  ReadMe.txt
```

```
  Makefile
```

```
  bmi2013.tex
```

```
  Old/
```



# Basic principles

- ▶ Develop your own system
- ▶ Put everything in a common directory
- ▶ Be consistent
  - directory structure; names
- ▶ Separate raw from processed data
- ▶ Separate code from data
- ▶ It should be obvious what code created what files, and what the dependencies are.
- ▶ No hand-editing of data files
- ▶ Don't use spaces in file names
- ▶ Use relative paths, not absolute paths
  - `../blah` not `~/blah` or `/users/blah`


## PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13  
20130227 2013.02.27 27.02.13 27-02-13  
27.2.13 2013.II.27.  $27\frac{1}{2}$ -13 2013.158904109  
MMXIII-II-XXVII MMXIII  $\frac{\text{LVII}}{\text{CCCLXV}}$  1330300800  
 $((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$  2013  
10/11011/1101 02/27/20/13  $\begin{matrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ & 5 & 6 & 7 & 8 \end{matrix}$  

# Painful bits

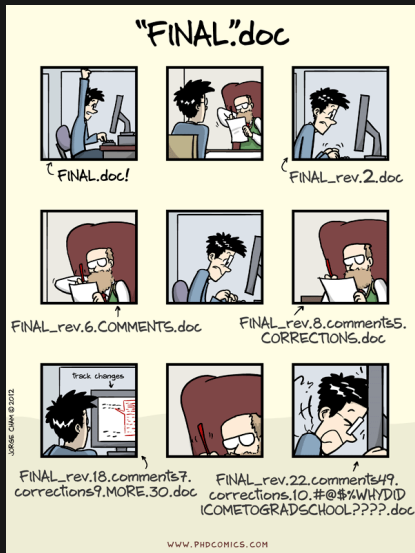
- ▶ Coming up with good names for things
  - Concise but informative
  - Code as verbs; data as nouns
  - Avoid spaces; avoid symbols except - and \_
- ▶ Stages of data cleaning
- ▶ Going back and redoing stuff
- ▶ Clutter of old stuff that you no longer need
- ▶ Keeping track of the order of things
  - dependencies; what gave rise to what

# Problem: 80 million side projects

```
$ ls ~/Projects/Attie
```

AimeeNullSims/	Deuterium/	Ping/
AimeeResults/	ExtractData4Gary/	Ping2/
AnnotationFiles/	ForFirstPaper/	Ping3/
Brian/	FromAimee/	Ping4/
Chr10adipose/	GoldStandard/	Play/
Chr6_extrageno/	HumanGWAS/	Proteomics/
Chr6hotspot/	Insulin/	R/
ChrisPlaisier/	Islet_2011-05/	RBM_PlasmaUrine/
Code4Aimee/	Lusis/	R_adipose/
CompAnnot/	MappingProbes/	R_islet/
CondScans/	Microarrays/	Rawdata/
D20_2012-02-14/	MultiProbes/	Scans/
D20_Nrm_2012-02-29/	NewMap/	SimsRePower/
D20_cellcycle/	Notes/	Slco1a6/
D20corr/	NullSims/	StudyLineupMethods/
Data4Aimee/	NullSims_2009-09-10/	eQTLPaper/
Data4Tram/	PepIns_2012-02-09/	transeQTL4Lude/

# Keep track of versions of things



[bit.ly/PhDComics\\_notFinal](http://bit.ly/PhDComics_notFinal)

# 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/	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_nqrnk_final.RData
adipose_mlratio_nqrnk_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_nqrnk_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_nqrnk_final.RData	liver_mlratio_final.RData
gastroc_prcomp.RData	liver_mlratio_nqrnk_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_nqrnk_final.RData	pmark.RData
hypo_mlratio_nqrnk_final_old.RData	rbm_final.RData
hypo_omit.RData	

And don't forget...

Backups



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

– Keith Baggerly