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/qtl_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`

xkcd.com/1179

Painful bits

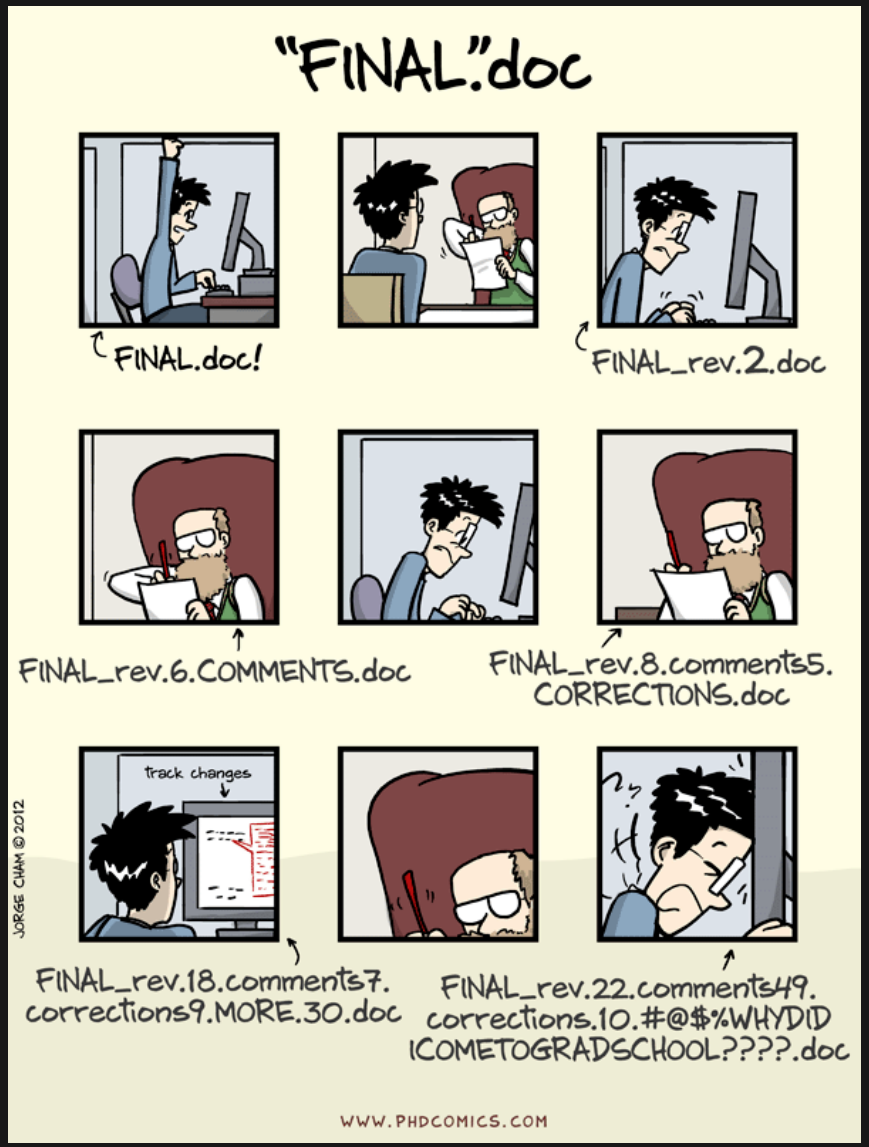
- Coming up with good names for things
  - Concise but informative
  - Code as verbs; data as nouns
  - Avoid spaces; avoid symbols except - and _
- Stages of data cleaning
- Going back and redoing stuff
- Clutter of old stuff that you no longer need
- Keeping track of the order of things
  - dependencies; 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/
D2O_2012-02-14/         MultiProbes/            Scans/
D2O_Nrm_2012-02-29/     NewMap/                 SimsRePower/
D2O_cellcycle/          Notes/                  Slco1a6/
D2Ocorr/                NullSims/               StudyLineupMethods/
Data4Aimee/             NullSims_2009-09-10/    eQTLPaper/
Data4Tram/              PepIns_2012-02-09/      transeQTL4Lude/
```

Keep track of versions of things



bit.ly/PhDComics_notFinal

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
```

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
```

And don’t forget...

Backups

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

– Keith Baggerly