Collaborating reproducibly

Karl Broman

Biostatistics & Medical Informatics
Univ. Wisconsin–Madison

kbroman.org
github.com/kbroman
@kwbroman

Slides: bit.ly/rrcollab

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?

Which image goes with which experiment?

“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

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. License your software

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
```

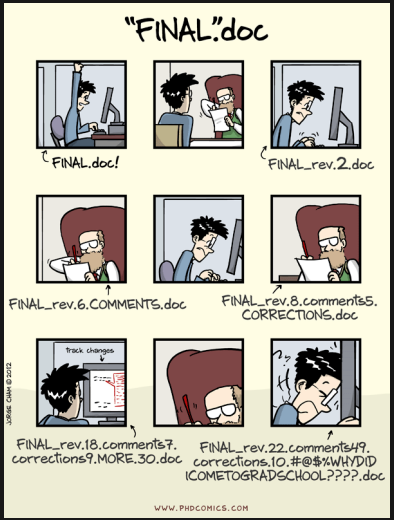
Organize your project

```
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
```

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/
```

No “final” in file names



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.
```

kbroman.org/steps2rr

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. License your software

Collaboration

Collaboration

- Do more, by working in parallel
- Do more, through diversity of ideas and skills
- Reproducible pipelines have immediate advantages
- Tests of reproducibility
- Code review

Challenges in collaborations

- Shared vision?
- Compromise
- Coordination
- Communication
- Sharing code and data
- Synchronization

Challenges in collaborations

- Shared vision?
- Compromise
- Coordination
- Communication
- Sharing code and data
- Synchronization
- Weakest link?

Genetics of metabolic disease in mice

Alan Attie, UW-Madison, Biochemistry

Karl Broman, UW-Madison, Biostat & Med Info

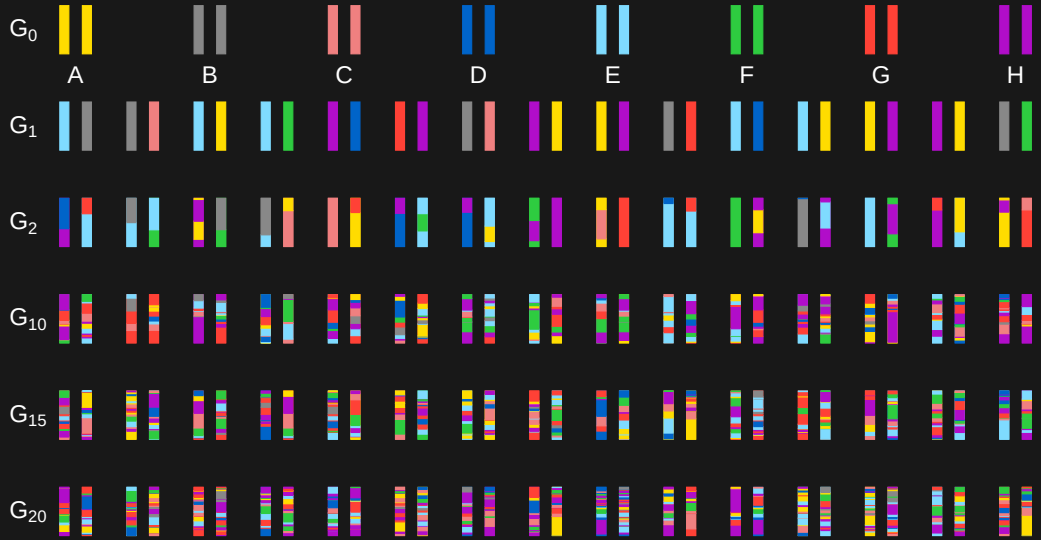
Gary Churchill, Jackson Lab

Josh Coon, UW-Madison, Chemistry

Federico Rey, UW-Madison, Microbiology

Brian Yandell, UW-Madison, Statistics

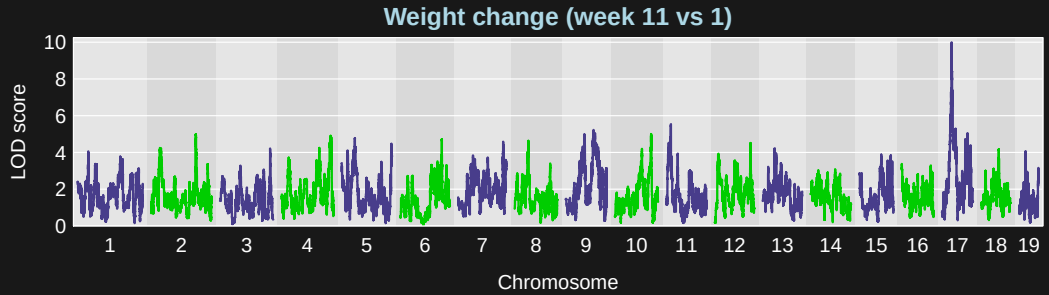
Diversity outbred mice

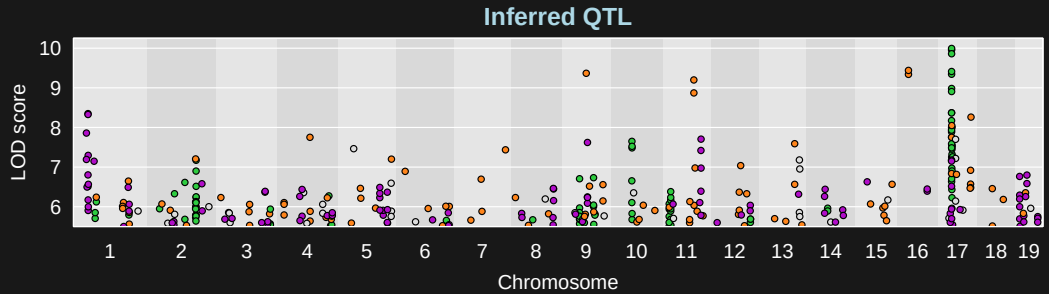


Data

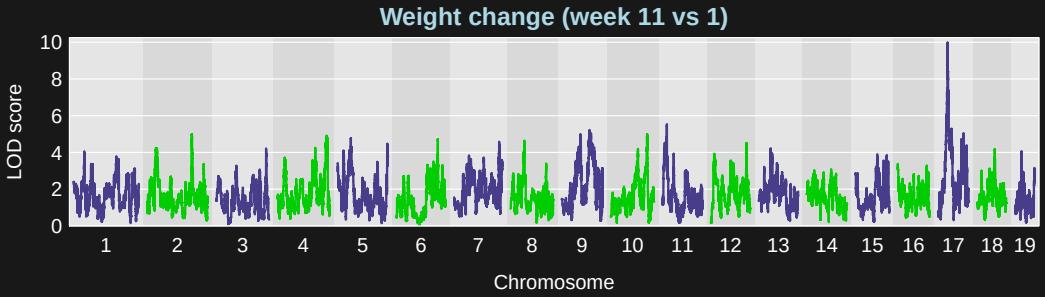
- 500 DO mice
  - generations 17–23
  - high fat, high sugar diet
- GigaMUGA SNP arrays
  - 140k SNPs
- Clinical traits
  - Weekly body weight
  - Glucose tolerance test
  - Longitudinal serum samples
  - ex vivo islet insulin secretion
- Islet gene expression by RNA-seq
- Proteins by mass spec
- Lipids by mass spec
- Gut microbiome
  - 16S RNA
  - metagenomic data

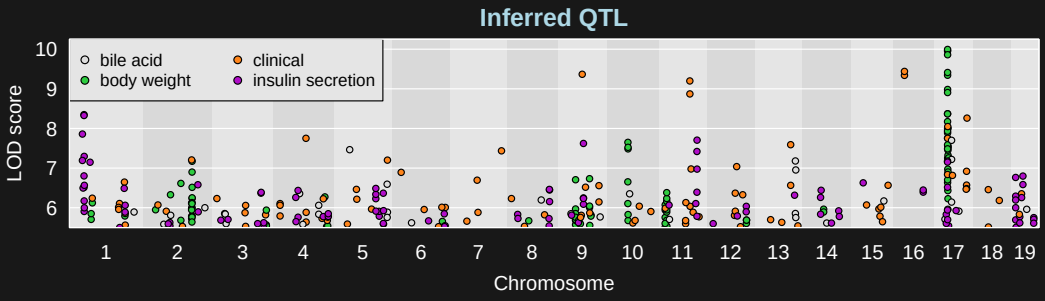
Genome scans





Genome scans





Challenges

(totally hypothetical)

“Could we meet to talk about the data file structure?”

“Could we meet to talk about the data file structure?”

“No.”

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

“Wait, these results seem to be based on the older SNP map.”

“Could you write the methods section?”

“But I didn’t do the work,
and we don’t have the code that was used.”

“My data analyst has taken a job at Google.”

“Could you do these analyses? X said they would, but they’re not responding to my emails.”

Shared vision

- Publication
- Code & data sharing
- Who will do what
- Timeline
- Ongoing sharing of methods, results

Shared workspace

- Project structure
- Data and metadata formats
- Software environment
- Automated sync (or it won't happen)

Technology for sharing

- Data
  - figshare
  - dropbox / box / google drive
- Code
  - github / bitbucket
- Pipeline / workflow
  - make / drake / snakemake / rake
- Full environment
  - docker containers
  - `mybinder.org` / `wholetale.org`

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

– Keith Baggerly

The second-most important tool is training.

– me

Slides: bit.ly/rrcollab

kbroman.org

github.com/kbroman

@kwbroman