Data management is fun
What is science?

The study of the natural world through observation and experiment.

Reproducible study.
“Prove it” isn’t an insult, it is a compliment.

Somebody cares enough about your result that they want to be sure it is right.
Is it OK to do work that isn’t reproducible?

No.
Reasons for lack of reproducibility include:

Poor organization.
Poor communication.
Fraud.
Reasons for lack of reproducibility include:

Poor organization.
Data management is the most critical aspect of organization.
Data management isn’t part of wrapping a project up.

It is a critical part of getting a project done.
Many scientists keep detailed records of data acquisition, but not data management or analysis.
Effective data management strategies
GET ALL OF THE DATA OUT OF YOUR LAB RIGHT NOW.
Why?
When most people think of data management they think first of data storage.
Step 1 - Calculate need

Storage need = (dataset size) x (number of datasets)
Step 2 - Calculate number of hard drives

Number of hard drives = (storage need) / (hard drive size)
Step 3 - Buy hard drives
Step 4 - Stack hard drives on desk
Step 5 - Buy more hard drives
Step 6 - Stack more hard drives on desk
Step 8 - Stack more hard drives on desk
Why doesn’t this work?

The user (ie, you) has to take care of a lot of details

Track hard drives
Make sure they are backed up
Make sure they aren’t stolen
Archive them long term
This is a bad data management plan.
You can’t do as good a job as CCV, Amazon, Google, the Library, github, NCBI, FigShare, DataDryad, or the thousand of other cloud services at managing your data.
Data management isn’t just about storage!
- Backup
- Versioning
- Metadata
- Accessibility
- Integration with analysis tools
Our data analysis ecosystem

- Oscar
- Google docs
- git
Analyses and storage on Oscar
Constrained file tree

All raw data go into a central repository

Common analyses (e.g., assembly) go into a central repository

Users have their own folders for specialized stuff
Shared access

Set permissions so that all users have access to all files in data directory
Always be thinking about how to minimize your disk footprint

Storage is about $1 per GB

Never copy data files to analysis directories

Don’t save analyses you don’t need

Delete intermediate analysis files

Far cheaper to keep scripts and regenerate analyses than to keep analyses
Documentation on Google Docs

<table>
<thead>
<tr>
<th>Title</th>
<th>Owner</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly Notebook.doc</td>
<td>me</td>
<td>10:50 pm</td>
</tr>
<tr>
<td>To order</td>
<td>me</td>
<td>Oct 17</td>
</tr>
<tr>
<td>Mollusc Taxon Sampling</td>
<td>me</td>
<td>Oct 17</td>
</tr>
<tr>
<td>StellaExpressionOctober2011</td>
<td>Rebecca Helm</td>
<td>Oct 17</td>
</tr>
<tr>
<td>Smith Lab Expression</td>
<td>Rebecca Helm</td>
<td>Oct 17</td>
</tr>
<tr>
<td>Deep sequencing run stats</td>
<td>me</td>
<td>Oct 17</td>
</tr>
<tr>
<td>rna extractions</td>
<td>me</td>
<td>Oct 17</td>
</tr>
<tr>
<td>Specimen data</td>
<td>me</td>
<td>Oct 17</td>
</tr>
<tr>
<td>specimen data</td>
<td>Dunn Lab</td>
<td>Oct 17</td>
</tr>
</tbody>
</table>
Documentation

Data and analyses are a liability rather than an asset if they aren’t well documented.

Documentation should be realtime, not something that is done after analyses.

Good documentation is a powerful teaching and learning tool.
Create a collection in google docs, share it with your lab members

Create new analysis notebooks for each project

Add each notebook to the lab collection
Paste commands, notes, plots, etc, right into the doc
Code management with git
git is a:

- Distributed software revision control system
- Allows you to organize all lab software in a single central repository
- Can write and use software in the repository on any computer
git allows you to:

- Revert to an earlier version
- Store verbose comments on your code development
- Fork and merge projects
<table>
<thead>
<tr>
<th>Subject</th>
<th>Author</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>cleanup for snume_rastqc.py</td>
<td>Casey Dunn</td>
<td>2011-08-16 13:22:35</td>
</tr>
<tr>
<td>cleanup for shuffle_fastqc.py</td>
<td>Casey Dunn</td>
<td>2011-08-16 13:17:41</td>
</tr>
<tr>
<td>Merge branch 'master' of <a href="http://cdunn.cc.ccn.edu/git/labcode">http://cdunn.cc.ccn.edu/git/labcode</a></td>
<td>Casey Dunn</td>
<td>2011-08-16 13:17:41</td>
</tr>
<tr>
<td>more edits</td>
<td>Stephen Smith</td>
<td>2011-08-09 10:59:59</td>
</tr>
<tr>
<td>adding some graph things</td>
<td>Stephen Smith</td>
<td>2011-08-09 10:57:32</td>
</tr>
<tr>
<td>Merge <a href="http://cdunn.cc.ccn.edu/git/labcode">http://cdunn.cc.ccn.edu/git/labcode</a></td>
<td>Stephen Smith</td>
<td>2011-08-09 10:56:18</td>
</tr>
<tr>
<td>more graphs</td>
<td>Stephen Smith</td>
<td>2011-08-09 10:56:18</td>
</tr>
<tr>
<td>more graphs</td>
<td>Stephen Smith</td>
<td>2011-08-04 10:58:31</td>
</tr>
<tr>
<td>cleaned up shuffle_fastqc.py</td>
<td>Casey Dunn</td>
<td>2011-07-21 13:17:09</td>
</tr>
<tr>
<td>Merge branch 'master' of <a href="http://cdunn.cc.ccn.edu/git/labcode">http://cdunn.cc.ccn.edu/git/labcode</a></td>
<td>Casey Dunn</td>
<td>2011-07-25 16:41:45</td>
</tr>
<tr>
<td>added some cluster and model things</td>
<td>Stephen Smith</td>
<td>2011-07-25 16:15:30</td>
</tr>
<tr>
<td>gene count generation for illumina 1.6 and SOLID SAGE alignments.</td>
<td>Casey Dunn</td>
<td>2011-07-25 16:15:17</td>
</tr>
<tr>
<td>cleaned up shuffle_fastqc.py</td>
<td>Stephen Smith</td>
<td>2011-07-25 16:15:17</td>
</tr>
<tr>
<td>Merge branch 'master' of <a href="http://cdunn.cc.ccn.edu/git/labcode">http://cdunn.cc.ccn.edu/git/labcode</a></td>
<td>Casey Dunn</td>
<td>2011-07-20 23:38:03</td>
</tr>
<tr>
<td>more raxml phylobayes scripts</td>
<td>Stephen Smith</td>
<td>2011-07-14 19:18:34</td>
</tr>
<tr>
<td>just a nex to phy script</td>
<td>Stephen Smith</td>
<td>2011-07-11 11:04:21</td>
</tr>
<tr>
<td>added qseq2fastaq.pl</td>
<td>Casey Dunn</td>
<td>2011-07-01 02:37:07</td>
</tr>
<tr>
<td>expression_on_trees.R - added functionality to prune tips from</td>
<td>Casey Dunn</td>
<td>2011-07-01 02:37:07</td>
</tr>
<tr>
<td>finished get_taxa function in expression_on_trees.R</td>
<td>Casey Dunn</td>
<td>2011-07-01 16:58:20</td>
</tr>
<tr>
<td>Added barebones of sff_trim.py, which will reset trim points in sff</td>
<td>Casey Dunn</td>
<td>2011-06-30 21:15:12</td>
</tr>
<tr>
<td>454_adaptor_trim_file_investment.py fixed so that front trim point is</td>
<td>Casey Dunn</td>
<td>2011-06-20 14:18:43</td>
</tr>
<tr>
<td>454_adaptor_trim_file_investment.py updated to include all 454 MID</td>
<td>Casey Dunn</td>
<td>2011-06-20 14:18:43</td>
</tr>
<tr>
<td>added template for get_taxa function</td>
<td>Casey Dunn</td>
<td>2011-06-24 16:47:48</td>
</tr>
<tr>
<td>expression_on_trees.R modified so that all fields in</td>
<td>Casey Dunn</td>
<td>2011-06-23 12:03:09</td>
</tr>
<tr>
<td>minor save to expression_on_trees</td>
<td>Casey Dunn</td>
<td>2011-06-22 17:07:29</td>
</tr>
<tr>
<td>fixed assignment of phydata list to dataframe in expression_on_trees</td>
<td>Casey Dunn</td>
<td>2011-06-22 17:01:32</td>
</tr>
<tr>
<td>some phylobayes things</td>
<td>Stephen Smith</td>
<td>2011-06-21 15:42:08</td>
</tr>
<tr>
<td>minor updates</td>
<td>Stephen Smith</td>
<td>2011-06-01 16:10:48</td>
</tr>
<tr>
<td>had not stored last record, fixed that shuffle_fastqc.py</td>
<td>cdunn</td>
<td>2011-05-27 15:26:28</td>
</tr>
<tr>
<td>progress on shuffle_fastqc.py</td>
<td>cdunn</td>
<td>2011-05-27 15:11:15</td>
</tr>
<tr>
<td>pulled over round-support.py from old repository; it hadn't been</td>
<td>cdunn</td>
<td>2011-05-26 15:22:50</td>
</tr>
</tbody>
</table>

108 commits loaded
bitbucket.org is great for hosting git central repositories.
practical computing
for biologists

Steven H. D. Haddock
The Monterey Bay Aquarium Research Institute,
and University of California, Santa Cruz

Casey W. Dunn
Department of Ecology and Evolutionary Biology,
Brown University
Parting thoughts...

Much of science is secretive, even though we claim to strive for openness.

Both the costs of secrecy and the benefits of openness are often underestimated.
Parting thoughts...

Data management, analysis, and communication are becoming seamless.

This will reshape the way we conduct and communicate science.
Why is effective data management fun?

- Many fewer headaches
- Real-time project overview
- Faster discovery
Further reading
Retraction Watch

A Serbian Sokal? Authors spoof pub with Ron Jeremy and Michael Jackson references
with 18 comments

What do porn star Ron Jeremy, Max Weber and Michael Jackson have in common? Very little — except the three names appear in the list of references for a recent hoax paper by a group of Serbian academics who, fed up with the poor state of their country's research output, scammed a Romanian magazine by publishing a completely fabricated article.

The paper is replete with transparent gimmicks — obvious, that is, had anyone at the publication been paying attention — including a reference to the scholarship of Jackson, Weber, Jeremy and citations to new studies by Bernoulli and Laplace, both dead more than 180 years (Weber died in 1920). They also throw in references to the "Journal of Modern Illogical Studies," which to the best of our knowledge does not and never has existed (although perhaps it should), and to a researcher named, dubiously, "A.S. Hole." And, we hasten to add, the noted Kazakh polymath B. Sagdiyev, otherwise known as Borat. Read the rest of this entry »

Should science put up with sloppiness?
with 16 comments

That's the question we pose in our newest column in LabTimes, based on some recent cases we've covered:

"The implication seems to be that as long as researchers can pass off their mistakes as sloppiness, rather than intentional misconduct, they should be forgiven and carry on their work. We're with that logic, to a point; after all, we've argued before that due process is much too important, no matter how apparently damning the evidence is. And as long as corrections and retraction notices are detailed, telling the whole story, science and the public are served."

Read the rest of this entry »

Dutch anthropologist Mart Bax faked 61 papers, says university
Misconduct is often shrouded in sloppiness.

It can therefore be difficult to tell misconduct from sloppiness.
Herd immunity

Fraud by its nature needs secrecy.

Reducing secrecy (which requires effective data management) provides fewer shadows for fraud to hide in.