Data Management and Accessibility

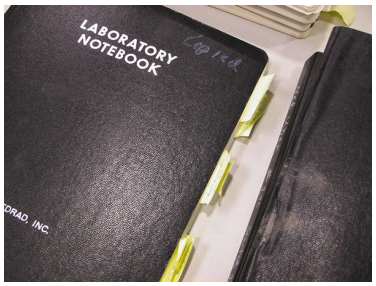
Deanna M. Church
Senior Director of Genomics and Content
Personalis, Inc

Disclosures

Deanna Church works for Personalis, Inc. A company that provides whole genome and whole exome sequencing and analysis services.

The nature of data management has changed

Was…



- Notes
- Pictures of gels
- Autoradiograms

Now…

- Notes
- Data images
- Sequence files

http://www.sxc.hu/photo/1072645
http://www.hcii.cmu.edu/M-HCL/2006/MEDRADProject/images/notebook.jpg

Protocols

Preparation of Electrocompetent *E. coli* (i.e. DH5α)

revised 2/24/96

Before starting procedure, prepare/chill the following:

1 L	2XYT media (no antibiotics!), store at RT
1 L	chilled autoclaved dH_20, stored in 4° cold room.
100ml	chilled 10% glycerol / dH_20 solution, store at 4°C.
4	Corning 250ml pointed bottles (w/ orange caps) with white adapters chilled at 4°C. (used in Beckman tabletop GPR with GH-3.7 rotor, Church 64) Rinse bottles first with 95% EtOH, then really well with dH_20. No soap!

Step 1. Preferably, select single colony of *E. coli* from fresh LB plate for inoculating a 10 ml 2XYT overnight (O/N) starter culture. Alternatively, streak out frozen glycerol stock of bacterial cells onto LB plate, grow plate O/N, and then select single colony for starter culture. Grow 10 ml starter culture O/N in 37°C shaker (250rpm).

Step 2. Inoculate 1L of 2XYT media and place culture in 37° shaker. Grow cells and measure OD_{600} every 45min-1hr. When the OD_{600} equals 0.6-0.9 (log phase growth), remove the cells from the shaker and place on ice.

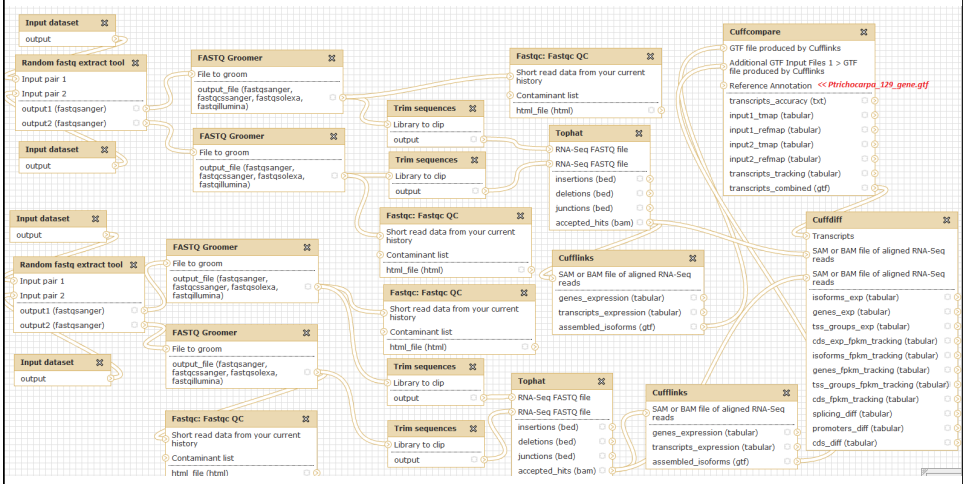
NOTE: It very important to keep the cells at 4°C (or on ice) for the remainder of the procedure.

Step 3. Split the 1L culture into four equal parts by pouring ~250ml of culture into each chilled 250ml Corning pointed bottle.

Step 4. Spin (#1) in GPR centrifuge at 4000rpm, 25min at 4°C.
(if you chose to use the J6/ JS-4.2 rotor (E. Davidson Lab), use 1L bottles , fill half full, spin 4000rpm, 20min, at 4°C.)

http://www.its.caltech.edu/~bjorker/protocols.html

Protocols: electronic workflows

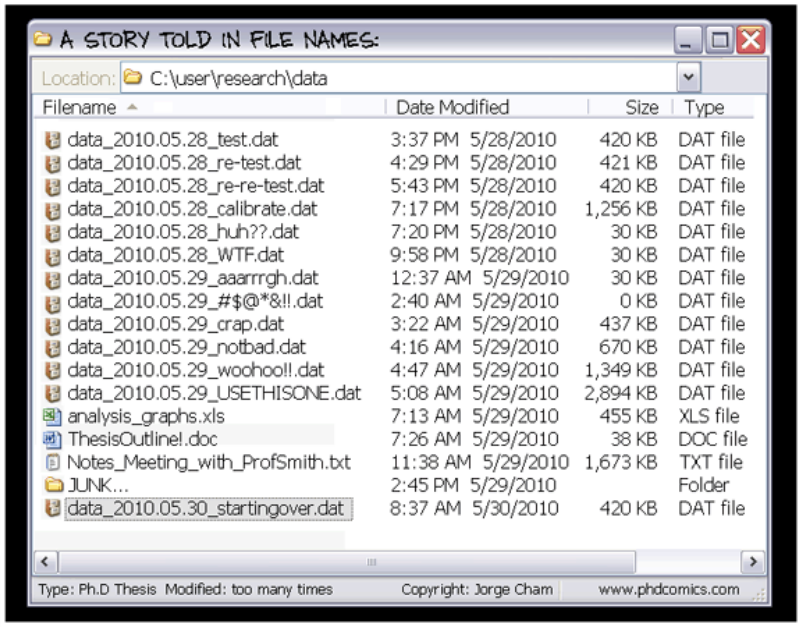


http://galaxy.popgenie.org

Electronic protocols: workflows

- Often require a Unix environment
- Often run on the command line
- Not always easy to make work

Getting a handle on data



http://www.phdcomics.com/archive.php?1323

Getting a handle on data: version control

SVN example

Directory revision: 406708 (of 406823)
Sticky Revision: Set

File	Rev.	Age	Author	Last log entry
Parent Directory				
DBAction.pl	404199	3 weeks	rafanovi	ctgoverlap migrated from CVS;JIRA:SVN-55
FindBestHit.pl	404199	3 weeks	rafanovi	ctgoverlap migrated from CVS;JIRA:SVN-55
GetJiraIssues.pl	405887	6 days	church	CG-1028: Issue asn.1 now created with issue mapping
GetResolvedIssues.pl	404199	3 weeks	rafanovi	ctgoverlap migrated from CVS;JIRA:SVN-55
MakeBigChrGraphic.pl	404199	3 weeks	rafanovi	ctgoverlap migrated from CVS;JIRA:SVN-55
MakeCertXML.pl	404199	3 weeks	rafanovi	ctgoverlap migrated from CVS;JIRA:SVN-55
MakeRegionReport.pl	404199	3 weeks	rafanovi	ctgoverlap migrated from CVS;JIRA:SVN-55
MakeReportConfig.pl	404199	3 weeks	rafanovi	ctgoverlap migrated from CVS;JIRA:SVN-55

Getting a handle on data: version control

SVN example

Links to HEAD: (view) (download) (as text) (annotate)
Sticky Revision: [ ] Set

Revision **405887** - (view) (download) (as text) (annotate) - [select for diffs]
Modified *Tue Jul 9 13:19:25 2013 EDT* (6 days, 23 hours ago) by *church*
File length: 46179 byte(s)
Diff to previous 405859

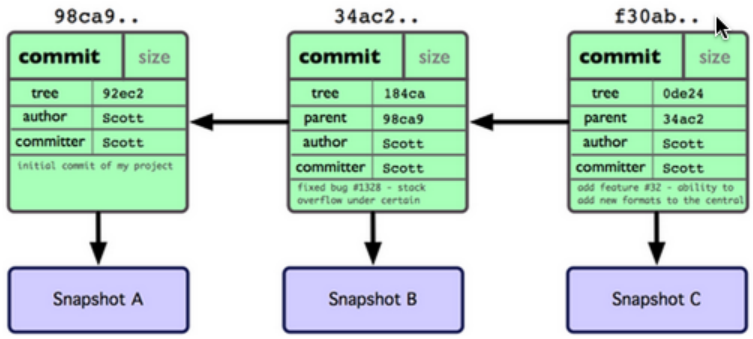
```
CG-1028: Issue asn.1 now created with issue mapping
```

Revision **405859** - (view) (download) (as text) (annotate) - [select for diffs]
Modified *Tue Jul 9 10:54:36 2013 EDT* (7 days, 1 hour ago) by *church*
File length: 45341 byte(s)
Diff to previous 405839

```
fixed CG-1488
Found bug related to missed mappings
```

Getting a handle on data: version control

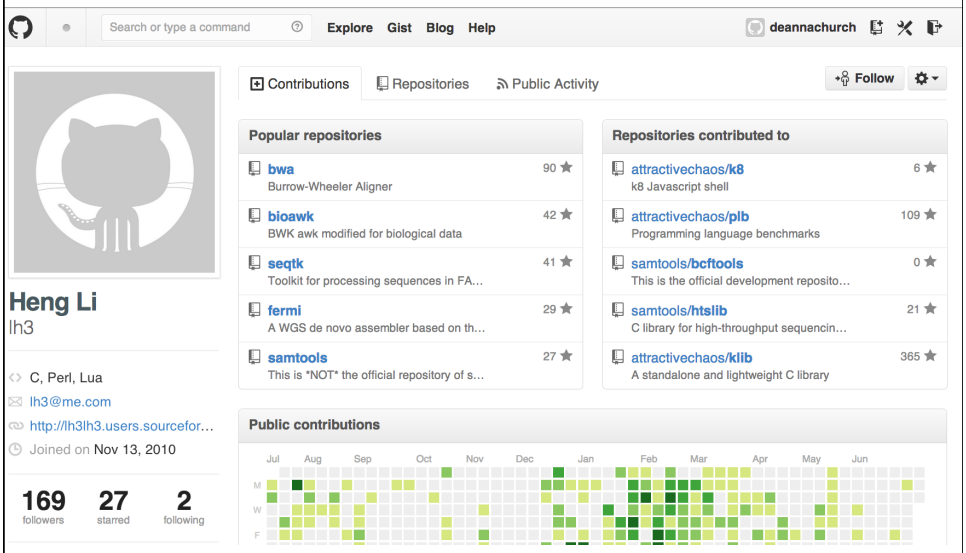
Git



http://nyuccl.org/pages/GitTutorial/

Getting a handle on data: socially

GitHub example

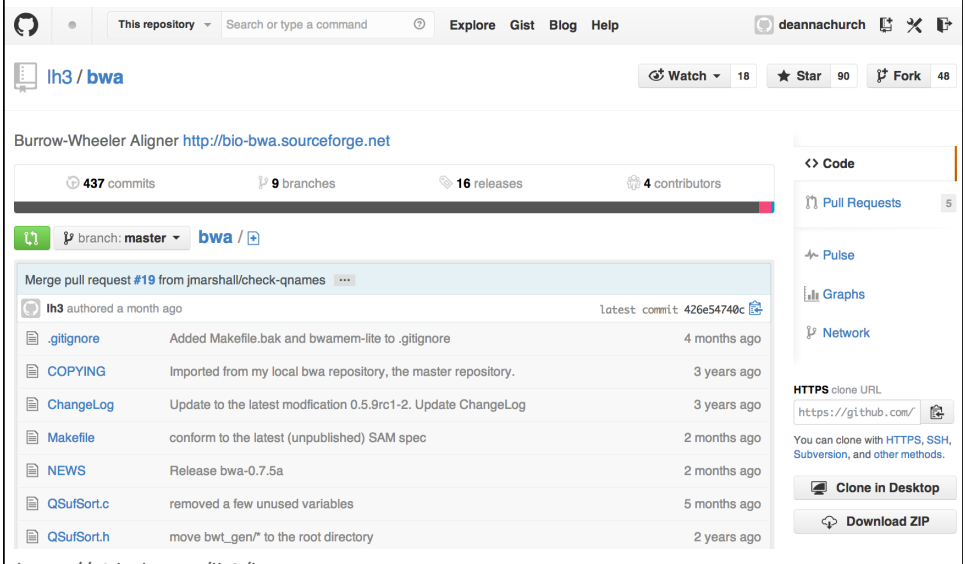


https://github.com/lh3

Bitbucket, Google Code

Getting a handle on data: version control

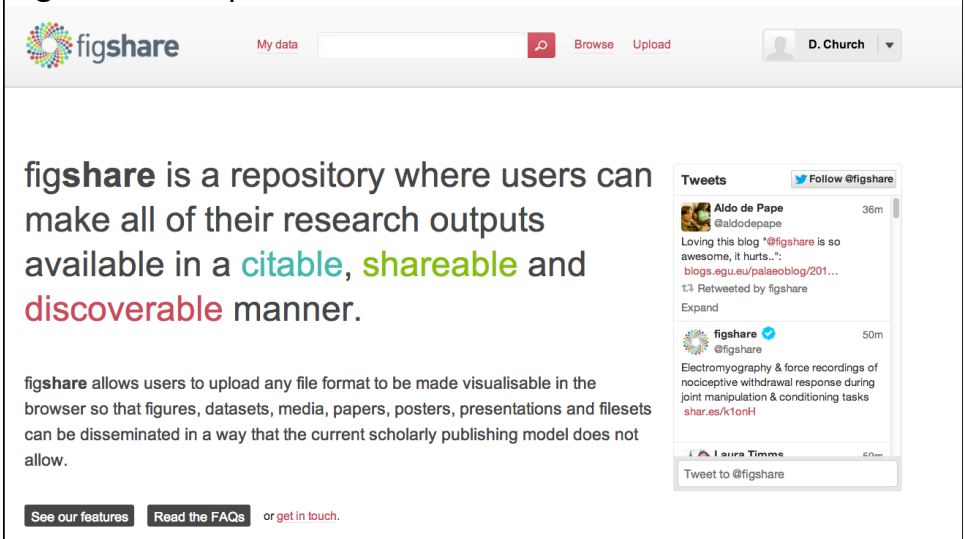
GitHub example



https://github.com/lh3/bwa

Data citing and sharing

Figshare example



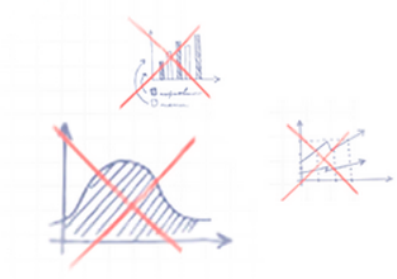
https://figshare.com/about

Data citing and sharing

Figshare example

Upload all formats

all research outputs welcomed – images, graphs, videos, datasets

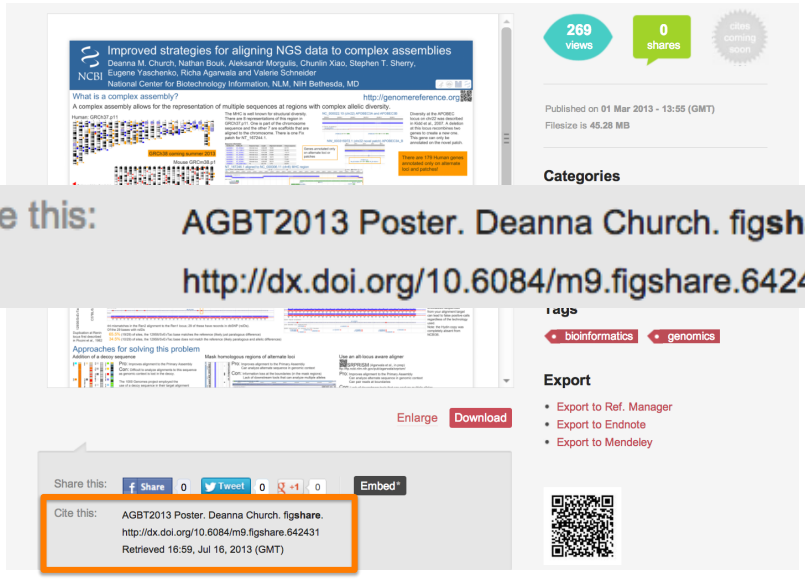


Publish negative data

all published research is citable

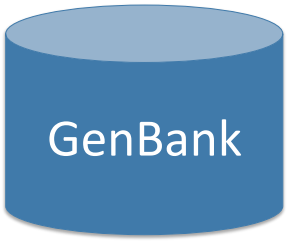
Data citing and sharing

Figshare example



Data citing and sharing

INSDC submission



INSDC
NCBI
DDBJ
ENA
European Nucleotide Archive

- Data in a common format
- Data in a single location (and mirrored)
- Most quality checked prior to deposition
- Robust data tracking mechanism (accession.version)
- Data owned by submitter

What is next?

- Talk to colleagues about their best practices
  - Solutions may exist in your institution
- Learn to program
  - Software carpentry (http://software-carpentry.org)
  - Codecademy (http://www.codecademy.com)
- Watch this video
  http://www.youtube.com/watch?v=N2zK3sAtr-4