




Data Management and Accessibility

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

 @deannachurch

Short Course in Medical Genetics 2014

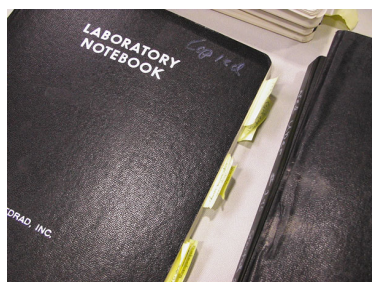
Disclosures

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

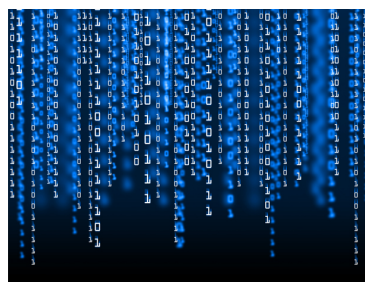


The nature of data management has changed

Was...



Now...



- 🌿 Notes
- 🌿 Pictures of gels
- 🌿 Autoradiograms

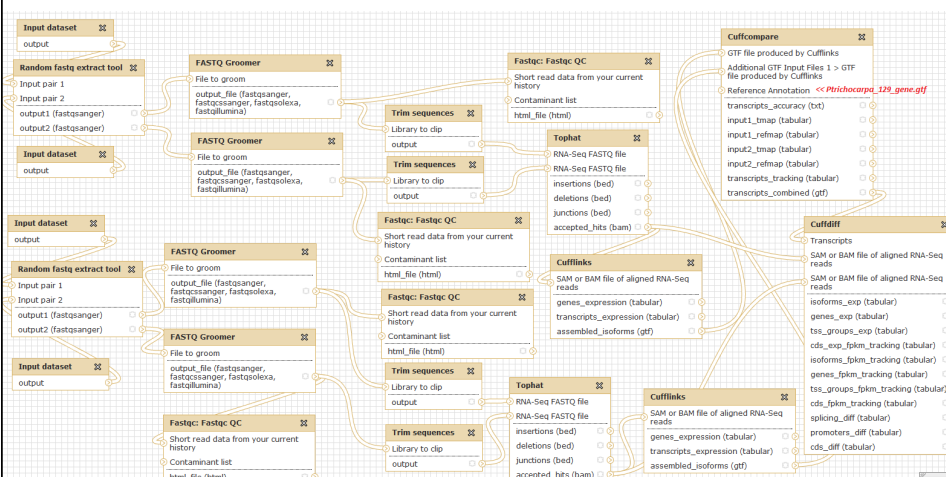
- 🌿 Notes
- 🌿 Data images
- 🌿 Sequence files

<http://www.sxc.hu/photo/1072645>

<http://www.hcii.cmu.edu/M-HCL/2006/MEDRADProject/images/notebook.jpg>

Personalis.

Electronic protocols: workflows



<http://galaxy.popgenie.org>

Personalis.

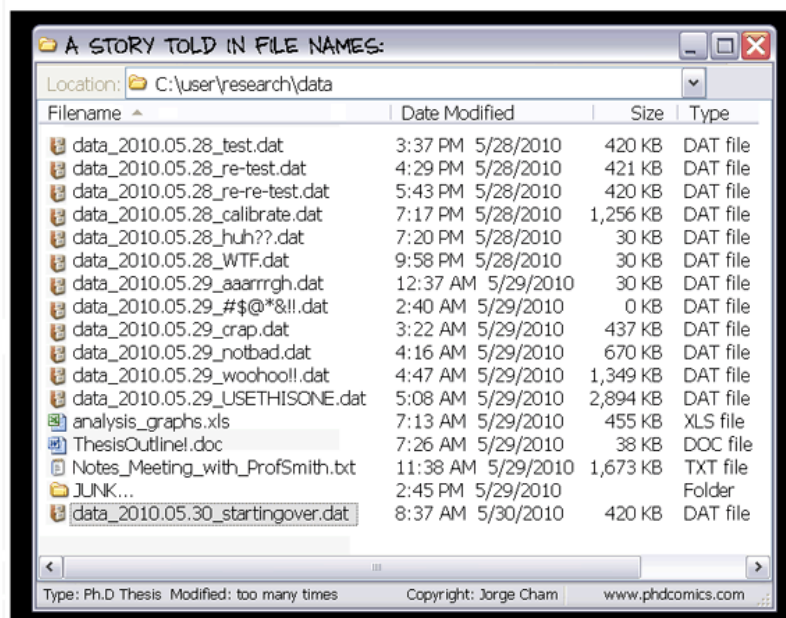
Electronic protocols: workflows

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



Personalis.

Getting a handle on data



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

alis.

Getting a handle on data: version control

SVN example

Directory revision: [406708](#) (of [406823](#))

Sticky Revision:

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:

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

GitHub example

Search or type a command

Explore Gist Blog Help

deannachurch

Contributions Repositories Public Activity

Follow

Heng Li
lh3

C, Perl, Lua
lh3@me.com
http://lh3lh3.users.sourceforge.net
Joined on Nov 13, 2010

169 followers 27 starred 2 following

https://github.com/lh3

Popular repositories

Repository	Stars
bwa Burrow-Wheeler Aligner	90 ★
bioawk BWK awk modified for biological data	42 ★
seqtk Toolkit for processing sequences in FA...	41 ★
fermi A WGS de novo assembler based on th...	29 ★
samtools This is "NOT" the official repository of s...	27 ★

Repositories contributed to

Repository	Stars
attractivechaos/k8 k8 Javascript shell	6 ★
attractivechaos/plib Programming language benchmarks	109 ★
samtools/bcftools This is the official development reposito...	0 ★
samtools/htslib C library for high-throughput sequencin...	21 ★
attractivechaos/klib A standalone and lightweight C library	365 ★

Public contributions

Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun

M
W
F

Personalis.

Getting a handle on data: version control

GitHub example

This repository Search or type a command

Explore Gist Blog Help

deannachurch

lh3 / bwa

Watch 18 Star 90 Fork 48

Burrow-Wheeler Aligner <http://bio-bwa.sourceforge.net>

437 commits 9 branches 16 releases 4 contributors

branch: master bwa

Merge pull request #19 from jmarshall/check-qnames

lh3 authored a month ago latest commit 426e54740c

File	Commit Message	Time
.gitignore	Added Makefile.bak and bwamem-lite to .gitignore	4 months ago
COPYING	Imported from my local bwa repository, the master repository.	3 years ago
ChangeLog	Update to the latest modification 0.5.9rc1-2. Update ChangeLog	3 years ago
Makefile	conform to the latest (unpublished) SAM spec	2 months ago
NEWS	Release bwa-0.7.5a	2 months ago
QSufSort.c	removed a few unused variables	5 months ago
QSufSort.h	move bwt_gen/" to the root directory	2 years ago

Code

Pull Requests 5

Pulse

Graphs

Network

HTTPS clone URL
<https://github.com/>

You can clone with HTTPS, SSH, Subversion, and other methods.

Clone in Desktop

Download ZIP

Personalis.

https://github.com/lh3/bwa

Data citing and sharing

Figshare example

figshare is a repository where users can make all of their research outputs available in a **citable**, **shareable** and **discoverable** manner.

figshare allows users to upload any file format to be made visualisable in the browser so that figures, datasets, media, papers, posters, presentations and filesets can be disseminated in a way that the current scholarly publishing model does not allow.

[See our features](#) [Read the FAQs](#) or [get in touch.](#)

<https://figshare.com/about>

Personalis.

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

Personalis.

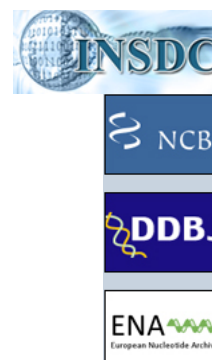
Data citing and sharing

Figshare example

The screenshot shows a Figshare poster page. At the top, the poster title is 'Improved strategies for aligning NGS data to complex assemblies' by Deanna M. Church, Nathan Bouk, Aleksandr Morgulis, Chunlin Xiao, Stephen T. Sherry, Eugene V. Yatsenko, Richa Agarwala, and Valerie Schneider. The poster content includes a diagram of a complex assembly and text explaining the challenges of aligning NGS data to such assemblies. On the right side of the poster, statistics show 269 views and 0 shares. Below the poster, a grey box contains the citation information: 'Cite this: AGBT2013 Poster. Deanna Church. figshare. http://dx.doi.org/10.6084/m9.figshare.642431'. Below this, there are social media sharing options (Facebook, Twitter, etc.) and a QR code. The bottom right corner of the slide shows the Personalis logo.

Data citing and sharing

INSDC submission



- 🌐 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>