MSBI 32400 – LAB 2

LARRY HELSETH PHD &
JASON EDELSTEIN

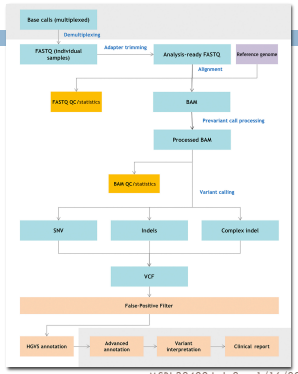
January 16, 2019

Why document?

- Titus Brown's manifesto on reproducibility: "How we make our papers replicable", July 15, 2014
  - http://ivory.idyll.org/blog/2014-our-paper-process.html
- Cited by MacArthur Lab when they released code to reproduce all the figures in their ExAC paper
  - https://macarthurlab.org/2016/03/17/reproduce-all-the-figures-a-users-guide-to-exac-part-2/
- **CAP, CLIA requirements to document bioinformatics workflow as a part of NGS analysis**
  - Roy, et al. "Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists." J Mol Diagn. 2018 Jan;20(1):4-27. doi: 10.1016/j.jmoldx.2017.11.003. Epub 2017 Nov 21. **PMID**: 29154853

Next-generation sequencing (NGS) bioinformatics pipeline

Roy, et al.
J Mol Diagn. Jan 2018



Documentation for Dry Lab Biology

- ~~Hand-written lab notebooks?~~
- Both textbooks reference the same recommendation for organizing your bioinformatics research/results in project folders with documentation inside each project folder.

```
student@MSBI32400Lab1:/data
File  Edit  View  Search  Terminal  Help
[student@MSBI32400Lab1 data]$ tree myproject/
myproject/
├── bin
├── data
├── doc
│   └── README_larry.md
├── results
└── src

5 directories, 1 file
```

Simple documentation using Markdown

5

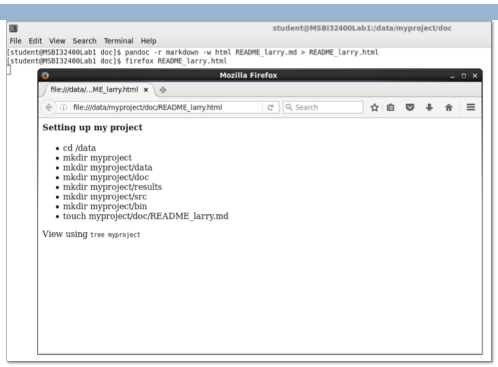
- See- Buffalo, pgs 31-35
  - http://proquestcombo.safaribooksonline.com.proxy.uchicago.edu/book/bioinformatics/9781449367480/2dot-setting-up-and-managing-a-bioinformatics-project/ch02_markdown_html
- Use simple characters for headers, bullets, hyperlinks, etc
- Convert from markdown to HTML using pandoc

```
[student@MSBI32400Lab1 doc]$ pandoc -r markdown -w html README_larry.md
<h4 id="setting-up-my-project">Setting up my project</h4>
<ul>
<li>cd /data</li>
<li>mkdir myproject</li>
<li>mkdir myproject/data</li>
<li>mkdir myproject/doc</li>
<li>mkdir myproject/results</li>
<li>mkdir myproject/src</li>
<li>mkdir myproject/bin</li>
<li>touch myproject/doc/README_larry.md</li>
</ul>
<p>View using <code>tree myproject</code></p>
[student@MSBI32400Lab1 doc]$
```

Better-Redirect to an HTML file using
pandoc –r markdown -w html README_larry.md > README_larry.html

View README_larry.html in Firefox

6



View example README_installation

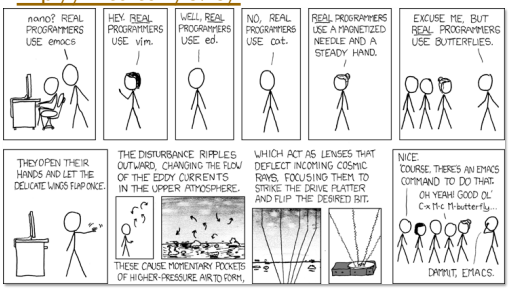
7

- I left a summary of the installation notes from configuring the VM in the student home directory
  - *Copy also available on Canvas Files/Lab 1*
- You can view the .md in an editor or run:
  - firefox /home/student/README_install_notes.html to view
- Be sure to add date created/installed somewhere in the document; if anyone edits the file it will look like it was newly created and you won't remember when you did what you did

XKCD's view on Linux editors

8

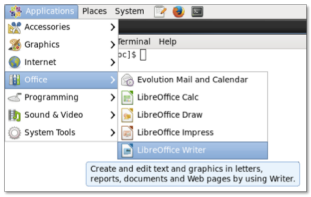
- http://xkcd.com/378/



Your work

9

- I don't care which editor you use, just don't use Microsoft Word or LibreOffice Writer



  - You can use gedit today, but remember that, at some point, you'll be connected to a remote Linux server (Amazon?) that doesn't have Gnome installed so learn some basic VIM or Nano commands!

The following editors are installed on VM

10

- VIM (vi) - Esc : h for help
- Nano - Ctrl-G for help
- Emacs – Help in menu bar
- Gedit – Help in menu bar

Capturing Linux commands is simple

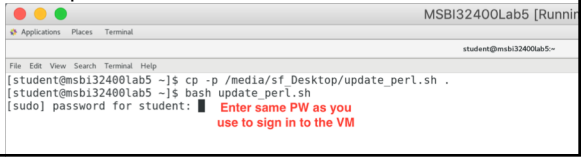
11

- Linux keeps a command history
  - Scroll back through commands with up/down cursor
  - Type 'history' to view most recent ~1000 commands
- Can cut and paste from console
- Better, use 'echo'
  - Scroll back to a command, then add echo and quotes
  - echo 'mkdir myproject/doc' >> README_larry.md
  - NB-Single > creates a new file (ERASING what was there!) so use >> to append
- Can type 'history >> README_larry.md' then edit lines you don't want
- Don't assume the history will "be there the next time"….

UPDATING VM FOR PERL

12

- Perl v5.16.3 installed on CentOS 7.5 VM but dependencies missing (https://www.bioinformatics.org/wiki/Edirect_Perl_library_dependency_problem)
- Install CPAN (Comprehensive Perl Archive Network software manager) and run a few updates
- Download Files/Lab2/**update_perl.sh** to your host desktop and run as shown below:



Answer some prompts

13

- Answer 'y' when asked or press **return** key to accept defaults. 'quit' to exit CPAN and continue

```
Dependencies Resolved

================================================================================
 Package                 Arch          Version                 Repository    Size
================================================================================
Installing:
 perl-CPAN               noarch        1.9800-293.el7          base         293 k
Installing for dependencies:
 gdbm-devel              x86_64        1.10-8.el7              base          47 k
 libdb-devel             x86_64        5.3.21-24.el7           base          38 k
 perl-Digest             noarch        1.17-245.el7            base          23 k
 perl-Digest-SHA         x86_64        1:5.85-4.el7            base          58 k
 perl-ExtUtils-Install   noarch        1.58-293.el7            base          74 k
 perl-ExtUtils-MakeMaker noarch        6.68-3.el7              base         275 k
 perl-ExtUtils-Manifest  noarch        1.61-244.el7            base          31 k
 perl-ExtUtils-ParseXS   noarch        1:3.18-3.el7            base          77 k
 perl-devel              x86_64        4:5.16.3-293.el7        base         453 k
 perl-local-lib          noarch        1.008010-4.el7          base          64 k
 systemtap-sdt-devel     x86_64        3.3-3.el7               base          74 k

Transaction Summary
================================================================================
Install  1 Package (+11 Dependent packages)

Total download size: 1.5 M
Installed size: 4.2 M
Is this ok [y/d/N]:
```


14

- Quit out of CPAN

```
Use of uninitialized value $deactivating in numeric eq (==) at /usr/share/perl5/vendor_perl/local/lib.pm line 381.
Use of uninitialized value $deactivating in numeric eq (==) at /usr/share/perl5/vendor_perl/local/lib.pm line 383.
Use of uninitialized value $options{"interpolate"} in numeric eq (==) at /usr/share/perl5/vendor_perl/local/lib.pm line 424.
Use of uninitialized value $options{"interpolate"} in numeric eq (==) at /usr/share/perl5/vendor_perl/local/lib.pm line 424.
Use of uninitialized value $options{"interpolate"} in numeric eq (==) at /usr/share/perl5/vendor_perl/local/lib.pm line 424.
export PERL_LOCAL_LIB_ROOT="$PERL_LOCAL_LIB_ROOT:/home/student/perl5";
export PERL_MB_OPT="--install_base /home/student/perl5";
export PERL_MM_OPT="INSTALL_BASE=/home/student/perl5";
export PERL5LIB="/home/student/perl5/lib/perl5:$PERL5LIB";
export PATH="/home/student/perl5/bin:$PATH";

Would you like me to append that to /home/student/.bashrc now? [yes]


commit: wrote '/home/student/.cpan/CPAN/MyConfig.pm'

You can re-run configuration any time with 'o conf init' in the CPAN shell
Terminal does not support AddHistory.

cpan shell -- CPAN exploration and modules installation (v1.9800)
Enter 'h' for help.

cpan[1]> quit
```

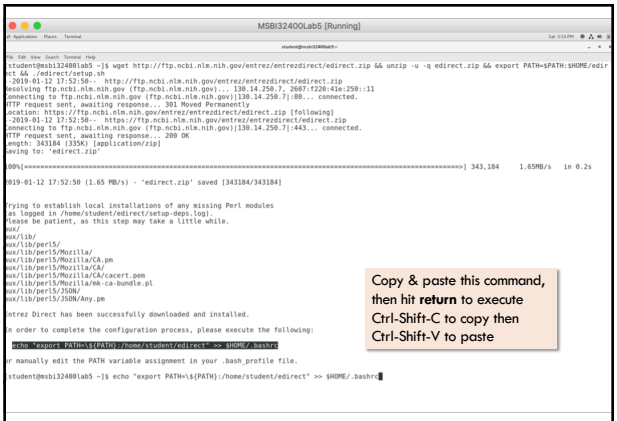

Installing NCBI Command Line Tools

15

- Go to your home folder on VM ('cd ~')
  - *You would have to install in /data if other users need access to these tools*
  - Enter the following:

wget http://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/edirect.zip && unzip -u -q edirect.zip && export PATH=$PATH:$HOME/edirect && ./edirect/setup.sh

REMEMBER TO ADD THESE COMMANDS TO YOUR README



Copy & paste this command, then hit **return** to execute
Ctrl-Shift-C to copy then
Ctrl-Shift-V to paste

Try a few of the examples using tools

17

- Create a new directory: mkdir –p /data/Lab2/doc
- From your /data/Lab2 folder (**should work if your PATH is right; if not, add ~/esearch/<command>**):
  - esearch –db pubmed –query "helseth dl AND collagen" | efetch –format pubmed > doc/example1.txt
  - esearch -db pubmed -query "bioinformatics [MAJR] AND software [TIAB]" | efetch -format xml | xtract -pattern PubmedArticle -block Author -sep " " -tab "\n" -element LastName,Initials | sort-uniq-count-rank > doc/bioinformatics_authors.txt
  - esearch -db protein -query 'NP_000509.1' | efetch -format fasta > doc/hbb.fasta

CAUTION-Copy & paste from **Microsoft PowerPoint** substitutes the wrong kind of dash (-- instead of -) so check carefully.

REMEMBER TO ADD THESE COMMANDS TO YOUR README

Sickle Cell Disease – 1 SNP

18

- https://www.nhlbi.nih.gov/health/health-topics/topics/sca
- Single mutation in HBB subunit causes the hemoglobin tetramer to aggregate when deoxygenated, forming strands within the red blood cells.
- A Glutamic acid ('E') is changed to a Valine ('V'), altering the way hemoglobin molecules interact

Viewing HBB in 1000 Genome Data

19

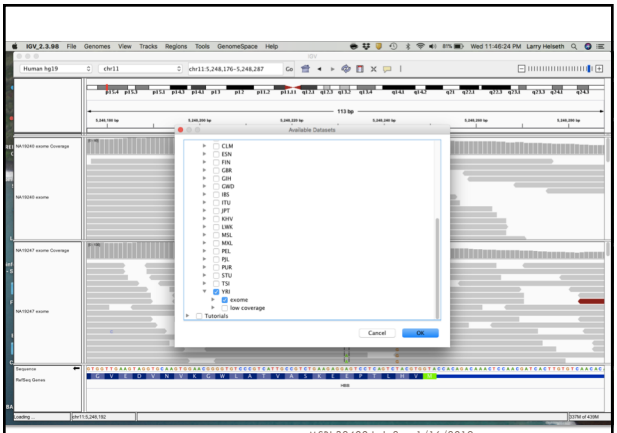
- YRI Trio (Yoruban) Trio from Ibadan, Nigeria
- NA19238, NA19239 and NA19240 are the YRI trio: mother, father, & daughter respectively
  - http://www.internationalgenome.org/data-portal/sample/NA19240
- 1000 Genome data, and reference links like dbSNP, are available to add "From server" in IGV

A Auton *et al. Nature* **526**, 68-74 (2015) doi:10.1038/nature15393

1000 Genome & dbSNP in IGV

20

- Click **File/Load from server** then expand **1000 Genome**
- Expand **Alignments,** scroll down and expand **YRI**
- Expand **exome**
- Scroll down and select **NA19238, NA19239** and **NA19240** then click **OK**
- Enter **HBB** in IGV then zoom in on exon 1
- Add **dbSNP** with **File/Load from server,** expand **Annotations,** expand **Variations and Repeats,** select **dbSNP 1.4.7** and click **OK**



Include in your write-up

22

- Who has variant at E7?
- Expand the dbSNP track. Which rsID(s) are associated with this variant at the E7 position?

Using NCBI tools to visualize HBSc

23

- This will not work on the Virtual Machine so you'll need to use your laptop.
- Go to OMIM.org and search for "sickle cell disease", then click on the second link (+ 141900. HEMOGLOBIN--BETA LOCUS; HBB)
- Click on the "Table View" in the left menu, then search the web page (Command-F on Mac, Ctrl-F Win) for "sickle", and click on the left link ('.0243')
- Note the first rs#. Open dbSNP (https://www.ncbi.nlm.nih.gov/snp/) and search for that rs#

Visualizing HBSc (cont)

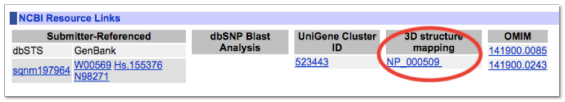
24

- Click on the 'Protein 3D' link below the sequence coordinates and HGVS entries
- Install Cn3D if it's not already installed, then click the "View Structure and Alignment in Cn3D" button to view the single HBB chain.
- Highlight Glu 6 ("E"), and note where it appears on the surface (spin molecule if needed). Use File/Export PNG to save to your Desktop (<username>_hbb.png)

NCBI is changing dbSNP

25

- Look for the NCBI Resource Links section:



- Follow the 3D structure link (not always available) then click to launch the CN3D viewer

Viewing final HBSc

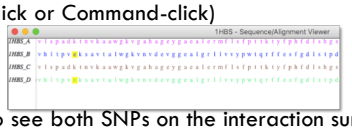
26

- Hemoglobin is tetrameric, so we need a different crystal structure. Go to Google and search "ncbi 1HBS" and click the first link
  - https://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?uid=1hbs
- Below "Interactions" on the right side Download Structure Data in "ASN.1 (Cn3D)" format

Viewing final HBSc (cont)

27

- Click on the 'e' in position 7 of the 2nd & 4th lines (Ctrl-click or Command-click)



- Spin to see both SNPs on the interaction surface.
- Use File/Export PNG to save your image as <username>_hbsc.png

Homework

28

- E-mail Jason (jasone@uchicago.edu) with "**Lab #2**" in the subject line
  - Your README_<your net id>.md
  - Your list of top bioinformaticians & your hbb.fasta
  - What you found about the YRI Trio
  - A screen shot after running 'ls' or 'tree' on your project directory
    - Either print screen or use Linux Applications/Take Screenshot (Can use VM Firefox to mail to yourself)
  - Cn3D image showing both Valines highlighted
- Please e-mail Jason before next class