MSBI 32400 – LAB 2

LARRY HELSETH PHD &
JASON EDELSTEIN

June 28, 2017

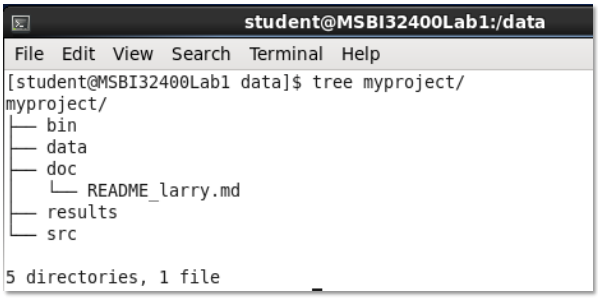
Why document?

- 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

Documentation for Dry Lab Biology

3

- ~~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.



Simple documentation using Markdown

4

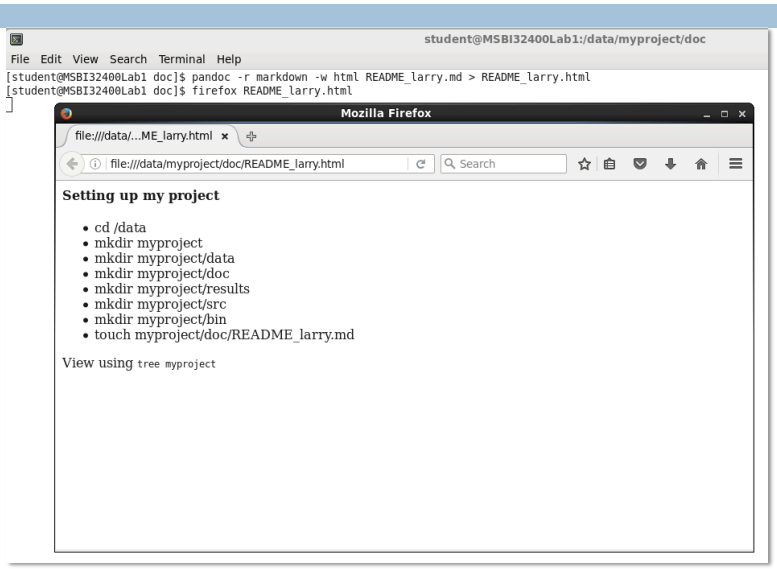
- See- Buffalo, pgs 31-35
  - http://proquestcombo.safaribooksonline.com.proxy.uchicago.edu/book/bioinformatics/9781449367480/2dot-setting-up-and-managing-a-bioinformatics-project/ch02_markdown_html
- Use simple characters for headers, bullets, hyperlinks, etc
- Convert from markdown to HTML using pandoc

```
[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

5



View example README_installation

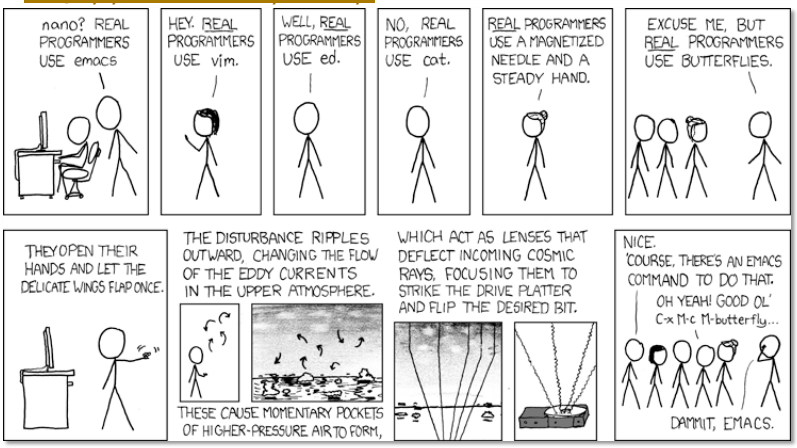
6

- ~~I left a summary of the installation notes from configuring the VM in /home/student/Downloads/~~
  - ***Copy available on Canvas Files/Lab 1***
- You can view the .md in an editor or run:
  - firefox /home/student/Downloads/ 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

7

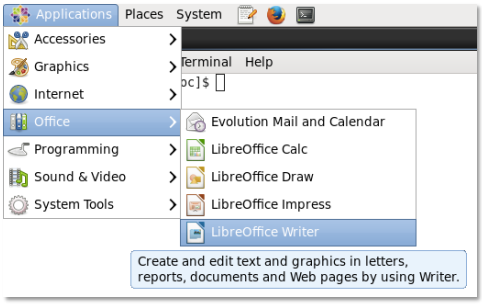
- http://xkcd.com/378/



Your work

8

- 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

9

- 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

10

- 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"….

Installing NCBI Command Line Tools

11

- Cf- Pevsner, Box 2.4, pgs 45-49. Download pevsner_box_2.sh from files/Lab 2 to your host
- Go to your home folder on VM ('cd ~')
  - Install in /data if other users need access to these tools
- Enter the following (copy from "pevsner_box_2.sh") :

```
perl -MNet::FTP -e \
  '$ftp = new Net::FTP("ftp.ncbi.nlm.nih.gov", Passive => 1); $ftp->login;
   $ftp->binary; $ftp->get("/entrez/entrezdirect/edirect.zip");'
unzip -u -q edirect.zip
rm edirect.zip
export PATH=$PATH:$HOME/edirect
./edirect/setup.sh
```

REMEMBER TO ADD THESE COMMANDS TO YOUR README

Try a few of the examples in the book

12

- From your myproject folder:
  - ~/edirect/esearch −db pubmed −query "pevsner j AND gnaq" | ~/edirect/efetch −format pubmed > doc/example1.txt
  - ~/edirect/esearch -db pubmed -query "bioinformatics [MAJR] AND software [TIAB]" | ~/edirect/efetch -format xml | xtract -pattern PubmedArticle -block Author -sep " " -tab "\n" -element LastName,Initials | sort-uniq-count-rank > doc/bioinformatics_authors.txt
  - ~/edirect/esearch -db protein -query 'NP_000509.1' | ~/edirect/efetch -format fasta > doc/hbb.fasta

  CAUTION-Sometimes copy & paste from Windows substitutes the wrong kind of dash (-- instead of -) so check carefully.

REMEMBER TO ADD THESE COMMANDS TO YOUR README

Sickle Cell Disease – 1 SNP

13

- 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

Using NCBI tools to visualize HBSc

14

- 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)

15

- 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)

Viewing final HBSc

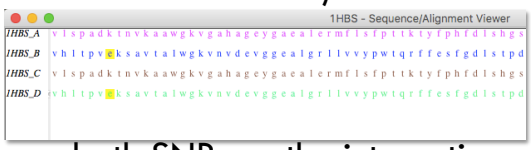
16

- 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)

17

- 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

18

- 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
  - A screen shot after running 'ls' or 'tree' (have Larry install) 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