Sharing Scripts & Data

C. Ryan Campbell

Duke University

c.ryan.campbell@duke.edu

28 Sept 2017

Overview

1. Goals
2. Gitting
   - Branching
   - Commit, Push & Pull
3. Data
   - Commandline Tools
   - SRA

Today's Goals

- Set up a group git

Today's Goals

- Set up a group git
  - branch

Today's Goals

- Set up a group git
  - branch
  - push and pull

Today's Goals

- Set up a group git
  - branch
  - push and pull
  - access files in cluster folder

Today's Goals

- Set up a group git
  - branch
  - push and pull
  - access files in cluster folder
- Download project data

Today's Goals

- Set up a group git
  - branch
  - push and pull
  - access files in cluster folder
- Download project data
  - wget, curl, sratoolkit

Data

- Many journals and grants require data to be public

Data

- Many journals and grants require data to be public
- PLOS, NSF, NIH

Data

- Many journals and grants require data to be public
- PLOS, NSF, NIH
- The information for where data can be found is often in:

Data

- Many journals and grants require data to be public
- PLOS, NSF, NIH
- The information for where data can be found is often in:
  - The “Methods” section of a paper

Data

- Many journals and grants require data to be public
- PLOS, NSF, NIH
- The information for where data can be found is often in:
  - The “Methods” section of a paper
  - An appendix

Data

- Many journals and grants require data to be public
- PLOS, NSF, NIH
- The information for where data can be found is often in:
  - The “Methods” section of a paper
  - An appendix
  - The author’s lab website

Data

- Many journals and grants require data to be public
- PLOS, NSF, NIH
- The information for where data can be found is often in:
  - The “Methods” section of a paper
  - An appendix
  - The author’s lab website
- We’ll talk about how to get data from each of these places

Downloand Commands

- `curl`

Downloand Commands

- `curl`
- `wget`

Downloand Commands

- `curl`
- `wget`
- `fastq-dump` - subset of the SRA Toolkit

Downloand Commands

- `curl`
- `wget`
- `fastq-dump` - subset of the SRA Toolkit
- Each is useful in different situations

curl

- Downloads a given url site

```
Usage:  curl [options...]  <url>
```

- Options include username & password
- Useful for sftp servers
- An example:

```
curl -o theESPNpage.html http://www.espn.com/
```

wget

- Downloads a given url site

```
Usage:  wget [OPTION]...  [URL]...
```

- Same options as curl
- An example:

```
wget http://www.bzip.org/1.0.6/bzip2-1.0.6.tar.gz
```

- Often used for downloading software (you'll see later...)

Sequence Read Archive

- Holds raw sequence data from published articles

Sequence Read Archive

- Holds raw sequence data from published articles
- Sorted by Experiment and Project

Sequence Read Archive

- Holds raw sequence data from published articles
- Sorted by Experiment and Project
- Use SRA Toolkit to access the files

Sequence Read Archive

- Often directly reference in papers

Sequence Read Archive

- Often directly reference in papers
- e.g. “data are available in SRR#########”

Sequence Read Archive

- Often directly reference in papers
- e.g. “data are available in SRR#########”
- Sometimes they are well organized, sometimes not

Sequence Read Archive

- Often directly reference in papers
- e.g. “data are available in SRR#########”
- Sometimes they are well organized, sometimes not
- You should be able to follow the paper’s methodology to separate the data into samples (if it wasn’t kept that way on SRA)

SRA Toolkit

- So, how do we download the data?

SRA Toolkit

- How do we get the software onto the cluster?

SRA Toolkit

- How do we use this file?

SRA Toolkit

- How do we use the toolkit?
- What is the toolkit (file, command, etc)?

Group Work

- You should have everything you need now to download your own data

Group Work

- You should have everything you need now to download your own data
- TIPS:

Group Work

- You should have everything you need now to download your own data
- TIPS:
  - Run the line of code in the terminal first so you can troubleshoot it

Group Work

- You should have everything you need now to download your own data
- TIPS:
  - Run the line of code in the terminal first so you can troubleshoot it
  - Instead of downloading the whole file use `head` to check output

Group Work

- You should have everything you need now to download your own data
- TIPS:
  - Run the line of code in the terminal first so you can troubleshoot it
  - Instead of downloading the whole file use `head` to check output
  - Once you're happy THEN submit to the cluster

Group Work

- You should have everything you need now to download your own data
- TIPS:
  - Run the line of code in the terminal first so you can troubleshoot it
  - Instead of downloading the whole file use `head` to check output
  - Once you're happy THEN submit to the cluster
  - This will save a lot of headaches and waiting

The End