Welcome to The Carpentries Etherpad!

This pad is synchronized as you type, so that everyone viewing this page sees the same text. This allows you to collaborate seamlessly on documents.

Use of this service is restricted to members of The Carpentries community; this is not for general purpose use (for that, try https://etherpad.wikimedia.org).

Users are expected to follow our code of conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

---

Data Carpentry - Genomics

Instructor Info

Amanda Devine
Email: DevineA@si.edu
Twitter: @amdvine (not super active...)

Mike Trizna
Email: triznam@si.edu
Twitter: @miketrizna

Workshop Participants

Please list your name, affiliation, and one recent accomplishment (data related or otherwise!)

- Argit (Participant): NHLBI fellow, learing Computation through online classes
- Steve Tsang, NIAID/Office of Data Science and Emerging Technologies (ODSET) - completed Architecting on AWS training last week.
- Esther Asaki (Helper) NIH/Center for Information Technology) -Taking Google Cloud classes
- Abhijit (Helper): NIH, Carpentries certified instructor, updated a paper revision in 5 minutes using literate programming
- Grace: NCI Division of Cancer Treatment and Diagnosis, helped finalize accurate real-time clinical trial accrual reports for internal tracking.
- Kyle Webb: NIAID/ODSET, data science fellow, conducted NLP analysis on HPO assignment for clinical records

- Chris G, NHGRI Social and Behavioral Research Branch, recently tweeted links about sexism in science without getting trolled.
- Komi Messan: NIAID/Office of Data Science and Emerging Technologies, data science fellow - finished a manuscript revision.
- Ben Krajacich - NIAID/NIH Postdoctoral Fellow - managed to get the 'reticulate' package functioning in R/RStudio
- Leo Meister - NIAID/ODSET, data science fellow, trained computer vision and neural network models on tuberculosis chest x-rays
- Seher Anjum - NIAID staff clinician
- Jeff Robinson - NCI postdoctoral fellow. I designed and fabricated a simple circuit board for motor control for an AI-driven robot.
- Eric Cox - NCBI Refseq Scientist. signed up for this class!
- Gladys Wang - NIH/NINDS, Data science fellow.
- Noffisat Oki - NIAID/Office of Data Science and Emerging Technologies, AAAS Science and Technology Policy Fellow
- Palchamy Elango - NIA, Baltimore - Data Management Coordinator
- ying zhang_NIAID
- Hoai Nguyen: NIAID, Laboratory of Bacteriology, Bld 50, Rm 6250
- Qing Chen :NICHD

Pre-Workshop Survey!

https://carpentries.typeform.com/to/wi32rS?slug=2020-01-13-nih

Project organization and management

Workshop Lesson Materials (follow along): https://datacarpentry.org/genomics-workshop/

Workshop files

Ecoli_metadata_composite_messy.xlsx: https://github.com/datacarpentry/organization-genomics/raw/gh-pages/files/Ecoli_metadata_composite_messy.xlsx

sample_submission.txt: https://datacarpentry.org/organization-genomics/files/sample_submission.txt

sequencing_results_metadata.txt:
https://datacarpentry.org/organization-genomics/files/sequencing_results_metadata.txt
sequencing_results_metadata.xls:
https://datacarpentry.org/organization-genomics/files/sequencing_results_metadata.xls

European Nucleotide Archive: https://www.ebi.ac.uk/ena

Dropbox link to sample data: https://www.dropbox.com/s/cehpb8fteyvj014/dc_sample_data.tar.gz?dl=1

Paper: https://www.nature.com/articles/nature18959

Discussion questions

What kinds of data and information have you generated before you sent your DNA/RNA off for sequencing?

- data about the specimens used for the sequencing
- concentration of DNA/information about DNA extracts
- age, race, gender; demographic data about samples
- clinical relevant data

This is some potential spreadsheet data generated about a sequencing experiment. With the person next to you, for about 2 minutes, discuss some of the problems with the spreadsheet data shown above. You can look at the image, or download the file to your computer via this link and open it in a spreadsheet reader like Excel.

- Don't load multiple tables of data in a single spreadsheet tab
- Either combine sheets or make two separate files (.csv)
- Combine all three tables into one, with citations as a column
- Standardize column names and labels
- Avoid using special character, e.g. "+", ensure data consistency, plus instead of +
- Pull other metadata into a new column, i.e. paired/single end reads/read length
- Don't use colors to show different data -- not as machine readable
- Put "NA" in blank fields
- Before journal submission, read the guide to authors to see what they require
- See this article about how Excel can change a gene name to a date, e.g. SEP7, OCT10
- "Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics"
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC459209/

1. What challenges do you think you'll face (or have already faced) in working with a large sequence dataset?

- DISK SPACE
- deal with bugs
- sequence QC

2. What is your strategy for saving and sharing your sequence files?
3. How can you be sure that your raw data have not been unintentionally corrupted?
4. Where/how will you (did you) analyze your data - what software, what computer(s)?

sequencing_results_metadata
1. How are these samples organized?

- by sample ID

2. If you wanted to relate file names to the sample names submitted above (e.g. wild type…) could you do so?
yes by comparing the "tube_barcode" column in metadata to the sample_id column from sequencing

results
3. What do the _R1/_R2 extensions mean in the file names?
different runs? forward/reverse paired-end reads?
4. What does the '.gz' extension on the filenames indicate?
compressed file - "gzip" file
5. What is the total file size? what challenges in downloading and sharing these data might exist?
1113GB, bandwidth limitations, storage amounts. Actual file size is larger since these are compressed files

At NIH: Biowulf is the way to go. https://hpc.nih.gov/. There are regular training for Biowulf through the HPC folks (https://hpc.nih.gov/training/#upcoming)
For trainees, getting an account goes through your PI, since there is a nominal annual fee involved.
Also https://hpc.nih.gov/dashboard will tell you about your account and allows you to request more space

Link to publication -
https://www.nature.com/articles/nature18959
SRA - https://www.ncbi.nlm.nih.gov/bioproject/295606
https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=295606
Run info table - https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP064605&o=acc_s%3Aa
NCBI Documentation on downloading metadata - https://www.ncbi.nlm.nih.gov/books/NBK242621/
Linux command to download the RunInfo Table -
```
wget -O ./SRP064605_info.csv 'http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?save=efetch&db=sra&rettype=runinfo&term= SRP064605'
```

1. What strain of E. coli was used in this experiment?
REL606
2. What was the sequencing platform used for this experiment?
Illumina HiSeq 2xxx
3. What samples in the experiment contain paired end sequencing data?
SAMN04096045, SAMN04096046, SAMN04096055, SAMN04096145, SAMN04096056, SAMN04096057, SAMN04096166, SAMN04096058, SAMN04096186, SAMN04096187
4. What other kind of data is available?
Number of bases,
5. Why are you collecting this kind of information about your sequencing runs?
Reproducibility

SRR2589044

Additional resources

Setting up lab notebooks: https://www.hhmi.org/sites/default/files/Educational%20Materials/Lab%20Management/Making%20the%20Right%20Moves/moves2_ch8.pdf

Genomics metadata standards: https://press3.mcs.anl.gov/gensc/projects/

Hadley Wickham - Tidy Data: https://www.jstatsoft.org/article/view/v059i10

SRA Documentation: http://www.ncbi.nlm.nih.gov/Traces/sra/

SRA Toolkit: http://www.ncbi.nlm.nih.gov/Traces/sra/?view=toolkit_doc

Lunch from 11:40 to 12:45!!

Introduction to the command line

PuTTY (Windows)

Download: https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html
Host Name: your AMI address
Port: 22
Connection type: SSH

If you don't have admin rights on Windows, you can download from https://portableapps.com/apps/internet/putty_portable. This installs without the admin requirement, just make sure it is installing in a folder you have write priviledges on

WinSCP portable: https://winscp.net/download/WinSCP-5.15.9-Portable.zip

Filezilla (if you do have Admin rights): https://filezilla-project.org/download.php?show_all=1

- (works on Mac, Windows, and Linux)

Terminal (Mac/Linux)

ssh dcuser@your-amazon-instance

Amazon Machine Images

Please put your name next to one of the instances listed below. This will be your access point for the duration of the workshop.

ec2-52-90-167-208.compute-1.amazonaws.com - Amanda Devine
ec2-3-86-68-25.compute-1.amazonaws.com - Ben Krajacich
ec2-54-205-70-175.compute-1.amazonaws.com - Steve Tsang
ec2-54-166-149-220.compute-1.amazonaws.com - Palchamy Elango
ec2-54-165-22-151.compute-1.amazonaws.com Chris
ec2-18-232-63-151.compute-1.amazonaws.com Jeff R

ec2-3-91-24-164.compute-1.amazonaws.com - Dev
ec2-52-54-186-178.compute-1.amazonaws.com - Kyle Webb
ec2-34-239-49-71.compute-1.amazonaws.com - Noffisat Oki
ec2-18-212-133-7.compute-1.amazonaws.com- Gladys Wang
ec2-54-162-152-166.compute-1.amazonaws.com- Komi Messan
ec2-54-224-25-182.compute-1.amazonaws.com - Grace Mishkin
ec2-34-235-144-86.compute-1.amazonaws.com - Leo Meister
ec2-54-81-251-195.compute-1.amazonaws.com- Hoai Nguyen
ec2-34-229-60-156.compute-1.amazonaws.com - Eric Cox
ec2-3-81-139-59.compute-1.amazonaws.com - argit
ec2-52-205-28-54.compute-1.amazonaws.comYan
ec2-54-81-48-187.compute-1.amazonaws.com- Qing Chen
ec2-34-230-14-54.compute-1.amazonaws.comSeher
ec2-54-172-173-19.compute-1.amazonaws.com - Mike
ec2-54-162-81-1.compute-1.amazonaws.comying zhang
ec2-54-173-123-231.compute-1.amazonaws.com
ec2-3-81-219-68.compute-1.amazonaws.comAbhijit
ec2-3-81-103-120.compute-1.amazonaws.com
ec2-18-234-55-56.compute-1.amazonaws.com
ec2-54-196-98-114.compute-1.amazonaws.com
ec2-54-80-200-125.compute-1.amazonaws.com Esther
ec2-52-87-199-0.compute-1.amazonaws.com
ec2-18-212-131-234.compute-1.amazonaws.com
ec2-18-206-239-122.compute-1.amazonaws.com
ec2-3-84-174-39.compute-1.amazonaws.com
ec2-3-81-184-103.compute-1.amazonaws.com
ec2-3-90-30-72.compute-1.amazonaws.com
ec2-18-212-158-199.compute-1.amazonaws.com
ec2-3-81-215-23.compute-1.amazonaws.com
ec2-54-80-119-58.compute-1.amazonaws.com
ec2-54-225-42-112.compute-1.amazonaws.com
ec2-3-90-155-122.compute-1.amazonaws.com
ec2-35-173-216-14.compute-1.amazonaws.com
ec2-54-152-68-123.compute-1.amazonaws.com

Challenge questions

Use the -l option for the ls command to display more information for each item in the directory. What is one piece of additional information this long format gives you that you don’t see with the bare ls command?

- file permissions
- file/dir size (in bytes)
- dates
- username

First navigate to the shell_data directory. There is a hidden directory within this directory. Explore the

options for ls to find out how to see hidden directories. List the contents of the directory and identify the name of the text file in that directory.
Hint: hidden files and folders in Unix start with ., for example .my_hidden_directory

ls -a or ls -All: .hidden
youfoundit.txt

ls -f

ls -la
cd .hidden
ls

"Here I am"

Navigate to your home directory. From there, list the contents of the untrimmed_fastq directory.

Take a look at the file system diagrammed here:
https://datacarpentry.org/shell-genomics/fig/filesystem-challenge.svg
If pwd displayers /Users/thing, what will ls ../backup display?

original, pnas_final, pnas_sub

Do each of the following tasks from your current directory using a single ls command for each:
- List all of the files in /usr/bin that start with the letter ‘c’.
- List all of the files in /usr/bin that contain the letter ‘a’.
- List all of the files in /usr/bin that end with the letter ‘o’.

Extra credit: List all of the files in /usr/bin that contain the letter ‘j’ or the letter ‘q’. This requires a Unix wildcard that we haven’t talked about yet. Try searching the internet for information about Unix wildcards to find what you need to solve this problem.

ls -l /usr/bin*j* || /usr/bin/*q*
ls /usr/bin/{*j*,*q*}
ls /usr/bin*j* | ls /usr/bin/*q*
ls /usr/bin*j* /usr/bin/*q*

/usr/bin/*[jq]*
also produce same output as for ls /usr/bin/{*j*,*q*}

Find the line number in your history for the command that listed all the .sh files in /usr/bin. Rerun that command.

Start at the beginning of the file, and serach for the first instance of TTTTT. What are the next three nucleotides (characters) following this first instance?
CAC

Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJK
                  | | | | |
Quality score: 0.......10.......20.......30.......40..

Permissions: https://datacarpentry.org/shell-genomics/fig/rwx_figure.svg

Starting in the shell_data/untrimmed_fastq/ directory, do the following:
1. Make sure that you have deleted your backup directory and all files it contains.
2. Create a backup of each of your FASTQ files using cp. (Note: You'll need to do this individually for each of the two FASTQ files. We haven't learned yet how to do this with a wildcard.)
3. Use a wildcard to move all of your backup files to a new backup directory.
4. Change the permissions on all of your backup files to be write-protected. (This can be done with one command!)

1. Search for the sequence GNATNACCACTTCC in the SRR098026.fastq file. Have your search return all matching lines and the name (or identifier) for each sequence that contains a match.
2. Search for the sequence g in both FASTQ files. Have your search return all matching lines and the name (or identifier) for each sequence that contains a match.

How many sequences in SRR098026.fastq contain at least 3 consecutive Ns?

249

Print the file prefix of all of the .txt files in our current directory.

Remove _2020 from all of the .txt files.

We want the script to tell us when it's done.
1. Open bad-reads-script.sh, add the line echo "Script finished!" after the grep command, and save the file.
2. Run the updated script.

ftp://ftp.ensemblgenomes.org/pub/release-37/bacteria/species_EnsemblBacteria.txt

scp dcuser@ip.address:/home/dcuser/shell_data/untrimmed_fastq/scripted_bad_reads.txt ~/Downloads

To move a file called temp from the home directory in AWS to Desktop
sftp dcuser@ip.address:temp C:/Users/tsang/Desktop

Windows Update 1803 (April 2018) added scp, ssh, sftp, tar, curl by default: https://twitter.com/Oddvarmoe/status/996177052599545857

DAY TWO

Participants: What is one thing you learned yesterday that you will use in the future?

- Chris G: How to open a terminal window and remote in to a specific environment
- Ben Krajacich -- 'history' command to export command histories to generate reproducible workflows.
- Jeff R - for loops
- Argit Marishta - How to manipulate files through the terminal
- What is unix and how to use it
- Grace Mishkin - accessing data from NLM
- Leo Meister - Using grep for searching through files
- Komi Messan - files management using the command lines
- Qing Chen:linux command lines to manipulate data
- Noffisat Oki - Windows now has built in ssh
- Eric - how to write a for loop
- Gladys Wang - refreshing the linux commands and searching genomic data
- Amanda: That Windows now includes some of the same commands that were traditionally UNIX-only
- Mike - How to import a CSV file into Excel using the Data tab
- Kyle - bash shell scripting

Lesson Materials: https://datacarpentry.org/wrangling-genomics/

Metadata Table: https://github.com/datacarpentry/wrangling-genomics/blob/gh-pages/files/Ecoli_metadata_composite.csv

Challenge Questions

Based on the metadata, can you answer the following questions?

1. How many different generations exist in the data? 25
2. How many rows and how many columns are in this data? 12 columns, 63 rows (one of which is the header)
3. How many citrate+ mutants have been recorded in **Ara-3**? 10 10
4. How many hypermutable mutants have been recorded in **Ara-3**? 6

What is the last read in the SRR2584863_1.fastq file? How confident are you in this read?
tail -n 4 SRR2584863_1.fastq, read is not too bad

Which samples failed at least one of FastQC's quality tests? What test(s) did those samples fail?

```
cp ~/.miniconda3/pkgs/trimmomatic-0.38-0/share/trimmomatic-0.38-0/adapters/
NexteraPE-PE.fa .

trimmomatic PE SRR2589044_1.fastq.gz SRR2589044_2.fastq.gz \
            SRR2589044_1.trim.fastq.gz SRR2589044_1un.trim.fastq.gz \
            SRR2589044_2.trim.fastq.gz SRR2589044_2un.trim.fastq.gz \
            SLIDINGWINDOW:4:20 MINLEN:25 ILLUMINACLIP:NexteraPE-PE.fa:2:40:15
```

Use the output from your Trimmomatic command to answer the following questions.
1) What percent of reads did we discard from our sample? Dropped: 2548 (0.23%) so 0.23%.

2) What percent of reads did we keep both pairs?
79.96%

```
for infile in *_1.fastq.gz
do
  base=$(basename ${infile} _1.fastq.gz)
  trimmomatic PE ${base}_1.fastq.gz ${base}_2.fastq.gz \
              ${base}_1.trim.fastq.gz ${base}_1un.trim.fastq.gz \
              ${base}_2.trim.fastq.gz ${base}_2un.trim.fastq.gz \
              SLIDINGWINDOW:4:20 MINLEN:25 ILLUMINACLIP:NexteraPE-
PE.fa:2:40:15
done
```

https://github.com/miketrizna/2020-01-13-nih

Lunch 12pm - 1pm!

Post-workshop survey:

https://carpentries.typeform.com/to/UgVdRQ?slug=2020-01-13-nih

Changing your command line prompt with PS1: https://ss64.com/bash/syntax-prompt.html
```
PS1='\w $ '
```

Copying FastQC HTML of trimmed fastq to local desktop:
```
scp
dcuser@ec2-54-172-173-19.compute-1.amazonaws.com:/home/dcuser/dc_workshop
/data/trimmed_fastq/SRR2584863_2.trim_fastqc.html ~/Desktop/
```

Downloading reference E. coli genome:
```
curl -L -o data/ref_genome/ecoli_rel606.fasta.gz
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/017/985/GCA_000017985.1_ASM179
8v1/GCA_000017985.1_ASM1798v1_genomic.fna.gz
```

Downloading small subset of FASTQ files for aligning to ref genome and variant calling:
```
curl -L -o sub.tar.gz https://ndownloader.figshare.com/files/14418248
```

When using rm, a good practice can be to use the -i flag (like rm -i). This will prompt you to confirm whether you would like to remove each file that you are attempting to rm.

Use the grep and wc commands you've learned to assess how many variants are in the vcf file.

- 767

Downloading final all-together script:
curl -O https://raw.githubusercontent.com/datacarpentry/wrangling-genomics/gh-pages/files/run_variant_calling.sh

Answering question about how the various bioinformatics tools we used were installed:
https://bioconda.github.io/

This is the R lesson we breezed through in the final hour of the workshop:
https://datacarpentry.org/genomics-r-intro/05-data-visualization/index.html

Use what you just learned to create a scatter plot of mapping quality (MQ) over position (POS) with the samples showing in different colors. Make sure to give your plot relevant axis labels.

```
ggplot(data=variants, aes(x=POS, y=MQ, color=sample_id)) +
geom_jitter(alpha=0.5) + labs(x="Base Pair Position", y="Quality")
```