- 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

---

USEFUL LINKS

- Link to workshop (25-26 June): https://us02web.zoom.us/j/82589653111?pwd=RlJRUFhteVlvb0FsUVZlRWdITEJOQT09
  - Password: sent in email on 22 June
- Workshop webpage: https://edcarp.github.io/2020-06-25-edinburgh-dc-genomics/
- Doc with useful info: https://docs.google.com/document/d/1XWINNFh75Nfxuv3lT3XtjIfO9oFDpqxvPfj2RpvBJjQ
- PuTTY: https://www.putty.org
- PSCP: http://the.earth.li/~sgtatham/putty/latest/x86/pscp.exe
- shellshare: https://shellshare.net
- https://zoom.us/download#client_4meeting

LESSON 1: Project Organization and Management for Genomics (Edward Wallace)

Edward's opening presentation:
https://drive.google.com/file/d/1IWioQBDCa3HL5D_8nYfERjT5aXdloto-/view?usp=sharing
https://datacarpentry.org/organization-genomics/

Discussion 1: Your experiment

- **What's your name?** What kind of sequencing data are you collecting? What question is your experiment answering? What kinds of data and information have you generated before you sent your DNA/RNA off for sequencing?

Example

- **Edward:** CLIP-seq/CRAC. Asking which RNA is bound by my yeast protein (Ssd1) and how that changes as cells grow. Have data on cell growth (timepoints and optical density) in text format...

- Justin ChIP seq and RNA seq data for epigenetic changes etc-metadata was provided at submission to Ed genomics. Other data on various mutants as Q PCR analyses, photos, microscopy images etc
- Samir Watson - Nanopore Direct RNA sequencing, RNA seq, CLIP effect of RNA modifiactions on splicing Associated data includes Wester blotting, qPCR I just use cell lines so not that much more metadata.
-Jemma Dunn I have not started collecting genomic sequencing data yet but I would like to move into transctriptomics within the tumours I am researching. My background is proteomic screening so I wanted to get an overview of how to go about working with genomic data if I perfrom any sequencing experiments. We collect all clinical background data on all patient speciments inc mutational status, recurrence, treatments to enable stratification and ideally look at transcriptional changes between stratified groups. We then perform validation inc Western blotting, IHC, qPCR and knockdown.
-Caroline Orr - mulitple projects ongoing all using 16S rRNA sequencing. Always collect large amounts of environmental meta data in form of excel sheets and often also try to link qPCR data. A specific example at the moment would be taking soils which show unusual levels of preservation at archaeological sites and trying to identify unusual environmental and microbial signatures. Metadata in this example would be precise information about location of samples, depth, sampling dates, soil moisture, pH, elemental composition etc. Some can be collected retrospecitvely but we always keep additional samples to go back in case we miss something.
-Alex Blain - WES/WGS/RNA seq. Identifying somatic cancer variants in lymphoma samples and trying to use RNA to identify potential changes in the tumour microenvironment between cases, especially those who relapse vs those that don't relapse. Metadata includes as much clinical data as possible, e.g. age at diagnosis, relapse status, translocations present idenified by FISH, stage at diagnosis
-Liz Hughes - RNA seq to look at the transcriptome changes in Cryptococcus grown under different environmental conditiond and relate these to the cell cycle and virulence factors/transcription networks. Not yet started. I will have bioreps and techreps with related microscopy images and the different sample time points. So I will need to collect meta data on strain, environmental conditions, related microscopy, multiplexing.
-Dom Gilchrist - DNA looking at identifying possible transporter gene families. Possibly confocal images, Raman spectra. Just starting really.
-Rochelle Hockney- 16S rRNA sequencing, mainly of low biomass female reproductive tract, placenta and fetal membranes, using extracted DNA to understand the microbiota of these locations liked to negative conditions/infections. Metadata includes patient characteristic data, clinical information, sample processing data, pre-processing data, DNA quality and quantity. Also, qPCR data, potenial for Raman spectroscopy and microscopy.
-
-Alex Newman - whole exome sequencing, whole genome sequencing, RNA-seq - Looking for somatic mutations and events in lymphoma associated with higher risk of relapse and other clinical factors.
-Lucy Butler - I will be analysing DNA genetic engineering for production of natural products
-Amy Burgess- I will be monitoring viral evolution through sequencing of enviromental samples. I
- Maria O'Hanlon - I want to analyse RNAseq to look at transcriptome changes between wildtype and Parkinson's mutant Drosophila melanogaster. I'm yet to start looking at this but I think my first step will be to look at publically available data on fly brains.
-Matt Edmundson- comparing differenes between several genomes. I have a spreadsheet of all the quality control parameters for all the samples I've sent for sequencing (DNA concentration, A260/280, etc).
-Cassy Ross - 16s rRNA sequence gene. Looking at bacterial DNA within the gut of infants. Metadata includes patient information, DNA quality and pre-processing data.
-Seán Dunphy - I have not collected any data yet but ChIP-seq, Hi-C and CLIP-seq data are becoming common in the lab.
-Peixun Zhou: WES and RNA-seq, analysing somatic variants and differentially expressed genes in

lymphoma patient biopsies to identify potential therapeutic targets and biomarkers to stratify high-rish patient

- Natalie Hajduga: Whole exome sequenced data (provided by NHC generated by illumina novaseq) and patience metadata (such as age/sex/prognosis provided by CCLG), producing a R/python program to highlight mutations or similarities between adult and paediatric samples that may be used as targets for novel drugs in B cell lymphoma treatments.
-Sylvie Vagena- collecting data from climbing assays (not having them yet) and rna sequencing to look about any transcriptomic changes in drosophila regaring parkinson's disease. at the moment i would like to have a look on public data
-Naiming Chen - I have not collected data yet, but I am interested in ChIP-seq and Hi-C data which I would need in the future.
-Behnaz Nowrouzi-I'm going to analyse genomic dataset
-Sunil: Hi-C data to analyse genome organization in yeast.
-Mapping genes underlying variation between species by Pool-seq of hybrids.
-Ning ZHAO - I have been done Small RNA sequence before, and want to know how to collect that and analyse in a proper way.
-Koray Malci - we plan to do WGS on yeast genome. We have some engineered strains including many integrations and some deletions so we plan to find all the alterations over the yeast genome.
-Denise: RNAseq data to determine which enzymes could be present in cancer metabolic networks to facilitate reactions

Discussion 2: Problems with spreadsheet data.

Link to the spreadsheet data (image):
https://datacarpentry.org/organization-genomics/fig/01_tidiness_datasheet_example_messy.png

Link to the spreadsheet data :
https://github.com/datacarpentry/organization-genomics/raw/gh-pages/files/Ecoli_metadata_composite_messy.xlsx

Comments?
symbols and text used within a column - uniformity

Mutiple tables in one tab
Comments and data in one cell

Inconsistencies with upper case and lower case letters within the same column.
Multiple columns with the same variable, with values randomly filled in one column or the other.

Discussion 3: Working with sequence data

- What challenges do you think you'll face (or have already faced) in working with a large sequence dataset?
- What is your strategy for saving and sharing your sequence files?
- How can you be sure that your raw data have not been unintentionally corrupted?
- Where/how will you (did you) analyze your data - what software, what computer(s)?

Comments from breakout discussions:

-Breakout room 9
Q.What challenges do you think you'll face (or have already faced) in working with a large sequence dataset?

- finding storage space
- Processing require large computing power , using a compute cluster

Q.How can you be sure that your raw data have not been unintentionally corrupted?

- md5checksum. On some systems this may be called just md5sum.

Q. Where/how will you (did you) analyze your data - what software, what computer(s)?

- linux , comand line, R bioconductor, eddie computer

-Breakout Room 3: might be given other people's data, so might not be familiar with what problems in the data might look like, or might not be able to tell what corrupted data looks like? How can we tell?

- use shared google drive sometimes to save and share raw sequences between the team
- Q: is there a good place to store sequence data you've produced / save it & upload somewhere?
- We used to send DNA to another university, they do the sequencing and send an OTU (operational taxonomic unit) table. Now we have to do that "magic bit in the middle" ourselves. Strategy: make friends with people who know how, and ask them questions.
- Not knowing where the case of the problem is?

-BR3: Look at the structure of your data and see if it is as expected...i.e. correct number of colums and rows
Breakout Room 7 - Actually having the computers that can run the amount of data you have without the computers crashing. Difficult to know how to store your data well whether on external hardrives or cloud etc and how to keep a good log of it for others.
Breakout Room 2: have a central database that is accessible from the browser, avoids duplication. Try to avoid storing EVERYTHING (too big), store enough to make it reproducible.

Discussion 4: Sample submission sheet

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

Questions:

1. What are some errors you can spot in the data? Typos, missing data, inconsistencies?
2. What improvements could be made to the choices in naming?
3. What are some errors in the spreadsheet that would be difficult to spot? Is there any way you can test this?

Answers:
-theres a mix of lower and upper case under client sample ID a A

-the data are not aligned well so they are hard to read also different formating of dates
-are a and A the same thing?
-Client sample Id inconsistency

Date formats are differnt in separate columns
Wierd characters in column names
-wildtype vs wild type
-There are inconsistencies with the dates - some contain only numbers, others abbreviate the month.
- Decimal points aren't consistent.
-Dates are inconsistent
-
-Should it be well "A01" etc?
Dates are not the same format
630ul looks wrong possibly a typo, similarly 0.5 probably wrong
spaces not underscores
prep_date and ship_date are under volume and concentration columns (this is because it is a tab-separated document; it would still be read correctly by some spreadsheet and command line applications (as long as the correct delimiter was used—tab, not whitespace))

Discussion 5: Getting data back from the facility

- Sequencing results - text
  https://datacarpentry.org/organization-genomics/files/sequencing_results_metadata.txt
- Sequencing results - Excel
  https://datacarpentry.org/organization-genomics/files/sequencing_results_metadata.xls

Questions:

1. How are these samples organized?
2. If you wanted to relate file names to the sample names submitted above (e.g. wild type…) could you do so?
3. What do the _R1/_R2 extensions mean in the file names?
4. What does the '.gz' extension on the filenames indicate?
5. What is the total file size - what challenges in downloading and sharing these data might exist?

Comments:
-3. reads in both directions?
5. 1113.6 GB?
-3: ends of paired-end?
-4 zipped

-4. gzipped meaning it's a compressed file
-2: Organized by sample id
-
--5: As the file is over 1TB, download may be taxing and take time as would sharing thsi large data file. Would take hours.

Discussion 6: Review the SraRunTable

Dataset link: https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP064605

Discuss:

1. What strain of *E. coli* was used in this experiment?
2. What was the sequencing platform used for this experiment?
3. What samples in the experiment contain paired end sequencing data?
4. What other kind of data is available?
5. Why are you collecting this kind of information about your sequencing runs?

-1. Escherichia coli B str. REL606 (Organism tab) What is the strain tab?
Escherichia coli B str. REL606

2-ILLUMINA
2 - Illumina platform
2. Illumina HiSeq 2000/2500 or Genome Analyzer IIx

-3. entries with PAIRED under LibraryLayout tab
-

Feedback

Write one thing that you liked and one thing that you think could be improved.

Thing I liked

- Really nice style and pace thank you :)
- Everything was very clearly explained, thank you for your time :)
- Very well explained at a good pace so we could easily follow.
- well paced and informative, thankyou
- Nice use of hypotheticals to simplify tools
- Very informative and the right pace.

Thing that could be improved

- A little bit more hands on would be good to practice what we're learning. I understand time contraints though.

LESSON 2: Introduction to the Command Line for Genomics (Flic Anderson)

Information for the lesson here:
https://datacarpentry.org/shell-genomics/

Good guide with pictures for how to login to the instances:

https://datacarpentry.org/cloud-genomics/02-logging-onto-cloud/#logging-onto-a-cloud-instance

IP address: below your name in the list of participants at the bottom of this document. NOTE: everyone's login IP address is different & yours won't have the same numbers as the instructor's.

The username is the same for all of you: dcuser
Password : data4Carp (it is key sensitive!!!!!)

ssh dcuser@your_IP_address

Reference document for commands we have been using, et cetera:

https://git.ecdf.ed.ac.uk/s1770238/computing-tutorials/blob/master/EdCarp_Linux_Reference.pdf

Questions:
Do you have any questions about the introduction to the command line material? Please add them here & I'll keep an eye on it during the session & throughout the workshop and follow up :)

Commands:

clear: clears the screen. It doesn't erase anything
pwd: It prints "where you are"
ls: you can see "what things" are in your current location - **look-see** command according to Flic's friend :)
cd "name the directory that you want to visit": allows you to navigate in your file system

ls -F: Using a flag you can "tweak" the output of the command. -F helps you to distinguish the folders from the files!

man comand opens the manual of each command, allows you to see the list of the flags. There is a man command for everything! Even for man itself!
If you tyme "man man" you will see it! :)

In case that your system doesn't support "man" command, you can give the "--help" after each command.
e.g. ls --help

Use the arrows in your keyboard to navigate the previous commands that you have used. You can avoid the typing!!!!

cat: The concatenation command can be used for printing the content of a file on the screen. (It doesn't do only this! :))
syntax: cat filename.txt

less: less command can be used to open a file and interact with the content. It is more effective than the cat command because it doesn't load the whole content on the screen.

Syntax: less filename.txt

quit less: press q

"head" and "tail" are very important commands, especially if you want to check if big sequence files are corrupted. (you can use the tail to see if the last lines are keeping the standards of the file format).

cp "source" "destination"
copy command take two inputs. The file that you want to copy and the place that you want to copy.

mkdir: for creating a new directory

mv: for moving (or renaming) files. The syntax is the same with the "cp" command

UNIX commands cheat sheet:
https://files.fosswire.com/2007/08/fwunixref.pdf

~~~

Regarding the file system:
An absoulte path start always from the root "/" which includes everything.
The relative one starts from your current directory.

Think the file system as an upside down tree, where the top directory is the root, untill the final leaves that are the files.

Exercise 2.1: Relative Path Resolution:

Using the filesystem diagram (you can see it at the link below), if pwd displays /Users/thing, what will ls ../backup display?

1. ../backup: No such file or directory
2. 2012-12-01 2013-01-08 2013-01-27
3. 2012-12-01/ 2013-01-08/ 2013-01-27/
4. original pnas_final pnas_sub

- https://datacarpentry.org/shell-genomics/fig/filesystem-challenge.svg

Please place a star next to the answer you're voting for '*' :
Answer 1: **
Answer 2:
Answer 3: **
Answer 4: **********

Exercise 3.1: Wildcards:

Do each of the following tasks from your current directory using a single ls command for each:

1. List all of the files in /usr/bin that start with the letter ‘c’.
2. List all of the files in /usr/bin that contain the letter ‘a’.
~~~

3. List all of the files in /usr/bin that end with the letter ‘o’.
4. Bonus: List all of the files in /usr/bin that contain the letter ‘a’ or the letter ‘c’.

Hint: The bonus question 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 the bonus problem.
1. ls /usr/bin/c*
2. ls /usr/bin *a*
3. ls /usr/bin/ *o
4. ls /usr/bin/ | grep "a/c" or ls /usr/bin/*[ac]* (on linux) also ls /usr/bin/{*a*,*c*}

Exercise 3.2: Examining Files:

1. Print out the contents of the ~/shell_data/untrimmed_fastq/SRR097977
2. .fastq file. What is the last line of the file?

C:CCC::CCCCCCCC<8?6A:C28C<608'&&&,'$
+SRR097977.249 209DTAAXX_Lenski2_1_7:8:3:441:292 length=36
C:CCC::CCCCCCCC<8?6A:C28C<608'&&&,'$
c
+SRR097977.249 209DTAAXX_Lenski2_1_7:8:3:441:292 length=36
C:CCC::CCCCCCCC<8?6A:C28C<608'&&&,'$

1. From your home directory, and without changing directories, use one short command to print the contents of all of the files in the ~/shell_data/untrimmed_fastq directory.
2. cat ~/shell_data/untrimmed_fastq/*

Exercise 3.3: File Manipulation

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.

Exercise 4.1: Searching Files

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 AAGTT in both FASTQ files. Have your search return all matching lines and the name (or identifier) for each sequence that contains a match.

Exercise 4.2: Redirecting Output:

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

Exercise 4.3: Using Basename in For Loops:

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

Exercise 4.4: Using Basename in For Loops:

Remove .fastq from all of the .fastq files (keep the file names).

```
for filename in *.fastq
> do
> name=$(basename ${filename} .fastq)
>echo $name
> done
```

Exercise 5.1: Writing Scripts:
We want the script to tell us when it's done.

1. Open bad-reads-script.sh and add the line echo "Script finished!" after the grep command and save the file.
2. Run the updated script.

Exercise 6.1: Getting Your Project Started

Use the mkdir command to make the following directories:

- dc_workshop
- dc_workshop/docs
- dc_workshop/data
- dc_workshop/results

Exercise 6.2: Documenting Your Activity on the Project:

Using your knowledge of the shell, use the append redirect >> to create a file called dc_workshop_log_XXXX_XX_XX.sh (Use the four-digit year, two-digit month, and two digit day, e.g. dc_workshop_log_2020_06_25.sh)

Feedback

Write one thing that you liked and one thing that you think could be improved.

Thing I liked

- At times could be quite complicated but by working through it on puTTY this made it easier to understand the concepts. Great presentation, thanks.
- Great presentation and fairly easy to follow for a complete beginner! And great to also have the website to follow at the same time incase you miss anything.

Really clear, perfect pace.
Everything nice.. well explaned.. many thanks..
Well explained and I enjoyed very much
The explanations were well done.
Really clear, great presentation. Learnt a lot.
Clearly explained and nice presentations. I've learnt a lot, thanks!
Nice tempo for beginners and experienced users. Very logically built.

Thing that could be improved

- No dissing botanists!
- I didnt follow the loop section so well but having the website means you can read through so it's not so bad! Thanks!
- Having a list of commands on a separate sheet would be quite useful as I've never used a command promt before and got a bit lost.
- Perhaps just a little bit more time.

LESSON 3: Data Wrangling and Processing (Vasilis Lenis)

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

---

1. Assessing Read Quality

The sequencing data of 3 samples (Raw data):

curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/004/SRR2589044/SRR2589044_1.fastq.gz
curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/004/SRR2589044/SRR2589044_2.fastq.gz
curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/003/SRR2584863/SRR2584863_1.fastq.gz
curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/003/SRR2584863/SRR2584863_2.fastq.gz
curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/006/SRR2584866/SRR2584866_1.fastq.gz
curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/006/SRR2584866/SRR2584866_2.fastq.gz

Commands:

mkdir -p ~/dc_workshop/data/untrimmed_fastq/
cd ~/dc_workshop/data/untrimmed_fastq

gunzip SRR2584863_1.fastq.gz

fastqc *.fastq*

mv *.html ~/dc_workshop/results/fastqc_untrimmed_reads/
mv *.zip ~/dc_workshop/results/fastqc_untrimmed_reads/

Process for running FastQC on the .fastq.gz files:

fastqc *.fastq*

mkdir -p ~/dc_workshop/results/fastqc_untrimmed_reads
ls
(NOTE: should see .html files, .zip files and .fastq.gz files)

mv *zip ~/dc_workshop/results/fastqc_untrimmed_reads/
mv *.html ~/dc_workshop/results/fastqc_untrimmed_reads/
ls
(NOTE: should see only .fastq.gz files left now in the ~/dc_workshop/data/untrimmed_reads folder)

cd ~/dc_workshop/results/fastqc_untrimmed_reads/
ls
(NOTE: after moving to this newly created folder, should see the .zip and .html files here in this folder)

Exercise 1:
What is the last read in the SRR2584863_1.fastq file? How confident are you in this read?
it looks good qual
Better quality at the end of the sequence than the previous but still variable.
It looks good, better than the last one at the end

@SRR2584863.1553259 HWI-ST957:245:H73R4ADXX:2:2216:21048:100894/1
CTGCAATACCACGCTGATCTTTCACATGATGTAAGAAAAGTGGGATCAGCAAACCGGGTGCTGCTGTGGCTAGTTGCAGCAAACCATGCAGTGAACCCGCCTGTGCTTCGCTATAGCCGTGACTGATGAGGATCGCCGGAAGCCAGCCAA
+
CCCFFFFFHHHHGJJJJJJJJHGIJJJIJJJJIJJJJIIIIJJJJJJJJJJJJIIJJJHHHHHFFFFFEEEEEDDDDDDDDDDDDDDDDDCDEDDBDBDDBDDDDDDDDDBDEEDDDD7@BDDDDDD>AA>?B?<@BDD@BDC?BDA?

Exercise 2:

How many reads are in the SRR2584863_1.fastq file?
6213036/4
wc -l
grep '^@' SRR2584863_1.fastq | wc -l, but the quality line could also start with a @...
grep SRR2584863 SRR2584863_1.fastq | wc -l
Fastqc Manual:
https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

Breakroom:
Discuss your resultsin the breakroom. Which sample(s) looks the best in terms of per base sequence quality? Which sample(s) look the worst?

Exercise 3:
Which samples failed at least one of FastQC's quality tests? What test(s) did those samples fail?
SRR2584863_2.fastq has 2 fails
 SRR2589044_2.fastq also on seq qual
  SRR2589044_2.fastq adapter qual
FAIL Per base sequence quality SRR2584863_2.fastq
FAIL Per tile sequence quality SRR2584863_2.fastq
FAIL Per base sequence content SRR2584863_2.fastq
FAIL Per base sequence quality SRR2584866_1.fastq
FAIL Per base sequence content SRR2584866_1.fastq
FAIL Adapter Content SRR2584866_1.fastq
FAIL Adapter Content SRR2584866_2.fastq
FAIL Adapter Content SRR2589044_1.fastq
FAIL Per base sequence quality SRR2589044_2.fastq
FAIL Per tile sequence quality SRR2589044_2.fastq
FAIL Per base sequence content SRR2589044_2.fastq

2. Trimming and Filtering

Trimmomatic Manual:
http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf

Check that tool is installed:

$ trimmomatic

 To check version:
$ trimmomatic -version

Change directory to where the data is
$ cd ~/dc_workshop/data/untrimmed_fastq/

To copy the adapter file (note the space between the file name and the . noting the current directory):

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

To run trimmomatic:
$ trimmomatic PE SRR2584863_1.fastq SRR2584863_2.fastq.gz SRR2584863_1.trim.fastq SRR2584863_1.un.trim.fastq SRR2584863_2.trim.fastq SRR2584863_2.un.trim.fastq SLIDINGWINDOW:4:20 MINLEN:25 ILLUMINACLIP:NexteraPE-PE.fa:2:40:15

Just make sure you've got the .fastq and .fastq.gz file extensions specified - I tried unzipping all the .fastq.gz files, so for me I need to make sure I edited the command when I pasted it in

Check that trimmed files are in the directory
$ ls -lh
You should see the output files

To gzip the unzipped file for consistency:
$ gzip SRR2584863_1.fastq
Check that the six fastq files are zipped:
$ ls *.fastq.gz
To check file names for iteration over read 1 files
$ ls *_1.fastq.gz

This is the loop to run trimmomatic on all our .fastq.gz sample files

for infile in *_1.fastq.gz; dobase=$(basename ${infile} _1.fastq.gz); echo "$base"; trimmomatic PE ${infile} ${base}_2.fastq.gz ${base}_1.trim.fastq.gz ${base}_1.un.trim.fastq.gz ${base}_2.trim.fastq.gz ${base}_2.un.trim.fastq.gz SLIDINGWINDOW:4:20 MINLEN:25 ILLUMINACLIP:NexteraPE-PE.fa:2:40:15 -threads 16; done

(the same loop code split up, so you can put it in line by line if you want to:)
for infile in *_1.fastq.gz
do
base=$(basename ${infile} _1.fastq.gz)
echo "$base"; trimmomatic PE ${infile} ${base}_2.fastq.gz ${base}_1.trim.fastq.gz ${base}_1.un.trim.fastq.gz ${base}_2.trim.fastq.gz ${base}_2.un.trim.fastq.gz SLIDINGWINDOW:4:20 MINLEN:25 ILLUMINACLIP:NexteraPE-PE.fa:2:40:15 -threads 16
done

NOTE: it'll take a little while (only a few minutes) to run, so bear with it until you get your prompt back :)

Move trimmed reads to new driectory called trimmed_fastq:

mv *.trim* ../trimmed_fastq/

Exercise 1:
Use the Trimmomatic output:

- 1) What percent of reads did we discard from our sample?
- 2) What percent of reads did we keep both pairs?

Exercise 2:
Can you check the quality of the trimmed reads?

3. Variant Calling Workflow

Exercise 1:
We saved this file as data/ref_genome/ecoli_rel606.fasta.gz and then decompressed it. What is the real name of the genome?

Exercise 2:
how many variants are in the vcf file?

Exercise 3:
What variant is present at position 4377265? What is the canonical nucleotide in that position?

Reference 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_ASM1798v1/GCA_000017985.1_ASM1798v1_genomic.fna.gz

Trimmed data (small):
curl -L -o sub.tar.gz https://ndownloader.figshare.com/files/14418248

$ tar xvf sub.tar.gz
$ mv sub/ ~/dc_workshop/data/trimmed_fastq_small

Mapping commads:

****** cd = ~/dc_workshop******

1. Indexing

- bwa index data/ref_genome/ecoli_rel606.fasta

2. **Mapping**

bwa mem data/ref_genome/ecoli_rel606.fasta data/trimmed_fastq_small/SRR2584866_1.trim.sub.fastq data/trimmed_fastq_small/SRR2584866_2.trim.sub.fastq > results/sam/SRR2584866.aligned.sam

3. **SAM to BAM**

samtools view -S -b results/sam/SRR2584866.aligned.sam > results/bam/SRR2584866.aligned.bam

ls -lh results/bam

4**. BAM sorting**

samtools sort -o results/bam/SRR2584866.aligned.sorted.bam results/bam/SRR2584866.aligned.bam

5. Alignment stats

samtools flagstat results/bam/SRR2584866.aligned.sorted.bam

6. Coverage calculation

bcftools mpileup -O b -o results/bcf/SRR2584866_raw.bcf -f data/ref_genome/ecoli_rel606.fasta results/bam/SRR2584866.aligned.sorted.bam

7. Variant calling

bcftools call --ploidy 1 -m -v -o results/bcf/SRR2584866_variants.bcf results/bcf/SRR2584866_raw.bcf

8. bcf to vcf format

vcfutils.pl varFilter results/bcf/SRR2584866_variants.bcf > results/vcf/SRR2584866_final_variants.vcf

4. Automating a Variant Calling Workflow

read_qc.sh script content: (can read more about it & see it here: https://datacarpentry.org/wrangling-genomics/05-automation/index.html)

- set -e
- cd ~/dc_workshop/data/untrimmed_fastq
- echo "Running FastQC..."
- fastqc *.fastq*
- mkdir -p ~/dc_workshop/results/fastqc_untrimmed_reads
- echo "Saving FastQC results..."
- mv *.zip ~/dc_workshop/results/fastqc_untrimmed_reads
- mv *.html ~/dc_workshop/results/fastqc_untrimmed_reads
- cd ~/dc_workshop/results/fastqc_untrimmed_reads
- echo "Unzipping..."
- for filename in *.zip

- do
- unzip $filename
- done
- echo "Saving summary..."
- cat */summary.txt > ~/dc_workshop/results/docs/fastqc_summaries.txt

For downloadin the second script:
curl -O https://raw.githubusercontent.com/datacarpentry/wrangling-genomics/gh-pages/files/run_variant_calling.sh

....regarding SAM/BAM format:
https://academic.oup.com/bioinformatics/article/25/16/2078/204688

...regarding VCF format:
https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format

To build an index: (from ~/dc_workshop/data/ref_genome:)
bwa index ecoli-rel606.fasta

To make results output directories:
mkdir -p results/sam results/bam results/bcf results/vcf

Or, from home directory:
$ mkdir -p ~/dc_workshop/results/bam ~/dc_workshop/results/sam ~/dc_workshop/results/bcf

To look up bwa usage:
bwa mem

To run alignment, ensure you are in ~/dc_workshop, then:
bwa mem data/ref_genome/ecoli_rel606.fasta data/trimmed_fastq_small/SRR2584866_1.trim.sub.fastq data/trimmed
_fastq_small/SRR2584866_2.trim.sub.fastq

bwa mem data/ref_genome/ecoli_rel606.fasta data/trimmed_fastq_small/SRR2584866_1.trim.sub.fastq data/trimmed_fastq_small/SRR2584866_2.trim.sub.fastq > results/sam/SRR2584866.aligned.sam

To convert sam file to bam file (from ~/dc_workshop/data/ref_genome:) :
$ samtools view -S -b results/sam/SRR2584866.aligned.sam >
results/bam/SRR2584866.aligned.bam

To sort the bam files;
samtools sort -o results/bam/SRR2584866.aligned.sorted.bam results/bam/SRR2584866.aligned.bam
To get bam stats:
samtools flagstat results/bam/SRR2584866.aligned.sorted.bam
s

Exercise 1:
Can you put some comments in your code?

Exercise 2:
Can you automate the process that it counts the variants of all the vcf files?

Feedback

Write one thing that you liked and one thing that you think could be improved.

Thing I liked

- very useful even through zoom questions were answered
- Fast, but helpers really helped.
- Great to get an overview of how to run and create a script from scratch. Brilliant help even over zoom. Always someone to help when you get lost. Very patient and non judgemental of complete novices!
- Clearly explained, extremely patient with lovely guides and helpers! Comfortable and non-judgemental team/enviroment
- So useful, thank you so much. So patient and very clear :)
- Helpers were very helpful when lost
- The workshop was great, all the intructors were really helpfull. Loved learning how to write scripts.

Thing that could be improved

- maybe give a liltle more time by the instructors before hit enter after the commands
- Warning to stick to given directory names and structure at the start.
- For complete beginner the final script was large and with more time I would probably have grasped more but with the online info can go back and read through.
- The delivery was fast but I kept up by following along on the datacarpentry.org file, perhaps this could be highlighted at the start? I got the feeling other were not doing the same. Following the documentation may help keep people on track and save time.

LESSON 4: Introduction to Cloud Computing for Genomics (Vasilis Lenis)

https://datacarpentry.org/cloud-genomics/

Breakout room:

In small groups or on your own, plot out your next bioinformatics project. With guidance from your instructors and the above references, try to determine not only what types of resources you'll need, but what platform will best suit your project.
Some things to consider:

- How much data do you have?
- What computational steps will it need?
    - What is the *largest* computational step?
    - Can any steps be done in parallel?
- What is your timeframe?
- Who will be doing most of the computational work?
    - What computational skills do they have?
    - Do you need to share the data across many labs?
- How many times will you need to run this pipeline?

Feedback

Write one thing that you liked and one thing that you think could be improved.
Covered a lot, which was great,
helpers were really useful when going under!
Sometimes hard to type and listen

Thing I liked

great delivery, great helpers, and excellent detail thanks very much.

- Really excellent teaching and explanations.

Thing that could be improved

- Some updates to check which directory you are working in could have been helpful.

LIST OF PARTICIPANTS

Giacomo Peru, University of Edinburgh, EPCC, Project Coordinator (organiser)
ec2-3-86-3-57.compute-1.amazonaws.com

Felicity 'Flic' Anderson, University of Edinburgh, School of Biological Sciences (instructor / helper)
ec2-54-165-83-147.compute-1.amazonaws.com

Bailey Harrington, University of Edinburgh, Institute of Genetics and Molecular Medicine (helper)
ec2-54-156-35-19.compute-1.amazonaws.com

Vasilis Lenis, Teesside University, School of Health and Life Science (instructor / helper)

https://shellshare.net/r/WsYUfQupqoyawF9DSf

Edward Wallace, University of Edinburgh, School of Biological Sciences (instructor/helper)
ec2-18-212-138-132.compute-1.amazonaws.com

Graeme Grimes, University of Edinburgh, Institute of Genetics and Molecular Medicine (helper)
ec2-54-162-233-23.compute-1.amazonaws.com

Alessandro Felder, University College London, Research Software Development (helper)
ec2-54-224-92-227.compute-1.amazonaws.com

David Perez-Suarez, University College Lodon, Research Software Development (helper)
ec2-34-227-88-123.compute-1.amazonaws.com

Mario Antonioletti, University of Edinburgh, EPCC (helper)
ec2-54-80-212-229.compute-1.amazonaws.com

Caroline Orr, Teesside University, School of Health and Life Sciences (participant)
ec2-54-225-10-208.compute-1.amazonaws.com

Rochelle Hockney, Teesside University, School of Health and Life Sciences
ec2-18-232-95-198.compute-1.amazonaws.com

Alex Newman, Teesside University, School of Health and Life Sciences
ec2-54-208-123-8.compute-1.amazonaws.com

Natalie Hajduga, Teesside University, School of Health and Life Sciences
ec2-18-215-234-183.compute-1.amazonaws.com

Peixun Zhou, Teesside University, School of Health and Life Sciences
ec2-54-146-44-121.compute-1.amazonaws.com
https://shellshare.net/r/8YduW5qNaOmxj5IPGr

Mirza Mohidullah Baig, Tesside University, School of health and Life Sciences.
ec2-54-221-173-97.compute-1.amazonaws.com

Hema Narayanasamy Ramakrishnan, Teesside University, School of Health and Life Sciences
ec2-35-173-240-25.compute-1.amazonaws.com

Matt Edmundson, University of Edinburgh, School of Biological Sciences
ec2-52-205-167-104.compute-1.amazonaws.com

Andrew Hudson, University of Edinburgh, School of Biological Sciences
ec2-52-200-85-203.compute-1.amazonaws.com

Sylvie Vagena, Teesside University, School of Health and Life Science
ec2-54-210-78-175.compute-1.amazonaws.com

Amy Burgess, Teesside University, School of Health and Life Sciences

ec2-34-239-185-102.compute-1.amazonaws.com

Lucy Butler, Teesside University, School of Health and Life Sciences
ec2-54-242-189-223.compute-1.amazonaws.com

Liz Hughes, University of Edinburgh, School of Biological Sciences
ec2-18-234-177-109.compute-1.amazonaws.com

Alex Blain, University of Newcastle, Institute of Translational and Clinical Research
ec2-3-90-224-150.compute-1.amazonaws.com

Koray Malci, The University of Edinburgh, Institute for Bioengineering
ec2-3-80-176-220.compute-1.amazonaws.com

Sunil Nahata, University of Edinburgh, Institute of Cell Biology
ec2-54-167-208-152.compute-1.amazonaws.com

Yue Fei, University of Edinburgh, School of Biological Sciences
ec2-54-159-79-231.compute-1.amazonaws.com

Seán Dunphy, University of Edinburgh, School of Biological Sciences
ec2-3-81-22-185.compute-1.amazonaws.com

Dominic Gilchrist, Teesside University, School of Health and Life Sciences
ec2-54-161-48-100.compute-1.amazonaws.com

Maria O'Hanlon, Teesside University, School of Health and Life Sciences
ec2-34-227-98-221.compute-1.amazonaws.com

Enatha Mukantwari,University of Rwanda ,College of Medicine and Health Sciences
ec2-34-238-150-46.compute-1.amazonaws.com

Naiming Chen, University of Edinburgh, School of Biological Sciences
ec2-54-211-166-44.compute-1.amazonaws.com

Yixi Chen, University of Edinburgh, School of Biological Sciences
ec2-52-205-137-15.compute-1.amazonaws.com

Anwar Azad, University of Edinburgh, Hugh Robson building
ec2-54-165-132-45.compute-1.amazonaws.com

Cassy Ross Teesside University
ec2-54-174-208-210.compute-1.amazonaws.com

Jemma Dunn, University of Plymouth
ec2-54-235-0-64.compute-1.amazonaws.com

Behnaz Nowrouzi, University of Edinburgh

ec2-3-90-89-63.compute-1.amazonaws.com

Ning ZHAO, University of Edinburgh, School of Biological Sciences
ec2-54-81-158-121.compute-1.amazonaws.com

Basanta Dhungana, Teesside University
ec2-54-196-94-67.compute-1.amazonaws.com

Samir Watson University of Edinburgh / University of Aarhus
ec2-3-92-57-50.compute-1.amazonaws.com

Justin Goodrich, University of Edinburgh Institute of Molecular Plant Science
ec2-34-229-116-236.compute-1.amazonaws.com

Denise Thiel, Imperial College London
ec2-18-234-54-2.compute-1.amazonaws.com

Lora Boteva, Univeristy of Edinburgh, School of Biological Sciences
ec2-35-173-131-6.compute-1.amazonaws.com