lsWelcome 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

---

Post workshop survey: https://survey.mdc-berlin.de/875143?lang=en

Instructor & helper team

- Project organization and management for genomics - Annajiat Alim Rasel (instructor), Somesh Sai (helper till 11am), Alison Meynert (helper 11am-1pm)
- Introduction to the command line - Martino Sorbaro (instructor), Somesh Sai (helper)
- Data wrangling and processing - Alison Meynert (instructor), Somesh Sai & Thomas Gladytz (helpers)
- Cloud computing for genomics - Annajiat Alim Rasel (instructor), Somesh Sai & Thomas Gladytz (helpers)

Workshop wepage

https://carpentriesitalia.github.io/2023-02-21-mdc-berlin-online/

Zoom

https://mdc-berlin.zoom.us/j/97006149883
Meeting ID: 970 0614 9883

AWS instances

https://docs.google.com/spreadsheets/d/1zhHqPY_9WQrECppvRV9EEjgcpPZSq4IJag8MrS1zZ8s/edit#gid=0

Project organization and management for genomics

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

Data tidiness

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

- Concentrations/molarity, sample names, species, nucleic acid type, library fragment size distribution, illumina adapter sequences
- name project, samples ID, genotypes, adapters, single end/paired end
- gRNA sequence, primer sequence,
- Concentration, sample size (kb), purity, primer concentration and sequence
- metadata(mapping data), concentration of the isolated DNA/RNA
- Sequence information of the PCR primers
- gRNA sequence

Exercise: What are some of the problems with this spreadsheet?

- each line contains more than one type of information, also for one column different formats of data
- doesnt indicate precise read length for each sample
- different table formats
- different ways to name one thing (in CIT column), one table below other
- Multiple tables in one sheet make it difficult to load the data into other software (e.g. python)
- No consistent use of titles e.g. strain vs. STRAIN
- spaces in heading

Planning for NGS projects

Exercise: Working with sequence data

Exercise: Sending samples to the facility

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

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?

- client_sample_id not unique
- Naming Volume with capital V
- inconsistencies in replicate col, volume and concentrations units cannot be read
- inconsistent date format
- Difference in date format in prep date and ship date
- Coding system for replicates unclear

Exercise: interpreting sequencing metadata received from the facility

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

- R1 is the fwd and R2 the rev sequencing, gz means it is compressed,
- use sample id or barcode to match with the sample name
- r1-read1 and r2-read2
- organized by sample ID
- Total: 1113,60 GB

Examining data on the NCBI SRA database

Exercise: locate the Run Selector Table

1. Go to https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA294072

2. Once on the BioProject page, scroll down to the table under “This project encompasses the following 14 sub-projects:”.
3. In this table, select subproject *“PRJNA295606 SRA or Trace Escherichia coli B str. REL606 E. coli genome evolution over 50,000 generations (The University of Texas at…)”*.
4. This will take you to a page with the subproject description, and a table “Project Data” that has a link to the 224 SRA files for this subproject.
5. Click on the number “224” next to “SRA Experiments” and it will take you to the SRA page for this subproject.
6. For a more organized table, select “Send results to Run selector”.

Exercise: download metadata from SRA

1. Go to https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP064605
2. Notice on this page there are three sections. “Common Fields” “Select”, and “Found 312 Items”. Within “Found 312 Items”, click on the first Run Number (Column “Run” Row “1”).
3. This will take you to a page that is a run browser. Take a few minutes to examine some of the descriptions on the page.
4. Use the browser’s back button to go back to the ‘previous page’. As shown in the figure below, the

second section of the page (“Select”) has the Total row showing you the current number of “Runs”, “Bytes”, and “Bases” in the dataset to date. On 2020-04-27 there were 312 runs, 109.58 Gb data, and 177.17 Gbases of data.
5. Click on the “Metadata” button to download the data for this lesson. The filename is “SraRunTable.txt” and save it on your computer Desktop.

Exercise: review the SRA Run Table

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?

- Strains: 1. REL762B 2. REL762C... (Illumina HiSeq 2000 both)
- REL606
- Paired end information in column: LibraryLayout

Introduction to the command line for genomics

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

Command cheat sheet

clear - clear the screen
echo - echos input to terminal
PS1='$ ' - set command prompt
pwd - get current directory location
cd - change directory location
ls - list directory contents

- -F - adds trailing / to directories, to executable programs
- -l - shows file permissions, date created, and size
- -a - shows hidden directories

~ - navigation shortcut for your home directory
.. - navigation shortcut for one directory level up from current directory
* - wildcard to match a sequence of any characters of any length
Ctrl + C - cancel current command
Ctrl + L - clear the screen
Ctrl + R - reverse search history
history - numbered list of recent commands
!<number> - repeat command from history
cat - print contents of file to command line
less - scrollable & searchable content

- <space> - go forward

- b - go backward
- g - go to beginning
- G - go to end
- q - quit
- / - search forwards
- ? - search backwards
- / + <enter> - repeat search forwards from current location
- ? + <enter> - repeat search backwards from current location

head - print first 10 lines of file

- -n change number of lines

tail - print last 10 lines of file

- -n change number of lines

cp - copy file
mkdir - make new directory
mv - move or rename file
chmod - change permissions

- + - add permission
- - - remove permission
- r - read permission
- w - write permission
- x - execute permission (files), transit permission (directories)

rm - remove files

- -r - recurse into directories & remove directories

grep - find text in a file

- -c - count number of matches
- -n - include line numbers of matches
- -v - invert search (everything that doesn't match search term)
- -A<N> - output N lines after match
- -B<N> - output N lines before match
- -h - don't include filenames in output when grepping against multiple files

> - write output
>> - append output
| - pipe output to next command
wc - count characters, words, and lines in a file

- -l only return number of lines

basename - returns filename minus specified extension
var=$(command) - assign result of command to variable $var
cut - extracts fields from a delimited text file

- -f specify which fields in numerical order, e.g. 1 or 1,6 or 3-5
- -d specify delimiter character, e.g. "," (default is tab)

sort - sorts a text file alphanumerically

- -n sort numerically
- -k specify which field (key) to sort on

uniq - returns unique lines if a text file is sorted

- -c - returns counts of how many times each line occurs

Nano cheat sheet

ctrl+space or ctrl+-> - move forward one word
alt+space or esc+space or ctrl+<- - move back one word
ctrl+a - move to beginning of current line
ctrl+e - move to end of current line
ctrl+w - search
ctrl+k - delete current line
ctrl+o - write to current file
ctrl+x - exit

Data Wrangling and Processing for Genomics

Metadata location
- View https://github.com/datacarpentry/wrangling-genomics/blob/gh-pages/files/Ecoli_metadata_composite.csv
- Download link https://raw.githubusercontent.com/datacarpentry/wrangling-genomics/gh-pages/files/Ecoli_metadata_composite.csv

Background and Metadata

https://datacarpentry.org/wrangling-genomics/01-background/index.html
Challenge:

1. How many different generations exist in the data?
2. How many rows and how many columns are in this data?
3. How many citrate+ mutants have been recorded in Ara-3?
4. How many hypermutable mutants have been recorded in Ara-3?

Commands relevant to answering questions above:
1. $ cut -d ',' -f 2 metadata.csv | sort -u | wc -l
Subtract one due to the header
2. Counting rows: $ wc -l metadata.csv
subtract one due to the header
Counting columns $ head -n 1 metadata.csv | sed 's/,/\n/g' | wc -l
3. $ cut -d ',' -f 12 metadata.csv | grep -c 'plus'
4. $ cut -d ',' -f 6 metadata.csv | grep -c 'plus'

Assessing Read Quality

https://datacarpentry.org/wrangling-genomics/02-quality-control/index.html
What is the last read in the SRR2584863_1.fastq file? How confident are you in this read?

Exercise

Use the output from your Trimmomatic command to answer the following questions.
1) What percent of reads did we discard from our sample? 2) What percent of reads did we keep both pairs?

Exercise

We trimmed our fastq files with Nextera adapters, but there are other adapters that are commonly used. What other adapter files came with Trimmomatic?

Variant Calling Workflow

```
cp -R .backup/ref_genome data/
cp -R .backup/trimmed_fastq_small data/
bwa index data/ref_genome/ecoli_rel606.fasta

mkdir -p results/sam results/bam results/bcf results/vcf

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

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

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

```
cd results/bam
samtools index SRR2584866.aligned.sorted.bam
samtools tview --reference ../../data/ref_genome/ecoli_rel606.fasta SRR2584866.aligned.sorted.bam

bcftools stats results/vcf/SRR2584866_final_variants.vcf | less -S
```

```
cd
mkdir dc_workshop
mv data docs results scripts dc_workshop/
```

Script for variant calling workflow

```
set -e

# Move into results folder
cd ~/dc_workshop/results

# Set reference genome
genome=~/dc_workshop/data/ref_genome/ecoli_rel606.fasta

# Index reference genome
bwa index $genome

# Make output folders
mkdir -p sam bam bcf vcf

# Iterate over samples
for fq1 in ~/dc_workshop/data/trimmed_fastq_small/*_1.trim.sub.fastq
    do
    echo "working with file $fq1"

    # Extract sample name from file name
    base=$(basename $fq1 _1.trim.sub.fastq)
    echo "base name is $base"

    # Assign input and output file paths to variables
    fq1=~/dc_workshop/data/trimmed_fastq_small/${base}_1.trim.sub.fastq
    fq2=~/dc_workshop/data/trimmed_fastq_small/${base}_2.trim.sub.fastq
    sam=~/dc_workshop/results/sam/${base}.aligned.sam
    bam=~/dc_workshop/results/bam/${base}.aligned.bam
    sorted_bam=~/dc_workshop/results/bam/${base}.aligned.sorted.bam
    raw_bcf=~/dc_workshop/results/bcf/${base}_raw.bcf
    variants=~/dc_workshop/results/bcf/${base}_variants.vcf
    final_variants=~/dc_workshop/results/vcf/${base}_final_variants.vcf

    # Align paired end reads to reference genome
    bwa mem $genome $fq1 $fq2 > $sam

    # Convert output to BAM format
    samtools view -S -b $sam > $bam

    # Sort aligned reads by chromosome and position
```

```
    samtools sort -o $sorted_bam $bam

    # Index aligned reads BAM file
    samtools index $sorted_bam

    # Create coverage pileup for variant calling
    bcftools mpileup -O b -o $raw_bcf -f $genome $sorted_bam

    # Call variants
    bcftools call --ploidy 1 -m -v -o $variants $raw_bcf

    # Filter variants
    vcfutils.pl varFilter $variants > $final_variants

    done
```

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

Visualize the alignment of the reads for our SRR2584866 sample. What variant is present at position 4377265? What is the canonical nucleotide in that position?

Automating a Variant Calling Workflow

1) It is a good idea to add comments to your code so that you (or a collaborator) can make sense of what you did later. Look through your existing script. Discuss with a neighbor where you should add comments. Add comments (anything following a # character will be interpreted as a comment, bash will not try to run these comments as code).

2) The samples we just performed variant calling on are part of the long-term evolution experiment introduced at the beginning of our variant calling workflow. From the metadata table, we know that SRR2589044 was from generation 5000, SRR2584863 was from generation 15000, and SRR2584866 was from generation 50000. How did the number of mutations per sample change over time? Examine the metadata table. What is one reason the number of mutations may have changed the way they did?
Hint: You can find a copy of the output files for the subsampled trimmed FASTQ file variant calling in the ~/.solutions/wrangling-solutions/variant_calling_auto/ directory.

Day 3 feedback

One thing we did well

- Nice to apply the principles we have learned so far on real data sets
- it's quite practical and clearly organised
- The script is very useful also for after the workshop
- course material is very good.

One thing we could improve

- The pace was sometimes too high to code at the same time and eventually fix things not working
- I got lost for the biological meaning of trimmomatic, bwa and sam... maybe it has been explained but I missed, it can be found by google, but hope some interpreting if possible
- For me also the pace was a tiny bit too fast
- would be nice to get more info about the rational of using different aligners, and their use in different cases such as sequencing DNA or RNA
- possible improvement in course material: make folders before downloading

Day 4

export PS1_backup=$PS1
echo $PS1_backup
export PS1='$ '
export PS1='\n$

MDC Training Evaluation
Please fill in the **MDC training evaluation** to help us improve the MDC training portfolio:

https://survey.mdc-berlin.de/875143?lang=en

please also fill up up
https://carpentries.typeform.com/to/UgVdRQ?slug=2023-02-21-mdc-berlin-online

Day 4 feedback

One thing we did well

-
-
-
-
-

One thing we could improve

-
-
-
-
-