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

---

Zoom Link: https://ksu.zoom.us/j/93785525724?pwd=Y3oGaRnj46OZZjCEeZOlFSQWnW8BaF.1
Meeting ID: 937 8552 5724
Passcode: 629512
Do not connect to audio

Wireless: "KSU Guest"

Day 1: Attendees
Kendra Spahr - Academic Services Librarian, Kansas State University
David Molik - Computational Biologist, USDA ARS Arthropod-borne Animal Disease Research Unit (ABADRU)
Carol Sevin - Academic Services Librarian, Kansas State University
Gwendolyn Sibley - Scholarly Communications and Copyright librarian - Helper
Tom Misilo - Systems Administrator, Kansas State University
Carolyn Jackson Scholarly Communications & OER Librarian, K-State Libraries Center for Scholarly Publishing
Julia Eilert, KSU GRA in HWWGRU. Genetics masters program.
Ramgopal Prajapati, Postdoc, Plant molecular biology, Plant Pathology, K-State
Raissa Debacker - Postdoc, Plant Pathology/Kstate
Luiza Adami - Grad Student, Genetics/Plant Path, Kansas State University
Erin Schirtzinger - Research Assistant Professor, Diagnostic Medicine/Pathobiology, K-State
James Zhu USDA-ARS, Foreign Animal Disease Research Unit, NBAF
Patricia Assato - Research Associate, Diagnostic Medicine & Pathobiology, K-State
Amy Bernardo - USDA Genotyping Lab, HWWGRU
Raymond Deiser - Systems Librarian at Kansas State University. I maintain the libraries catalog and discovery layer (Search It -- https://k-state.primo.exlibrisgroup.com/discovery/search?vid=01KSU_INST:NewUI)
Rose Scott- Undergraduate Entomology - USDA-ARS Shults & Gerken labs - Interest in Arthropod borne diseases, genetics, and nonmodel species research. https://www.linkedin.com/in/rose-scott/
Sydney Andreano- VBS Graduate Student, department of Anatomy & Physiology in College of Veterinary Medicine at Kansas State University
Derek Samarian - Foreign Animal Disease Research Unit, NBAF
Neethu Francis- Visting Research Scholar, Genetics & Plant Breeding
Venkate

Dr Dave Turner - Beocat application scientist
Jacqueline Maille- Phd Student, Department of Entomology, Kansas State University
Ruolin Bian - Post Doc, Agronomy
Wei Zhao-visting scholar, Agronomy
Rupinder Singh -Phd Student, Department of Entomology, Kansas State University

Today's Setup

- Pre-workshop Survey - https://carpentries.typeform.com/to/wi32rS?slug=2024-08-05-genomics-workshop
- Download workshop files
    - Make a copy of the data file on your local system by going to https://github.com/carpentries-incubator/life-sciences-workshop and clicking 'Clone or download' followed by 'Download zip'. Extract the files in 'life-sciences-workshop-gh-pages\data' to a suitable directory
- make sure you have access to spreadsheet software
- make sure you have access to RStudio
- About BeoCat / SciNET
    - Dr. Dave Turner, K-State BeoCat (documentation:https://support.beocat.ksu.edu/Docs/Main_Page)
        - Rocky Linux with the slurm batch scheduler
        - Over 10k Interl/ADM cores on 310 compute nodes
        - 53 TeraBytes of RAM memory
        - Low-latency 30-100 Gbps InfiniBand/RoCE network
        - 1/10/40 Gbps Ethernet to a 1 PetaByte file server
        - 290 TB fast scratch space/fastscratch
        - 170 32-bit NVIDIA GPUs and 4 64-bit NCODOA P100s
        - 310 Compute Nodes access through Slurm
        - 4 Interactive nodes accessed only through OnDemand
    - BeoCat Group - any of them might answer support tickets
        - Director, Dan Andresen
        - Application Scientist, Dave Truner
        - Lead System Admin, Adam Tygar
        - System Admin, Nathan Wells
    - Tips
        - if you're accessing beocat on a computer that's not already logged into your beocat account, ssh *username*@headnode.beocat.ksu.edu
        - useful command: kstat --help (kstat is the monitoring tool, this command provides information about the jobs that are running, kstat --me will show your jobs)
        - use BeoCat when your computer is too slow or you don't have enough memory to run your work
    - Dr. David Molik, SciNet (https://scinet.usda.gov/guides/)
        - Need HPC like SciNet/BeoCat for bioinformatics tools that assemble genomes
        - slurm - queue manager
- Accessing BeoCat/SciNet
        - to login on a Mac, terminal ssh command
        - to login on a PC, PuTTY ssh command
        - SciNet: ssh login https://scinet.usda.gov/guides/access/ssh-login, first install smallstep or you can use open ondemand: https://scinet.usda.gov/guides/access/web-based-login

        - BeoCat: ssh <k-state eid>@beocat.ksu.edu or you can use open ondemand: https://support.beocat.ksu.edu/Docs/OpenOnDemand
        - A useful tool for windows machines: https://mobaxterm.mobatek.net/
    - Globus - to transfer files between institution and HPC (BeoCat/SciNet) accounts
        - USDA, ORCID/PID card to login for 48 hours
        - K-State, https://support.beocat.ksu.edu/Docs/Globus

Life Sciences Workshopow to use best practices for managing your research data, along with case studies and examples to help you use these techniques. https://guides.library.stanford.edu/data-best-practices

- File naming conventions vary with operating systems (ex. space in WIndows is fine, in Linux any space becomes '/')
- Linux commands help sort by last modification date so no need to include the date in the name
- Use a folder to separate/protect raw data from report/presentation files (ex. separate folders for archive, literature, reports, processed data, and raw data)

Data Management: Meta-data

- some data repositories generate DOI (figshare, zenodo, osf, Ag Data Commons)
- to archive code - GitHub and Zenodo link-in will generate a DOI for your GitHub code

Raw data - protect it! Make it a read only file, analyze/make changes in a copy
Data Management and Planning

- Look for guidance from funding agencies ex. https://www.nal.usda.gov/data/data-management-planning
- dmptool.org

Spreadsheets

- Professor Daniel Lemire from the University of Quebec, in a popular blog post titled **'You shouldn't use a spreadsheet for important work (I mean it)'**, states that *"spreadsheets are good for quick and dirty work, but they are not designed for serious and reliable work"*.
    - prone to errors, not reproducible, not designed for advanced statistics/large amounts of data/scalability
    - https://eusprig.org/research-info/horror-stories/
- Guiding principles
    - save and don't edit raw data
    - metadata, could be a spreadsheet but don't use tabs in excel spreadsheets (.csv is non-proprietary/more accessible format doesn't have tabs)
    - consistent treatment of empty cells described in dictionary
    - one cell, one datum (don't mix data ex. location, age in one cell)
    - rectangular data - one observation (row) x one variable (column), non-rectangular might have additional information like other information and tables in some columns/rows
    - dates
    - color
    - validation > data tab > data tools

R Introduction

- 'wide' data - lots of columns and fewer rows
- Tidy or 'long' data - one column contains every value, remaining columns give context1
- R is a programming language great for open science publishing methods/graphics for researchers

- RStudio - application/interface for coding in R
- up arrow - go back through things you've previously run
- na is handled differently by different scientific packages (also good to check how different versions handle things like 'na')

R Function notes:
sample - Imagine you have a bag of marbles, and you reach in for a handful. That is what the code will do with a set of data, like vectors or values.
Numeric: Any number that is not an integer, basically it has decimal places.
c() is used to create matrices

If your R console changes to a +, then hit the ESCAPE button and it will retern to >.

Rbind: is used to stack dataframes, and is dependent on the number of columns in two data frames

Rep: replicate. Used to replicate. Written as rep('value name', #) The # indicates how many times the value should replicate.

When you merge data frames, you have to indicate a value that it can merge from. It is best to pick a column header that is in all data frames you are combining.

You can use the $ to access certain cells under a column header.

Tuesday 2024-08-06
Attendees:
Carolyn Jackson, Scholarly Communication & OER Librarian, Instructor
Tom Misilo, K-State Libraries
Sydney Andreano- GRA

Library Resources and Research Support Tools
https://lib.k-state.edu/

List of Library Databases: https://guides.lib.k-state.edu/az/databases

Request Reference Help - https://kstate.qualtrics.com/jfe/form/SV_3PExB7S1B05rA4B (Both K-State and community)

Lib Guides - https://guides.lib.k-state.edu/

O'Reilly (database with tech books/tutorials) @ K-State https://go.oreilly.com/kansas-state-university
O'Reilly @ Manhattan Public Library https://links.mhklibrary.org/oreilly
Manhattan Public Library also has LinkedIn Learning https://links.mhklibrary.org/linked-in-learning

Project Organization and Management
Howard Hughes Medical Institute “Making the Right Moves: A Practical Guide to Scientific Management for Postdocs and New Faculty” section on Data Management and Laboratory Notebooks.

https://www.hhmi.org/sites/default/files/2023-10/making-the-right-moves-second-edition.pdf

The Digital Curation Center maintains a list of metadata standards (https://www.dcc.ac.uk/resources/metadata-standards/listand) some that are particularly relevant for genomics data are available from the Genomics Standards Consortium (https://www.gensc.org/pages/projects.html).

FAIR Standards https://www.go-fair.org/fair-principles/ FAIR data is Findable, Accessible, Interoperable, and Reusable.

Messy Data:
https://datacarpentry.org/organization-genomics/files/Ecoli_metadata_composite_messy.xlsx
https://raw.githubusercontent.com/datacarpentry/wrangling-genomics/gh-pages/files/Ecoli_metadata_composite.tsv

Library resources on R published since 2020:

- https://k-state.primo.exlibrisgroup.com/discovery/search?query=sub,equals,R%20(Computer%20program%20language)&tab=Everything&search_scope=MyInst_and_CI&vid=01KSU_INST:NewUI&facet=searchcreationdate,include,2020%7C,%7C2024&offset=0
- https://search.nal.usda.gov/discovery/search?query=sub,exact,R%20(Computer%20program%20language),AND&tab=LibraryCatalog&search_scope=MyInstitution&sortby=rank&vid=01NAL_INST:MAIN&facet=searchcreationdate,include,2020%7C,%7C2024&mode=advanced&offset=0

Project organization and management - Lesson
https://datacarpentry.org/organization-genomics/

Thought this UNIX Tutorial would be worth sharing: https://info-ee.surrey.ac.uk/Teaching/Unix/

Introduction to the Command Line for Genomics - Lesson
https://datacarpentry.org/shell-genomics/
Glossary and reference links - https://datacarpentry.org/shell-genomics/instructor/reference.html

For this section you will need a directory called "shell_data"
For Scinet:

On both ceres atlas this directory is located at:
/90daydata/shared/david.molik/shell_data
You can copy these with the following command:
cp -R /90daydata/shared/david.molik/shell_data ~

For Beocat
The space then '~' at the end abcp -rp /homes/daveturner/shell_data ~ove will put the shell_data into your home directory

Alternatively:

You can also download the files directly:
wget -O shell_data.tar.gz https://figshare.com/ndownloader/files/14417834
tar -xvf shell_data.tar.gz

https://datacarpentry.org/shell-genomics/instructor/02-the-filesystem.html#relative-path-resolution

Wednesday 2024-08-07
To move files for Genomics on USDA ARS systems:
cd ~
cp /90daydata/shared/david.molik/dc_genomics.sif .

To move files for Genomics on K-State **beocat**:
cd ~

cp -r /opt/beocat/shared/containers/dc_genmics.sif .

cd ~
wget -O dc_workshop.tar.gz https://figshare.com/ndownloader/files/14502617
tar -xvf dc_workshop.tar.gz -C dc_workshop --strip-components=1

Data wrangling:
https://datacarpentry.org/wrangling-genomics/
https://datacarpentry.org/wrangling-genomics/files/Ecoli_metadata_composite.csv

Running an interative job on our resources:
srun --nodes=1 --time=8:00:00 --mem=1G --pty --overlap bash
this allocates a node for you to use, scient/beocat relevant documentation tells you how to add more resources

Running a container:
(ceres and atlas): module load apptainer
beocat does not need to load the module

Next we run fastqc
apptainer exec /90daydata/shared/david.molik/dc_genomics.sif fastqc -h

but, you will run fastqc out of your directory:
apptainer exec /homes/<userid>/<optional>/dc_genomics.sif fastqc -h

KSU
apptainer exec /opt/beocat/shared/containers/dc_genomics.sif trimmomatic -h
you can also copy the container to your directory
apptainer exec /homes/.YOUR_LOGIN_NMAE/dc_genomic.sif trimmomatic -h

(1) download adaptor:

wget raw.githubusercontent.com/timflutre/trimmomatic/master/adapters/NexteraPE-PE.fa

(2) KSU Beocat:
apptainer exec /homes/YOURLOGIN/dc_genomics.sif 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

check your Trimmomatic run

ls SRR2589044*

human eye

ls SRR2589044* -l -h

Compression of file
gzip SRR2584863_1.fastq

USDA
```
for infile in *_1.fastq.gz
do
   base=$(basename ${infile} _1.fastq.gz)
   apptainer exec /90daydata/shared/david.molik/dc_genomics.sif trimmomatic PE \
   ${infile} ${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
```

KSU
```
for infile in  *_1.fastq.gz
do
   base=$(basename ${infile} _1.fastq.gz)
   apptainer exec /homes/YOUR_LOG_IN/dc_genomics.sif trimmomatic PE
done
```

Download 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$ gunzip data/ref_genome/ecoli_rel606.fasta.gz - ftp times out without downloading - FTP is closed at K-State for security reasons - can ask for individual site ftp to be opened ie

NCBI

KSU
cd data/ref_genome
Copy from the dorectory below:

wget -o 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- ftp times out without downloading (Works if following cd ~/dc_workshop & then mkdir -p data/ref_genome before using)
$ gunzip data/ref_genome/ecoli_rel606.fasta.gz

Download small data

curl -L -o sub.tar.gz https://ndownloader.figshare.com/files/14418248

tar xvf sub.tar.gz - makes a sub directory in data directory

$ mv sub/ ~/dc_workshop/data/trimmed_fastq_small - sub directory in data will disappear

Results note

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

reference genome upzip

Go to reference genome directory

gunzip ecoli_rel606.fasta.gz

Index the genome

General information for container
usda

apptainer exec /90daydata/shared/david.molik/dc_genomics.sif

KSU
apptainer exec /homes/YOURLOGIN/dc_genomics.sif
apptainer exec /homes/daveturner/genomic_workshop/dc_genomics.sif
apptainer exec /opt/beocat/shared/containers/dc_genomics.sif

your command

bwa index data/ref_genome/ecoli_rel606.fasta

Aligner run

Add USDA or KSU specific dc_genomics.sif file PATH

For #USDA
apptainer exec /90daydata/shared/david.molik/dc_genomics.sif bwa mem data/ref_genome/ecoli_rel606.fasta data/trimmed_fastq_small/sub/SRR2584866_1.trim.sub.fastq data/trimmed_fastq_small/sub/SRR2584866_2.trim.sub.fastq > results/sam/SRR2584866.aligned.sam

For KSU
apptainer exec /homes/YOURLOGIN/dc_genomics.sif
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

For August 08 , 2024 Notes:::

Sorting of BAM file by Co-ordinates
apptainer exec ../dc_genomics.sif samtools sort -o results/sam/SRR2584866.aligned.sorted.bam results/bam/SRR258486.aligned.bam

For USDA

Go to your terminal

Go to dc_workshop Directory

module load apptainer

General information for container command

apptainer exec YOUR-dc_genomics.sif COMMAND .....

Run this command

apptainer exec /90daydata/shared/david.molik/dc_genomics.sif samtools sort -o results/bam/SRR2584866.aligned.sorted.bam results/bam/SRR2584866.aligned.bam

KSU

wGo to your terminal

Go to dc_workshop Directory

- # If you need the dc_workshop directory:
- cd ~
- wget -O dc_workshop.tar.gz https://figshare.com/ndownloader/files/14502617
- tar -xvf dc_workshop.tar.gz -C dc_workshop --strip-components=1
- # Copy container to your home directory
- cp /opt/beocat/shared/containers/dc_genomics.sif ~

General information for container command

apptainer exec /homes/YOURLOGIN/dc_genomics.sif

run this command

apptainer exec /homes/YOURLOGIN/dc_genomics.sif samtools sort -o results/bam/SRR2584866.aligned.sorted.bam results/bam/SRR2584866.aligned.bam

Learning about the BAM file

USDA

apptainer exec /90daydata/shared/david.molik/dc_genomics.sif samtools flagstat results/bam/SRR2584866.aligned.sorted.bam

KSU

apptainer exec /homes/YOURLOGIN/dc_genomics.sif samtools flagstat results/bam/SRR2584866.aligned.sorted.bam

VARIANT CALLING

Step 1 --- Calculate read coverage

USDA

apptainer exec /90daydata/shared/david.molik/dc_genomics.sif bcftools mpileup -O b -o results/bcf/SRR2584866_raw.bcf \-f data/ref_genome/ecoli_rel606.fasta results/bam/SRR2584866.aligned.sorted.bam

KSU
apptainer exec /homes/YOURLOGIN/dc_genomics.sif bcftools mpileup -O b -o results/bcf/SRR2584866_raw.bcf \-f data/ref_genome/ecoli_rel606.fasta results/bam/SRR2584866.aligned.sorted.bam

Step 2 --- Detect single nucleotode variant (SNV)

USDA

apptainer exec /90daydata/shared/david.molik/dc_genomics.sif bcftools call --ploidy 1 -m -v -o results/vcf/SRR2584866_variants.vcf results/bcf/SRR2584866_raw.bcf

KSU375

apptainer exec /homes/YOURLOGIN/dc_genomics.sif bcftools call --ploidy 1 -m -v -o results/vcf/SRR2584866_variants.vcf results/bcf/SRR2584866_raw.bcf

Step 3: Filter and report the SNV variants in variant calling format (VCF)

USDA

apptainer exec /90daydata/shared/david.molik/dc_genomics.sif vcfutils.pl varFilter results/vcf/SRR2584866_variants.vcf > results/vcf/SRR2584866_final_variants.vcf

KSU

apptainer exec /homes/YOURLOGIN/dc_genomics.sif vcfutils.pl varFilter results/vcf/SRR2584866_variants.vcf > results/vcf/SRR2584866_final_variants.vcf

FILTERING data (SRR2584866_variants.vcf vs SRR2584866_final_variants.vcf)

USDA
Before Filter

apptainer exec /90daydata/shared/david.molik/dc_genomics.sif bcftools stats results/vcf/SRR2584866_variants.vcf | grep TSTV

After Filter

apptainer exec /90daydata/shared/david.molik/dc_genomics.sif bcftools stats results/vcf/SRR2584866_final_variants.vcf | grep TSTV

KSU
Before Filter

apptainer exec /homes/YOURLOGIN/dc_genomics.sif bcftools stats
results/vcf/SRR2584866_variants.vcf | grep TSTV

After Filter

apptainer exec /homes/YOURLOGIN/dc_genomics.sif bcftools stats
results/vcf/SRR2584866_final_variants.vcf | grep TSTV

Explore the VCF format:

less -S results/vcf/SRR2584866_final_variants.vcf

Assess the alignment (visualization) - optional step

USDA

Index of bam file

exec /90daydata/shared/david.molik/dc_genomics.sif samtools index
results/bam/SRR2584866.aligned.sorted.bam

Viewing with tview

apptainer exec /90daydata/shared/david.molik/dc_genomics.sif samtools tview
results/bam/SRR2584866.aligned.sorted.bam data/ref_genome/ecoli_rel606.fasta

KSU

Index of bam file

apptainer exec /homes/YOURLOGIN/dc_genomics.sif samtools index
results/bam/SRR2584866.aligned.sorted.bam

Viewing with tview

apptainer exec /homes/YOURLOGIN/dc_genomics.sif samtools tview
results/bam/SRR2584866.aligned.sorted.bam data/ref_genome/ecoli_rel606.fasta

Viewing with IGV

OPEN A NEW terminal in your laptop (powershell in Dell and Terminal in MacBook/Linux)

make a Directory

```
mkdir ~/Desktop/files_for_igv
```

change to that Directory

```
cd ~/Desktop/files_for_igv
```

Copy your files here for Viewing in igv

USDA

For ceres

<SCINetID>=Firstname.Lastname (FOR ME siddhartha.kanrar)

sorted BAM file
scp
<SCINetID>@ceres-dtn.scinet.usda.gov:~/dc_workshop/results/bam/SRR2584866.aligned.sorted.bam .

Index sorted BAM file
scp
<SCINetID>@ceres-dtn.scinet.usda.gov:~/dc_workshop/results/bam/SRR2584866.aligned.sorted.bam.bai .

reference genome

scp <SCINetID>@ceres-dtn.scinet.usda.gov:~/dc_workshop/data/ref_genome/ecoli_rel606.fasta .

final Variant file

scp <SCINetID>@ceres-dtn.scinet.usda.gov:~/dc_workshop/results/vcf/SRR2584866_final_variants.vcf .

For atlas

sorted BAM file
scp
<SCINetID>@atlas-dtn.hpc.msstate.edu:~/dc_workshop/results/bam/SRR2584866.aligned.sorted.bam .

Index sorted BAM file
scp

<SCINetID>@atlas-dtn.hpc.msstate.edu:~/dc_workshop/results/bam/SRR2584866.aligned.sorted.bam.bai .

reference genome

scp <SCINetID>@atlas-dtn.hpc.msstate.edu:~/dc_workshop/data/ref_genome/ecoli_rel606.fasta .

final Variant file

scp <SCINetID>@atlas-dtn.hpc.msstate.edu:~/dc_workshop/results/vcf/SRR2584866_final_variants.vcf .

Scripts

use scripts to automate a commonly used set of commands
write scripts in editor like nano
#! shebang tells computer which interpreter to use
#! /bin/bash

comments

ex. apptainer

- #aliases container for cleanliness
- alias run_c="apptainer exec /homes/USER/dc_genomics.sif"

ex. shortcut for ls -la

- alias ll= "ls -la"

A script for running **fastqc**

- #!/bin/bash
- # "#" denotes a comment line
- # USDA ARS requires appatiner to be loaded
- echo "Loadding apptainer"
- module load apptainer
- # Beocat does not require appatiner
- # Aliasing our container for cleanliness
- run_c="apptainer exec /90daydata/shared/david.molik/dc_genomics.sif"
- set -e
- cd /90daydata/shared/david.molik/dc_workshop/data/untrimmed_fastq/
- echo "Running FastQC ..."
- $run_c fastqc *.fastq*
- mkdir -p /90daydata/shared/david.molik/dc_workshop/results/fastqc_untrimmed_reads
- echo "Saving FastQC results..."
- mv *.zip /90daydata/shared/david.molik/dc_workshop/results/fastqc_untrimmed_reads/
- mv *.html /90daydata/shared/david.molik/dc_workshop/results/fastqc_untrimmed_reads/
- cd /90daydata/shared/david.molik/dc_workshop/results/fastqc_untrimmed_reads/
- # This portion unzips the fastqc .zip files
- echo "Unzipping..."
- for filename in *.zip

- do)
- unzip $filename
- done
- echo "Saving summary..."
- mkdir -p /90daydata/shared/david.molik/dc_workshop/docs
- cat */summary.txt > /90daydata/shared/david.molik/dc_workshop/docs/fastqc_summaries.txt
- R Code 8/8/24 :

https://figshare.com/articles/dataset/Data_Carpentry_Genomics_beta_2_0/7726454
data <- read.csv('/location/to/data/combined_tidy_vcf.csv')

OR

data <- read.csv(file.choose())

subset

subset <- data.frame(variants[,c(1:3, 6)])

alt_alleles <- subset$ALT

snps <- c(alt_alleles[alt_alleles == "A"],
alt_alleles[alt_alleles == "T"],
alt_alleles[alt_alleles == "G"],
alt_alleles[alt_alleles == "C"])

factor_snps <- factor(snps)

plot(factor_snps)

ordered_factor_snps <- factor(factor_snps, levels = names(sort(table(factor_snps))))

plot(ordered_factor_snps)

SRR258863_variants <- variants[variants$sample_id == "SRR2584863",]

snp_positions_2 <- c("8762685", "66560624", "67545785", "154039662")

install.packages("BiocManager")
BiocManager::version()
BiocManager::install("vcfR")

install.packages("dplyr") ## installs dplyr package
install.packages("tidyr") ## installs tidyr packagei
install.packages("ggplot2") ## installs ggplot2 package
install.packages("readr") ## install readr package

```
library("dplyr")          ## loads in dplyr package to use
library("tidyr")          ## loads in tidyr package to use
library("ggplot2")           ## loads in ggplot2 package to use
library("readr")          ## load in readr package to use

variants <- read.csv("https://raw.githubusercontent.com/datacarpentry/genomics-r-intro/main/episodes/data/combined_tidy_vcf.csv")
```