See these examples from 2019 for help as well. https://github.com/biodataprog/GEN220_2019_examples/tree/master/Bioinformatics_8

DNA Sequence Read alignment

Align genomic DNA reads from three different experiments to the genome.

We will align SARS-CoV-19 genomes against the reference genome.

1. The genome is available at NCBI RefSeq Accession NC_045512.2 - this is already downloaded for you in `/bigdata/gen220/shared/data/SARS-CoV-2`
2. I selected 3 random SARS-CoV SRA accessions `SRR11587604`, `SRR11140748`.
3. Align reads to the genome with `bwa` (or other tools you choose if you prefer). You'll need to index the genome - remember you want to make a symlink to this file.
4. Create the BAM files for this alignment using `samtools`.
5. Use `samtools` and the subcommand `flagstat` (or other tools if you want) to get a count for the number of reads which map to the genome.

I highlight reccomend reading the htslib tools docs if you want to see more on `samtools` and later on SNP calling with `bcftools`.

Additional things you can explore:

See other options for samtools - such as try the option to retrieve reads which are unmapped `samtools view -f 4`

Try using the `samtools fastq` option to dump out reads which are unmapped. For example.

SNP calling

Call SNPs from this same dataset to explore how variants are called to create a VCF file. Follow example from class.

Generate a table of SNP locations using `bcftools view` to reformat.

RNAseq and comparisons

Reanalyze data in this published paper Baker et al 2014 "Slow growth of Mycobacterium tuberculosis at acidic pH is regulated by phoPR and host associated carbon sources"

Data are downloaded to `/bigdata/gen220/shared/data/M_tuberculosis`

The Transcriptome file is also in the folder as `M_tuberculosis.cds.fasta` - I have already renamed the sequences to be the LOCUS names. It was downloaded from https://www.ncbi.nlm.nih.gov/assembly/GCF_000008585.1/ and the specific file is linked here

There is a sra_info.tab file which lists the sample accessions and their metadata so you can see what are the data sets. This is from the BioProject PRJNA226557 and the SRA Project SRP032513

Compare gene expression between two sets of conditions. - pH5.7 - pH7

And growth carbon source - Glycerol - Pyruvate

1. Run Kallisto to get the gene expression calculated from each sample - you will need the file `M_tuberculosis.cds.fasta` as the database and each of the 8 `.fastq.gz` files in the folder. You can make links to these files (`ln -s /bigdata/gen220/shared/data/M_tuberculosis/*.fastq.gz`. You do not need to uncompress the files, Kallisto can read gzip compressed files.

The goals here to run the results from this with DESeq2. You may need to install DEseq2 - you can do this by starting up an R terminal (eg on cmdline do `R`)

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("DESeq2")
```

If you have run kallisto - here is already written script that will generate a figure for you.

```
Rscript kallisto_DESeq.R
```

extra credit / extra

2. Run pfam analysis to get the Protein domains found in each protein - you will need the file `M_tuberculosis.pep.fasta`
3. Construct a tab delimited file which lists on each line

- The Gene (LOCUS) name
- The Protein length
- An average TPM across replicates for each condition (eg there will be 4 conditions, two replicates per condition)
- The Pfam Protein domains, separated by comma found in each Protein

FYI - to process the file and move the locus_tags as the sequence names I ran this regular expression (in Perl)

```
perl -p -e 's/>(\S+).+(\[locus_tag=([^\]]+)\])/>$3 $1 $2/' GCF_000008585.1_ASM858v1_cds_fro
```

I made the protein file of sequences using script from BioPerl. `bp_translate_seq.pl M_tuberculosis.cds.fasta > M_tuberculosis.pep.fasta`