For this demo, we will process RNA-seq data using kallisto/sleuth pipeline

Getting the data

1. This dataset contains paired-end RNA-seq of Saccharomyces cerevisiae under aerobic and anaerobic conditions, each with 2 replicates (r1 and r2).
2. Get all the files from https://drive.google.com/drive/folders/1iVyRCoYxFOpnhbkpiv5rf4ryjgf8PvhA?usp=sharing

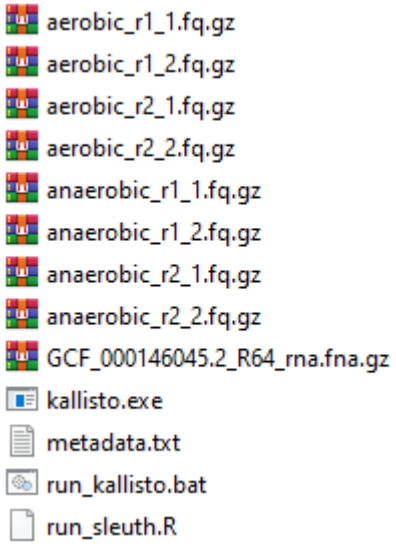
Setting up software

1. R
    a. At the time of the demo, R version 4.2.1 was the latest version.
    b. R repository in Thailand is at http://mirrors.psu.ac.th/pub/cran/
2. RStudio
    a. At the time of the demo, RStudio version 2022.07.1+554
    b. https://www.rstudio.com/products/rstudio/download/
3. sleuth
    a. Don't follow the guide on https://pachterlab.github.io/sleuth/download because those instructions are for outdated
    b. install.packages("BiocManager")
    c. BiocManager::install("rhdf5")
    d. **Note**: If you run into prompt: Update all/some/none? [a/s/n], it is safe to choose "n"
    e. install.packages("devtools")
    f. devtools::install_github("pachterlab/sleuth")
    g. Test that sleuth can be loaded using library(sleuth)
4. kallisto
    a. kallisto is a standalone software that run outside R
    b. The latest version is 0.46.1
    c. Download from https://pachterlab.github.io/kallisto/download
    d. Unzip
    e. To test that kallisto can be run, you need to open command prompt (CMD) on Windows or terminal on Mac OS, use command cd to move to the directory where you place kallisto, and then type kallisto to run as shown below

```
C:\Users\Sira\Downloads\kallisto>kallisto
kallisto 0.46.1

Usage: kallisto <CMD> [arguments] ..

Where <CMD> can be one of:

    index         Builds a kallisto index
    quant         Runs the quantification algorithm
    bus           Generate BUS files for single-cell data
    pseudo        Runs the pseudoalignment step
    merge         Merges several batch runs
    h5dump        Converts HDF5-formatted results to plaintext
    inspect       Inspects and gives information about an index
    version       Prints version information
    cite          Prints citation information

Running kallisto <CMD> without arguments prints usage information for <CMD>
```

Running the demo

1. Move the files so that you have everything inside a folder, as shown below



2. Click to launch run_kallisto.bat. You should see a pop-up command prompt like this.
   a. The first step in kallisto is to index the transcriptome database

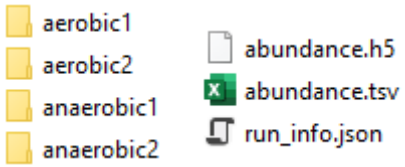
```
C:\Users\Sira\Downloads\yeast_data>kallisto index -i yeast_rna GCF_000146045.2_R64_rna.fna.gz

[build] loading fasta file GCF_000146045.2_R64_rna.fna.gz
[build] k-mer length: 31
[build] counting k-mers ... done.
[build] building target de Bruijn graph ...  done
[build] creating equivalence classes ...  done
[build] target de Bruijn graph has 11192 contigs and contains 8200305 k-mers
```

   b. Then, kallisto will perform pseudoalignment and quantify each transcript in each sample

```
C:\Users\Sira\Downloads\yeast_data>kallisto quant -i yeast_rna --bias -b 20 -o
fq.gz

[quant] fragment length distribution will be estimated from the data
[index] k-mer length: 31
[index] number of targets: 6,125
[index] number of k-mers: 8,200,305
[index] number of equivalence classes: 7,426
[quant] running in paired-end mode
[quant] will process pair 1: aerobic_r1_1.fq.gz
                             aerobic_r1_2.fq.gz
[quant] finding pseudoalignments for the reads ... done
[quant] learning parameters for sequence specific bias
[quant] processed 7,321,658 reads, 6,898,749 reads pseudoaligned
[quant] estimated average fragment length: 177.412
[   em] quantifying the abundances ... done
[   em] the Expectation-Maximization algorithm ran for 523 rounds
[bstrp] running EM for the bootstrap: 20
```

   c. At the end, you will see Press any key to continue . . .

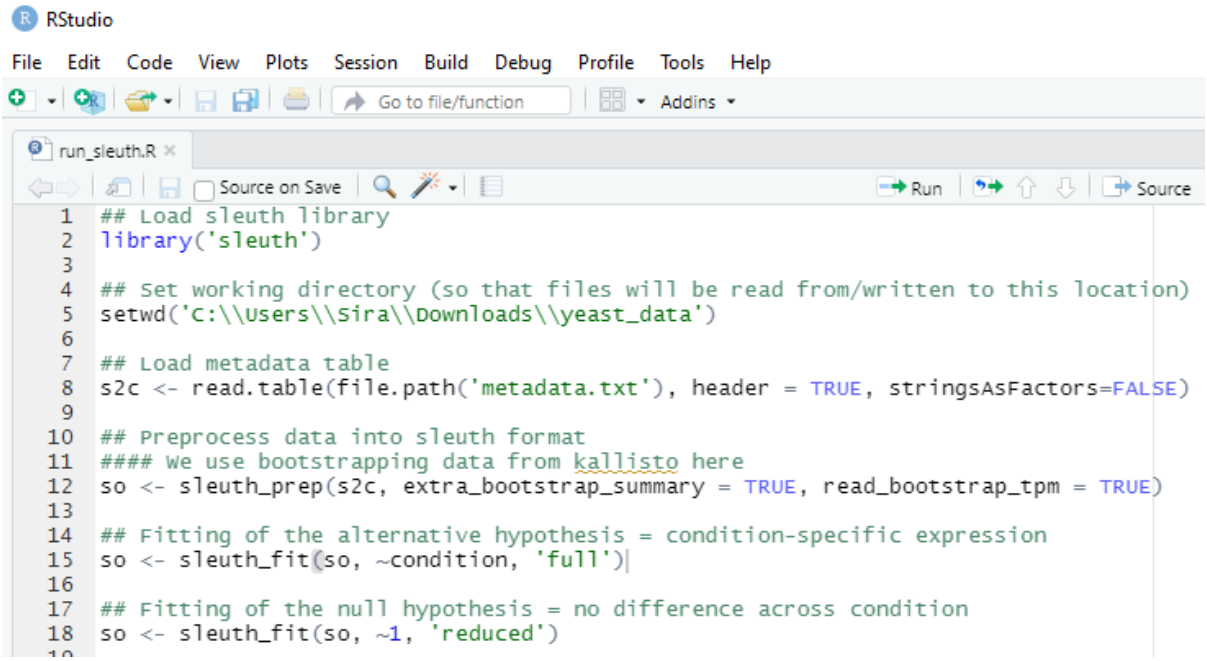
3. If kallisto ran successfully, you should get 4 output folders. Inside each, you should find 3 files as shown below. These contain transcript expression across 20 bootstraps from each sample.



4. Next, we will update the metadata.txt in excel to make sure that the path column matches the location of files on your computer

	A	B	C
1	sample	condition	path
2	aerobic1	aerobic	C:\Users\Sira\Downloads\yeast_data\aerobic1
3	aerobic2	aerobic	C:\Users\Sira\Downloads\yeast_data\aerobic2
4	anaerobic1	anaerobic	C:\Users\Sira\Downloads\yeast_data\anaerobic1
5	anaerobic2	anaerobic	C:\Users\Sira\Downloads\yeast_data\anaerobic2

5. Finally, we open RStudio and load the run_sleuth.R file
    a. Update the 5th row: setwd('PATH') to match where you placed the data files



6. We will go over the meaning of each command in class