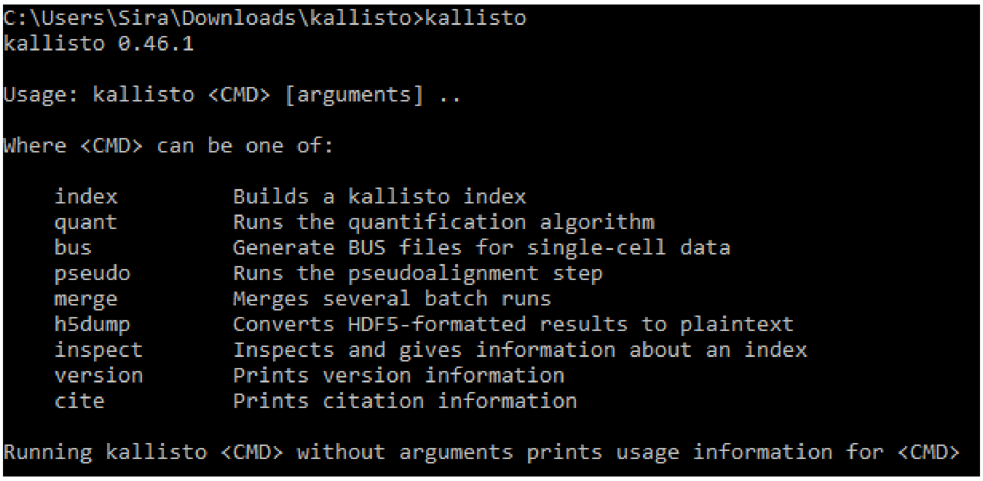
For this demo, we will analyze RNA-seq data with the kallisto-sleuth pipeline

Getting the data and script

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 the files from FigShare at https://figshare.com/articles/dataset/Yeast_RNA-seq_data_and_transcriptome_for_kallisto-sleuth_demo_session/24182520
3. The template R script for running sleuth for differential expression is provided at https://github.com/cmb-chula/comp-biol-3000788/blob/main/demo/run_sleuth.R

Setting up required software

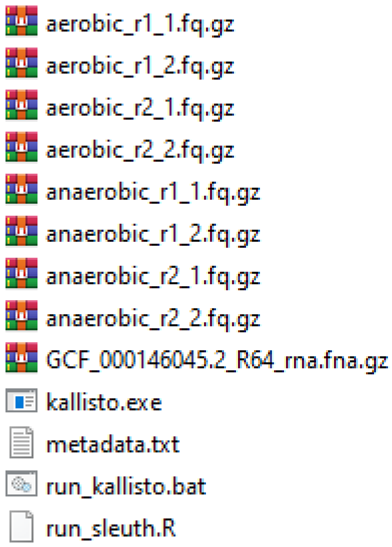
1. **R**: http://mirrors.psu.ac.th/pub/cran/
2. **RStudio**: https://www.rstudio.com/products/rstudio/download/
3. **sleuth**
   a. Don't follow the guide on https://pachterlab.github.io/sleuth/download because those instructions are for outdated!
   b. Type the following command into RStudio, one at a time.
      i. install.packages("BiocManager")
      ii. BiocManager::install("rhdf5")
      iii. install.packages("devtools")
      iv. devtools::install_github("pachterlab/sleuth")
   c. To test that **sleuth** can be loaded, type the command library(sleuth) in RStudio.
   d. If you run into prompt: Update all/some/none? [a/s/n], it is safe to choose "n".
4. **kallisto**
   a. Go to https://pachterlab.github.io/kallisto/download
   b. Look for the prebuilt software in the **Releases** section.
   c. Unzip the package.
   d. To test that **kallisto** can be run, you need to open **command prompt** (CMD) on Windows or **terminal** on Mac OS, and then type kallisto 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>
```

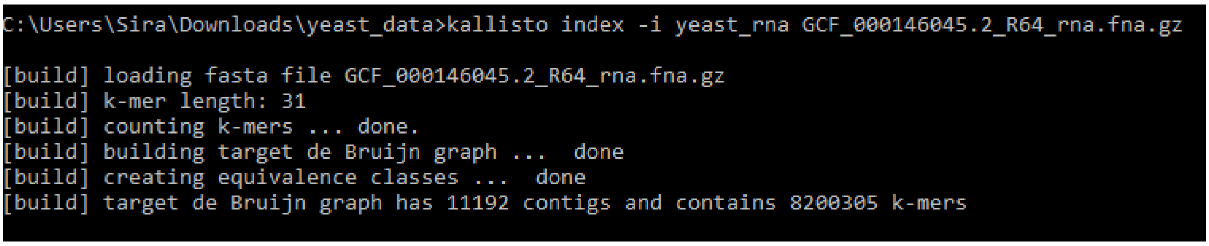
Running the demo

1. Move the files so that you have everything inside a folder, as shown below. There are some files that you will create during the class.



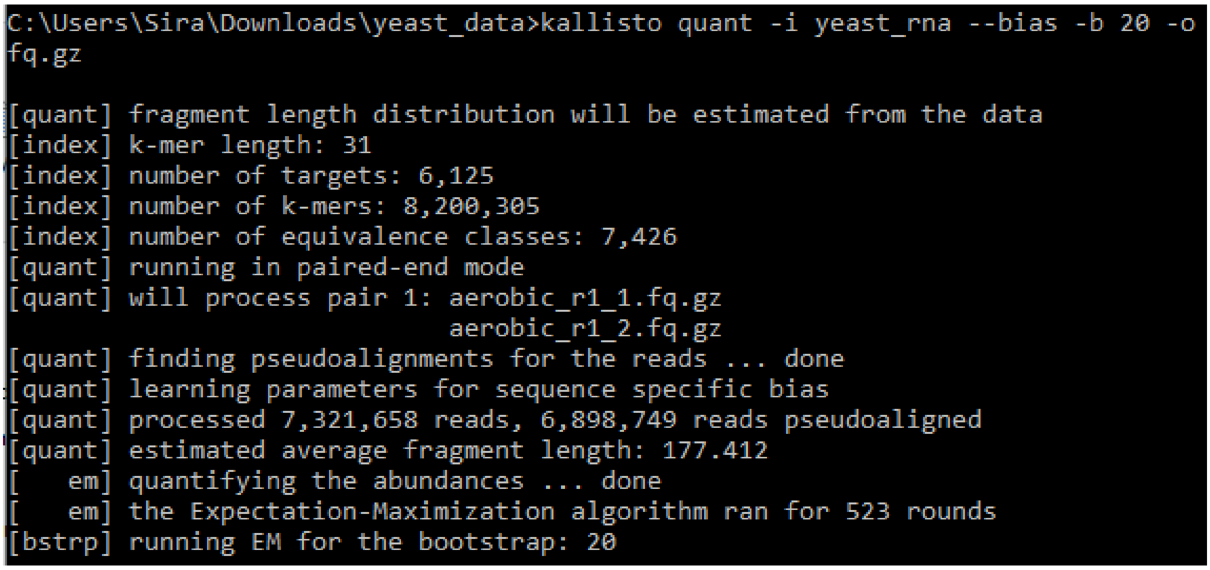
```
aerobic_r1_1.fq.gz
aerobic_r1_2.fq.gz
aerobic_r2_1.fq.gz
aerobic_r2_2.fq.gz
anaerobic_r1_1.fq.gz
anaerobic_r1_2.fq.gz
anaerobic_r2_1.fq.gz
anaerobic_r2_2.fq.gz
GCF_000146045.2_R64_rna.fna.gz
kallisto.exe
metadata.txt
run_kallisto.bat
run_sleuth.R
```

2. Use **kallisto** to map RNA-seq reads to the reference transcriptome.
   a. First, **index** the reference transcriptome.



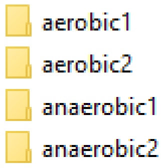
```
C:\Users\Sira\Downloads\yeast_data>kallisto index -i yeast_rna GCF_000146045.2_R64_rna.fna.gz

[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. Next, perform **pseudoalignment and quantify** transcript abundance.



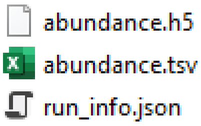
```
C:\Users\Sira\Downloads\yeast_data>kallisto quant -i yeast_rna --bias -b 20 -o
fq.gz

[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. If **kallisto** ran successfully, you should get **4 output folders**, one for each sample.



aerobic1
aerobic2
anaerobic1
anaerobic2

d. Inside each folder, you should find 3 files as shown below. If you don't have abundance.h5 file, there might be a problem with your **kallisto** (this is known if you install **kallisto** without using the pre-built software).

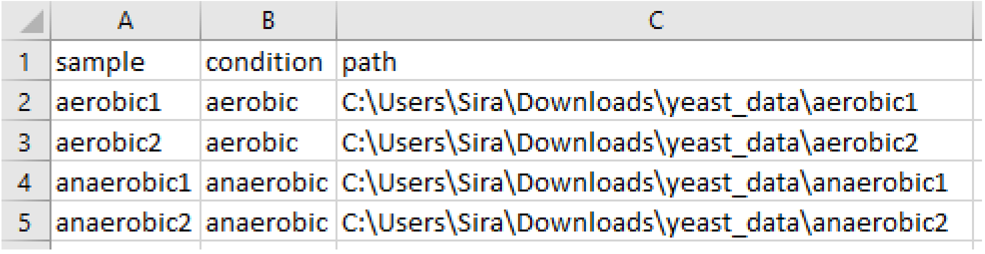


abundance.h5
abundance.tsv
run_info.json

3. Use **sleuth** to perform differential expression analysis.

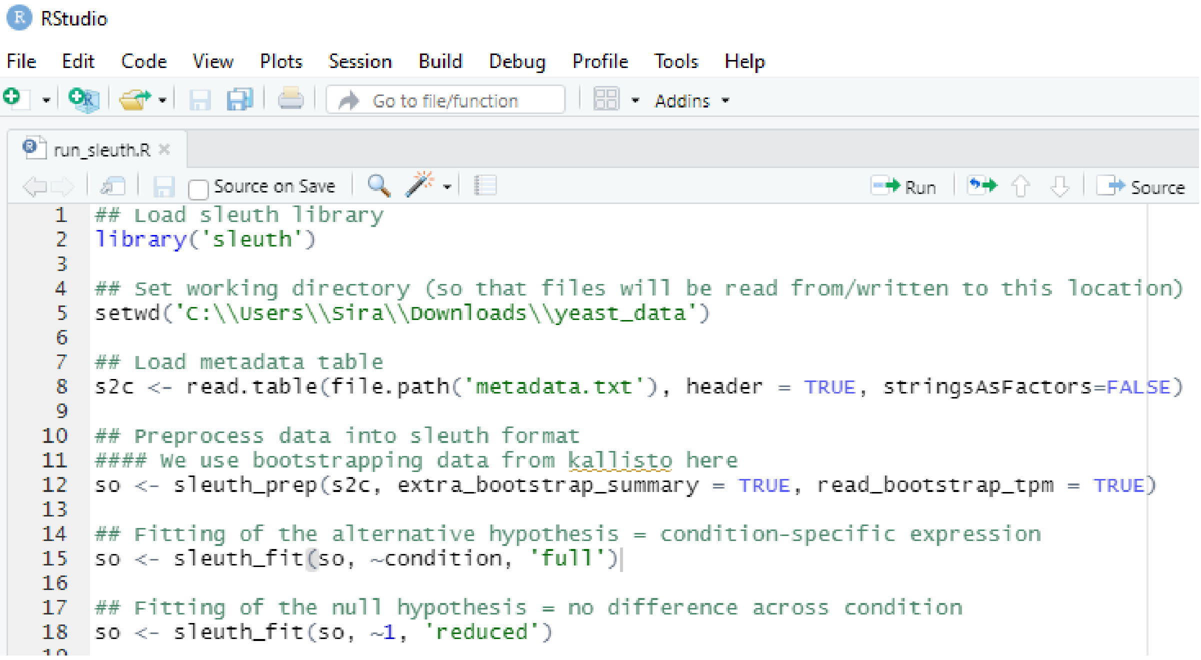
a. First, we create the metadata.txt table in excel to summarize the sample, metadata, and the path to the output folders from **kallisto**.

b. Make sure that the path column matches the location 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

c. Open RStudio and load the run_sleuth.R file. We will edit this script in class.



```r
## Load sleuth library
library('sleuth')

## Set working directory (so that files will be read from/written to this location)
setwd('C:\\Users\\Sira\\Downloads\\yeast_data')

## Load metadata table
s2c <- read.table(file.path('metadata.txt'), header = TRUE, stringsAsFactors=FALSE)

## Preprocess data into sleuth format
#### We use bootstrapping data from kallisto here
so <- sleuth_prep(s2c, extra_bootstrap_summary = TRUE, read_bootstrap_tpm = TRUE)

## Fitting of the alternative hypothesis = condition-specific expression
so <- sleuth_fit(so, ~condition, 'full')

## Fitting of the null hypothesis = no difference across condition
so <- sleuth_fit(so, ~1, 'reduced')
```