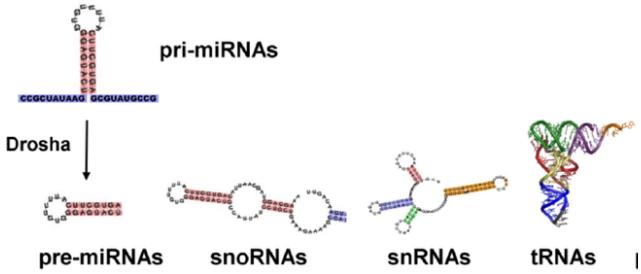
Small RNAseq exercises



Group work

Work in groups.

The group exercises

Exercise 1 – Small RNA size distribution
Exercise 2 – Differential expression of small RNAs

The computer environment - reminder

- Login to FOX:

```
ssh ec-username@fox.educloud.no
```

Use 2 factor authentication as described in link above.
Now you are in the login nodes. Here you can do small operations such as read, copy and move files etc.

- Login to the interactive nodes

```
ssh int-1
ssh int-2
ssh int-3
ssh int-4
```

We will assign you to the nodes to divide the work load.

Home folder:

```
/fp/homes01/u01/ec-trinro
```

Exercise folder:

```
/projects/ec34/in-biosx000/smallRNA
```

Make a directory with your username at the exercise result folder and write your results in there.

```
cd /projects/ec34/in-biosx000/smallRNA/results
mkdir username
```

Task 1: Size distribution

AIM:
Visualize the size distribution (lengths) of RNAs from the small RNAseq data

Data for this lecture are here:
`Cd /cluster/projects/nn9989k/BIOS_5410/smallRNA/data`

Files:
Sample10_clipped_single.fq
Sample11_clipped_single.fq
Sample12_clipped_single.fq

The files are trimmed and uncompressed
Choose one!

Lets look at the fastq format and use AWK for getting an overview of the RNA sequence lengths.

The fastq format:

```
@D00132:185:C9BFAANXX:4:2206:1479:2204 1:N:0:GGCTAC
GCCATAGACGGTGATAGTCCGGTAGACGAAAACTCA
+
CCCCCGGCGGGGGGGGGGGGGGGGGGGGGGGGGGG
@D00132:185:C9BFAANXX:4:2206:1566:2216 1:N:0:GGCTAC
GGCTGGTCCGATGGTAGTGGGTTATCAGAAA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGG
```

We have the sequence in every fourth line, starting with the second

We can extract every fourth line using for example AWK
Awk is programming language designed for text processing
https://www.gnu.org/software/gawk/manual/gawk.html
Awk has built-in Variables

Awk NR gives you the total number of records being processed or line number
awk 'NR%2==1' file.txt – prints every second line starting with the first in file.txt

Try this in the folder with the fastq files
awk 'NR%4==2' FASTQ.FILE

Length of sequence
awk length(string) calculates the **length** of a string

You need to specify the string
AWK treats tab or whitespace for file separator by default
$0 is the whole line, $1 for first field…$n for nth field

Try
awk '{if(NR%4==2) print length($0)}' FASTQ.FILE

Piping in Unix or Linux |

A pipe is a form of redirection (transfer of standard output to some other destination) that is used in Linux and other Unix-like operating systems to send the output of one command/program/process to another command/program/process for further processing.

Sort the length

SORT command is used to sort a file, arranging the records in a particular order. By default, the sort command sorts file assuming the contents are ASCII. Using options in sort command, it can also be used to sort numerically. https://www.geeksforgeeks.org/sort-command-linuxunix-examples/

awk '{if(NR%4==2) print length($0)}' FASTQ.FILE | sort

Count the lengths

The uniq command in Linux is a command line utility that reports or filters out the repeated lines in a file. -c option count these lines.
https://www.geeksforgeeks.org/uniq-command-in-linux-with-examples/?ref=rp

awk '{if(NR%4==2) print length($0)}' FASTQ.FILE | sort | uniq -c

Print to file

The work of any command is either taking input or gives an output or both. ">" is redirecting to standard output.
https://www.geeksforgeeks.org/input-output-redirection-in-linux/

awk '{if(NR%4==2) print length($0)}' FASTQ.FILE |sort |uniq -c > ../USERNAME/ FASTQ_len.txt

2. Plot the size distribution in R studio

To start R studio on Educloud:

1. Browse to https://ondemand.educloud.no/
2. Log in with Educloud credentials (e.g. ec-username, OTP, password)
3. Click on "RStudio Server Containerized"
4. Select options, e.g.:
   - Educloud project: ec34
   - Resources: Small (8 cores, 16GB RAM)
   - Runtime: 3 hour
   - Container: Custom
   - Additional modules: (none)
5. Path to R module: /projects/ec34/biosin5410/rstudio/rstudio-extra.sif
6. Click Launch

7. Wait until the button "Connect to RStudio Server" appears. The job may be waiting in the queue for some time.
8. Click "Connect to RStudio Server" appears.
9. You are inside Rstudio!

To quit:
1. Enter q() in the R terminal, or select "Quit session" from the File menu inside Rstudio.
2. Close the browser window with RStudio.
3. Click "Cancel" on the RStudio instance in the list of "My interactive Sessions". This will free the reserved resources.

I found it difficult to access files in the project folder (/projects/ec34) from within RStudio. I therefore created a link to that folder from my home directory. To do that:

1. Log in to Fox by ssh, e.g. "ssh ec-username@fox.educloud.no"
2. Make the link: "ln -s /projects/ec34 ~"
3. Files in the project folder (e.g. in the course folder biosin5410) can now be accessed from within RStudio in the folder ec34 in your home directory.

Go to your folder with the length distribution file.
getwd()

Read the size distribution files and check the data
len<-read.table("FILENAME.TXT")
Use head() and str()

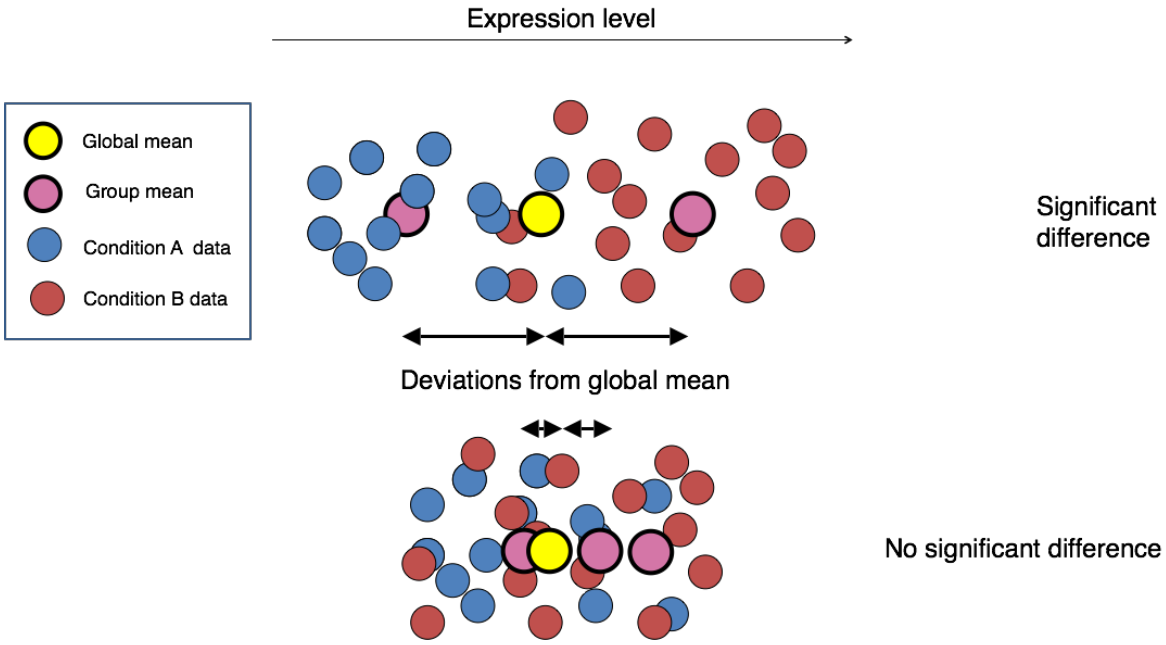
Change the columnames
colnames(len)<-c("counts","lengths")

Plots in R
R is a powerfull tool for making graphs. There are many ways of creating very nice plots. There are many packages for plotting. We will use basic plotting, but I recommend looking into the packages ggplot2, Lattice and Ploty and RColorBrewer for nice colours.

Plot the distribution in a scatterplot
plot(len$lengths, len$counts)

Make it into a line plot
plot(len$lengths, len$counts, type="l")

Make it nice, add tittle and axis labels.
plot(len$lengths, len$counts, type="l", main="RNA length distribution", xlab="length", ylab="counts")



3. Differential expression of small RNAs

AIM: Identify differential expressed circulating RNAs between serum samples from lung cancer and healthy individuals.

I have provided RNA sequencing counts for three small RNA classes miRNA, piRNA and tRF – fragments of tRNAs. These data are synthetic small RNA count data derived from the project https://www.kreftregisteret.no/en/Research/Projects/Identification-of-early-cancer-biomarkers/ and the publications:

Sinan U Umu, Hilde Langseth, Verena Zuber, Åslaug Helland, Robert Lyle and Trine B Rounge, 2022. "Serum RNAs can predict lung cancer up to 10 years prior to diagnosis".

Umu SU, Langseth H, Keller A, Meese E, Helland Å, Lyle R, Rounge TB, 2019. "A 10 year prediagnostic followup study shows that serum RNA signals are highly dynamic in lung carcinogenesis"

Hilde Langseth, Sinan Ugur Umu, Cecilie Bucher-Johannessen, Ronnie Babigumira, Magnus Leithaug, Marianne Lauritzen, Paolo Vineis, Giske Ursin, Robert Lyle, Trine B. Rounge, 2021. "Data Resource Profile: thousands of circulating RNA profiles of pre-clinical samples from the Janus Serum Bank Cohort" (preprint)

We are not allowed to share the real small RNA data with you since it may contain personal health information. However, the synthetic data contain the same data structures as the real data set.

- We will be using DEseq2 for differential expression analyses of data derived from lung cancer and health individuals. Start with miRNAs. The two others are for extras challenges.

- I have only given the commands to use here, use the DEseq2 tutorial to find out how to use them.
- If you are stuck, I have provided a R script with the correct commands available too – but please, try first.

Manual and tutorial:

DESeq2 vignett:
https://www.bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html
DESeq2 manual:
https://bioconductor.org/packages/release/bioc/manuals/DESeq2/man/DESeq2.pdf

This is an exercise in using manuals and the internet as help to solve a task. Therefore, I have not here provided you with a detailed R code here. First see if you can solve this yourself and understand what you do! If not, there is a solution file on Educloud in the smallRNA folder and you can always ask us for help ☺.

R packages
First you need to install packages you need.
This takes time and can be challenging due to dependencies (other packages) that don't match. Therefore we have loaded the packages you need with R on Educloud.

Load the package you need
Use library() or require()
Load DEseq2

Read in the data you need
Locate the small RNA files in R

- lc_dataset.csv
- lc_mirna_counts_LC_new.csv

Optional files:

- lc_tRF_counts.csv
- lc_piRNA_counts.csv

Use getwd(), setwd(), list.files()

Read the csv files:
Be aware of separators and headers
Use read.csv() option sep and header to separate the columns with "," and tell R that the first line is the header.

- What is in the lc_dataset.csv file? How many men and women? How many lung cancer cases and how many controls? Which types of lung cancer is present in this dataset?

Data preparation: check and clean the data
Use the str() function to check your data
The differential expression analyses will only accept integers in the count tables

- Change rowname to RNA names and remove the RNA name column
  - Use rownames()

    - And remove column dataframe[-1]
- What is in the lc_mirna_counts.csv file?
    - What is the dimensions of the file, dim()?
    - Which miRNA has the highest, lowest and average expression in the dataset, RowSums(), min(), max(), rowMeans()?
    - Which sample has the highest and lowest expression in the dataset, colSums()?
- Extra: Can you plot the miRNA expression and sample counts?

Differential Expression (DE) analyses
Carefully set up your design variable
Use DESeqDataSetFromMatrix
countData - your filtered count data frame
colData – the dataframe with the contrast groups
design ~ will contrast lung cancer cases vs controls

```
dds <- DESeqDataSetFromMatrix(countData = miRNA_counts, colData =
                  lc_diag, design =~ condition
```

Normalise and analyse the count file using DESeq2
DESeq()

```
dds_process <- DESeq(dds)
```

This will take a bit of time

- What does the `DESeqDataSetFromMatrix() and DESeq()` functions do? Read the manual?

Extract the results from the DE analyses
results()
summary()

Identified the differential expressed small RNAs
- Extract the results with alpha (q value) less than 0.05 as a criteria for significance
- res_05 <- results()
- summary()

Write a table of the significant miRNA to a file
- write.table()

Ekstra:
- Plot the the log2foldchange vs -log10 pvalue
- Do the same analyses for one of the other small RNA class and compare counts and DE.