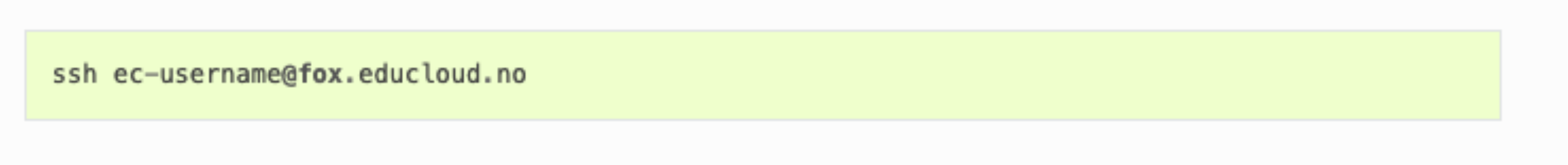
Practical work:

Small RNA sizes and differential expression

Task 1

Size distribution

Login - reminder



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

- One-time password and password
- https://uio-in-biosx000.readthedocs.io/en/latest/Educloud/index.html
- https://www.uio.no/english/services/it/research/platforms/edu-research/help/fox/index.md
- Work on the interactive nodes
- ssh int-<N> choose the one with the least load
- Files are availeble here:
  - `/projects/ec34/in-biosx000/smallRNA`

P.1 Size distribution

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

- Data for this lecture are here:
  - `/projects/ec34/in-biosx000/smallRNA/fastq`

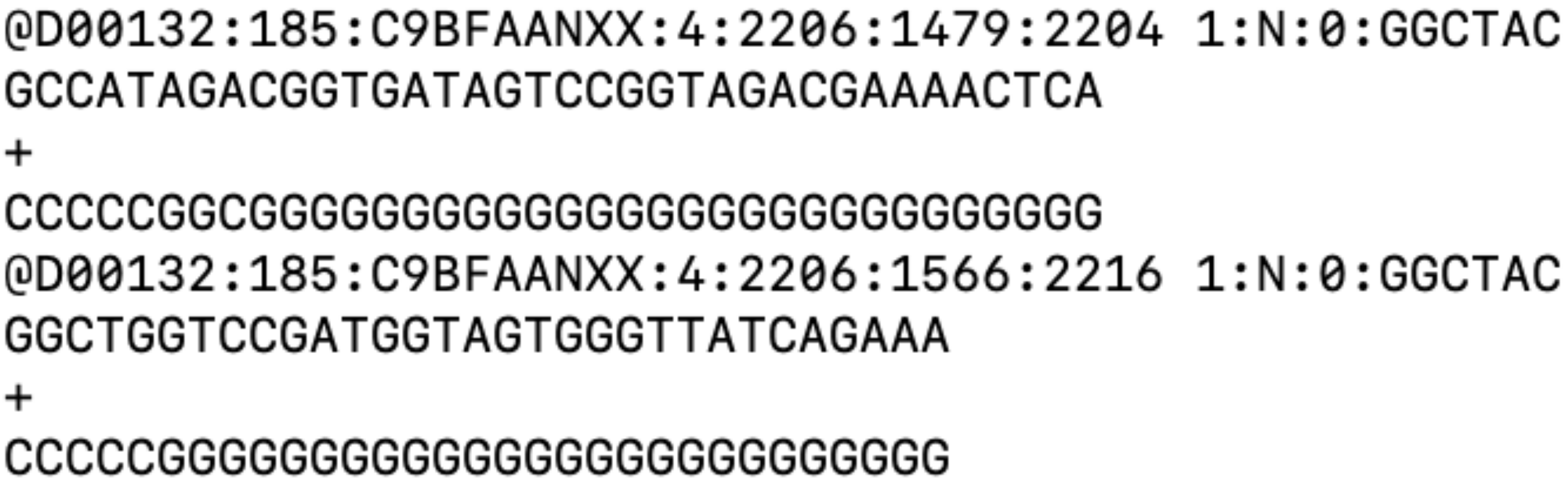
- Files
  - Sample10_clipped_single.fq
  - Sample11_clipped_single.fq
  - Sample12_clipped_single.fq

- The files are trimmed and uncompressed

- Choose one!

Size distributions

- The fastq format:



```
@D00132:185:C9BFAANXX:4:2206:1479:2204 1:N:0:GGCTAC
GCCATAGACGGTGATAGTCCGGTAGACGAAAACTCA
+
CCCCCGGCGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@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

Size distributions

- We can extract every fourth line using for example AWK
  - programming language designed for text processing
  - Awk Built-in Variables
  - **Awk NR** gives you the total number of records being processed or line number
  - awk 'NR%2==1' filename.fq – prints every second line starting with the first in file.txt

- awk 'NR%4==2'

Size distributions

- We need the length of the sequence
  - **awk length**(string) calculates the **length** of a string

  - Length of what ?
  - AWK treats tab or whitespace for file separator by default
  - $0 is the whole line, $1 for first field…$n for nth field

- awk '{if(NR%4==2) print length($0)}' Sample10_clipped_single.fq

Size distributions

- Sort the length
  - awk '{if(NR%4==2) print length($0)}' Sample10_clipped_single.fq|sort

- Count the lengths
  - awk '{if(NR%4==2) print length($0)}' Sample10_clipped_single.fq|sort |uniq -c

- Print to file

  awk '{if(NR%4==2) print length($0)}' Sample10_clipped_single.fq|sort |uniq -c >
  home/Sample10_len.txt

Size distributions

- Plot the size distributions

R

- Programming language
- Free software environment for statistical computing

- Long video introduction
  - https://www.youtube.com/watch?v=_V8eKsto3Ug&ab_channel=freeCodeCamp.org
- Short video introduction:
  - https://www.youtube.com/watch?v=SWxoJqTqo08&list=PLjgj6kdf_snYBkIsWQYcYtUZiDpam7ygg&ab_channel=DataCamp

R on Educloud

- https://www.uio.no/english/services/it/research/platforms/edu-research/help/fox/installing-software-r.md
- module spider Bioconductor
- module load R-bundle-Bioconductor/3.9-foss-2019a-R-3.6.0
- R

# Size distributions

## Size distributions of small RNA seq data

- R
- setwd("PATH")
  - setwd("/fp/homes01/u01/ec-trinro/smallRNA/test1”) # example

- ## Read the size distribution files

len<-read.table("filename.txt")

len

colnames(len)<-c("counts","lengths")

Size distributions

```
## In R - plot the distribution
plot(len$lengths, len$counts)

## with lines
plot(len$lengths, len$counts, type="l")

## make it nice
plot(len$lengths, len$counts, type="l", main="RNA length distribution",
xlab="length", ylab="counts")
```

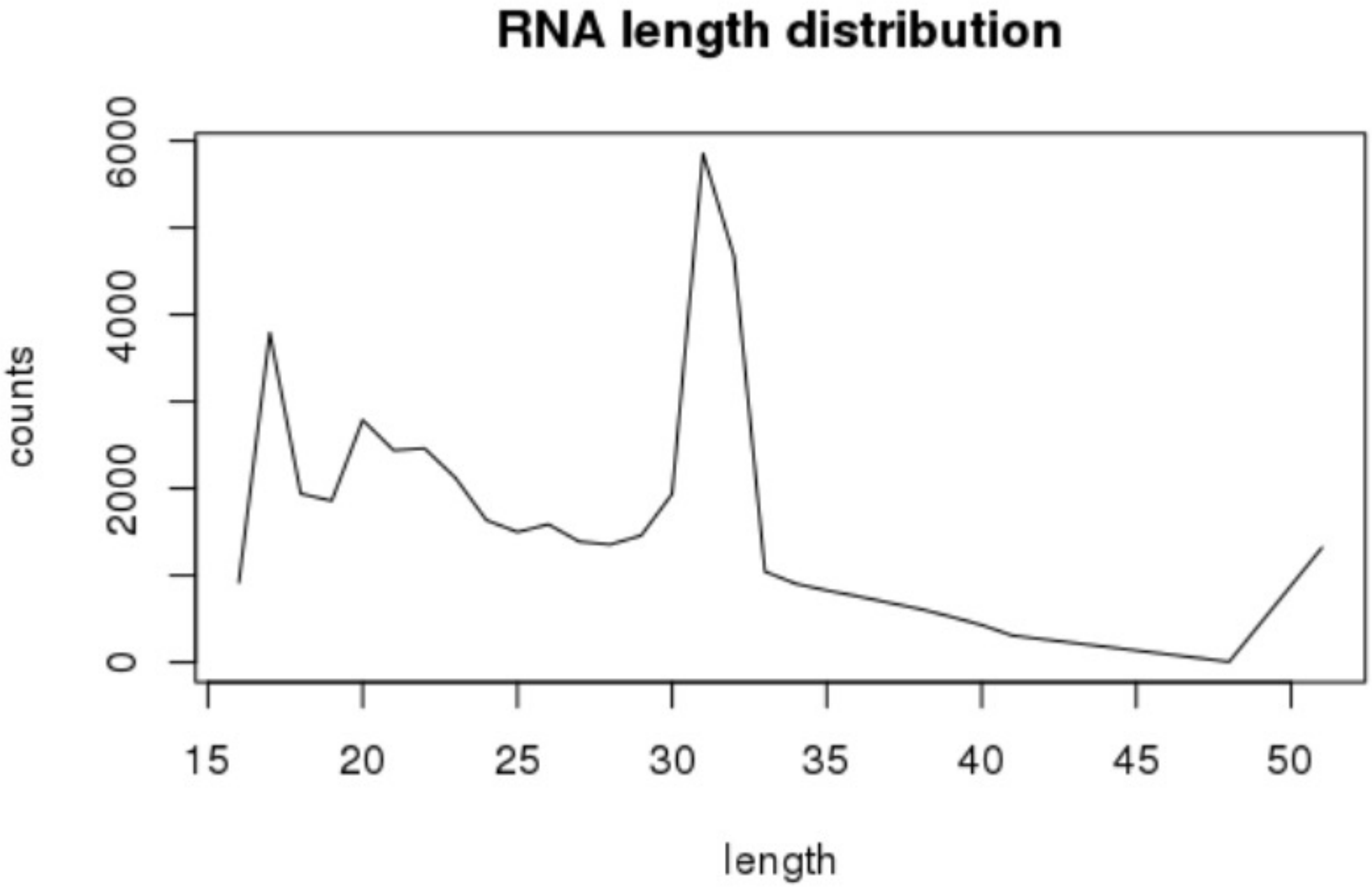
Size distributions

```
## In R

# Open a pdf file
pdf("rplot.pdf")

# plot
plot(len$lengths, len$counts, type="l", main="RNA length distribution", xlab="length", ylab="counts")

# Close the pdf file
dev.off()
```


RNA length distribution
counts
0
2000
4000
6000
15
20
25
30
35
40
45
50
length

Task 2

Differential expression

Task

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

- The data for this task is here:
  - /work/IN-BIOSx/data/smallrna/de
- Select RNA class
  - miRNA
  - piRNA
  - tRF
- We will be using DEseq2 for differential expression analyses

Need help – new to R ?

Take a look at the script is you are stuck:

- https://drive.google.com/file/d/18BqumBzN6BHjKWuh2yMTBXddK35l2fEb/view?usp=sharing

R package

- Install packages you need (This takes time – so lets skip this)

```
# manually install Hmisc - old version due to issues
#install.packages("https://cran.r-project.org/src/contrib/Archive/Hmisc/Hmisc_3.9-3.tar.gz", repos = NULL, type="source",dependencies=TRUE)
# https://bioconductor.org/packages/release/bioc/html/DESeq2.html
#if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
```

- You will use libraries I already installed for you to avoid problems
- .libPaths("/work/IN-BIOSx/data/smallrna/Rlibs")

- Load the pachage you need
  - Use library() or require()
  - Load DEseq2

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

Files

- Locate the small RNA files
- Set working directory
  - Use getwd() and setwd()
- Read in the count table
  - Use read.csv()
  - Be aware of separators and headers
- Read in the metadata table with case vs control
  - Use read.csv(), remember separators and headers

Check your data

- Use the str() function to check your data

- The differential expression analyses will only axcept int in the count tables

- Change rowname to RNA names and remove the RNA name column
  - Use rownames()
  - And remove column dataframe[-1]

Check your data

## What is the average number of counts per RNA?

- rowMeans()
- plot(rowMeans())
- ## What is the average number of counts per sample?
- colMeans()
- plot(colMeans())

Filter your data

- To remove very low count RNAs
  - Reduce multiple testing
  - rowMeans

- Remove RNAs with mean counts less than 100
  - `df[ which(rowMeans(df)>100), ]`

Design the DE analyses

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

DE analyses

- Normalise and analyse the count file using DESeq2
  - DESeq()
  - This will take a bit of time
- Extract the results
  - results()
  - summary()

Identify RNAs that are DE

- # Extract the results with alpha (q value) less than 0.05 as a criteria for significance
  - res_05 <- results()
  - summary()

- # Extract significantly DE list and write them to a file
  - subset()
  - write.table()

Visulize the result

- Refer to DESeq2 manual for plot description
  - plotDispEsts(dds_process)
  - plotPCA(DESeqTransform(dds_process))
  - plotMA(dds_process)
  - sizeFactors(dds_process)

Extra: Evaluate your results

- Select one DE RNA
- What is the strenght of the result
- What is the biological meaning of the result
  - Use internett resourses such as
    - UCSC genome Browser, TargetScan, +++