

The goal of this homework is to familiarise you with some of the aspects of rna-seq, including:

1. Processing of RNA-seq
2. Quality Control in RNA-seq
3. Analysis of RNA-seq

**Any of these steps you can do using your computers or a computational cluster provided by us. If you decide to use your own laptops/computers make sure to install the packages that are used in the pipelines from the scripts.**

**Even if you decide to do this homework using your own computer, you can use all the scripts provided in the folder as a reference on how to use the tools.**

**If you decide to do this homework using your own computer, please, see file `download_reference.bash` to get the fasta/gtf and other files for mouse reference genome. If you are using our cluster, reference is already downloaded to folder `/opt/data/reference`.**

**Using our cluster you might see error/warning like**

`/opt/conda/lib/libtinfo.so.6: no version information available`

**Just ignore it.**

First thing to start, make sure you download the reference genome and create genome indexes for kallisto and Hisat2:

1. Download `index.bash`
2. Put it in the folder “`index`” in your home folder (create the folder, if it's not there)
3. Run `bash index.bash` (this command might take some time to run, please be patient)

Now you are ready to start running the pipeline, however, we have to know which datasets to process.

1. To do that, find your dataset at GEO omnibus ( I will use GSE116239 as an example)

The screenshot shows the NCBI GEO Accession Display page for GSE116239. The page header includes the NCBI logo and the GEO Gene Expression Omnibus logo. The main content area displays the following information:

- Series GSE116239** (with an UPDATE button)
- Query DataSets for GSE116239**
- Status:** Public on Aug 30, 2018
- Title:** Bulk RNA-seq of foamy and non-foamy intimal macrophages from ApoE -/- mouse
- Organism:** *Mus musculus*
- Experiment type:** Expression profiling by high throughput sequencing
- Summary:** Macrophages in atherosclerotic aorta are major population in lesion and contributes to lesion formation by becoming foam cells. To investigate in vivo transcriptome profiles of those macrophages, we extracted foamy and non-foamy macrophages from atherosclerotic aorta using lipid probe-based flow cytometry sorting. Our data indicates that intima non-foamy and foamy macrophages show different mRNA expressions. Rather than non-foamy macrophages, foamy macrophages expressed more genes related to cholesterol metabolism, oxidative phosphorylation, lysosome and so on. The non-foamy macrophages expressed more genes related to immune response (IL-1b related pathways, TNF, TLR signalling pathways) than foamy macrophages.
- Overall design:** Comparison of transcriptional profiles from two cell types: Intimal foamy and non-foamy macrophages in atherosclerotic ApoE-/- mice (n=6, WD for 28

- press Run SRA Selector link at the bottom of the page.

This SubSeries is part of SuperSeries:

[GSE116271](#) Transcriptome analysis reveals non-foamy rather than foamy plaque macrophages are pro-inflammatory in atherosclerotic murine models

**Relations**

BioProject [PRJNA477844](#)  
SRA [SRP151293](#)

**Download family**

[SOFT formatted family file\(s\)](#)  
[MINIML formatted family file\(s\)](#)  
[Series Matrix File\(s\)](#)

**Format**

[SOFT](#) [?](#)  
[MINIML](#) [?](#)  
[TXT](#) [?](#)

Supplementary file	Size	Download	File type/resource
<a href="#">GSE116239_RAW.tar</a>	19.6 Mb	<a href="#">(http)(custom)</a>	TAR (of TXT)
<a href="#">SRA Run Selector</a> <a href="#">?</a>			

Raw data are available in SRA  
Processed data provided as supplementary file

[NLN](#) | [NIH](#) | [GEO Help](#) | [Disclaimer](#) | [Accessibility](#)

- On the newly opened page press accession list button

NCBI SRA Run Selector

Accession

**Common Fields**

Center Name	GEO
DATASTORE filetype	FASTQ,SRA
DATASTORE provider	GS,NCBI,S3
DATASTORE region	gs.US,ncbl.public,s3.us-east-1
Instrument	Illumina HiSeq 3000
LibraryLayout	SINGLE
LibrarySelection	CDNA
LibrarySource	TRANSCRIPTOMIC
mouse_genotype	ApoE -/-
Organism	Mus musculus

**Select**

	Runs	Bytes	Bases	Download
Total	6	3.46 Gb	9.06 G	<a href="#">Metadata</a> or <a href="#">Accession List</a>
	0	0	0	<a href="#">Metadata</a> or <a href="#">Accession List</a> or <a href="#">JWT Cart</a>

- These are IDs for your datasets

SRR\_Acc\_List - Notepad

File Edit Format View Help

SRR7425017  
SRR7425018  
SRR7425019  
SRR7425020  
SRR7425021  
SRR7425022

You might not be familiar with any of the tools needed for analysis of RNA-seq data, but fear not. You might use file **pipeline.bash** which has all of the commands required to run the homework. You might use this file as a reference on how to run any individual tool, or you might use right away as a whole. You might also write your own pipeline, if you are willing to. This pipeline was tested with 4cpus and 8gb RAM, seems to work.

If you want to use this pipeline.bash file, please make sure that you change USERNAME (it appears several times in the script) to your actual username on the cluster (pipeline will resolve absolute paths to the files).

Once you have those IDs you can put them in the file called **pipeline.bash** after TAG=

```
1 #!/bin/bash
2
3 ##### Step 1: preparation
4 # starting with only one sample
5 # downloading data from SRA
6
7 TAG=SRR8193349 ← here
```

Once you put your ID, you can save the pipeline.bash and in terminal run  
**bash pipeline.bash**

This will take some time to download and analyze the sample  
You have to do this with all the samples.

After you think, you have processed the samples, contact me, and we will double-check that by looking at your QC report.

Once you processed all the samples, proceed to DESeq2 analysis:

1. Combine counts into count matrix
2. Run VST for PCA, run PCA
3. Run differential expression
4. Find pathways for DE