*Chair of Experimental Bioinformatics*
*TUM School of Life Sciences Weihenstephan*

**Lecture: Advanced Bioinformatics**
**SoSe 2022**

# Exercise Sheet 04 - RNA-Seq Pseudo alignment

Total: 15.0 points 31.05.22 - 09.06.22

## 1. Measures of gene expression (2.0 points)

a) Find a formula that allows you to compute gene-level (i.e. not transcripts) TPM values from FPKM values (1.0).

Formula: $TPM_i = \left(\frac{FPKM_i}{\sum_j FPKM_j}\right) \cdot 10^6$

b) Why are TPM and FPKM considered within-sample measures of gene expression? Name two between-sample biases they do not capture. (1.0)

1. Not normalized: if one gene is very highly expressed, it increases the mean expression https://btep.ccr.cancer.gov/question/fa/what-is-the-difference-between-rpkm-fpkm-and-tpm/

2. Different read depth, or different sequencing platforms

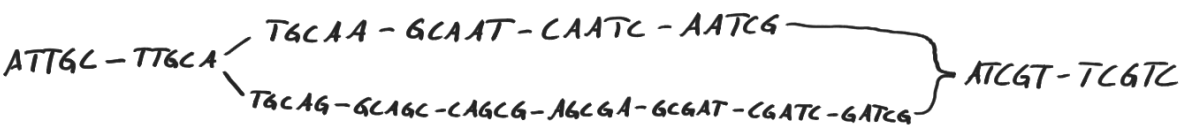
## 2. Pseudoalignment with Kallisto (6.0 points)

Given

```
Genome:        ATTGCATGCAAATTGCGATCGTC
transcript t1: ATTGCA-----------ATCGTC
transcript t2: ATTGCA -------GCGATCGTC

read 1:        ATTGCA
read 2:           GCA--------GCG
read 3:           GCA--------ATC
read 4:                       GATCGT
read 5:                        ATCGTC
```

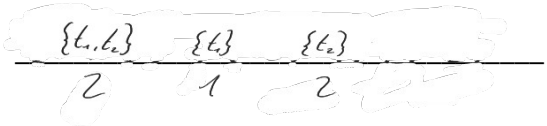
a) construct the transcriptome De-Bruijn graph with k-mers of size k = 5 (2.0)




b) which read is compatible with which transcript? (1.0)

| | t1 | t2 |
|---|---|---|
| r1 | 1 | 1 |
| r2 | 0 | 1 |
| r3 | 1 | 0 |
| r4 | 0 | 1 |
| r5 | 1 | 1 |

c) construct the equivalence classes and compute equivalence counts per class (1.0)




d) Hands-on: Use kallisto to build the kallisto-index and quantify transcript counts (2.0)

You will have to install kallisto first, one easy way is to use conda. In your command line type the following commands:

```
conda create -n kallisto -c bioconda kallisto
conda activate kallisto
kallisto -h
```

Now use kallisto to build the index for the GENCODE.v26 human reference transcriptome (find it in the LRZ sync+share link below).
Find the kallisto manual here: https://pachterlab.github.io/kallisto/manual

→ Use the `kallisto index` command with the correct parameters to build the index.

Next, you will quantify the transcript counts for a given set of sequencing samples. There are six samples obtained from a HepG2 liver cell line, with 2 x 3 replicates, namely control (CT) and overexpression (OE). You can find the zipped fastq files in a LRZ Sync+share folder here:

https://syncandshare.lrz.de/getlink/fi17ShyQLkRUhjpfGx42nhuc/ (Task2) (You will have to unpack these files first).

→ Use the `kallisto quant` command with the correct parameters to provide you with output files, which we will process in the following task 3.
Go through the final output files and **give the percentage of pseudo-aligned reads** for each of the 6 samples.

**Note:**
You will have to run kallisto once for each sample. With 6 given samples, this results in 6 separate kallisto runs; ideally you will store each result in its own folder named by the sample ID (HepG2_1_563, …). As no additional information regarding the experiment were provided, we assumed that the reads were single end and used the parameters -l=200 and -s=20 (as was done in the example from the linked manual)

1_563: 88.8%, 2_564: 91.2%

3_565: 90.9%, 4_566: 91.2%

5_567: 90.6%, 6_568: 91.8%

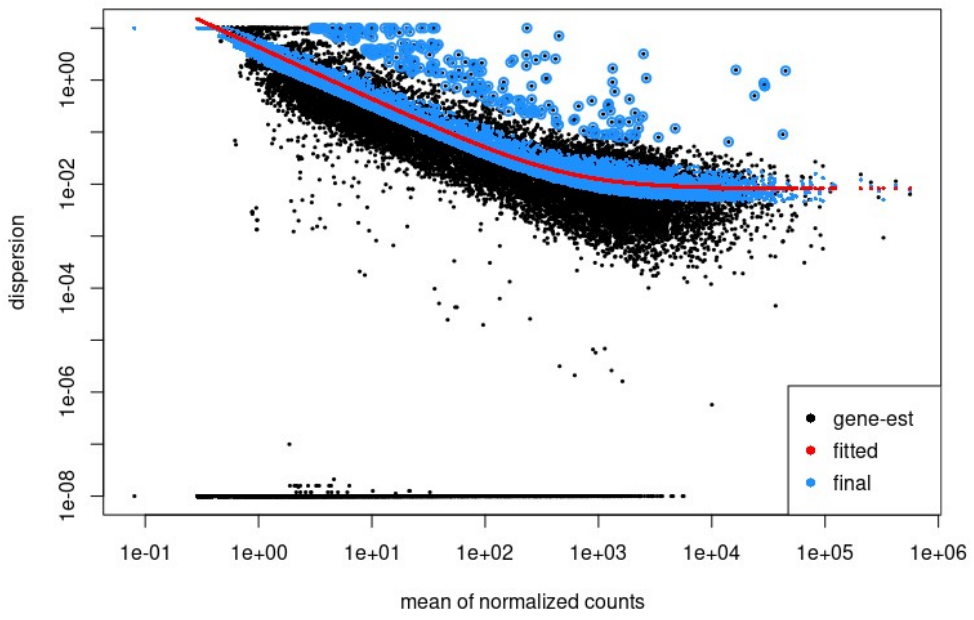
## 3. Differential Expression Analysis (4.0 points)

In the Moodle you will find an R script called DESeq2.R. Open it in a suitable editor such as RStudio.

We will use this script to analyze the transcript counts we generated in the previous task (2d).
You will notice the code is currently incomplete. It currently just reads the abundance files from kallisto and creates the DDS object used by DESeq2. Complete the code and run it such that you can answer the following questions. Use online tutorials for help, e.g.
http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

a) Create the dispersion plot (plotDispEsts(dds)) and give your opinion on the quality of the replicates (low shrinkage - high agreement between replicates). (2.0)

Plot:



Opinion on quality:

_____

_____

b) What is the gene with the highest log-fold change between CT and OE? (1.0) (Assuming no lfc Shrink)

Gene symbol: FSK (lfc = 22.29)

Ensembl gene id: ENSG00000202198

c) What type of gene is it? What is one of its known functions? (1.0)

Type: misc RNA (genecards)

Function: regulates transcription by controlling the positive transcription elongation factor P-TEFb (wikipedia)

# 4. Alternative Splicing (3.0 points)

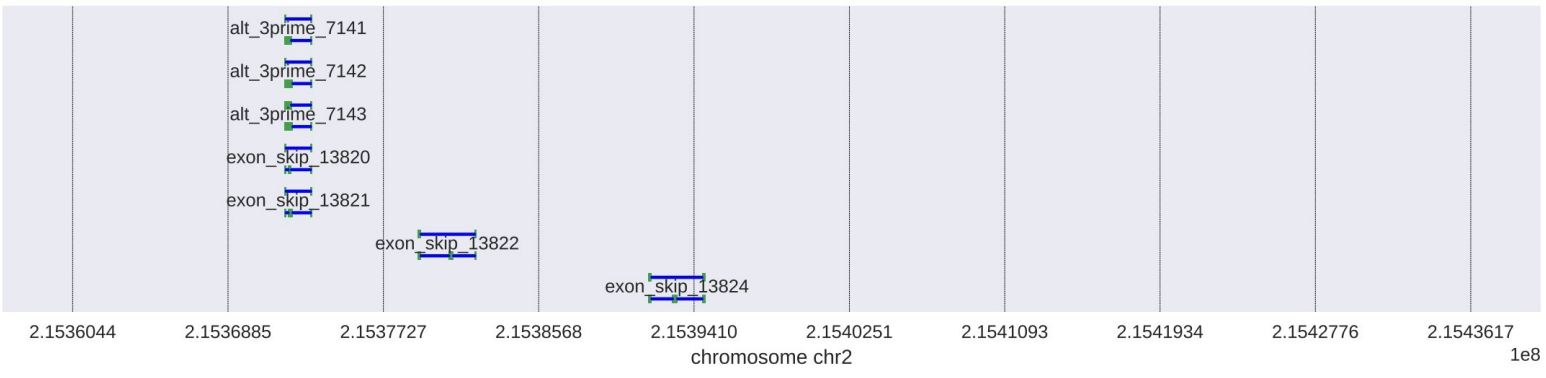
In the Sync+Share folder you will find the compressed output of the tool SplAdder. This bioinformatic tool can detect alternative splicing events using BAM files and a genome annotation file in gtf format. We ran SplAdder on the first sample of the above dataset (HepG2_1_563) and stored all of its outputs in the mentioned folder, as it takes quite some time to finish.
Find the outputs here (Task4): https://syncandshare.lrz.de/getlink/fi17ShyQLkRUhjpfGx42nhuc/

Install SplAdder and visualize all alternative splicing **events** for the gene ENSG00000115414.18; you will have to use the `spladder viz` command with the correct parameters for this.

Installation manual: https://spladder.readthedocs.io/en/latest/installation.html
spladder viz manual: https://spladder.readthedocs.io/en/latest/spladder_modes.html#the-viz-mode

a) Show the generated plot and give all alternative splicing event types (skipped exon, retained intron,...) for this gene along with the number of their occurrences. (1.0)

Plot:



AS events:

alt_3prime : 3

exon_skip : 4

b) Give the PSI values for all exon skipping events for this gene. Hint: check out the exon_skip_C3.confirmed.txt file. (1.0)

13820 : 0.73

13821 : 0.95

13822 : 0.89

13824 : 0.14

c) One exon skipping event has a significantly lower PSI value compared to the other two; explain what this means for this specific exon skipping event. (1.0)