BIO-463
Genomics and bioinformatics

Lecture 7:
Bulk gene expression analysis

Dr Raphaëlle Luisier

EPFL

Table of Content

Preliminary Information

Overview of the four modules

Module I

- Bulk RNA-sequencing I
- Unsupervised clustering analysis I

Module II

- Bulk RNA-sequencing II
- Unsupervised clustering analysis II

Module III

- Single-cell RNA-seq I
- Unsupervised clustering analysis III

Module IV

- Single-cell RNA-seq II
- Differential gene expression analysis

Preliminary Information

Objectives of the course

1. Extract knowledge from messy and noisy ata.
2. Understand the analysis, the existing tools and the publicly available data-bases.
3. Get practical hands on statistics and machine learning.
4. Develop a critical view on the results.
5. Stimulate the engineers to develop better or novel technologies.

Preliminary Information

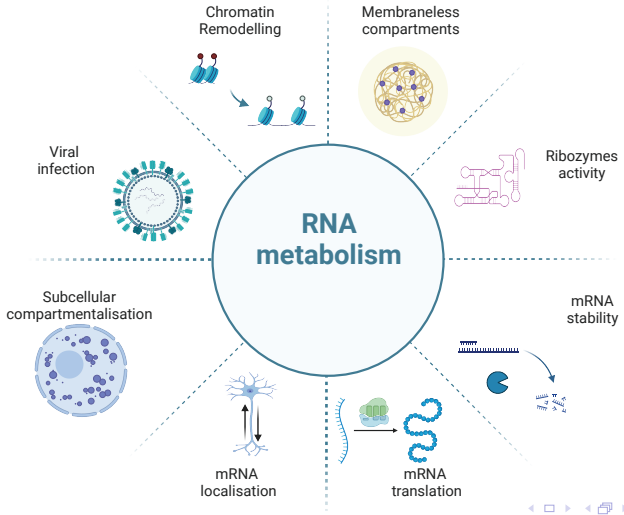
Structures and evaluations

- First period: biological background.
- Second period: machine learning.
- Exercises: R with more than what you need.
- Evaluation:
  - Evaluate your knowledge and critical thinking on a new data-set.
  - Adapt source code that has been done during exercises.
  - Upload your HTML file including source code.
  - Criteria: correct answer (50%), quality of the figure (25%), description of the results (25%).

Table of Content

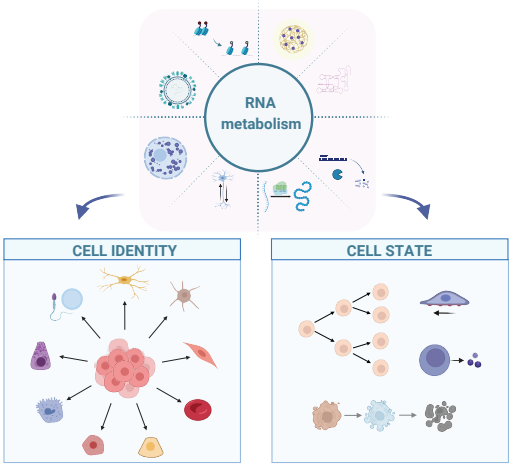
RNA Molecules

Central to all biological processes



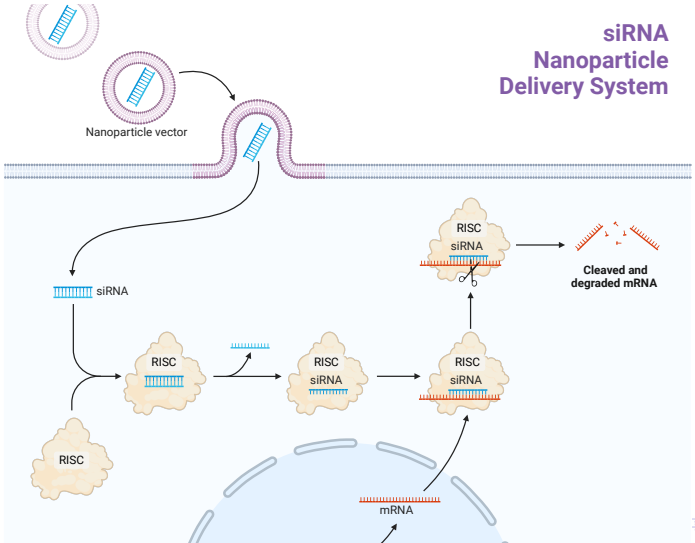
RNA Molecules

Underlying all cell behaviours and identities



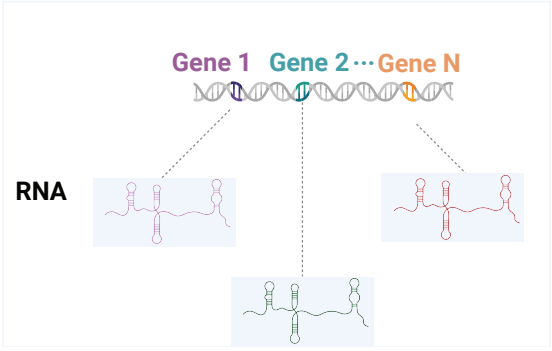
RNA Molecules

Strong therapeutics potential



RNA Molecules

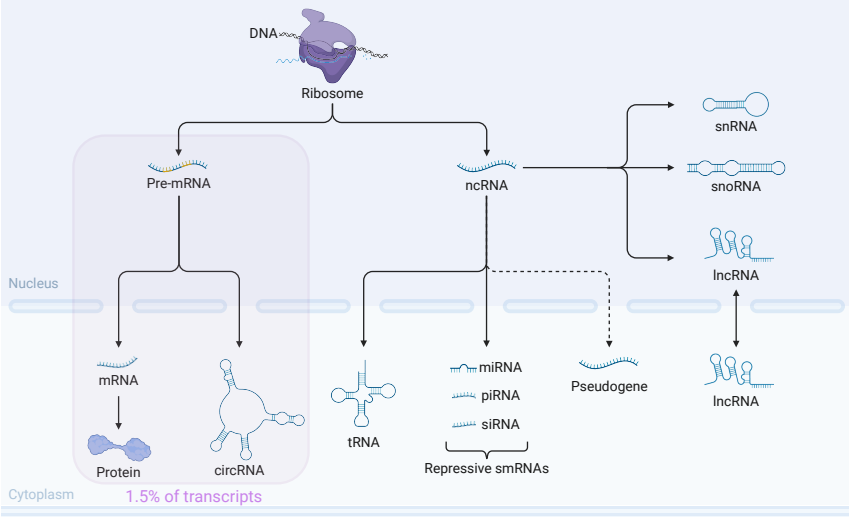
Copies of DNA segments composed of 4 components



Bases (adenine, cytosine, guanine, thymine)

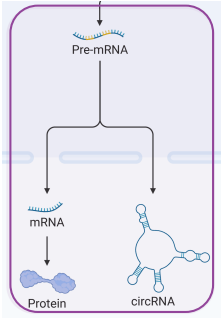
RNA Molecules

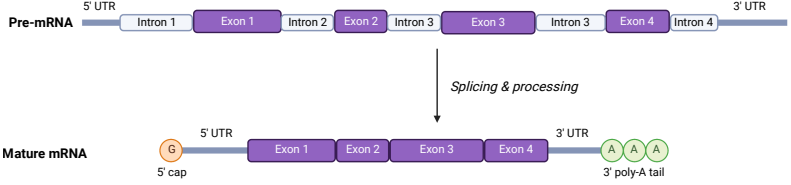
Only a small fraction encode for proteins



Versatile RNA Molecules

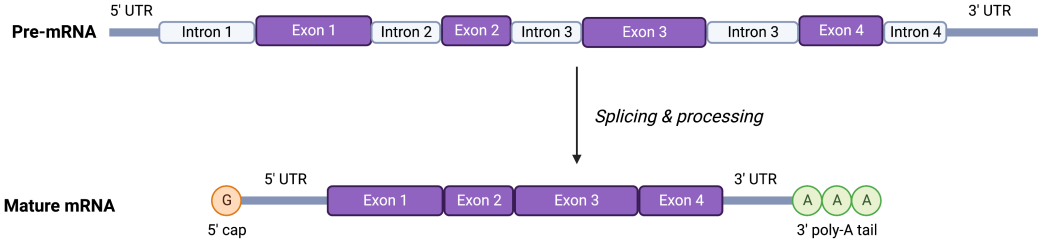
$>$ 95% *of RNA molecules are non-coding*





Versatile RNA Molecules

$>$ 95% *of RNA molecules are non-coding*

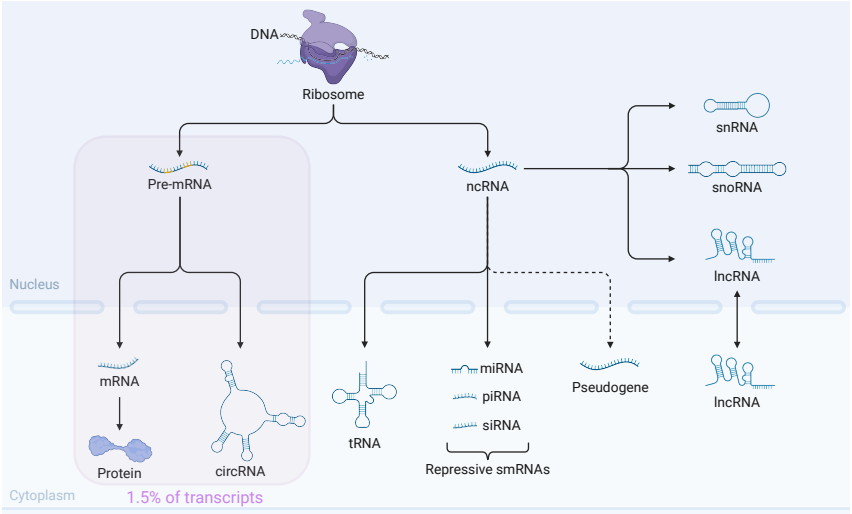


↓

Function beyond their own protein product?

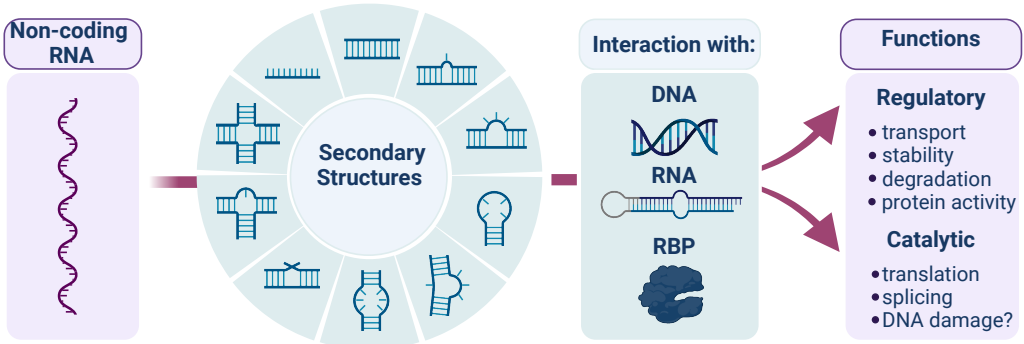
RNA Molecules

Only a small fraction encode for proteins



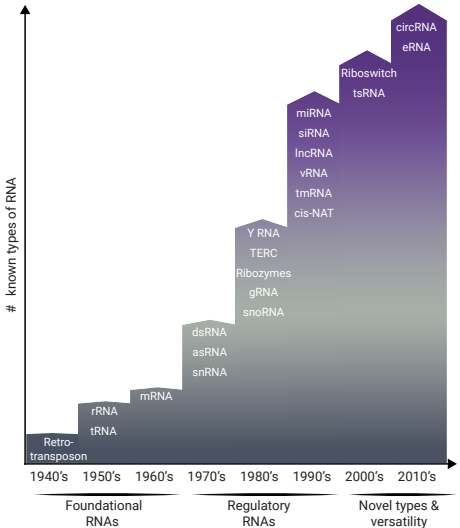
Canonical knowledge

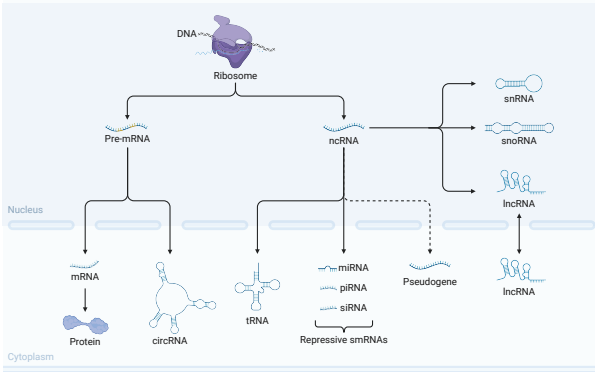
A structured and interacting multifunctional molecule



Three phases of discoveries in RNA

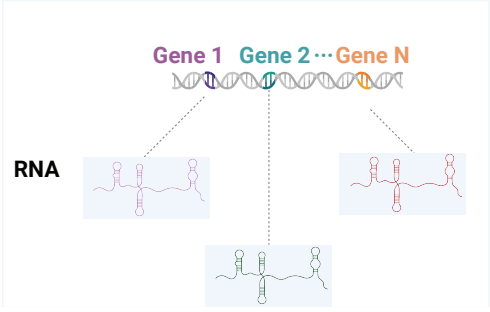
Decode RNA versatility with AI





RNA Molecules

Ubiquitous, versatile with strong therapeutic potential



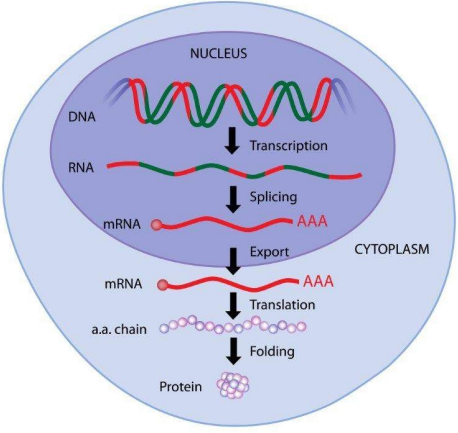
Bases (adenine, cytosine, guanine, thymine)

- Central to all biological process
- Implication in most human disorders
- Strong therapeutics potential (ASO for SMA)

Table of Content

RNA-sequencing data

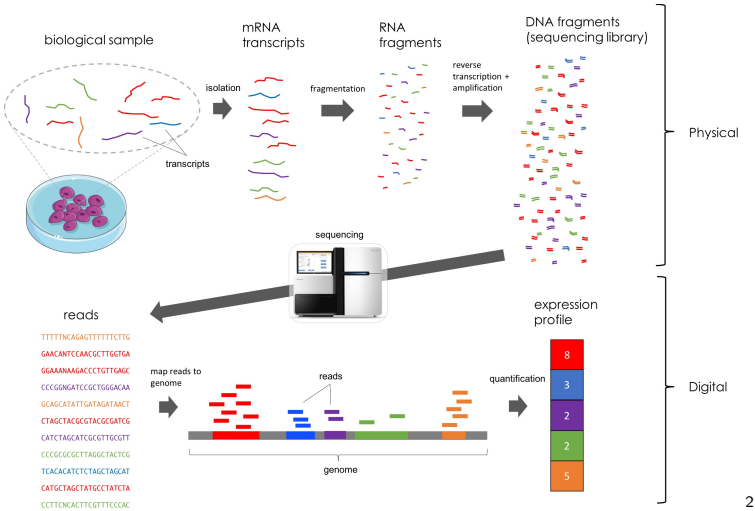
What are we looking at?



1

[1] Image fromhttps://chem.libretexts.org/ (CC BY 3.0).

RNA-sequencing



2

[2] https://mbernste.github.io/posts/rna_seq_basics/

RNA-sequencing

Information retrieved from such analysis

Qualitative - what type of molecule

- Identify expressed genes, transcript isoforms
- Identify transcript start and end boundaries

Quantitative - how much of each molecule

- Relative amount of mRNA produced in a cell, tissue, spatial location
- Enable sample clustering, differential gene expression analysis, differential splicing analysis, etc.

RNA-sequencing

High-throughput measurements of RNA and proteins.

Bulk RNA-sequencing

Can resolve the full complexity of the transcriptome, yet confound changes due to gene regulation with those due to shifts in cell type composition (Simpson's Paradox, 1951).

Single-cell RNA-sequencing

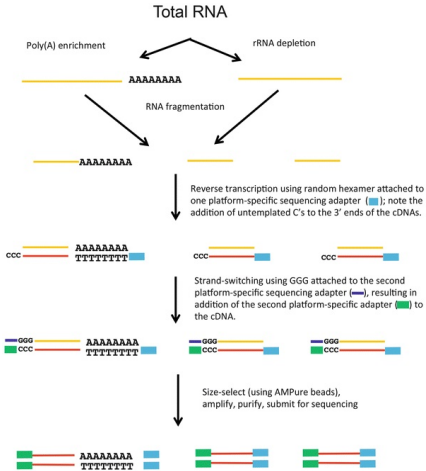
Resolution at the single-cell level however gene coverage is not even and therefore challenging to study gene structure.

Spatial RNA-sequencing

Mapping between location, form and gene expression. About 10 cells per spot. Not suitable for AS and APA analysis.

RNA-sequencing

High-throughput measurements of RNA and proteins.



- Different types of library preparation
- 1) ribosomal versus polA RNA enrichment
- 2) stranded versus non-stranded
- 3) paired-end versus single-end reads.

RNA-sequencing

Overview of a standard bioinformatic pipeline

1. Quality control (QC) of the fastq files (FASTQC)
2. Mapping of the reads onto a reference genome and/or transcriptome to obtain a count data table.
3. QC of the library
4. Pre-processing (statistical modelling)
5. Down-stream analysis
   - Unsupervised analysis (clustering,)
   - Supervised analysis (DGE, pathway analysis)

RNA-sequencing

FASTQ file store the reads

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Field 1 Sequence identifier

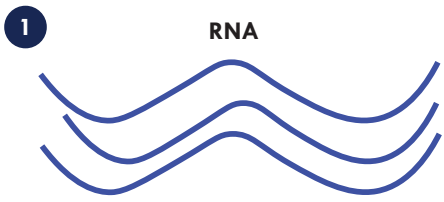
Field 2 Raw sequence letters

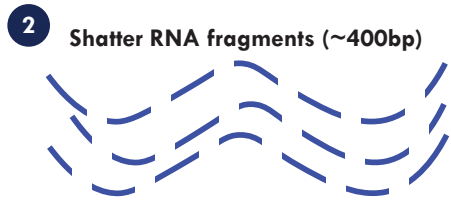
Field 3 + optionally followed by the same sequence identifier and any description

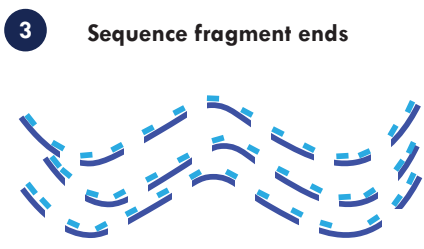
Field 4 Quality values for the sequence

RNA-sequencing

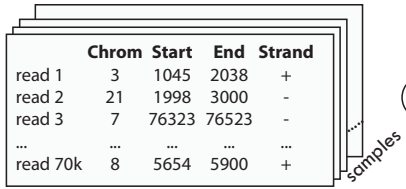
Relative quantification of mRNA content





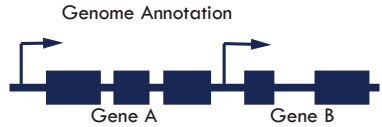


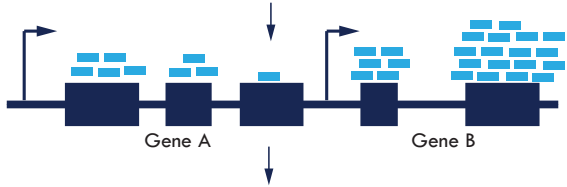
From the reads to gene expression count matrix



	Chrom	Start	End	Strand
read 1	3	1045	2038	+
read 2	21	1998	3000	-
read 3	7	76323	76523	-
...	...	...	...	...
read 70k	8	5654	5900	+

samples

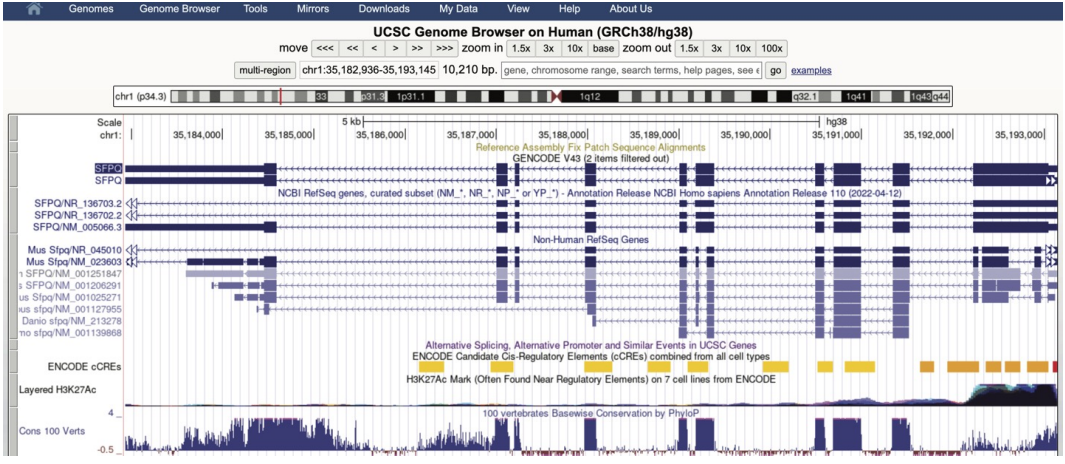




gene count matrix

	sample 1	sample 2		sample 30
gene 1	324533	3454		98
gene 2	33	6345		34532
gene 3	34555	966	...	62
...	...	...		...
gene 40k	3243	34		7544

UCSC Genome Browser



▸ UCSC

RNA-sequencing

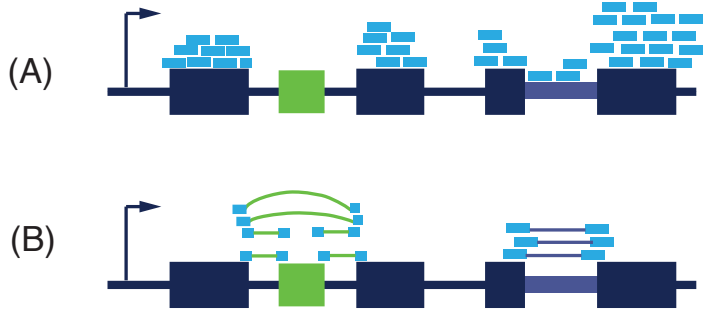
Mapping of the reads

(A)

(A) Reads mapped to genome.

RNA-sequencing

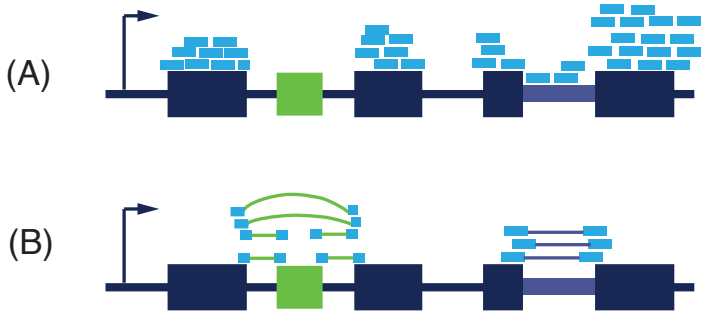
Mapping of the reads



(A) Reads mapped to genome.
(B) Spliced reads mapped to transcriptome and genome

RNA-sequencing

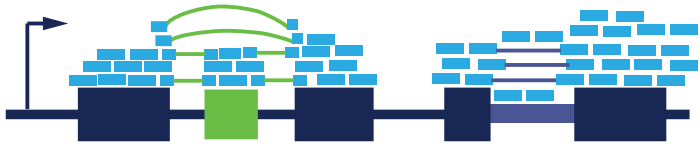
Mapping of the reads



Use of splice-aware alignment tools to resolve isoform complexity.

RNA-sequencing

Mapping of the reads



- Mature RNAs (mRNA) are spliced (without introns)
- Key info in the reads spanning the splice junctions)
- Mapping of the raw reads (fastq files) with an aligner that uses both a reference genome (fasta) and gene structure (gtf) information.

RNA-sequencing

Splice aware alignment tools

Open-source splice-aware aligners

MapSplice[3], HISAT2[4], STAR[5]

More info related to different aligners can be found HERE

Required files

- A fastq file
- A reference genome in fasta format which can be obtained from UCSC HG38 reference genome
- The gene structure or genome annotation as obtained from UCSC, Ensembl and NCBI HG38 Gencode
- Make sure that the annotation file (GTF) is exactly matched with the genome file (fasta)!

[3] Wang et al. 2010.
[4] Kim et al. 2019.
[5] Dobin et al. 2013.

RNA-sequencing

CookBook: ▸ *BAM/SAM files*

```
K00102:523:HK7FLBBXY:7:1228:16133:47032 355     chr21   5106425 3       101M    =       5106635 311     GACTGGCTGCTGTTTGTTTTTCTTGTGGCTGCAGTGGTTTTCTAAGAAGGTAACTCA
GTAAAAATAGTTACCATTCATGCCACACCTGC     AAFFFJJJJJFJFFJJ-AJJJJJJJJJJJJJJJJJF<JJJJJJJJJJJAFFJJJJJJJJJJJJJJJJJJJJFJJJJJJFJJJJJJJJJJJFJJJ   NH:i:2  HI:i:2  AS:i:
K00102:523:HK7FLBBXY:7:1201:29366:15170 403     chr21   5106558 3       101M    =       5106367 -292    CTTACACATGTGCACAGATACACATATCACACATGACTATGTGCAGACGTGTGCACA
CCCACAGGTGCGCTGCACACCCGCACACACAC     FFJJJF<JAF<JJJJJJJJJF<JAAJJJFFFJ<FJJJJJJJJJAJ7AJJJJJ<-7<JAFJJJJFJFJFJF-AJFJJJJJ<AF<JJJFJJJJJFFFFFAA   NH:i:2  HI:i:2  AS:i:
```

Read name — Alignement information: chromosome --- start coordinate --- Quality --- CIGAR

Col	Field	Type	Brief description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR string
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

CIGAR Code	BAM Integer	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

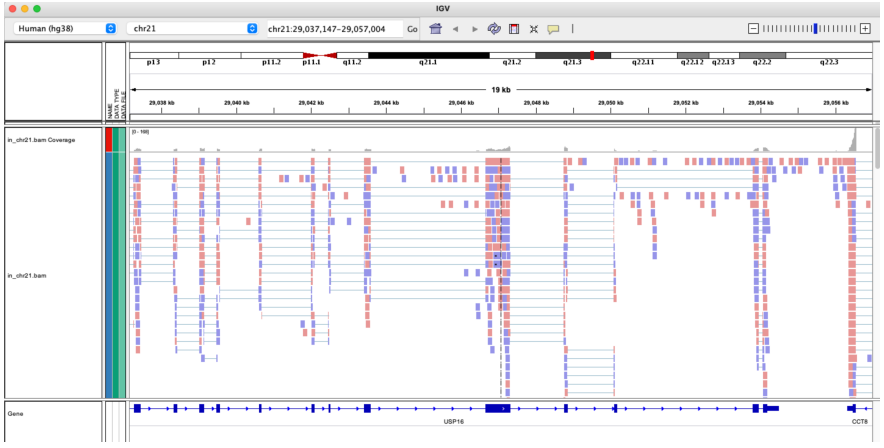
BAM file: compressed binary version of a SAM file

SAM file: represent aligned sequences

- Header section followed by Alignement section
- Alignment section have 11 mandatory fields

RNA-sequencing

IGV View of the mapped results



Identify AS events from RNA-seq data

CookBook:Example spliced reads

1. Open IGV
2. Load *in_chr*21.*bam*
3. Go to *chr*21 : 29013935 – 29053651
4. Sort alignements by start location
5. Group alignements by read strand

Use of IGV

↓

PRACTICE

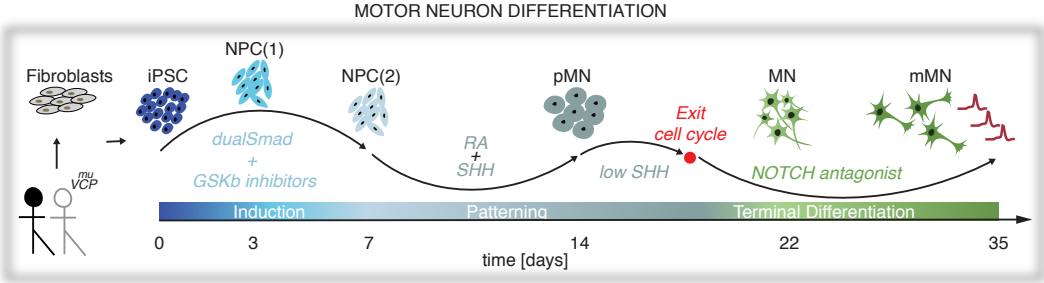
RNA-sequencing

QC of the library I

1. Strandedness: check fraction of reads mapping to correct strand
2. Relative fraction of reads mapping to intergenic regions
3. Relative fraction of reads mapping to each chromosome
4. RIN scores
5. 5'-3' coverage biases
6. Ribosomal contamination

Bulk RNA-sequencing Analysis of Differentiating Motor Neurons

Data-set for practical session



6

[6] Howe et al. 2022.

RNA-sequencing

Pre-processing of the read count table

1. Log-transformation
2. Filter-out lowly expressed genes
3. Normalisation

Pre-processing of the read count table

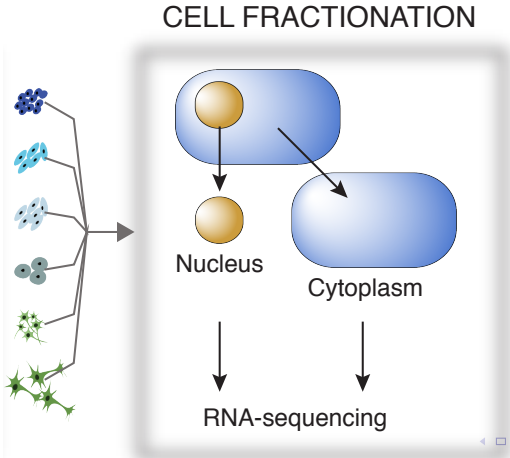
1. Log-transformation

Motivation

- To stabilize the variance which is quadratic in the mean.
- To converts multiplicative relative changes to additive differences.
- To get the sampled data in line with the assumptions of parametric statistics: the residuals from a model fit are normally distributed with a homogeneous variance.
- To deal with outliers.

Bulk RNA-sequencing Analysis of Differentiating Motor Neurons

Data-set for practical session



Pre-processing of the read count table

2. Filter lowly expressed genes

Motivation

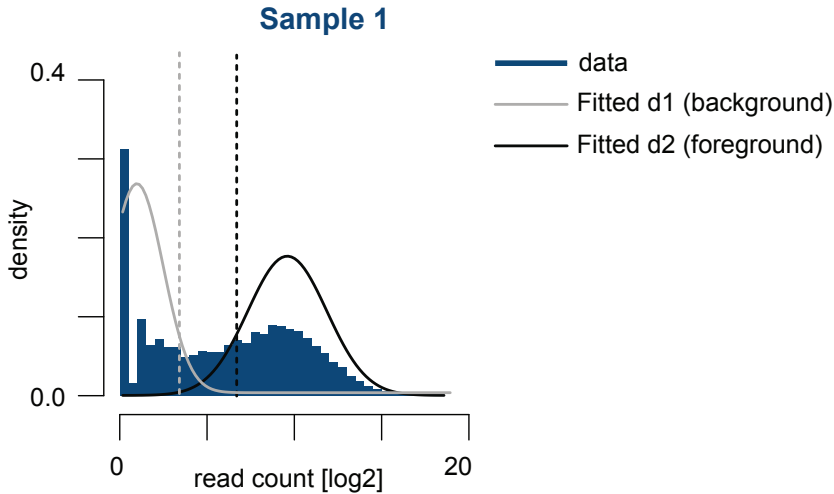
- Curse of dimensionality
- Sensitivity in differential gene expression analysis
- This value can be used for QC the data. .

Methods

- Based on fixed threshold (count per million)
- Select reliably expressed genes by fitting bimodal distribution

Pre-processing of the read count table

2. Filter lowly expressed genes



Pre-processing of the read count table

3. Normalisation: Why

Motivation

- Remove systematic technical artefacts:
  - Sequencing depth (total number of sequenced and mapped reads)
  - Library size
  - Gene length
  - Sequence composition due to PCR-amplification
- Essential for comparisons between samples and between genes.

RNA-sequencing Normalization

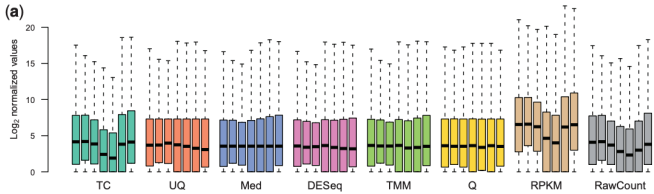
No unique solution.

BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 6. 671–683
Advance Access published on 17 September 2012
doi:10.1093/bib/bbs046

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

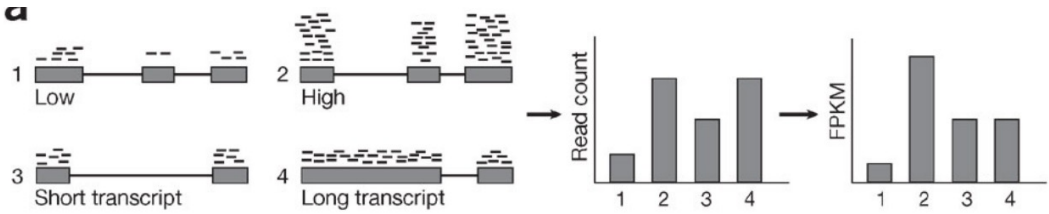
Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

Submitted: 12th April 2012; Received (in revised form): 29th June 2012



RNA-sequencing Normalization

Gene Length effect



RNA-sequencing Normalization

The Reads Per Kilobase per Million mapped reads (RPKM)

$$RPKM = \frac{x_{ij}}{N_j \times L_i} \times 10^9 \tag{1}$$

N_j = total number of reads sample j (in million)
L_i =gene length in kilobase x_ij=read count for a gene i in sample j

- Correct for gene length bias and sample to sample variation.
- To compare expression levels **between genes**.
- Some genes highly expressed may distord the signal (sink many reads)

RNA-sequencing Normalization

Library size might be biased by highly expressed genes

Upper quantile normalization (UQ)

$s_j = \frac{Q3_j}{\frac{1}{n}\sum_l Q3_l}$ with $Q3_j$ = upper quantile of sample j

Median Normalisation

$s_j = \frac{median_j}{\frac{1}{n}\sum_{\text{median } l}}$

Quantile Normalisation

Identical distribution of the read count across all samples.

Table of Content

Unsupervised clustering analysis

Hierarchical agglomerative clustering

Compute pair-wise distance between samples

- Flexible distance metrics between samples.
- Euclidean, correlation-based (Pearson or Spearman).

Agglomerative clustering of the samples

- Linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations.
- Linkage criterion influences shape of the clusters.
- The definition of shortest distance is what differentiates between the different agglomerative clustering methods.
- Complete-linkage tends to produce more spherical clusters than single-linkage.
- Single-linkage tends to produce long thin clusters in which nearby elements of the same cluster have small distances

Unsupervised clustering analysis

Hierarchical agglomerative clustering

Complete-linkage clustering

- Initially, each sample is in a cluster of its own.
- The clusters are then sequentially combined into larger clusters until all elements end up being in the same cluster.
- At each step, the two clusters separated by the shortest distance are combined.
- Shortest distance between clusters = the distance between the two samples **farthest** away from each other.
- $D(X, Y) = \max_{x \in X, y \in Y} d(x, y)$ where $d(x, y)$ is the distance between two elements of the clusters X and Y.

Unsupervised clustering analysis

Hierarchical agglomerative clustering

Single-linkage clustering

- Initially, each sample is in a cluster of its own.
- The clusters are then sequentially combined into larger clusters until all elements end up being in the same cluster.
- At each step, the two clusters separated by the shortest distance are combined.
- Shortest distance between clusters = the distance between the two samples **closest** away from each other.
- $D(X, Y) = \min_{x \in X, y \in Y} d(x, y)$ where $d(x, y)$ is the distance between two elements of the clusters X and Y.

Unsupervised clustering analysis

Hierarchical agglomerative clustering

UPGMA (unweighted pair group method with arithmetic mean)

- Initially, each sample is in a cluster of its own.
- The clusters are then sequentially combined into larger clusters until all elements end up being in the same cluster.
- At each step, the two clusters separated by the shortest distance are combined.
- Shortest distance between clusters = the average distance between pairs of object in X and Y.
- $D(X, Y) = \frac{1}{n_X n_y} \sum_{x \in X} \sum_{y \in Y} d(x, y)$

Unsupervised clustering analysis

Principal Component Analysis

- To transform a set of possibly correlated variables (genes) into "some more fundamental set of independent variables".
- To project a dataset from **many correlated** coordinates onto **fewer uncorrelated** coordinates
- The orthogonal coordinates are called principal components (PCs).
- PC retain most variability in the data.

Unsupervised clustering analysis

Principal Component Analysis

- To transform a set of possibly correlated variables (genes) into "some more fundamental set of independent variables".
- To project a dataset from **many correlated** coordinates onto **fewer uncorrelated** coordinates
- The orthogonal coordinates are called principal components (PCs).
- PC retain most variability in the data.

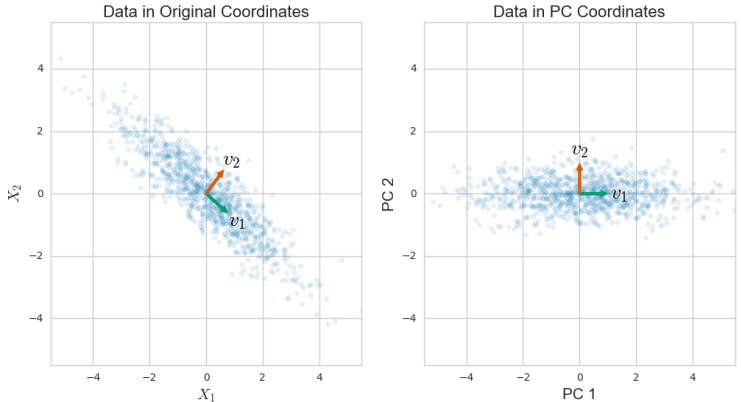
↓

- To reduce the dimensionality of data
- To extract essential information
- To characterize the structure of the data.

Principal Component Analysis

The goal of PCA is to find a collection of $k \leq d$ unit vectors $\vec{v}_i \in \mathbb{R}^n$ (for $i \in 1, \ldots, k$) called Principal Components, or PCs, such that

1. the variance of the dataset projected onto the direction determined by $\vec{v}_i$ is maximized
2. $\vec{v}_i$ is chosen to be orthogonal to $\vec{v}_1, \ldots, \vec{v}_{i-1}$



Principal Component Analysis

To find $\vec{v_1}$, the following conditions must be adressed:

1. $\|\mathbf{v}_1\| = 1$
2. The variance of $\mathbf{X}$ projected onto $\vec{v_1}$ must be maximised.

- The projection of a vector $\vec{x} \in \mathbb{R}^n$ onto $\vec{v_i}$ is $\vec{v_i}^T \vec{x}$
- The variance of $\mathbf{X}$ projected onto $\vec{v_1}$ is

$$\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{v}_1^T \mathbf{x_i} - \mathbf{v}_1^T \mu)^2 = \mathbf{v}_1^T \frac{1}{n-1}\mathbf{X}^T\mathbf{X}\mathbf{v}_1 \tag{2}$$

- Given a matrix $\mathbf{X}$, the covariance of the matrix around the mean can written as

$$Cov(\mathbf{X}) = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T \tag{3}$$

The solution to is therefore $Cov(\mathbf{X})\mathbf{v}_1 = \lambda_1\mathbf{v}_1 \Leftrightarrow \mathbf{v}_1^T Cov(\mathbf{X})\mathbf{v}_1 = \lambda_1$ i.e. $\mathbf{v}_1$ and λ_1 are an eigenvector and an eigenvalue respectively of $Cov(\mathbf{X})$.

Principal Component Analysis

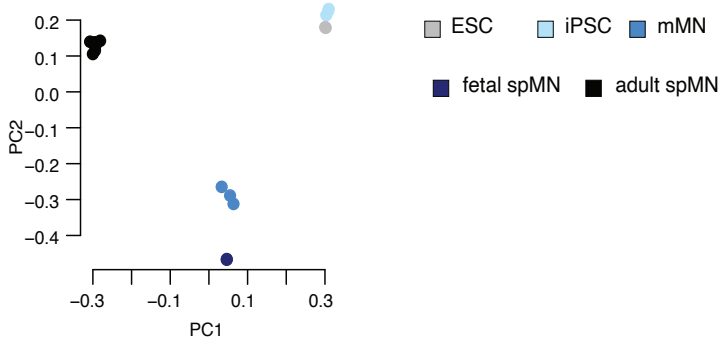
To find $\vec{v_1}$, the following conditions must be adressed:

1. $\|\mathbf{v}_1\| = 1$
2. The variance of $\mathbf{X}$ projected onto $\vec{v_1}$ must be maximised.

The solution to is therefore $Cov(\mathbf{X})\mathbf{v}_1 = \lambda_1\mathbf{v}_1 \Leftrightarrow \mathbf{v}_1^T Cov(\mathbf{X})\mathbf{v}_1 = \lambda_1$ i.e. $\mathbf{v}_1$ and λ_1 are an eigenvector and an eigenvalue respectively of $Cov(\mathbf{X})$.

$$Cov(\mathbf{X}) = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = V\Sigma V^T \tag{4}$$

Principal Component Analysis



- Most biologists focus on the clustering of the samples in PC1 and PC2.
- This is ok for simple experimental set-up (2 or 3 covariates).
- However what if more complex experiment?

↓

More subtle but biologically relevant signal might be captured in other components.

Bibliography I

- Dobin, Alexander et al. (2013). "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1, pp. 15–21.
- Howe, Marija Petrić et al. (2022). "Physiological intron retaining transcripts in the cytoplasm abound during human motor neurogenesis". In: *Genome Research* 32.10, pp. 1808–1825.
- Kim, Daehwan et al. (2019). "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype". In: *Nature biotechnology* 37.8, pp. 907–915.
- Wang, Kai et al. (2010). "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery". In: *Nucleic acids research* 38.18, e178–e178.