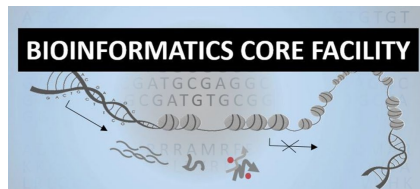


# Welcome to ABC.3

abc.au.dk



20. August 2024



Health  
Data Science  
Sandbox

# Agenda

- Introduction to bulk RNA-seq data analysis
- Workflow of analysis
- File structures
- Exercise with real bulk RNA-seq data – Alternatively use your own data.

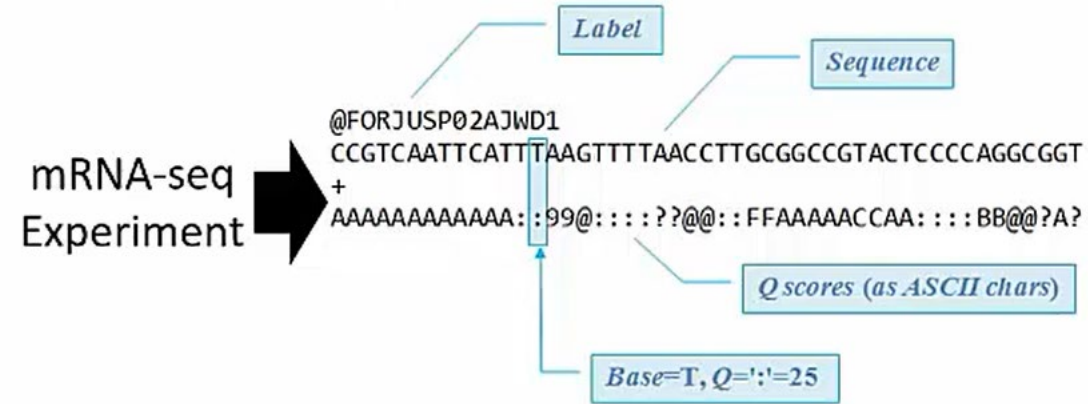
# For efficiency...

- We skip a few steps that require experience and knowledge of bash operations and work in terminals.
- Steps skipped:
  - Download of rawdata fastq files.
  - Download of reference genome.
  - Sequence alignment (mapping).
  - Quality control.
  - Count of features.

# Workflow: mRNA-seq Data Analysis

## FASTQ Data:

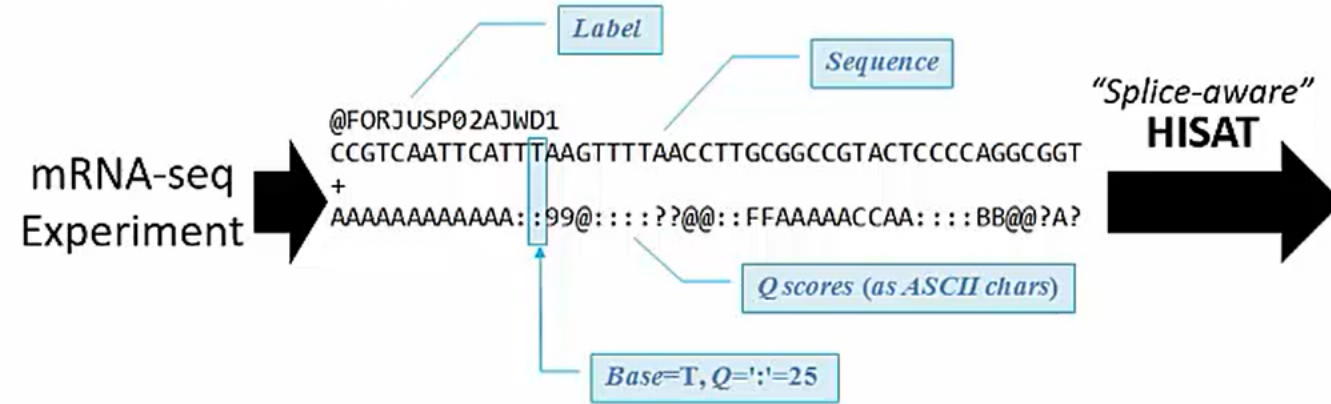
Individual reads & Quality scores  
Files usually contains 20-30M reads



# Workflow: mRNA-seq Data Analysis

## FASTQ Data:

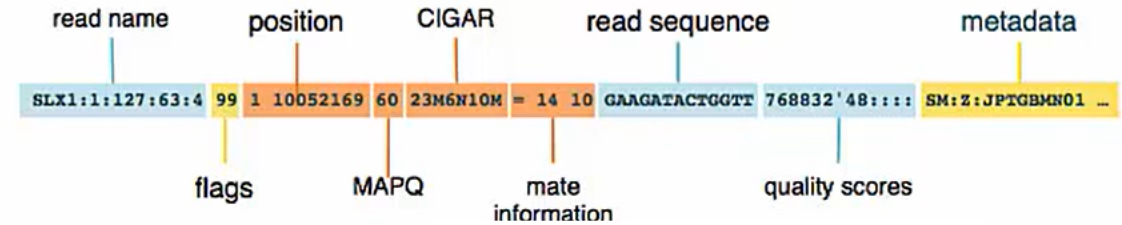
Individual reads & Quality scores  
Files usually contains 20-30M reads



## SAM/BAM:

Shows where each read aligns to a reference genome (e.g., hg38). SAM = Human readable, BAM = binary.

**HEADER** containing metadata (sequence dictionary, read group definitions etc)  
**RECORDS** containing structured read information (1 line per read record)

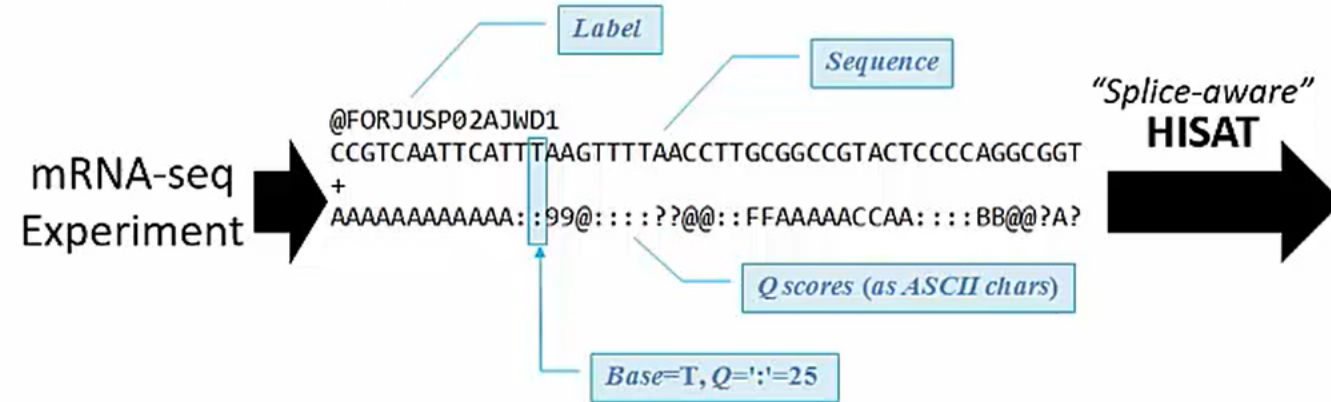


Format	Size_GB
SAM	7.4
BAM	1.9
CRAM lossless Q	1.4
CRAM 8 bins Q	0.8
CRAM no Q	0.26

# Workflow: mRNA-seq Data Analysis

## FASTQ Data:

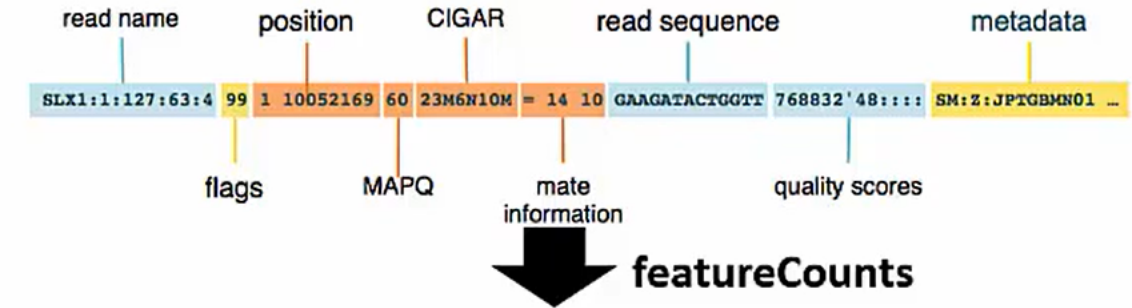
Individual reads & Quality scores  
Files usually contains 20-30M reads



## SAM/BAM:

Shows where each read aligns to a reference genome (e.g., hg38). SAM = Human readable, BAM = binary.

**HEADER** containing metadata (sequence dictionary, read group definitions etc)  
**RECORDS** containing structured read information (1 line per read record)



## Counts:

Lists of how many reads were detected per gene in each sample

Genelid	CON-1	CON-2	CON-3	TRAN-1	TRAN-2	TRAN-3
1	2	2	0	0	2	1
10	0	0	0	0	0	0
100	15	35	28	23	33	45
1000	223	231	170	278	270	277
10000	1	5	8	4	1	1
100008587	0	0	0	0	0	0



# Workflow: mRNA-seq Data Analysis

## Counts:

Lists of how many reads were detected per gene in each sample

Geneid	CON-1	CON-2	CON-3	TRAN-1	TRAN-2	TRAN-3
1	2	2	0	0	2	1
10	0	0	0	0	0	0
100	15	35	28	23	33	45
1000	223	231	170	278	270	277
10000	1	5	8	4	1	1
100008587	0	0	0	0	0	0

DESeq2

## Differentially Expressed Genes (DEGs)

Compare counts between types of samples to identify genes that were significantly up- or down-regulated.

geneid	BaseMean	log2FC	Std Error	Wald-stats	p	padj
4599	2881.924	-7.33044	0.198155	-36.9934	1.46E-299	1.84E-295
3434	2289.489	-5.70927	0.159381	-35.8216	5.09E-281	3.20E-277
3437	2011.254	-6.22573	0.1813	-34.33931	2.03E-258	8.51E-255
8638	1951.398	-6.58146	0.192536	-34.18301	4.32E-256	1.36E-252
4939	3073.582	-5.91225	0.175125	-33.76012	7.59E-250	1.91E-246
3433	3740.118	-5.53047	0.16426	-33.66909	1.64E-248	3.43E-245

For our purposes, we will define a DEG as a gene with:

- **Padj < 0.05** (statistically significant increase in exp.)
- **log2FC > 1** (at least 2-fold increase in expression)

# Workflow: mRNA-seq Data Analysis

## Counts:

Lists of how many reads were detected per gene in each sample

Geneid	CON-1	CON-2	CON-3	TRAN-1	TRAN-2	TRAN-3
1	2	2	0	0	2	1
10	0	0	0	0	0	0
100	15	35	28	23	33	45
1000	223	231	170	278	270	277
10000	1	5	8	4	1	1
100008587	0	0	0	0	0	0

Column Join  Excel

## Calculate Transcripts per Million (TPM)

Normalizes counts according to length of transcript and total number of reads per sample, thereby enabling us to make comparisons between genes and samples.

$$A_x = \frac{\text{total reads mapped to gene } X}{\text{length of gene } X \text{ transcript}}$$

$$TPM_x = \frac{A_x}{\sum A_{\text{all genes}}} \times 10^6$$

DESeq2 

## Differentially Expressed Genes (DEGs)

Compare counts between types of samples to identify genes that were significantly up- or down-regulated.

geneid	BaseMean	log2FC	Std Error	Wald-stats	p	padj
4599	2881.924	-7.33044	0.198155	-36.9934	1.46E-299	1.84E-295
3434	2289.489	-5.70927	0.159381	-35.8216	5.09E-281	3.20E-277
3437	2011.254	-6.22573	0.1813	-34.33931	2.03E-258	8.51E-255
8638	1951.398	-6.58146	0.192536	-34.18301	4.32E-256	1.36E-252
4939	3073.582	-5.91225	0.175125	-33.76012	7.59E-250	1.91E-246
3433	3740.118	-5.53047	0.16426	-33.66909	1.64E-248	3.43E-245

For our purposes, we will define a DEG as a gene with:

- **Padj < 0.05** (statistically significant increase in exp.)
- **log2FC > 1** (at least 2-fold increase in expression)

  
**Identify genes that are up/downregulated (DEGs) and expressed at a meaningful level (e.g., TPM > 10)**



# Structure of fastq files

[illegible]

# Structure of reference genome

```
>NC_000001.11 Homo sapiens chromosome 1, GRCh38.p14 Primary Assembly
ATGAAGAAGGTAAGTGCAGAGGCTATTCCTGGAATGAATCAACGAGTGAAACGAATAACTCTATGGTGACTGAATTCAT
TTTTCTGGGTCTCTCTGATTCTCAGGAAGTCCAGACCTTCCTATTTATGTTGTTTTTGTATTCTATGGAGGAATCGTGT
TTGGAAACCTTCTTATTGTCATAACAGTGGTATCTGACTCCACCTTCACTCTCCCATGACTTCCTGCTAGCCAACCTC
TCACTCATTGATCTGTCTGTCTTCAGTCACAGCCCCCAAGATGATTACTGACTTTTTAGCCAGCGCAAAGTCATCTC
TTTCAAGGGCTGCCCTTGTTCAGATATTTCTCCTTCACTTCTTTGGTGGGAGTGAGATGGTGATCCTCATAGCCATGGGCT
TTGACAGATATAGCAATATGCAAGCCCCCTACACTACACTACAATTATGTGTGGCAACGCATGTGTCGGCATTATGGCT
GTCACATGGGGAATTGGCTTTCTCCATTGCGTGAGCCAGTTGGCGTTTGCCGTGCACTTACTCTTCTGTGGTCCCAATGA
GGTCGATAGTTTTTATTGTGACCTTCCTAGGGTAATCAAACCTTGCCTGTACAGATACCTACAGGCTAGATATTATGGTCA
TTGCTAACAGTGGTGTGCTCACTGTGTGTTCTTTGTTCTTCTAATCATCTCATACACTATCATCCTAATGACCATCCAG
```

# Sequence alignment

```

      911      921      931      941      951      961      971      981      991      1001      1011      1021      1031      1041      1051      1061      1071
GTAGGTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTTCAAGTACCTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAATTGACTCTGT
GT   GTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAAC      ctgcttctgagattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttcc      ctctccattcaagacttaattgactctgt
GT   ATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC      tgcttctgagattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttcca      cctccattcaagacttaattgactctgt
GT   atttcatcttctaatttagaattcttgccaatcaagccctctcgaagttggcaatctataactcaac      GCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAATTGGTGTTATCGGGTCTTCCAA      cctccattcaagacttaattgactctgt
GT   atttcatcttctaatttagaattcttgccaatcaagccctctcgaagttggcaatctataactcaac      GCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAATTGGTGTTATCGGGTCTTCCAA      cctccattcaagacttaattgactctgt
GTAGGTTTAAT      aatcttgccaatcaagccctctcgaagttggcaatctataactcaacctctgcttctgagattcta      CTAGATGCCAAGTACATTACTATAAATTGGTGTTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAA      ctgt
GTAGGTTTAATTT      tcttgccaatcaagccctctcgaagttggcaatctataactcaacctctgcttctgagattctaag      CTAGATGCCAAGTACATTACTATAAATTGGTGTTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAA
GTAGGTTTAATTTTCATCTT      cttgccaatcaagccctctcgaagttggcaatctataactcaacctctgcttctgagattctaag      TTAGATGCCAAGTACATTACTATAAATTGGTGTTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAAT
GTAGGTTTAATTTTCATCTTC      TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC      ATGCCAAGTACATTACTATAAATTGGTGTTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAATTGAC
GTAGGTTTAATTTTCATCTCTAAT      TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC      GCCAAGTACATTACTATAAATTGGTGTTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAATTGACTC
gtaggtttaatttcatcttctaatttag      TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC      CATTACTATAAATTGGTGTTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAATTGACTCTGT
GTAGGTTTAATTTTCATCTCTAATTTAG      GCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACC      cattactataaattgggtttatcgggtcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAG      CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACC      tgttatcgggtcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAG      CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTT      ggggtcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAG      gccctctcgaagttggcaatctataactcaacctctgcttctgagattctaagtagctgcca      GGCTTCCAACCTCTCCATTCAAGACTTAATTGACTCTGT
GTAGGTTTAATTTTCATCTTCTAATTTAGAAT      CCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCA      ggtcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCT      ctctcgaagttggcaatctataactcaacctctgcttctgagattctaagtagctgccaag      ggtcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCT      CTGCAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTAC      GCTTCCAACCTCTCCATTCAAGACTTAATTGACTCTGT
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCT      CGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTAC      gcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCT      AAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATT      ctccaactcctccattcaagacttaattgactctgt
gtaggtttaatttcatcttctaatttagaattcttgcc      CAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAA      ctccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCA      CTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAATTGGTG      CTCCAACCTCTCCATTCAAGACTTAATTGACTCTGT
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAA      cttctgagattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaac      CTCCATTCAAGACTTAATTGACTCTGT
gtaggtttaatttcatcttctaatttagaattcttgccaatcaagcc      cttctgagattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaac      tccattcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCC      cttctgagattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaac      tccattcaagacttaattgactctgt
gtaggtttaatttcatcttctaatttagaattcttgccaatcaagccc      ttctgagattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaact      tccattcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCC      tgagattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaactcc      ccattcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTC      tgagattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaactcct      cattcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAG      tgagattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaactcct      tcaagacttaattgactctgt
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAG      gagattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaactcctc      AAGACTTAATTGACTCTGT
ATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC      agattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaactcctcc      ctttaattgactctgt
TTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCT      gattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaactcctcca      attgactctgt
gattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaactcctcca      gattctaagtagctgccaagtagacattactataaattgggtttatcgggtcttccaactcctcca      aagtagctgccaagtagacattactataaattgggtttatcgggtcttccaactcctcca
aagtagctgccaagtagacattactataaattgggtttatcgggtcttccaactcctccaattcaag      aagtagctgccaagtagacattactataaattgggtttatcgggtcttccaactcctccaattcaag
cttccaactcctccattcaagacttaattgactctgt      cttccaactcctccattcaagacttaattgactctgt
TTCCAACCTCTCCATTCAAGACTTAATTGACTCTGT      TTCCAACCTCTCCATTCAAGACTTAATTGACTCTGT
TCCAACCTCTCCATTCAAGACTTAATTGACTCTGT      caactcctccattcaagacttaattgactctgt
caactcctccattcaagacttaattgactctgt      caactcctccattcaagacttaattgactctgt
aactcctccattcaagacttaattgactctgt      aactcctccattcaagacttaattgactctgt
aactcctccattcaagacttaattgactctgt      tccattcaagacttaattgactctgt
ccattcaagacttaattgactctgt      ccattcaagacttaattgactctgt
ccattcaagacttaattgactctgt      ccattcaagacttaattgactctgt

```

# Exercise

- Download countmatrix from abc.au.dk
- Load the countmatrix into R
- Analyze the data using DESeq2  
<https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>
- Alternatively follow the instructions on  
<https://abc.au.dk/documentation/2024-08-20-ABC3.html>