ARISTOTLE UNIVERSITY OF THESSALONIKI

Introduction to RNA-seq Analysis: Basic Concepts in Gene Expression

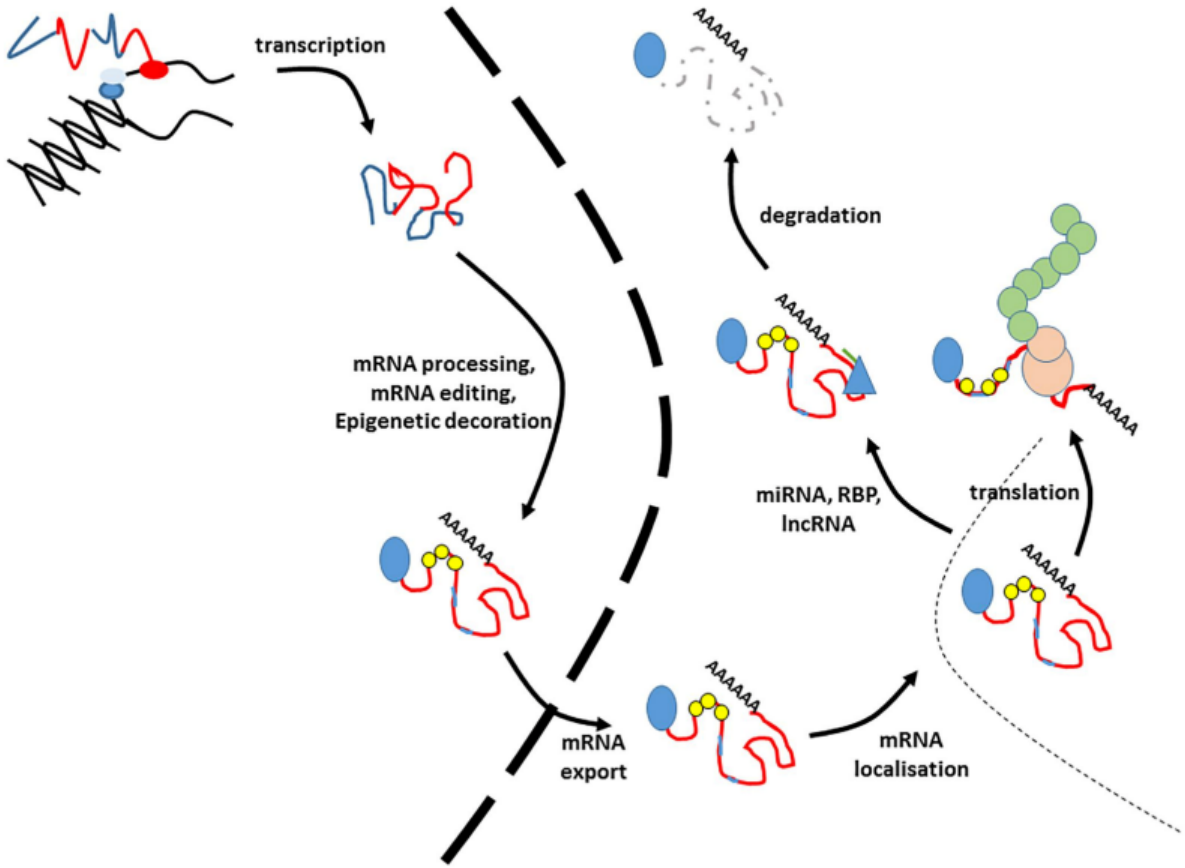
A Beginner-Friendly Introduction

Korina Karagianni

PhD Candidate in School of Biology,
Department of Genetics, Development and Molecular Biology

Supervisor:
Associate Professor Dimitra Dafou

The lifecycle of an RNA



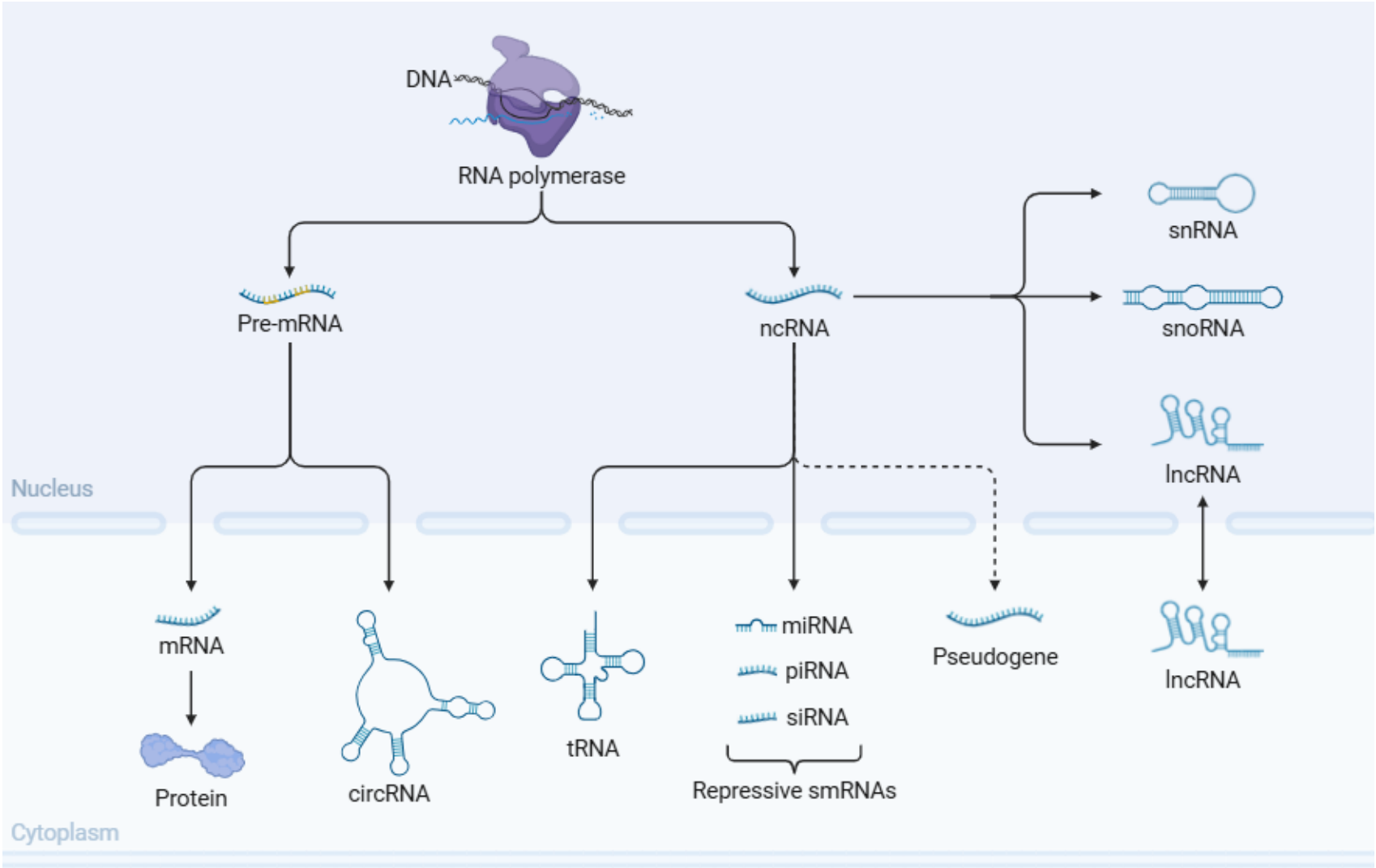
Basic Levels Complexity

Expression

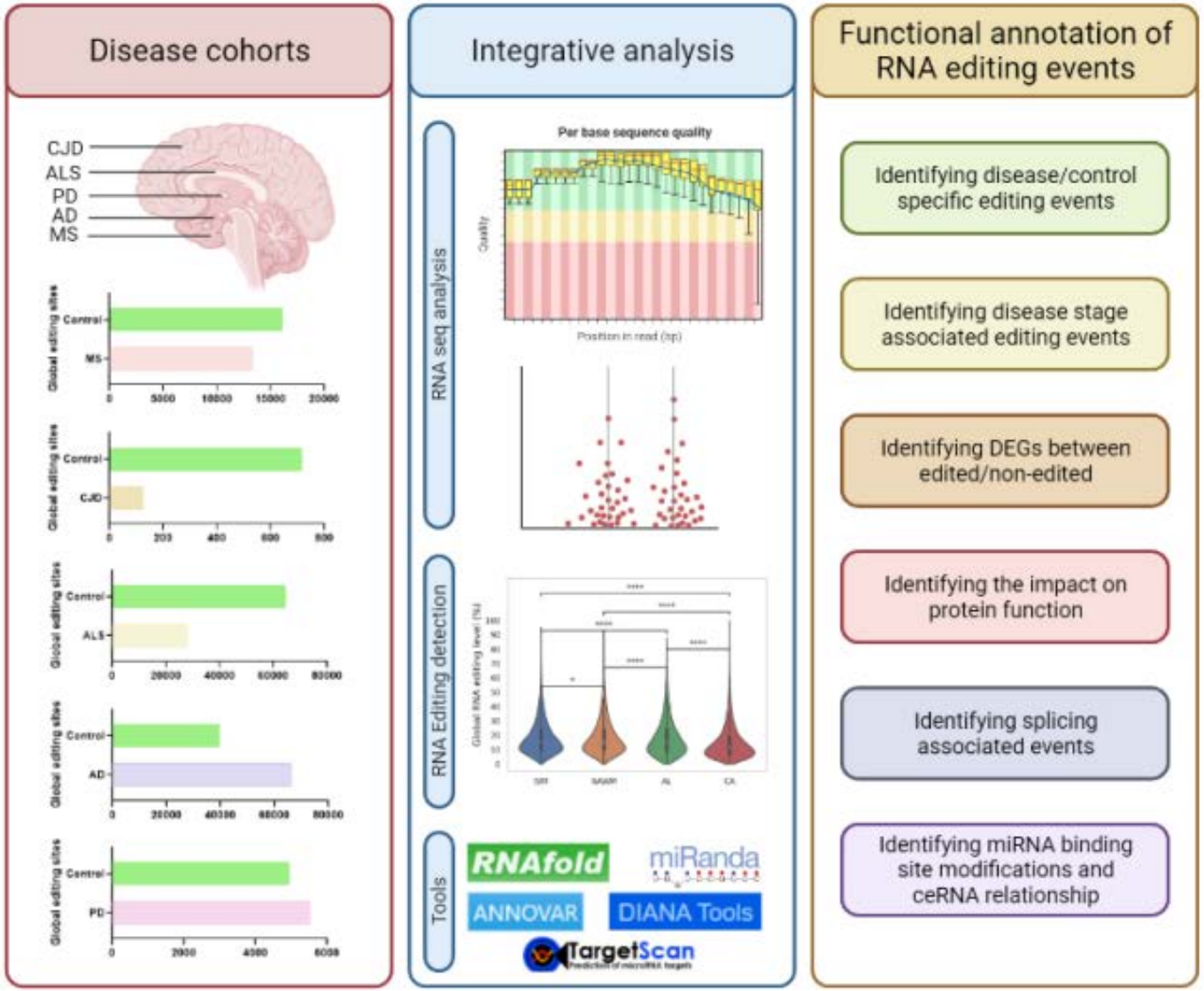
RNA Editing Patterns

Splicing

Types of Coding and non-coding RNA



Neurodegenerative Disease Research in Our Lab



Gene Expression

↕

RNA Editing Patterns

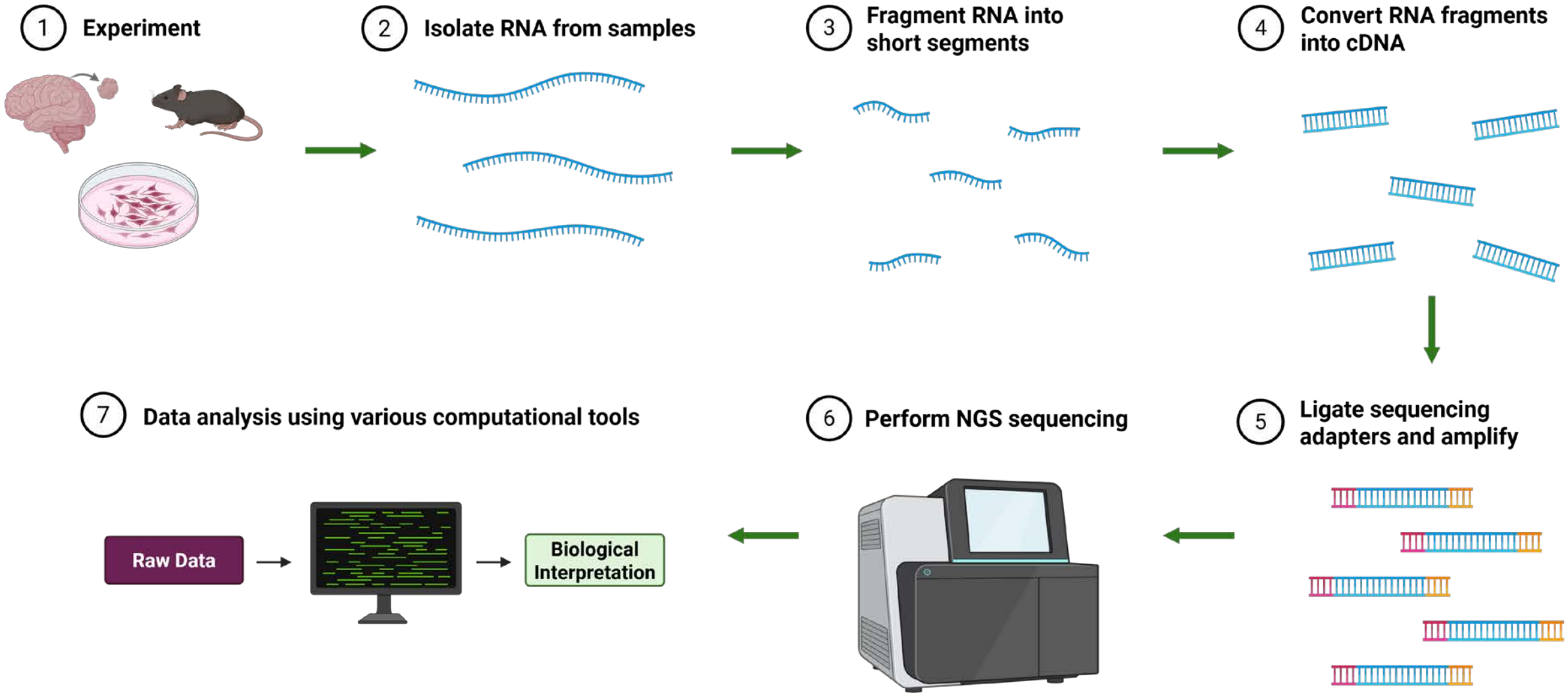
↕

Splicing

Analyzing RNA-Seq Data: What can gene expression tell us?

- Which genes are over/under-expressed in patients vs healthy controls?
- Which genes are correlated to disease progression?
- Can markers of hidden disease be found by sequencing plasma?
- Gene expression signatures for disease?

Workflow of (m)RNA-seq



Acquiring Relevant RNA-Seq Data

- In-house generated datasets
- Public resources: NCBI SRA, ENA, AMP-AD Knowledge Portal, GEO
- Combining public data with in-house datasets for meta-analysis

The ARISTOTLE HPC environment

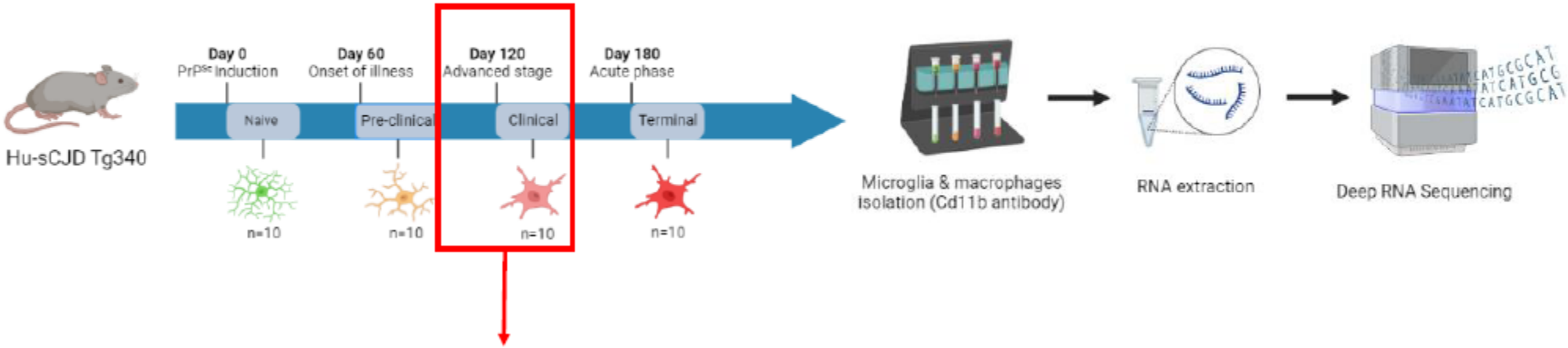
Large-scale data cannot be analyzed in a conventional personal PC

Large-scale data **Local Computer**

Benefits of the HPC:

- **Secure and Extensive Data Storage** – Easily accessible by the infrastructure's computing resources
- **Enhanced Computational Power** – Significantly faster data processing and analysis
- **Parallel Processing** – Efficient handling of large-scale datasets through parallel computing
- **Broad Software Availability** – Access to a wide range of tools and environments (R, Python, both command-line and graphical interfaces)
- **Comprehensive Documentation & Technical Support**
- **Active User Community** – Share knowledge, troubleshoot, and collaborate

Case study – Study Design



Sample Info

Sample Name	Encoding	Total Reads	Sequence Length	% QC
Mic_120_Cntr_01_S70_L001_R1_001.fastq	Sanger / Illumina 1.9	31.270.281	101	44
Mic_120_Cntr_01_S70_L002_R1_001.fastq	Sanger / Illumina 1.9	31.271.220	101	44
Mic_120_Cntr_03_S71_L001_R1_001.fastq	Sanger / Illumina 1.9	33.111.052	101	43
Mic_120_Cntr_03_S71_L002_R1_001.fastq	Sanger / Illumina 1.9	33.015.957	101	43
Mic_120_huCJD_01_S78_L001_R1_001.fastq	Sanger / Illumina 1.9	36.080.949	101	45
Mic_120_huCJD_01_S78_L002_R1_001.fastq	Sanger / Illumina 1.9	35.957.965	101	46
Mic_120_huCJD_02_S79_L001_R1_001.fastq	Sanger / Illumina 1.9	40.078.177	101	45
Mic_120_huCJD_02_S79_L002_R1_001.fastq	Sanger / Illumina 1.9	40.020.221	101	45

Control → 2 technical replicates, 2 biological replicates → 4

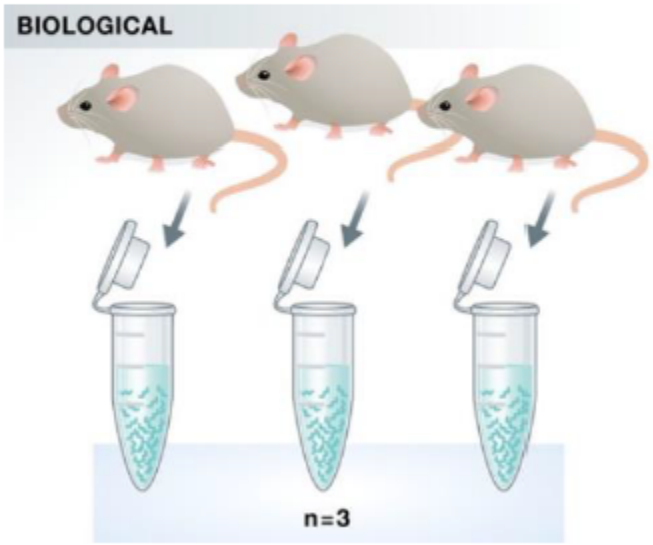
Disease → 2 technical replicates, 2 biological replicates → 4

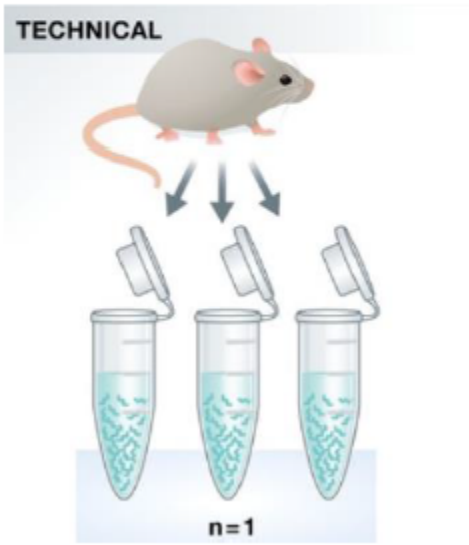
The importance of control groups in biological experiments

- **Establish a Baseline:** Provide a reference point to compare the effects of the experimental treatment.
- **Isolate Variables:** Help ensure that observed effects are due to the variable being tested, not external factors.
- **Increase Reliability**: Enhance the credibility and reproducibility of experimental results.
- **Identify Background Noise**: Help distinguish true biological effects from random fluctuations or technical artifacts.
- **Validate Experimental Setup:** Confirm that the methodology and reagents are working as expected.
- **Support Statistical Analysis:** Enable meaningful comparisons and robust statistical conclusions.

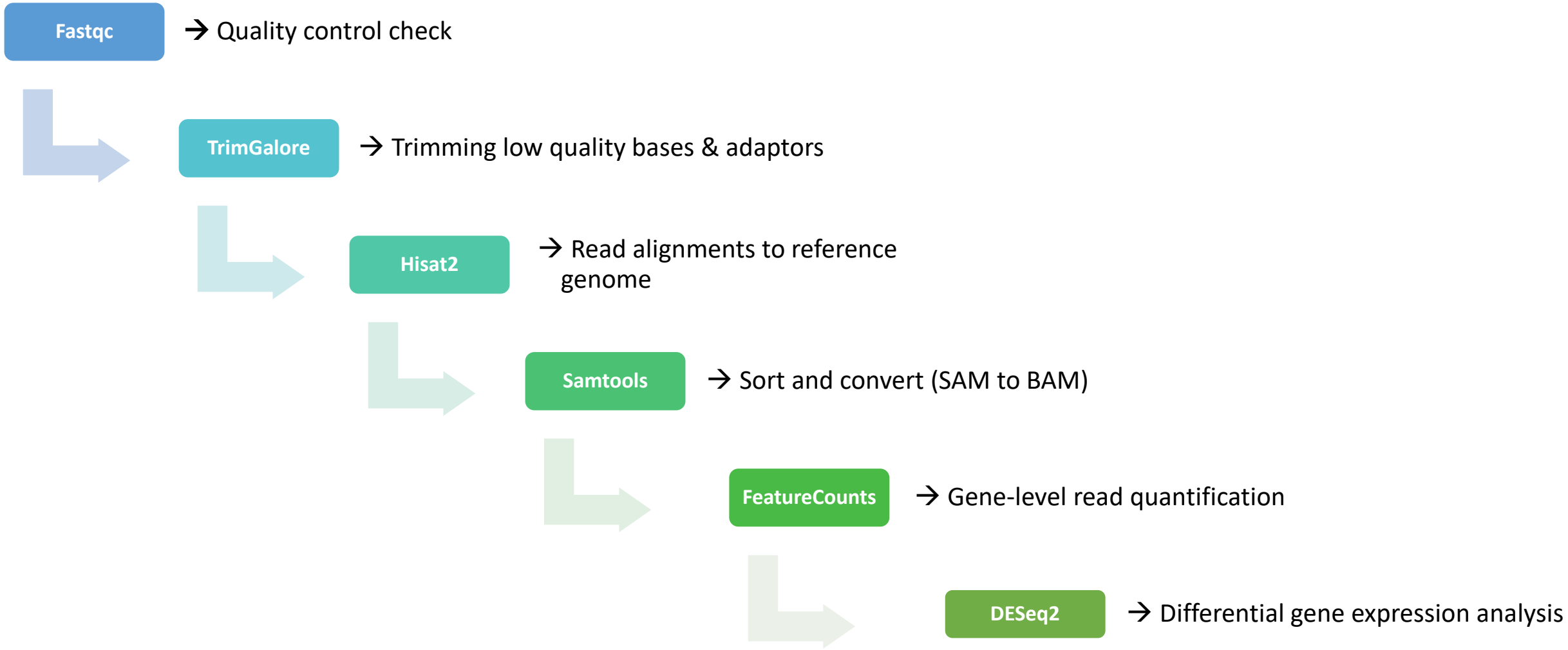
Why do we need replicates?

Replicates → assess and isolate sources of variation in measurements and limit the effect of spurious variation on hypothesis testing and parameter estimation.





Differential Gene Expression Analysis - Workflow

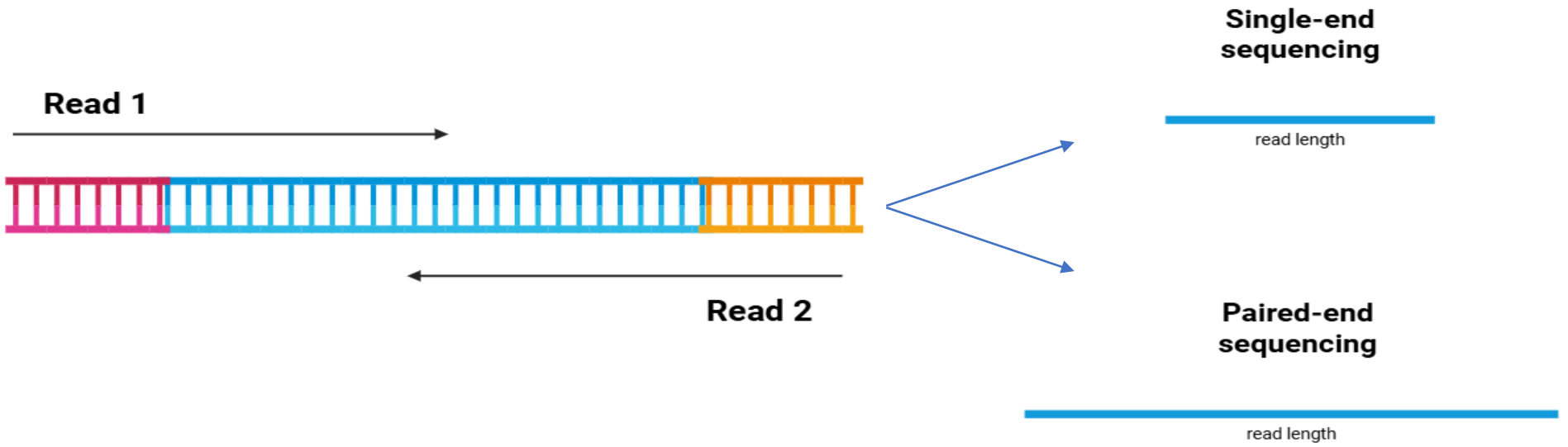


Details on the FASTQ format

What is a FASTQ file?

A fastq file is a text-based file for storing both a biological sequence and its corresponding quality scores.

- Single-read run → one Read 1 (**R1**) FASTQ file is created for each sample
- Paired-end run → one Read 1 (**R1**) and one Read 2 (**R2**) FASTQ file is created for each sample



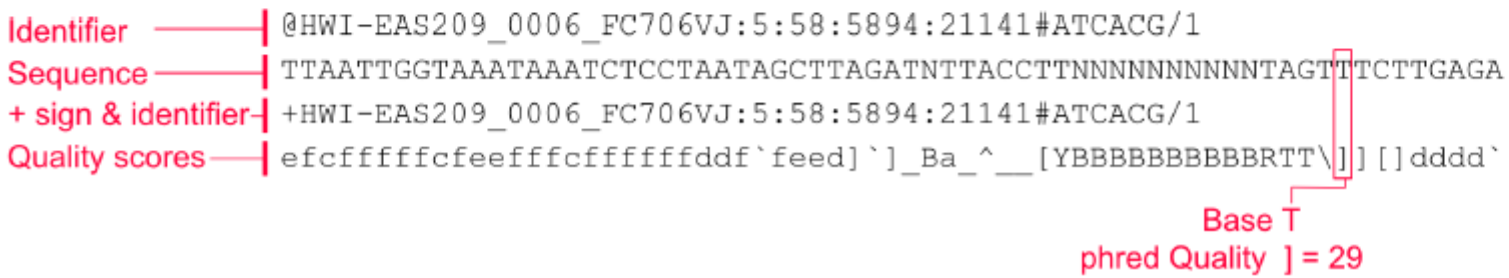
Details on the FASTQ format

What is a FASTQ file?

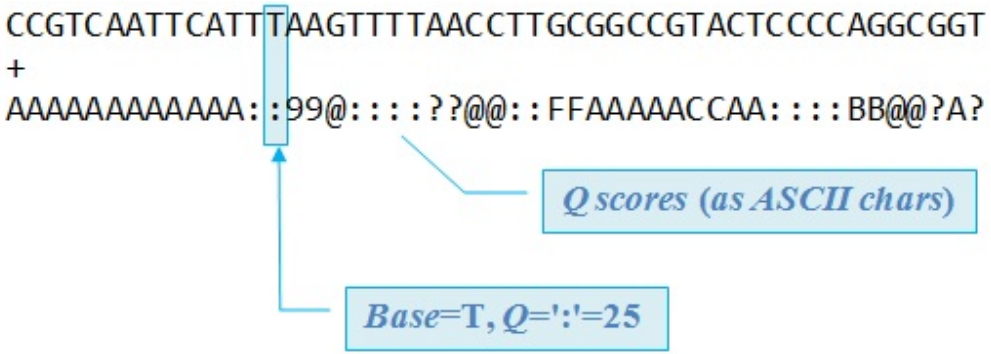
A fastq file is a text-based file for storing both a biological sequence and its corresponding quality scores.

- Single-read run → one Read 1 (**R1**) FASTQ file is created for each sample
- Paired-end run → one Read 1 (**R1**) and one Read 2 (**R2**) FASTQ file is created for each sample

What does a FASTQ file look like?



Sequence Quality: Phred Scores



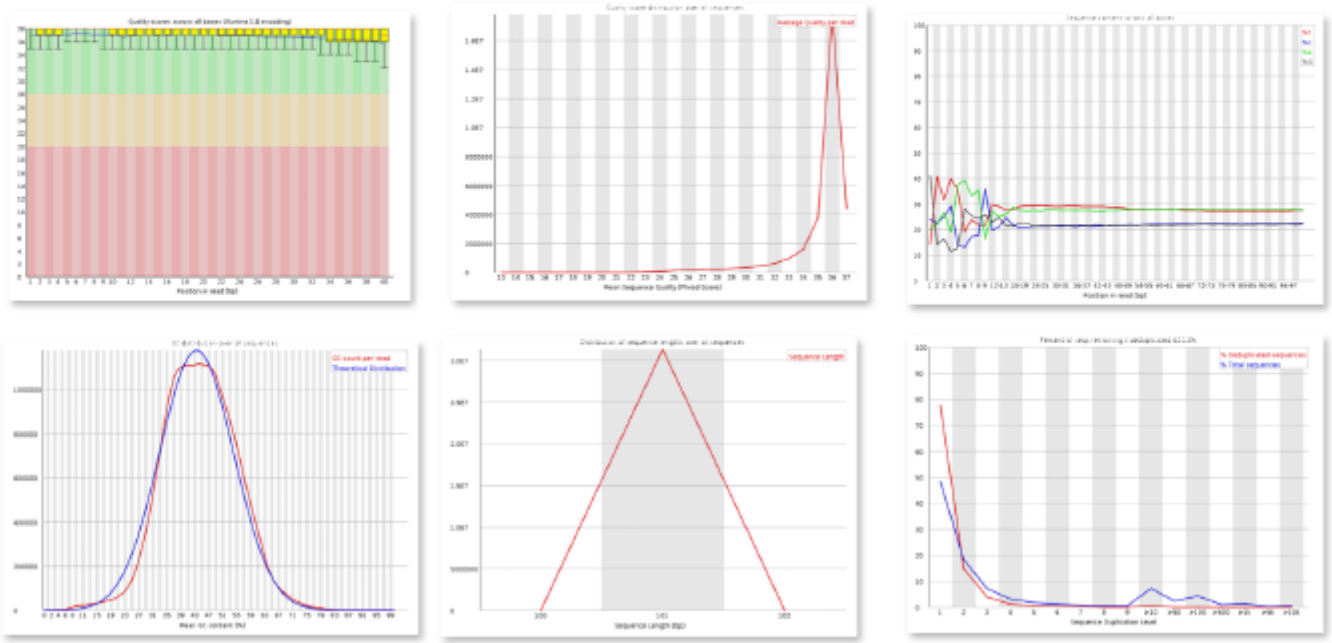
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Table 1 ASCII Characters Encoding Q-scores 0–40

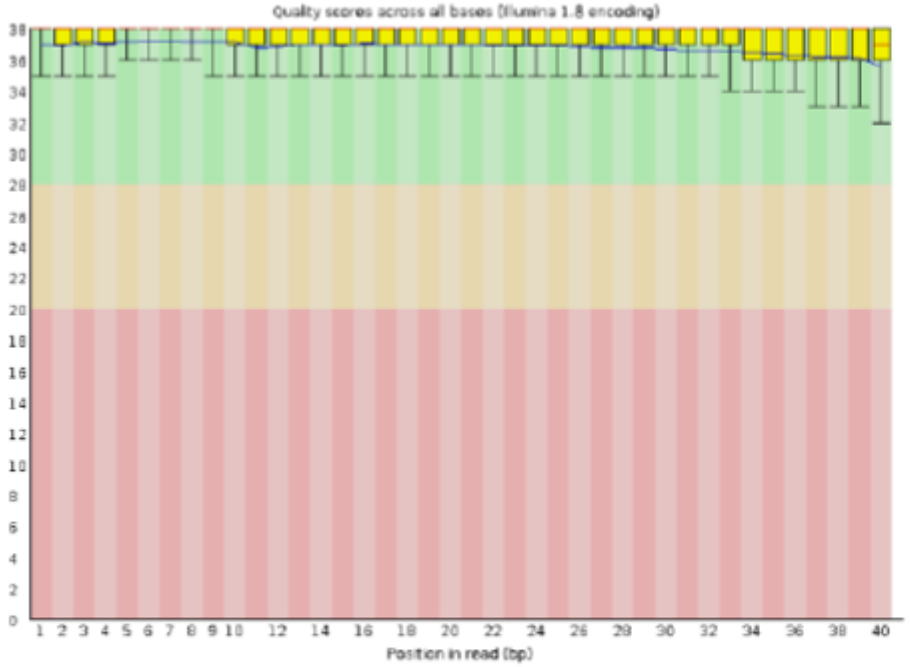
Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			

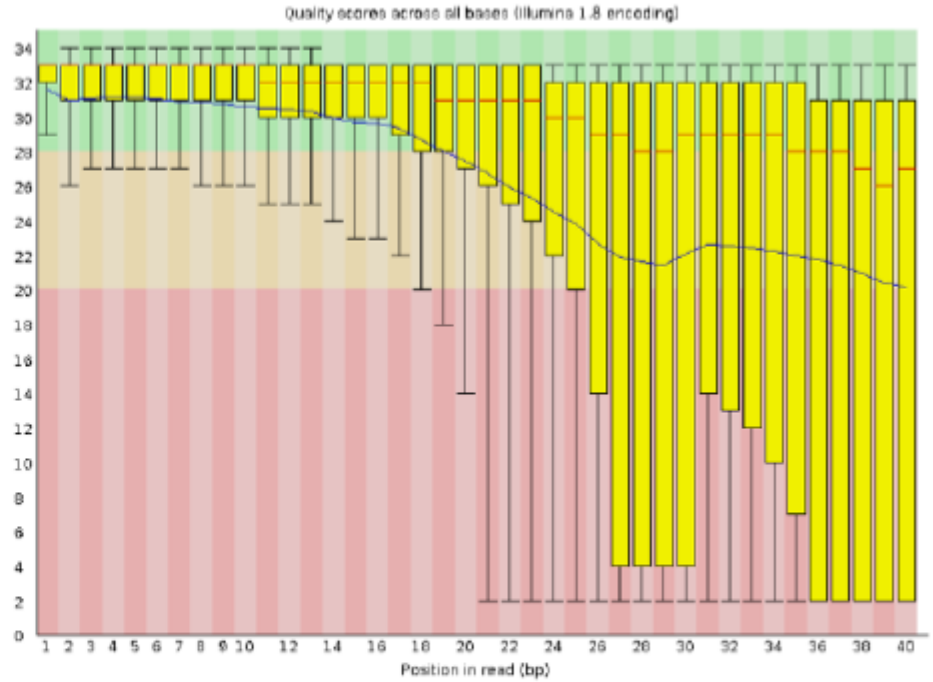
Assessing Read Quality: FASTQC

FastQC reads a set of sequence files and produces from each one a quality control report consisting of a number of different modules, each of which will help identify a different potential type of problem in your data.



What is a good read?





The yellow box shows the base-calling quality scores across all sequence reads. The blue line indicates the mean quality score. Q20 = 99% accuracy. Q30 = 99.9% accuracy...

Improving Read Quality – Trimming & Filtering

Which tool?

- Trimmomatic
- Cutadapt
- Trimgalore
- Fastp
- Scyckle
- Scythe
- Atropos

Which trimming threshold?

RNAseq DE Analysis:

✓Gentle trimming

✓Q > 5

X Too aggressive trimming → losing part of the dataset

RNA Editing Analysis:

✓ You need to be sure of the bases

✓Q > 20

TrimGalore

- **Step 1:** Quality trimming
- **Step 2:** Adapter trimming

```
Illumina:    AGATCGGAAGAGC
Small RNA:   TGGAATTCTCGG
Nextera:     CTGTCTCTTATA
```

- **Step 3:** Removing short Sequences

USAGE:

trim_galore [options] <filename(s)>

Basic Options:

--quality <INT> : Trim low-quality ends from reads.

--fastqc : Run FastQC in the default mode on the FastQ file once trimming is complete.

--adapter <STRING> : Adapter sequence to be trimmed.

--illumina : Trim Illumina universal adapter AGATCGGAAGAGC

--nextera : Trim Nextera adapter CTGTCTCTTATA

--small_rna : Trim Illumina Small RNA 3' Adapter TGGAATTCTCGG

--length <INT> : Discard reads that became shorter than a specified length

```
trim_galore --length 50 --fastqc --cores 4 --quality 25 --output_dir ${Output_Dir} ${RAW_READS}/SRR4447302.fastq
```

https://github.com/FelixKrueger/TrimGalore/blob/master/Docs/Trim_Galore_User_Guide.md

Read Alignment: Hisat2

HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads to a population of human genomes as well as to a single reference genome.

USAGE:

hisat2 [options] –x <hisat2-idx> {-1 <m1> -2 <m2> | -U <r> | --sra-acc <SRA accession number>} [-S <hit>]

Main arguments:

-x <hisat2-idx>: The basename of the index for the reference genome.
-1 <m1>: Comma-separated list of files containing mate 1s (filename usually includes _1).
-2 <m2>: Comma-separated list of files containing mate 2s (filename usually includes _2).
-U <r>: Comma-separated list of files containing unpaired reads to be aligned.
--sra-acc <SRA accession number>: Comma-separated list of SRA accession numbers.
-S <hit>: File to write SAM alignments to.

```
hisat2 -x ${Reference_Genome} -U ${TRIMMED_READS}/SRR4447292.fq  -S SRR4447292.sam
```

https://daehwankimlab.github.io/hisat2/manual/

Samtools

Samtools is a suite of programs for interacting and processing next-generation sequencing data.

Utilities for

- **Dara manipulation:** Efficiently handles large sequencing datasets.
- **Data Analysis:** Provides tools for alignment, sorting, indexing, and variant calling.
- **Integration:** Works seamlessly with other bioinformatics tools and pipelines**.**

Samtools supports various file formats essential for sequence data analysis:

- **SAM** (Sequence Alignment/Map): A text-based format for storing sequence alignment data.
- **BAM** (Binary Alignment/Map): A binary format that is more efficient and compact than SAM.
- **CRAM** (Compressed Reference-oriented Alignment/Map): A highly compressed format for storing alignment data).

Samtools

Samtools is a suite of programs for interacting and processing next-generation sequencing data.

USAGE (sort and convert SAM files to BAM):

samtools sort -o sorted_output.bam input.sam

Main arguments:

sort: Sort alignments by leftmost coordinates.
-o: specifies the file name of the BAM output file.
-@: specifies the number (n) of threads to be used.

```
samtools sort -@ 8 -o ${OUTPUT_OF_SAMTOOLS}/SRR4447292.bam ${OUTPUT_OF_HISAT}/SRR4447292.sam
```

https://www.htslib.org/doc/samtools.html

Gene-level Quantification: FeatureCounts

FeatureCounts is a highly efficient general-purpose read summarization program that counts mapped reads for genomic features such as genes, exons, promoter, gene bodies, genomic bins and chromosomal locations.

USAGE:

featureCounts -O -T n -a example_genome_annotation.gtf -o example_featureCounts_output.txt sorted_example_alignment.bam

Main arguments: .

-O: assigns reads to all their overlapping meta-features.

-T: specifies the number (n) of threads to be used.

-a: genome annotation file (in gtf format).

-o: specifies the name of the output file, which includes the read counts.

sorted_example_alignment.bam: the reads we want to count are aligned to the same genome as the annotation file.

```
featureCounts -O -T 4 -a ${GTF_FILE} -o SRR4447292_featurecounts.txt ${OUTPUT_OF_SAMTOOLS}/SRR4447292.bam
```

https://subread.sourceforge.net/SubreadUsersGuide.pdf

Differential Expression: DESeq2

DESeq2 is a widely used R/Bioconductor package for analyzing differential gene expression from RNA sequencing data.

🔬 What does DESeq2 do?

It helps you identify which genes are significantly upregulated or downregulated between different experimental conditions (e.g., treated vs. untreated, control vs. disease).

⚙️ How DESeq2 works?

1. Takes raw counts (not normalized) as input.
2. Normalizes the data to correct for sequencing depth and RNA composition.
3. Estimates dispersion (biological variability).
4. Fits a model (negative binomial GLM) for each gene.
5. Performs statistical tests to compare conditions.
6. Returns results, including Log2 fold changes (expression difference), p-values and adjusted p-values (FDR).

M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” Genome Biol., vol. 15, no. 12, p. 550, Dec. 2014

DESeq2 Results

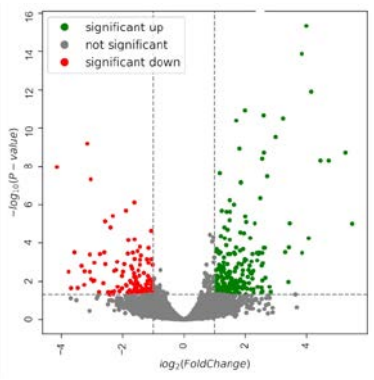
baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	ensembl	entrez	hgnc_symbol
3,088906664	0,17360386	1,330028704	0,130526401	0,896149968		ENSG00000279928		DDX11L17
2018,092198	-0,917883375	0,15340521	-5,983391161	2,18539E-09	3,99921E-07	ENSG00000142611	63976	PRDM16
1276,136007	0,089638059	0,150945089	0,593845478	0,552615469	0,906920199	ENSG00000157911	5192	PEX10
0						ENSG00000224340		RPL21P21
5,120093477	-1,458804625	0,993921711	-1,467725887	0,142178696		ENSG00000226374	105376672	LINC01345
0,227160238	0,750505043	4,55612132	0,164724552	0,869160791		ENSG00000229280		EEF1DP6
572,6635026	-0,092323219	0,286918039	-0,321775581	0,747622715	0,958581976	ENSG00000142655	5195	PEX14
0						ENSG00000232596		LINC01646
0						ENSG00000235054	284661	LINC01777
0						ENSG00000231510		LINC02782
5398,53925	-0,506210352	0,234528343	-2,158418665	0,030895296	0,254806092	ENSG00000149527	9651	PLCH2
684,3607612	-0,549461758	0,153885863	-3,570579821	0,000356192	0,010538345	ENSG00000171621	80176	SPSB1
1,191537866	-1,285993705	2,209136243	-0,582125122	0,560482405		ENSG00000142583	6518	SLC2A5
0						ENSG00000284674	105376680	LINC02781
0,731511629	-1,510303378	3,029456824	-0,498539331	0,618103955		ENSG00000224338		MTCYBP45
3,193664073	0,379589493	1,193031184	0,318172314	0,750354232		ENSG00000226457		RPL22P3
326,366812	0,026592669	0,169574632	0,156819851	0,875386827	0,978269215	ENSG00000173614	64802	NMNAT1
0						ENSG00000215720		MFFP1
0,176745512	-0,660961898	4,562152922	-0,144879382	0,884806105		ENSG00000233623		PGAM1P11
44,88914369	0,103445712	0,340813949	0,303525463	0,761489445	0,95876169	ENSG00000162592	148870	CCDC27
1337,362833	-0,445726517	0,251106711	-1,775048208	0,075889928	0,422025598	ENSG00000204624	57540	DISP3
424,2074237	-0,349026837	0,225657074	-1,546713474	0,121932353	0,537198046	ENSG00000142606	79258	MMEL1
4,801835555	-0,284488298	0,966672631	-0,294296424	0,768531405		ENSG00000171729	55092	TMEM51
31,50910723	0,301846973	0,599932804	0,503134637	0,614869616	0,924803147	ENSG00000279457		WASH9P
1762,813582	0,061000321	0,138665076	0,439911204	0,660001421	0,935260788	ENSG00000037637	54455	FBXO42
872,0717017	-0,164487615	0,233265197	-0,705152834	0,48071513	0,877414463	ENSG00000159423	8659	ALDH4A1
2796,059575	-0,020226542	0,104126924	-0,194248917	0,845980982	0,974171556	ENSG00000157916	11079	RER1

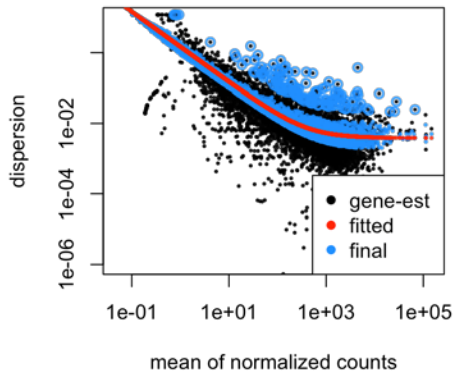
Differential Expression: DESeq2

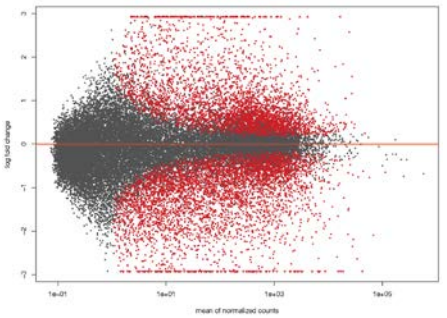
DESeq2 is a widely used R/Bioconductor package for analyzing differential gene expression from RNA sequencing data.

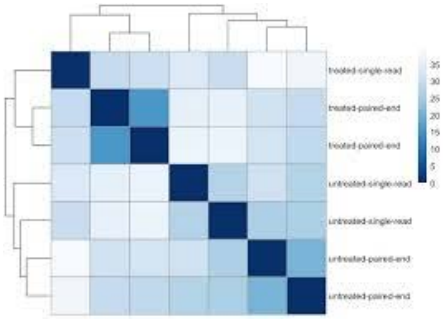
What are the key features of DESeq2?

1. Handles biological replicates
2. Adjusts for multiple testing (Benjamini-Hochberg)
3. Supports complex experimental designs
4. Offers tools for visualization (e.g., PCA, MA-plots, heatmaps)









M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” Genome Biol., vol. 15, no. 12, p. 550, Dec. 2014