BULK RNASEQ

MOHAMED HASSAN
PHD STUDENT

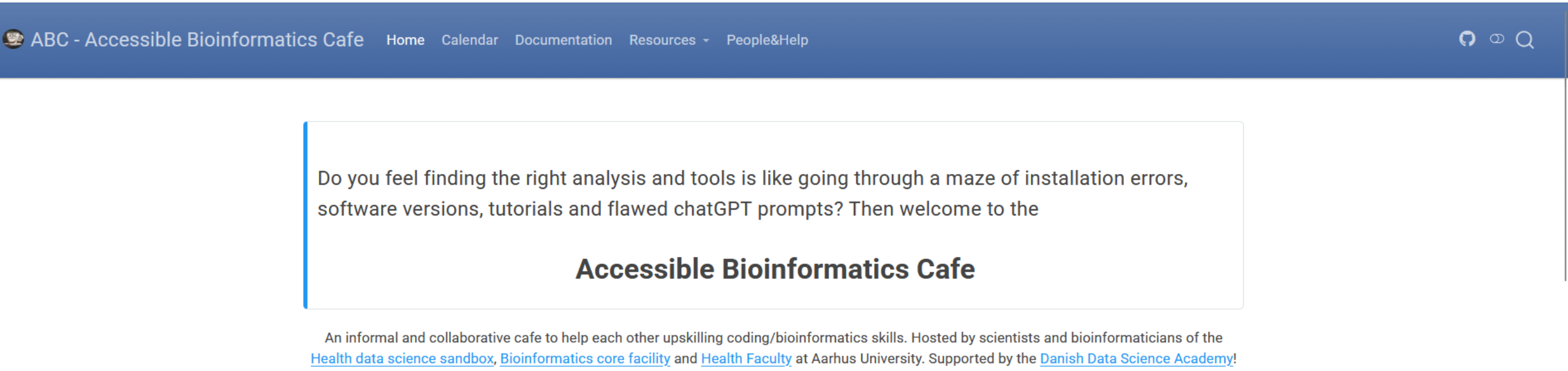
Other OMICS workshops and cafe

15/01	13.00-15.00 (@1150-115)	**ABC** (coding help and assistance – no signup needed)
22/01	10.00-14.00 (@1241-114)	Workshop: **GenomeDK** 1 (Intro to HPC and cluster basics)
29/01	13.00-15.00 (@1241-114)	**ABC** (coding café – no signup needed)
26/02	12.00-16.00 (@1241-114)	Workshop: **ABC OMICS** 1 (R and Tidyverse)
09/03	10.00-14.00 (@1241-114)	Workshop: **GenomeDK** 2 (Advanced usage)
12/03	13.00-15.00 (@1241-114)	**ABC** (coding café – no signup needed)
16/03	10.00-14.00 (@1241-114)	Workshop: **GenomeDK** 3 (Pipelines)
26/03	12.00-16.00 (@1241-114)	Workshop: **ABC OMICS** 2 (bulkRNA data)
09/04	13.00-15.00 (@1241-114)	**ABC** (coding café – no signup needed)
23/04	12.00-16.00 (@1264-104)	Workshop: **ABC OMICS** 3 (scRNA data)
07/05	13.00-15.00 (@1150-122)	**ABC** (coding café – no signup needed)
21/05	12.00-16.00 (@1241-114)	Workshop: **ABC OMICS** 4 (scRNA spatial data)
04/06	13.00-15.00 (@1150-122)	**ABC** (coding café – no signup needed)

JOIN THE DISCORD CHANNEL

ACCESS TO THE MATERIAL

https://abc.au.dk/



ACCESS TO WORKSHOP1 MATERIAL

https://abc.au.dk/documentation/2026-02-26-tidyR-omics1.html

ACCESS TO WORKSHOP2 MATERIAL

https://abc.au.dk/documentation/2026-03-26-bulkRNA-omics2.html

RECAP FROM WORKSHOP 1

MOHAMED HASSAN
PHD STUDENT

R + tidyr (Data Wrangling)

Mohamed Hassan

11 February 2026

R + tidyr (Data Wrangling)

Mohamed Hassan

11 February 2026

- Merging/Joining dataframes
    - left_join()
    - Debug: who did NOT match?
- Reshaping data (Wide vs Long data shape)
    - 1. pivot_longer(): wide -> long
    - 2. pivot_wider(): long -> wide
    - 3. Binding rows and columns
        - 3.1 rbind vs bind_rows (stacking rows)
        - 3.2 bind_cols() vs cbind() (gluing columns)
- Recap
    - 1. **Column-wise** functions
        - (modify or choose columns):
        - Create or modify columns
        - Identify columns using tidyselect helpers:
    - 2. **Row-wise** functions (filter, slice, row operations)
        - Keep or remove rows
        - Order rows
    - 3. **Group-wise** functions
        - (operate within groups)
        - Grouping changes how other functions behave:
    - 4. **Data-frame–wise** functions
        - (operate on whole table)
        - Reshape data:
    - 5. **Summary** functions
        - (usually inside summarise())
- Good to know: syntactic names

FUNCTIONS WE WILL USE TODAY

mutate

select

left_join

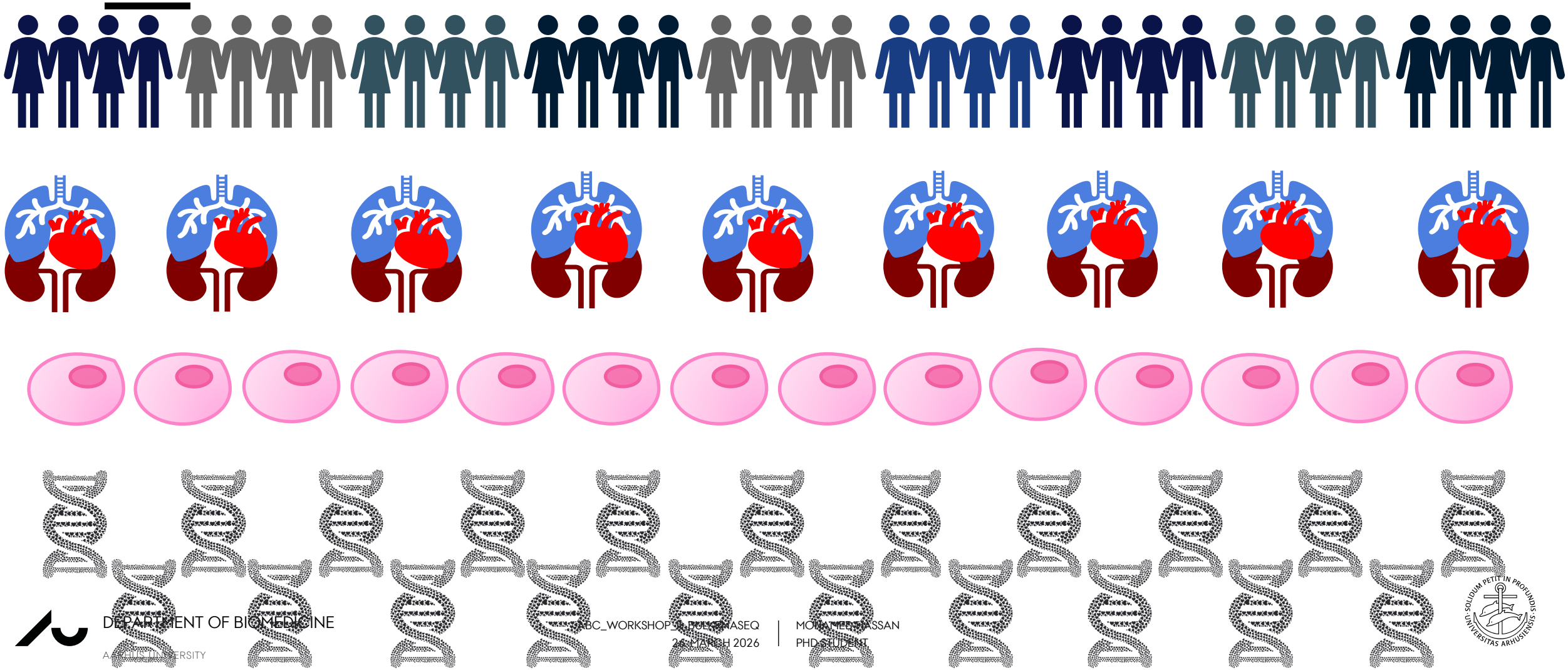
filter

%>% or |>

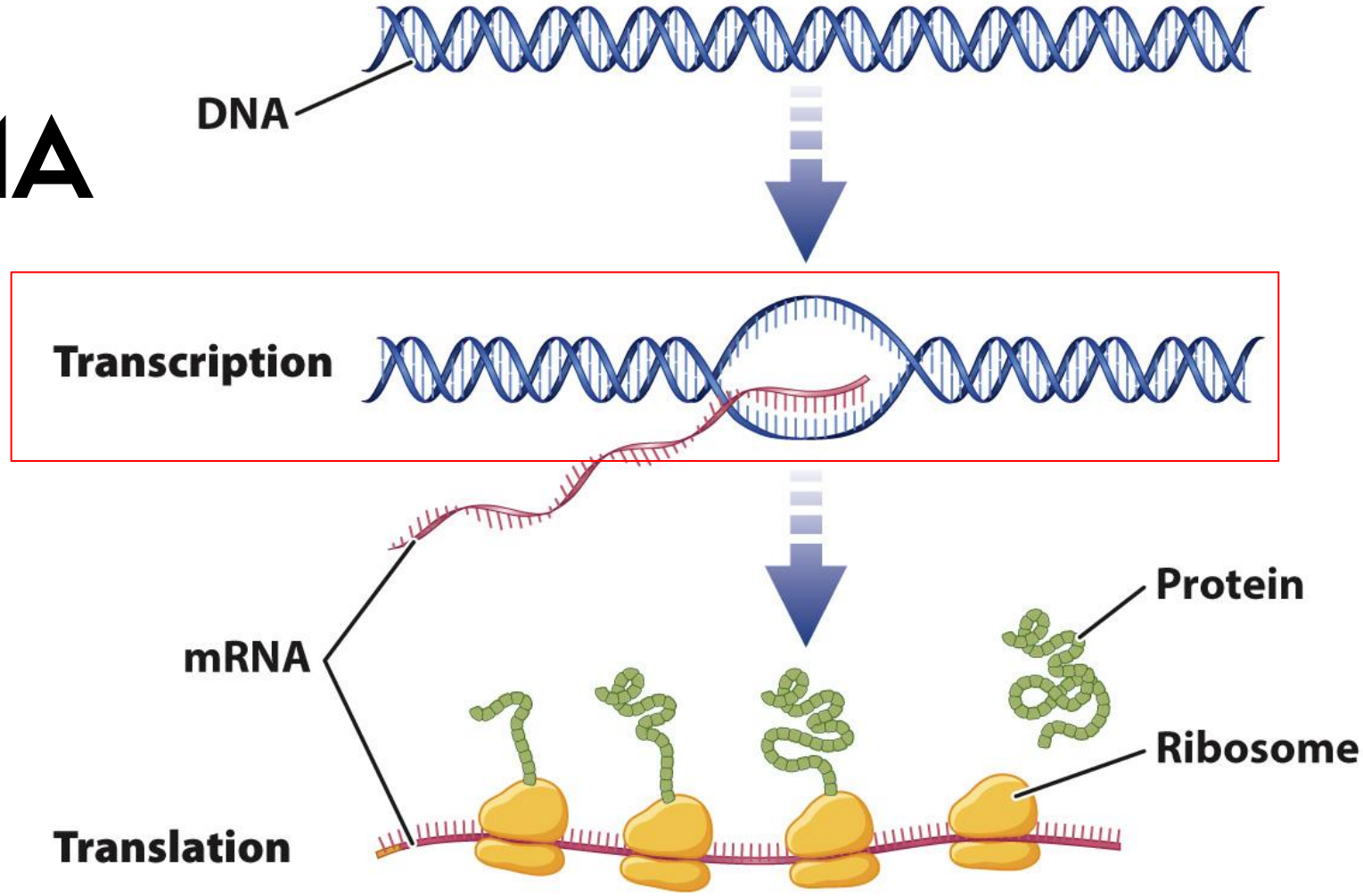
BULK RNASEQ

MOHAMED HASSAN
PHD STUDENT

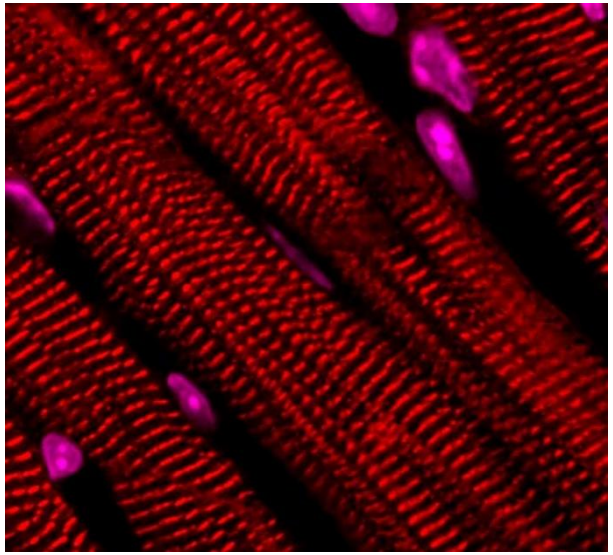
WHAT IS GENE EXPRESSION

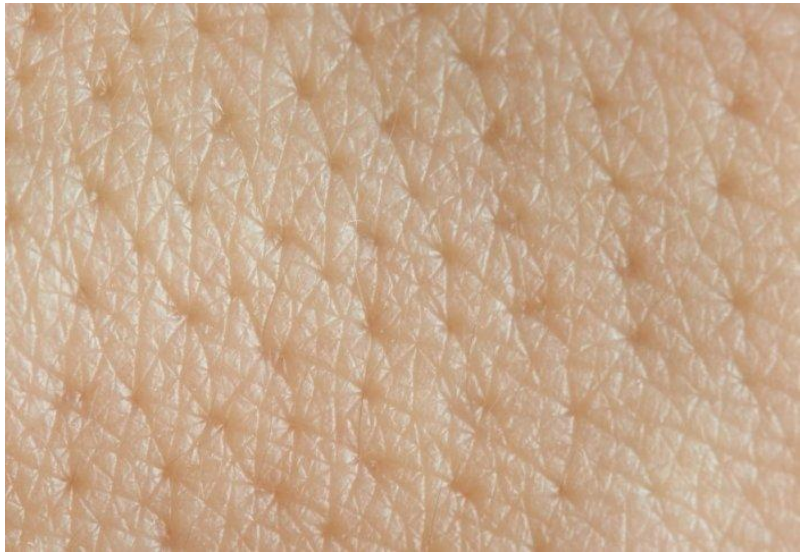


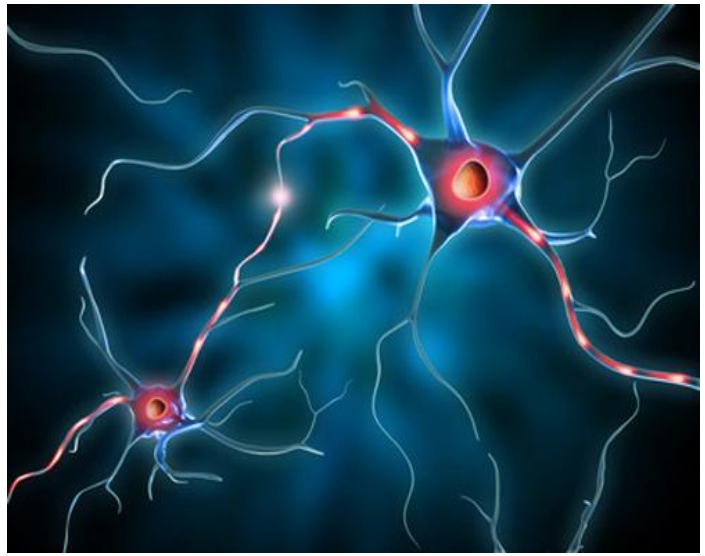
CENTRAL DOGMA

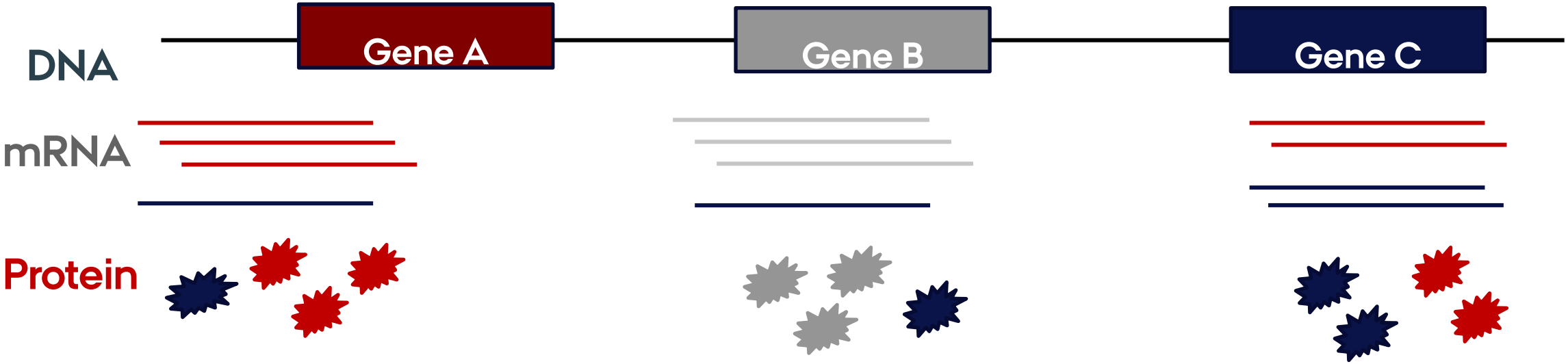


Proteins provide structure and carry out many essential activities in a cell.








DNA
Gene A
Gene B
Gene C
mRNA
Protein

TRANSCIPTOME PROFILE

is the total set of transcripts or mRNA molecules expressed in a cell at specific time **(snapshot of the cell)**, different cells show different patterns of gene expression.

These differences are responsible for the many different properties and behaviors of various cells and tissues, both in health and disease.

WHY RNA OVER DNA?

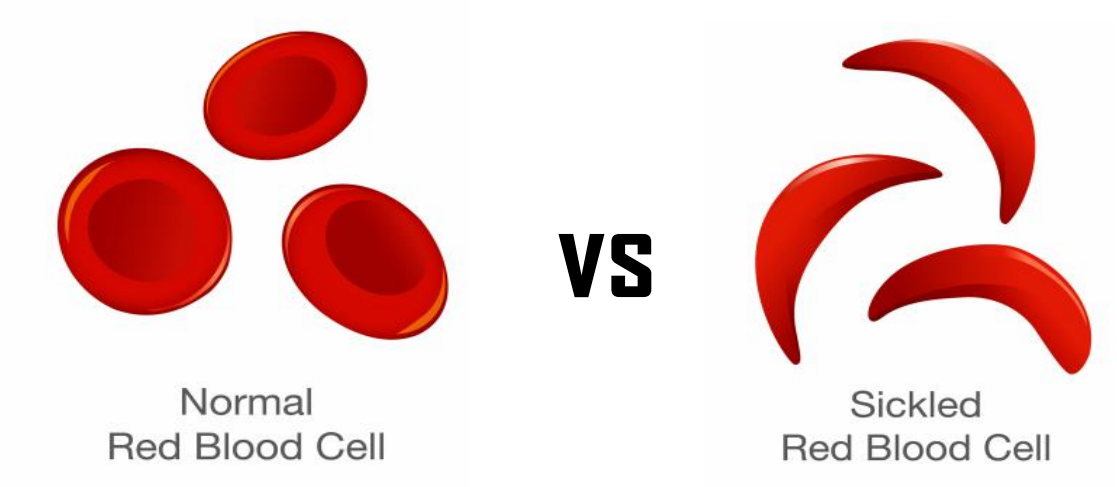
Functional studies

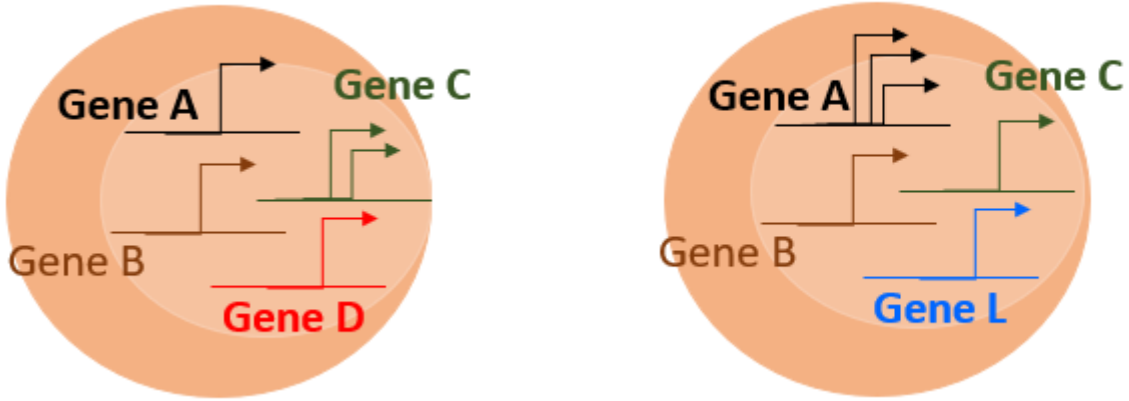
– Genome may be constant, but an experimental condition has a pronounced effect on gene expression (differential expression)

- e.g. Drug treated vs. untreated cell line
- e.g. Wild type versus knock out mice

Some molecular features can only be observed at the RNA level

– Alternative isoforms, fusion transcripts, RNA editing





Gene A is up regulated

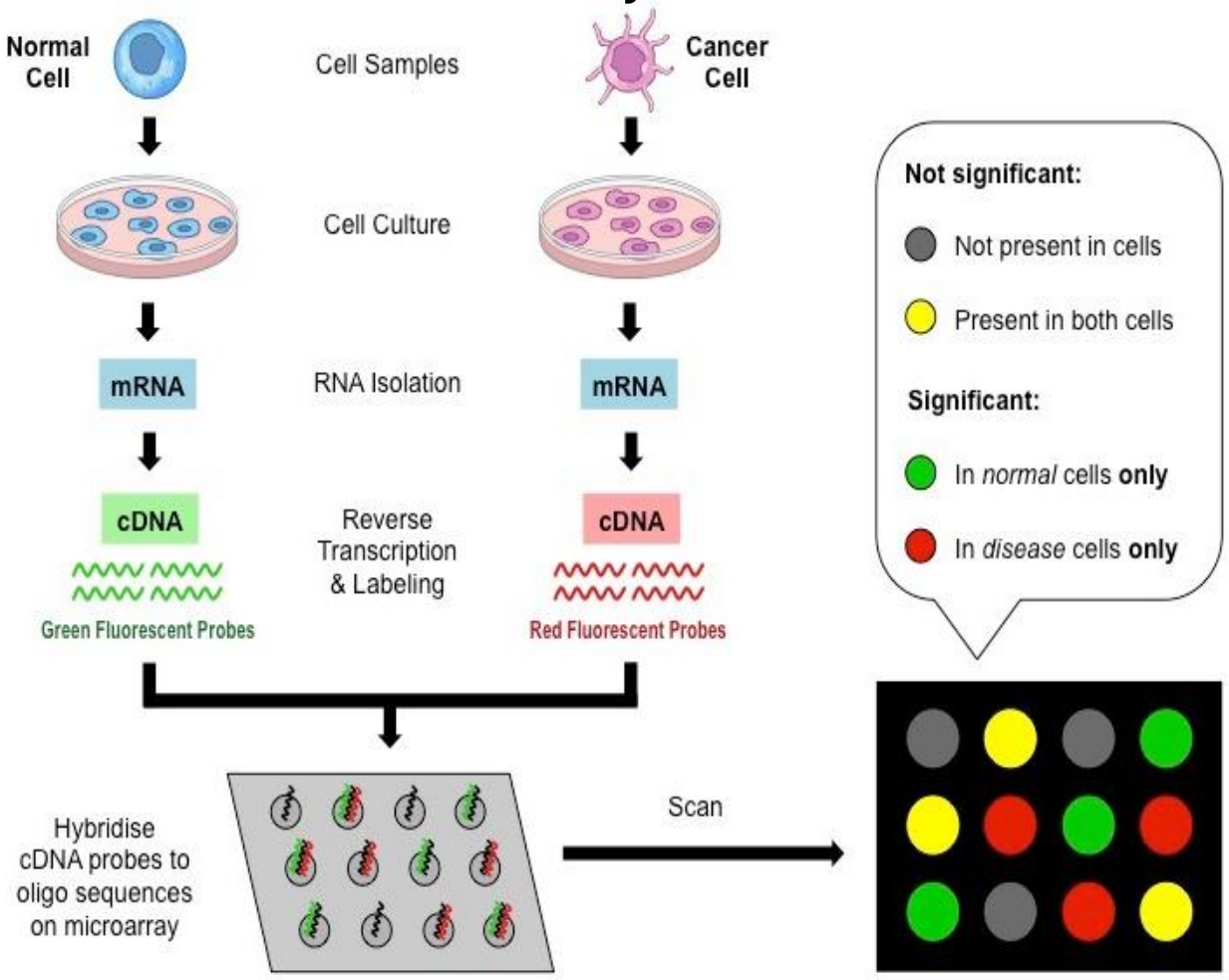
Gene C is down regulated

Gene D is turned off

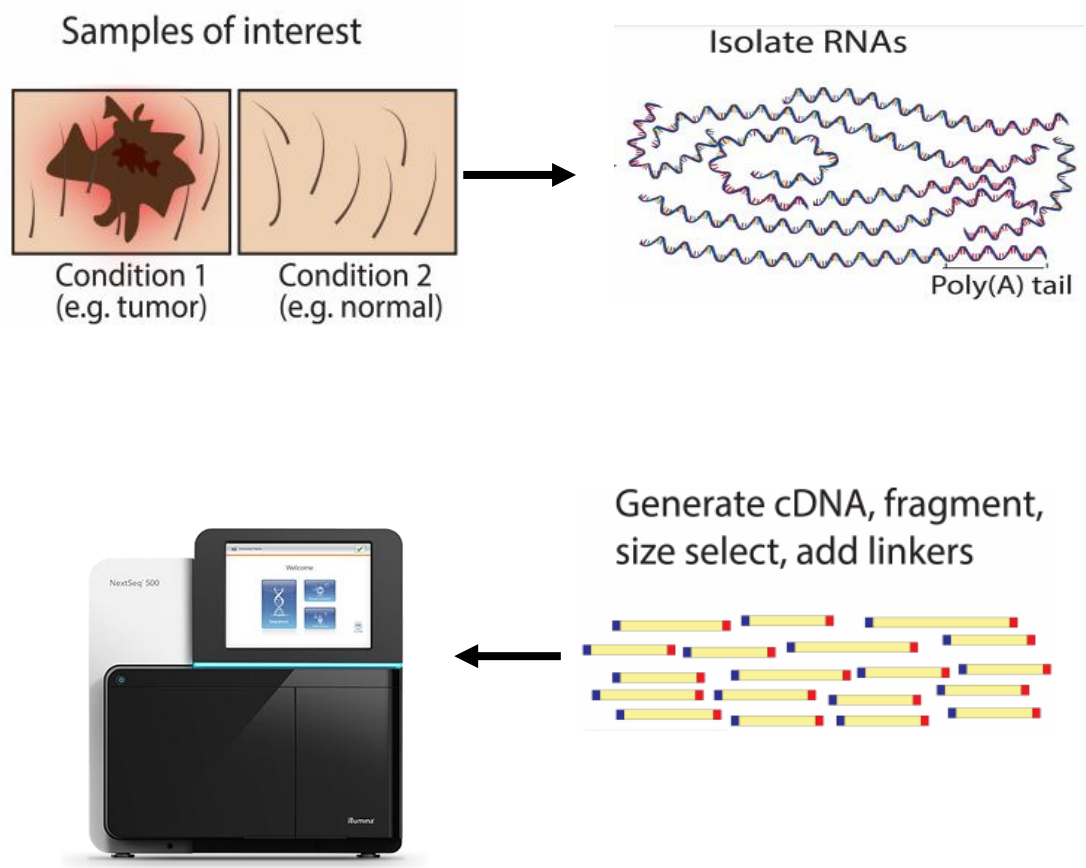
Gene L is turned on

GENE EXPRESSION PROFILING TECHNOLOGIES

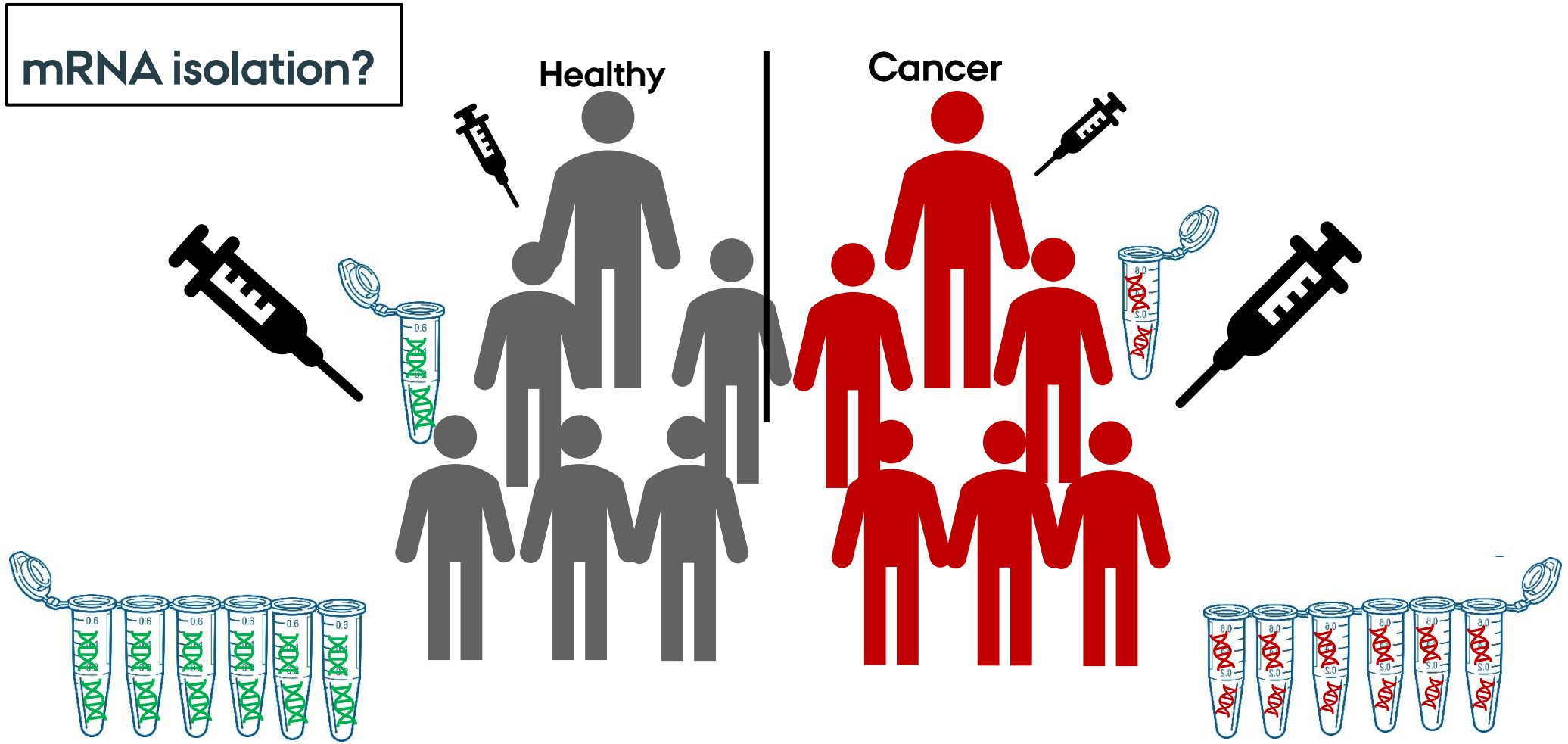
Microarray



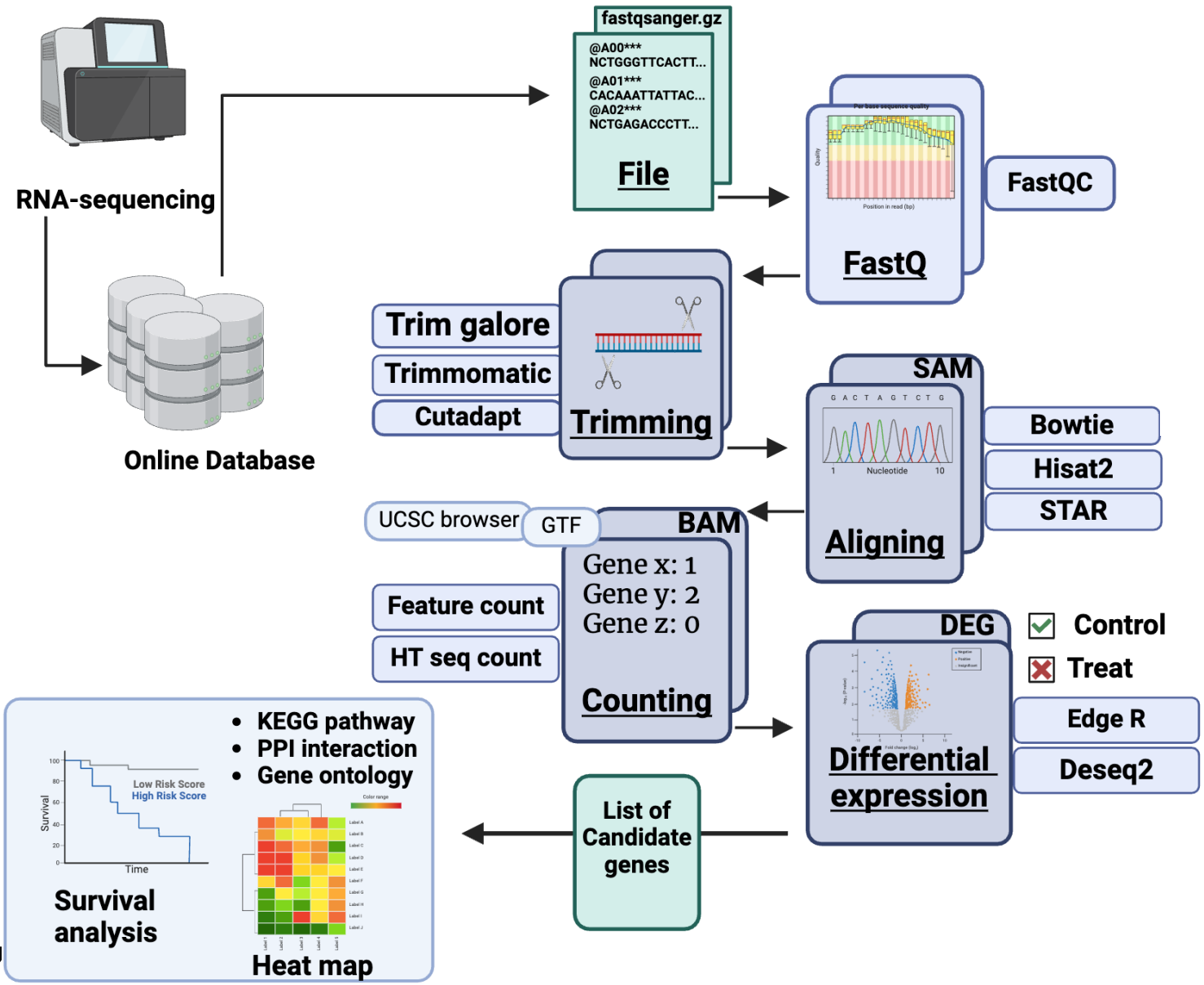
RNA-seq



RNA-SEQ EXPERIMENT OVERVIEW



WORKFLOW


RNA-sequencing
Online Database
fastqsanger.gz
@A00***
NCTGGGTTCACTT...
@A01***
CACAAATTATTAC...
@A02***
NCTGAGACCCTT...
File
FastQ
FastQC
Trim galore
Trimmomatic
Cutadapt
Trimming
SAM
Nucleotide
Bowtie
Hisat2
STAR
Aligning
UCSC browser
GTF
BAM
Gene x: 1
Gene y: 2
Gene z: 0
Feature count
HT seq count
Counting
DEG
Control
Treat
Edge R
Deseq2
Differential expression
List of Candidate genes
KEGG pathway
PPI interaction
Gene ontology
Low Risk Score
High Risk Score
Survival
Time
Survival analysis
Heat map

FASTQ FILES

B1943-E2porc-1_S1_L001_R1_001.fastq - Notepad

File Edit Format View Help

```
@M00877:474:000000000-K7HDK:1:1101:10371:2108 1:N:0:TCTCATGA+CTTCTAAT
TTTATGGCCTACGGGAGGCAGCAGTGAGGAATATTGGACAATGGAGGCAACTCTGATCCTGCCATTCCGCGTGTTTGTCTACGTCCCTATGTTCTTTTTTCTTCTTTTTTCTCTG
+
-ACCCGGGGGGGGGGGGGGGGGGGCFGFGF9FFADEF8EFE<A,,8,,:CFF,,;6;,,;;,;,,<,6+8+,,,,,,:,,,,,66:,:,,,,<,9,,+495,CCE,,+4,,,,
@M00877:474:000000000-K7HDK:1:1101:10833:2162 1:N:0:TCTCATGA+CTTCTAAT
TATTAAGCCTACGGGCGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCTGCCATGCCGCGTGTATGAAGAAGGCCTTCTGTTTTTTAAGTACTTTCTTCTTTGAC
+
@ACCCGGGGGGGGGGDGGGGGGGGGGGGGGGGGGGGGGGGGGGEC@+@76DF8,C,6,,C6C?,6C+7+8+B,C,,,,,,,,9:?E,,,C?,D+,,,:,:CEF,,,,,,,,
@M00877:474:000000000-K7HDK:1:1101:16190:2217 1:N:0:CCTCATGA+CGTCTAAT
CGCTAGGCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGAAAGCCTGATGCAGCAACGCCGCGTTAGCGATGAAGGCCTTCTGTTCGTTAAGCTCTTTCCTCAAGGAA
+
@BCCCGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGD77+@C<FGD8CC<,9C,,:++@+@+8+,,,4+:,,,,444C=F,,,C9+B,++,5AED,B55AA,,,,,
@M00877:474:000000000-K7HDK:1:1101:20697:2220 1:N:0:GCTCATGA+CGTCTAAT
TGACCTACCTACGGGAGGCTGCAGTGAGGAATATTGGTCAATGGTCGCGAGACTGATCCTGCCCTGTAGCGTTCATGTCGACTGCCCTCCGTTTTTTTTTCTGCTTTTTTTTTTG
+
```

@FORJUSP02AJWD1
CCGTACGTACGCGCGTACGATCGCGACTAGCTACGTCGATCGACGATCGTCG

Fill in the Quality score >>

$Q = -10 \log_{10}(\text{Probability of Error})$

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

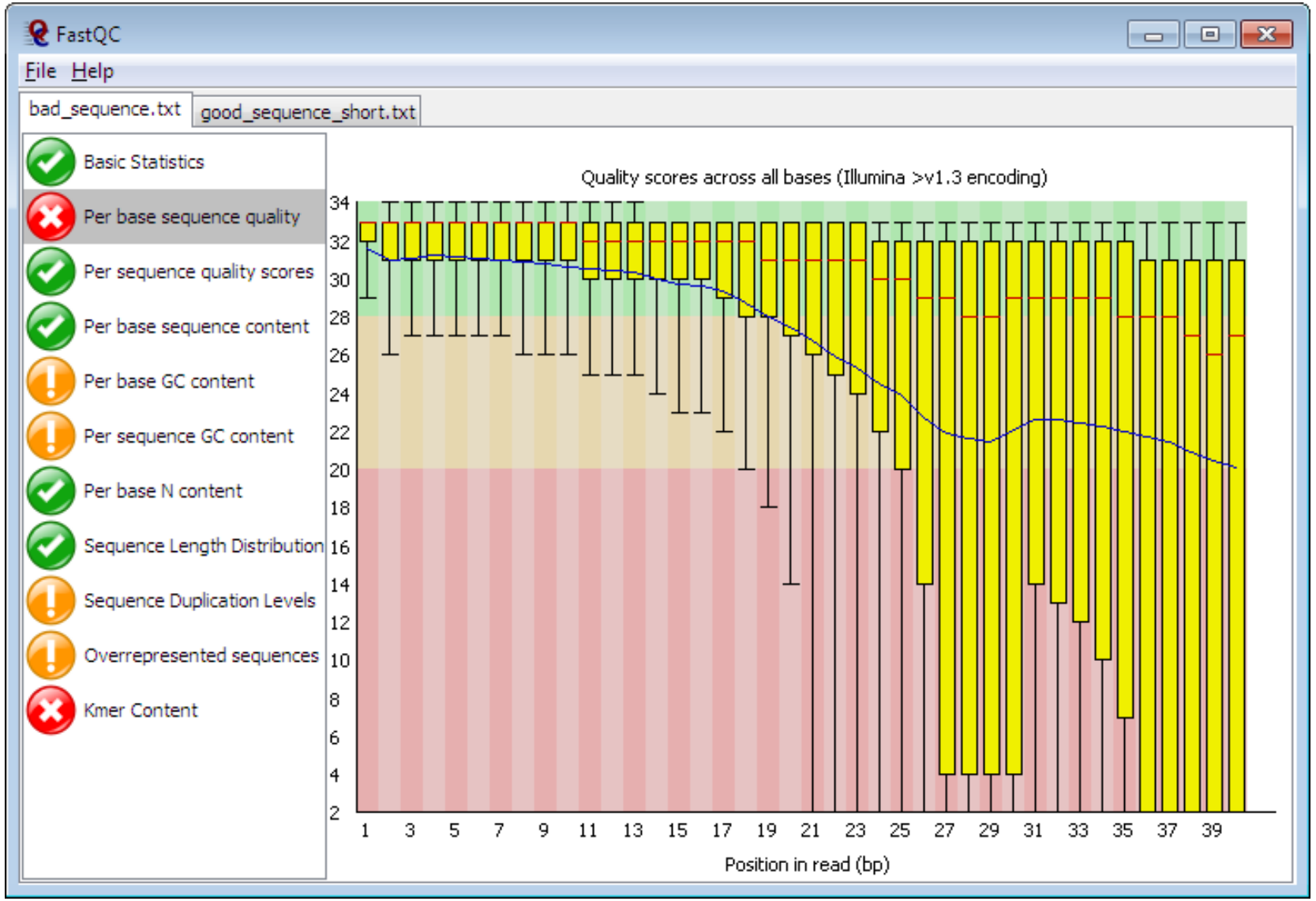
ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

N

FASTQC

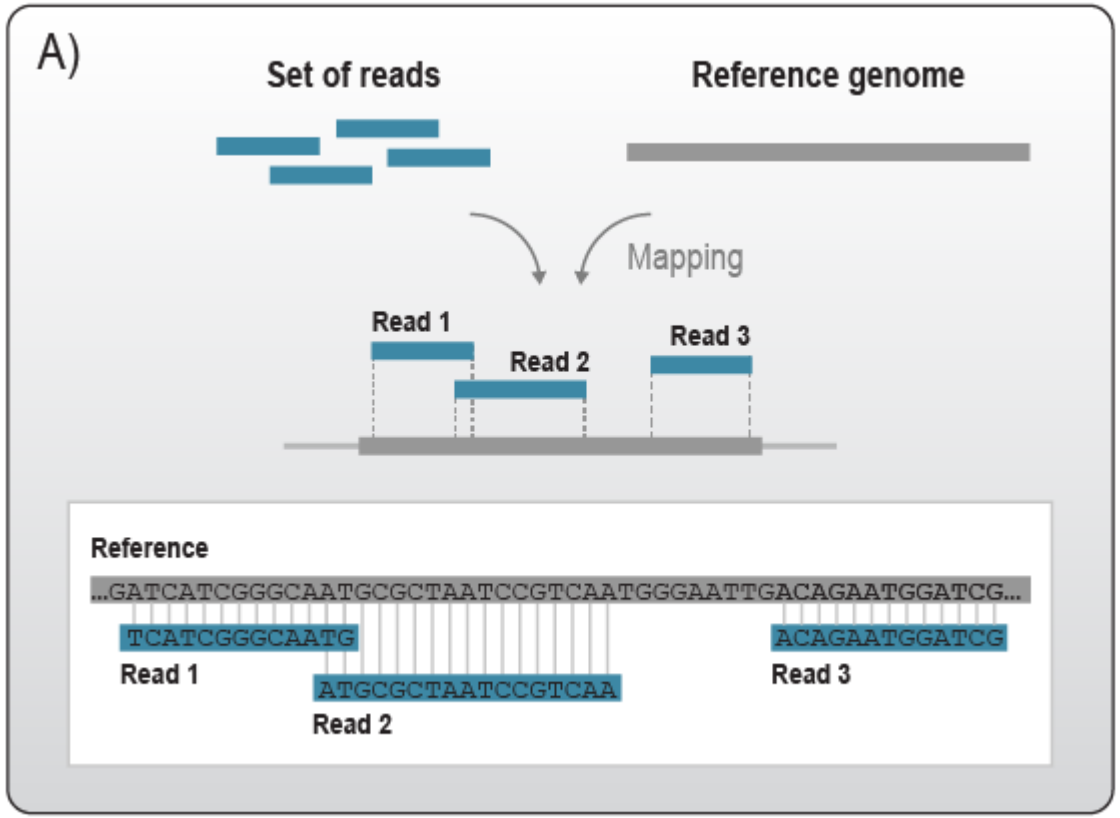
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/



Example Reports

- Good Illumina Data
- Bad Illumina Data
- Adapter dimer contaminated run
- Small RNA with read-through adapter
- Reduced Representation BS-Seq
- PacBio
- 454

ALIGNING THE READS TO THE REFERENCE GENOME

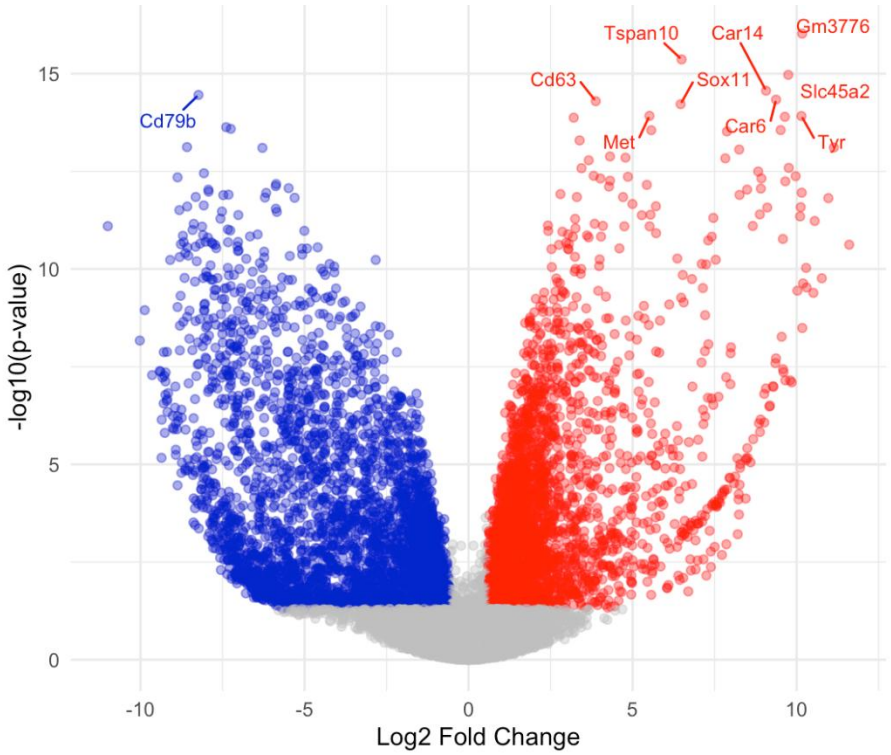


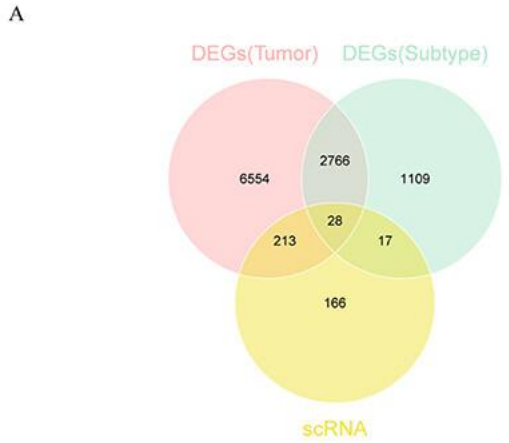
Each column is a sample

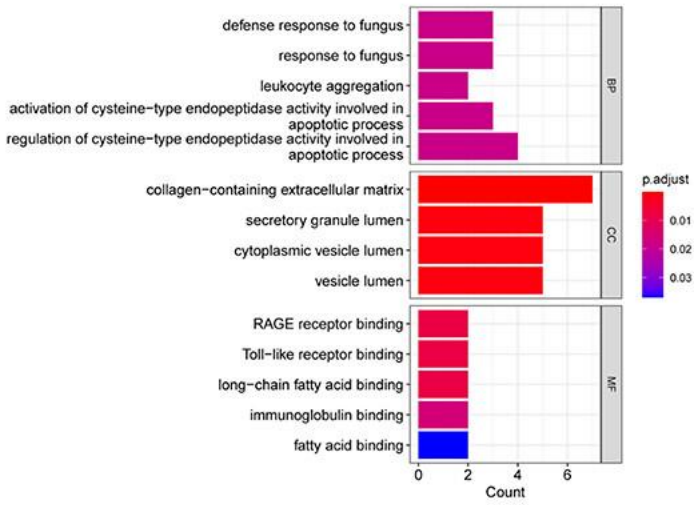
Each row is a gene

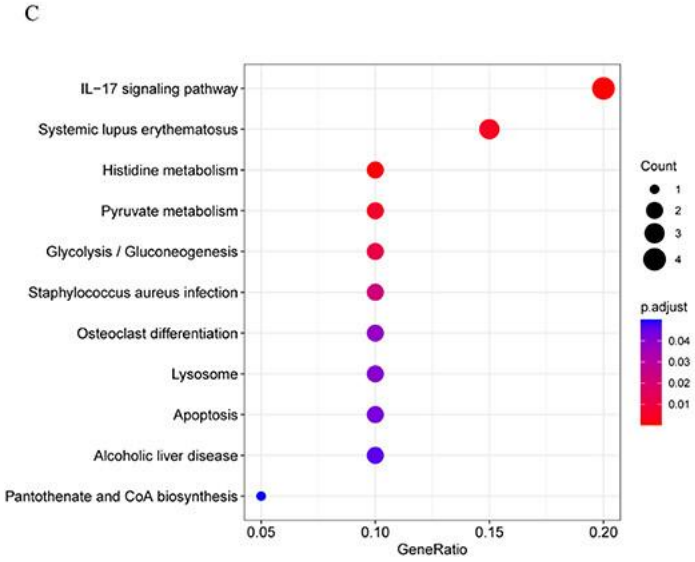
GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666

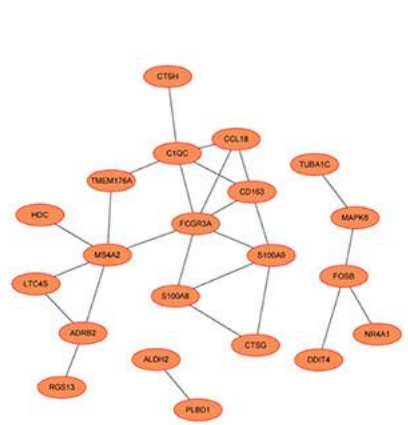
PROCESSING AND DATA ANALYSIS











GO TERMS

Gene ontology. It's a big database that groups genes into:

Biological processes (e.g. *apoptosis*)
Molecular functions (e.g. *ATP binding*)
Cellular components (e.g. *mitochondrion*)

GO enrichment asks:

Are certain biological functions overrepresented in my list of significant genes compared to what you would expect by chance?"

Imagine a bag of 20,000 marbles, 500 are green.

You pick 100 marbles.

If 30 of them are green, that's way more than expected

>> That's enrichment

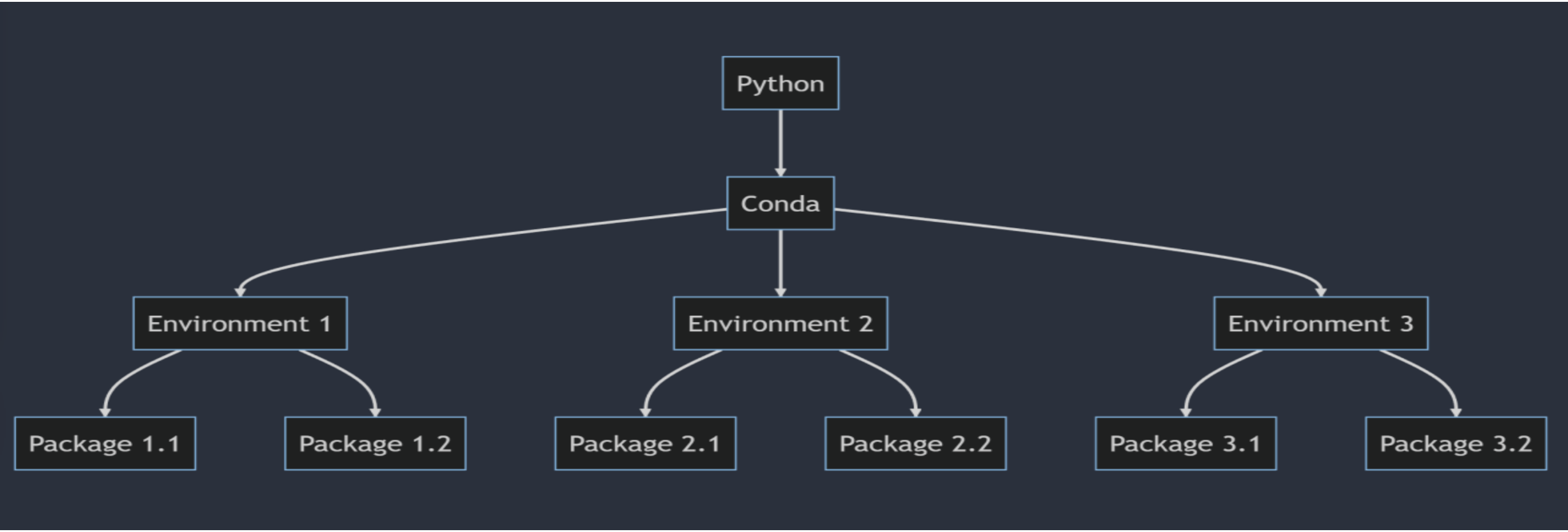
Imagine a bag of 20,000 genes, 500 are apoptosis-related.

You pick 100 DE genes

If 30 of them are apoptosis genes >> that's **way more than expected**
>> That's enrichment

HANDS-ON

CONDA ENVIRONMENT


Python
Conda
Environment 1
Environment 2
Environment 3
Package 1.1
Package 1.2
Package 2.1
Package 2.2
Package 3.1
Package 3.2

connect to GenomeDK from the terminal/MobaXterm (or use the desktop at desktop.genome.au.dk)

ssh yourusername@login.genome.au.dk

Enter your password

Downloading conda on genomedk

wget https://github.com/conda-forge/miniforge/releases/latest/download/Miniforge3-Linux-x86_64.sh -O miniforge.sh

Making it executable

chmod +x miniforge.sh

Run the installer

bash miniforge.sh -b

Initiate conda in the terminal

./miniforge3/bin/conda init bash

Configuring conda

conda config --append channels bioconda

conda config --append channels genomedk

conda config --set channel_priority strict

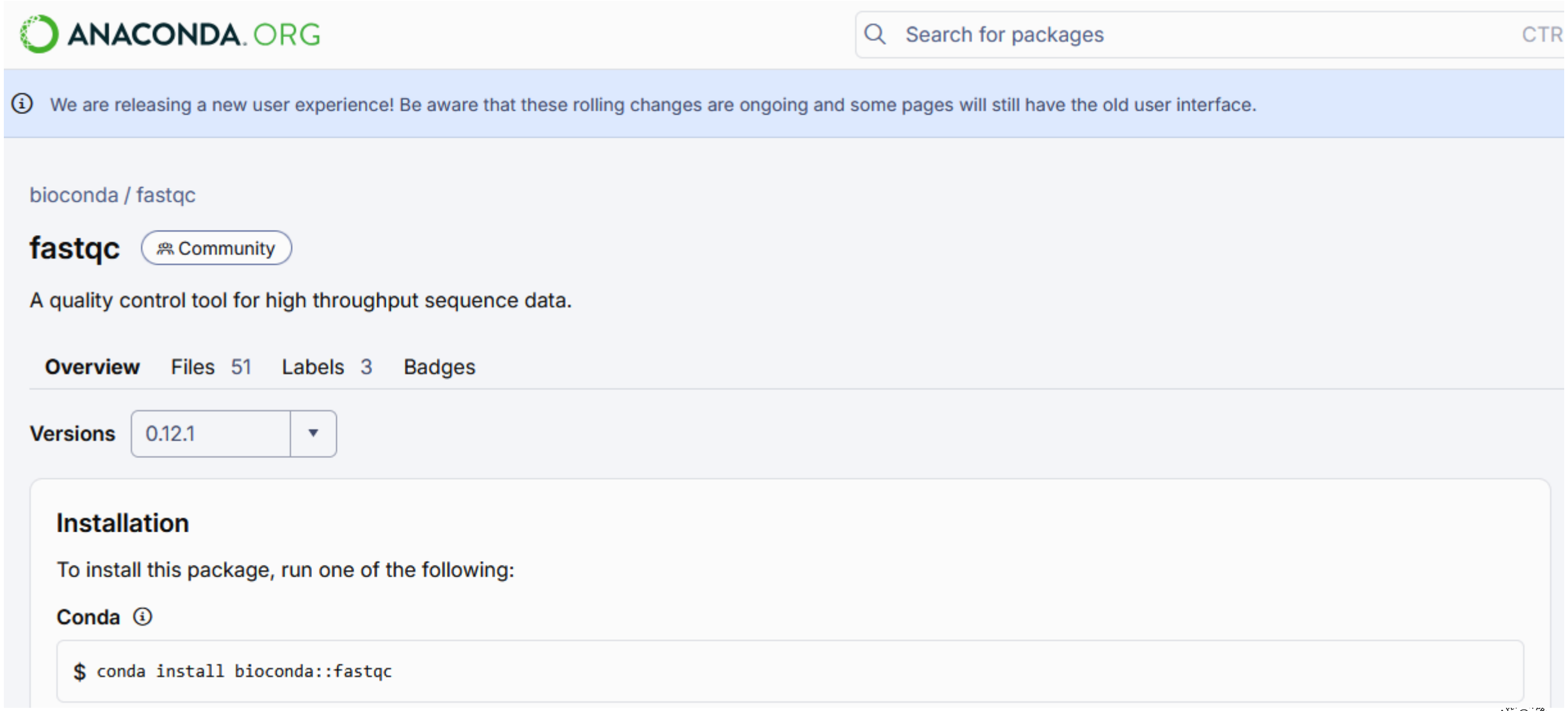
CREATING A CONDA ENVIRONMENT

conda create –n ‘env_name’ –y

conda activate env_name

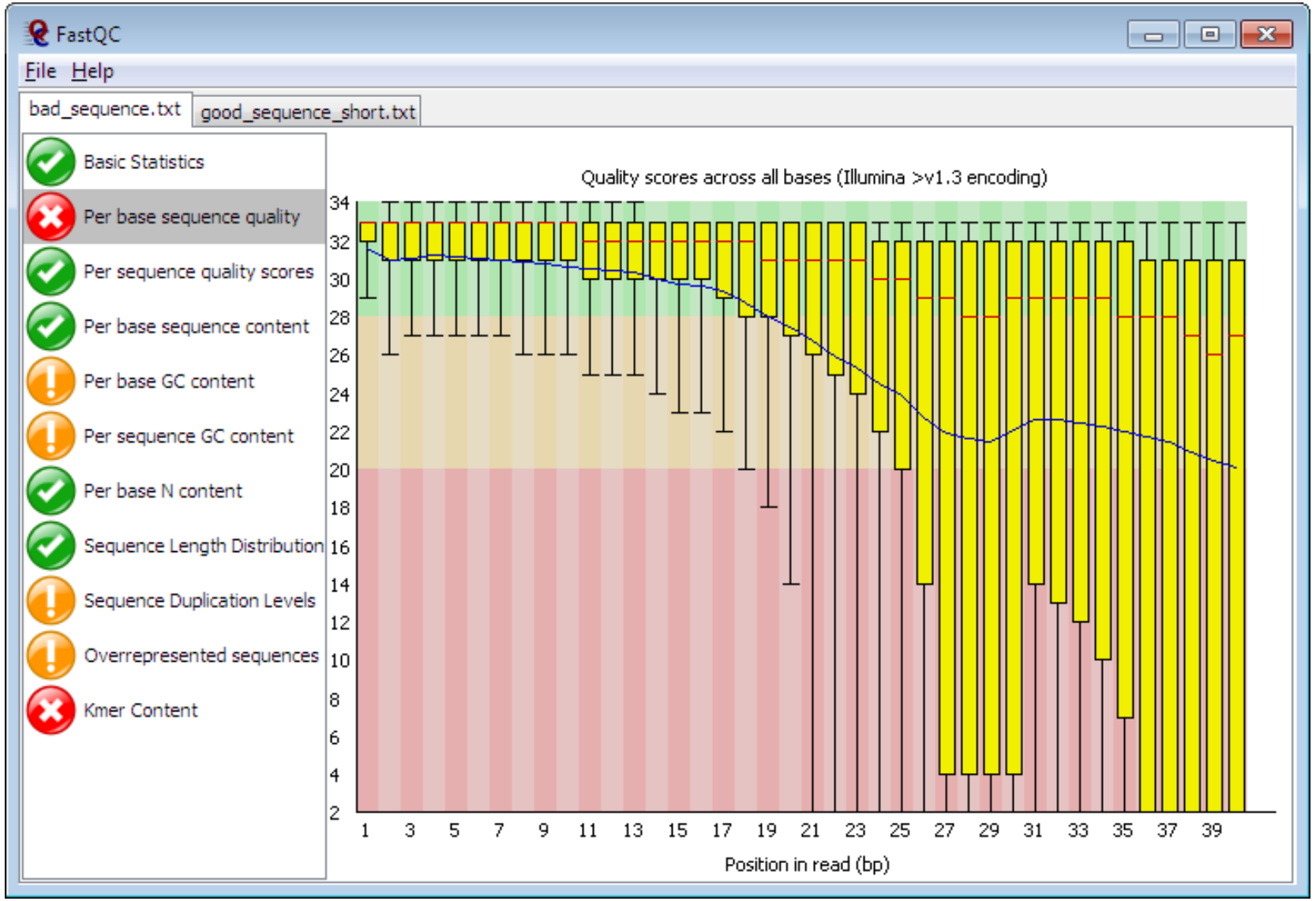
conda install <package_name> Ex: conda install r-ggplot2

CONDA INSTALL BIOCONDA::FASTQC



FASTQC

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/



Babraham Institute

Babraham Bioinformatics

About | People | Services | Projects | Training | Publications

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

Download Now

FastQC A quality control application for high throughput sequence data

- README
- Installation and setup instructions
- Release Notes Please read these before using the program.
- FastQC v0.12.1 (Win/Linux zip file)
- FastQC v0.12.1 (Mac DMG image)
- Source Code for the latest FastQC release

https://github.com/Morgan-Feuz/fastqc

BASH COMMANDS

pwd

ls

mkdir

cd

cat

mv

rm

head

tail

nano/gedit

|

>

*

HANDS-ON

DEPARTMENT OF BIOMEDIC