Differential Gene expression analysis by RNA-sequencing

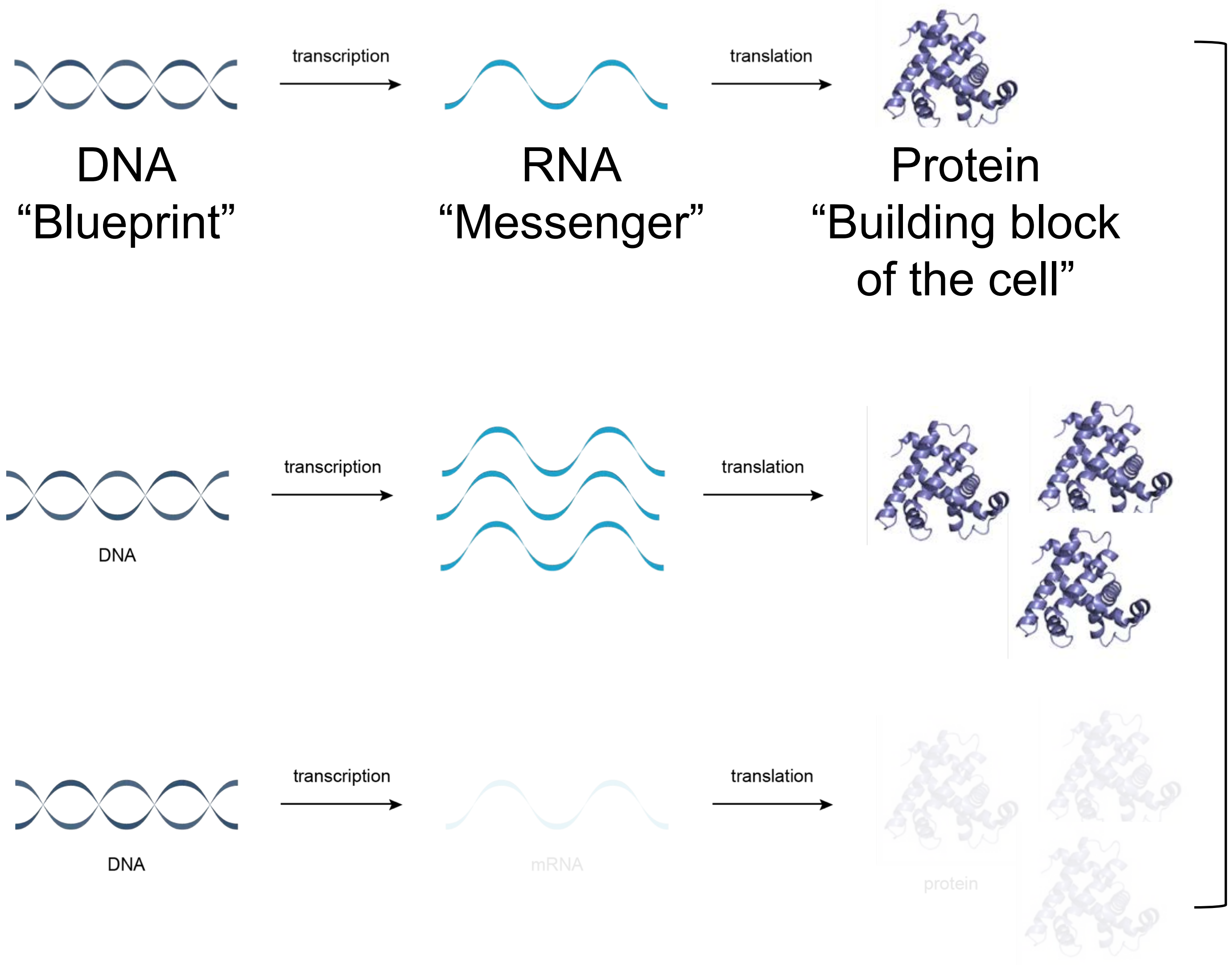
Webinar by Dr. Ildem Akerman

Society for Endocrinology, UK

UNIVERSITY OF
BIRMINGHAM

IMSR
INSTITUTE OF METABOLISM
AND SYSTEMS RESEARCH

What is gene expression ?



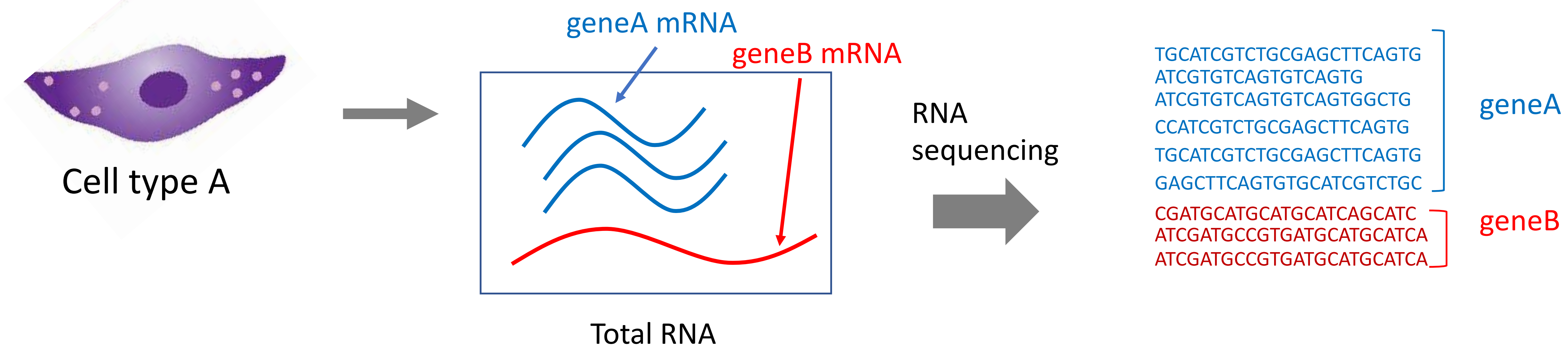
DNA
“Blueprint”

RNA
“Messenger”

Protein
“Building block
of the cell”

Differential
Gene
Expression

What is RNA-seq?

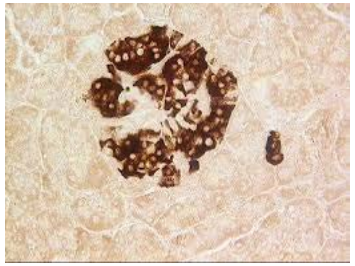


- RNA-sequencing is a method that helps quantify the amount of RNA in a given sample

Which comparisons can be made?

Normal subject pancreatic islet

Patient subject pancreatic islet
(T2 Diabetes)



Which genes / pathways are different in these cells?

How are genes specific to this tissue behave under these two conditions?

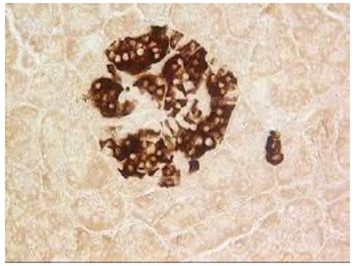
Which comparisons can be made?

Normal subject pancreatic islet

+drug

No drug

Patient subject pancreatic islet
(T2 Diabetes)



+drug

No drug

How do normal and patient samples respond to a drug/ Or other intervention.

Time course of drug/intervention response

Which comparisons can be made?

Wild type vs Knockout

Mouse models or Cell lines /models

-impact of tissue specific knockout,

-Impact of knockout on specific organs

-impact of drugs on tissues

In this online seminar series…

1. **Experimental Design**
   Design, controls, sample preparation, storage etc..

2. **RNA-sequencing technology overview**
   Library preparation & RNA-sequencing

3. **Analysis of RNA-sequencing data**
   Pipelines, how to learn

Experimental design and sample collection

Experimental design

- Number of samples (power)
- Controls and consistency between samples
- RNA extraction & library preparation
- Storage of samples

Number of samples

Cell lines

- Cell lines are isogenic (same genetic background) = Limited variation in expression

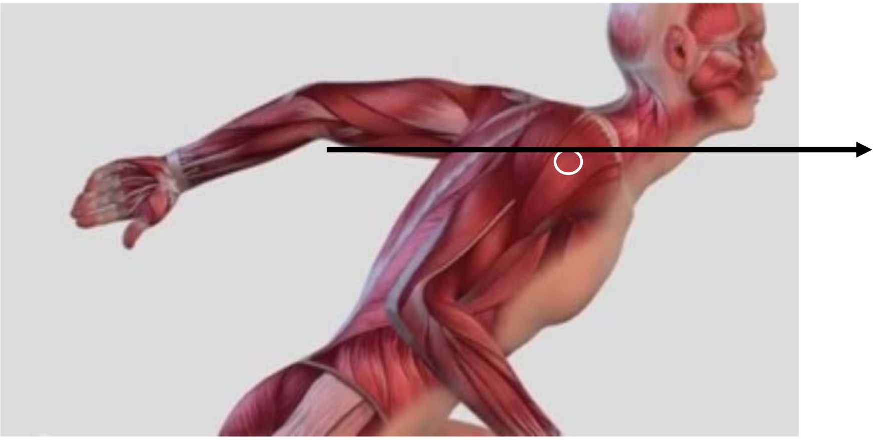
- n=3-4 if you expect large transcriptional changes i.e. transcription factor knockout

- n=5-6 if you expect subtle transcriptional changes i.e a mild drug treatment

! These are rough guidelines only: Best approach is to speak to a statistician

Mouse / Human tissue

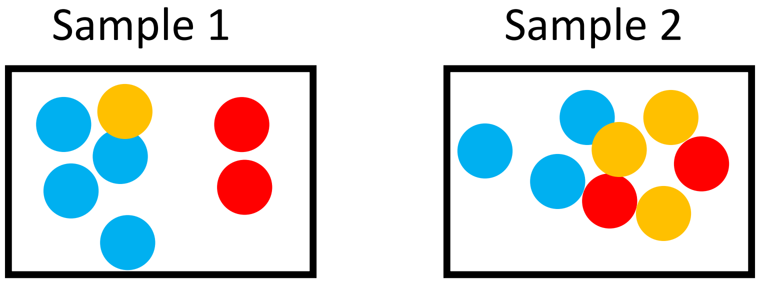
Number of samples



Muscle tissue

Muscle biopsy
- **muscle cells**
- blood vessels
- blood cells
- nerves
- connective tissue

Composition of these cells will vary slightly or significantly between samples



> “**pseudo-variations**” in gene expression

Number of samples

Tissue samples from human / animal models

- Tend to be more genetically heterogenous / environmental factors not exactly the same

- **Mouse tissue (more isogenic)**
  n=5-6 for moderate-to profound transcriptional changes
  n=8 for subtle transcriptional changes
- **Human tissue**
  n=6-8 for profound transcriptional changes
  (**a genetic disease**)
  n=8-14 for moderate transcriptional changes
  n=12-30 for subtle transcriptional changes

Controls and consistency between experiments/replicates

If you are comparing two conditions, all other conditions need to be kept the same!

1. Experimenter (nurse, surgeon isolating tissue etc..)
2. Harvest time (control and condition at the same time)
3. Time of the day (animals and some cells have a circadian rhythm)
4. Reagents (old vs fresh reagents may make a difference)
5. Duration of time that a sample has to **wait before** harvest
   (heat-shock genes are activated fast!)
6. Temperature of the environment
7. Gender (yes, sounds obvious!) , age, BMI,
   health status, race, diet, time of last food intake…
   time of menstrual cycle..
8. Sample purity

Sample collection: RNA extraction

1. Trizol/TriPure

(guanidinium thiocyanate-phenol-chloroform extraction)

Cheaper

Uniform extraction

(microRNAs and long RNAs)

Harder to use for beginners

2. Columns (Quiagen) purification kits

Easy to use

Best for beginners

Sample collection: RNA extraction

TIPS

- Some tissues (i.e. adipose) require specific extraction procedures, check before you start!
- Many tissue samples may need to be "pulverized" in LiqN2, before they can be extracted
- Both protocols also need GENOMIC DNA REMOVAL
- Never exceed column capacity/ or put too much tissue
  (each reagent will come with instructions)
  ~ sesame – rice grain of tissue/cells per 1 ml Trizol.
- Extract controls + samples together, don't overload!
  12-16 at a time..

Experimental design

Storage of samples after harvest

- **RNA-later** will keep most samples intact until harvest
(days at room temperature, weeks in fridge, months in freezer)

-Samples can be homogenized immediately and kept in **TRIZOL / or Quiagen** buffers at
-80.

-**RNA is always stored at -80, and shipped on dry ice**).

-All samples must be harvested and
stored the same way.

RNA sequencing overview

RNA quality controls

RIN : RNA integrity number

Once the cell structure is compromised, RNAses start to degrade the RNAs present in your sample

RIN measures the ratio of the two major RNA species in your sample (ribosomal RNAs)

RIN measure of 9-10 is excellent samples can be used for sequencing

RIN measure of 8 is usually acceptable, and can be used for sequencing

RIN measure of 6/7 is borderline – some facilities do not take RNA with this RIN.

RIN <=6 means your RNA is degraded:

However, some library prep kits will still accept RINs 2-6

(**i.e. Lexogen Quantseq**)

Library preparation

Strategy	**Type of RNA**	**Ribosomal RNA content**
Total **RNA**	All	High
PolyA selection	Coding	Low
rRNA depletion	Coding, noncoding	Low
RNA capture	Targeted	Low

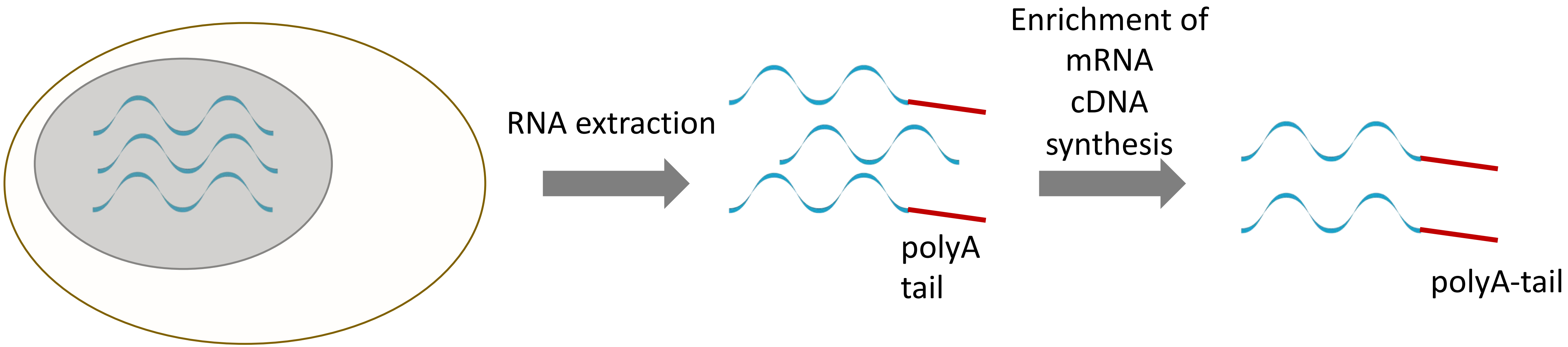
TIP: The choice of kit depends on **quantity and quality** of RNA you have. i.e. True-seq, Ultra low truseq, Lexogen Quantseq3 etc…

Best to discuss with your genomics facility

- 95% of the RNA isolated from any cell type will be ribosomal RNA. **Thus mRNA needs to be enriched.**
- We cannot sequence RNA directly, it needs to be converted to DNA
- Finally, DNA has to be amplified (expanded, more copies made) so that we can sequence it.

All of these steps are done as a part of a “library preparation kit”, usually by the facility.

Library preparation



RNA extraction

polyA tail

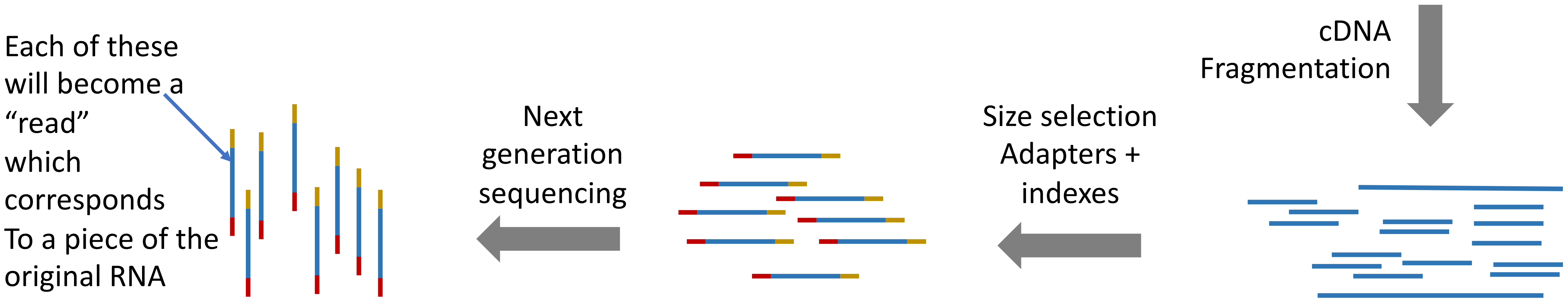
Enrichment of mRNA
cDNA synthesis

polyA-tail

- ➢ Since we are interested in differential gene expression, mRNA is enriched (<95% is mRNA, the rest is ribosomal RNA)

Ribosomal RNA depletion or
enrichment of mRNA (polyA tails).

RNA-sequencing methodology: Library preparation / sequencing



cDNA Fragmentation

Size selection Adapters + indexes

Next generation sequencing

Each of these will become a “read” which corresponds To a piece of the original RNA

- ➢ RNA is converted to cDNA

- ➢ Each sample gets a small DNA barcode called an **index** to identify which DNA fragments come from which sample.

- ➢ Usually samples are then “pooled” and sequenced.

https://www.youtube.com/watch?v=fCd6B5HRaZ8

Terminology & recommendations

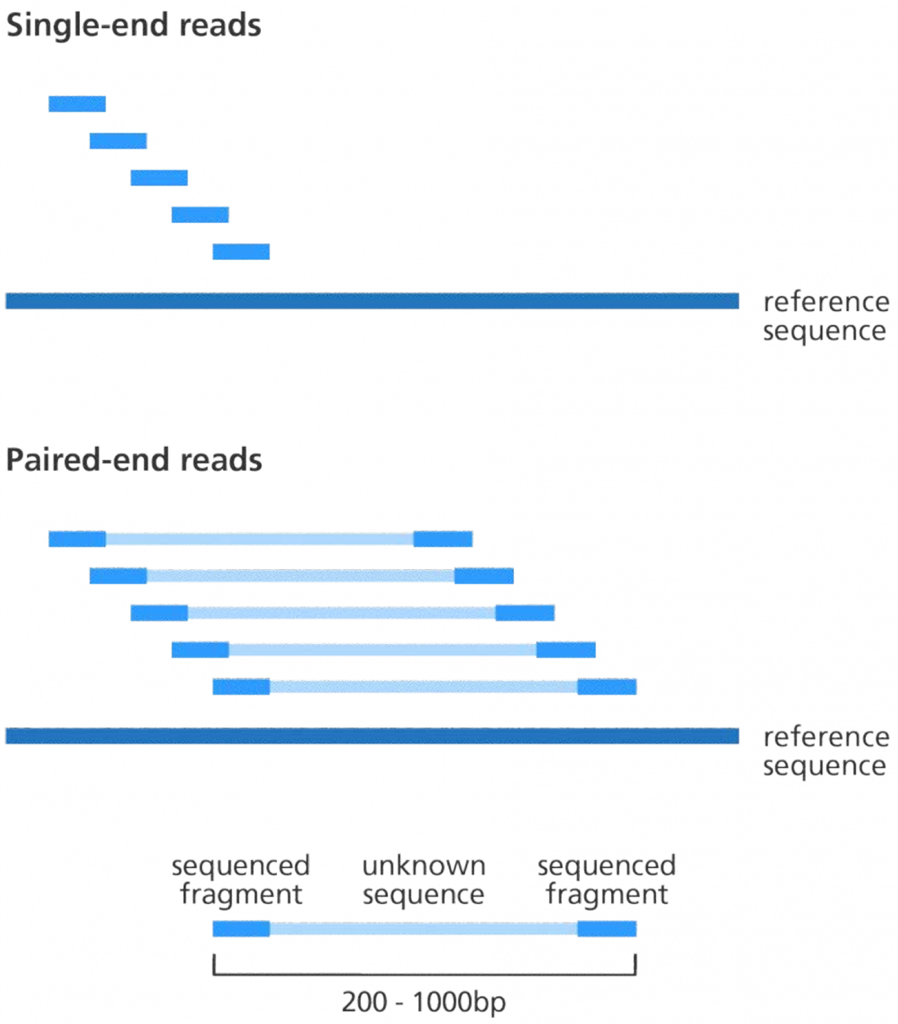
Read depth: refers to the number of reads obtained per sample

Read Depth (million reads/sample, on average)	Application
25-35 M	RNA-sequencing for differential gene expression
3-4 M	Lexogen Quantseq3 RNA-sequencing for differential gene expression ***
20-50 M	ChIP-seq (TF or histone mark?)
150 M	ATAC-seq

Read Length: refers to how much of the DNA fragments in the library are sequenced.; size of read.

above 50bp usually suitable for DGE

Terminology & recommendations



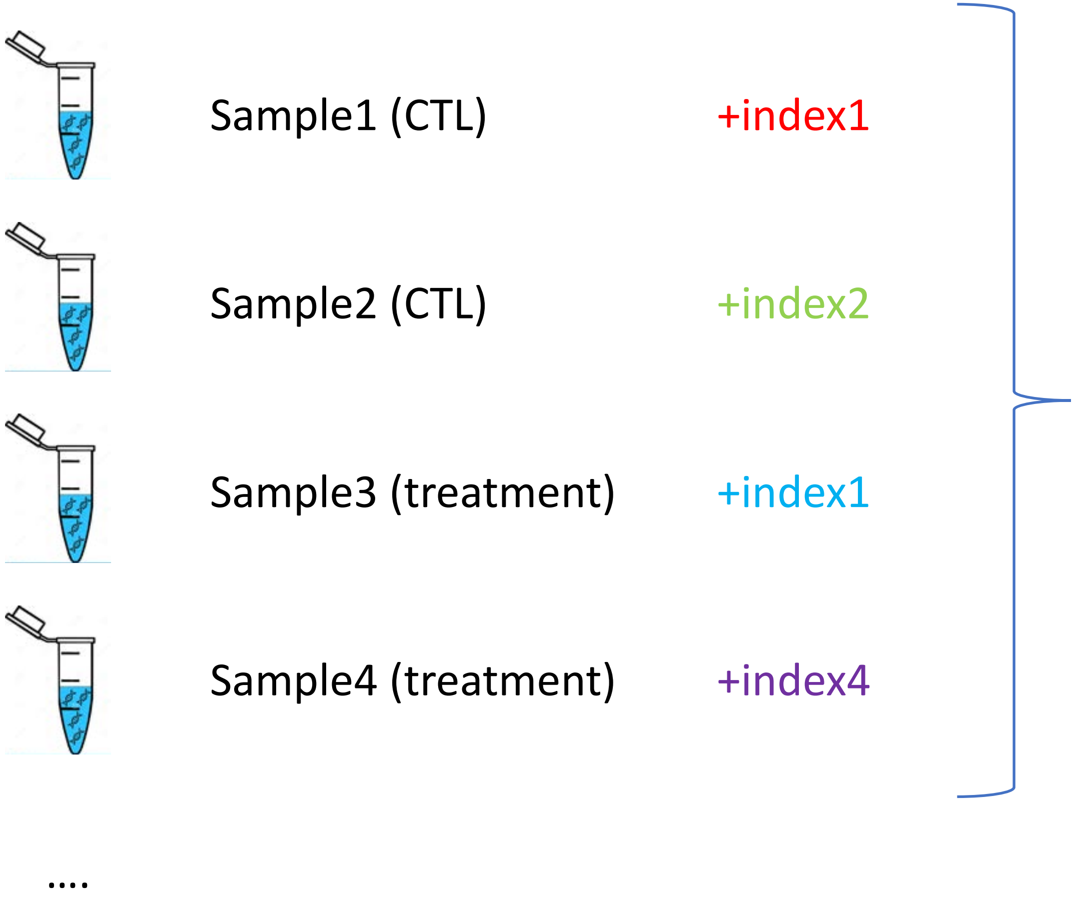
For differential gene expression, **single end** will work perfectly well…

Paired end: usually used for assembling genomes/ transcriptomes. (when one does not know the sequence of the genes in a cell, i.e. long non coding RNAs).

https://www.biostars.org/p/267167/

Your RNA-seq experiment for differential gene expression analysis

- ➢ 25-35M reads / sample
(or 3-4M for Quantseq, the cheaper option)
- ➢ 75 bp, single end reads


Sample1 (CTL) +index1

Sample2 (CTL) +index2

Sample3 (treatment) +index1

Sample4 (treatment) +index4

....

Mixing of the libraries
(Not the samples!!)

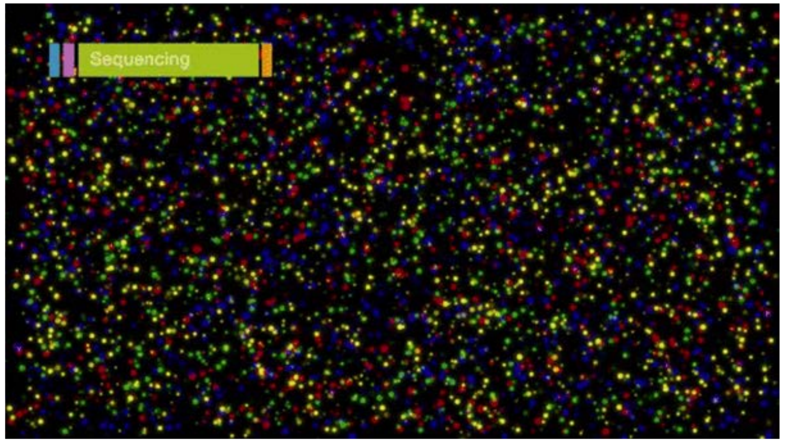
”pooling”

Analysis of RNA-sequencing data

Differential gene expression analysis

Differential gene expression analysis from RNA-seq data

Data analysis



illumina flowcell image

a Fastq file

```
TGCATCGTCTGCGAGCTTCAGTG
ATCGTGTCAGTGTCAGTG
ATCGTGTCAGTGTCAGTGGCTG
CCATCGTCTGCGAGCTTCAGTG
TGCATCGTCTGCGAGCTTCAGTG
GAGCTTCAGTGTGCATCGTCTGC
```

geneA

```
CGATGCATGCATGCATCAGCATC
ATCGATGCCGTGATGCATGCATCA
ATCGATGCCGTGATGCATGCATCA
```

geneB

25 Million reads!

Analysis pipeline

1	Fastq file (reads + quality scores)	A list of all the sequences from your sample + quality scores
2	Quality Controls Trimming	Any low quality bases? Any repeated reads? Any adapters and indexes NOT chopped off by the facility?
3	Alignment to the genome	Where in the genome do my reads come from?
4	Quantification	How many reads fall onto each gene?
5	Differential gene expression	How many of the genes have statistically different number of reads on them between control and my treatment samples?

Analysis pipeline

		Bioinformatic tools to use:
1	Fastq file (reads + quality scores)	
2	Quality Controls Trimming	FastQC (quality testing) Trimmomatic, TrimGalore, Bbduk etc (many exist!)
3	Alignment to the genome	For mRNA: **STAR** aligner (need to map splice junctions)
4	Quantification	**HTseq** / CountFeatures
5	Differential gene expression	**DeSeq2**, EdgeR, Limma…

Analysis pipeline

		File types generated
1	Fastq file (reads + quality scores)	.fastq Raw reads file
2	Quality Controls Trimming	.fastq Raw reads file
3	Alignment to the genome	.bam or .sam alignment file
4	Quantification	.htseq or .txt a text file with counts
5	Differential gene expression	list of genes usually can open in excel

how to learn?

1. Command line

- FASTER

-requires understanding of programming – takes long to learn (1 week course)

-best control over parameters

2. Online tools (usegalaxy.org)

-Slow, but usually manageable for few samples

-Easy to learn (follow online tutorial)

-Medium control over parameters

3. Pre-designed packages

-Fast and easy to use

-usually paid, but some come with library kit!

4. Collaborator / Bioinformatic company / facility

how to learn?

1. Command line

Online tutorials + Biostars.org is a great source for questions…

[Introduction to differential gene expression analysis using RNA-seq](http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf)

http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf

2. Online tools (usegalaxy.org)

usegalaxy.org

usegalaxy.eu (sometimes faster)

3. Pre-designed packages

Short video tutorials

i.e. Partek package

Quick look at the pipeline!

Differential gene expression analysis from RNA-seq data

“N” when it cannot decide which base it is

```
GGGCANCTTCCGGGAAACCAAAGTCTTTGGGTTCCGGGGGGANTATGGTTGCAAAAAAAAAAAAAAAAAAA
```

Sequence of the read

```
AAAAA#EEEAEEEEEEEEEEEEEEEEEE/EEEEEEEE#EEEA##AEEEEEEEEAEA/AEE//#<E<EE#//E<E
```

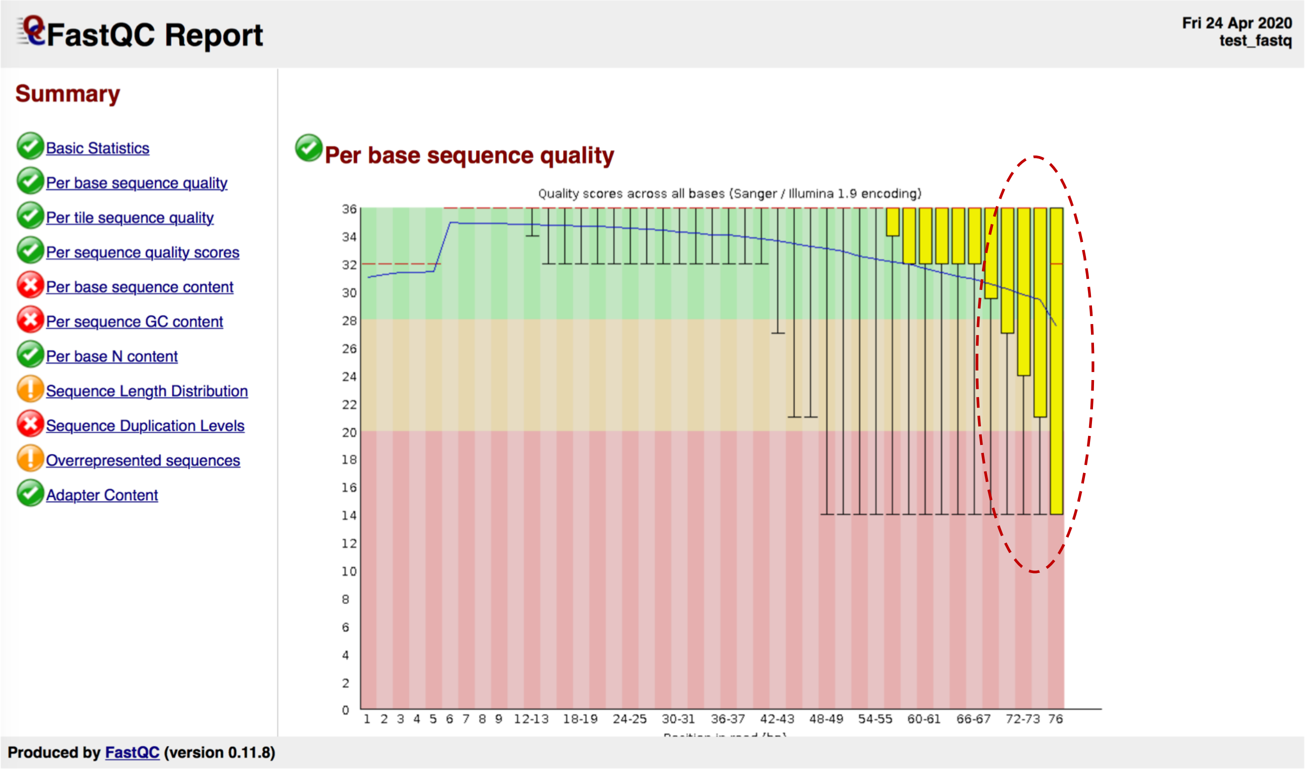
Quality score of the read

```
@NB501803:64:HVVN7BGX5:1:11101:12328:1062 1:N:0:ATACTG
```

Information about the read (i.e. index)

A typical “read” from illumina above…

1. Quality control: FastQC
2. Trimming



FastQC Report

Fri 24 Apr 2020
test_fastq

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content

✓ **Per base sequence quality**

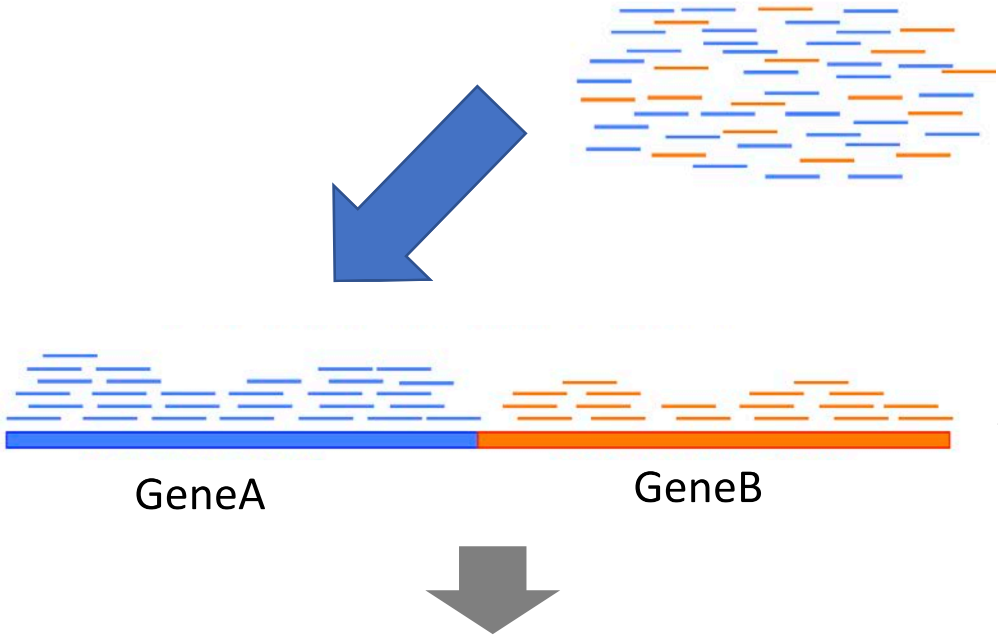
Produced by FastQC (version 0.11.8)

Last 5 bp : Low quality reads!

but still OK!

➢ If most of the read is in yellow/ red zone, ask your facility and get your money back!!

3. Alignment to genome



Reads (we know the sequence of the reads, but we do not know which gene they belong to!

The genome

ATACGAGTCTGTA   Chr1 0001000 - 0001100
ATACGAGTCTCTGGGTA   Chr1 0002000 - 0002100
….

An aligned read- we know its position in the genome !

3. Alignment to genome

- **RNA-seq reads Human / Mouse genomes**

% of reads mapped to the genome	Interpretation
>95%	Extremely good, perhaps too good?
80-95%	Very good alignment, good job trimming.
70-80%	Good
50-70%	Acceptable –may tweak trimming?
<50%	Poor sequencing/trimming. However, the data may still be usable:

!! Different genomes (Human, zebrafish, Xenopus) and different techniques (RNA-seq, ATAC-seq, ChIP-seq) may have different alignment rates!

- Do you have enough (uniquely aligned)reads aligned?

-Why do they not align?

3. Alignment to genome

- **Further reading on alignment**

https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/read-mapping-or

https://discoveringthegenome.org/discovering-genome/rna-sequencing-up-close-data/spliced-alignment

https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture_notes/STARmanual.pdf

4. Quantification

- Once we align our reads to the genome, this will result in a ".bam file"

- We now need to count all the reads that fall onto each gene using HT-seq

	Sample 1	Sample 2
Gene A	56	60
Gene B	0	0
Gene C	1203	3040
Gene D	50	50

…..

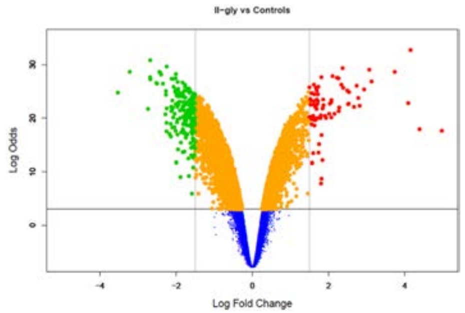
Human genome has > 18,000 genes!

- FURTHER reading: https://htseq.readthedocs.io/en/master/
  https://htseq.readthedocs.io/en/release_0.11.1/count.html

5. Differential gene expression analysis

- Differential gene expression (DGE) analysis is the application of statistical tools to determine which genes have (statistically) significant differences in expression (transcript levels) between two conditions (i.e. Control vs treatment).

  “Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution”



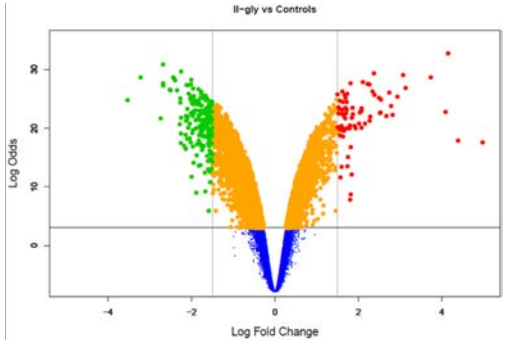
A volcano plot

- **DGE analysis tools correct for:**
  multiple testing (Benferroni).
  i.e. Given that we have a **large number of genes**, how much of the differences we see is truly significant?

  Up and down regulated genes at adj p-value < 0.05

5. Differential gene expression analysis: data visualisation

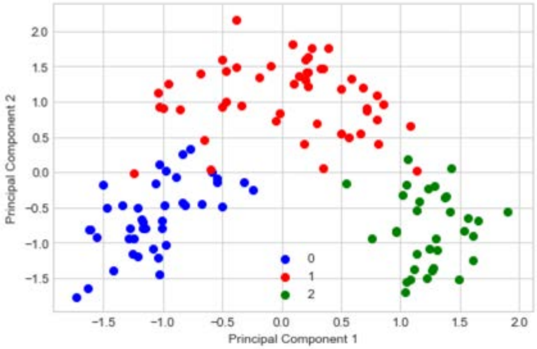
(i) Vocano plot



Customary to plot
-Log10 adj. p-value
vs
Log2 Fold change

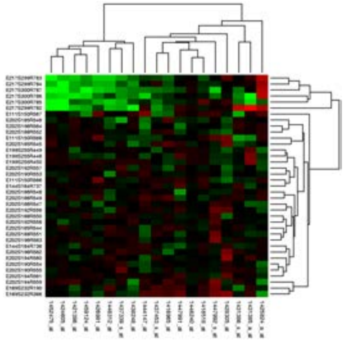
each dot is a gene, above a threshold, are differentially expressed.

(i) Principal component analysis



PCA analysis
Prin**cipal component analysis**
allows to visualize each sample in relation to other samples

(i) Gene expression heatmap



Heatmap of differential gene expression

Post-differential gene expression analysis

Downstream analysis

- **Gene ontology/ functional classification analysis**
  Isolate up and down regulated genes at adjusted p-value < 0.05 and perform a gene ontology analysis
  DAVID: https://david.ncifcrf.gov/summary.jsp
  Choose BP (biological process) or Panther/KEGG.
  There are many tools:
  https://bioinformaticsonline.com/blog/view/8798/list-of-gene-ontology-software-and-tools

- **Gene set enrichment analysis (GSEA):**
  GSEA https://www.gsea-msigdb.org/gsea/index.jsp
  Online, interface and command line

- **Other analysis:**
  -Isoform expression (not for Quantseq)
  -Transcriptional network analysis (100+ samples)
  -New gene identification
  etc….

Thanks for listening!

Dr. Ildem Akerman
Birmingham Fellow – Group leader
Pancreatic Beta Bell Gene Regulation Laboratory
Institute of Metabolism and Systems Research
University of Birmingham

@ildemAkerman
https://www.birmingham.ac.uk/staff/profiles/metabolism-systems/akerman-ildem.aspx

UNIVERSITY OF BIRMINGHAM | IMSR INSTITUTE OF METABOLISM AND SYSTEMS RESEARCH