Course overview: Introduction to single cell

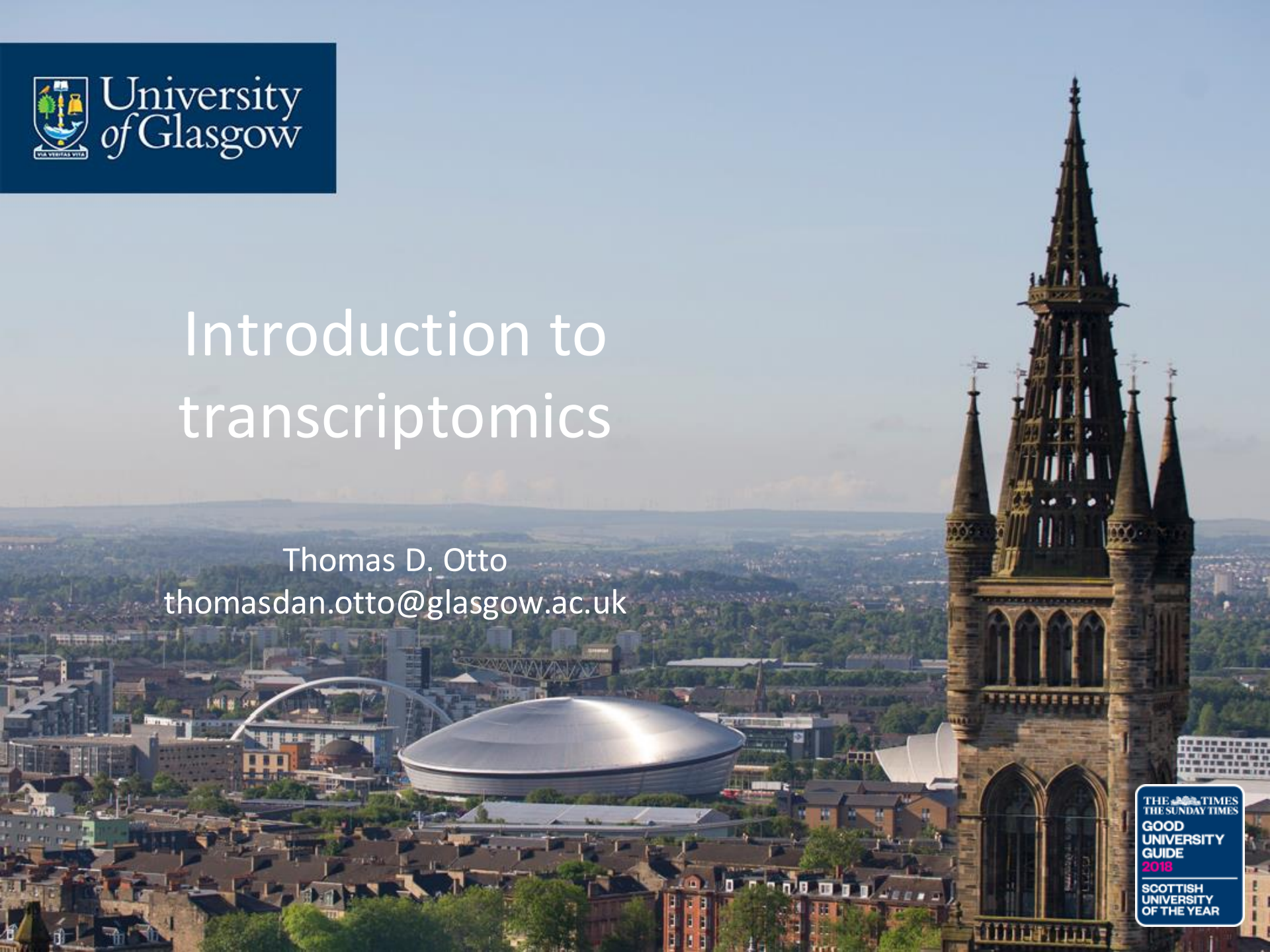
A. Introduction to the need to transcriptomics, from Qpcr to scRNA-Seq
B. Introduction to scRNA-Seq
C. Computational background
D. Pseudo time
E. More scRNA methods
F. Spatial
G. LLM & co – in exploration

Time

- Monday, Wednesday & Thursday 16-19h
- Break – when?
- Thursday 14h talk

Informal

- Small group, ask question when something is not clear


University of Glasgow
Introduction to transcriptomics
Thomas D. Otto
thomasdan.otto@glasgow.ac.uk
THE TIMES THE SUNDAY TIMES
GOOD UNIVERSITY GUIDE 2018
SCOTTISH UNIVERSITY OF THE YEAR

Learning outcomes

- Describe the benefit of RNA-Seq – when is a gene differentially expressed
- Develop a critical understanding of the need for replicates and a controlled experimental setup
- Perform (and understand) a t-test
- Reflect of the best use to represent any data
- Reflect on potential use of RNA-Seq for your future

Overview

- Why visualizing data?
- How and why to plot data
- Transcriptomics: From Q-PCR to Microarray / RNA-Seq – differential expression analysis
- Compare two conditions – what is different?

Overarching aim

- We want to compare two different samples groups, and find genes that are differently expressed – statistically significant

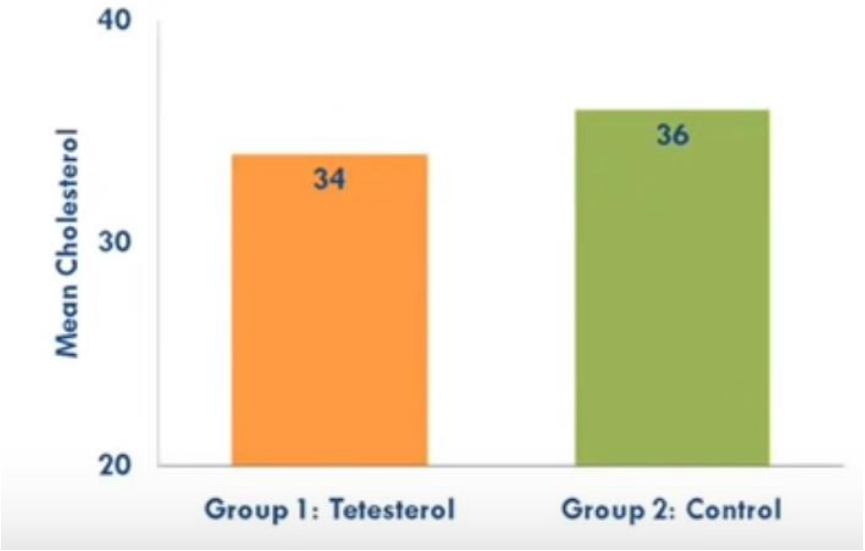
What is significance?

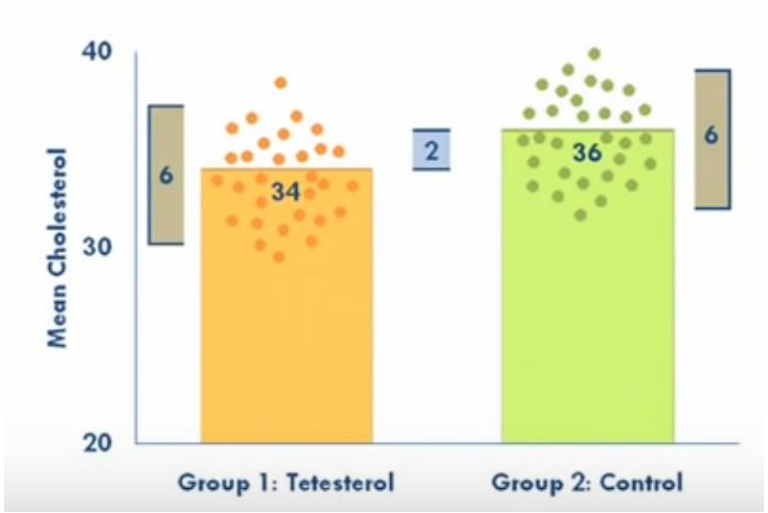
A t-test…

- … “checks” if something the average of two “groups” are “reliably” (or significantly) different
- That means, that the difference is **not** obtained by chance

https://www.youtube.com/watch?v=0Pd3dc1GcHc

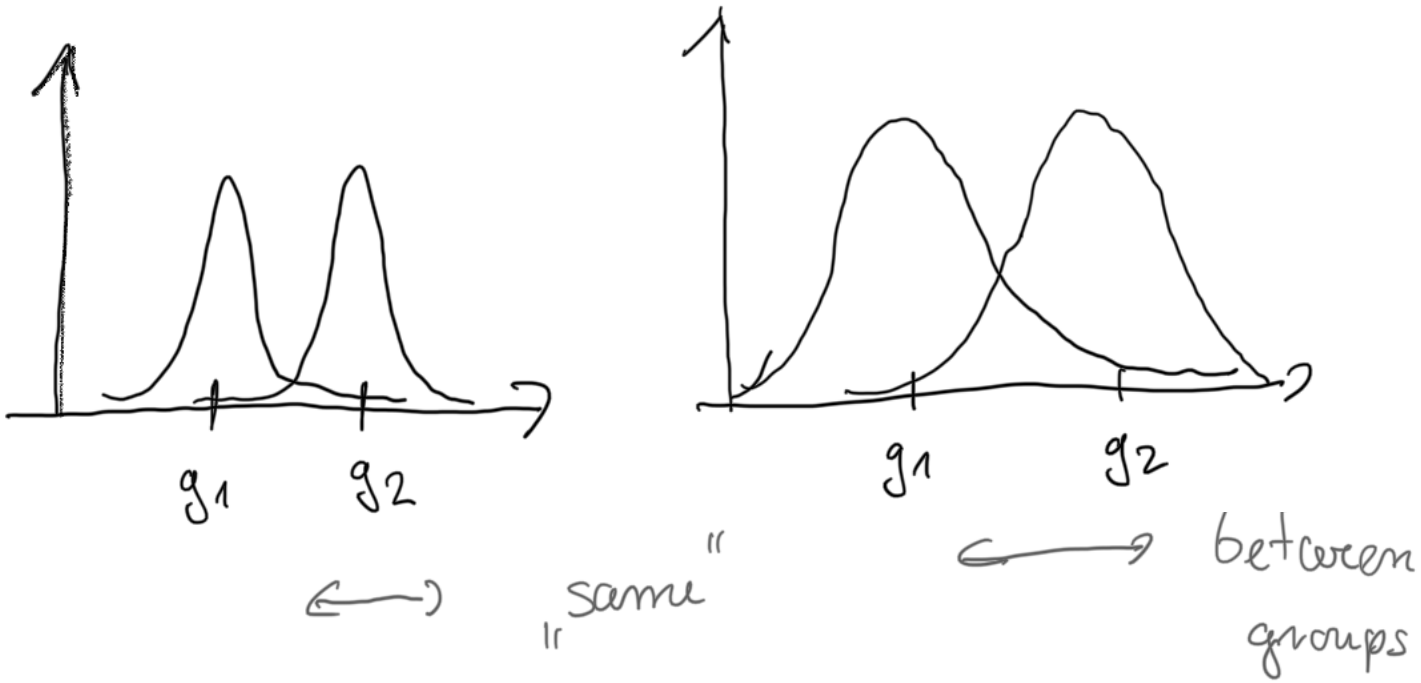
$$t = \frac{\text{variance between groups}}{\text{variance within groups}}$$





Difference between groups?

Intuition



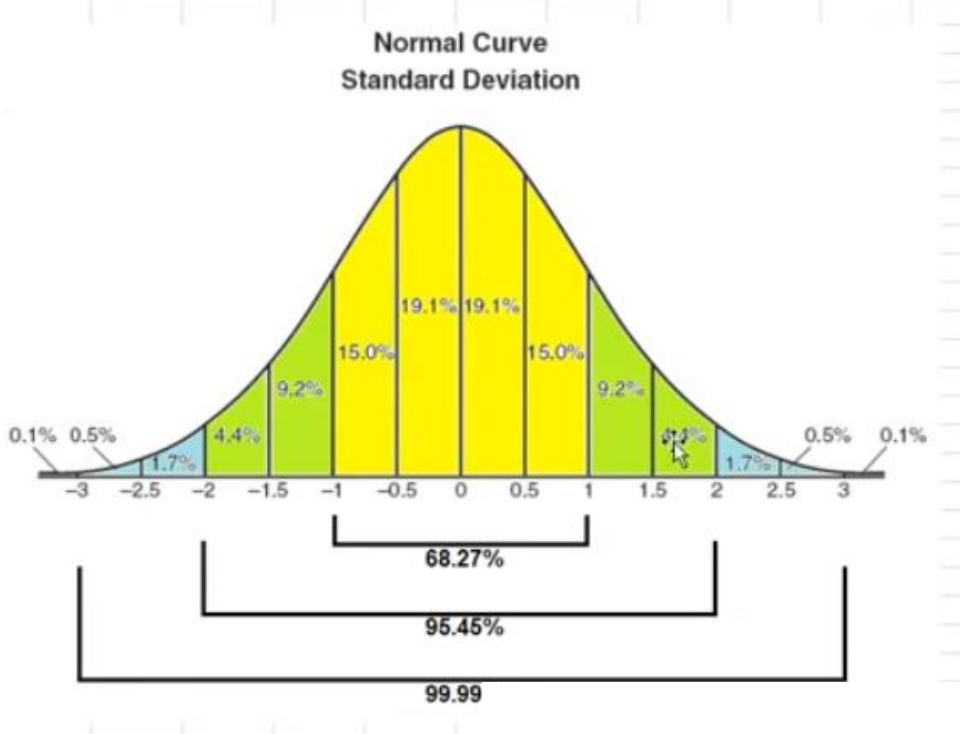
with in group?

p-value

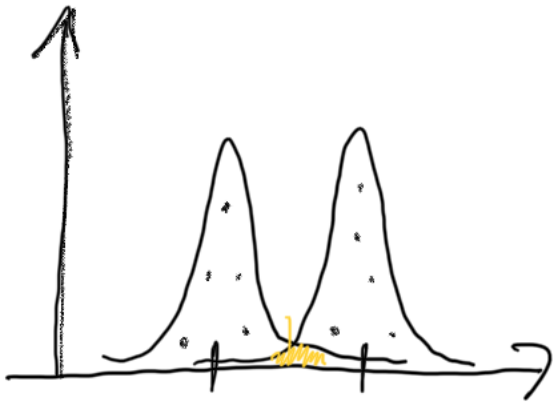
- each test has a p-value
- p-value tells us the likelihood that there is a real difference

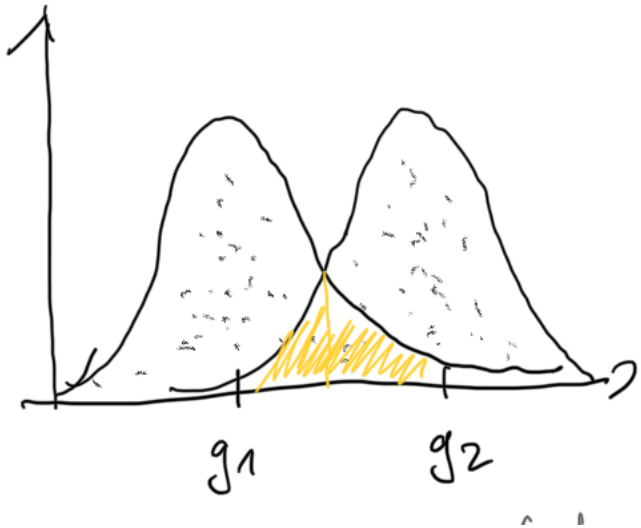
- **Specifically, the p-value is the probability that the pattern of data in the sample could be produced by random data**

- p=0.001 – there is a 0.1% chance that this result was obtained with random data



Gut feeling for P-value





95 percent confidence interval

https://www.youtube.com/watch?v=qvPWQ-e03tQ&t=64s

Sample size – biological replicates

- Why is the title in bold?
- The more the better – too small, less likely to find any difference
- “degrees of Freedome” (df) is equal the sample size minus one.

Example of T-test

In R:
t.test(data[,1],data[,2])

Welch Two Sample t-test

data: data[, 1] and data[, 2]
t = 2.4204, df = 4.3627**, p-value = 0.06745**
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.3269542 6.2469542
sample estimates:
mean of x mean of y
5.30 2.34

Limitations

1. Findings limited to type of experiment
2. Data need to be normal distributed!
3. Compare roughly that same number of datapoints for each group
4. Data should be independent to each other
5. Units to compare (not ranks)

How to write a t-test results?

"An independent-samples t-test was used to check the effectiveness of a rheumatoid arthristis drug t(99)=0.33, p=0.37, but no significant difference was found (Drug M=34; Control M=36)"

https://www.youtube.com/watch?v=0Pd3dc1GcHc

In a more formal way: t-test

- Null Hypothesis: What we are aim to disprove, that taking the drug does not change the expression of our genes of interest.
- Alternative Hypothesis: Our effect of interest, or the antithesis of the null hypothesis: The drug is working
- Significance level (α): Probability of incorrectly rejecting the null hypothesis. This is always established at the beginning of every hypothesis test. **$p<0.05$** (5% change that the result is due to chance alone)

Transcriptomics

- From Q-PCR
- to
- transcriptomics sequencing

Rheumatoid Arthritis

Rheumatoid arthritis (RA) is a long-term autoimmune disorder that primarily affects joints.

While the cause of rheumatoid arthritis is not clear, it is believed to involve a combination of genetic and environmental factors. The underlying mechanism involves the body's immune system attacking the joints.

The goals of treatment are to reduce pain, decrease inflammation, and improve a person's overall functioning.

Cascades of drugs: Steroids, disease-modifying antirheumatic drugs (DMARDs), biologic DMARDs (using biotechnology)

Wikipedia 2019

Imaging:

- A lab – where you do your master project – has a new drug that might “cure” RA!
- You have been asked to analyse Q-PCR data that where performed on one gene from a clinical trial. “*Blood treated with the drug*”
- The values are 5.5 and 1.8, for untreated and treated patient samples, respectively.
- What do you do?

What do you do?

- How could you plot the data?
- Does this experiment make any sense?

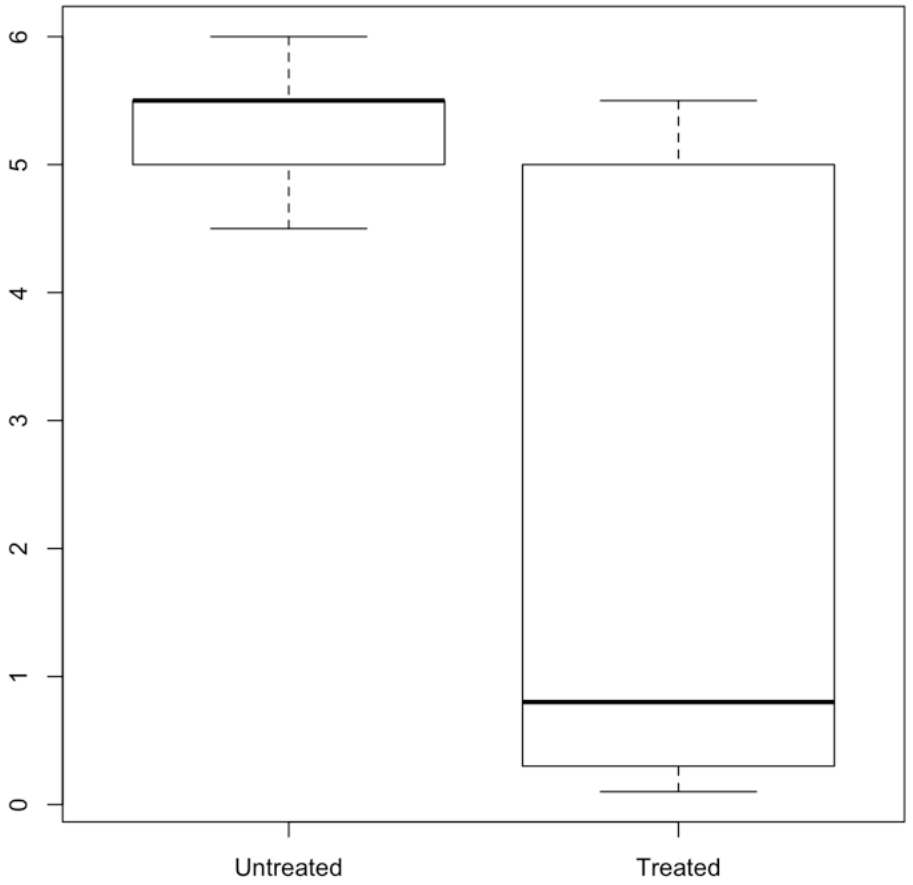
- If yes -> why?
- If no -> why not?

What about more data?

Expression fold change against 3 core genes from 10 patients	
Untreated	Treated
5.5	0.8
5	5
4.5	0.3
6	5.5
5.5	0.1

How can we re-present these?
Can we do in excel?

The boxplot

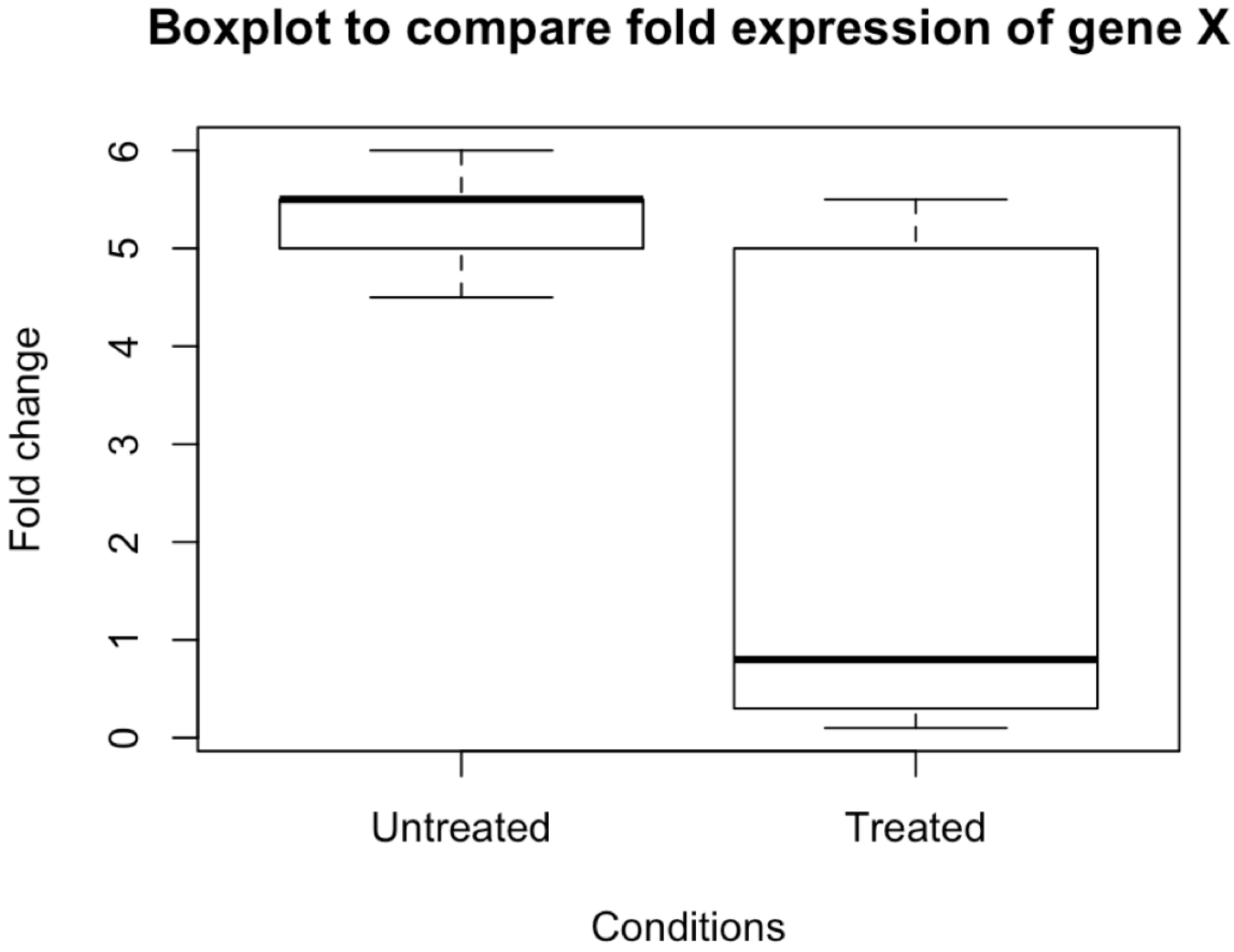


Is that a well annotated plot, yes or no?

- 1 1 1 2 100

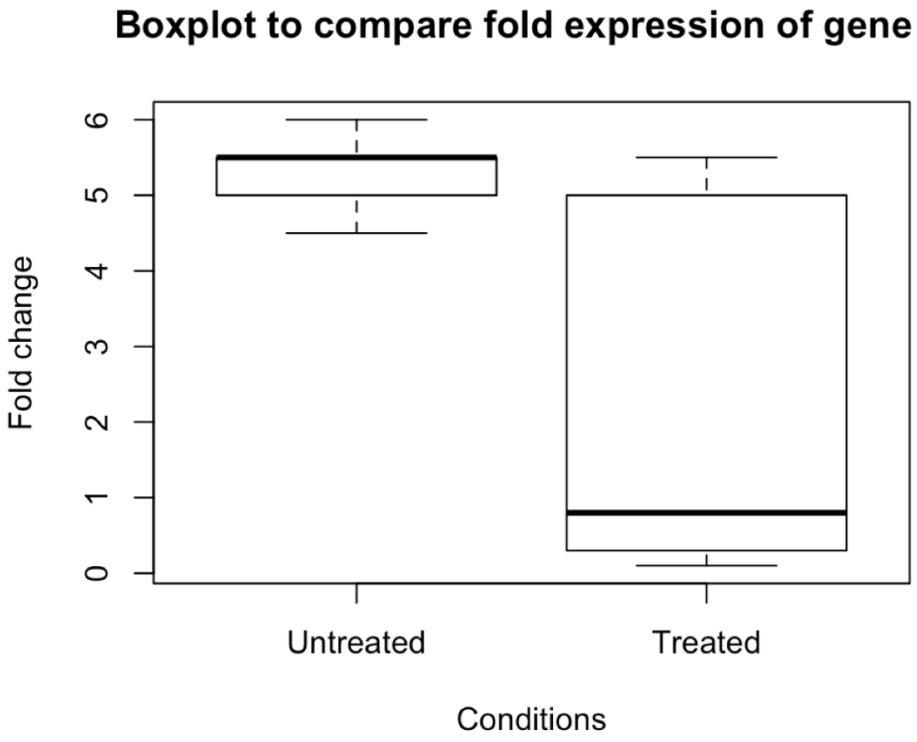
- What is the average?
- What is the median?

Is the annotation better?



boxplot(d[,1],d$Treated, main="Boxplot to compare fold expression of gene X",xlab=”Conditions",ylab="Fold change",names=c("Untreated",”Treated"))

The boxplot



In R:
t.test(data[,1],data[,2])

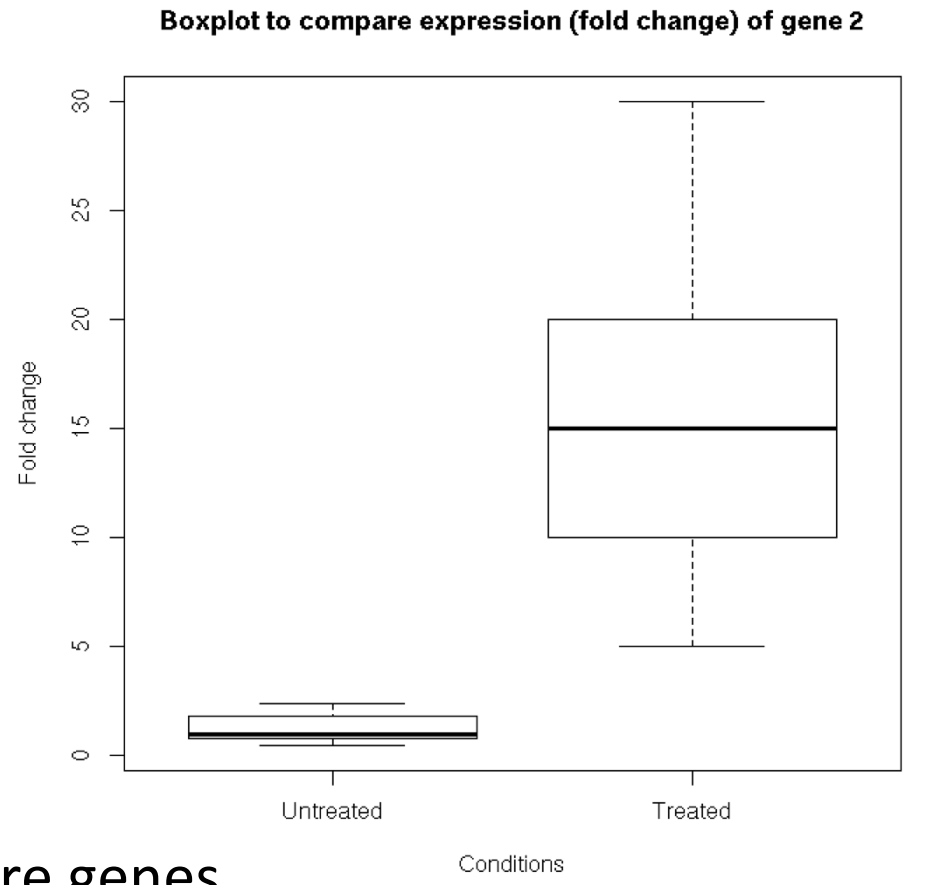
Welch Two Sample t-test

data: data[, 1] and data[, 2]
t = 2.4204, df = 4.3627**, p-value = 0.06745**
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.3269542 6.2469542
sample estimates:
mean of x mean of y
5.30 2.34

Let's take another gene

Untreated	Treated
1.8	10
0.8	15
2.4	20
1	5
0.5	30

p-value = 0.02657



- It is below **0.05**!
- So we don't have to do more genes…
- Is there something wrong here?
  - Correction for Multiple testing

Let’s split this group

- And prove that there is a real difference

Multiple testing, let's make an example

	Null hypothesis is True (H_0)	Alternative hypothesis is True (H_1)	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

Not today !

- m is the total number hypotheses tested
- m_0 is the number of true null hypotheses
- $m - m_0$ is the number of true alternative hypotheses
- V is the number of false positives (Type I error) (also called "false discoveries")
- S is the number of true positives (also called "true discoveries")
- T is the number of false negatives (Type II error)
- U is the number of true negatives
- R is the number of rejected null hypotheses (also called "discoveries")
- In m hypothesis tests of which m_0 are true null hypotheses, R is an observable random variable, and S, T, U, and V are unobservable random variables.

http://en.wikipedia.org/wiki/Multiple_comparisons_problem

Methods for multiple correction

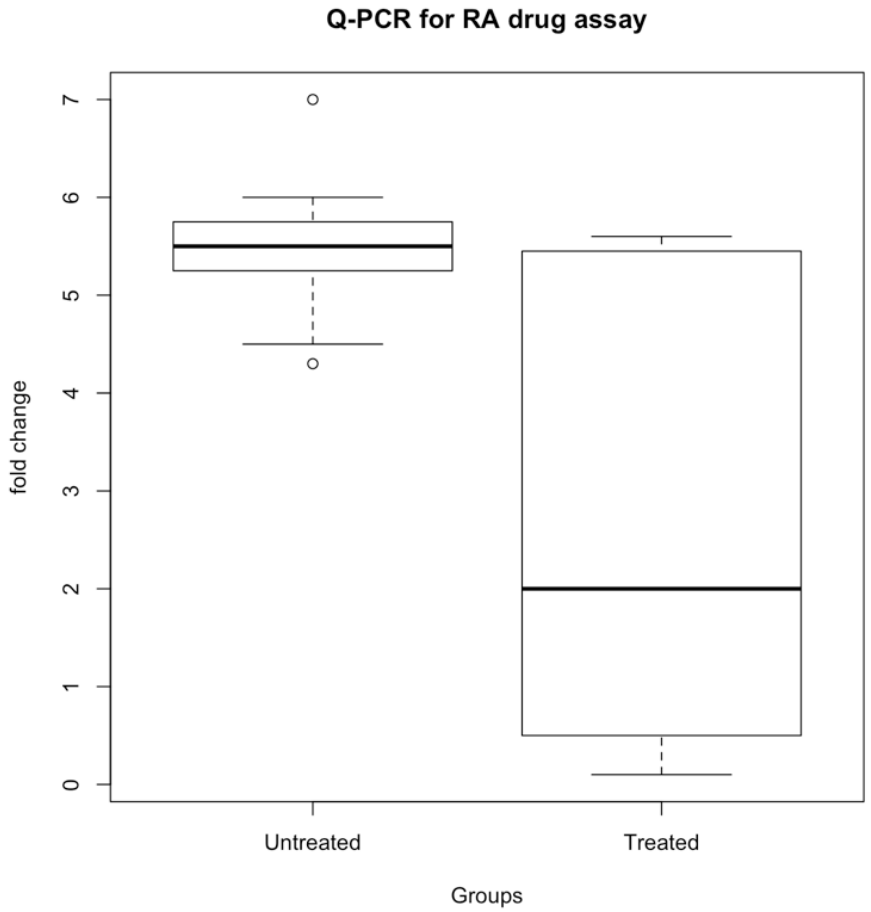
- **Bonferroni Correction**
  The most conservative of corrections, the Bonferroni correction is also perhaps the most straightforward in its approach. Simply divide α by the number of tests (*m*).
  - New α is 0.05/ 2 -> 0.025, so our test before is not significant ☹

- What can we do next?

https://towardsdatascience.com/
an-overview-of-methods-to-address-the-multiple-comparison-problem-310427b3ba92

OK, more measuring points

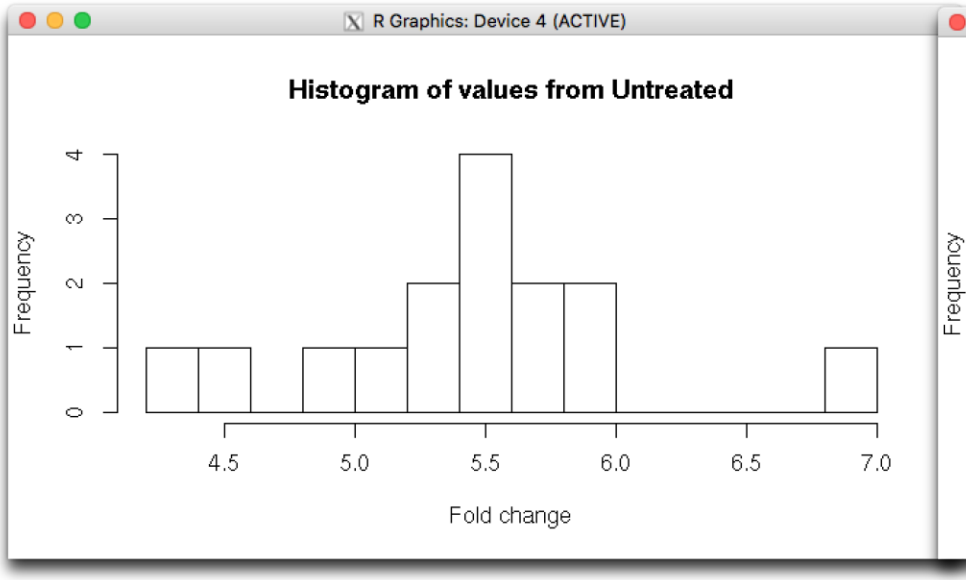
Untreated	Treated
5.5	0.8
5	5
4.5	0.3
6	5.5
5.5	0.1
4.3	5.6
5.8	0.2
5.9	5.4
5.5	0.5
7	5.5
5.3	0.5
5.6	5.6
5.4	0.6
5.7	5.2
5.2	2

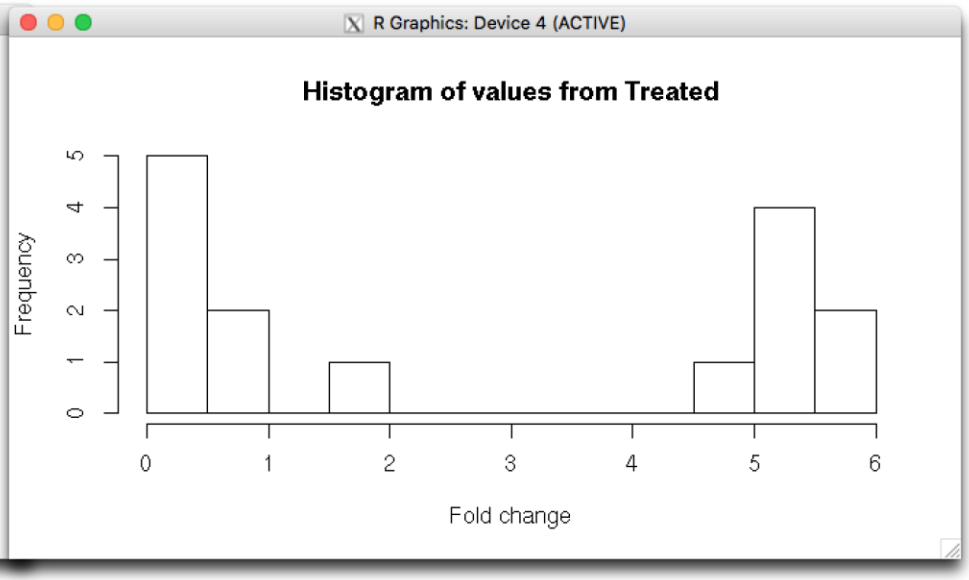


p-value = 0.001211

boxplot(data,xlab="Groups", ylab="fold change",main ="Q-PCR for RA drug assay")

Are we happy?



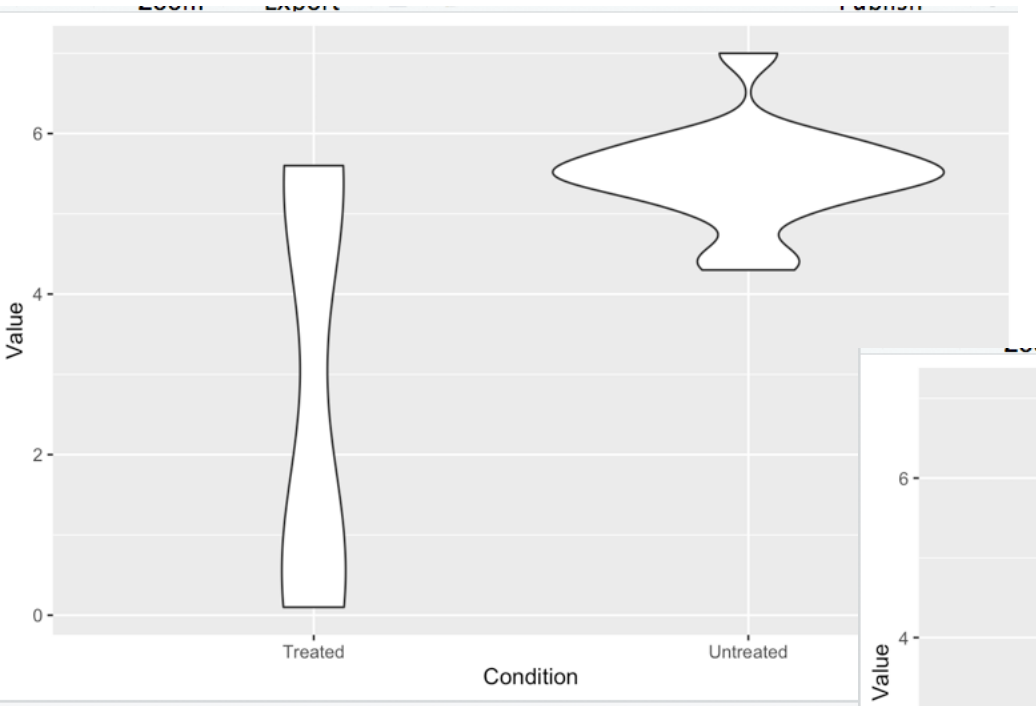


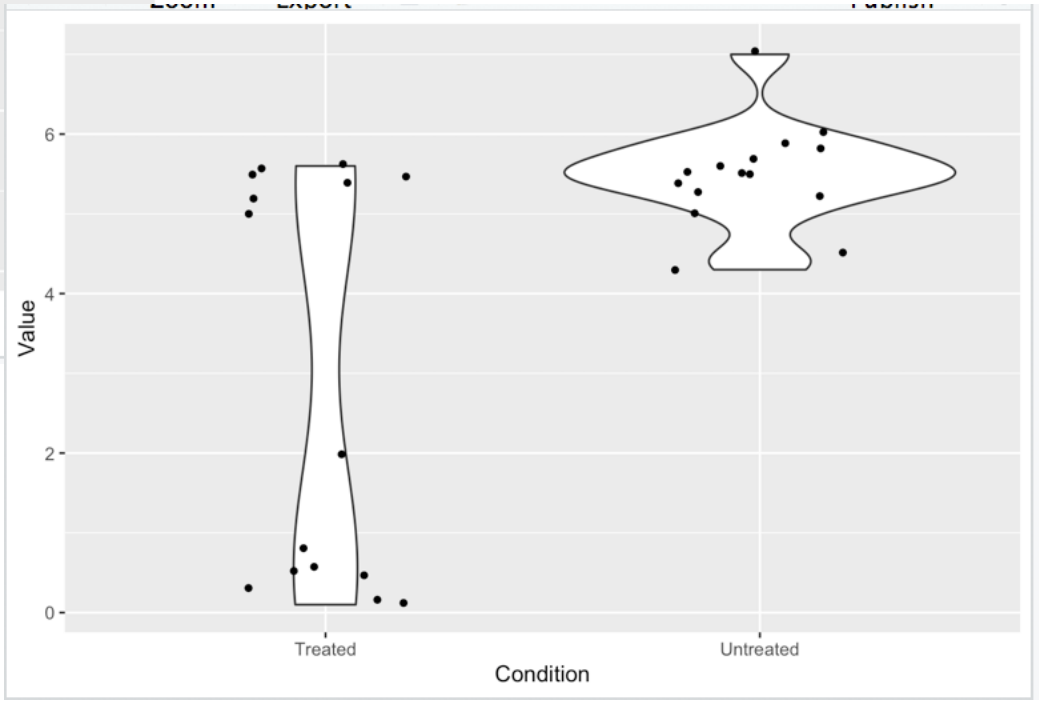
What could that immunological mean? (to see two peaks?)

Some of the individual are non-responders!

```
hist(data[,2],breaks=10,main="Histogram of values from Treated", xlab="Fold change")
```

Better: Violin plots - ggplot2






```
library(ggplot2)
data<-read.table("Set4.txt", header=T)
<- ggplot(ToothGrowth, aes(x=dose, y=len)) + geom_violin()
p + geom_jitter(shape=16, position=position_jitter(0.2))
```

Are you happy with the annotation of the graphs? No title, no y-axis label!

http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-r-software-and-data-visualization

Conclusions (2/n)

- Think about your null hypothesis and set Significance level (α)
- Look for significance (eg t-test) - have enough replicates
- Think about multiple test correction
- Visualise the data in different manners (histogram, boxplots, violin plots and many more)
- In a scientific project the **data** are the results, the heart of your experiment. The correct visualization is crucial.

Consideration

- From which part of the body did we take the sample? (Can we access all part of the body?)
- Is it a mixed or single cell type?
- Could that experiment be real?
- Include clinical data

- Are there other methods to access expression of genes?

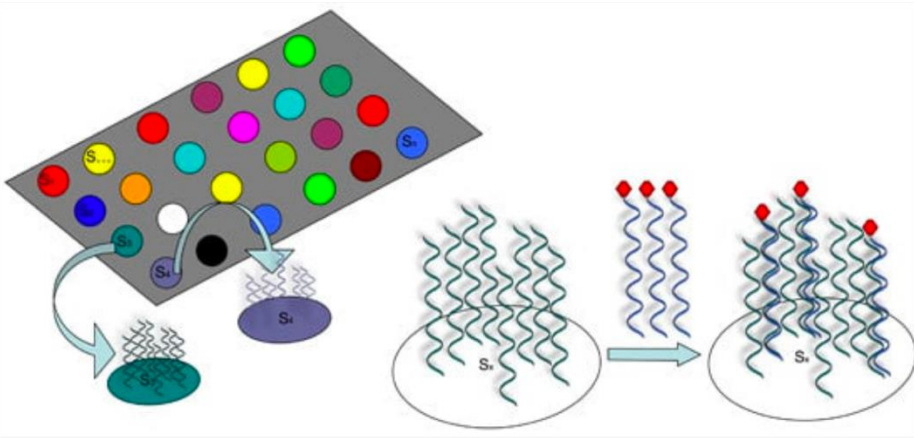
Part 2.2: Transcriptomics

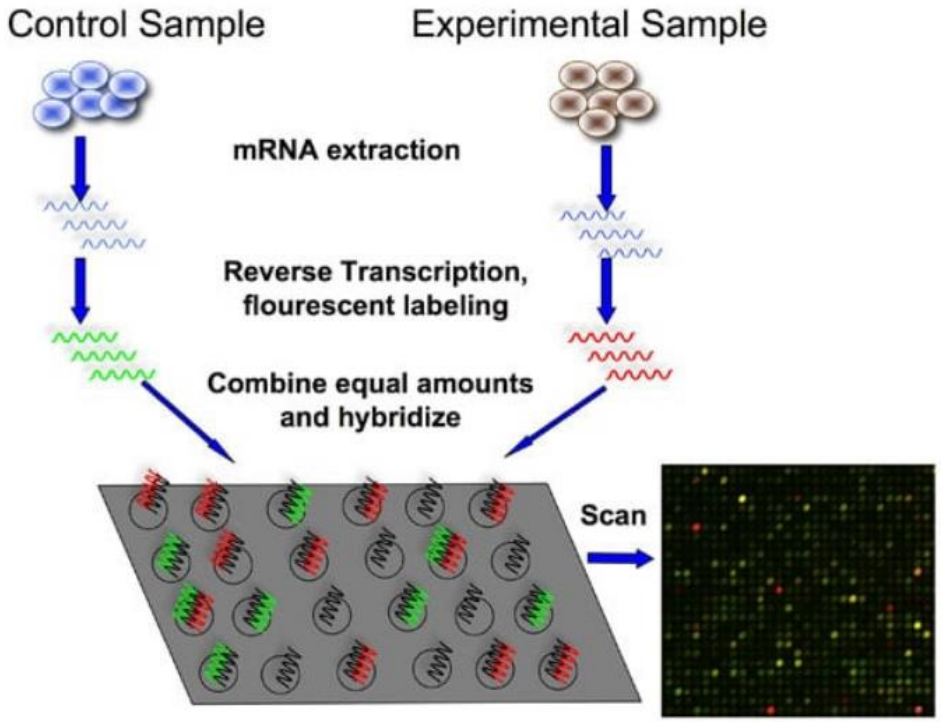
- Micro array – not doing this anymore
- RNA-Seq
  - General idea
  - Visualisation
  - Differential expression analysis in R

Why do we want to access the complete transcriptome?

Q-PCR versus microarray

- A microarray is a laboratory tool used to detect the expression of thousands of genes at the same time.





Microarray analysis is challenging

- Access expression of all genes (the one you have probes - can buy)
- Capture of signal (light) is not perfect, especial low abundance.

- With the RNA-Seq (sequencing of RNA / transcripts) most groups moved away from arrays.

Example: Identify markers and mechanisms of resistance to adalimumab therapy

Research article

Open Access

Gene expression profiling in the synovium identifies a predictive signature of absence of response to adalimumab therapy in rheumatoid arthritis

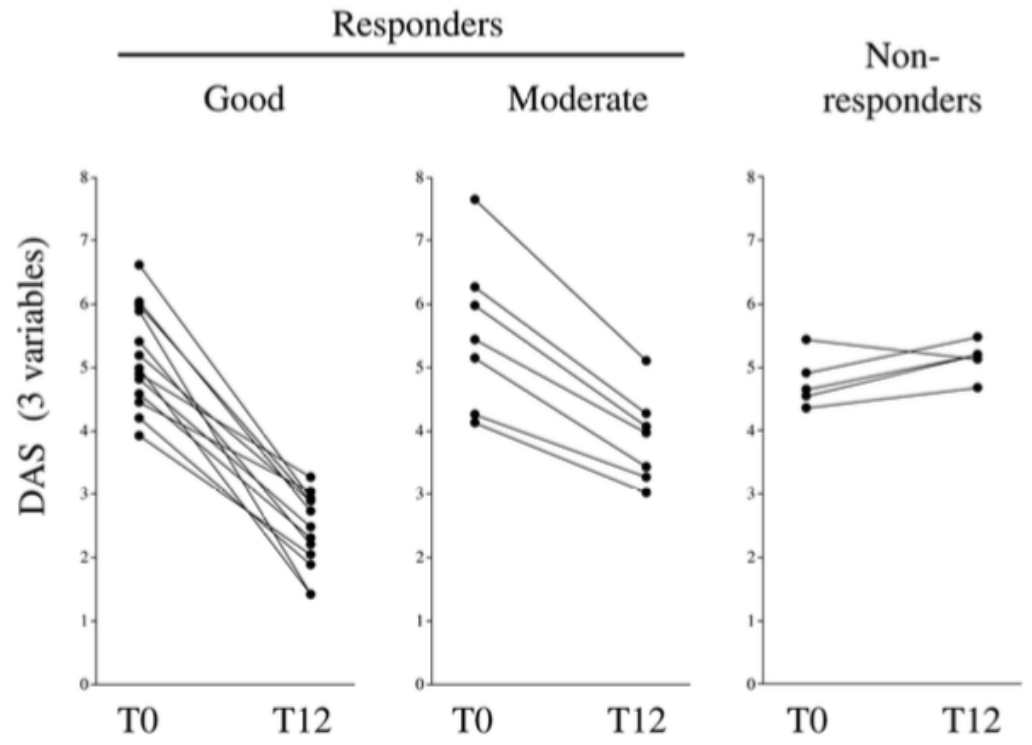
Valérie Badot[1,2], Christine Galant[3], Adrien Nzeusseu Toukap[1], Ivan Theate[3], Anne-Lise Maudoux[1], Benoît J Van den Eynde[4], Patrick Durez[1], Frédéric A Houssiau[1] and Bernard R Lauwerys[1]

- Paired synovial biopsies were obtained from the affected knee of 25 DMARD (disease-modifying antirheumatic drug)-resistant RA patients at baseline (T0) and 12 weeks (T12) after initiation of adalimumab therapy.

(PMID:19389237)

Group patients by responder

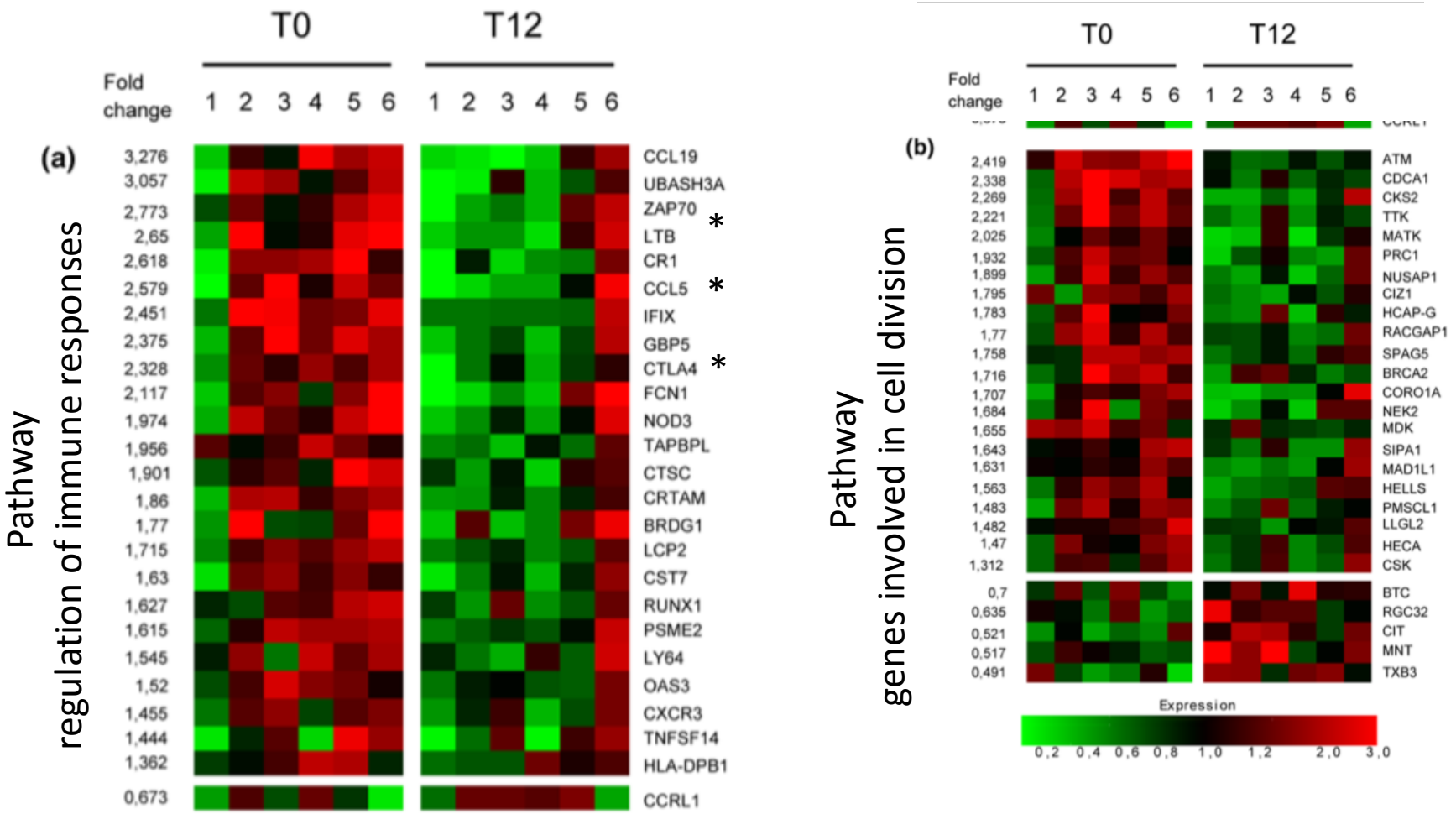
Figure 1



Evolution of disease activity score (DAS) (three variables) in 25 individual rheumatoid arthritis patients before (T0) and 12 weeks after (T12) initiation of adalimumab therapy. Patients are categorized into (good or moderate) responders or non-responders according to European League Against Rheumatism criteria.

- New type of graph
- Baseline versus later time point

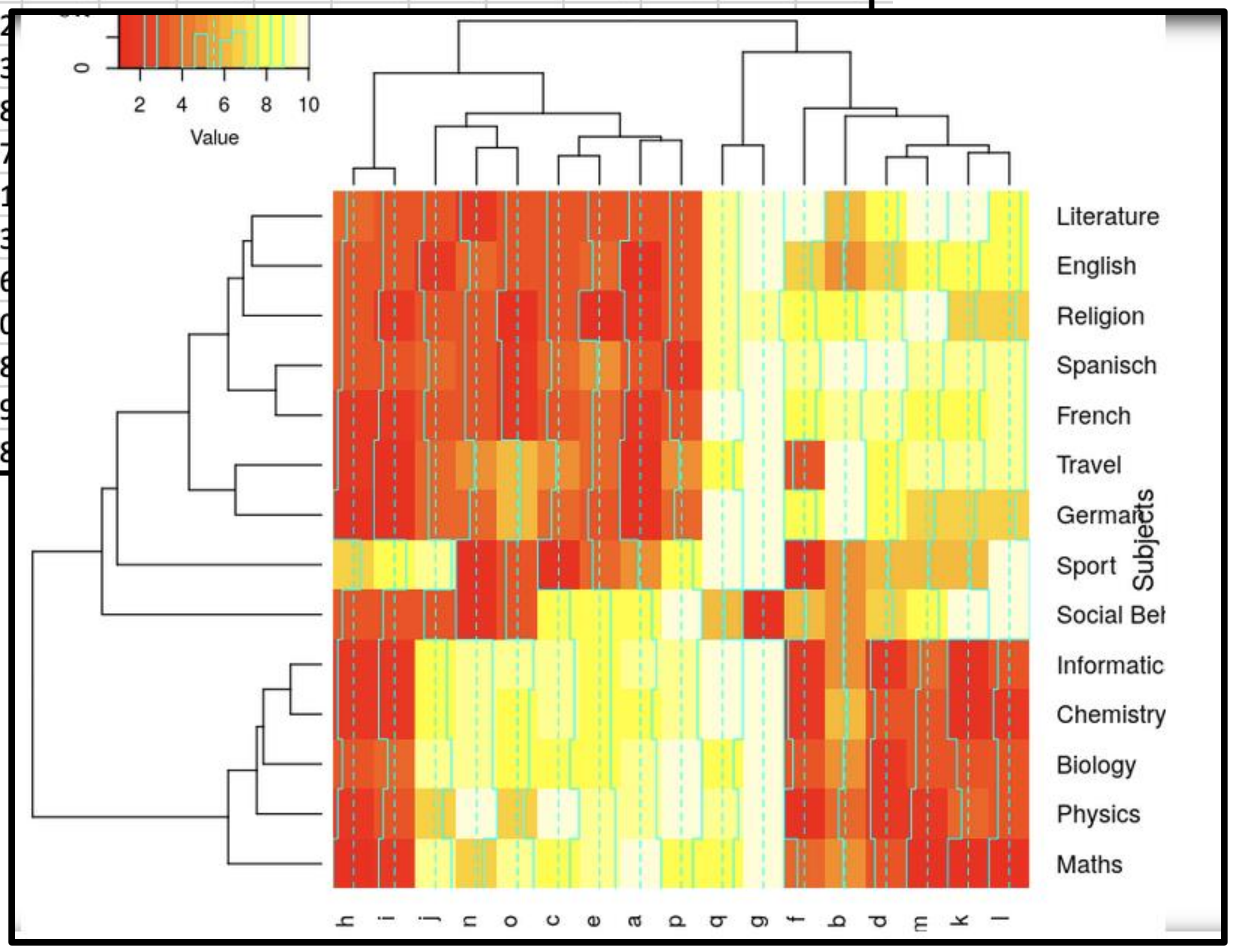
Differentially expressed genes before (T0) and 12 weeks after (T12) start of adalimumab in synovial biopsy specimens of rheumatoid arthritis patients who responded to therapy.



Paired Student *t* tests indicated that 632 (out of 54,675) genes displayed significant differences in expression between T0 and T12 in six synovial tissue samples obtained from RA patients who responded to adalimumab therapy

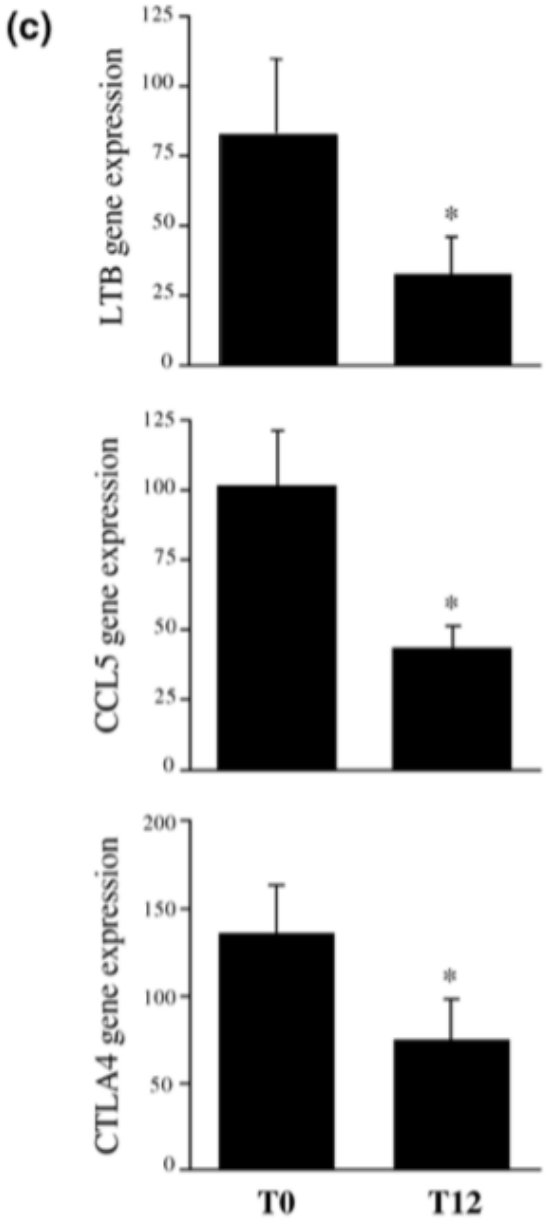
Do you know what a heatmap is?

Student	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q
Maths	10	5	8	3	9	4	10	1	2	9	1	1	1	7	9	8	8
Informatics	9	5	9	2	8	2	10	2	2	8	1	3	4	9	9	9	10
Physics	9	4	10	2	9	1	10	2	3	7	4	3	2	10	7	10	9
Chemistry	8	6	9	3	8	2											
Biology	9	5	8	2	8	3											
German	1	10	4	8	3	8											
English	1	5	3	7	4	7											
Sport	5	5	1	6	4	1											
Travel	1	10	5	8	4	3											
Social Behav	8	5	8	7	8	6											
Literature	3	6	3	8	3	10											
French	2	9	3	9	4	8											
Spanisch	3	10	4	10	5	9											
Religion	2	8	3	9	1	8											



Or explain better?

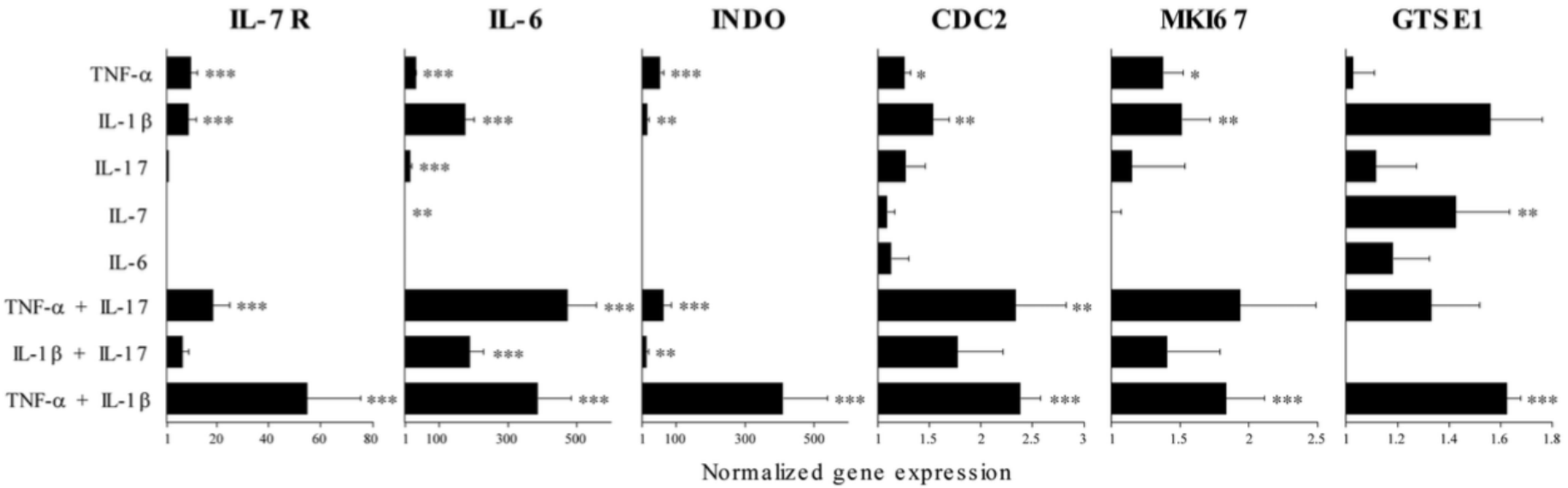
- /Users/thomasdan.otto/Library/CloudStorage/OneDrive-UniversityofGlasgow/Teaching/OtherWorkshops/2025_Uruguay_SC/Data/



Q-PCR confirmation

- QPCR (T0) (n = 10) and 12 weeks after (T12) (n = 8) initiation of adalimumab therapy
- Samples were loaded in triplicate
- $*P < 0.05$.
- CCL5, chemokine lig- and 5; CTLA4, cytotoxic T-lymphocyte-associated antigen 4; LTB, lymphotoxin beta.

fibroblast-like synovial cells cultured with of different cytokines



- evaluated in at least four different experiments
- mean fold change in gene expression and standard error of the mean, relative to the mean gene expression of the baseline condition normalized to
- $*P < 0.05$, $**P < 0.005$, $***P < 0.0005$ using Wilcoxon signed rank test

Summary of paper

- identified baseline markers of response to TNF blockade in a group of RA patients treated with adalimumab
- genes overexpressed in the poor responders are induced by TNF-α, but also by IL-1β
- “initiate larger studies in order to confirm the prognostic value of our markers “

Summary

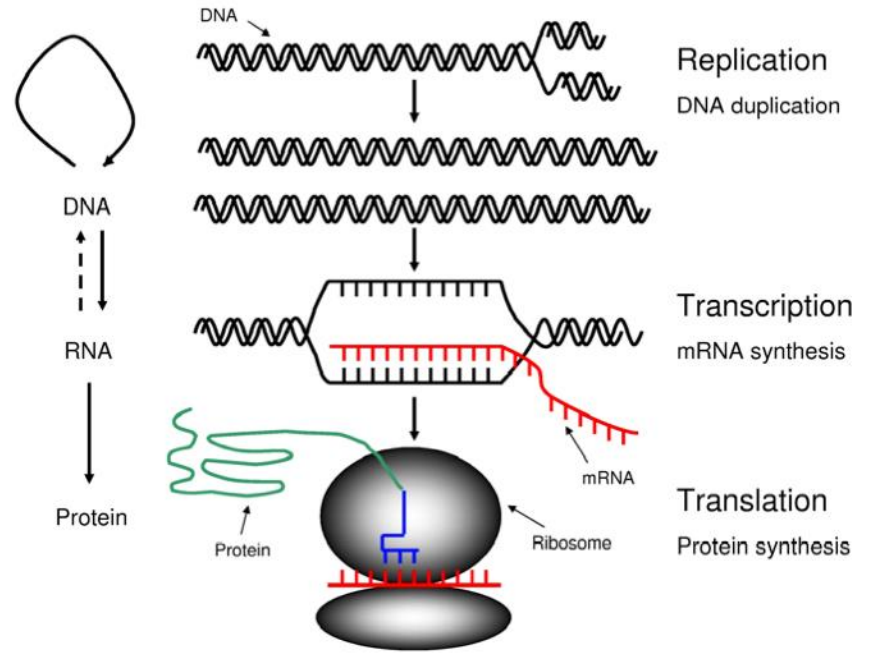
- Microarray analysis is challenging, need of many replicates
- Access expression of all genes (the one you have probes - can buy)
- Capture of signal (light) is not perfect, especial low abundance.

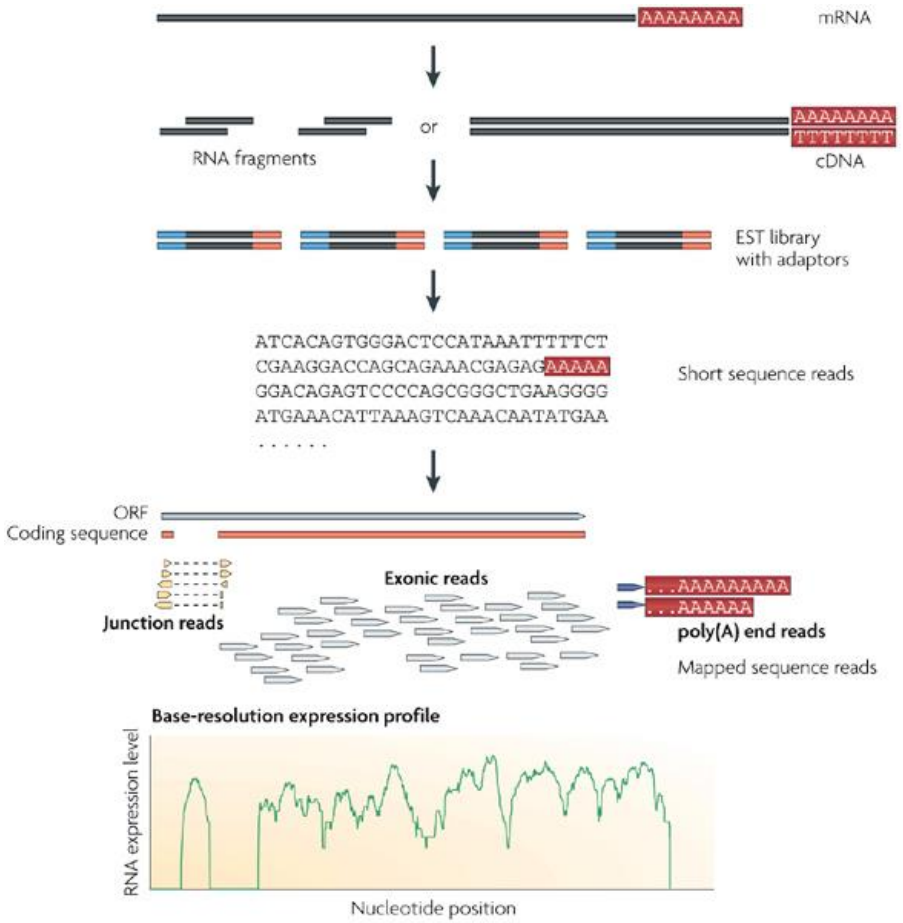
- With the RNA-Seq (sequencing of RNA / transcripts) most groups moved away from arrays.

Probes / Antibodies

- Concept will return

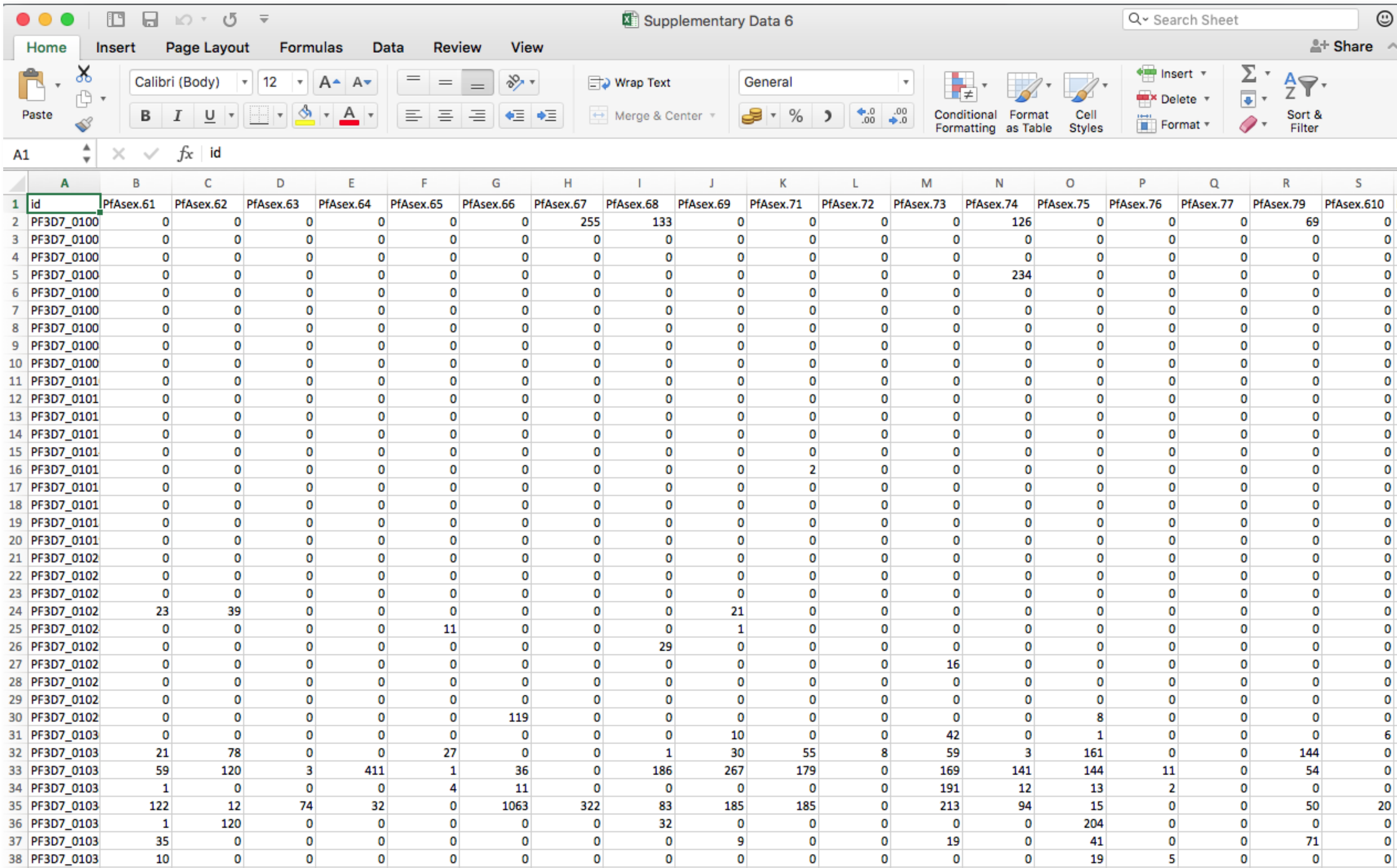
Transcriptome sequencing





What does a sequencing read represents?

How does the data looks like?



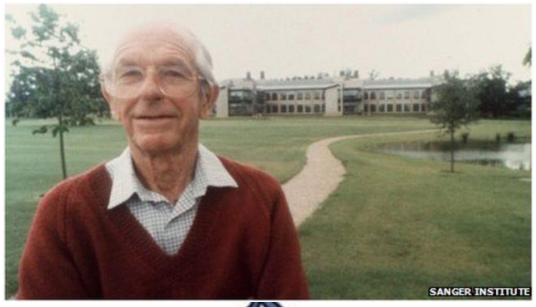
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	id	PfAsex.61	PfAsex.62	PfAsex.63	PfAsex.64	PfAsex.65	PfAsex.66	PfAsex.67	PfAsex.68	PfAsex.69	PfAsex.71	PfAsex.72	PfAsex.73	PfAsex.74	PfAsex.75	PfAsex.76	PfAsex.77	PfAsex.79	PfAsex.610	Pf
2	PF3D7_0100	0	0	0	0	0	0	255	133	0	0	0	0	126	0	0	0	69	0	
3	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0	234	0	0	0	0	0	
6	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	PF3D7_0100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	PF3D7_0101	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	
17	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	PF3D7_0101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	PF3D7_0102	23	39	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	
25	PF3D7_0102	0	0	0	0	11	0	0	0	1	0	0	0	0	0	0	0	0	0	
26	PF3D7_0102	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	
27	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	
28	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
29	PF3D7_0102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30	PF3D7_0102	0	0	0	0	0	119	0	0	0	0	0	0	0	8	0	0	0	0	
31	PF3D7_0103	0	0	0	0	0	0	0	0	10	0	0	42	0	1	0	0	0	6	
32	PF3D7_0103	21	78	0	0	27	0	0	1	30	55	8	59	3	161	0	0	144	0	
33	PF3D7_0103	59	120	3	411	1	36	0	186	267	179	0	169	141	144	11	0	54	0	
34	PF3D7_0103	1	0	0	0	4	11	0	0	0	0	0	191	12	13	2	0	0	0	
35	PF3D7_0103	122	12	74	32	0	1063	322	83	185	185	0	213	94	15	0	0	50	20	
36	PF3D7_0103	1	120	0	0	0	0	0	32	0	0	0	0	0	204	0	0	0	0	
37	PF3D7_0103	35	0	0	0	0	0	0	0	9	0	0	19	0	41	0	0	71	0	
38	PF3D7_0103	10	0	0	0	0	0	0	0	0	0	0	0	0	19	5	0	0	0	

Microarray vs. RNA-seq

	Microarray	RNAseq
Basis	Microarray platform required	Reference genome preferable
Interspecific comparison	Tricky	Relatively straightforward
Sample quantity	100 ng RNA per sample	1ug RNA per sample
Coverage	Low abundance transcripts not detectable	Single copy transcripts detectable with enough sequence coverage
Informatics	Relatively low requirements	Currently intense requirements
Cost	Fairly cheap once platform setup	Relatively expensive but falling.

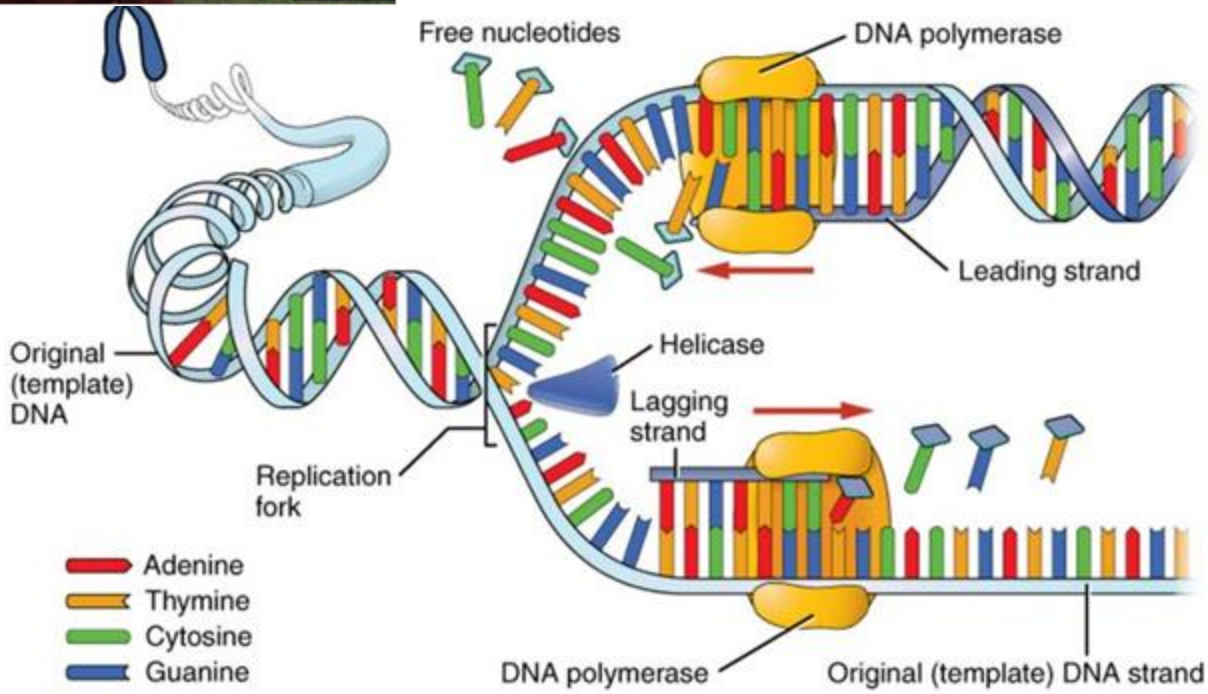
In fact RNA-seq is much more like EST sequencing, but truly global and quantitative

How Does Sequencing Work?



Sanger sequencing (1975)

- Carry out replication under controlled conditions
- Artificially slow down the reaction to see the order in which bases are incorporated



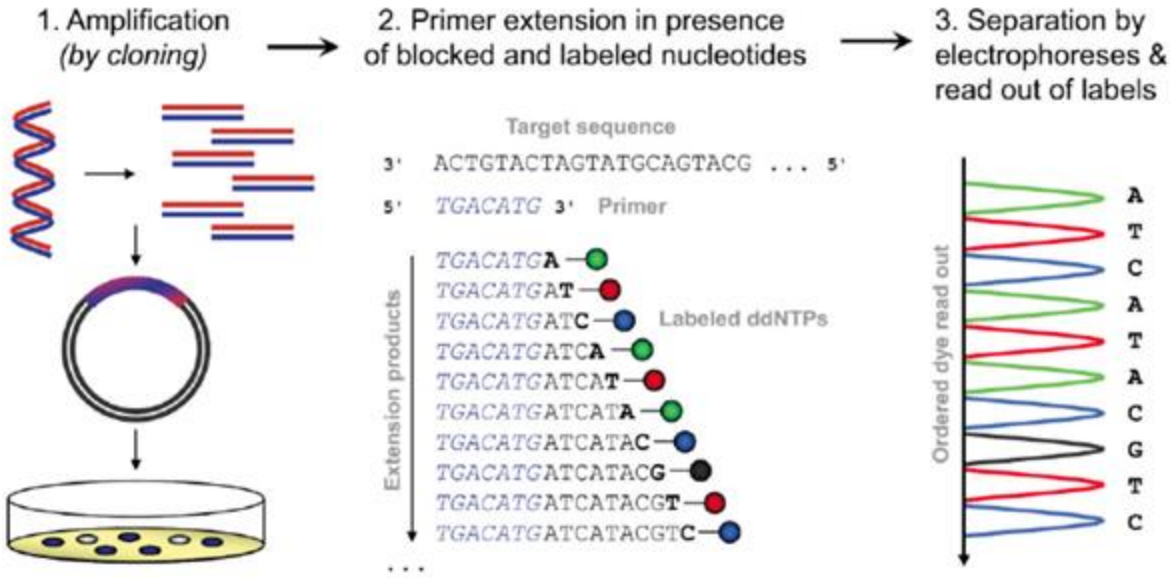
How Does Sequencing Work?

Sanger Sequencing (1975)

Uses modified nucleotides (didioxynucleotide - ddNTPs) that cannot be extended

Each ddNTP is labelled with a different dye so you can see the order in which they are incorporated

Long reads and low error rate, but low throughput



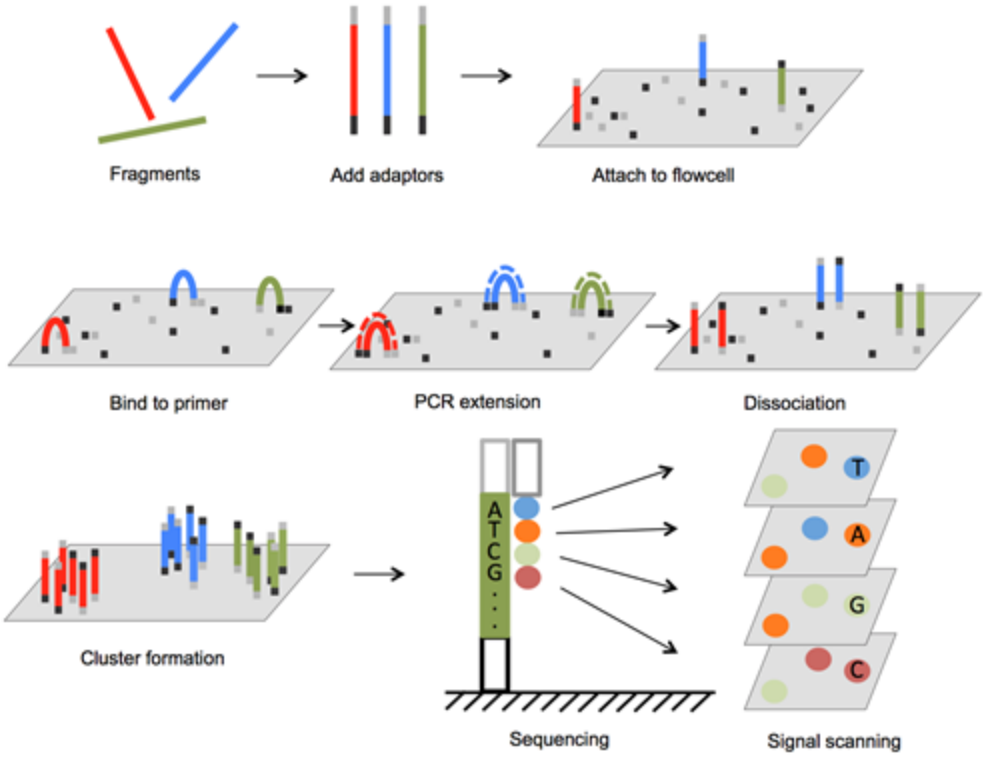
How Does Sequencing Work? Illumina Sequencing

Illumina Sequencing

DNA fragments are adaptor-ligated and attached to a flow cell

PCR is carried out in situ to form clusters

Sequencing can be carried out on millions of clusters simultaneously



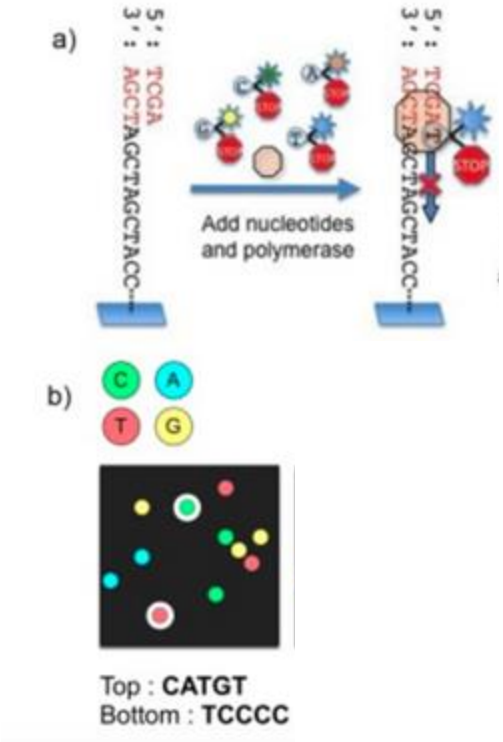
How Does Sequencing Work?

Illumina Sequencing

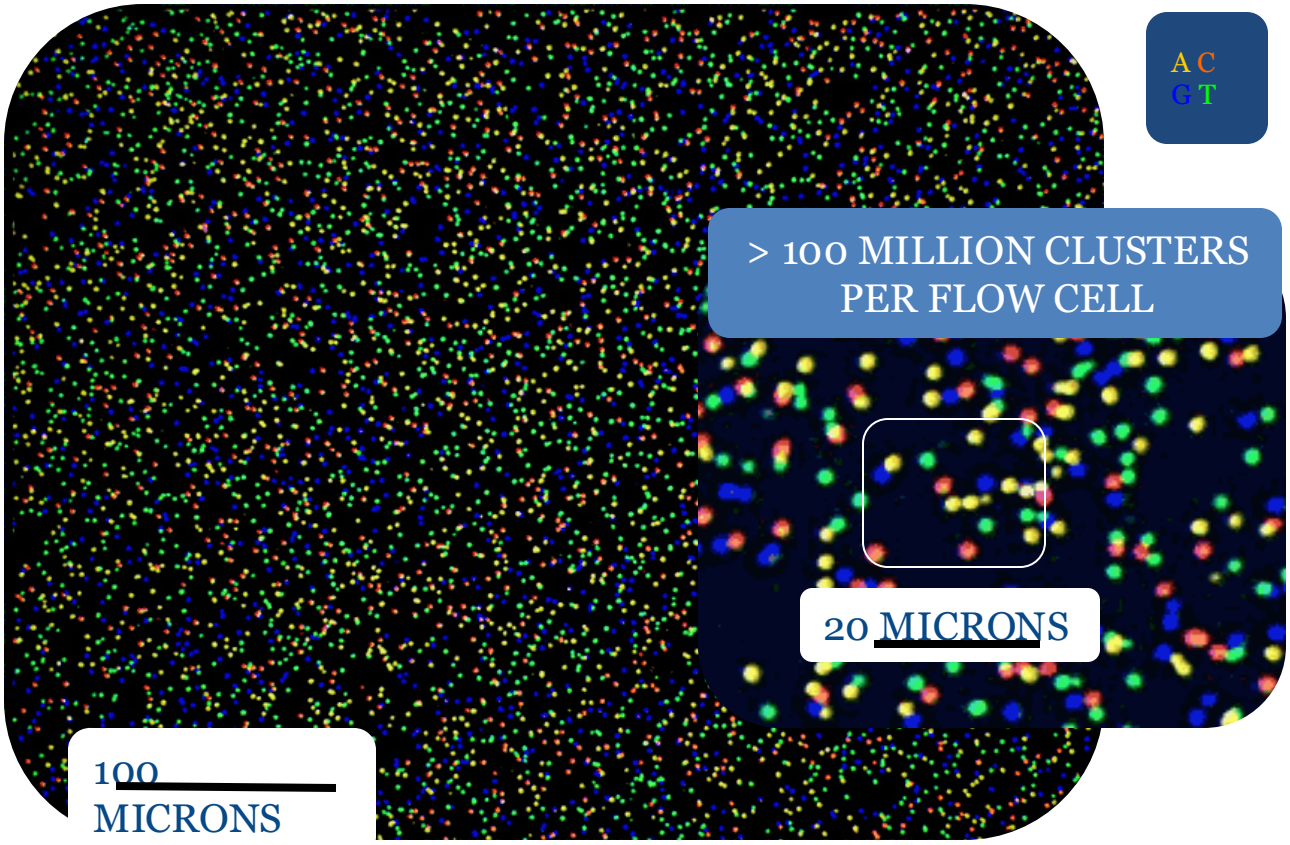
Blocked and labelled nucleotides are added

1 nucleotide is incorporated and an image is taken of the array

Label and block are removed and cycle repeats



The technology



Illumina reversible terminator

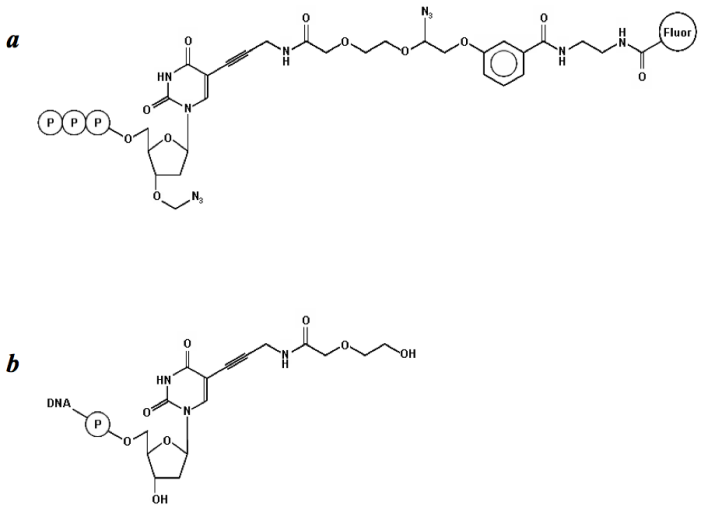


Figure S1. *a*. Structure of the reversible terminator 3'-*O*-azidomethyl 2'-deoxythymine triphosphate (T) labelled with a removable fluorophore. *b*. Structure of the incorporated nucleotide after removal of the fluorophore and terminator group. Each of the four nucleotides have an equivalent structure to the one shown here, except for the different base and a corresponding base-specific fluor.

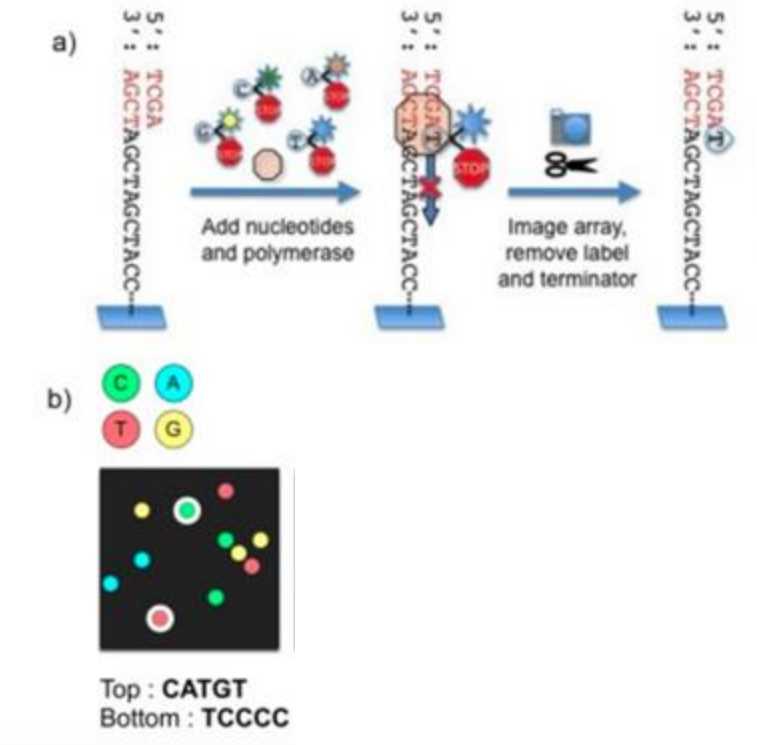
How Does Sequencing Work?

Illumina Sequencing

Blocked and labelled nucleotides are added

1 nucleotide is incorporated and an image is taken of the array

Label and block are removed and cycle repeats



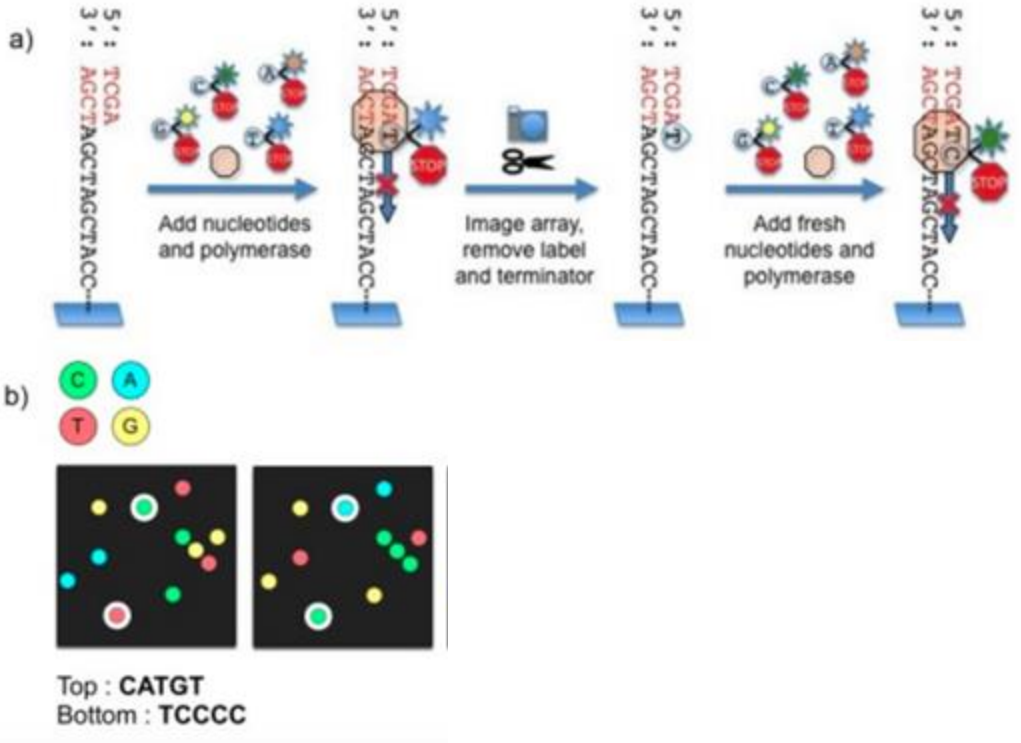
How Does Sequencing Work?

Illumina Sequencing

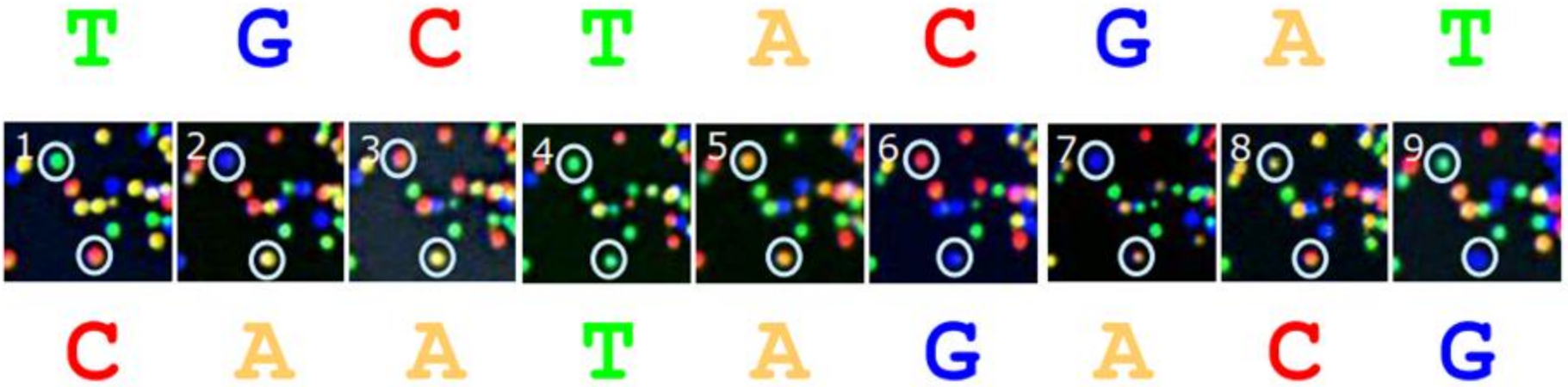
Blocked and labelled nucleotides are added

1 nucleotide is incorporated and an image is taken of the array

Label and block are removed and cycle repeats



Basecalling



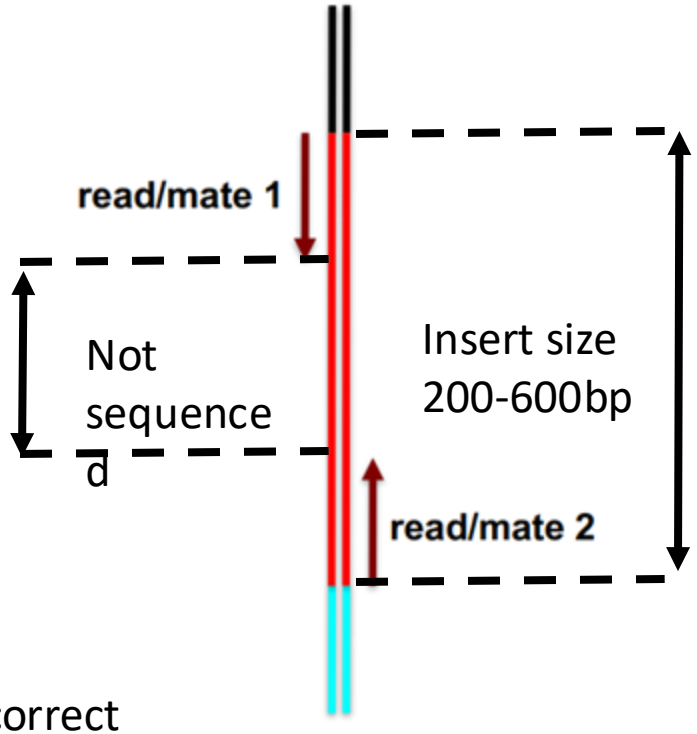
Paired End Sequencing

A short read is sequenced from each end of each fragment

Blue and black are adaptors

Insert or fragment size is the distance between adaptors

The region between the reads is not sequenced - it is covered by other fragments

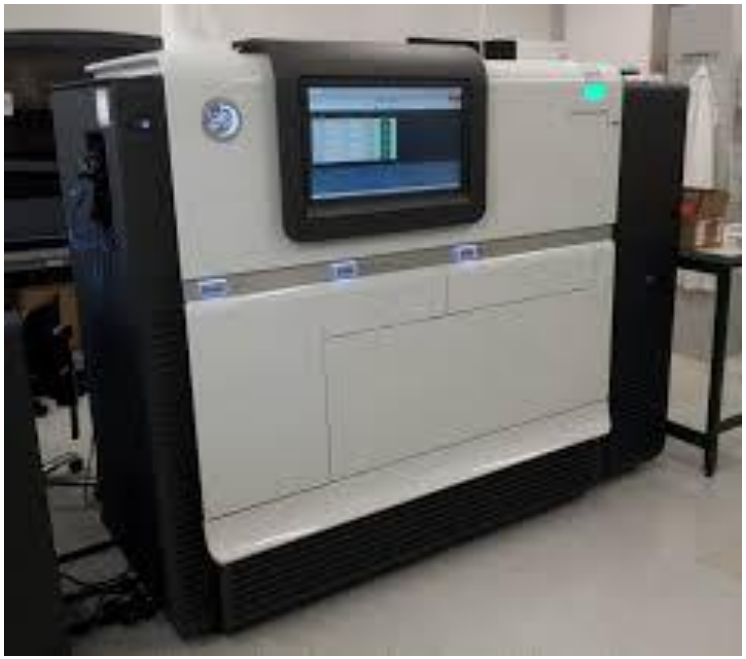


Reads that map in the correct orientation and the expected distance apart are “concordant” or “proper pairs”

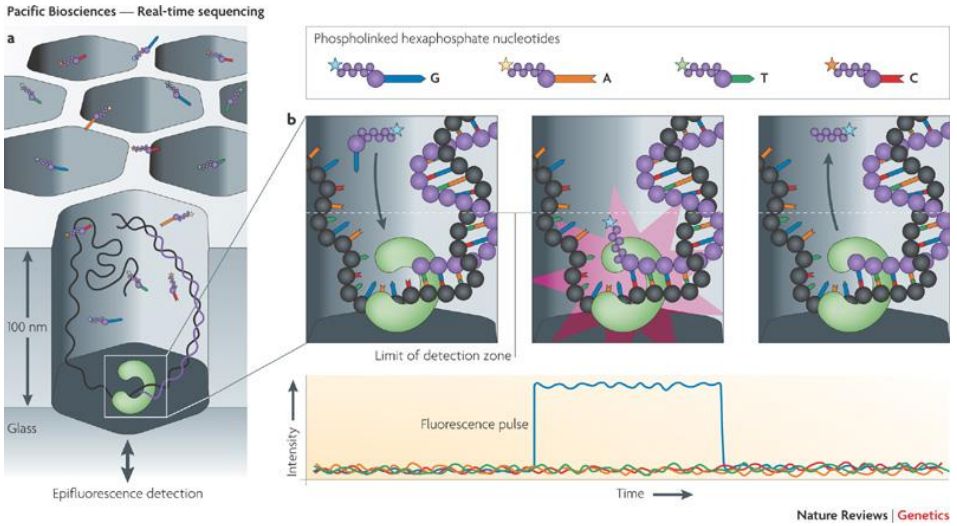
Concordant alignments are prioritised

Illumina HiSeq

Pacific Biosciences and Single Molecule Real Time (SMRT) sequencing

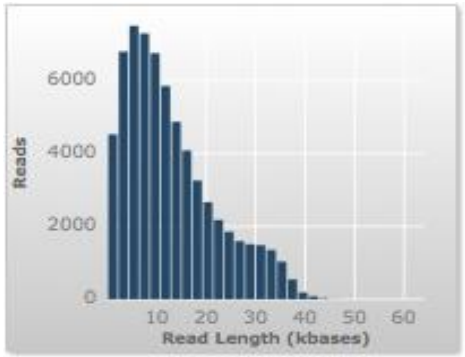


- Sequencing via light detection
- Light passing through channels is attenuated
  - “Zero Mode Waveguides”
  - Tiny reaction chambers
  - A single polymerase is monitored

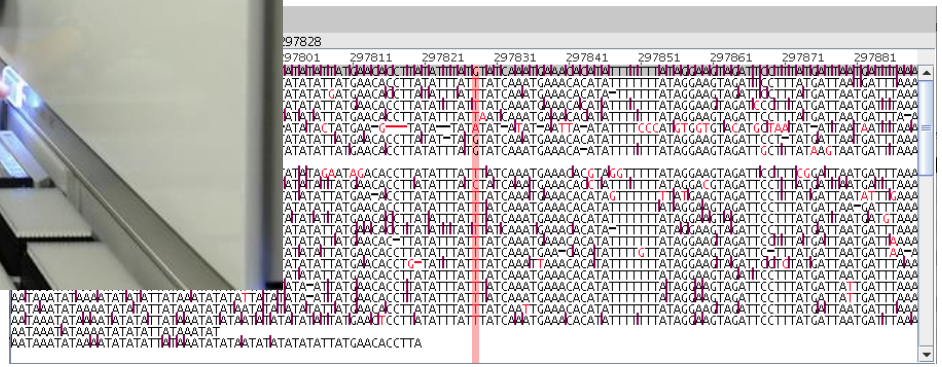


PacBio RS

raw reads



reads of insert



Vertical lines are insertions in read

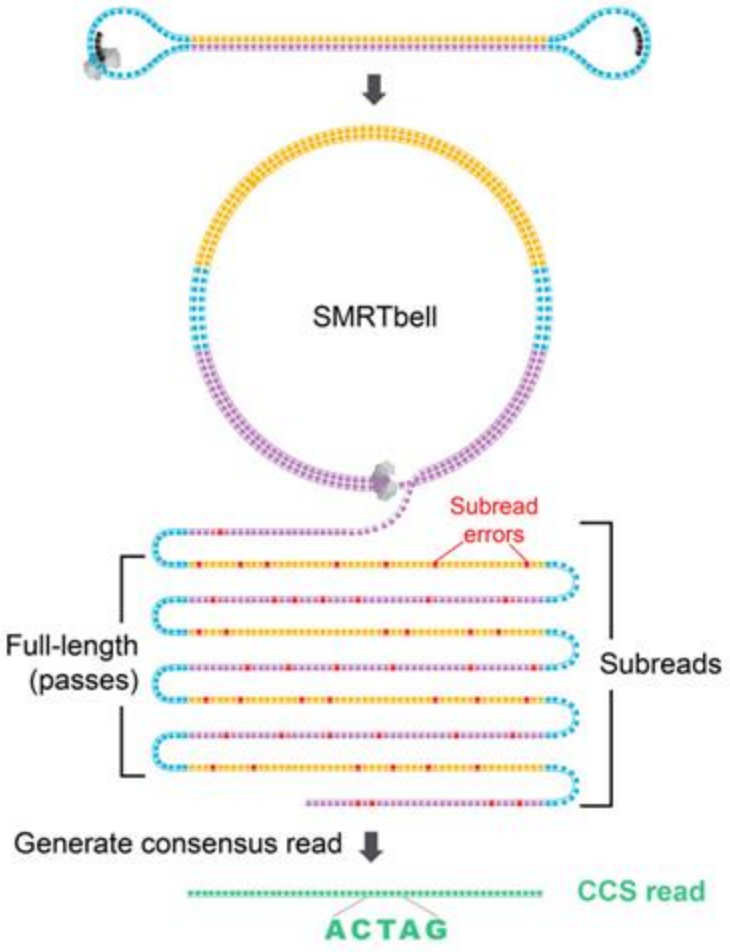
Hifi – High quality

Third Generation
PacBio

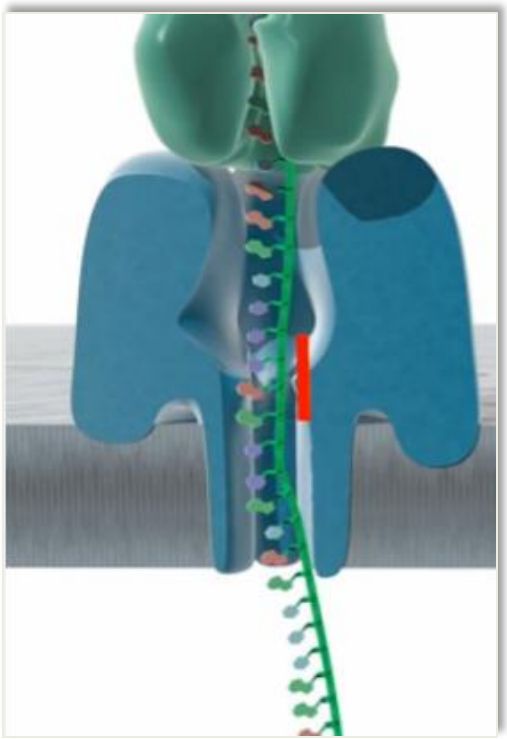
A polymerase trapped in a well on a plate synthesises DNA

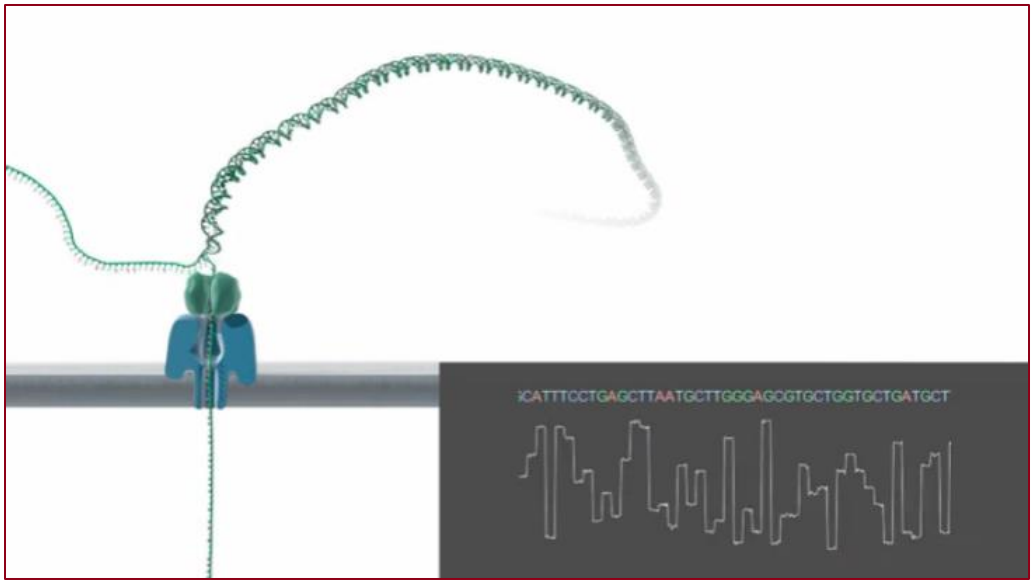
High stochastic error rate can be mitigated by circular consensus sequencing

This reduces the read length



Strand Sequencing by Oxford Nanopore



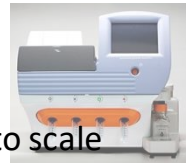


Which Technology to Use?

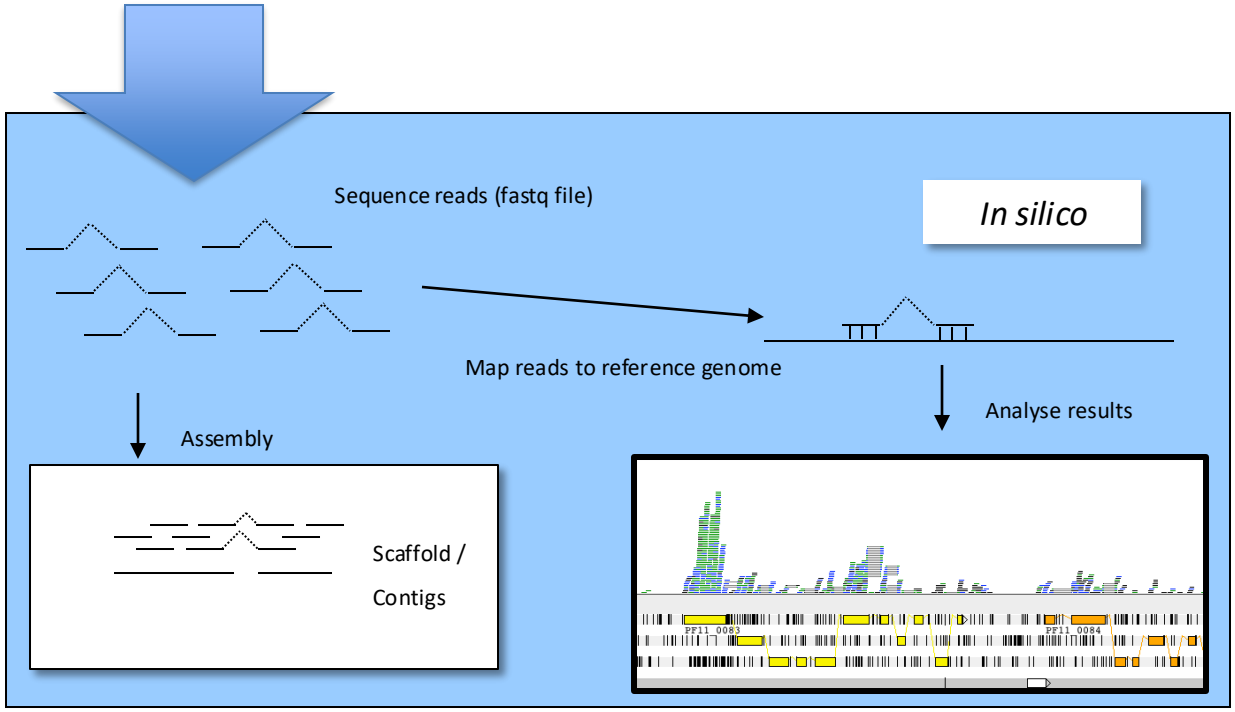
Sanger	**Illumina**	**PacBio**	**Nanopore**
Cheap	Mid	Expensive	Cheap
1000 bp	30 - 350 bp **BEST 150bp**	10,000 bp	100,000 bp (?)
>99.5% accuracy	>99% accuracy	85% accuracy Now 99%	70% accuracy Q20: 99%??
Low throughput			Portable
Sequence a mutant or construct	Genome wide SNV analysis, differential expression	Generating a new reference assembly	Reference assembly, field monitoring

Sequencing



Not to scale



What is an alignment?

- THOMAS
- T-OMAZ

THOMAS

TOMAZ

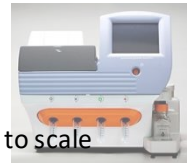
↓ „aligne"

```
THOMAS
|  | | |
T-OMAZ
```

Align the following two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

Not to scale



- Base calling
- Quality Control
- Read Trimming
- Alignment
- Variant calling

taggattg

atgagtag

ggtaaaaattgc

atgagtag

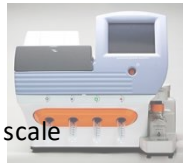
tgagtag

gtaaaaatt

gtagcatt

aaaatttttgacgaagcaggtaatgactagcattgacaggtaaaaaaaaattgcgcggcca

Not to scale



- Base calling
- Quality Control
- Read Trimming
- Alignment
- Variant calling

taggattg

atgagtag

ggtaaaaattgc

atgagtag

tgagtag

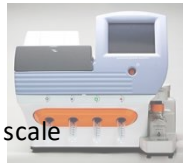
gtaaaaatt

gtagcatt

aaaatttttgacgaagcaggtaatgactagcattgacaggtaaaaaaaaattgcgcggcca

How long that take us for the human genome?

Not to scale



- Base calling
- Quality Control
- Read Trimming
- Alignment
- Variant calling

taggattg

atgag**tag**

atgag**tag**

tgag**tag**

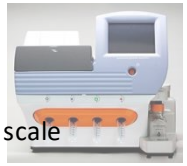
g**gta**aaaattgc

gtaaaaatt

g**tag**catt

aaaatttttgacgaagcaggtaatgac**tag**cattgacag**gta**aaaaaaaattgcgcggcca

Not to scale

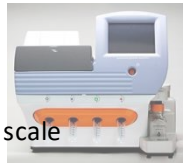


- Base calling
- Quality Control
- Read Trimming
- Alignment
- Variant calling

```
                                  atgagtag
                                  atgagtag
                                   tgagtag
                                       tagcattg    ggtaaaaa----ttgc
                                     gtagcatt       gtaaaaa----tt
aaaatttttgacgaagcaggtaatgactagcattgacaggtaaaaaaaaattgcgcggcca
```

Reference

Not to scale



Calling

- Base calling
- Quality Control
- Read Trimming
- Alignment
- Variant calling

```
                                atgagtag
                                atgagtag
                                 tgagtag
                                     tagcaatg    ggtaaaaa----ttgc
                                   gtagcatt       gtaaaaa----tt
aaaatttttgacgaagcaggtaatgactagcattgacaggtaaaaaaaaattgcgcggcca
```

Reference – Could the SNP be in the COVID spike protein?

How to map reads

reference: GGGTTAGCGATGGAGA

read1: GA

read2: TAGCG

read3: GGCG

How to map reads

Read

Reference

```
    TAGCG
    |||||
GGGTTAGCGATGGAGA
```

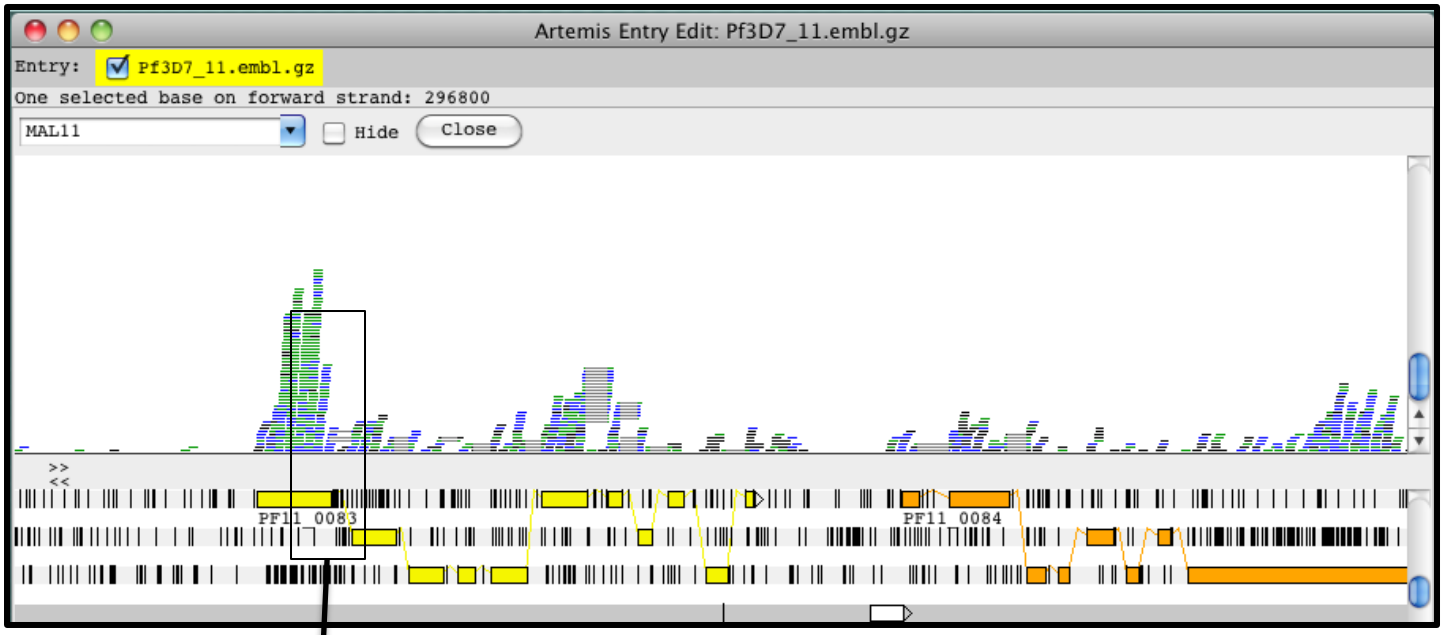
Read

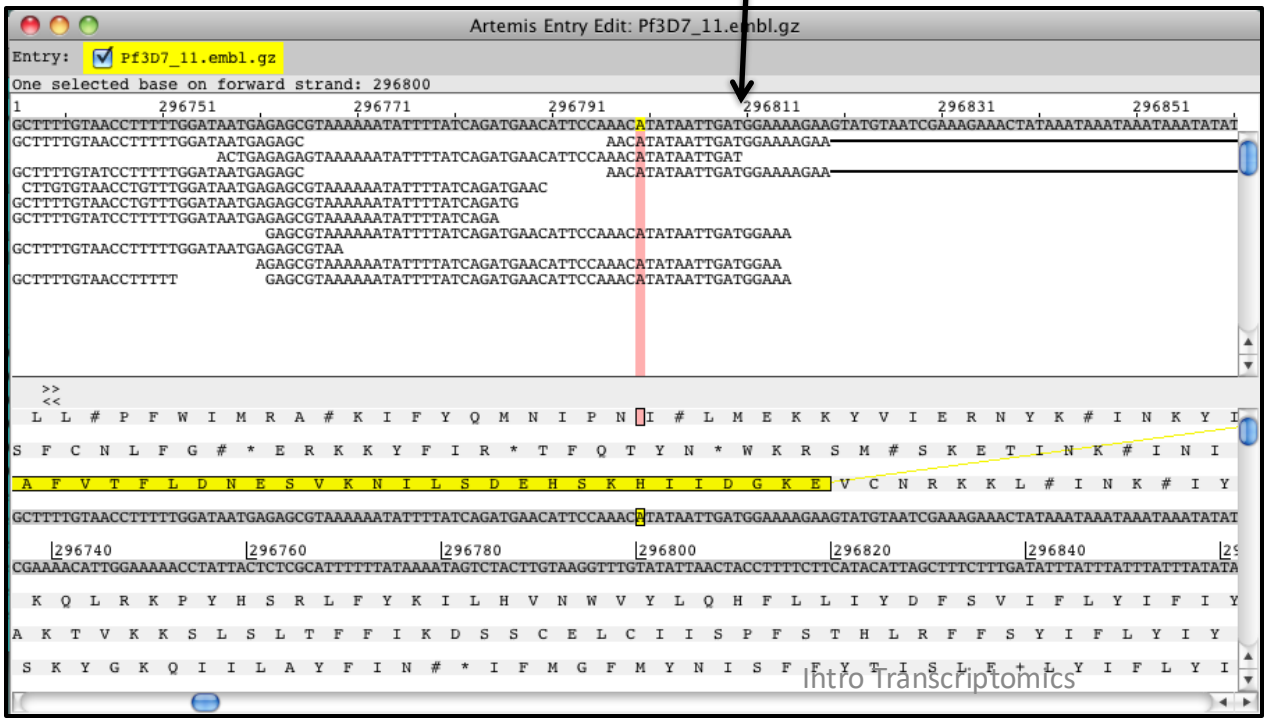
Reference

```
GG-----CG
||     ||
GGGTTAGCGATGGAGA
```

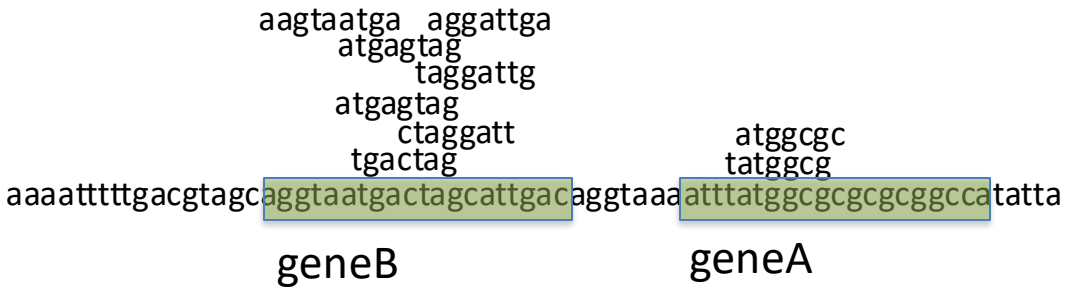
Splicing

Mapped reads in Artemis



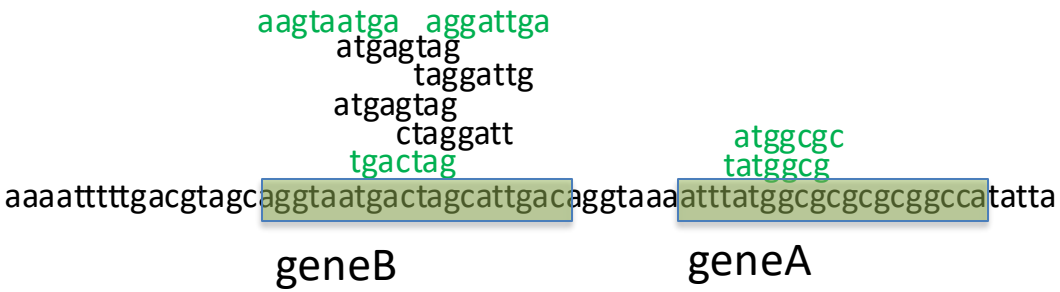


Mapped reads to read count matrix



read count table

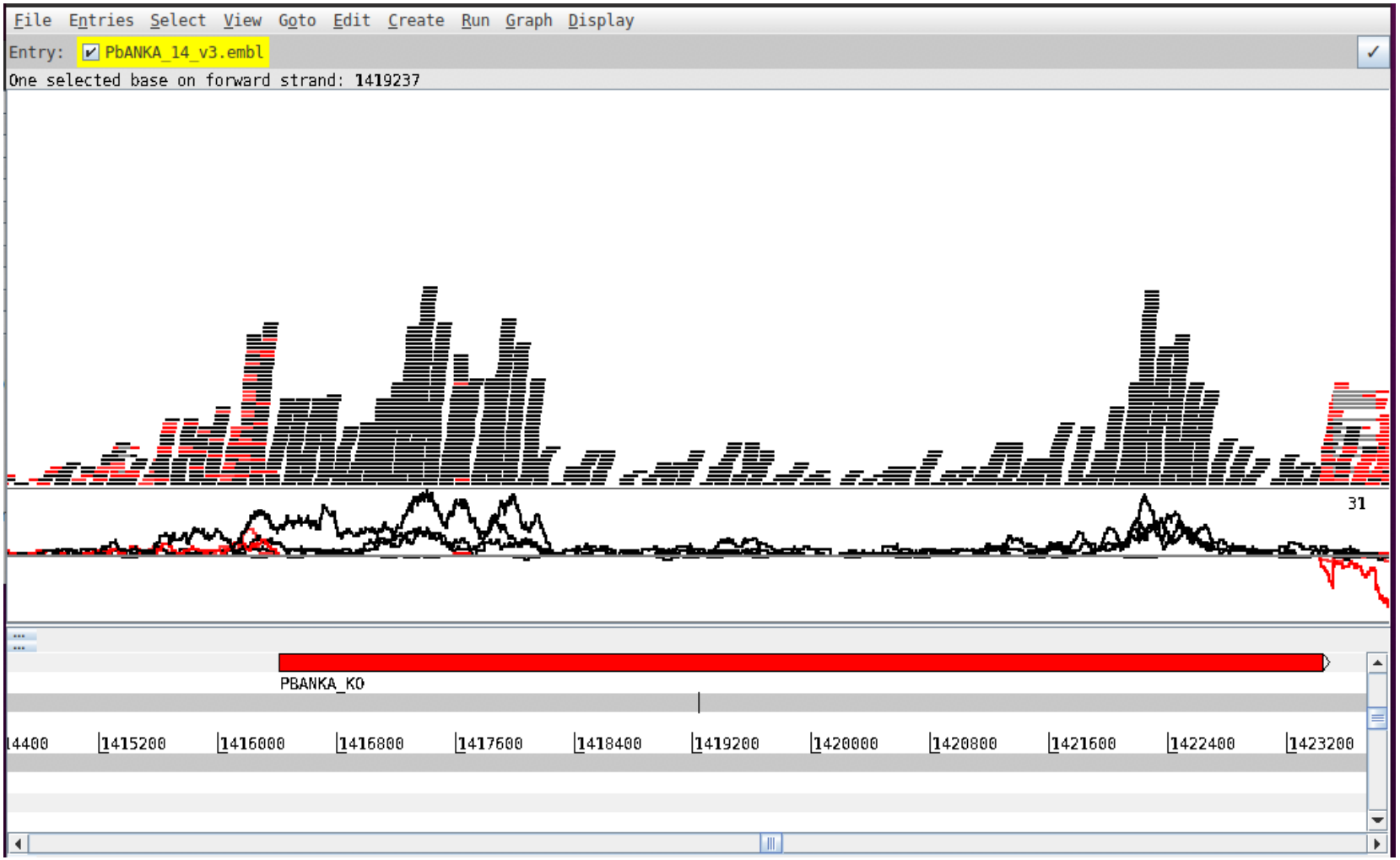
Genes	Condition1



read count table

Genes	Condition1	Cond 2

Counting the reads…



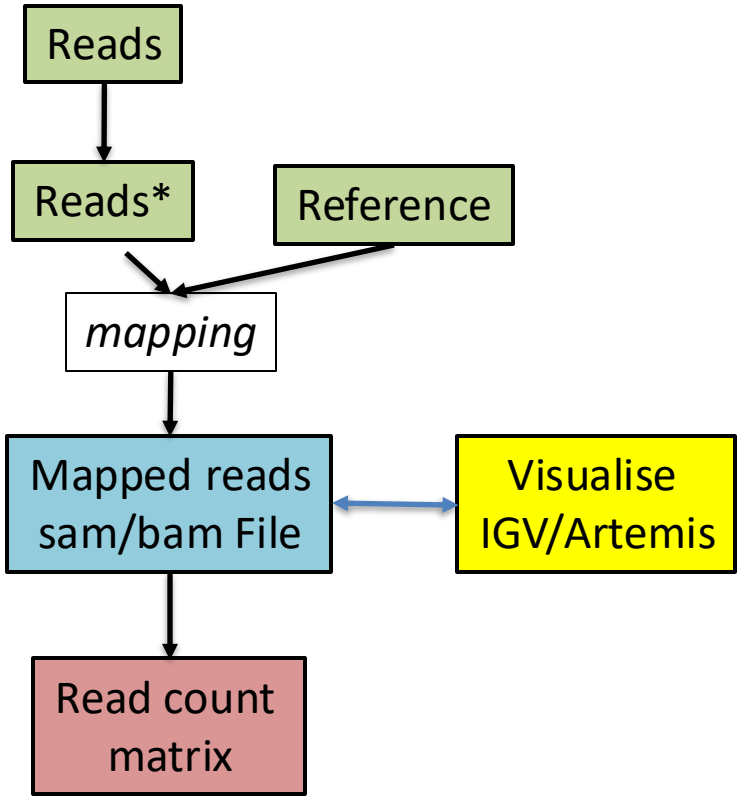
RNA-Seq pipeline

1. Quality control, Adapter trimming, low quality
fastqQC, trimOmatic

2. Mapping (reads on reference)
STAR, HiSat2

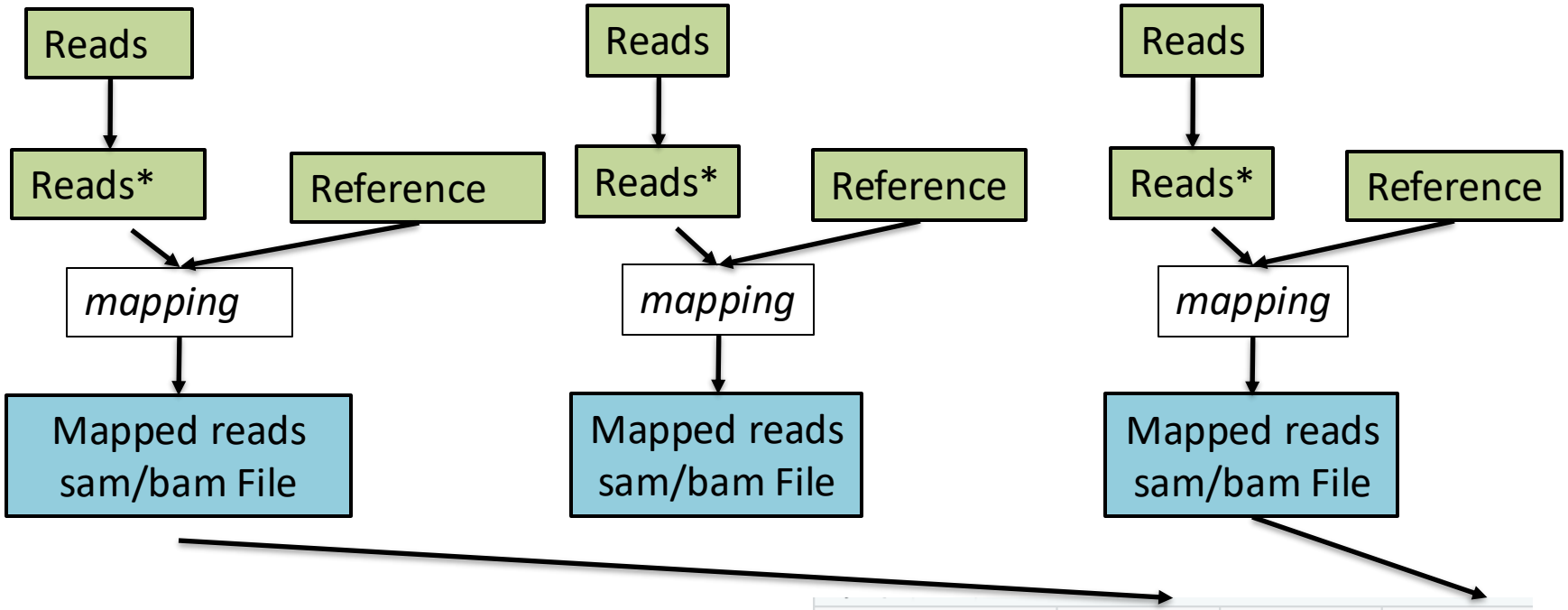
3. Transform mapping reads
- bam file (*samtools*)
- read count matrix (*featureCounts, HtSeq*)

4. Transform mapped reads (bam file)
- read count matrix (*featureCounts, HtSeq*)



	X107_d0
ENSG00000268903	531
ENSG00000269981	292
ENSG00000241860	74
ENSG00000279457	152
ENSG00000237094	220
ENSG00000225630	1915
ENSG00000237973	252
ENSG00000248527	6225

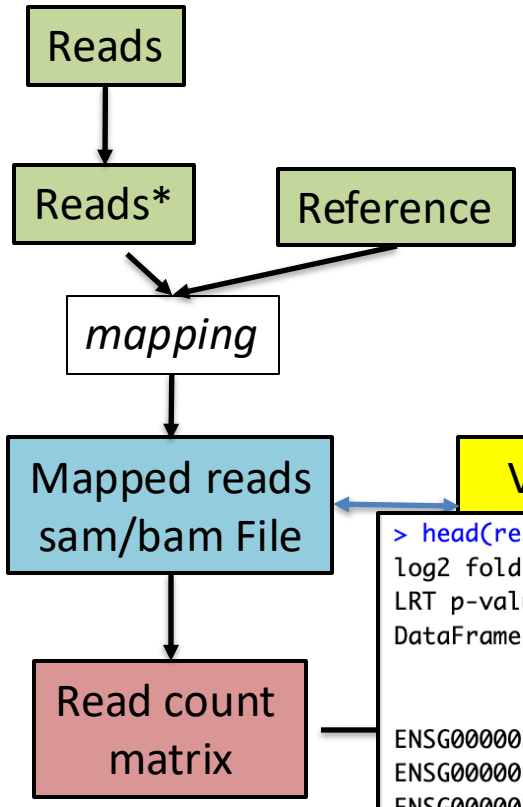
RNA-Seq Pipeline



	X107_d0	X107_d3	X107_d7
ENSG00000268903	531	280	349
ENSG00000269981	292	148	211
ENSG00000241860	74	33	91
ENSG00000279457	152	165	164
ENSG00000237094	220	76	155
ENSG00000225630	1915	1723	2483
ENSG00000237973	252	324	363
ENSG00000248527	6225	5439	7077
ENSG00000240618	72	27	82

4. Transform mapped reads (bam file)
- read count matrix (*featureCounts, HtSeq*)

RNA-Seq pipeline



Visualise

DeSeq2 in R

Design comparison

Quality control

Differential expression

Filtering data

critical result

```
> head(resSigGender[order(resSig$padj),],11)
log2 fold change (MLE): Day D1 vs D0
LRT p-value: '~ Day + Gender' vs '~ Gender'
DataFrame with 11 rows and 7 columns
                  baseMean log2FoldChange     lfcSE      stat       pvalue
                 <numeric>      <numeric> <numeric> <numeric>    <numeric>
ENSG00000234518    82.7813        1.98608 0.1383391  362.1162  3.54831e-78
ENSG00000123610  1615.9064        4.73725 0.3192287  361.9511  3.85287e-78
ENSG00000212283    49.0348        1.09099 0.1536103  164.6714  1.79851e-35
ENSG00000253276   802.7724        1.17711 0.1871856   46.5053  4.42838e-10
ENSG00000187764 14997.5395        1.07816 0.1126134  192.8330  1.49143e-41
ENSG00000110848   103.5402        1.72053 0.3353100   45.2773  8.07851e-10
ENSG00000232815   503.3275        2.80073 0.1576507  535.1428 1.15402e-115
ENSG00000166278   135.4917        2.93175 0.2424493  199.8576  4.52818e-43
ENSG00000116016   352.2069        3.51374 0.2060639  640.0369 2.10607e-138
ENSG00000196358  6179.2086        3.56994 0.2367376  342.2559  7.08674e-74
ENSG00000136827  1410.8754        1.69272 0.0900663  538.8147 1.84646e-116
```

padj	
9.03199e-42	C2
2.05956e-136	EPAS1
2.59209e-72	NTNG2
1.25613e-114	TOR1A

Quality Control

Trimming

Read Alignment

Read Counting

Differential Expression

Who spots the mistake?

Some definitions

- What does a RNA-Seq read represents?
- Why does that represent expression?
- What does the height of the coverage plot represents?

- 2 methods to count
  - **read counts (how many reads map to each gene)**
  - normalized : FPKM (fragments per kilobase of exon per million fragments mapped)

Overarching aim

- We try to detect genes that are differentially expressed between two groups that can explain observed “differences”

- We need good quality data that we can trust

Two examples

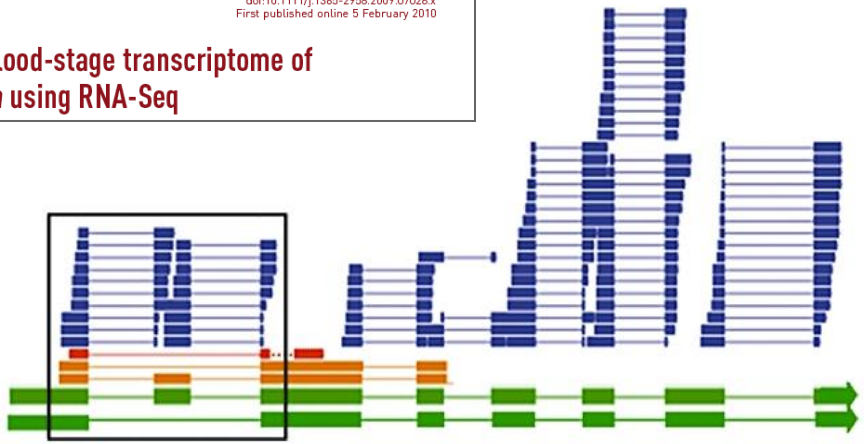
Plasmodium RNAseq

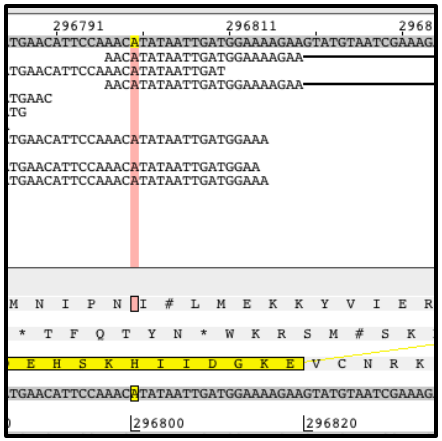
(2010)

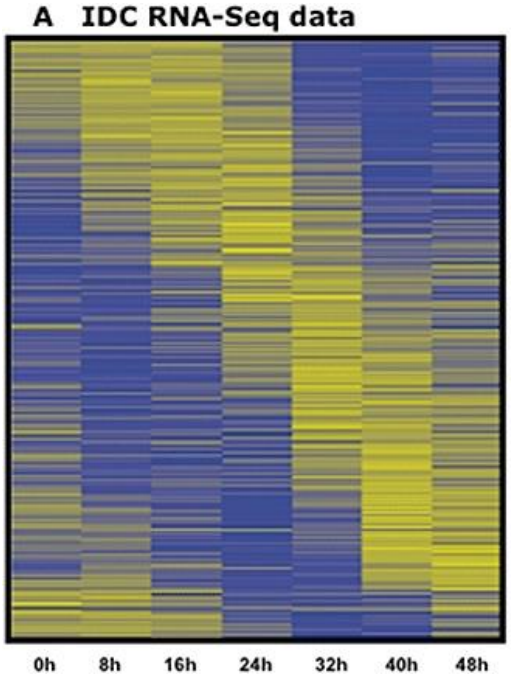
RNAseq

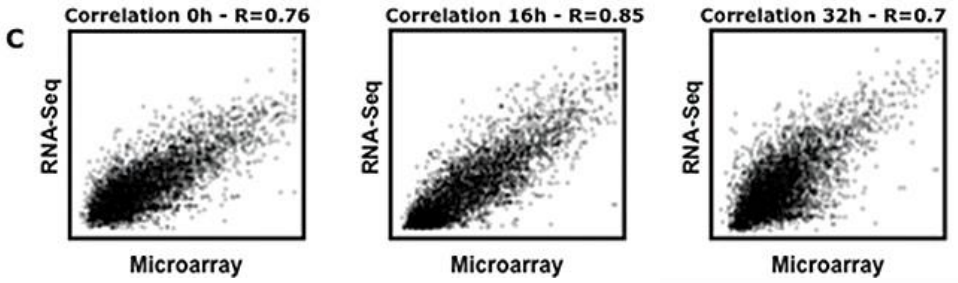
RNAseq

RT-PCR



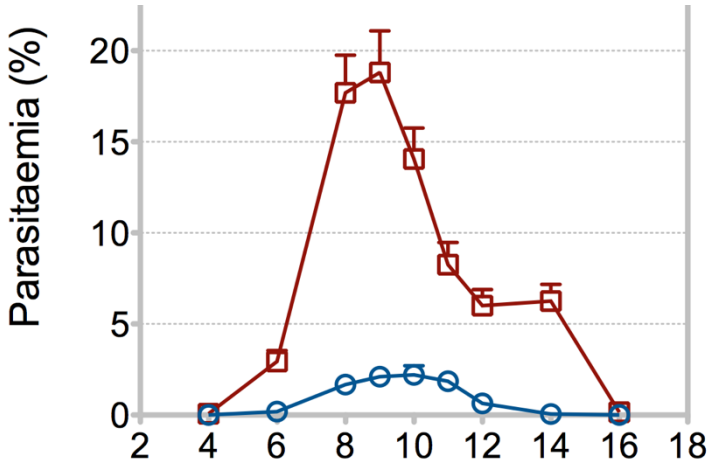






Differential expression: Vector transmission regulates immune control of virulence

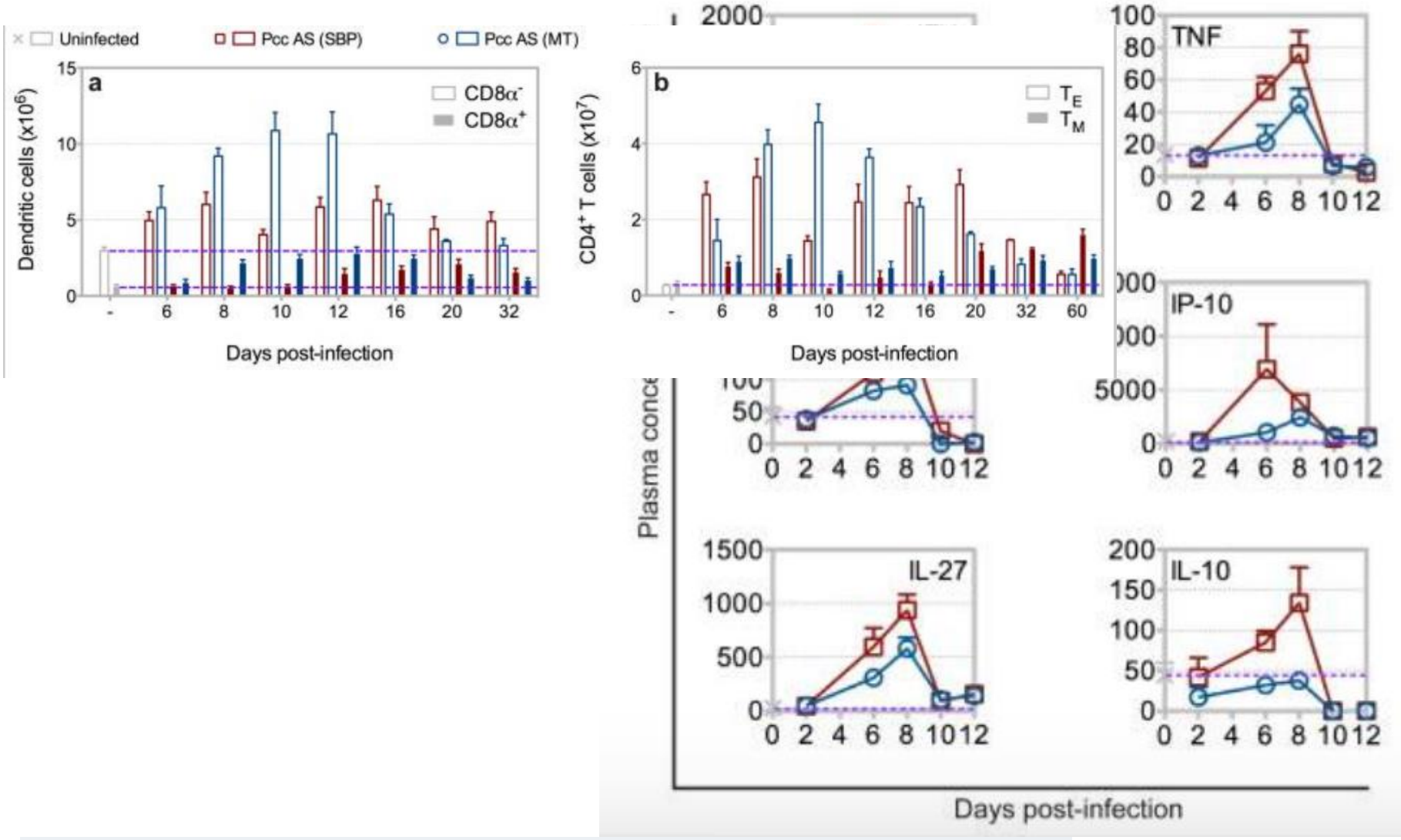
- Serial blood passage of parasites leads to increased virulence
- Phil Spence & Jean Langhorne developed routine mosquito transmission of *P.c. chabaudi*.



MT parasites show attenuated growth (not due to dose or pre-erythrocytic-stage of infection)

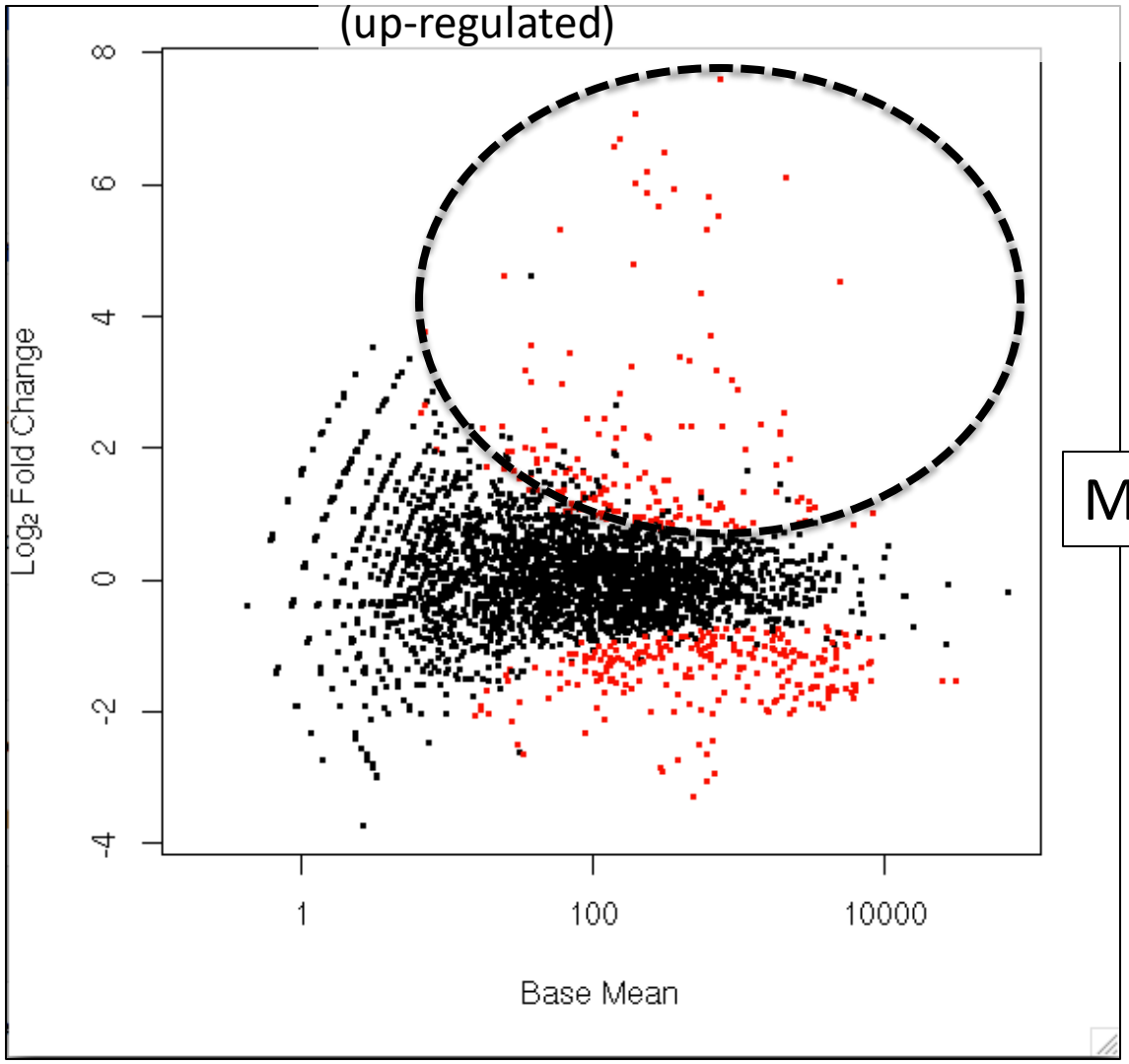
Spence et al. Nature 2013

The immune response is different



Differential expression

statistically significantly expressed genes (up-regulated)



MA-plot

3 biological replicates

Use read counts

assume negative binomial distribution of read counts

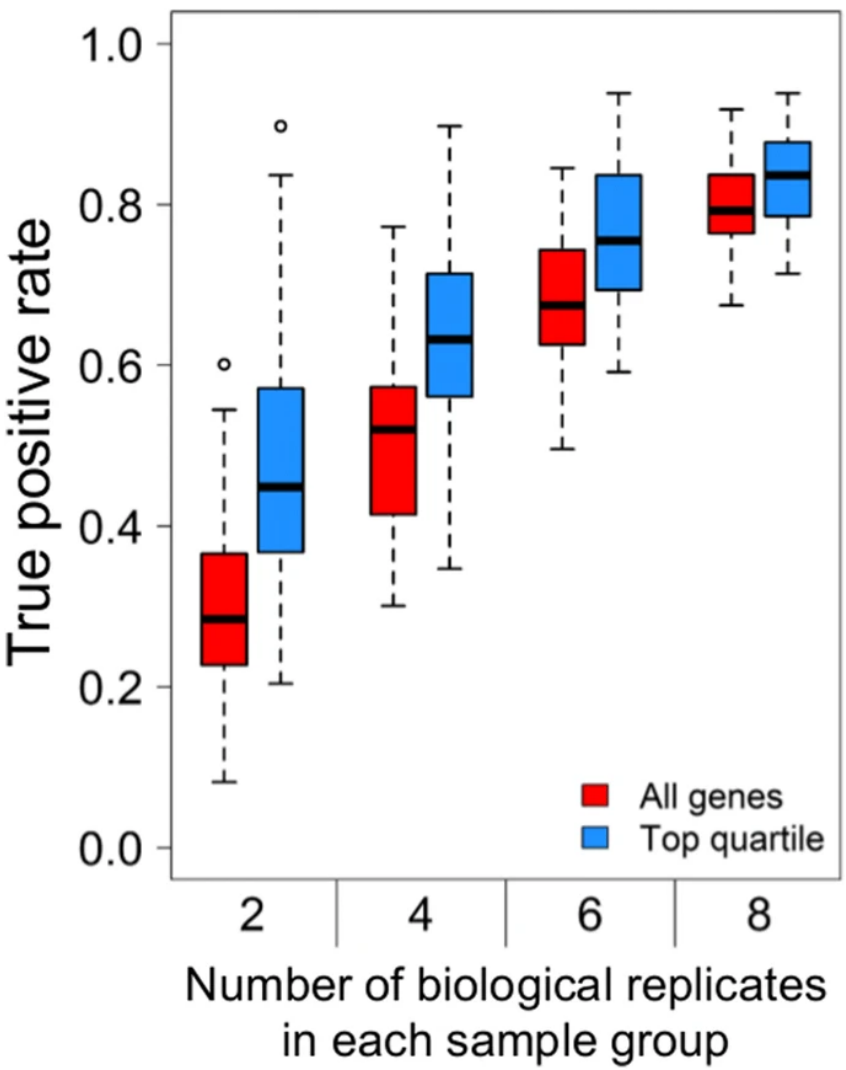
DESeq

Statistically differentially expressed?

- Need a test to understand which genes are “really” different
- Correction of multiple testing
- Tools: Sleuth, DEseq2, EDGER
- More robust results with more replicates

Statistically differ

- Need a test to unders
  “really” different
- Correction of multiple
- Tools: Sleuth, DEseq2
- More robust results w



Increasing numbers of sample replicates improves identification of schizont-stage genes varying in expression between different *P. falciparum* lines. Assessment of the proportion of genes captured as being differentially expressed between two different parasite clones (3D7 and D10), by taking 100 random samples of two, four, six and eight replicates of each (out of ten initially analysed replicates that identified 123 genes with $\log_2$ differences of > 2 in relative transcript levels between the two clones)

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-5257-x

Differential expression with DESeq2

- DESeq is an R package, part of Bioconductor
- Starts from read counts and does its own normalisation
- For each sample the size factor is calculated – how many reads
- Dispersion estimated (how tied are the expression values)
- For DE uses a negative binomial distribution, and “performs an advanced t-test”
- We will focus to run the tools and have an idea of the functionality, rather than becoming statistician’s!

How does the result looks like?

```
> head(resGender[order(resGender$padj),],11)
log2 fold change (MLE): Day D1 vs D0
LRT p-value: '~ Day + Gender' vs '~ Gender'
DataFrame with 11 rows and 7 columns
                 baseMean log2FoldChange     lfcSE      stat       pvalue         padj hgnc_symbol
                <numeric>      <numeric> <numeric> <numeric>    <numeric>    <numeric> <character>
ENSG00000117228  9186.124        4.85146 0.1558552   1421.24 7.24338e-308 1.02001e-303        GBP1
ENSG00000136689 17782.754        5.04210 0.1655905   1417.55 4.57195e-307 3.21911e-303       IL1RN
ENSG00000115267  4708.550        5.05549 0.1640378   1280.36 2.69209e-277 1.26367e-273       IFIH1
ENSG00000196116  1635.390        3.63350 0.1255556   1277.60 1.06827e-276 3.76085e-273       TDRD7
ENSG00000170581 23333.275        4.07187 0.1401291   1270.88 3.06152e-275 8.62246e-272       STAT2
ENSG00000138035  1222.625        4.55961 0.1589723   1259.13 1.08754e-272 2.55246e-269       PNPT1
```

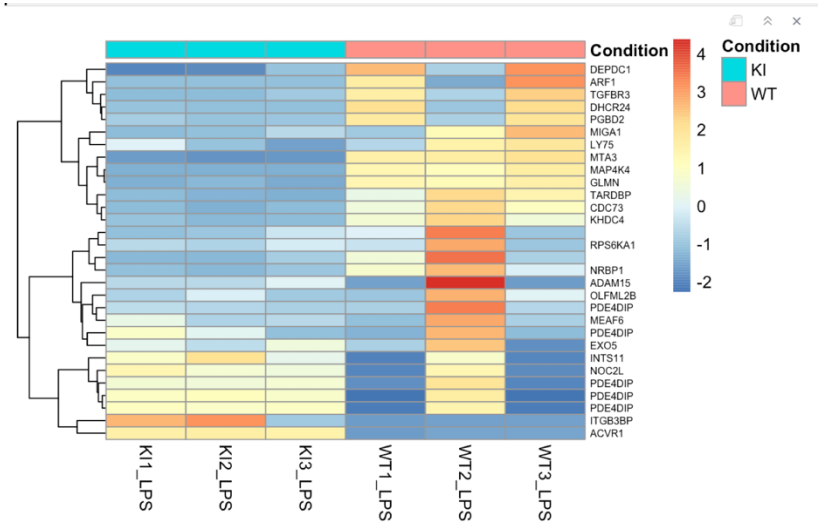
Most significant

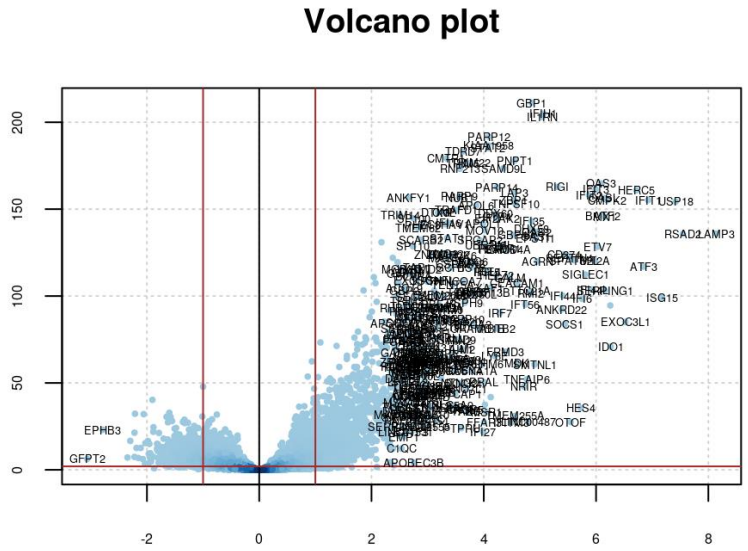
```
> head(resGender[order(resGender$log2FoldChange,decreasing = F),],11)
log2 fold change (MLE): Day D1 vs D0
LRT p-value: '~ Day + Gender' vs '~ Gender'
DataFrame with 11 rows and 7 columns
                 baseMean log2FoldChange     lfcSE      stat      pvalue         padj hgnc_symbol
                <numeric>      <numeric> <numeric> <numeric>   <numeric>    <numeric> <character>
ENSG00000131459    7.0171       -3.04647  0.582347   52.3370 2.53858e-11 1.08889e-10       GFPT2
ENSG00000182580   41.9070       -2.84842  0.260561  153.4997 4.63201e-33 6.75938e-32       EPHB3
ENSG00000125780  141.0757       -2.34984  0.419139   41.9035 4.20589e-09 1.42785e-08        TGM3
ENSG00000148600   46.3343       -2.32099  0.264634  113.6641 1.78515e-24 1.85661e-23       CDHR1
ENSG00000154764   70.7451       -2.26813  0.221951  137.4909 1.31352e-29 1.69232e-28       WNT7A
ENSG00000154102  142.0127       -2.20036  0.181780  148.4089 5.80734e-32 8.16158e-31    C16orf74
ENSG00000279447   36.9222       -2.19621  0.290009   70.4884 3.35492e-15 1.98839e-14
ENSG00000276600   10.6137       -2.18503  0.384055   44.4879 1.18876e-09 4.27478e-09       RAB7B
ENSG00000261150   46.0225       -2.14882  0.279827   70.2949 3.69071e-15 2.18006e-14       EPPK1
ENSG00000178947   22.8845       -2.08345  0.239324   94.6484 2.19719e-20 1.80518e-19   SMIM10L2A
ENSG00000158270    9.9380       -2.05730  0.488719   46.4807 4.48206e-10 1.68896e-09     COLEC12
```

Downregulated

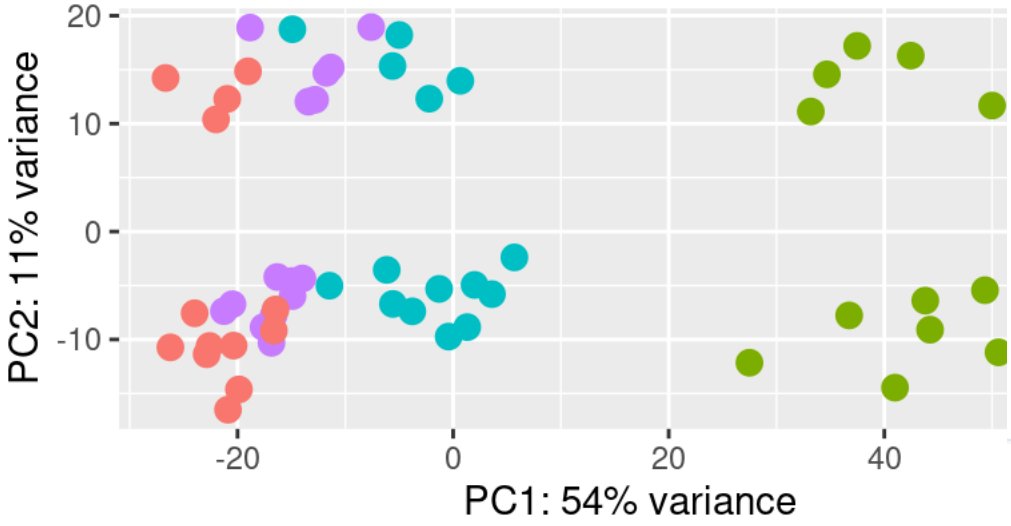
Visualisation

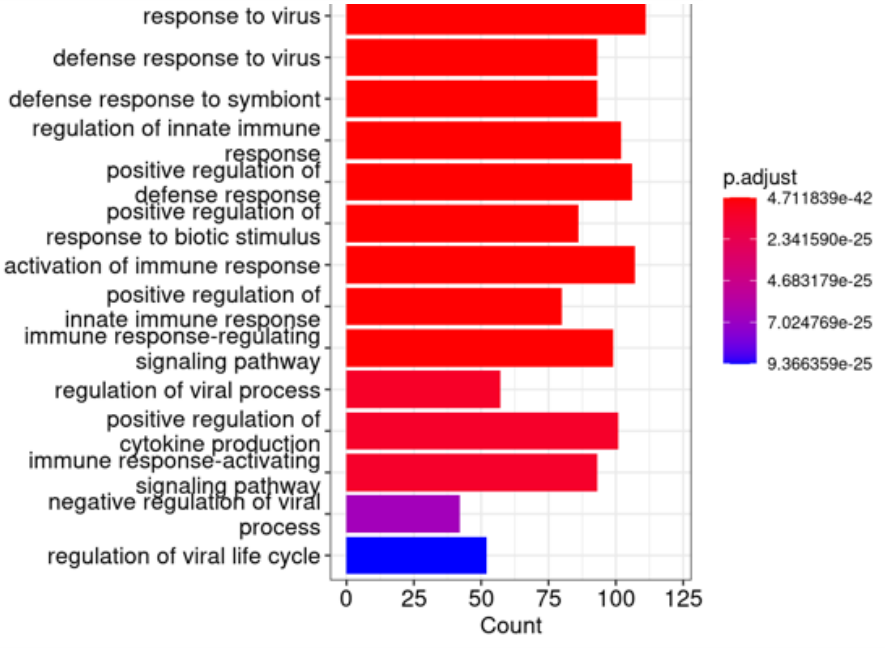
heatmap most expressed genes





PCA





GO enrichment (later)

Differential expression

- Good biological question
- Good sample preparation
- Enough biological replicates (get information through published studies)
- Sequencing -> Quality control, fastqc and visualize the results
- Differential expression (DeSeq2, EdgeR, Sleuth)
- Check FDR, enrichment test (GO)
- Interpretation of the results
- Write the paper

DE genes, what to do now?

- From Covid paper, CD14 cluster

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
ISG15	5.387767e-159	5.0588481	0.998	0.233	7.571429e-155
IFIT3	1.945114e-154	6.1124940	0.965	0.052	2.733468e-150
IFI6	2.503565e-152	5.4933132	0.965	0.076	3.518260e-148
ISG20	6.492570e-150	3.0549593	1.000	0.668	9.124009e-146
IFIT1	1.951022e-139	6.2320388	0.907	0.029	2.741772e-135
MX1	6.897626e-123	3.9798482	0.905	0.115	9.693234e-119
LY6E	2.825649e-120	3.7907800	0.898	0.150	3.970885e-116
TNFSF10	4.007285e-112	6.5802175	0.786	0.020	5.631437e-108
IFIT2	2.672552e-108	5.5525558	0.786	0.037	3.755738e-104
B2M	5.283684e-98	0.6104044	1.000	1.000	7.425161e-94
PLSCR1	4.634658e-96	3.8010721	0.793	0.113	6.513085e-92
IRF7	2.411149e-94	3.1992949	0.835	0.187	3.388388e-90
CXCL10	3.708508e-86	8.0906108	0.651	0.010	5.211566e-82
UBE2L6	5.547472e-83	2.5167981	0.851	0.297	7.795863e-79
PSMB9	1.716262e-77	1.7715351	0.937	0.568	2.411863e-73

Study them one by one

Enrichment analysis: what do we need?

It is a computational biology method that identifies biological functions that are overrepresented in a group of genes more than would be expected by chance

DE genes from your dataset

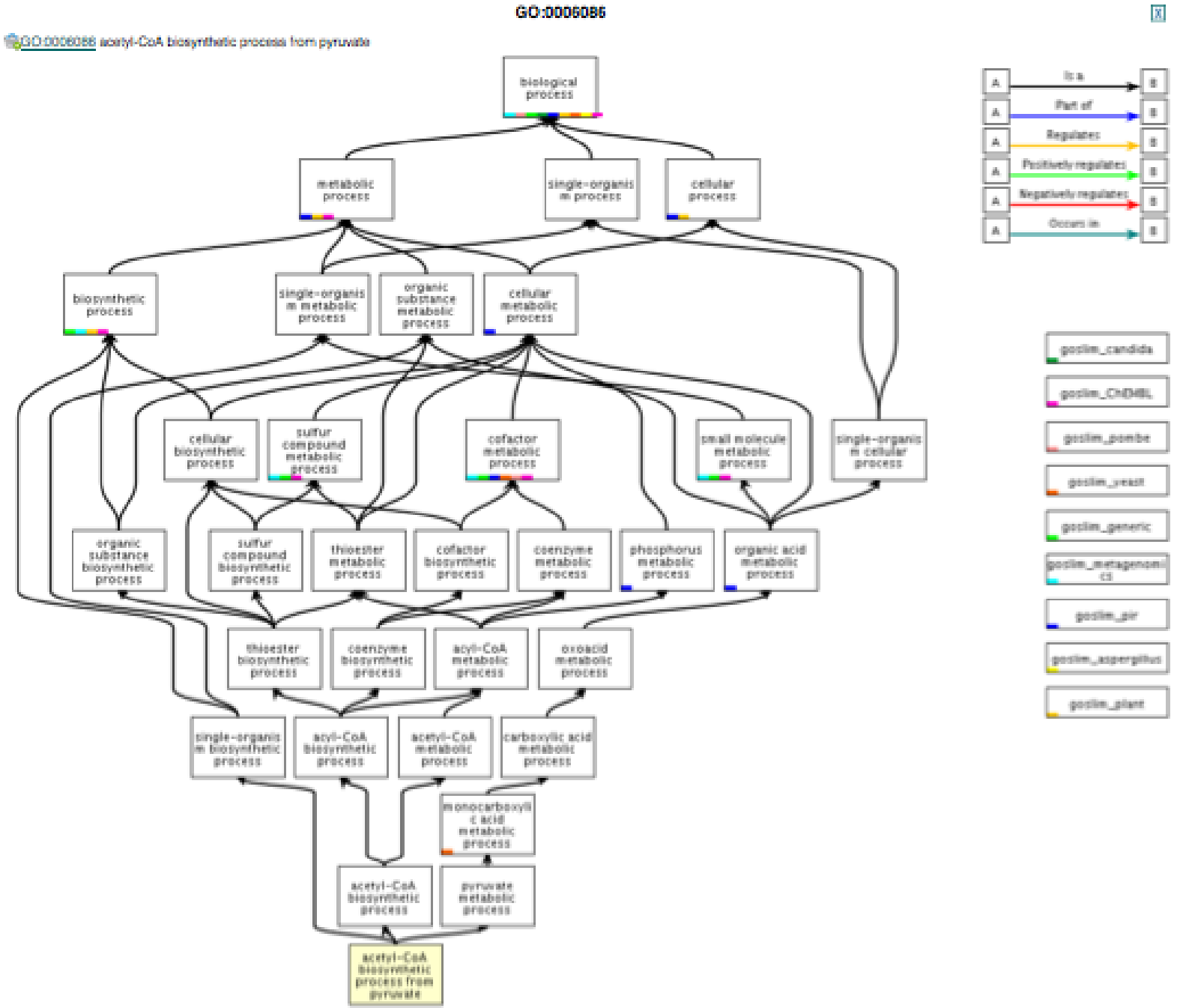
Where I can find info about “group of genes”?

- Databases:
  - KEGG
  - GO
  - StringDB
  - Reactome
  - DAVID knowledge
  - …

- From where are the info?
  - Literature
    - Automatic
    - Manual
  - Experimental
  - Modelling/Prediction

Stop here? what is GO?

- Gene Ontology
- Controlled vocabulary
- Saved as a graph
- The terms are grouped into three categories:
  - molecular function (describing the molecular activity of a gene)
  - biological process (describing the larger cellular or physiological role carried out by the gene, coordinated with other genes)
  - cellular process (describing the location in the cell where the gene product executes its function).


GO:0006086
GO:0006086 acetyl-CoA biosynthetic process from pyruvate
biological process
metabolic process
single-organism process
cellular process
biosynthetic process
single-organism metabolic process
organic substance metabolic process
cellular metabolic process
cellular biosynthetic process
sulfur compound metabolic process
cofactor metabolic process
small molecule metabolic process
single-organism cellular process
organic substance biosynthetic process
sulfur compound biosynthetic process
thioester metabolic process
cofactor biosynthetic process
coenzyme metabolic process
phosphorus metabolic process
organic acid metabolic process
thioester biosynthetic process
coenzyme biosynthetic process
acyl-CoA metabolic process
oxoacid metabolic process
single-organism biosynthetic process
acyl-CoA biosynthetic process
acetyl-CoA metabolic process
carboxylic acid metabolic process
monocarboxylic acid metabolic process
acetyl-CoA biosynthetic process
pyruvate metabolic process
acetyl-CoA biosynthetic process from pyruvate
is a
Part of
Regulates
Positively regulates
Negatively regulates
Occurs in
goslim_candida
goslim_chembl
goslim_pombe
goslim_yeast
goslim_generic
goslim_metagenomics
goslim_pir
goslim_aspergillus
goslim_plant

Or in R

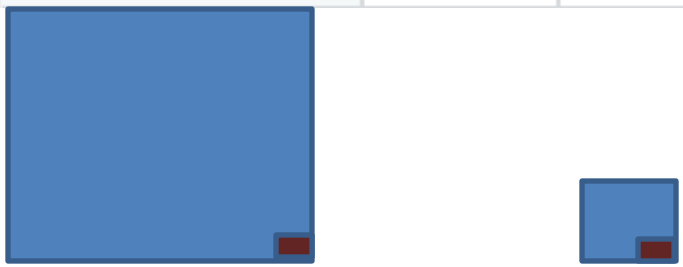
```
Good =matrix(data=c(100,900,500,11500),nrow=2)
colnames(Good) = c("DE", "Rest")
rownames(Good) =c("in Pathway","somewhere else")
fisher.test(Good)
```

- p-value < 2.2e-14

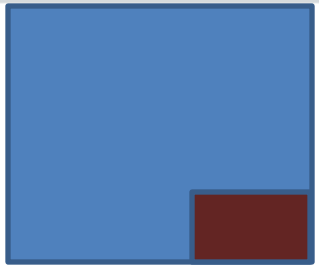
```
NotG
=matrix(data=c(100,900,1100,119000),nrow=2)
colnames(NotG) = c("DE", "Rest")
rownames(NotG) =c("in Pathway","somewhere
else")
fisher.test(NotG)
```

?

	DE	Rest
in Pathway	100	500
somewhere else	900	11500



	DE	Rest
in Pathway	100	1100
somewhere else	900	11900



Ratio is
The same

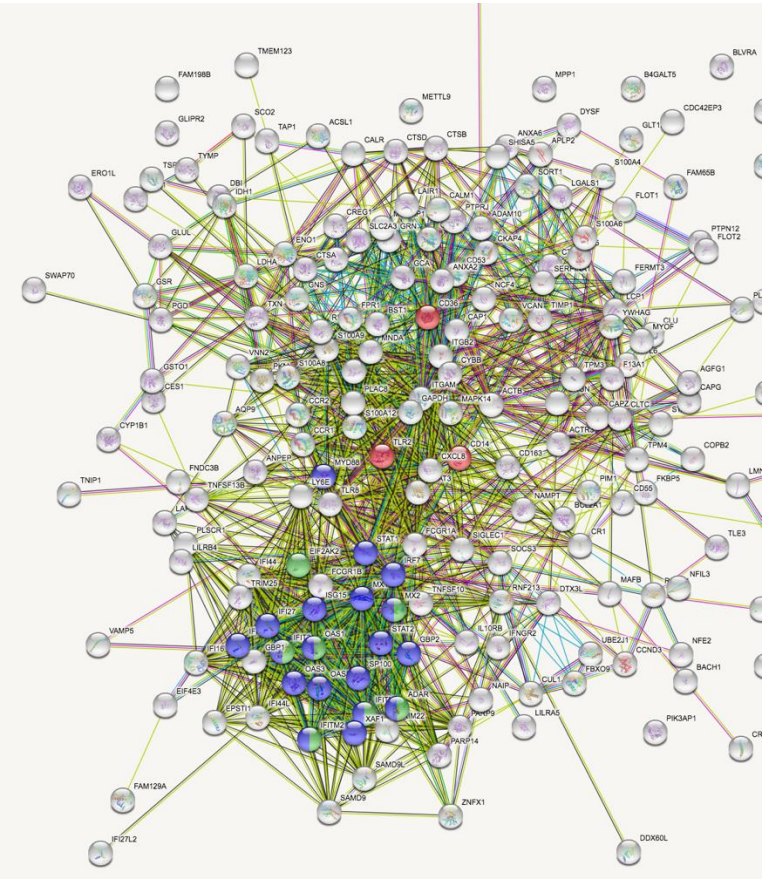
Multiple correction

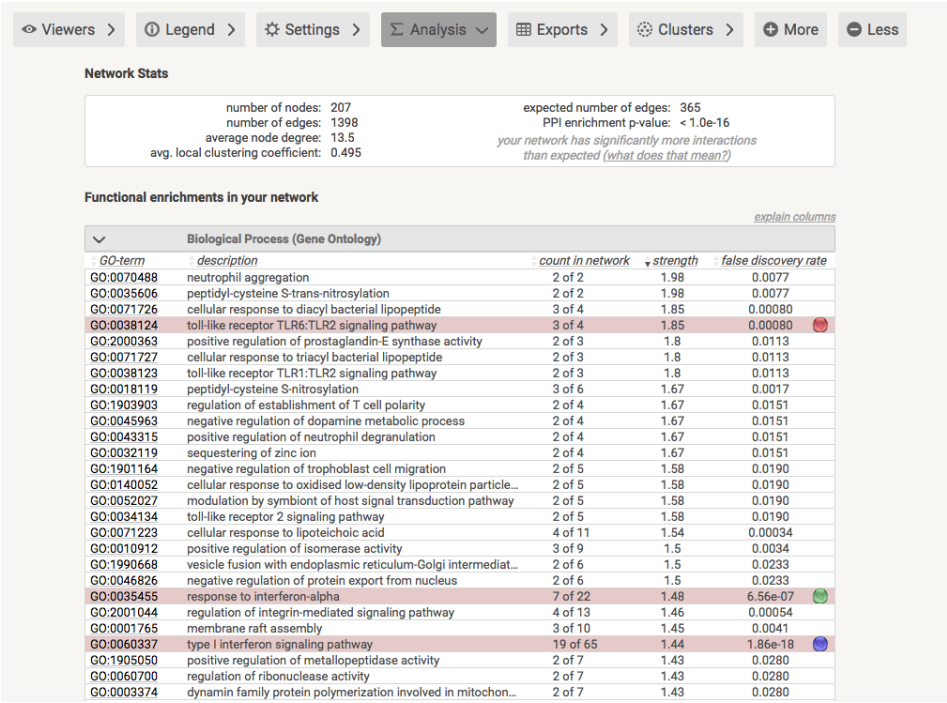
- If you do more test, you need to apply correct for multiple testing!!!

- If you have four conditions A B C D E and do
- A vs B; A vs C; A vs D; A vs E; B vs C; B vs D; B vs E; C vs D; C vs E & D vs E – you have to correct ALL P-values for the 10 amount of test.
- Most conservative method is to multiple so an adjusted P-value of 0.005 becomes 0.05 and is on the boardline of the false discovery rate!

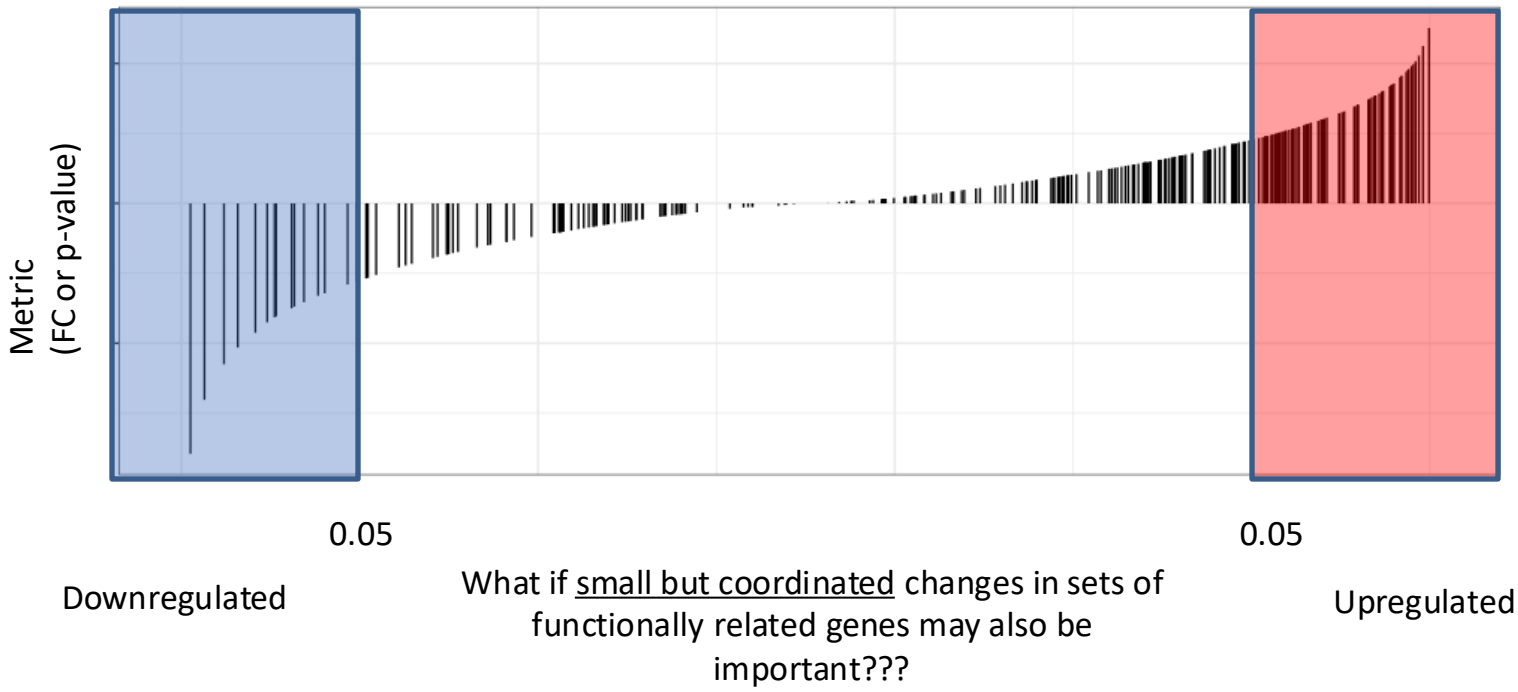
- Tools normally do multiple correction, but not between “experiments” of the user

Example 2 StringDB



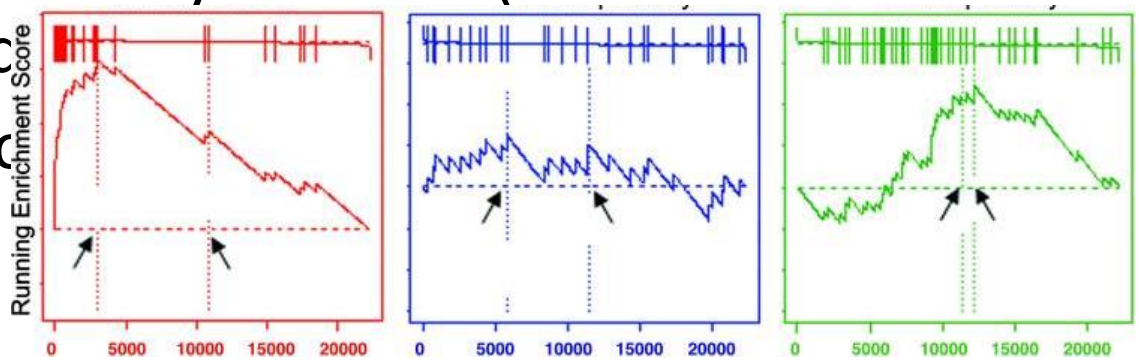


DE genes/p-value and pathways relationship…



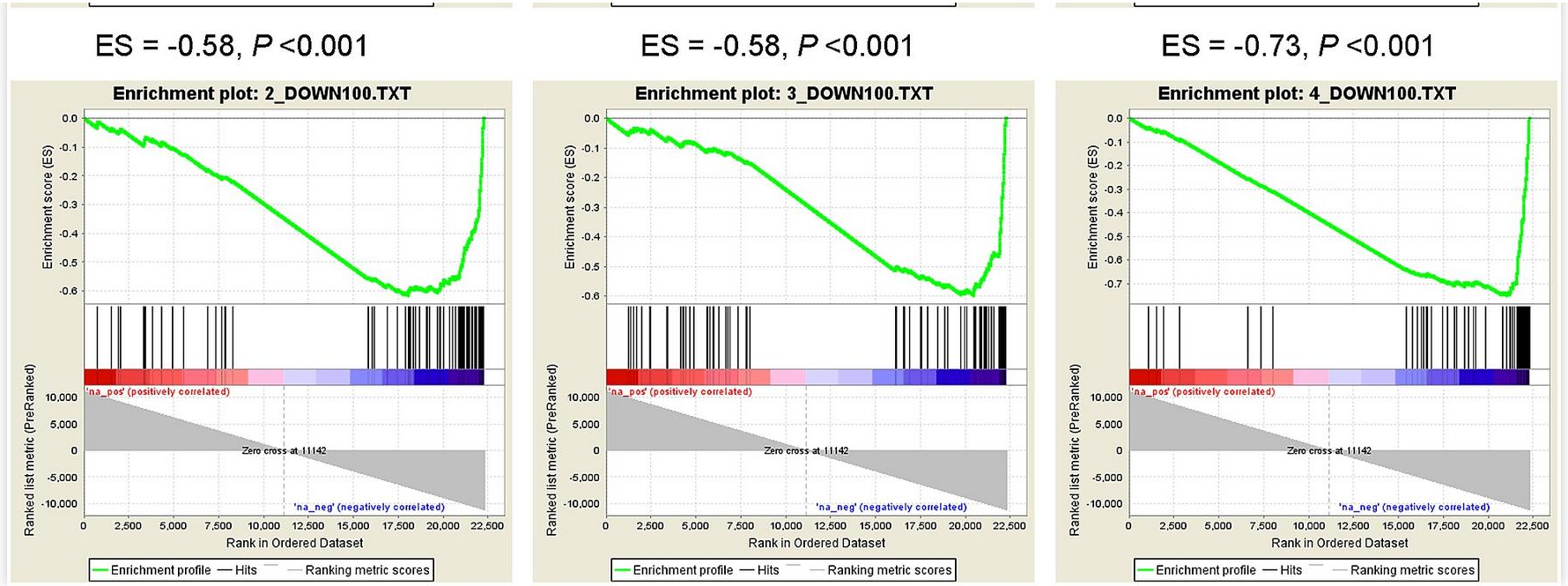
GSEA

- **Gene set enrichment analysis (GSEA)** (also **functional enrichment analysis**)
- Uses the complete gene list (by P-value)
- Orders genes by P-values (down regulatec
- Test the c terms of



https://www.gsea-msigdb.org/gsea/index.jsp

Examples



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4470683/

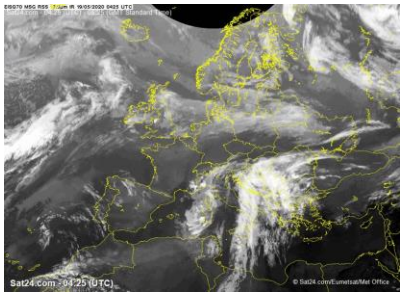
Questions

1. What is the power of RNA-Seq and DE analysis?
2. Why do we need good quality data?
3. What is the importance of biological replicates?
4. Why is statistical analysis DE / correct of P-values so important?

Consider

- How powerful are enrichments? In some cases these make more sense than in others
- Databases are based on existing knowledge
- Annotation is not evenly distributed: There is more research in T-cell in cancer than in a rare novel subtype in rheumatoid arthritis – this might distracted
- Enrichments gives you a general overview, but you need to then dig deeper
- Check that your ”most significant genes” for the gene list are also part of the enrichments

- In the end, it stills comes down to how much you and your collaborator know!



Conclusions (3/n)

- RNA-Seq is a powerful tool to access expression of all genes (and other elements miRNA)
- **Differential expression analysis to find genes changing expression due to different condition**
- Very population in Immunology
- In the exercise you are going to perform differential expression from a read count matrix with DESeq2
- You need good quality data and enough biological replicates to obtain statistically relevant results