RNA-seq gene level differential expression and clustering

Gilgi Friedlander

The Nancy & Stephen Grand Israel National Center for Personalized Medicine

Outline

- Introduction
- Quantification of gene expression
- Normalization
- Differential Expression
- Exploratory analysis

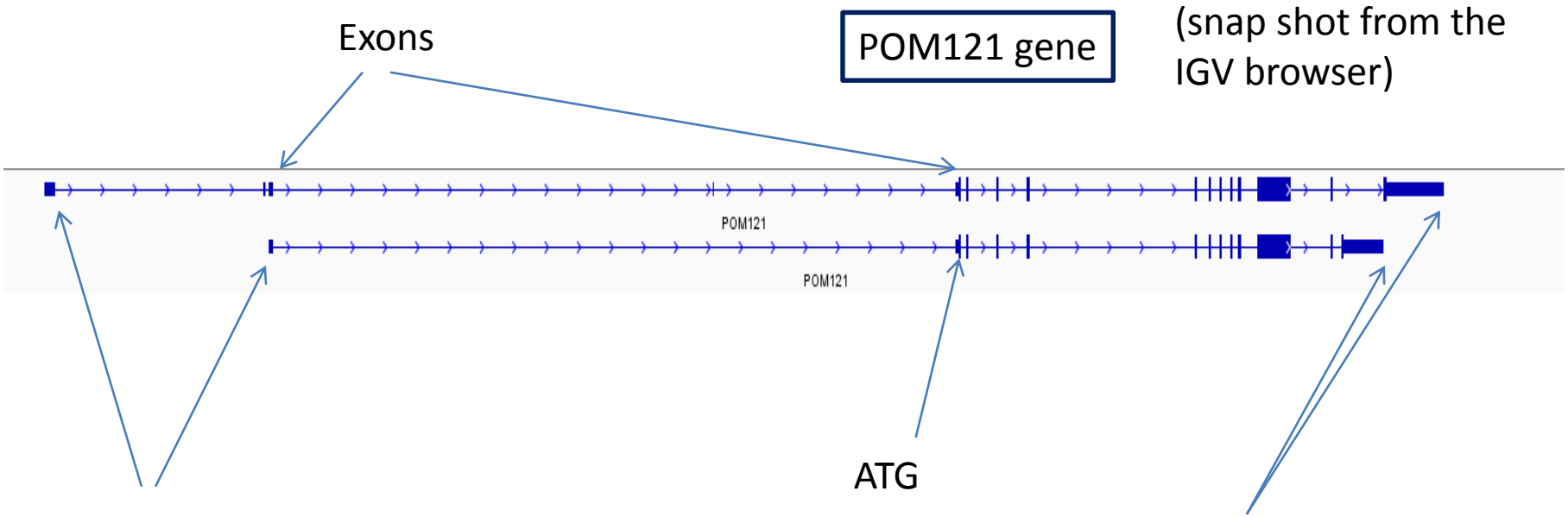
Introduction

In RNA-seq we measure the expression level of mRNAs in a given cell population

What information can we extract from RNA seq experiment?

Do comparison:
Given samples from different experimental conditions, we can find effects of the treatment on the gene expression

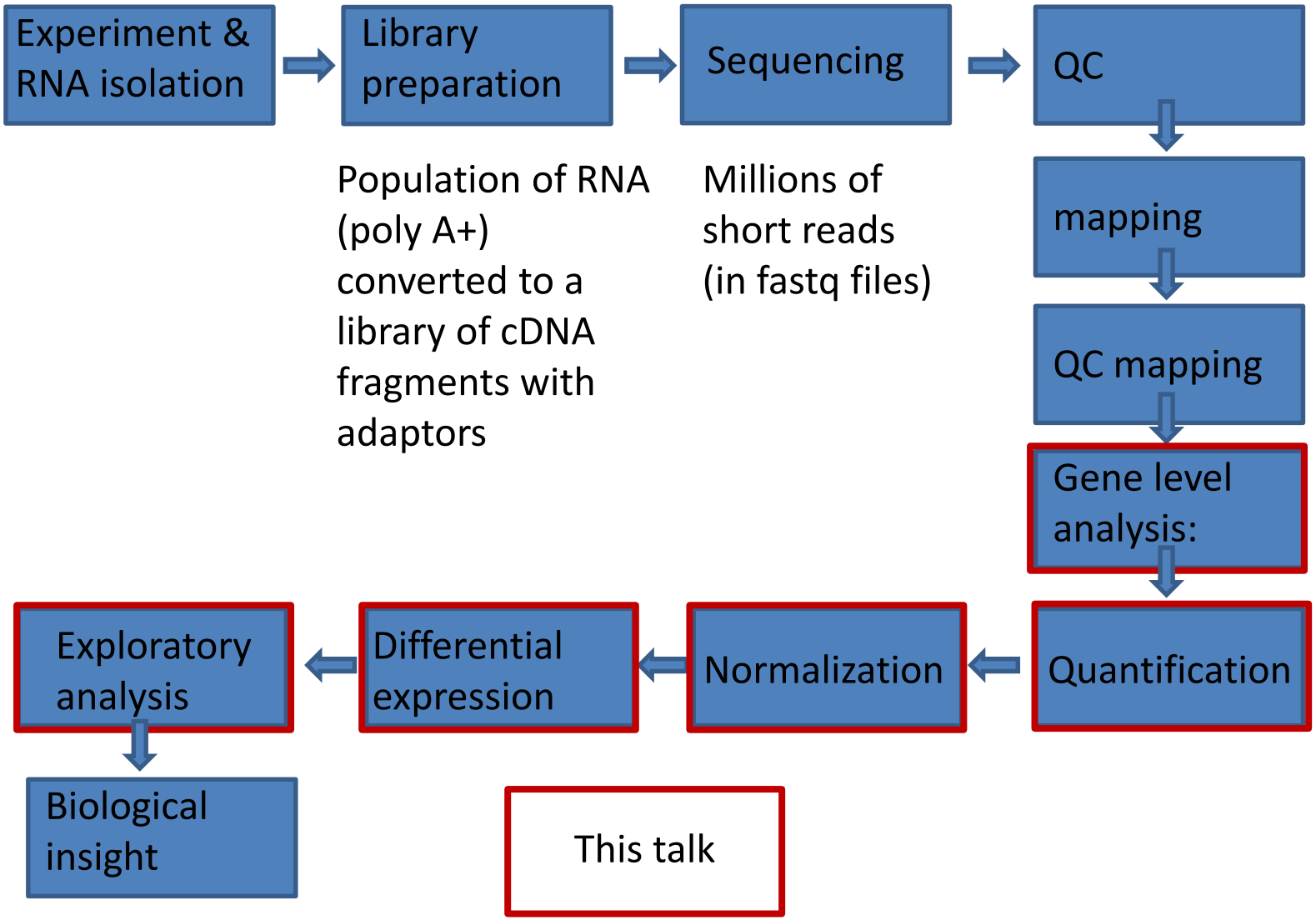
Alternative splicing



In RNA-seq experiments we can also learn:
Splice pattern and differential exon usage
Discover new transcripts: transcript boundaries and splice junctions

However in this talk we will focus on:
Differential expression at the **gene level**

The case: Available genome sequence and available gene annotations
The goal: Quantification and differential expression analysis at the gene level

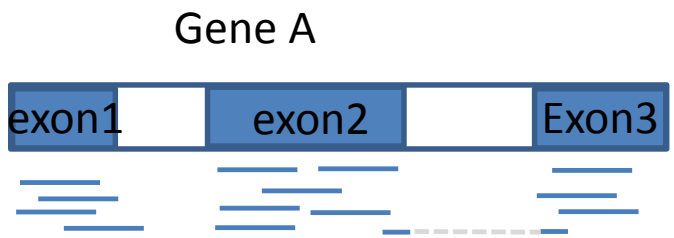


Visualization and characteristics

Outline

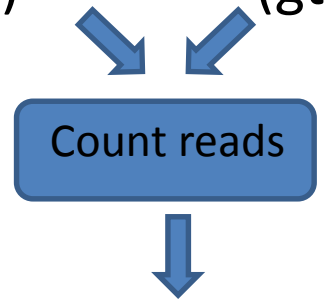
- Introduction
- Quantification of gene expression
- Normalization
- Differential Expression
- Exploratory analysis

Quantification



Input: alignment files (bam)

Input: gene annotations (gtf file)



Technically possible to do with many tools.

HTSeq-count *

Output: For each sample:

Gene	Read Count
Gene1	985
Gene2	23
Gene3	1900
.	
.	

* Simon Anders, EMBL Heidelberg

Quantification: strict counting rules

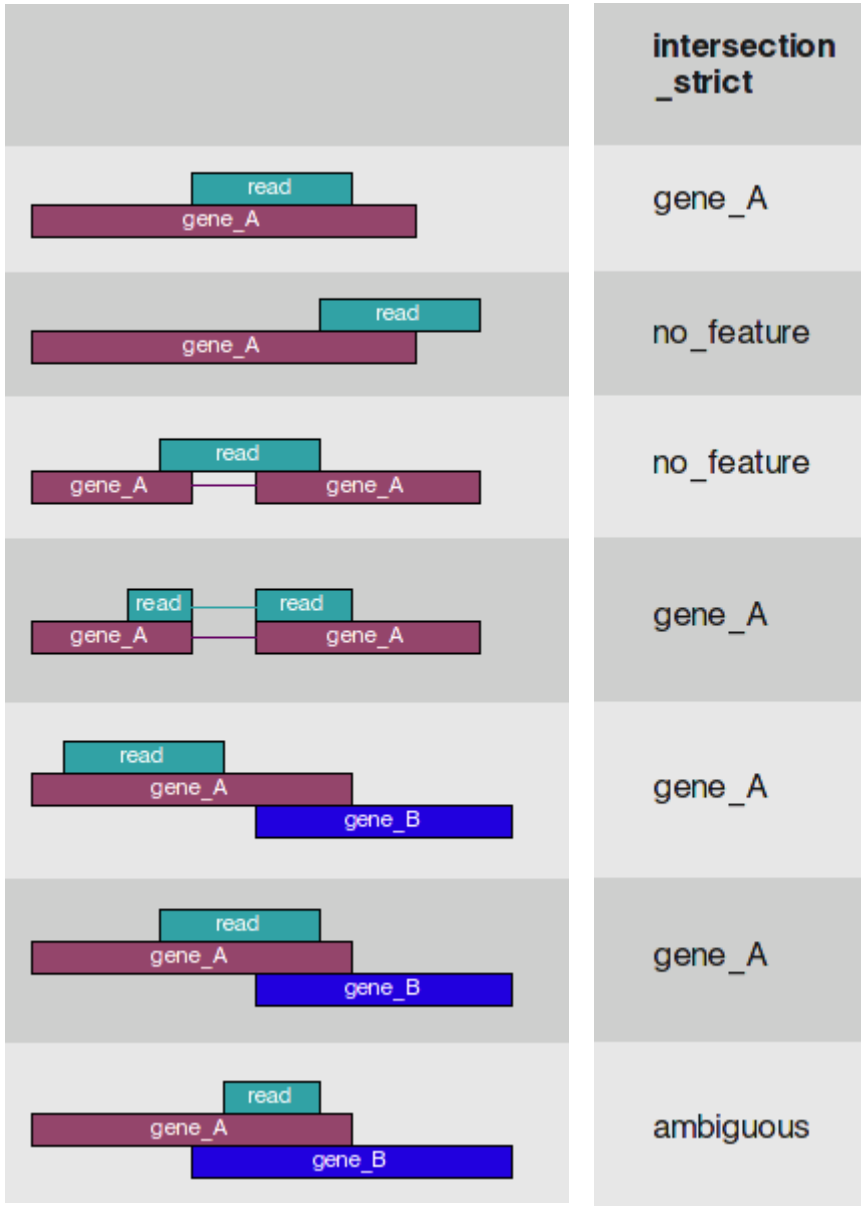
Count reads, not bases

Discard a read if it cannot be uniquely mapped (Large genomes are typically rich in sequence repeats, duplications and families of paralogous genes)

Its alignment overlaps with several genes

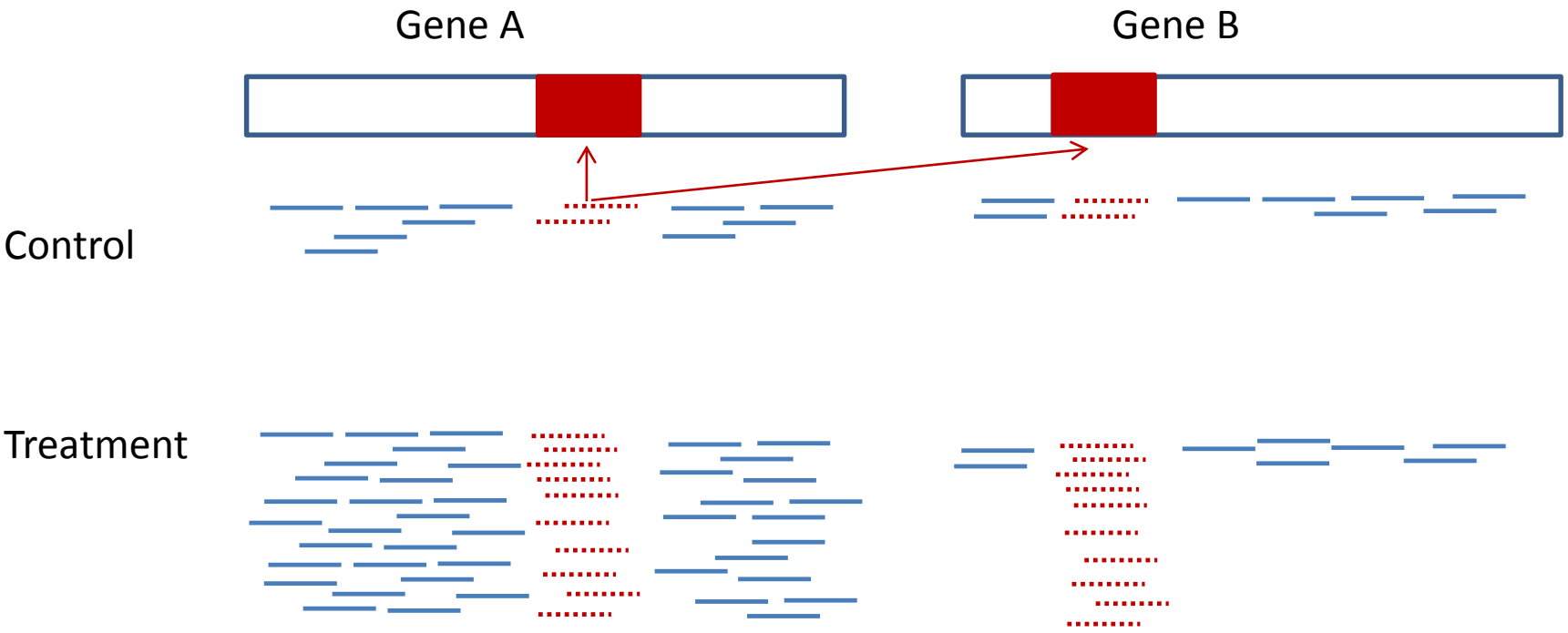
For paired end reads: the mates do not map to the same gene

Strict counting with HTSeq count



http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

Why should we discard non unique alignments?



Reads can align uniquely (that is, there is only one possible place in the genome that they could have originated from) ——

or

They can align non-uniquely (they could originate from more than one location, such as when a gene has a very close paralog in the genome) ·····

Count data table

Gene	Sample 1	Sample 2	Sample 3
Gene1	985	20	8000
Gene2	23	50	3000
Gene3	1900	1750	25

The number of reads that align to each gene provides a quantification of how many RNA transcripts of that gene were in the sample.

Compare the same gene across two conditions

Outline

- Introduction
- Quantification of gene expression
- Normalization
- Differential Expression
- Exploratory analysis

Normalization

Normalization for library size

If sample A has been sampled deeper than sample B, we expect counts to be higher.

Naive approach: Divide by the total number of mapped reads per sample

RPKM (Reads per kilobase of transcript per million reads of library)

$$RPKM = gene\ count \times \frac{10^3}{gene\ length\ (bp)} \times \frac{10^6}{total\ library\ size}$$

Mortazavi et al. Nature Methods 2008

RPKM

Corrects for total library coverage

Corrects for gene length

Most widely used measure in the beginning of the RNA-seq era

Simple, easy to understand

Has problems:
Genes that are strongly and differentially
expressed may distort the ratio of total reads

RPKM has problems!

What alternative normalizations can be applied to RNA-seq data?

RNA-seq normalization

Gene	Sample 1	Sample 2	Sample 3
G1	985	1000	1200
G2	421	450	550
G3	1900	1750	1891

- There are various alternatives
- DESeq alternative
- DESeq* is an R package for normalization and differential expression of RNA seq data
- Based on the hypothesis that most genes are not differentially expressed

* DESeq author: Simon Anders, EMBL Heidelberg

RNA-seq normalization

Gene	Sample 1	Sample 2	Sample 3	Geom. mean
G1	985	1000	1200	1057.3
G2	421	450	550	470.5
G3	1900	1750	1891	1845.7

- There are various alternatives
- DESeq alternative
- DESeq* is an R package for normalization and differential expression of RNA seq data
- Based on the hypothesis that most genes are not differentially expressed
- For each gene: calculate the ratio of its read count over its geometric mean across all lanes.
The underlying idea is that non-DE genes should have similar read counts across samples, leading to a ratio of 1
- Assuming most genes are not DE, the median of this ratio should be 1
- A normalizing factor is calculated for each sample to fulfill the hypothesis

* DESeq author: Simon Anders, EMBL Heidelberg

Outline

- Introduction
- Quantification of gene expression
- Normalization
- Differential Expression
- Exploratory analysis

Differential expression

Statistical models: Is a gene differentially expressed?

Microarrays traditionally used continuous statistical tests (t-test ANOVA etc)

- Can we use this also for RNA seq?
- **NO!!**
- Because the data is:
  discrete data
  not continuous
  do not have equal variance across all the read counts

Poisson distribution in RNA-seq

The Poisson distribution turns up whenever things are counted

The number of reads mapped to a gene from a sample can be modeled as independent Poisson random variables

Poisson distribution

The variance is equal to the mean

It can only describe the counting noise and nothing else.

Gene	Sample 1	Sample 2
Gene1	1	2
Gene2	100	200

Less certain

In RNA Seq, noise depends on count level.

Why?

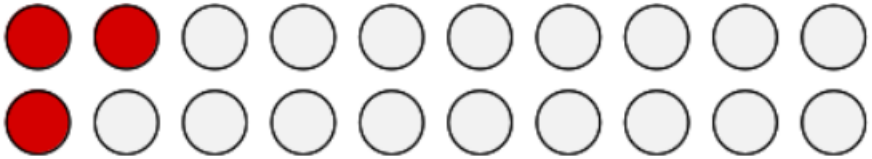
It is more difficult to detect small fold changes in genes with low read counts than in those with high read counts

This bag contains very many marbles, 10% of which are red

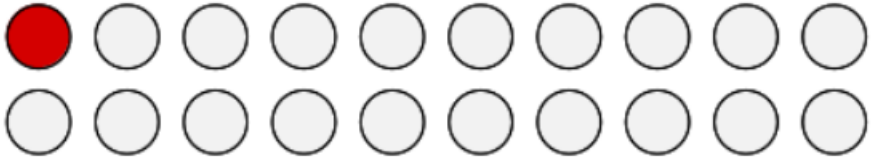
Several experimenters are tasked with determining the percentage of red balls

Each of them is permitted to draw 20 marbles out of the bag, without looking

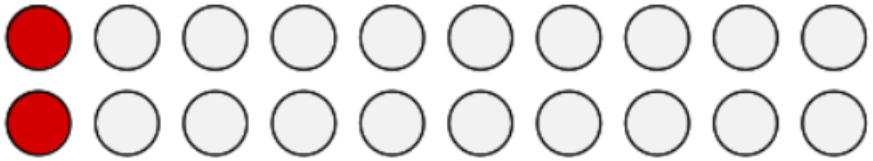
http://www.bioconductor.org/help/course-materials/2013/CSAMA2013/tuesday/morning/Anders__DESeq_DEXSeq.pdf



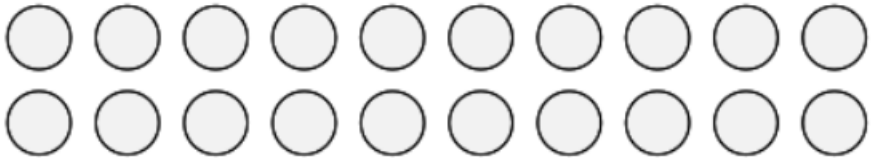
3 / 20 = 15%



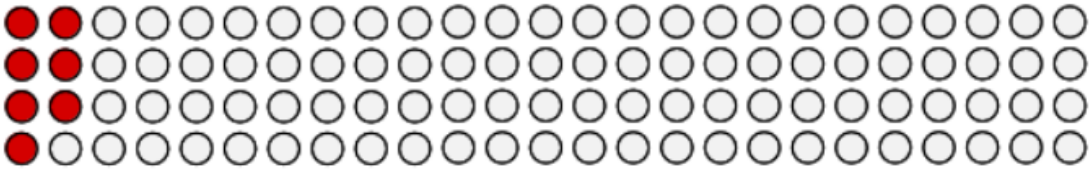
1 / 20 = 5%



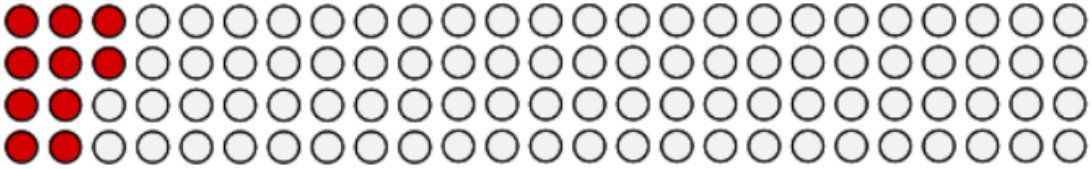
2 / 20 = 10%



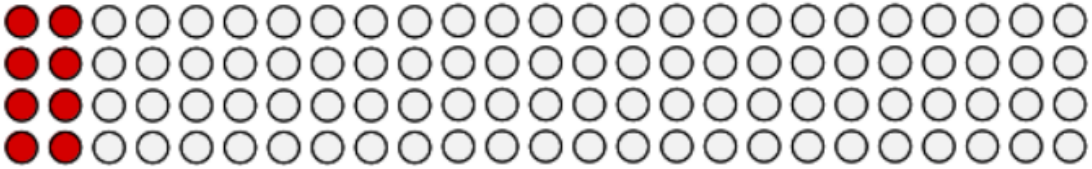
0 / 20 = 0%



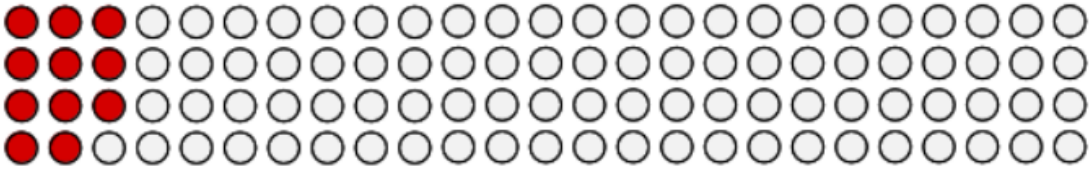
$7 / 100 = 7\%$



$10 / 100 = 10\%$



$8 / 100 = 8\%$



$11 / 100 = 11\%$

Instead of a red marble – gene…

Noise in RNA-seq data

Shot Noise

Unavoidable, appears also with perfect replication (Poisson distribution)

Can be calculated

Biological noise and Technical noise (sample preparation; negligible)

Need to be estimated from the data

The Poisson distribution doesn't model real RNA seq data very well: also have biological noise and technical artifact.

➡ Use negative binomial distribution better fit than Poisson distribution

The negative binomial distribution

The negative binomial distribution has 2 parameters

In RNA-seq data the sample variance exceeds the sample mean

Use the negative binomial distribution instead

DESeq (as well as other packages) uses the negative binomial distribution for the statistics for RNA-seq

Input for DESeq: gene counts

DESeq: If you have two different experimental conditions, with replicates
For a given gene: the change in the expression strength between the two conditions is large as compared to the variation within each group (based on the negative binomial distribution).

DESeq output

For each comparison between two conditions:

Fold change
P-value

The P-value indicates the probability that the observed difference between treatment and control or an even stronger one, is observed even though the there is no true treatment effect

<u>Adjustment of the p-value to multiple testing:</u>
If we are testing ~30,000 genes for differential gene expression, and we use a significance cut off of $p<0.01$, then we should expect to call approximately 300 genes differentially expressed by random chance.

Methods to correct for multiple comparisons

After DESeq:

Treat / Control

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Raw counts				DESeq normalized counts				DESeq statistics			
2	Gene	Control1	Control2	Treat1	Treat2	Control1	Control2	Treat1	Treat2	Ratio	log2 Ratio	pval	padj
3	ABHD4	610	6154	10142	6903	660.512	3523.5	10229.4	10603.5	0.20084	-2.3159	9.44E-06	0.00108
4	ACBD7	1892	4557	403	228	2048.67	2609.13	406.473	350.224	6.15544	2.62186	2.29E-05	0.00234
5	ADAM10	8371	6821	19763	19112	9064.18	3905.39	19933.3	29357.3	0.26312	-1.92618	0.00029	0.02139
6	ADI1	3050	2279	6381	4175	3302.56	1304.85	6435.99	6413.09	0.35858	-1.47964	0.00054	0.03555
7	AIM1	6428	17335	577	345	6960.28	9925.23	581.973	529.944	15.186	3.92467	1.81E-12	5.89E-10
8	A1BG	439	590	307	192	475.352	337.807	309.646	294.925	1.34502	0.42763	0.70702	1
9	A1BG-AS1	200	347	123	101	216.561	198.676	124.06	155.143	1.48723	0.57262	0.65709	1
10	.	.	.	.	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.	.	.	.	.	.	.

How to Choose genes:

Fold change and p-value

DESeq has much more to offer

Modeling of the experimental design

Transformations of the counts for further exploratory analysis

Above the scope of today's talk…

Validations

Important to perform some sort of validation to ensure that the experimental findings are correct

qPCR
On the same sample: This provides an external measurement, but only speaks to the technical accuracy of the experiment

Biological validation: necessary to perform validation on biological samples that are independent of those used in the original experiment. Not always practical, but provides better support

Outline

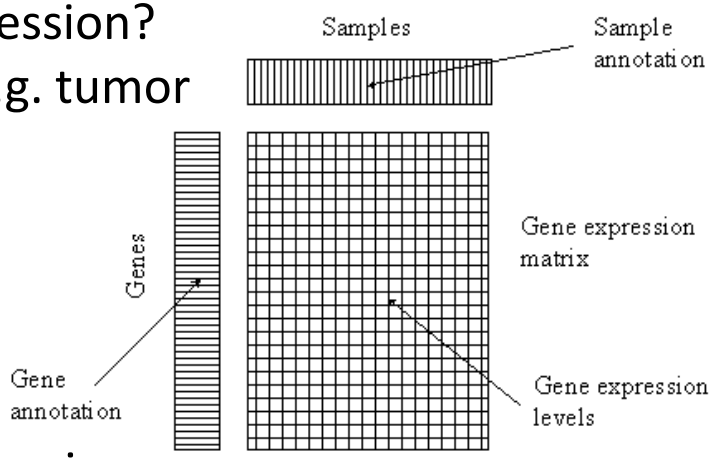
- Introduction
- Quantification of gene expression
- Normalization
- Differential Expression
- Exploratory analysis

Exploratory analysis

The goal: understanding general characteristics of the data
Visualizing data

Classification of samples:
Quality control: Are the replicates similar to one another?
Batch effects are systematic non-biological experimental variation
Do we have batch effects?
Which treatments/samples have similar expression?
Identify new classes of biological samples (e.g. tumor subtypes)

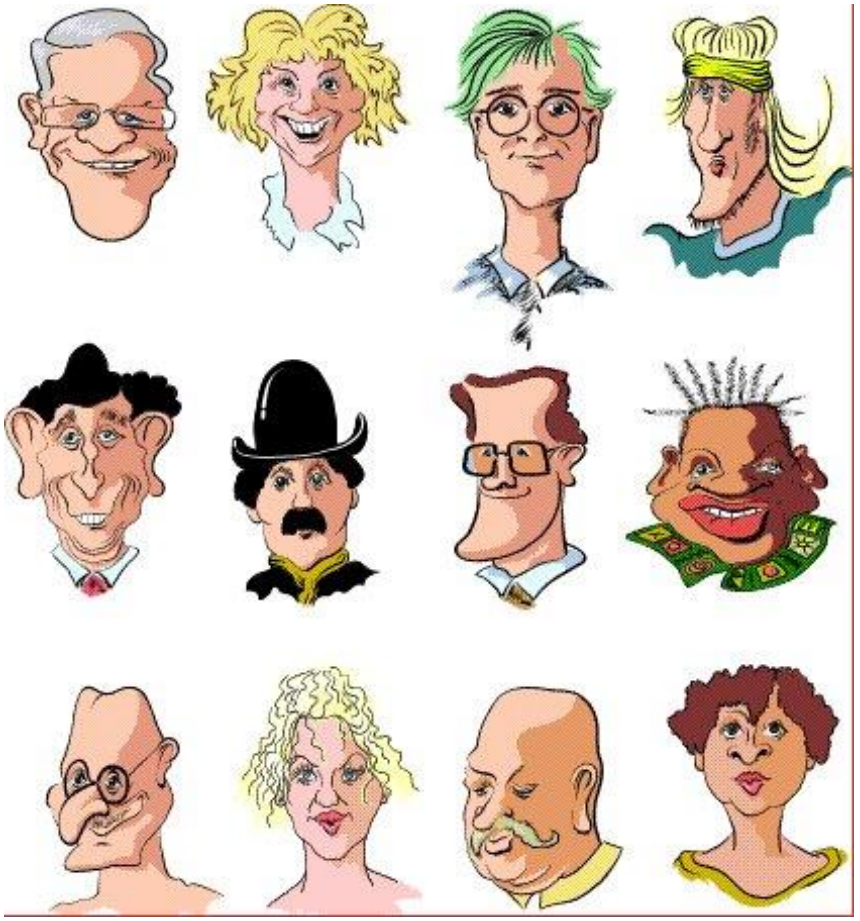
Classification of genes:
(We will filter the genes before doing so)
Which genes are regulated together?
Identify typical temporal or spatial gene expression patterns (e.g. cell cycle data).



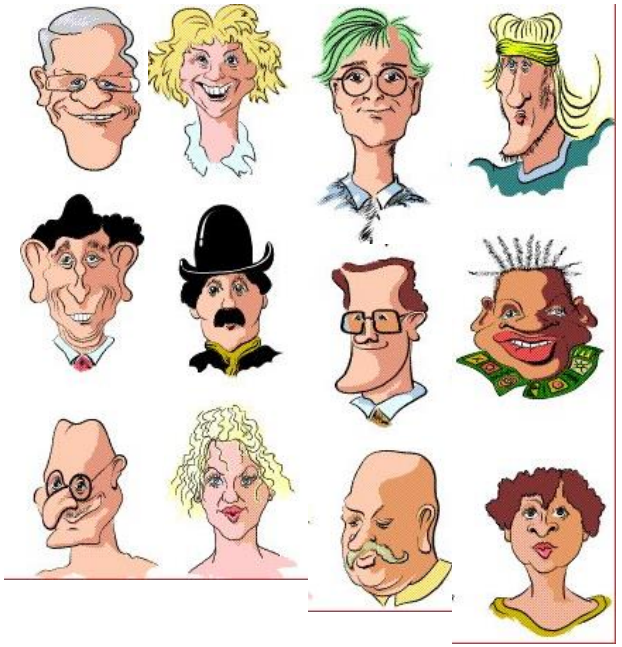
What is clustering?

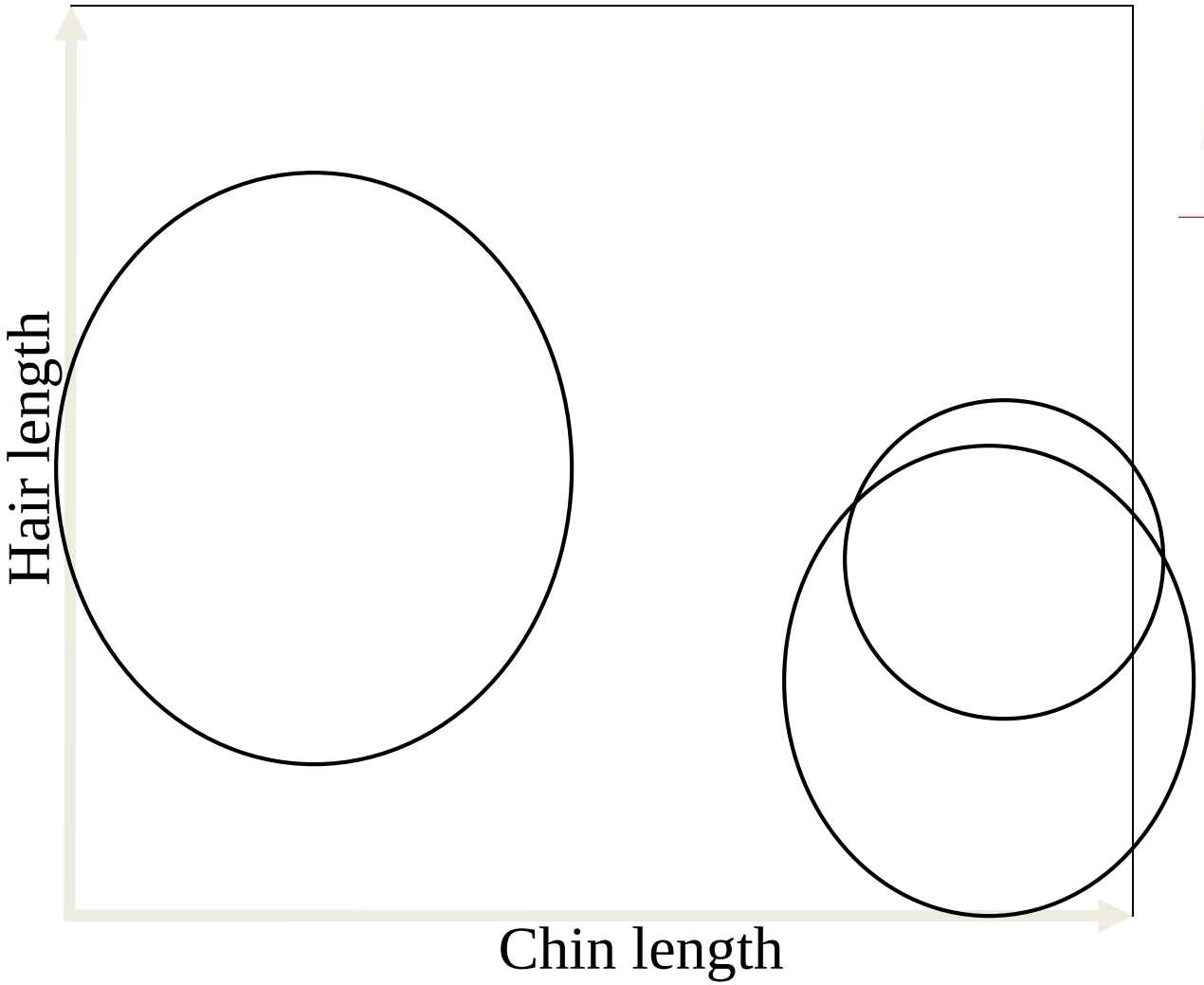
Clustering is a method to classify objects into groups according to their similarity

The objects within each cluster are more **closely related** to each other than to objects in other clusters



Similarity Matrix





People in n-dimentional characteristics space

Characters

People

	Chin	Hair	Hat	Nose	Glasses	Neck
Person1	5	0	0	5	3	2
Person2						
Person3						
Person4						
Person5						
Person6						
Person7						
Person8						
Person9						

Genes in n-dimentional experimental conditions space

RNA samples

Genes

	Heart	Uterus	Liver	Kidney	Pancreas	Muscle
Gene1	5.72	9.36	4.12	4.85	4.75	5.51
Gene2						
Gene3						
Gene4						
Gene5						
Gene6						
Gene7						
Gene8						
Gene9						

Cluster analysis

Generally, cluster analysis is based on two ingredients:

- **Distance measure**: Quantification of similarity of objects.
- **Cluster algorithm:** A procedure to group objects.
  Aim: small within-cluster distances, large between-cluster distances.

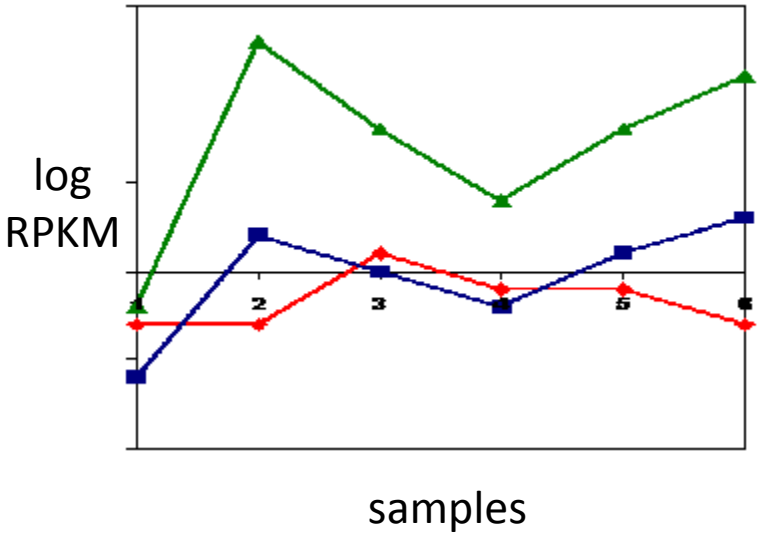
Distance measure – 2 main families:

Euclidean - depends on the point to point differences and accounts for absolute differences. It measures the average distance between vectors

Pearson correlation - that accounts for trends

Which distance measure to use?

- The choice of distance measure should be based on what you are interested in. What sort of similarities would you like to detect?
- In many situations correlation has more biological meaning



Standardization

- Some times we would like to apply **standardization** to the observations:

  For each gene:
  Subtract mean and divide by standard deviation:

$$x \mapsto \frac{x - \bar{x}}{\hat{\sigma}_x}$$

- After standardization, Euclidean and correlation distance are equivalent

Cluster analysis

Generally, cluster analysis is based on two ingredients:

- **Distance measure**: Quantification of (dis-)similarity of objects.
- **Cluster algorithm:** A procedure to group objects. Aim: small within-cluster distances, large between-cluster distances.

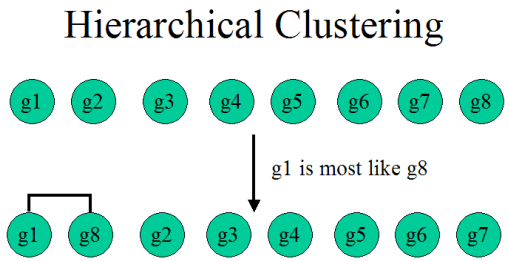
Cluster algorithm

Hierarchical

K-means

Others…

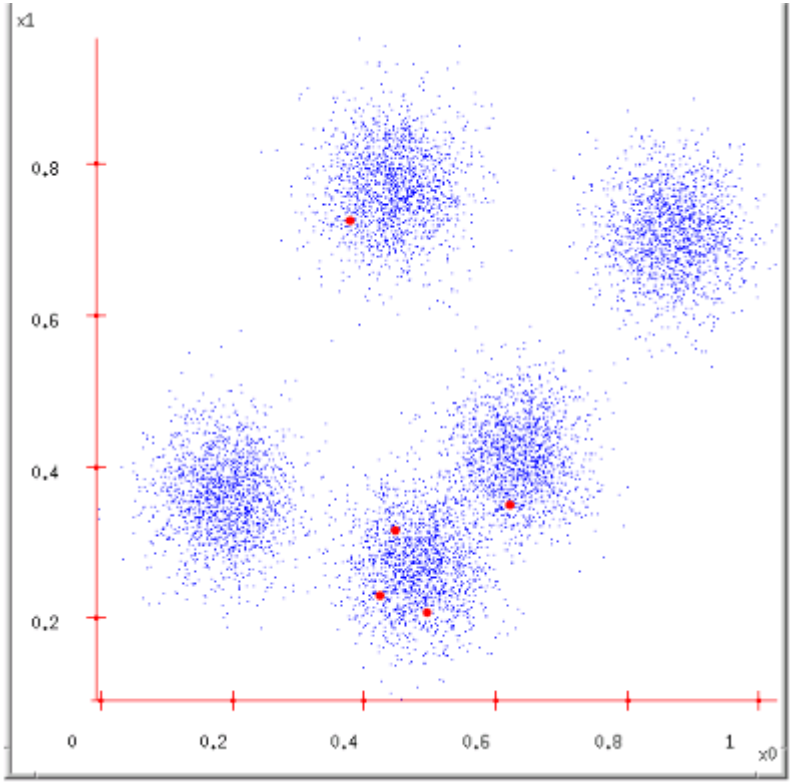
Hierarchical cluster algorithms



- In each iteration, merge the two clusters with the minimal distance from each other - until you are left with a single cluster comprising all objects.

K-Means clustering – partitioning clustering

The users need to choose the number of clusters
(e.g. k=5)

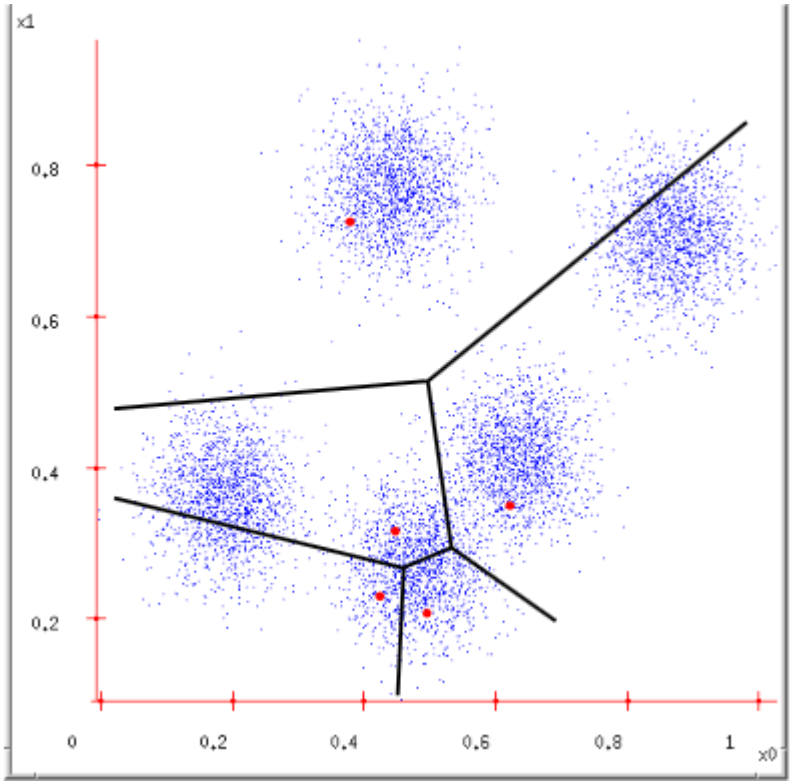


Randomly guess k cluster Center locations

Example taken from http://www.autonlab.org/tutorials/kmeans11.pdf

K-Means clustering – partitioning clustering

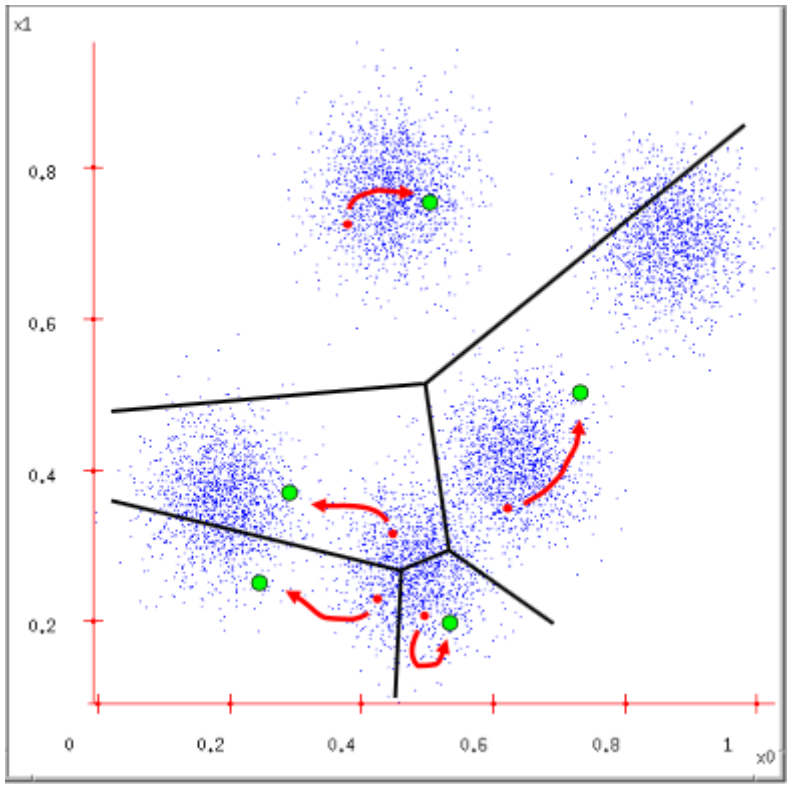
Each datapoint finds out which Center it's closest to (distance measure) (Thus each Center "owns" a set of datapoints)



Example taken from http://www.autonlab.org/tutorials/kmeans11.pdf

K-Means clustering – partitioning clustering

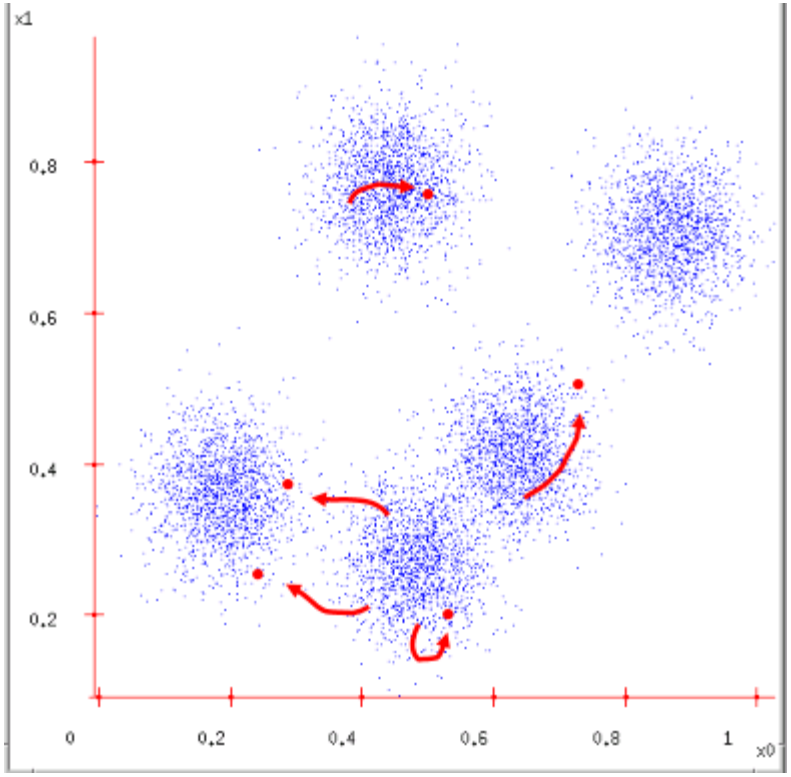
Each Center finds the centroid of the points it owns



Example taken from http://www.autonlab.org/tutorials/kmeans11.pdf

K-Means clustering – partitioning clustering

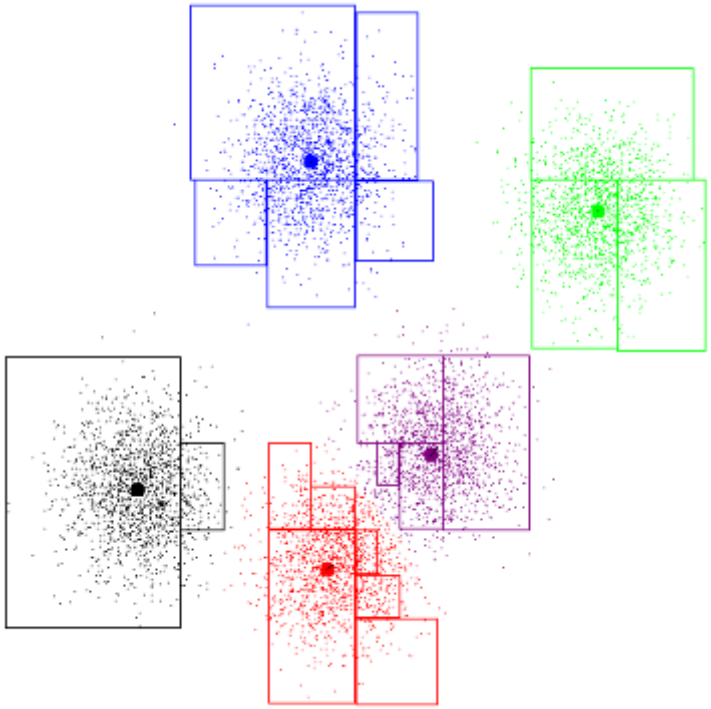
...and jumps there



This process is repeated until terminated

Example taken from http://www.autonlab.org/tutorials/kmeans11.pdf

At the end of the process:



Cluster algorithm

Hierarchical

K-means

Others…
CLICK * – utilizes graph theoretic and statistical techniques to identify groups of highly similar elements
Advantage: the user chooses homogeneity of the cluster; no need to choose the number of clusters

Partitioning clustering – good for creating groups that will be used for downstream pathway analysis

* Prof. Ron Shamir, TAU

Biological example

4 cell lines:

1. Leukemia cell line 1 with a specific translocation
2. Leukemia cell line 2 with a specific translocation
3. Leukemia cell line 3 (without the translocation)
4. Leukemia cell line 4 (without the translocation)

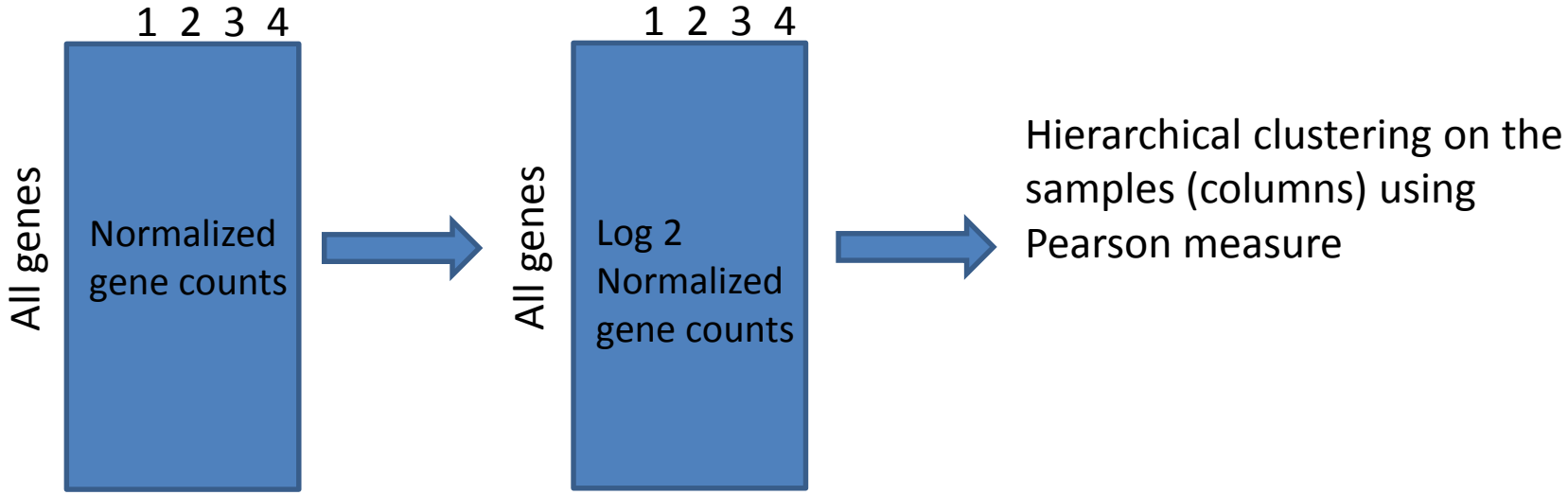
After mapping the reads, counting the reads on the features (genes) and DESeq *

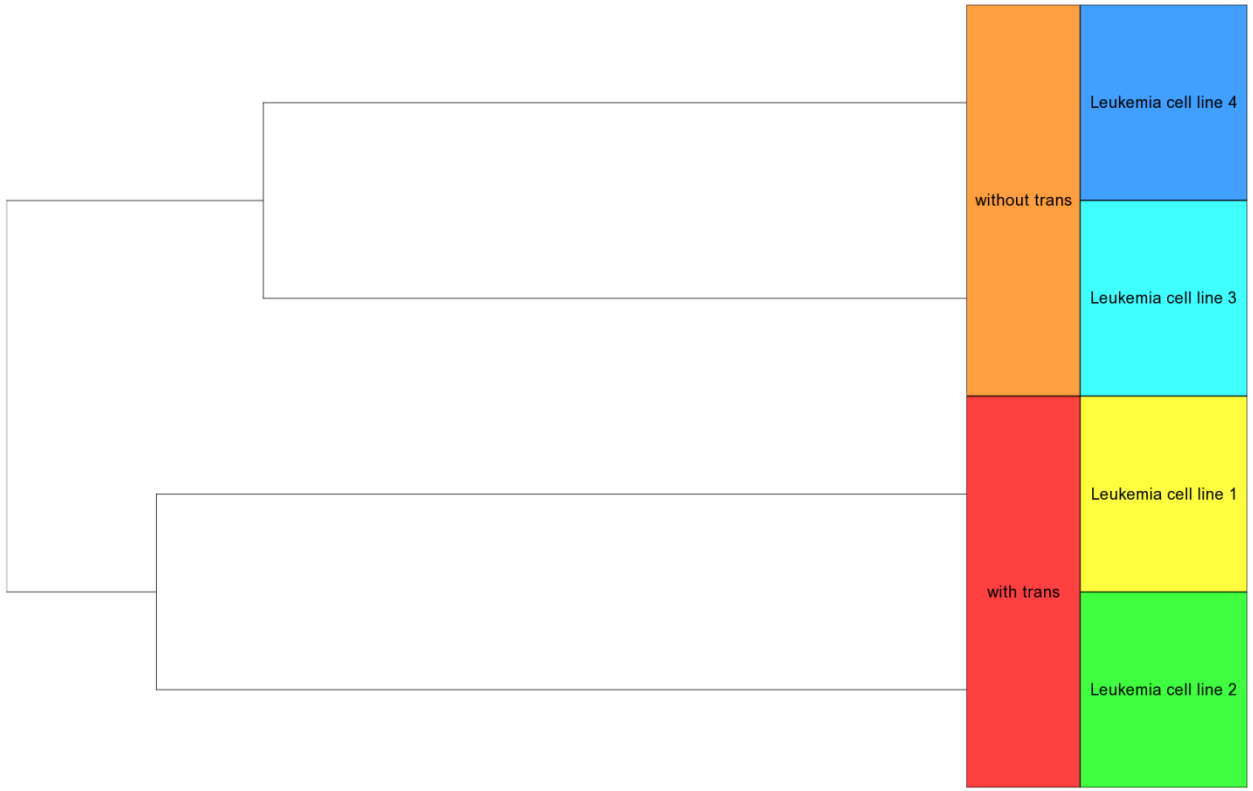
$$\frac{\text{Average(cancer1 with trans, cancer 2 with trans)}}{\text{Average(cancer3 without trans, cancer 4 with trans)}}$$

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Raw counts				DESeq normalized counts				DESeq statistics			
2	Gene	htseq count cancer 1	htseq count cancer 2	htseq count cancer 3	htseq count cancer 4	cancer 1 with trans	cancer 2 with trans	cancer 3	cancer 4	Ratio	log2 Ratio	pval	padj
3	ABHD4	610	6154	10142	6903	660.512	3523.5	10229.4	10603.5	0.20084	-2.3159	9.44E-06	0.00108
4	ACBD7	1892	4557	403	228	2048.67	2609.13	406.473	350.224	6.15544	2.62186	2.29E-05	0.00234
5	ADAM10	8371	6821	19763	19112	9064.18	3905.39	19933.3	29357.3	0.26312	-1.92618	0.00029	0.02139
6	ADI1	3050	2279	6381	4175	3302.56	1304.85	6435.99	6413.09	0.35858	-1.47964	0.00054	0.03555
7	AIM1	6428	17335	577	345	6960.28	9925.23	581.973	529.944	15.186	3.92467	1.81E-12	5.89E-10
8	A1BG	439	590	307	192	475.352	337.807	309.646	294.925	1.34502	0.42763	0.70702	1
9	A1BG-AS1	200	347	123	101	216.561	198.676	124.06	155.143	1.48723	0.57262	0.65709	1
10	.	.	.	.	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.	.	.	.	.	.	.

* Analyzed by Dr. Dena Leshkowitz

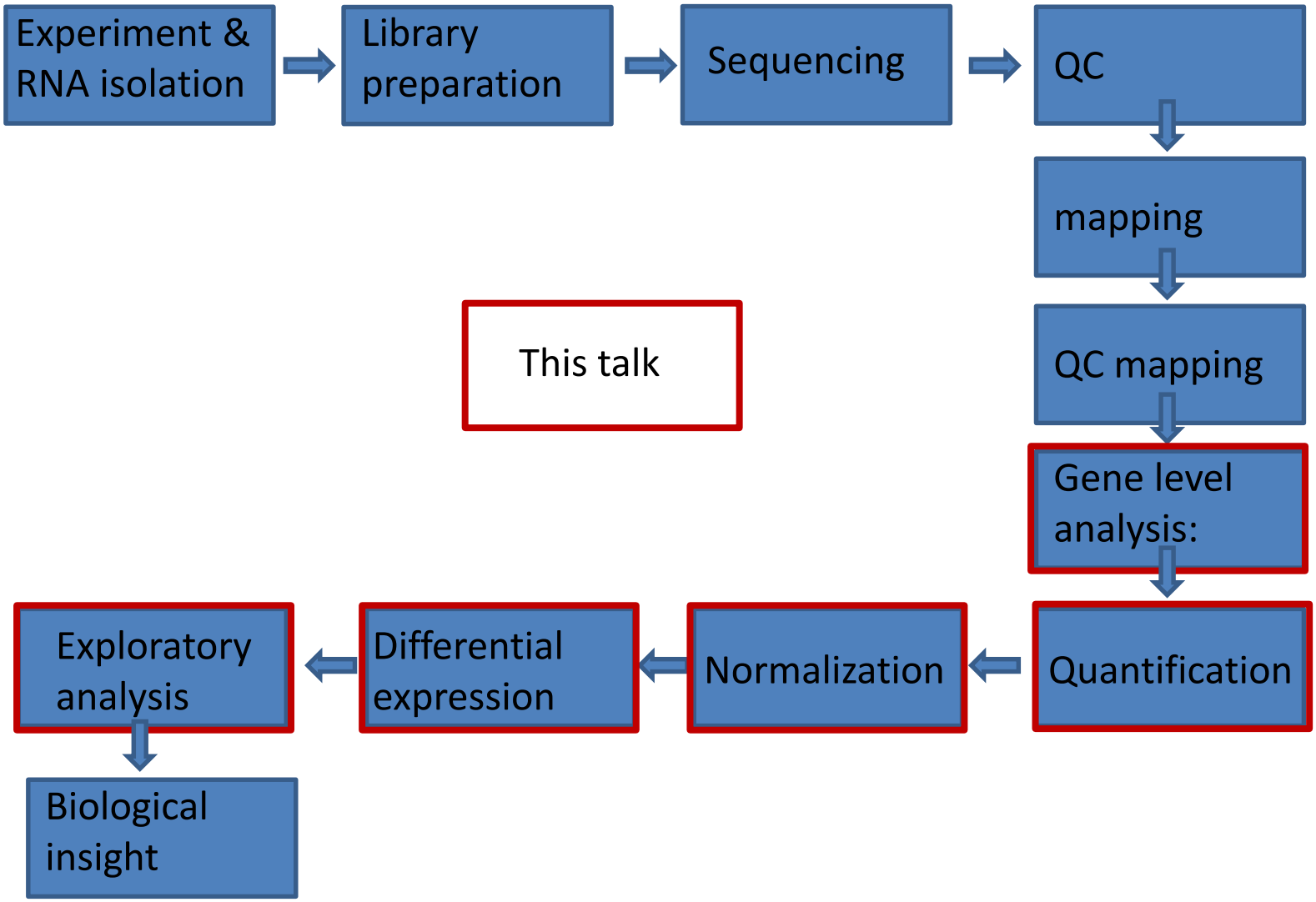
Are the cell lines that carry the translocation are more similar to each other, compared to the cell line without the translocation?





The Leukemia cell lines cluster according to the existence of the translocation

Summary


Experiment & RNA isolation
Library preparation
Sequencing
QC
mapping
QC mapping
Gene level analysis:
Quantification
Normalization
Differential expression
Exploratory analysis
Biological insight
This talk

Thanks

Dr. Ester Feldmesser
for sharing slides

Tutorial

In the tutorial, you will cluster the genes in the Leukemia data-set

\\ngs001\Open_Data\Course2015-exercise1\RNA_seq_cluster

ClusteringExercise_June_2015.pdf


THANK YOU!