Vall d'Hebron
Institut de Recerca

HANDS ON: Introduction to RNA-Seq Differential Expression Analysis

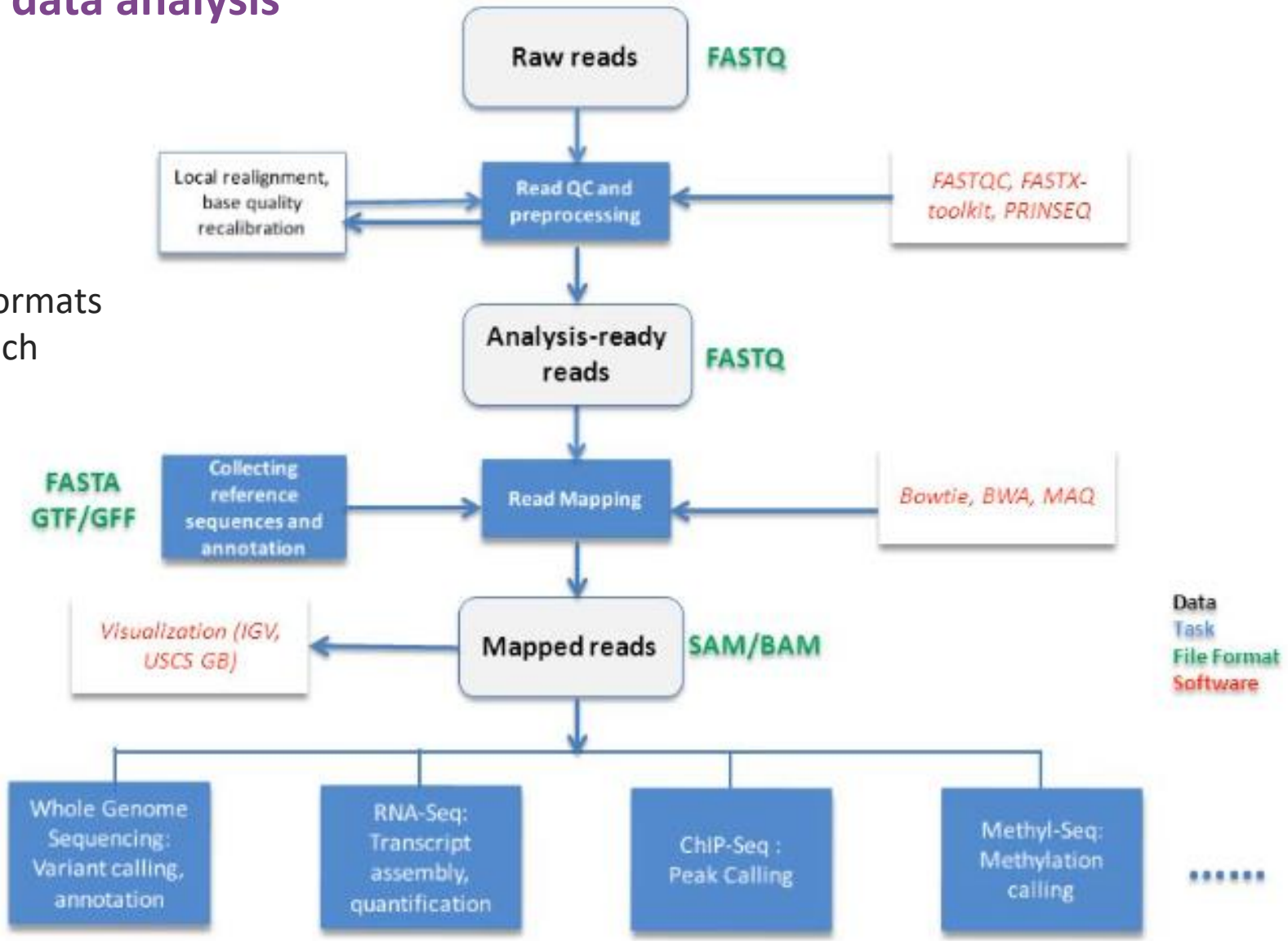
Bioinformàtica per a la Recerca Biomèdica

Mireia Ferrer[1] , Esther Camacho[1], Berta Miró[1], Álex Sánchez[1,2]
1 Unitat d'Estadística i Bioinformàtica (UEB) VHIR
2 Departament de Genètica, Microbiologia i Estadística, UB

Vall d'Hebron Hospital · Generalitat de Catalunya Departament de Salut · Institut Català de la Salut · UAB Universitat Autònoma de Barcelona · Generalitat de Catalunya Departament d'Economia i Coneixement · Institució CERCA Centres de Recerca de Catalunya · Instituto de Salud Carlos III

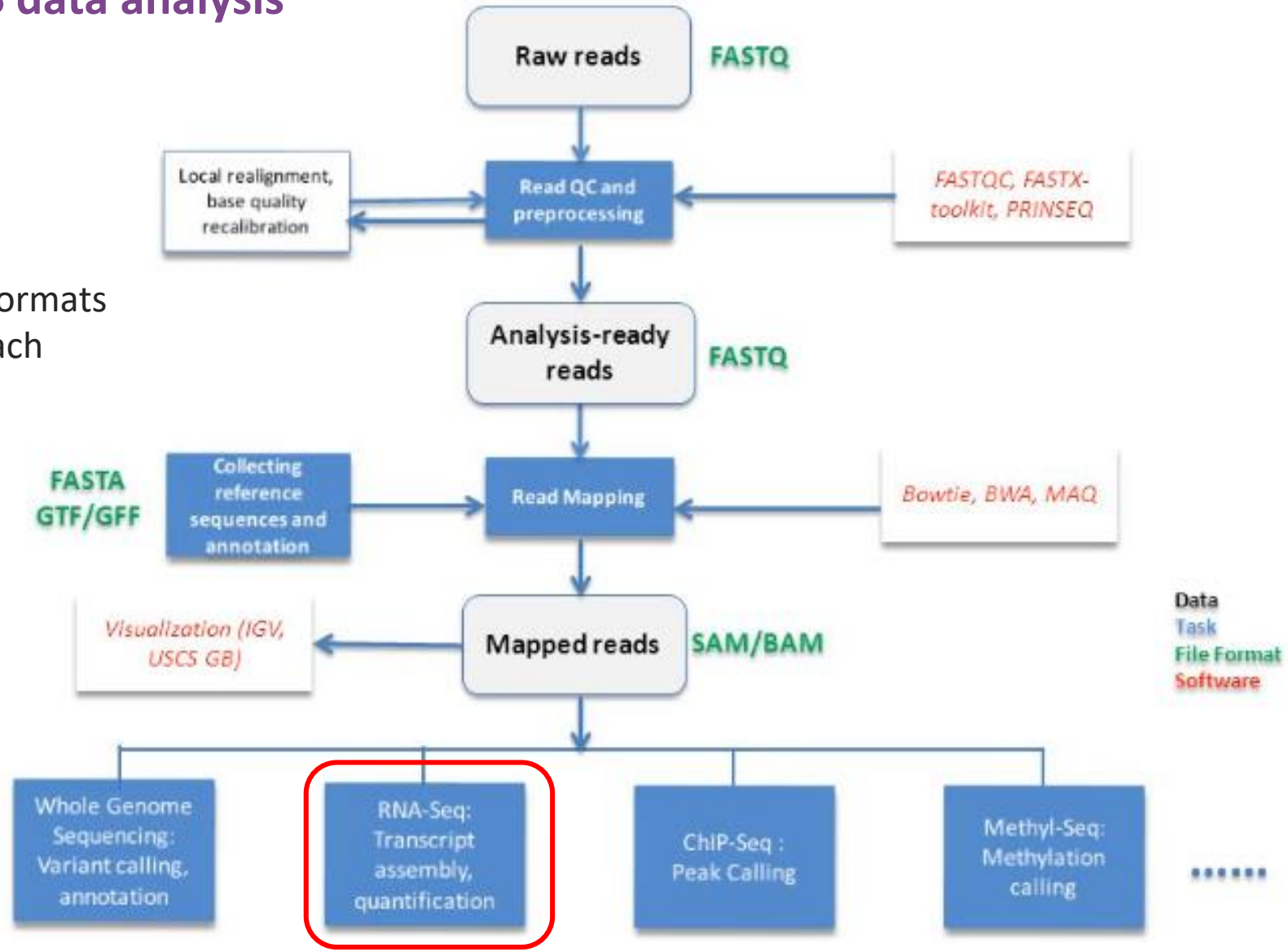
Steps in NGS data analysis

- We will have different data formats and tools for each step

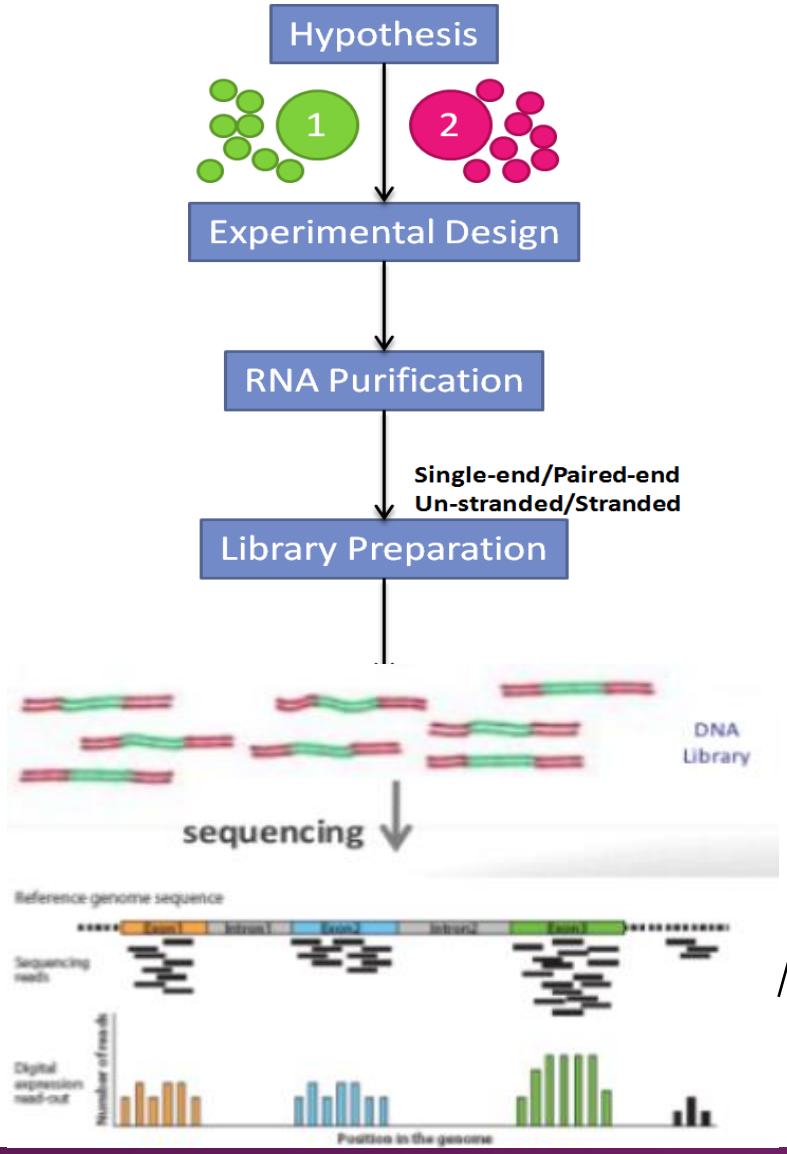


Steps in NGS data analysis

- We will have different data formats and tools for each step



RNA-seq Expression Analysis Workflow



Differential expression analysis

Table of counts

Normalization

- Between sample: TMM, upper quartile
- Within sample: RPKM, quantile

DE testing

- Poisson test: DEGseq
- Negative binomial test: edgeR, baySeq, DEseq

List of differentially expressed genes

Systems Biology

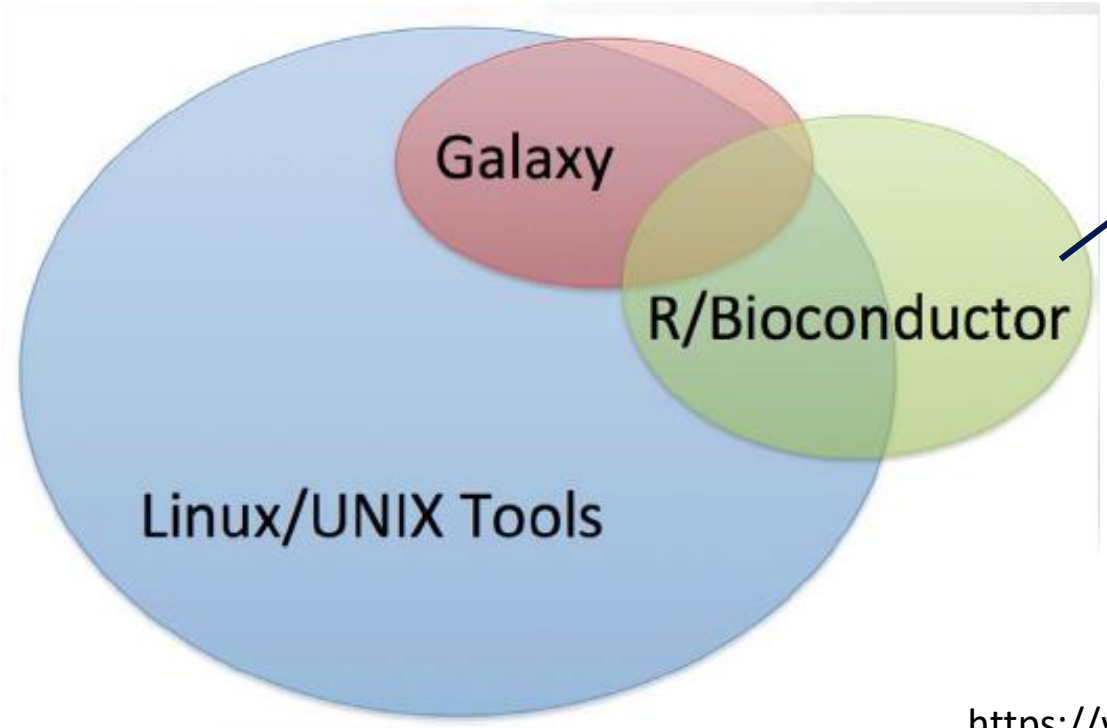
- Test for enriched categories/ pathways: GOseq
- Infer networks and integrate with other data

Biological insight

Wednesday

Tools for NGS data analysis

Highly efficient and fast processing tools are required to handle large volume of datasets



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It

https://www.bioconductor.org/

Today's practicum

Steps covered by the tutorial:

- ~~Start with the FASTQ files (how they are aligned to the genome)~~

1. Summarization/Quantification of aligned reads: obtaining the counts matrix
2. Data pre-processing and exploratory analysis
3. Differential gene expression analysis
4. Visually explore the results

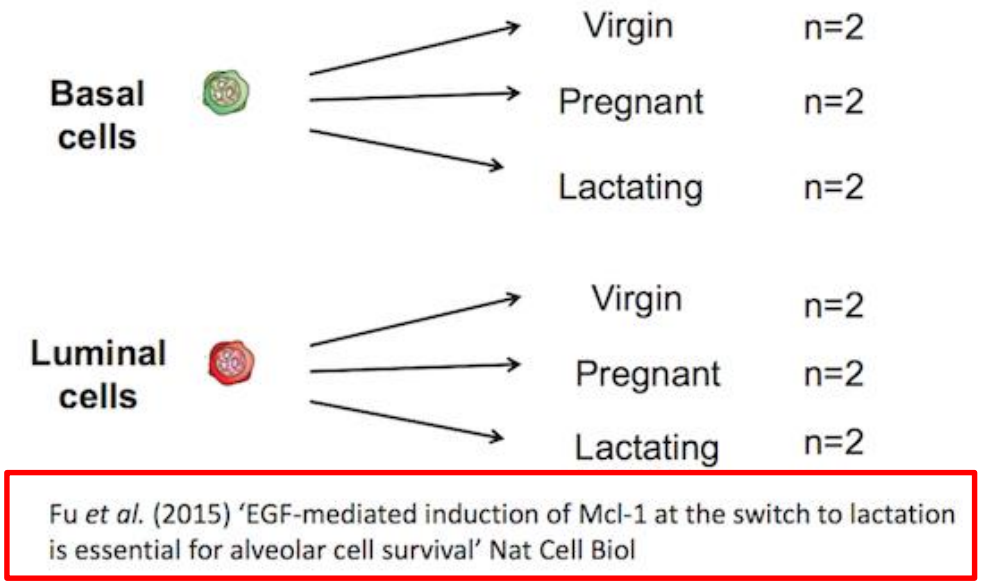
This material has been created using the following sources:

- https://combine-australia.github.io/RNAseq-R/
- https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html

Study overview

- **Objective: We want to identify genes differentially expressed in the lactating versus pregnant mammary gland**

Dataset: RNA-seq data of mouse mammary gland (GSE60450)



Fu *et al.* (2015) 'EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival' Nat Cell Biol

Note: two biological replicates are used here, however three replicates is usually recommended as a minimum requirement for RNA-seq.

GSE60450

Scope: Self Format: HTML Amount: Quick GEO accession: GSE60450

Series GSE60450 Query DataSets for GSE60450

Status	Public on Jan 19, 2015
Title	Transcriptome analysis of luminal and basal cell subpopulations in the lactating versus pregnant mammary gland
Organism	Mus musculus
Experiment type	Expression profiling by high throughput sequencing
Summary	To identify genes specifically expressed in lactating mammary glands, the gene expression profiles of luminal and basal cells from different developmental stages were compared.
Overall design	Comparison of gene expression in luminal and basal cells harvested from the mammary glands of virgin, 18.5 day pregnant and 2 day lactating mice (2 mice per stage).
Contributor(s)	Fu NY, Lun A, Smyth GK, Visvader JE
Citation(s)	Fu NY, Rios AC, Pal B, Soetanto R et al. EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival. *Nat Cell Biol* 2015 Apr;17(4):365-75. PMID: 25730472

Platforms (1)	GPL13112 Illumina HiSeq 2000 (Mus musculus)
Samples (12)	GSM1480291 Luminal virgin #1
More...	GSM1480292 Luminal virgin #2
	GSM1480293 Luminal 18.5 dP #1

Relations

BioProject	PRJNA258286
SRA	SRP045534

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60450

Organism: Mus musculus

Library:

Instrument: Illumina HiSeq 2000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: SINGLE
Construction protocol: FACS-sorted cells were
for sequencing using standard Illumina protoco

Vall d'Hebron
Institut de Recerca

Getting started

Hands on!

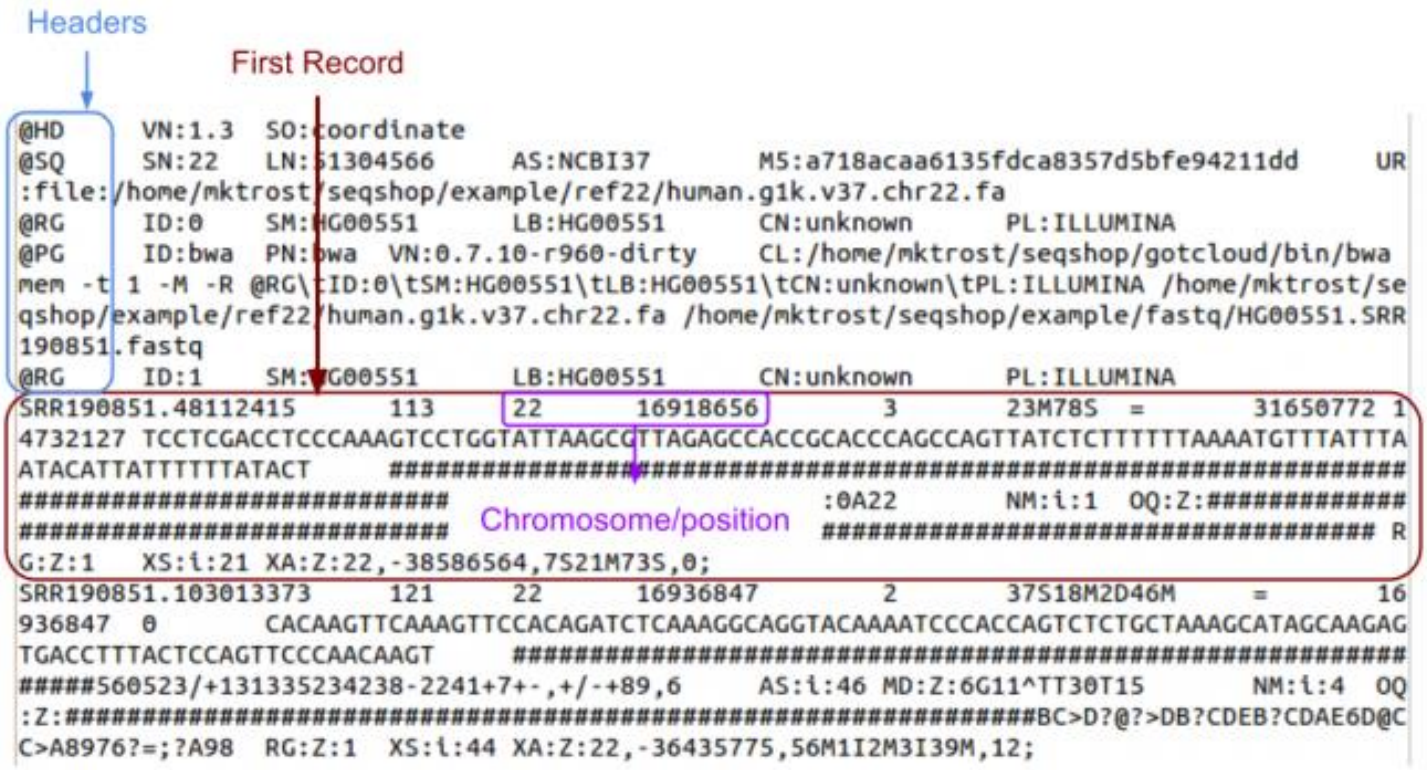
... follow **sections 1-2 of the Rmd file**

Steps covered in this tutorial:

1. **Summarization/Quantification of aligned reads: obtaining the counts matrix**
2. Data pre-processing and exploratory analysis
3. Differential gene expression analysis
4. Visually explore the results

Step 1: Summarization/Quantification of reads: obtaining the counts matrix

- The alignment produces a set of **BAM** files, where each file contains the read alignments for each sample.
- In the BAM file, there is a chromosomal location for every read that was mapped.

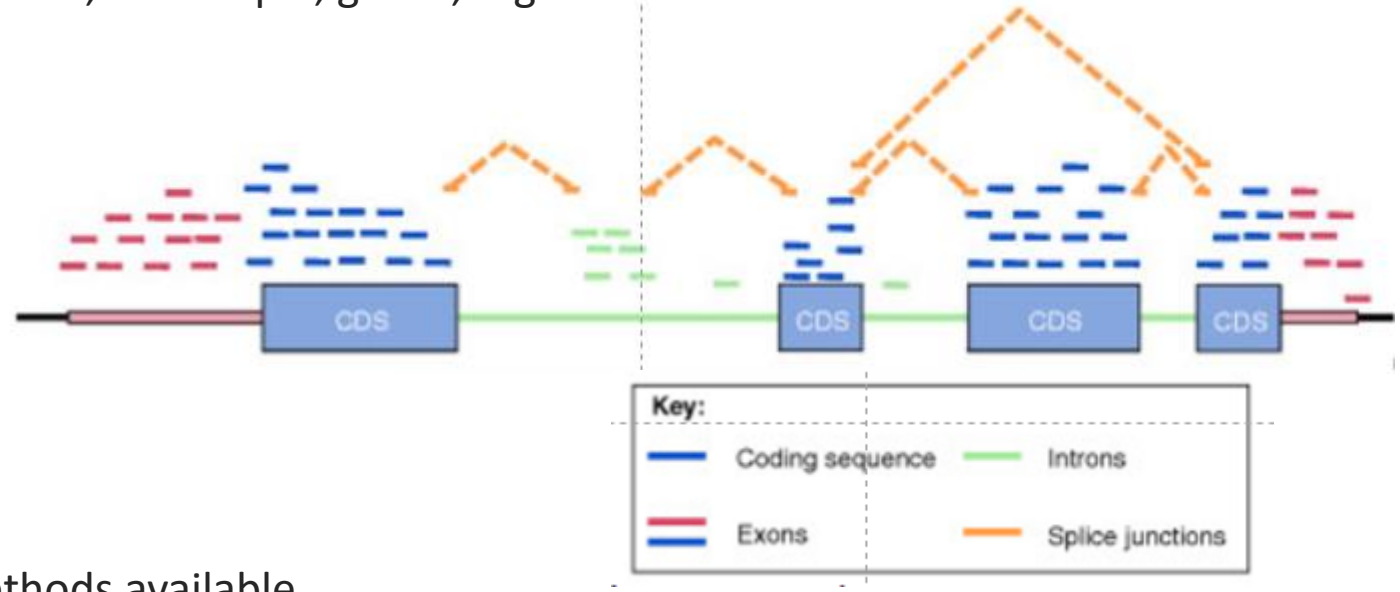



```
@HD     VN:1.3  SO:coordinate
@SQ     SN:22   LN:51304566     AS:NCBI37       M5:a718acaa6135fdca8357d5bfe94211dd     UR
:file:/home/mktrost/seqshop/example/ref22/human.g1k.v37.chr22.fa
@RG     ID:0    SM:HG00551      LB:HG00551      CN:unknown      PL:ILLUMINA
@PG     ID:bwa  PN:bwa  VN:0.7.10-r960-dirty    CL:/home/mktrost/seqshop/gotcloud/bin/bwa
mem -t 1 -M -R @RG\tID:0\tSM:HG00551\tLB:HG00551\tCN:unknown\tPL:ILLUMINA /home/mktrost/se
qshop/example/ref22/human.g1k.v37.chr22.fa /home/mktrost/seqshop/example/fastq/HG00551.SRR
190851.fastq
@RG     ID:1    SM:HG00551      LB:HG00551      CN:unknown      PL:ILLUMINA
SRR190851.48112415      113     22      16918656        3       23M78S  =       31650772 1
4732127 TCCTCGACCTCCCAAAGTCCTGGTATTAAGCGTTAGAGCCACCGCACCCAGCCAGTTATCTCTTTTTTAAAATGTTTATTTA
ATACATTATTTTTTATACT     ###############################################################
############################                    :0A22       NM:i:1  OQ:Z:##############
############################                    ##################################### R
G:Z:1   XS:i:21 XA:Z:22,-38586564,7S21M73S,0;
SRR190851.103013373     121     22      16936847        2       37S18M2D46M     =       16
936847  0       CACAAGTTCAAAGTTCCACAGATCTCAAAGGCAGGTACAAAATCCCACCAGTCTCTGCTAAAGCATAGCAAGAG
TGACCTTTACTCCAGTTCCCAACAAGT      ###############################################################
#####560523/+131335234238-2241+7+-,+/-+89,6      AS:i:46 MD:Z:6G11^TT30T15       NM:i:4  OQ
:Z:#############################################################BC>D?@?>DB?CDEB?CDAE6D@C
C>A8976?=;?A98  RG:Z:1  XS:i:44 XA:Z:22,-36435775,56M1I2M3I39M,12;
```

http://samtools.github.io/hts-specs/SAMv1.pdf

Step 1: Summarization/Quantification of reads: obtaining the counts matrix

- We need to summarize and aggregate reads over some biologically meaningful unit, such as exons, transcripts, genes, regions …



- Many methods available

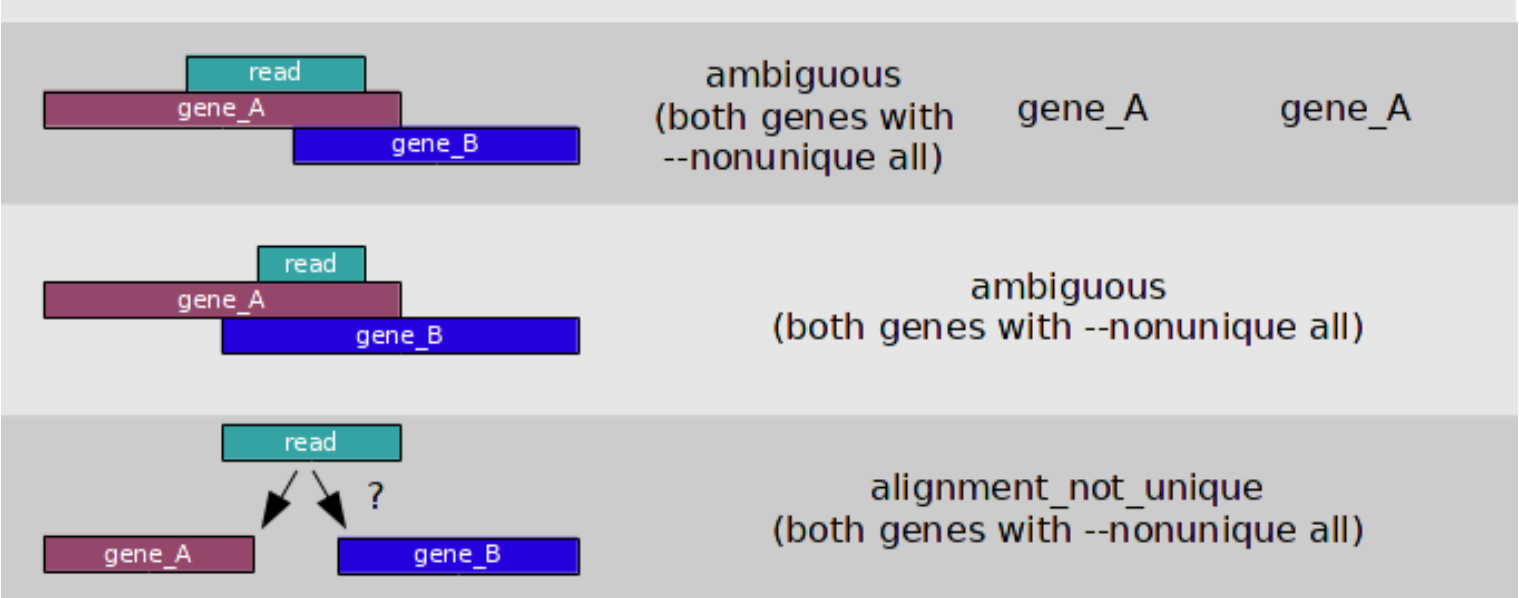
Step 1: Summarization/Quantification of reads: obtaining the counts matrix

- Requires gene annotation specifying the genomic start and end position of each exon of each gene . Usually this is contained in a data frame in *GTF* format for each organism.

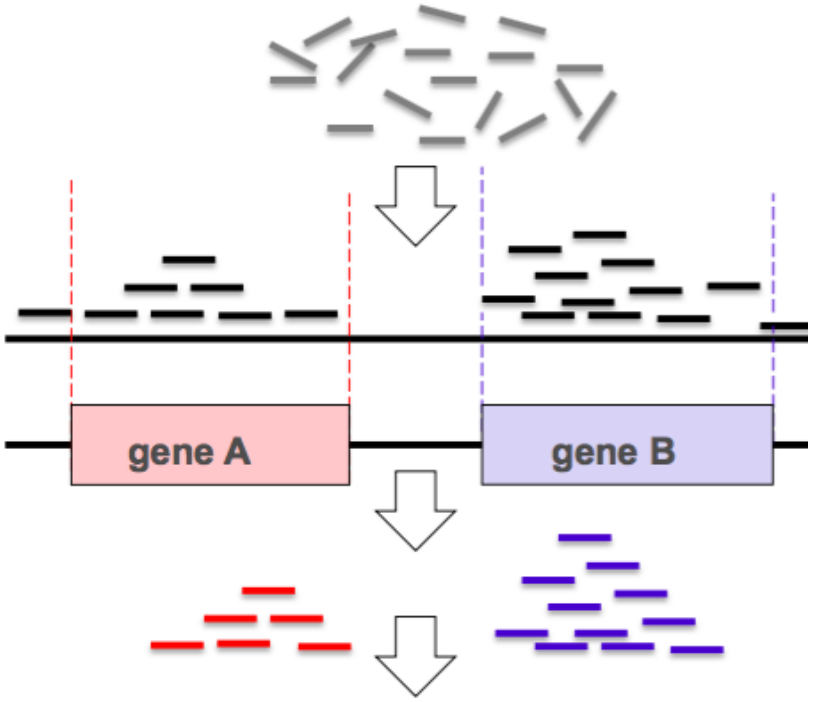
Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr4	protein_coding	CDS	24053	24477	.	+	0	exon_number "1"; gene_id "FBgn0040
chr4	protein_coding	exon	24053	24477	.	+	.	exon_number "1"; gene_id "FBgn0040
chr4	protein_coding	CDS	24979	25153	.	+	1	exon_number "2"; gene_id "FBgn0040
chr4	protein_coding	exon	24979	25153	.	+	.	exon_number "2"; gene_id "FBgn0040
chr4	protein_coding	CDS	25218	25450	.	+	0	exon_number "3"; gene_id "FBgn0040
chr4	protein_coding	exon	25218	25450	.	+	.	exon_number "3"; gene_id "FBgn0040
chr4	protein_coding	CDS	25501	25618	.	+	1	exon_number "4"; gene_id "FBgn0040
chr4	protein_coding	exon	25501	25621	.	+	.	exon_number "4"; gene_id "FBgn0040
chr4	protein_coding	stop_codon	25619	25621	.	+	0	exon_number "4"; gene_id "FBgn0040
chr4	pseudogene	exon	26994	27101	.	-	.	exon_number "7"; gene_id "FBgn0052
chr4	pseudogene	exon	27167	27349	.	-	.	exon_number "6"; gene_id "FBgn0052
chr4	pseudogene	exon	28371	28609	.	-	.	exon_number "5"; gene_id "FBgn0052

Step 1: Summarization/Quantification of reads: obtaining the counts matrix

- Reads that map to exons of genes are added together to obtain the count for each gene
- There may be some ambiguities



Step 1: Summarization/Quantification of reads: obtaining the counts matrix



	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3
Gene A	6	5	7	17	10	11
Gene B	11	11	10	3	4	2
Gene C	200	150	355	50	1	3
Gene D	0	1	0	2	0	1

- The summarized RNA-seq data is widely known as a ***count matrix***

Step 1: Summarization/Quantification of reads: obtaining the counts matrix

Hands on!

... follow **sections 3-4 of the Rmd file**

Steps covered in this tutorial:

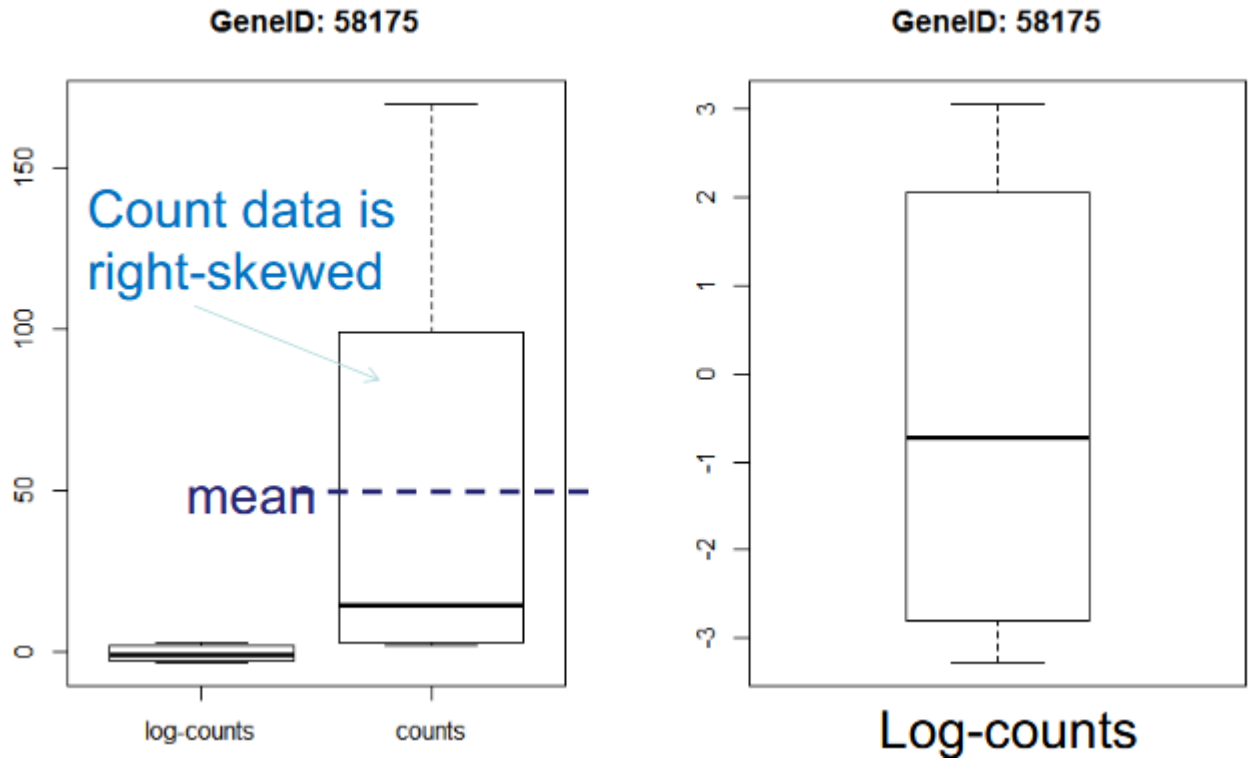
1. Summarization/Quantification of aligned reads: obtaining the counts matrix
2. **Data pre-processing and exploratory analysis**
3. Differential gene expression analysis
4. Visually explore the results

Step 2: Data pre-processing and exploratory analysis

- How my data looks like?
- Is it of enough quality for analysis?
- Are there some outlier samples that should be removed?
- Are samples grouped according to the experimental conditions?
- What are the main sources of variability in the data?

Step 2: Data pre-processing and exploratory analysis

- RNA-seq data is
  - discrete
  - has non-constant mean-variance trend



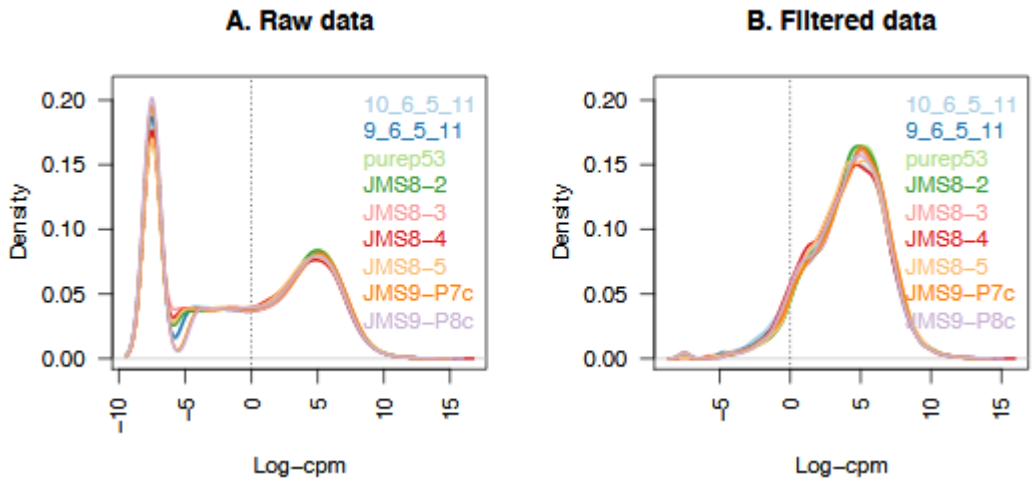
Non-normal distribution of count data

Step 2: Data pre-processing and exploratory analysis

Filtering

- It is recommended to filter for lowly expressed genes before differential expression testing.

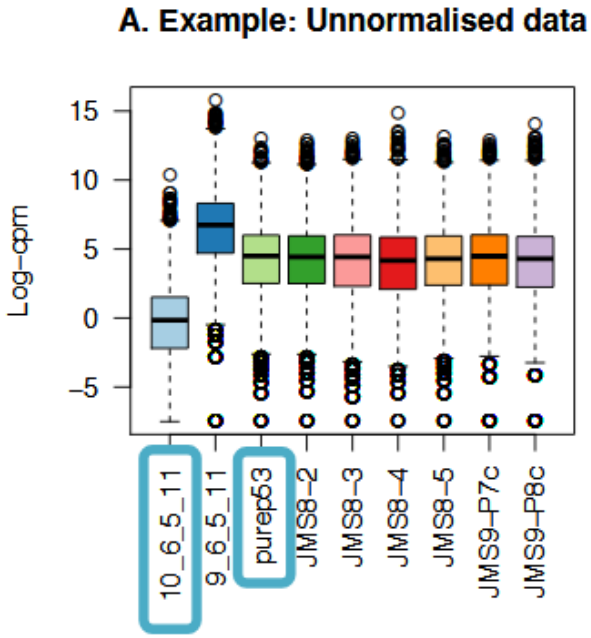
  - provide little evidence for differential expression and they interfere with some of the statistical approximations used
  - Add to the multiple testing burden when estimating FDR, reducing power to detect differential expressed genes.



Step 2: Data pre-processing and exploratory analysis

Scaling and normalization

- The counts of mapped reads for each gene is proportional to the expression of RNA ("interesting") in addition to many other factors ("uninteresting").
- Normalization is the process of scaling raw count values to account for the "uninteresting" factors.



If we ran a DE analysis on Sample 1 and Sample 3, almost all genes will be down-regulated in Sample 1!!

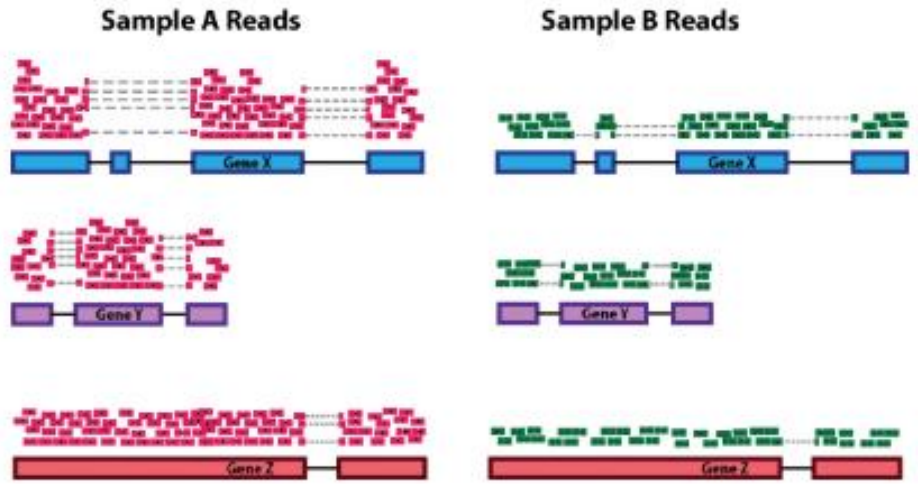
https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Step 2: Data pre-processing and exploratory analysis

Scaling and normalization

Main factors often considered during normalization are:

1. **Sequencing depth:** total number or reads mapped to the genome



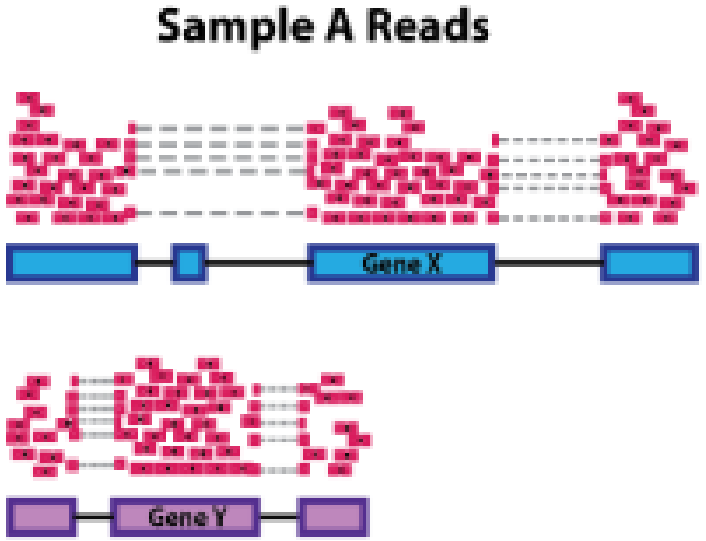
https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Step 2: Data pre-processing and exploratory analysis

Scaling and normalization

Main factors often considered during normalization are:

2. **Gene length:** Accounting for gene length is necessary for comparing expression between different genes within the same sample.



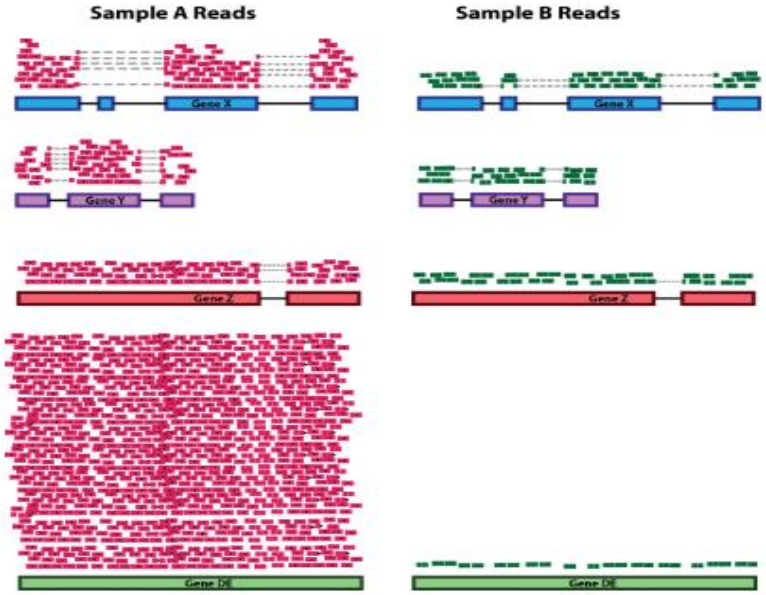
https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Step 2: Data pre-processing and exploratory analysis

Scaling and normalization

Main factors often considered during normalization are:

3. **RNA composition**: A few highly differentially expressed genes between samples, differences in the number of genes expressed between samples, or presence of contamination can skew some types of normalization methods.



https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Step 2: Data pre-processing and exploratory analysis

Scaling and normalization

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Step 2: Data pre-processing and exploratory analysis

Exploratory analysis

Unsupervised-separation methods based on data, without prior knowledge of experimental design, can be used to get an overview of the data

- Do samples separate by experimental groups?
- Where the greatest sources of variation in the data come from?
- Are there any outliers?

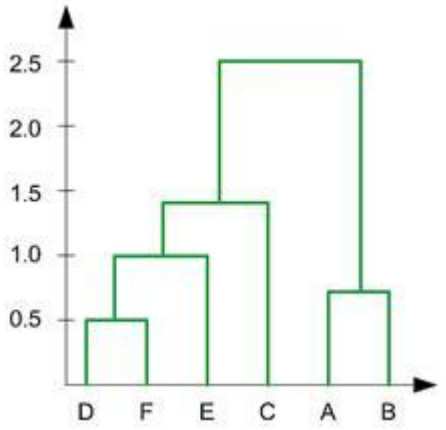
Step 2: Data pre-processing and exploratory analysis

Exploratory analysis

Hierarchical clustering

- Hierarchical clustering is typically based on pairwise comparisons of individual samples, which are grouped into “neighborhoods” of similar samples. The basis of hierarchical clustering is therefore a matrix of similarity metrics.

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

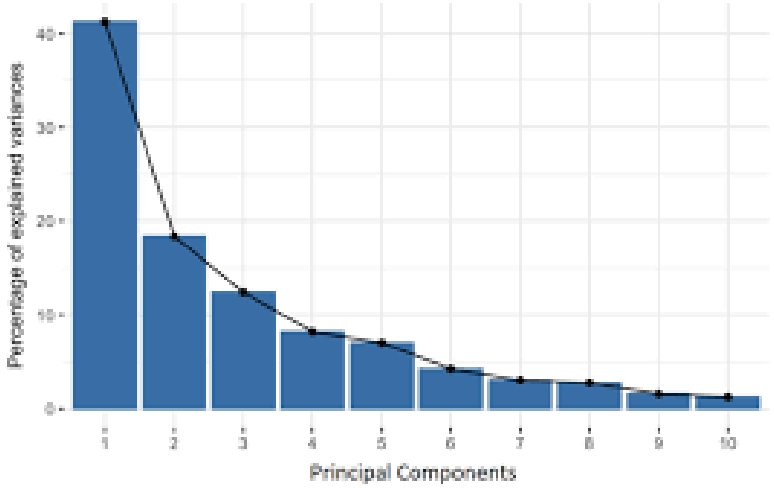


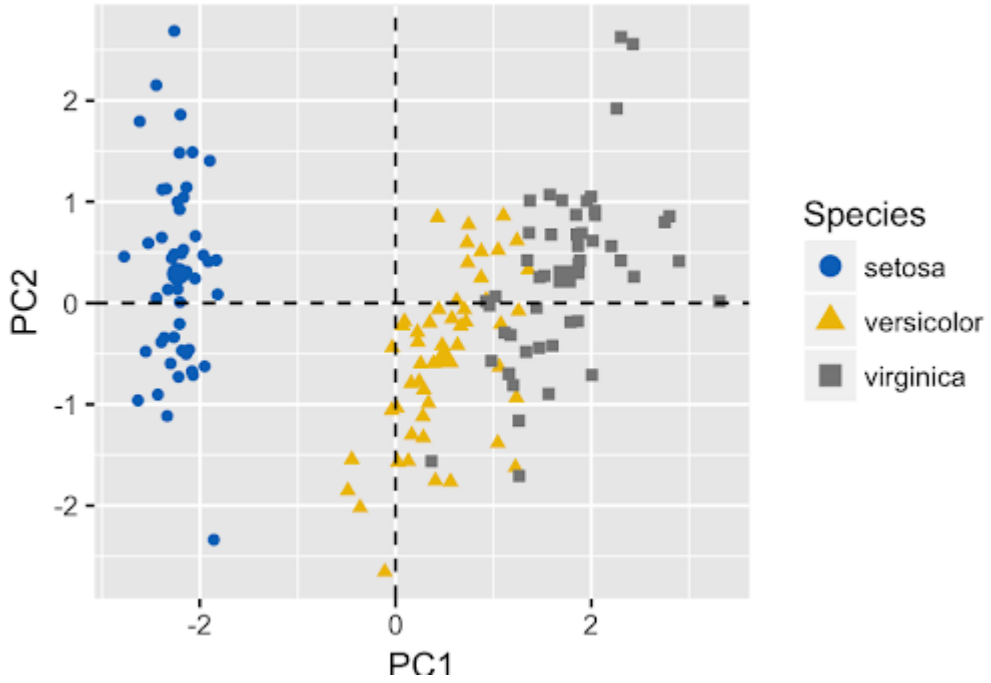
Step 2: Data pre-processing and exploratory analysis

Exploratory analysis

Principal Components Analysis (PCA)

- A dimensionality reduction approach that aims to find groups of features (e.g., genes) that have something in common (e.g., certain patterns of expression across different samples)
- Few dimensions (components) can be used to represent the information from thousands of mRNAs.



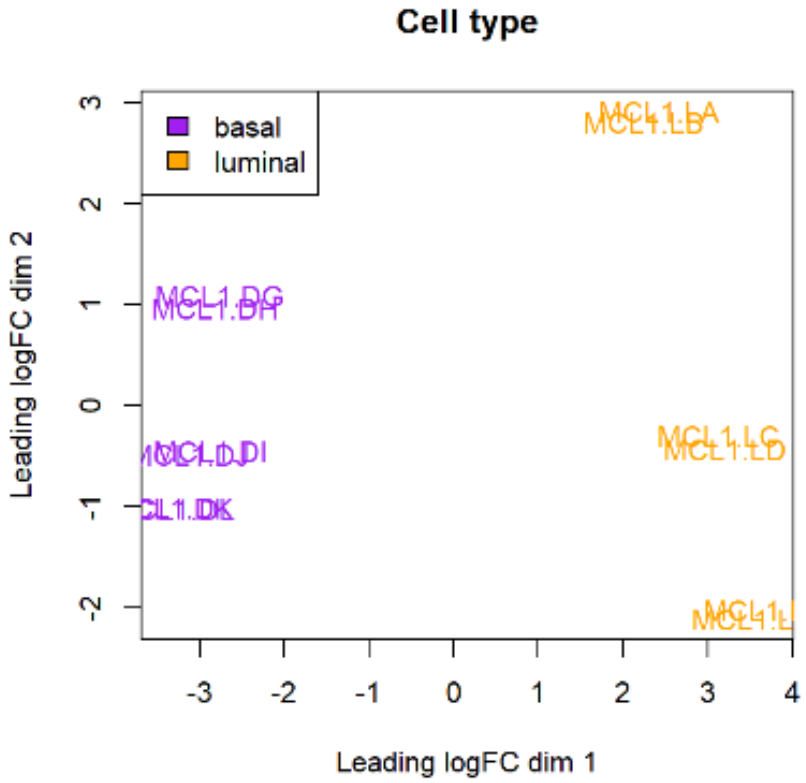


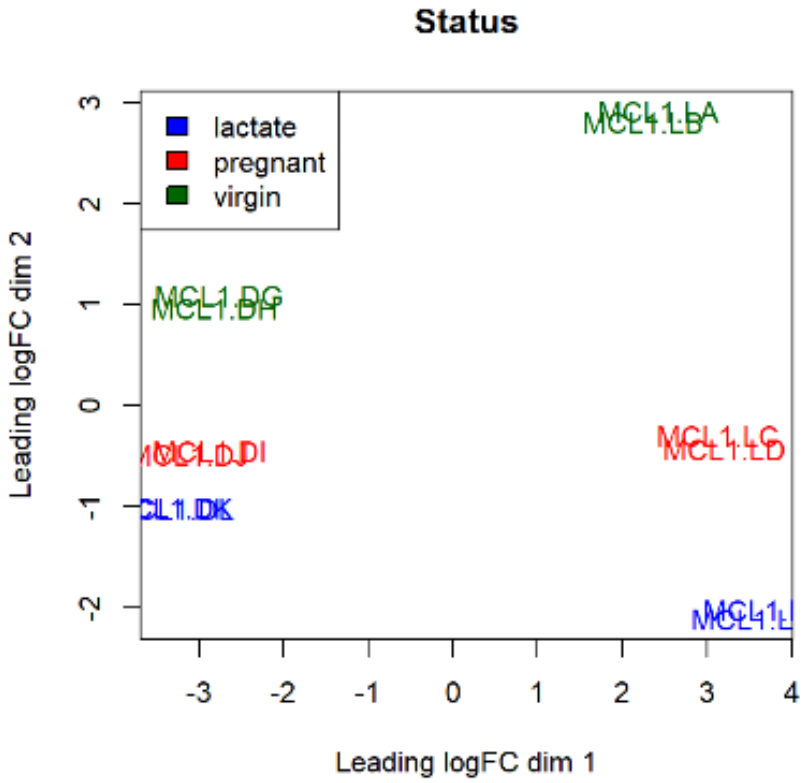
Step 2: Data pre-processing and exploratory analysis

Hands on!

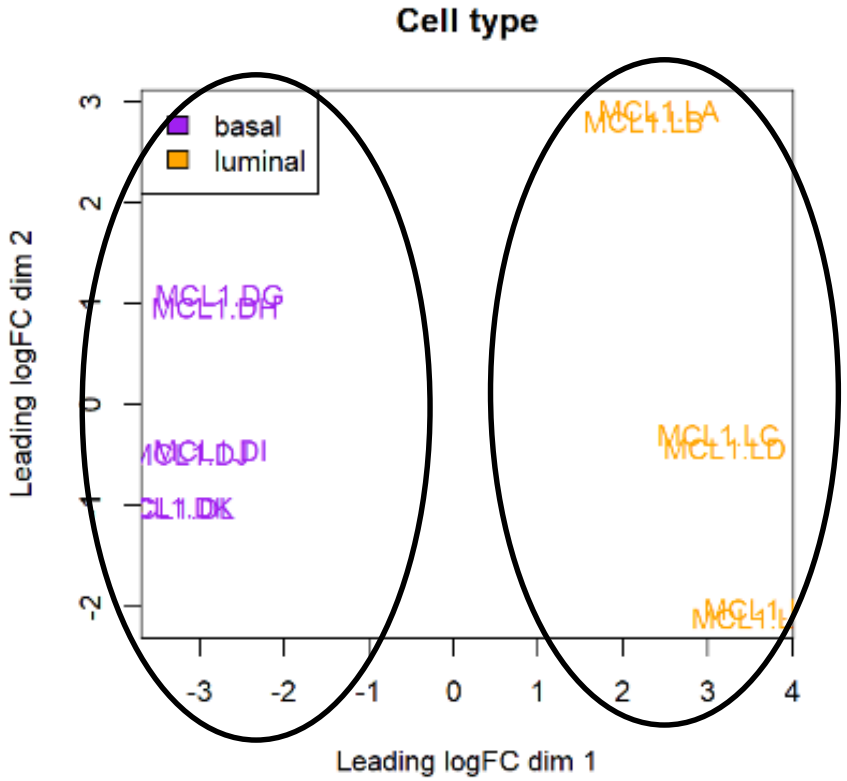
... follow **section 5 of the Rmd file**

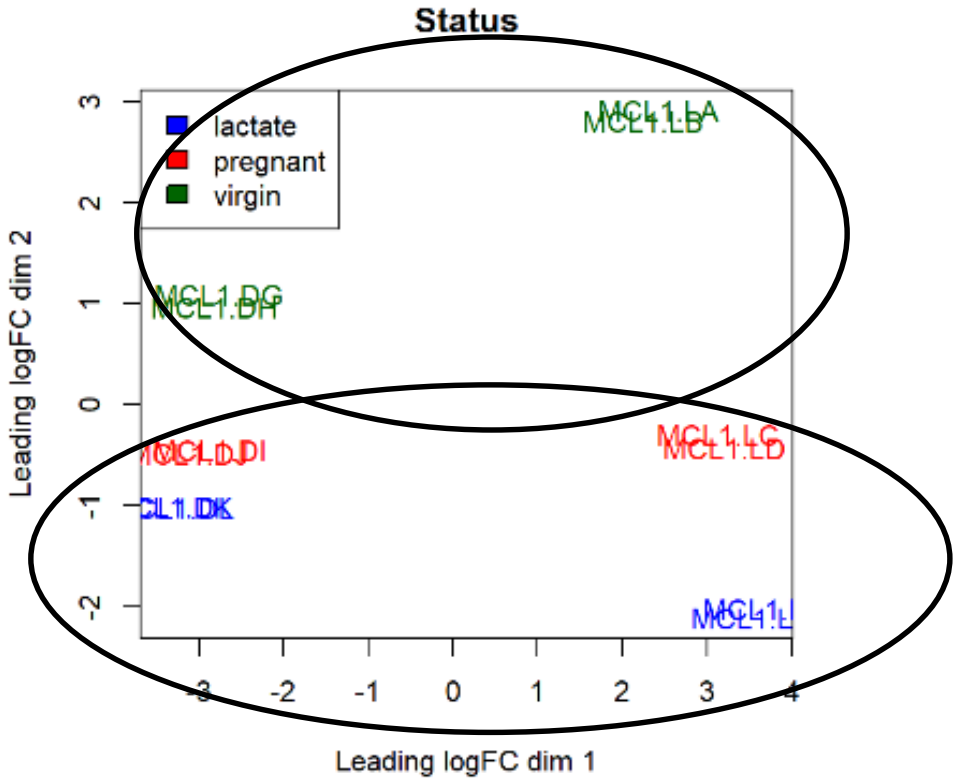
Step 2: Data pre-processing and exploratory analysis





Step 2: Data pre-processing and exploratory analysis

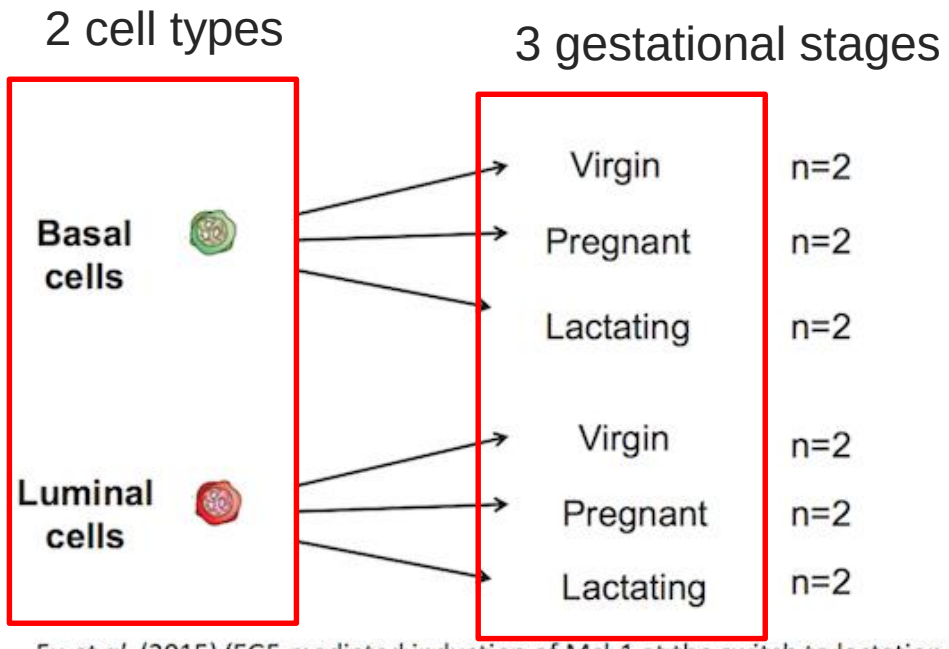




Steps covered in this tutorial:

1. Summarization/Quantification of aligned reads: obtaining the counts matrix
2. Data pre-processing and exploratory analysis
3. **Differential gene expression analysis**
4. Visually explore the results

Step 3: Differential expression analysis



Fu *et al.* (2015) 'EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival' Nat Cell Biol

6 Groups:

Basal.Virgin
Basal.Pregnant
Basal.Lactating

Luminal.Virgin
Luminal.Pregnant
Luminal.Lactating

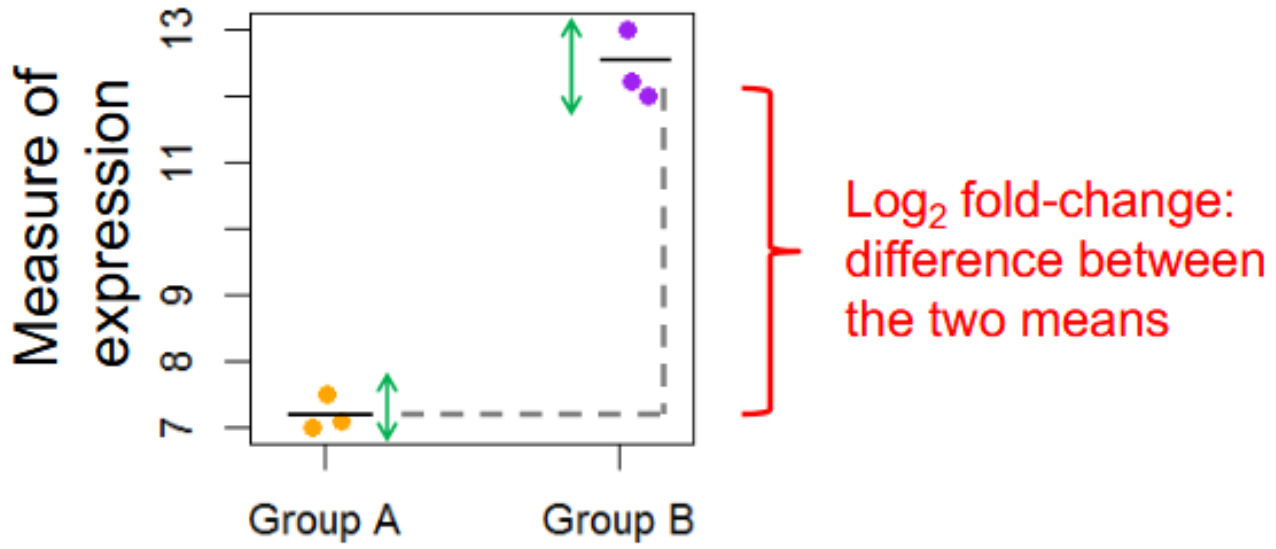
We will analyze the following comparisons:

- Basal.Lactating vs. Basal.Pregnant
- Luminal.Lactating vs. Luminal.Pregnant

Step 3: Differential expression analysis

What do we need to perform a statistical test?

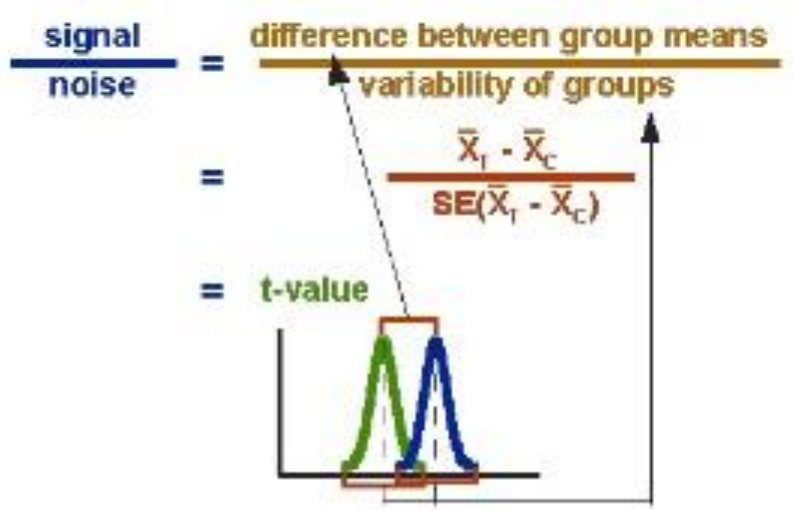
- Measure of average expression
- Measure of variability



•Finally we will be assigning a p-value for each test/gene.

Step 3: Differential expression analysis

Couldn't we just use a Student's t test for each gene?



http://www.socialresearchmethods.net/kb/stat_t.php

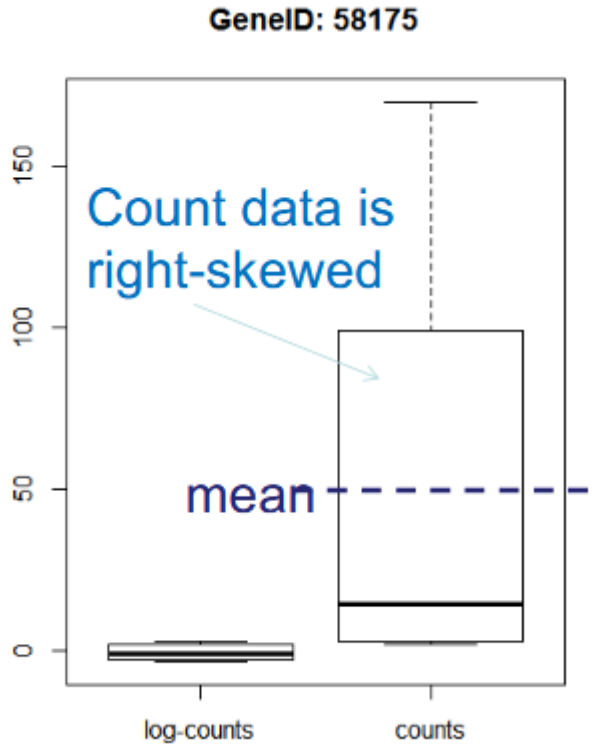
Problems with this approach:

- May have **few replicates**

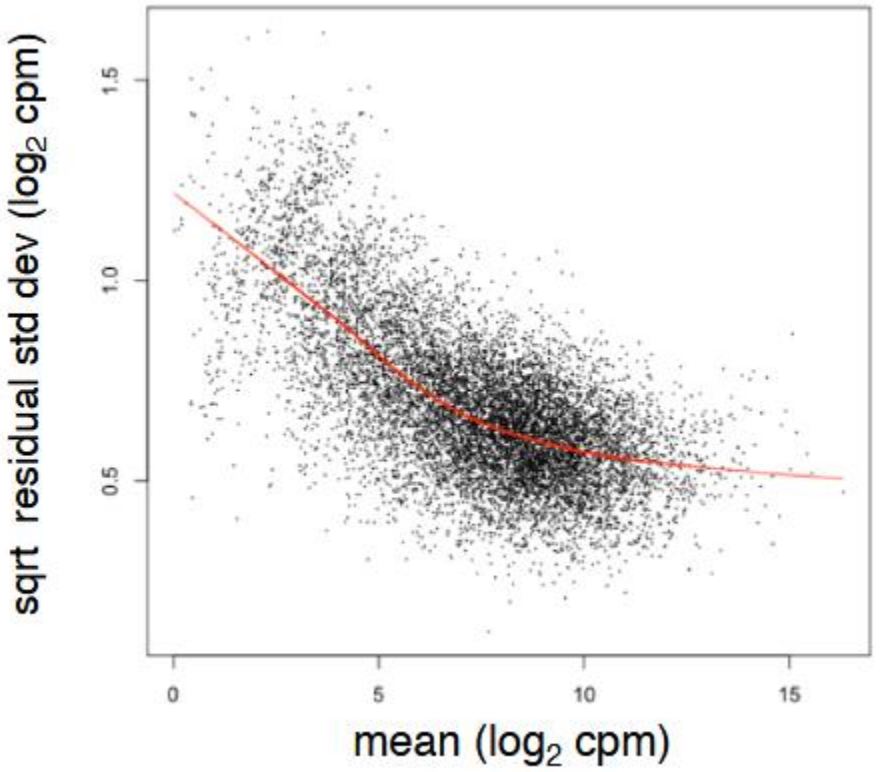
- Distribution is **not normal**

- **Multiple testing** issues

- RNA-seq data is
  - discrete
  - has non-constant mean-variance trend



Mean-variance relationship



Non-normal distribution of count data

Step 3: Differential expression analysis

- Different software use different approaches to deal with the “t test issues”

Distributional issue: Solved by variance stabilizing transform in limma – voom() function

edgeR and DESeq model the count data using a *negative binomial distribution* and use their own modified statistical tests based on that.

Multiple testing issue: All of these packages report false discovery rate (corrected p values). For SAMseq based on resampling, for others usually Benjamini-Hochberg corrected p values.

Variance estimation issue: edgeR, DESeq2 and limma (in slightly different ways) “borrow” information across genes to get a better variance estimate. One says that the estimates “shrink” from gene-specific estimates towards a common mean value.

Step 3: Differential expression analysis

Feature	DESeq2	edgeR	limmaVoom	Cuffdiff
Seq. depth normalization	Sample-wise size factor	Gene-wise trimmed median of means (TMM)	Gene-wise trimmed median of means (TMM)	FPKM-like or DESeq-like
Dispersion estimate	Cox-Reid approximate conditional inference with focus on maximum *individual* dispersion estimate	Cox-Reid approximate conditional inference moderated towards the *mean*	squeezes gene-wise residual variances towards the global variance	
Assumed distribution	Neg. binomial	Neg. binomial	*log*-normal	Neg. binomial
Test for DE	Wald test (2 factors); LRT for multiple factors	exact test for 2 factors; LRT for multiple factors	*t*-test	*t*-test
False positives	Low	Low	Low	High
Detection of differential isoforms	No	No	No	Yes
Support for multi-factored experiments	Yes	Yes	Yes	No
Runtime (3-5 replicates)	Seconds to minutes	Seconds to minutes	Seconds to minutes	Hours

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

Step 3: Differential expression analysis

Feature	DESeq2	edgeR	limmaVoom	Cuffdiff
Seq. depth normalization	Sample-wise size factor	Gene-wise trimmed median of means (TMM)	Gene-wise trimmed median of means (TMM)	FPKM-like or DESeq-like
Dispersion estimate	Cox-Reid approximate conditional inference with focus on maximum *individual* dispersion estimate	Cox-Reid approximate conditional inference moderated towards the *mean*	squeezes gene-wise residual variances towards the global variance	
Assumed distribution	Neg. binomial	Neg. binomial	*log*-normal	Neg. binomial
Test for DE	Wald test (2 factors); LRT for multiple factors	exact test for 2 factors; LRT for multiple factors	*t*-test	*t*-test
False positives	Low	Low	Low	High
Detection of differential isoforms	No	No	No	Yes
Support for multi-factored experiments	Yes	Yes	Yes	No
Runtime (3-5 replicates)	Seconds to minutes	Seconds to minutes	Seconds to minutes	Hours

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

Step 3: Differential expression analysis

Hands on!

... follow **sections 6.1-6.4 of the Rmd file**

Steps covered in this tutorial:

1. Summarization/Quantification of aligned reads: obtaining the counts matrix
2. Data pre-processing and exploratory analysis
3. Differential gene expression analysis
4. **Visually explore the results**

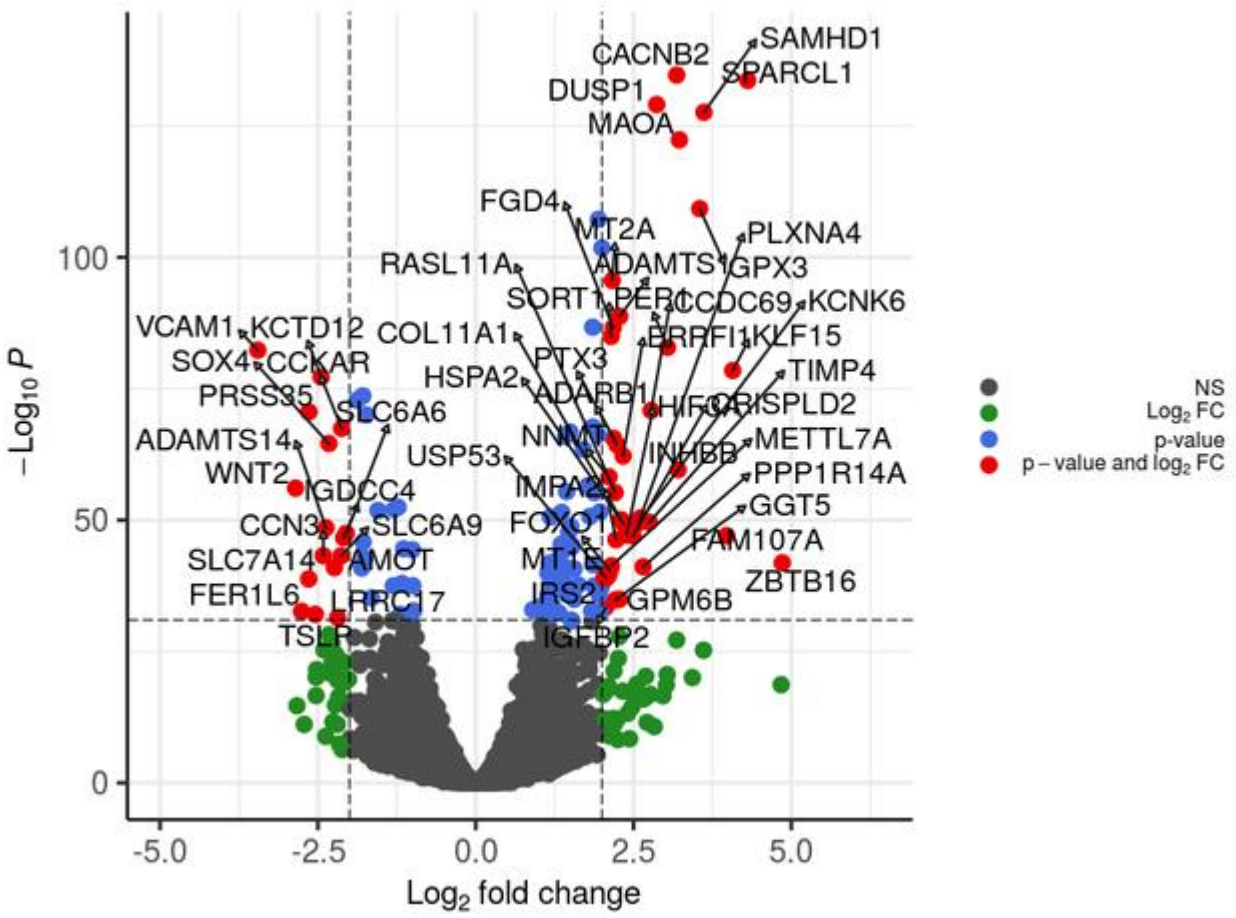
Step 4: Visual exploration of results

- For each comparison, the output of a differential expression analysis gives a **top table**
  - The logFC column gives the log2-fold change in expression of a gene between the two groups tested
  - The AveExpr column gives the average log2-expression level for that gene across all the samples
  - Column t is the moderated t-statistic.
  - Column P.Value is the associated p-value
  - adj.P.Value is the p-value adjusted for multiple testing. The most popular form of adjustment is “BH” which is Benjamini and Hochberg’s method to control the false discovery rate.
  - The B-statistic is the log-odds that the gene is differentially expressed

ENTREZID	logFC	AveExpr	t	P.Value	adj.P.Val	B
21953	-5.82324479390654	0.302071586676651	-11.4850006047225	3.51591878726469e-06	0.0254971055316036	3.85618421189551
67111	-2.53794287419365	3.29074574604293	-11.243908562307	4.11691826569324e-06	0.0254971055316036	4.554304142791
72515	1.93521071350637	6.45103453810206	10.6629295310258	6.09925372132422e-06	0.0254971055316036	4.56862100492472
232016	-2.59587128401098	5.00471484116641	-10.3045614809449	7.84745919853886e-06	0.0254971055316036	4.32545272264372
329739	-1.5207496111552	4.18813046970528	-10.2660361673776	8.06666208921906e-06	0.0254971055316036	4.29825625640396
211064	1.44554832890089	3.94643017156436	9.71625024434493	1.20740489824308e-05	0.0266846381792226	3.76475152189045
20319	3.38255617819199	2.77994069969196	9.41906713347006	1.51402912307052e-05	0.0266846381792226	3.09620289541151
67619	1.25007158752396	4.93701234604145	9.40799351731106	1.52702975887503e-05	0.0266846381792226	3.6853371946635
23790	-1.17888226679082	4.58546158856824	-9.32785193444217	1.62492332371926e-05	0.0266846381792226	3.65615865031246
16835	-1.23984127902741	7.48442251346755	-9.12538398952684	1.90501153595595e-05	0.0266846381792226	3.49313064462358
69237	1.03836392641394	6.06105681180188	9.10016923069938	1.9435221657598e-05	0.0266846381792226	3.4834523886179
20482	1.88759238559317	8.49892507013266	8.98770966816343	2.12620612422923e-05	0.0266846381792226	3.3918497852994
229595	-1.96924610875708	4.21407207130784	-8.92198428561004	2.24180823381668e-05	0.0266846381792226	3.33100037257509
14287	1.42446736142347	5.14381951119619	8.80850266980197	2.45830187089409e-05	0.0266846381792226	3.24487070648194
211577	-5.14520172661396	-0.936833491489371	-8.65346712795871	2.81368540250546e-05	0.0266846381792226	2.34229652309425
235542	-1.10867905046616	6.5010594105959	-8.6240600533168	2.86177920994012e-05	0.0266846381792226	3.10838403175677

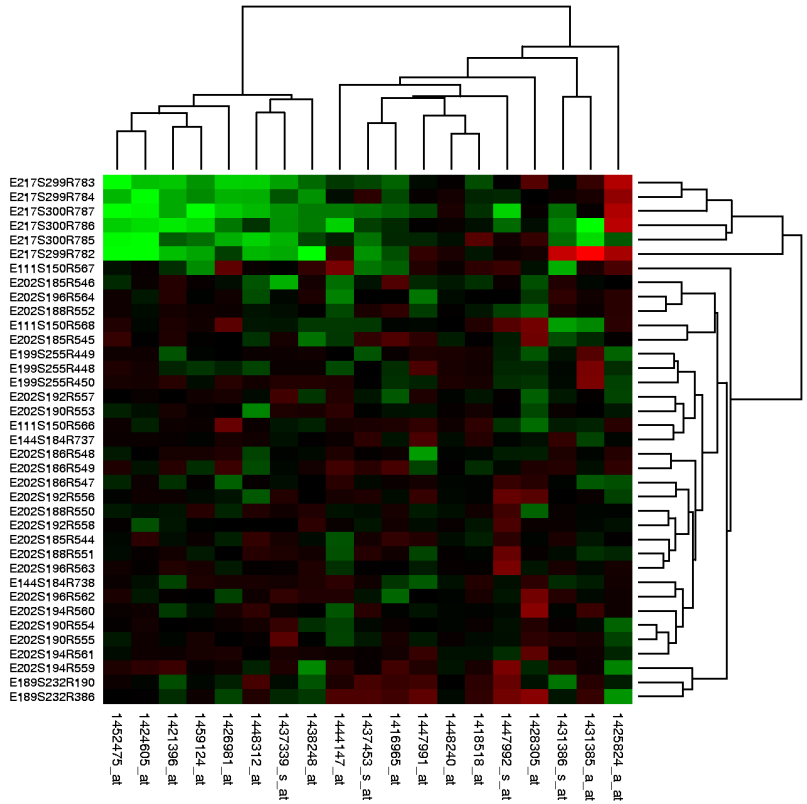
Step 4: Visual exploration of results

- A common visualisation is the **Volcano Plot** which display a measure of significance on the y-axis and fold-change on the x-axis



Step 4: Visual exploration of results

- Heatmap is a very useful tool for quick representation of quantitative differences in expression levels of biological data.
    - Each gene is represented as a row and is color-coded to represent the intensity of its variation (either positive or negative) relative to a reference value.
    - Biological samples are represented as columns in the grid.
- Heatmap representations are also combined with clustering methods to group genes and/or samples based on their expression patterns.



Step 4: Visual exploration of results

Hands on!

... follow **section 6.5 of the Rmd file**

Next step…

- We have performed analysis on gene level
- It would be interesting to look for how these genes act together

Analysis of biological significance