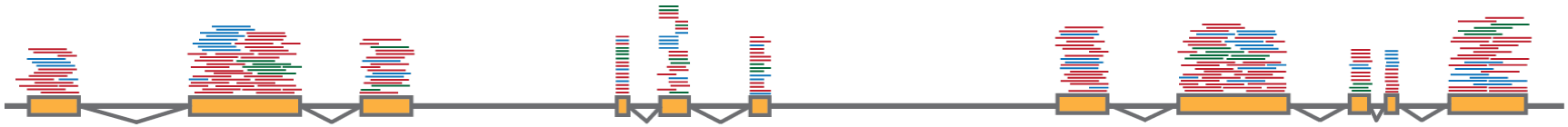


Session IV. RNAseq

Armando Reyes-Palomares

Department Biochemistry and Molecular Biology

Complutense University of Madrid

armandorp@ucm.es

Practical Sessions

1. Introduction to R
2. Bioconductor: BiomaRt and GenomicRanges
3. Whole-Genome Analysis – SARS-CoV2
4. Transcriptome analysis (e.g. RNAseq)

armandorp@ucm.es

Table of contents

1. Survey (educational innovation project).
2. Nextclade
3. Gene expression resources
4. Align and quantification of raw data from RNA-seq assays
5. Differential Expression Analysis using DESeq2

armandorp@ucm.es

Parte 1ª. Fase PREVIA a las sesiones Análisis Genómico (GAG)

Responde considerando tus conocimientos y estatus previo a estas sesiones de Análisis Genómico en GAG en el Máster BBMBiomed. Encuesta anónima relacionada con diversos métodos asociados al proyecto de innovación docente relacionado con el área de bioinformática de procesado y el análisis de datos en biología molecular y bioquímica.

armreyes@ucm.es (not shared) Switch account

¿Tenías conocimientos en algún lenguaje de programación el curso académico anterior?

- ○ Si
- ○ No

Nextstrain

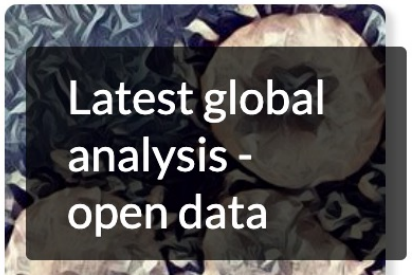
https://clades.nextstrain.org

Nextstrain

Real-time tracking of pathogen evolution

SARS-CoV-2 (COVID-19)

We are incorporating SARS-CoV-2 genomes as soon as they are shared and providing analyses and situation reports. In addition we have developed a number of resources and tools, and are facilitating independent groups to run their own analyses.



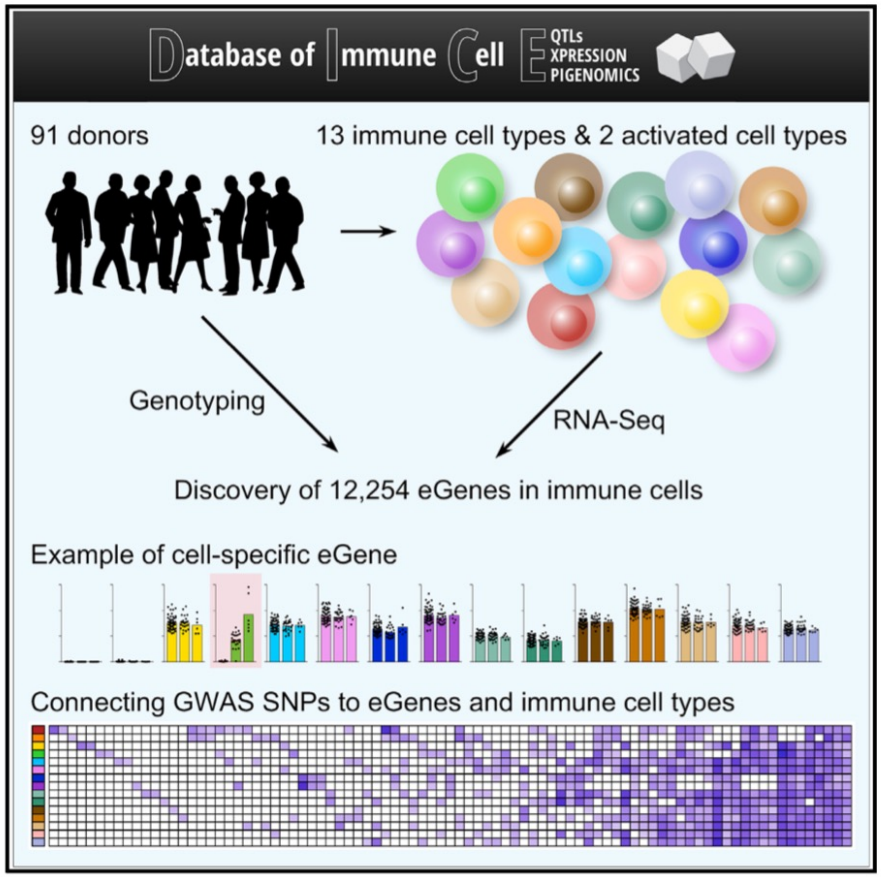
armandorp@ucm.es

Resource

Cell

Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression

Graphical Abstract



Authors

Benjamin J. Schmiedel, Divya Singh, Ariel Madrigal, ..., Mitchell Kronenberg, Bjoern Peters, Pandurangan Vijayanand

Correspondence

vijay@lji.org

In Brief

Surveying gene expression and SNP genotypes across immune cell types from healthy humans reveals cis-eQTLs affecting over half of all expressed genes and demonstrates that variant effects often manifest in cell types other than those with highest gene expression.

GEO Datasets (NCBI)

GEO DataSets | GEO Profiles

This database stores original submitter-supplied study descriptions, as well as curated gene expression DataSets. DataSets form the basis of GEO's advanced data display and analysis tools, including gene expression profile charts and clusters.

Search Examples:

Search by...	Search text
Free text	**smoking cancer**
Keywords and species	**(smok* OR diet) AND (mammals[organism] NOT human[organism])**
Studies in the **NIH Roadmap Epigenomics project**	**"roadmap epigenomics"[Project]**
Study type	**"expression profiling by high throughput sequencing"[DataSet Type]**
Studies with between 100 and 500 samples	**100:500[Number of Samples]**
Studies with CEL files	**"cel"[Supplementary Files]**
DataSets that have 'age' as an experimental variable	**"age"[Subset Variable Type]**
Author	**smith a[Author]**
Published between January and June 2007	**2007/01:2007/06[Publication Date]**
Platform accession	**GPL570**
Studies with PubMed identifiers	**"gds pubmed"[Filter]**

armandorp@ucm.es

ArrayExpress (EMBL-EBI)

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI Hinxton

ArrayExpress

Search

Examples: E-MEXP-31, cancer, p53, Geuvadis | advanced search

Home | Browse | Submit | Help | About ArrayExpress | Contact Us | Login

ArrayExpress – functional genomics data

ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.

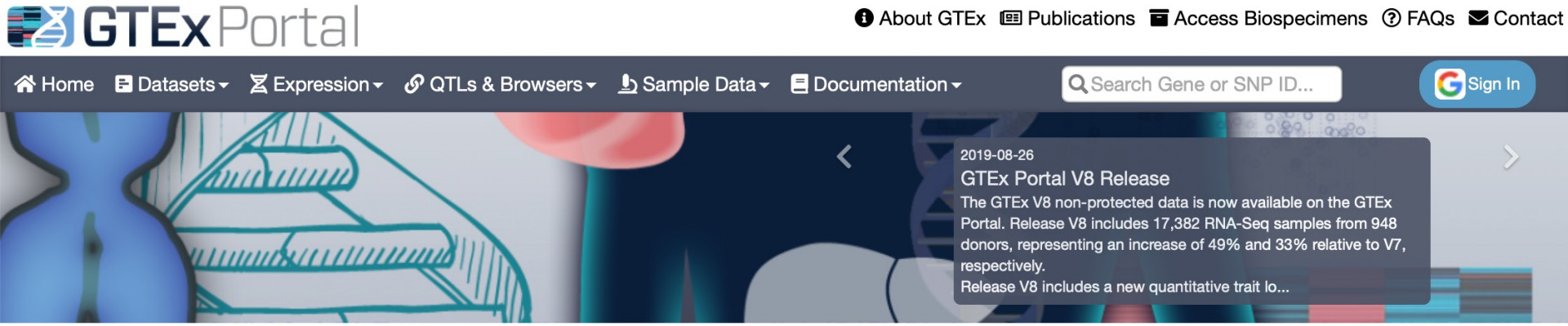
Browse ArrayExpress

Data Content

Updated today at 02:00

- 74979 experiments
- 2569020 assays
- 61.29 TB of archived data

GTEx portal (gtexportal.org)



Resource Overview

Current Release (V8)

Tissue & Sample Statistics
Tissue Sampling Info (Anatomogram)
Access & Download Data
Release History
How to cite GTEx?

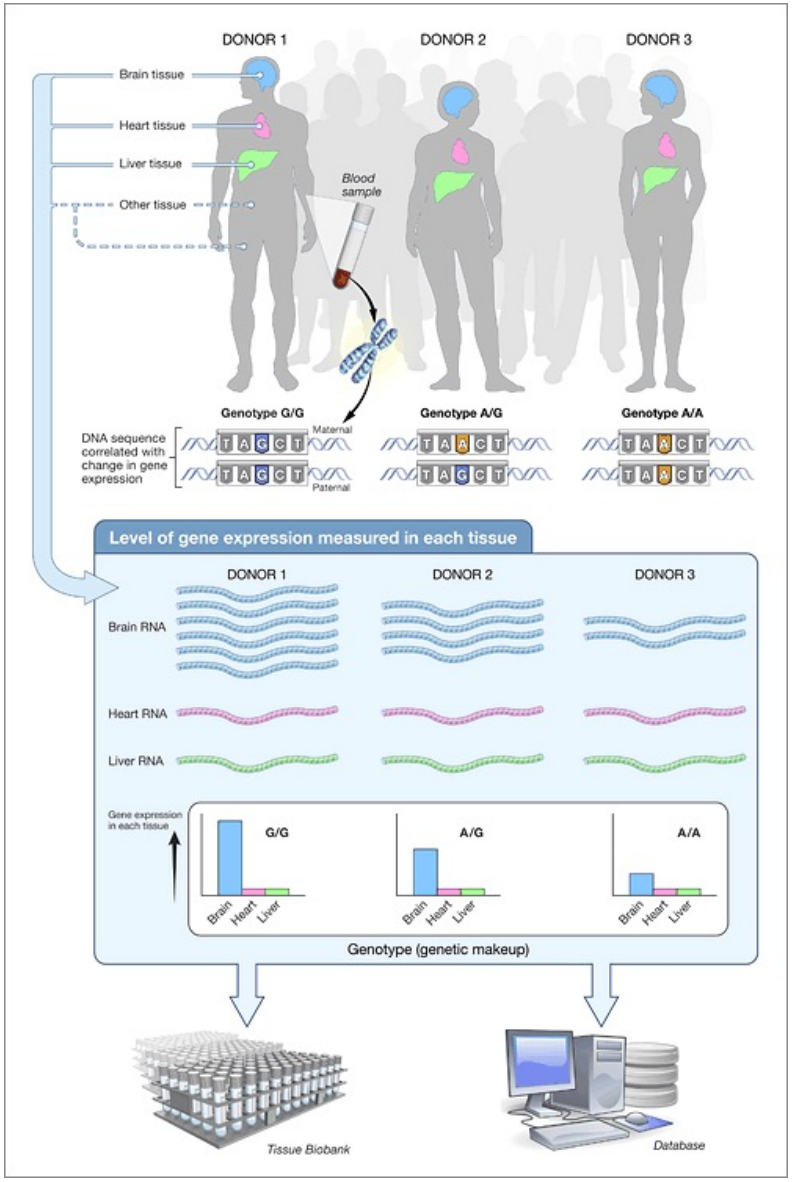
The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

Explore GTEx

Browse	By gene ID	Browse and search all data by gene
	By variant or rs ID	Browse and search all data by variant
	By Tissue	Browse and search all data by tissue
	Histology Image Viewer	Browse and search GTEx histology images
Expression	Multi-Gene Query	Browse and search expression by gene and tissue
	Top 50 Expressed Genes	Visualize the top 50 expressed genes in each tissue
	Transcript Browser	Visualize transcript expression and isoform structures

UNIVERSIDAD COMPLUTENSE MADRID

armandorp@ucm.es



armandorp@ucm.es

Expression Atlas (EMBL-EBI)

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI

Expression Atlas

Gene expression across species and biological conditions

Query single cell expression

To Single Cell Expression Atlas

Home | Browse experiments | Download | Release notes | FAQ | Help | Licence | Also in this section

Search across **65 species**, **4,169 studies**, **139,128 assays**

Ensembl 99, Ensembl Genomes 46, WormBase ParaSite 14,

Animals | Plants | Fungi

Homo sapiens

1518 experiments

Baseline: 79

Differential: 1439

Mus musculus

1185 experiments

Baseline: 49

Differential: 1136

Rattus norvegicus

152 experiments

Baseline: 3

Differential: 149

Drosophila melanogaster

142 experiments

Baseline: 4

Differential: 138

Gallus gallus

39 experiments

Baseline: 4

Differential: 35

Caenorhabditis elegans

30 experiments

Baseline: 1

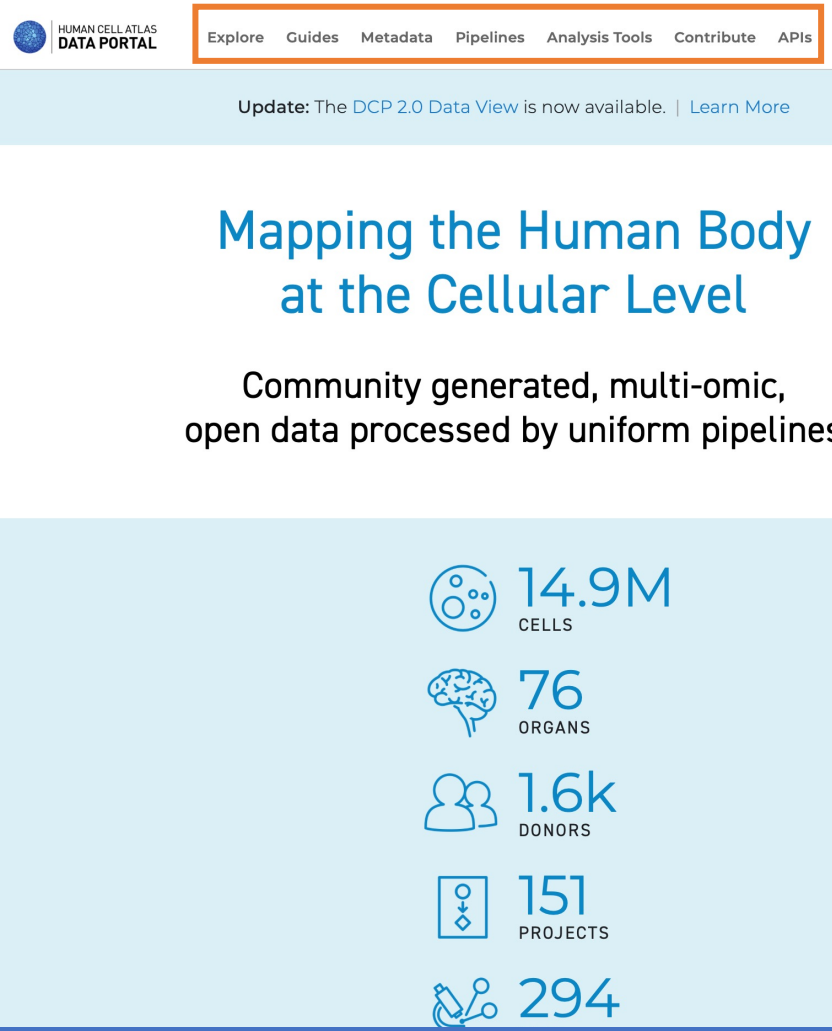
Differential: 29

UNIVERSIDAD COMPLUTENSE MADRID

armandorp@ucm.es

The Human Cell Atlas

www.humancellatlas.org / data.humancellatlas.c



About HCA:

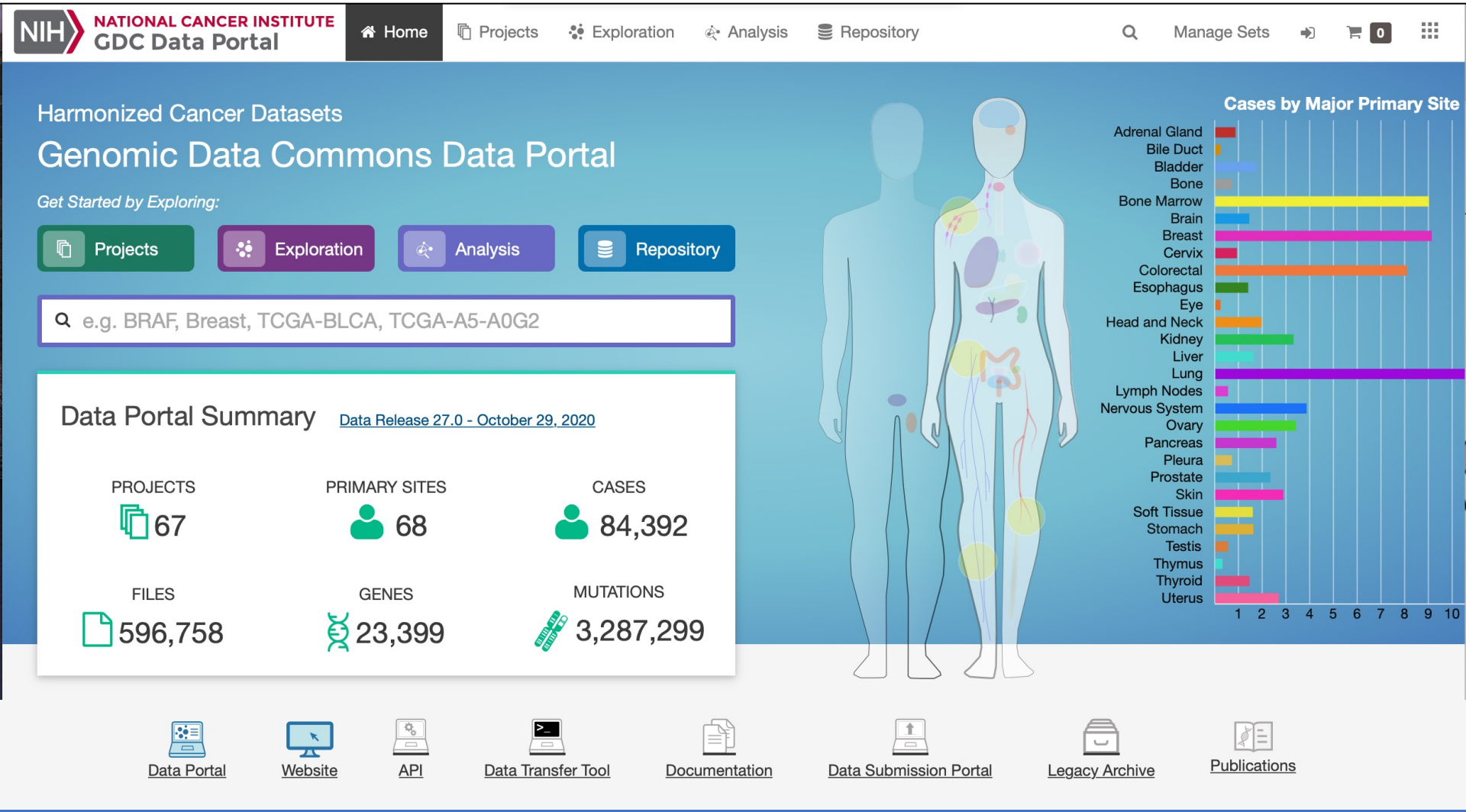
- From 2016, > 1,000 Institutions, 75 countries
- Catalog all cell types and sub-types
- Map cell types to their location within tissues
- Distinguish cell states
- Study transitions, such as activation or differentiation.

UNIVERSIDAD COMPLUTENSE MADRID

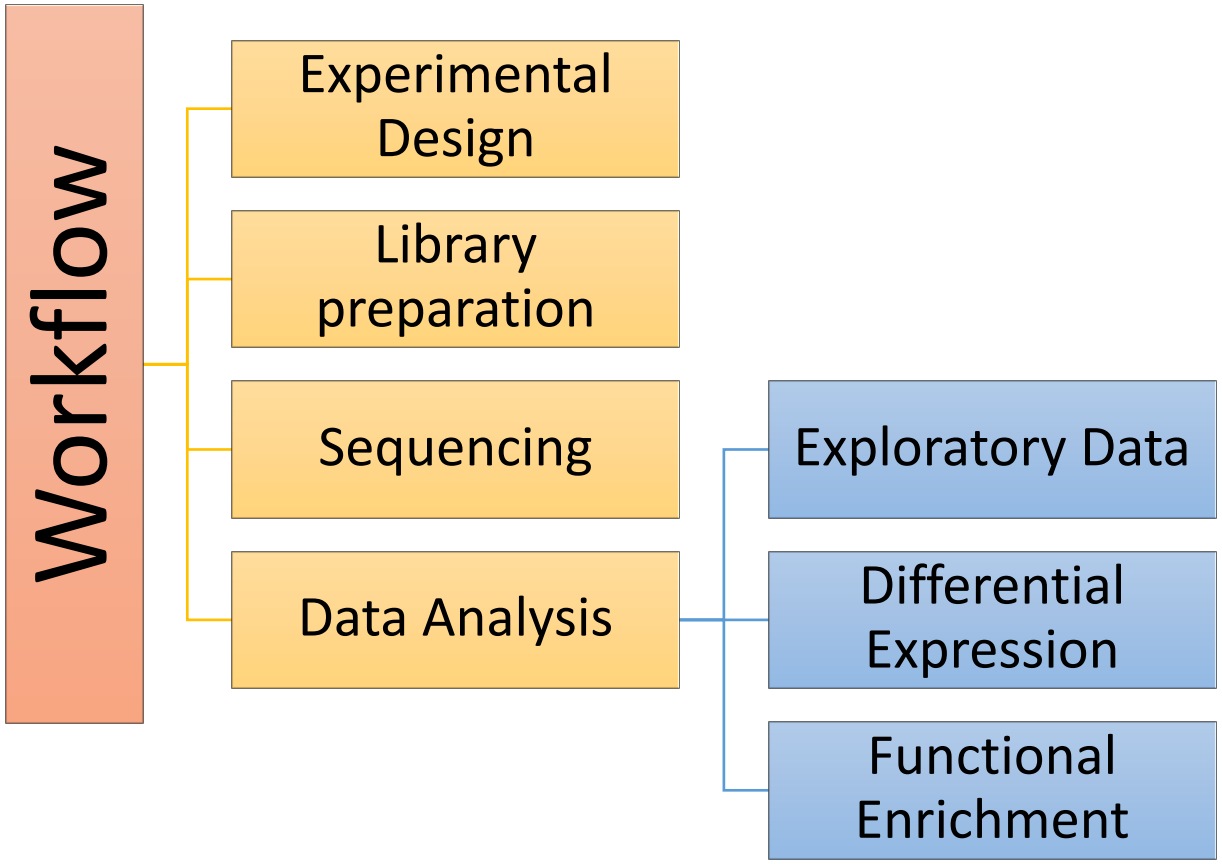
armandorp@ucm.es

The Cancer Genome Atlas

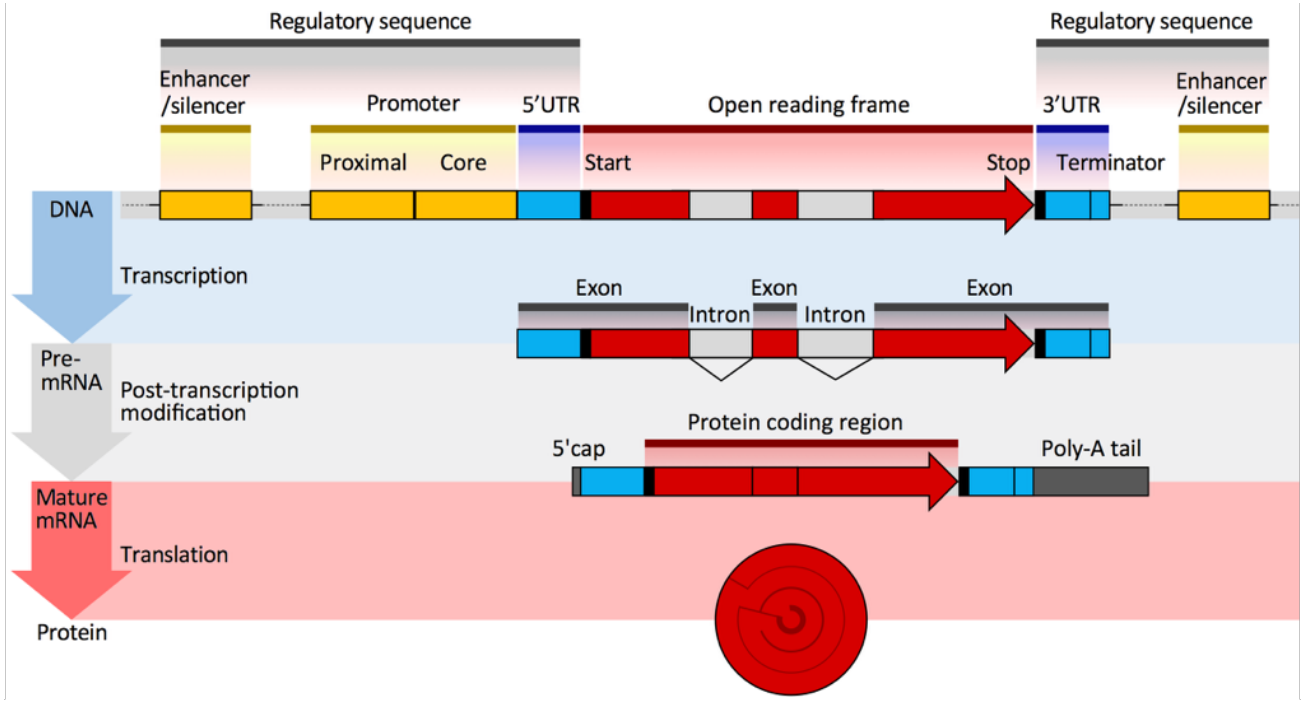
portal.gdc.cancer.gov



RNA-seq Workflow



Experimental Design: **what do we want?**



While **mRNA transcripts have a polyA tail**, many of the non-coding RNA transcripts do not as the post-transcriptional processing is different for these transcripts.

Experimental Design: **number and type of replicates**

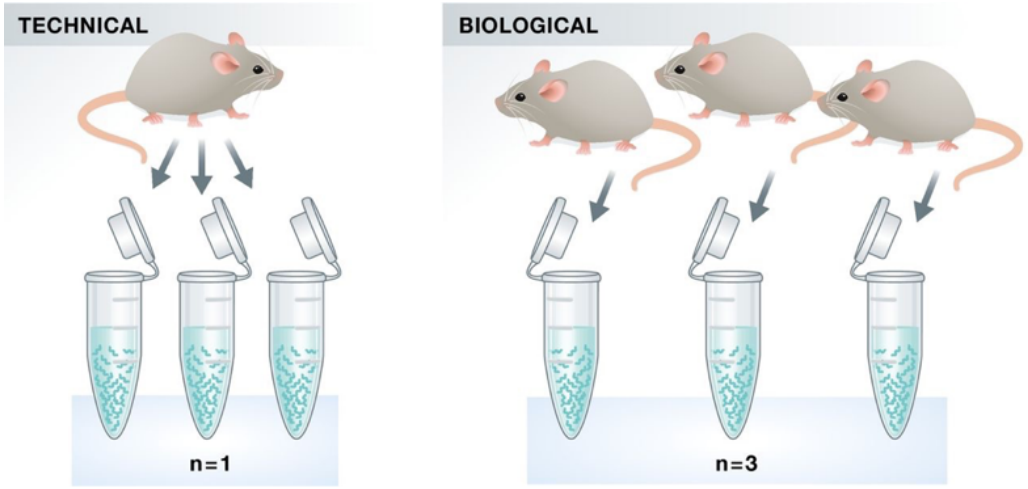


*Image credit: Klaus B., EMBO J (2015) **34**: 2727-2730*

- **Technical replicates:** use the same biological sample to repeat the technical or experimental steps in order to accurately measure technical variation and remove it during analysis (**Microarray**)
- **Biological replicates** use different biological samples of the same condition to measure the biological variation between samples.

Experimental Design: number and type of replicates

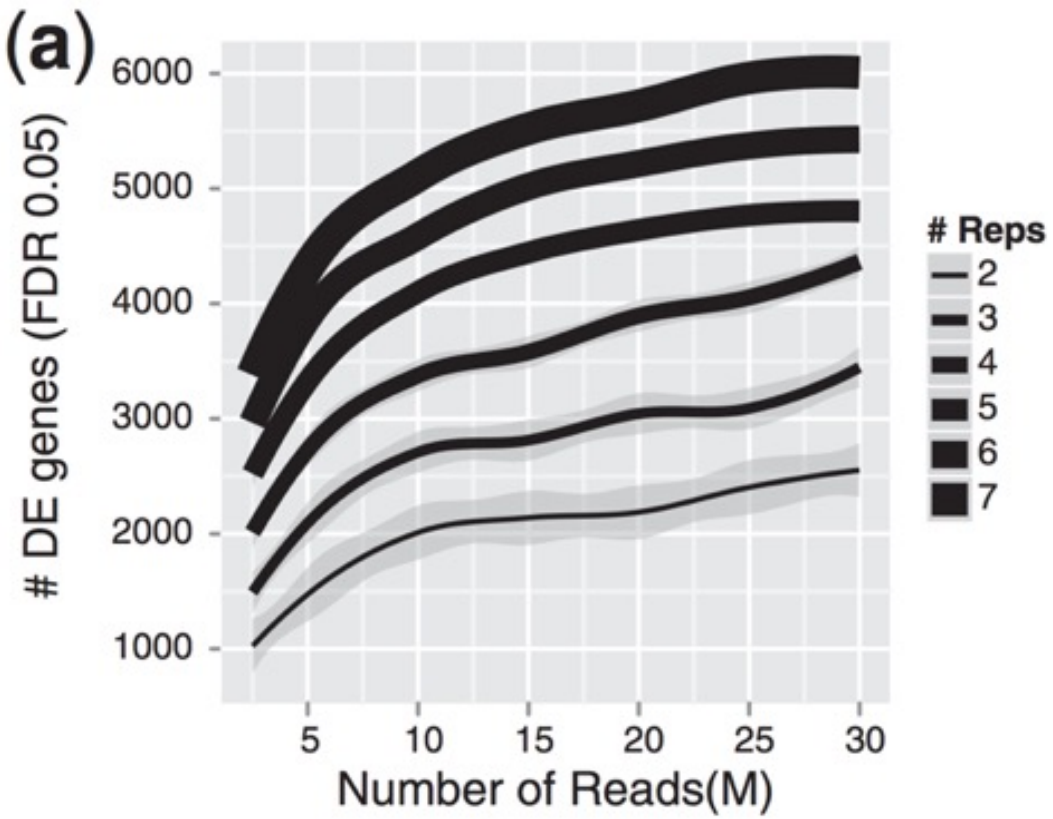
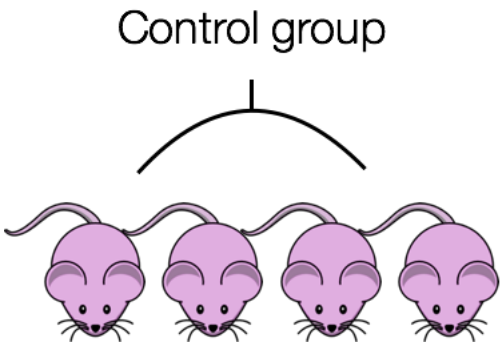


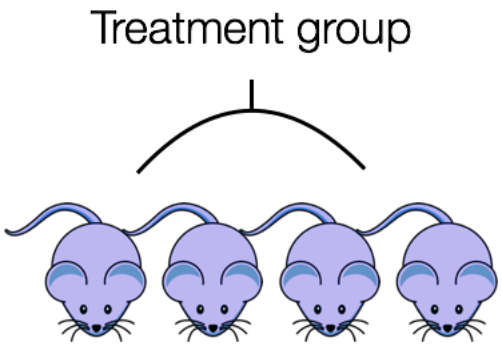
Image credit: [Liu, Y., et al., Bioinformatics (2014) **30**(3): 301–304]

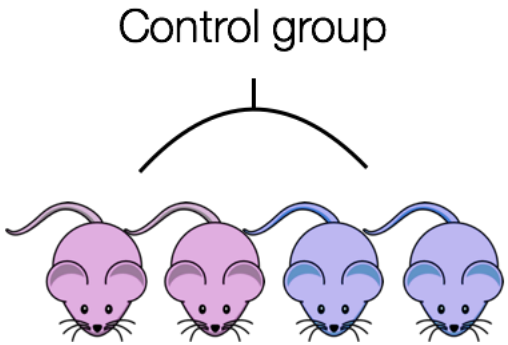
Experimental Design: **confounding factors**

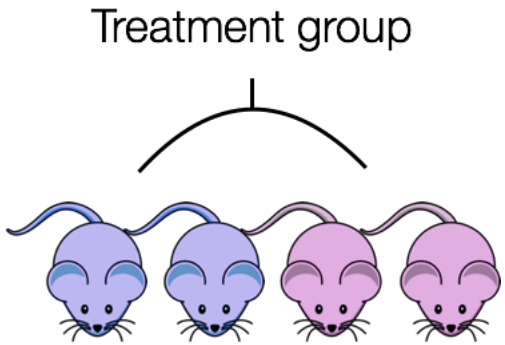
female

male

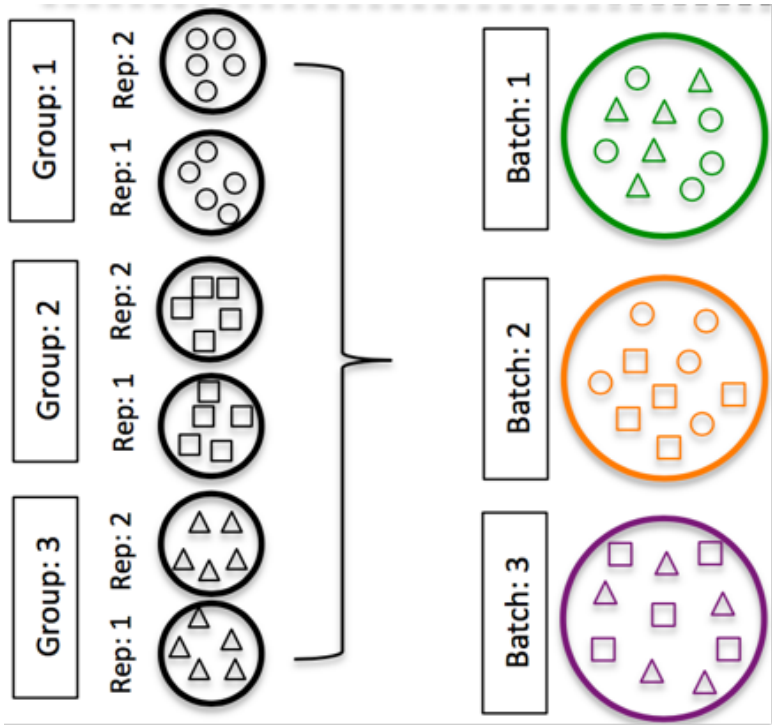


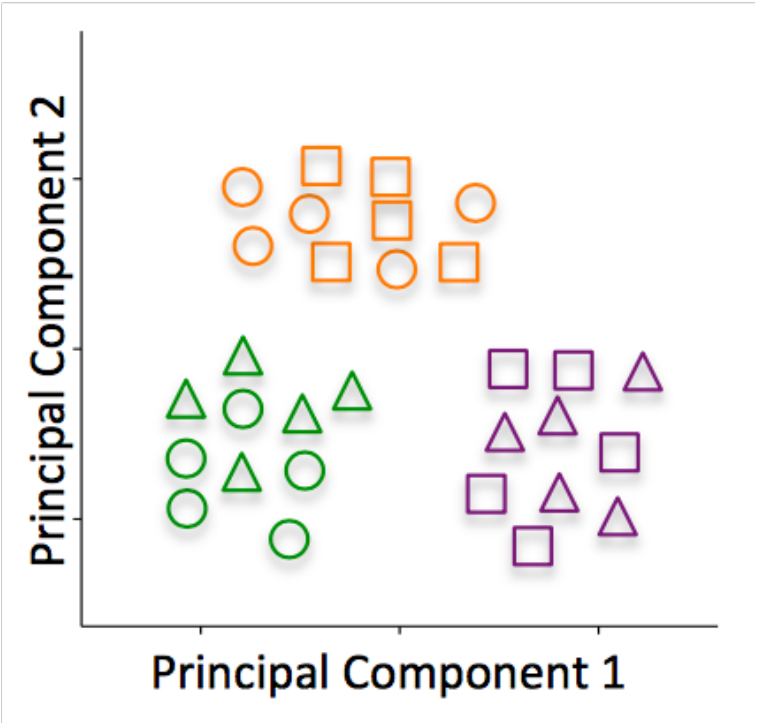






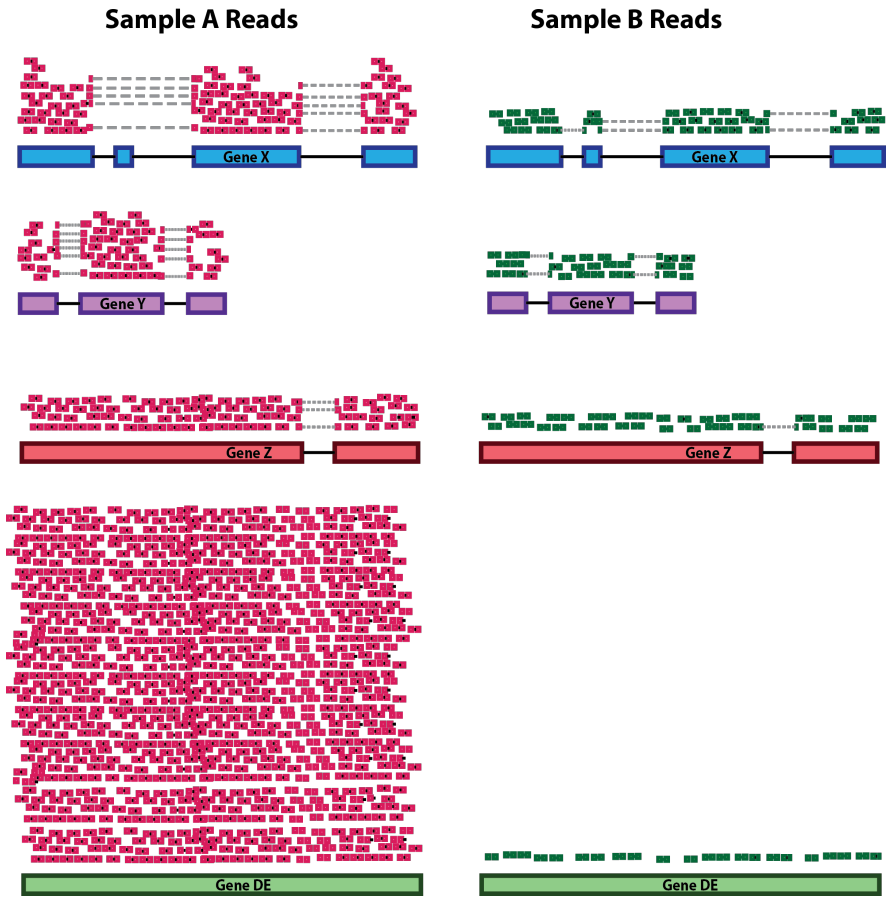
Experimental Design: **Batch effects**





armandorp@ucm.es

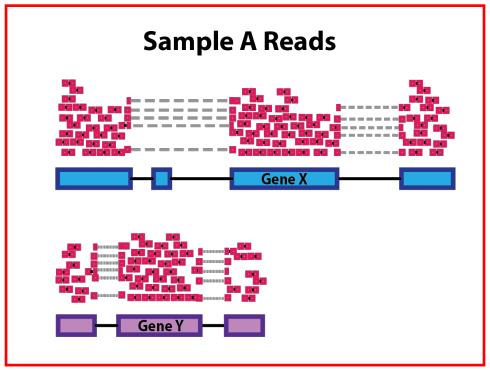
Technical variation (Library preparation) Normalization factors



armandorp@ucm.es

Technical variation (Library preparation) Normalization factors

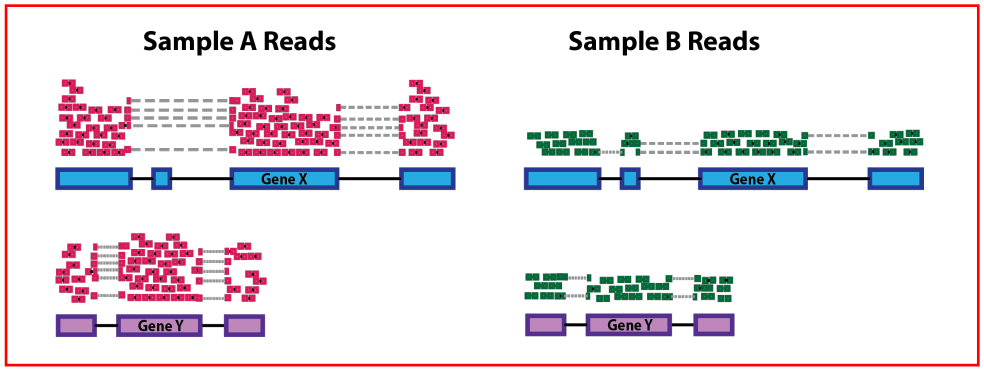
Gene length →



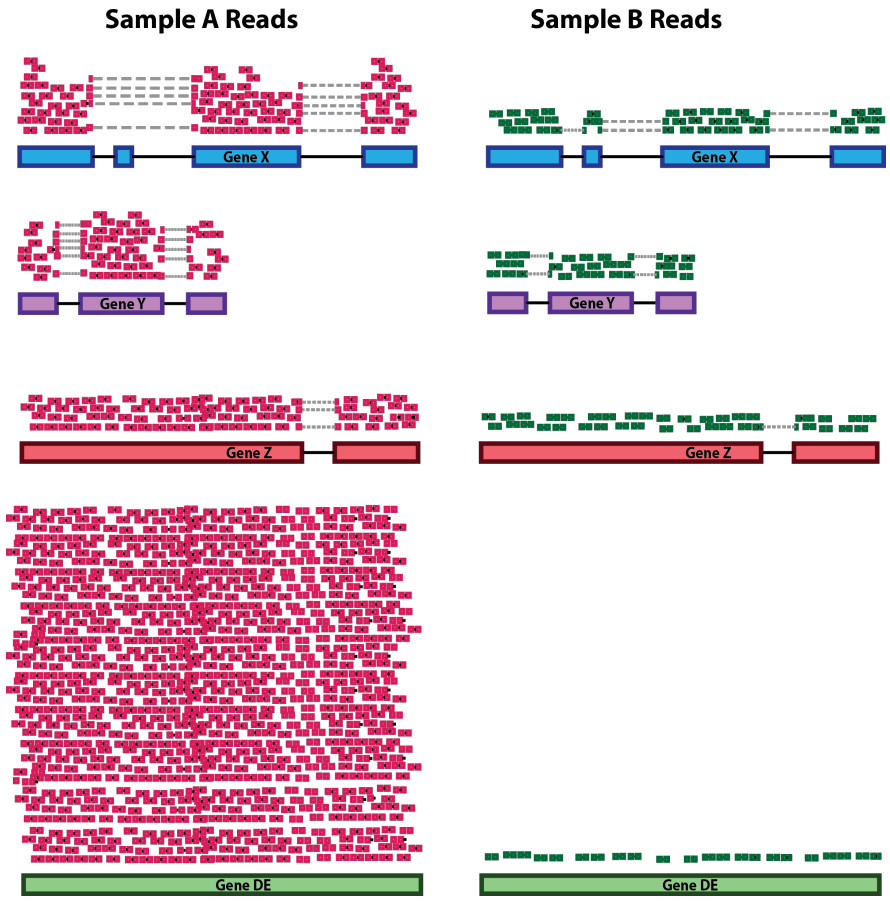
armandorp@ucm.es

Technical variation (Library preparation) Normalization factors

Sequence depth (Library size) →



Technical variation (Library preparation) Normalization factors



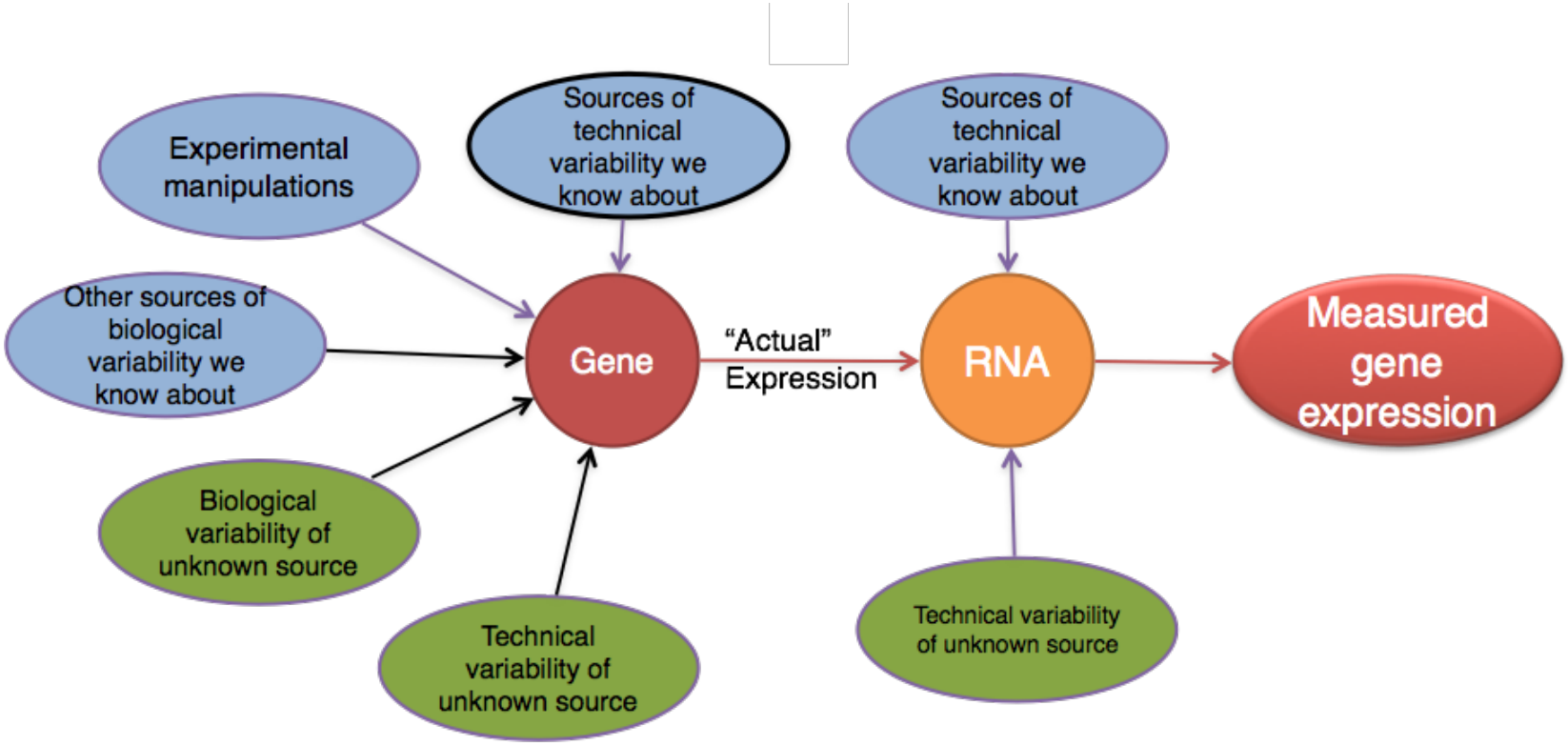
Minimizing the effect of highly expressed genes→

Normalization factors in RNAseq data

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**

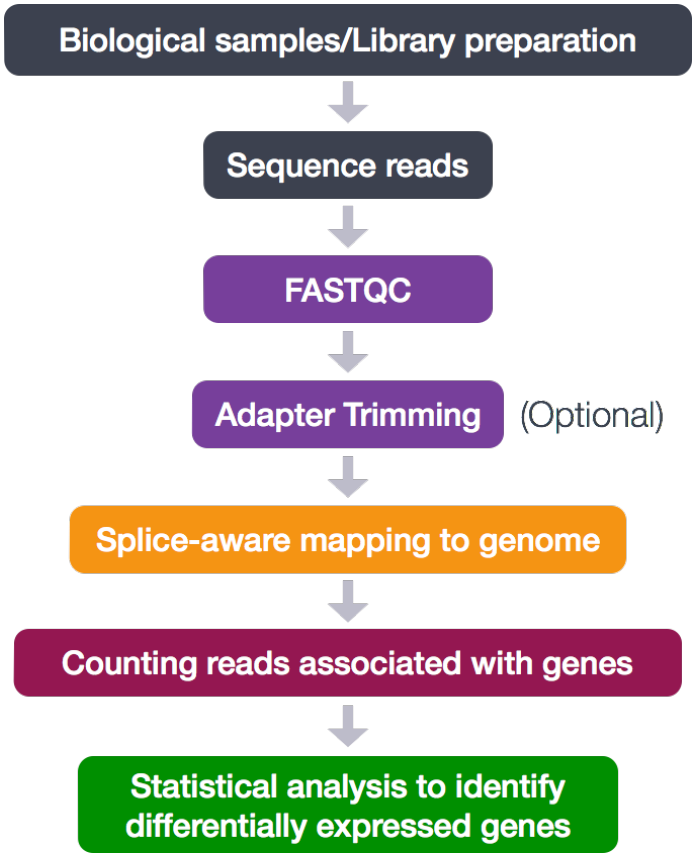
armandorp@ucm.es

Sources of variation in assays



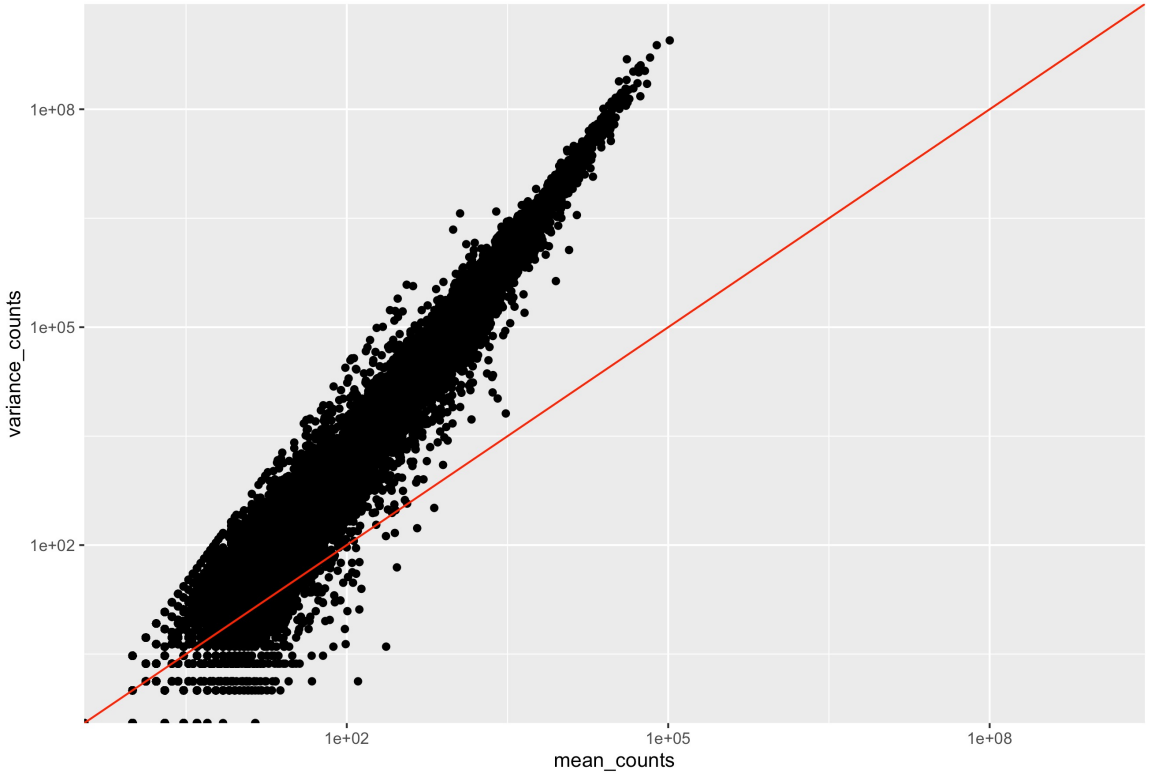
Courtesy of Paul Pavlidis, UBC

Raw RNAseq data and quantifications

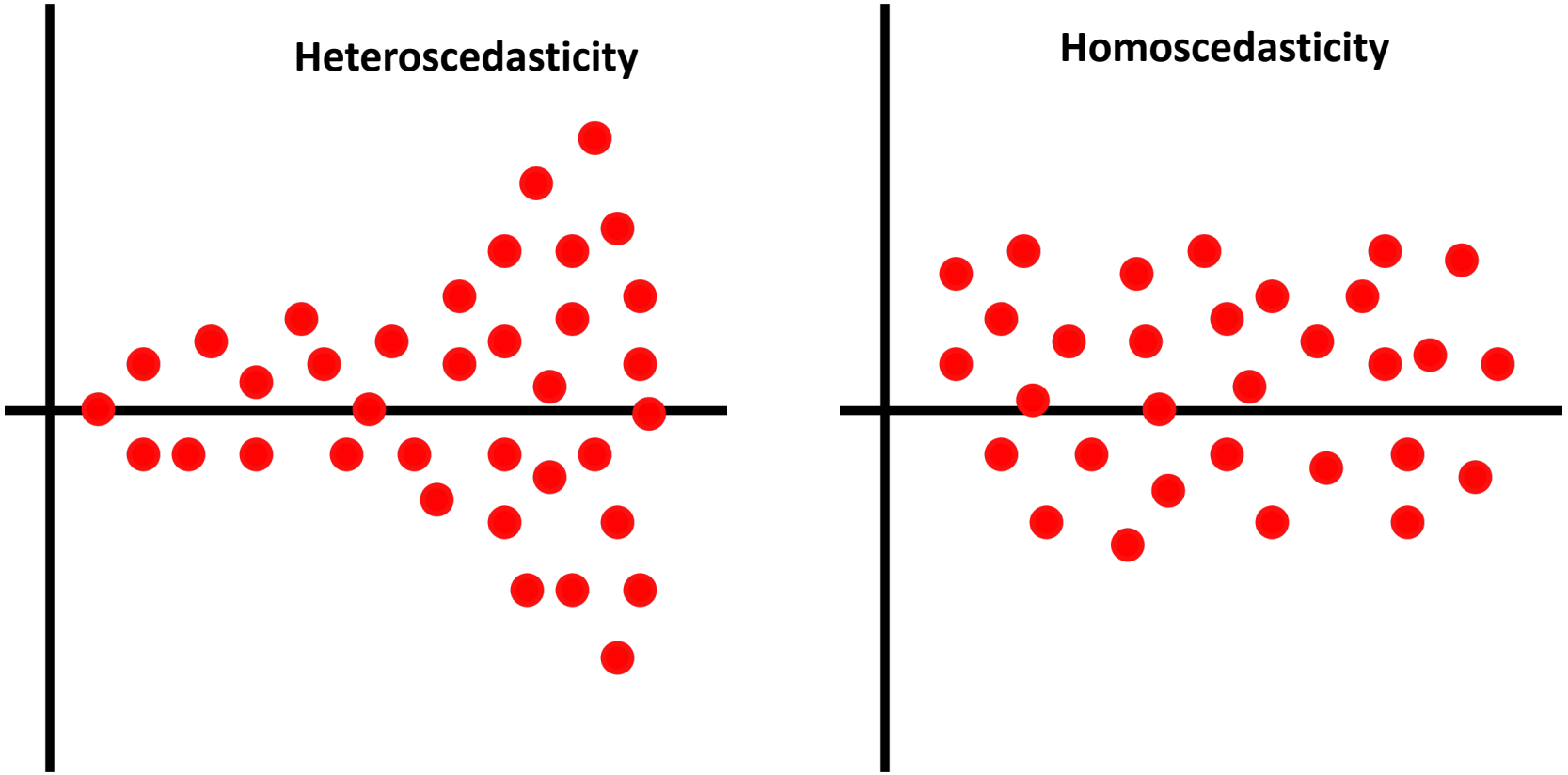


armandorp@ucm.es

Variance increases with mean expression



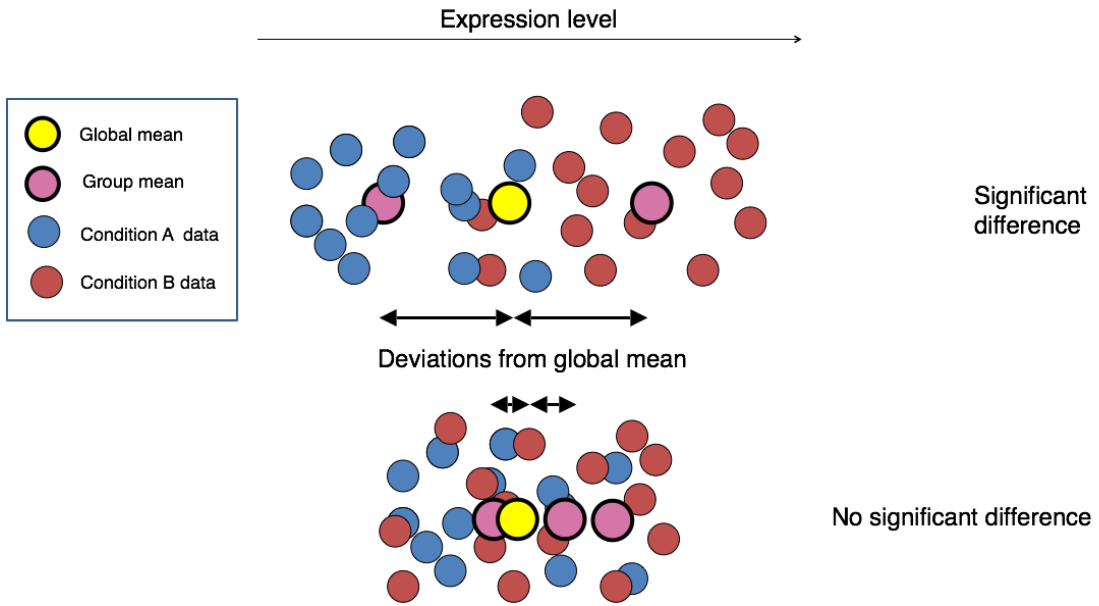
Raw counts in RNAseq



armandorp@ucm.es

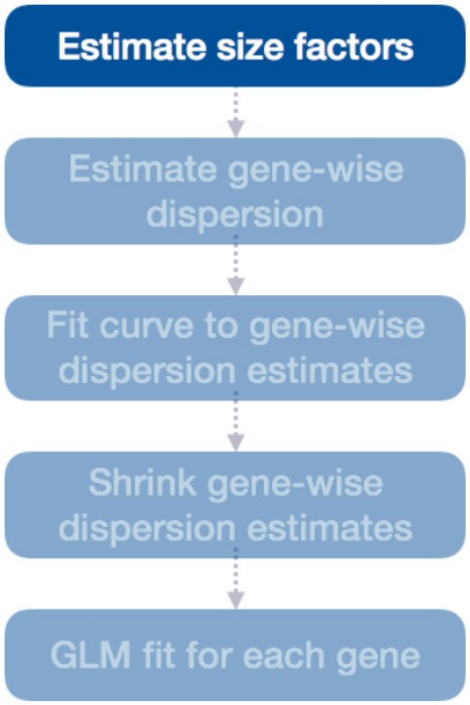
Differential Expression Analysis (DEA) steps with DESeq2

1. Estimate size factors
2. Estimate gene-wise dispersions
3. Fit curve to gene-wise dispersion estimates
4. Shrink gene-wise dispersion estimates
5. GLM fit for each gene



armandorp@ucm.es

DESeq2: (1) Size Factors



$$\sqrt{1489 \cdot 906} = 1161.5$$

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
...	...	...	...		

DESeq2: (2-4) Gene wise dispersions

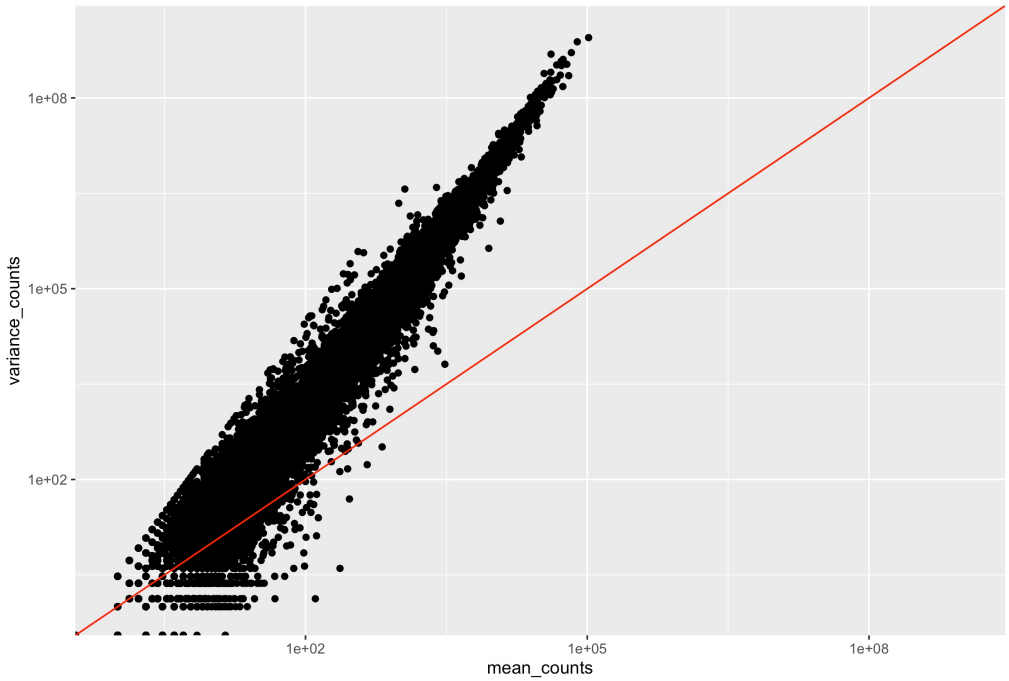
Estimate size factors

Estimate gene-wise dispersion

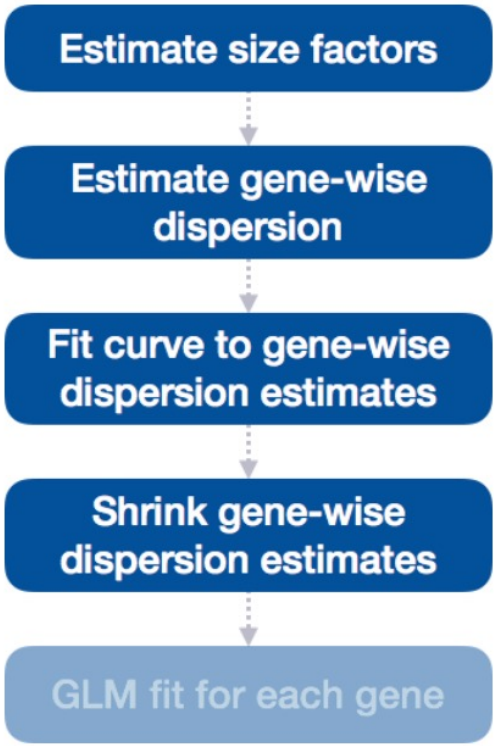
Fit curve to gene-wise dispersion estimates

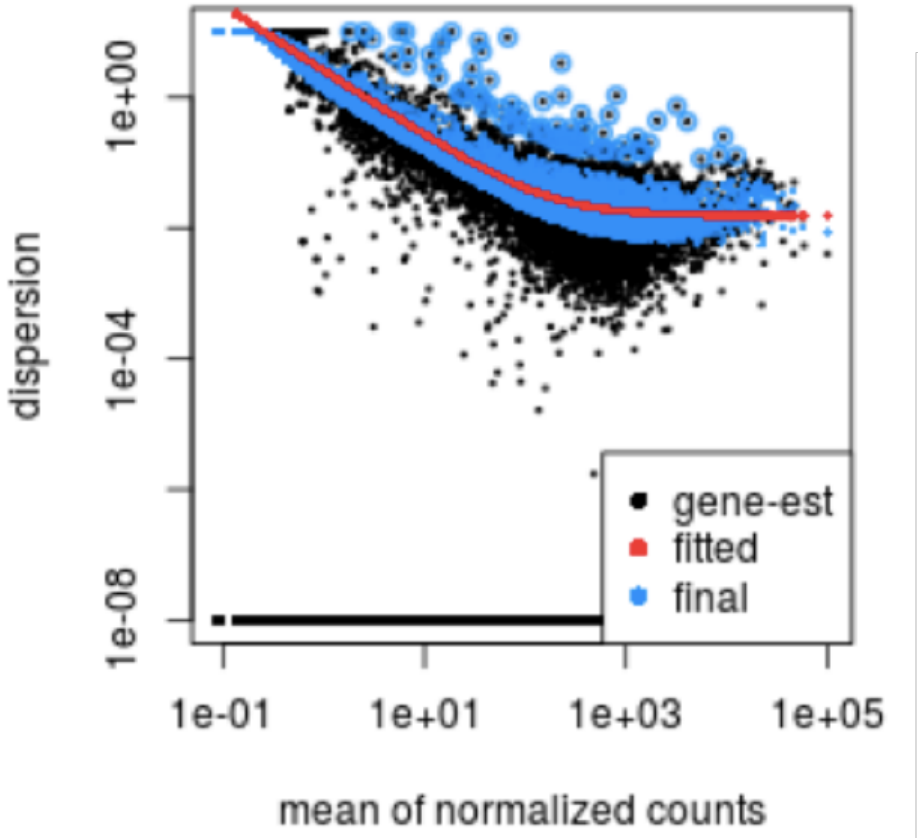
Shrink gene-wise dispersion estimates

GLM fit for each gene

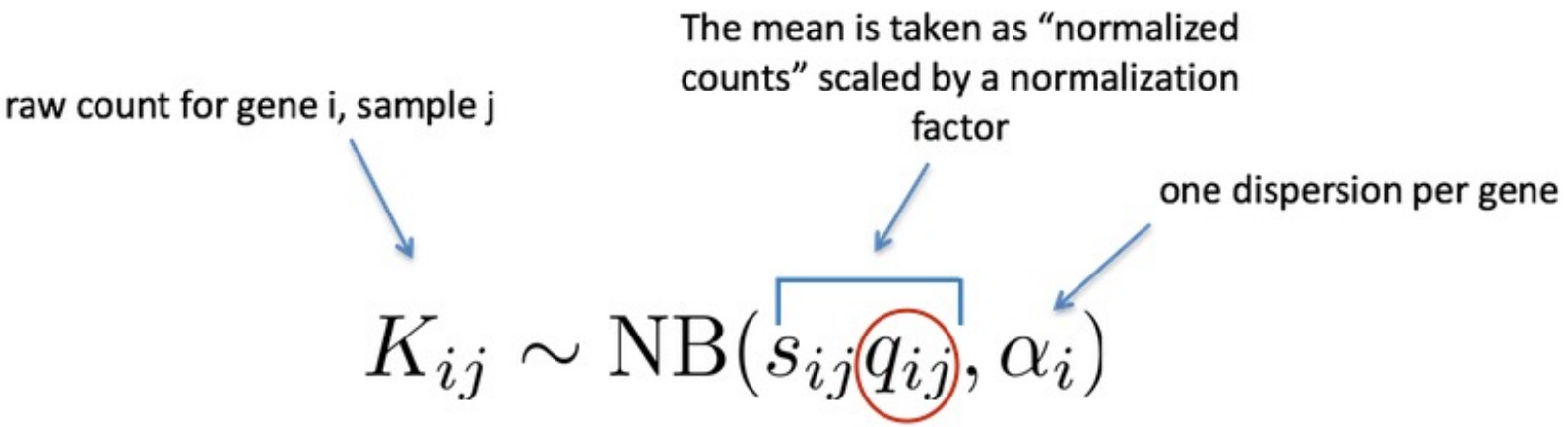


DESeq2: (2-4) Gene wise dispersions





DESeq2: (5) Negative Binomial Model

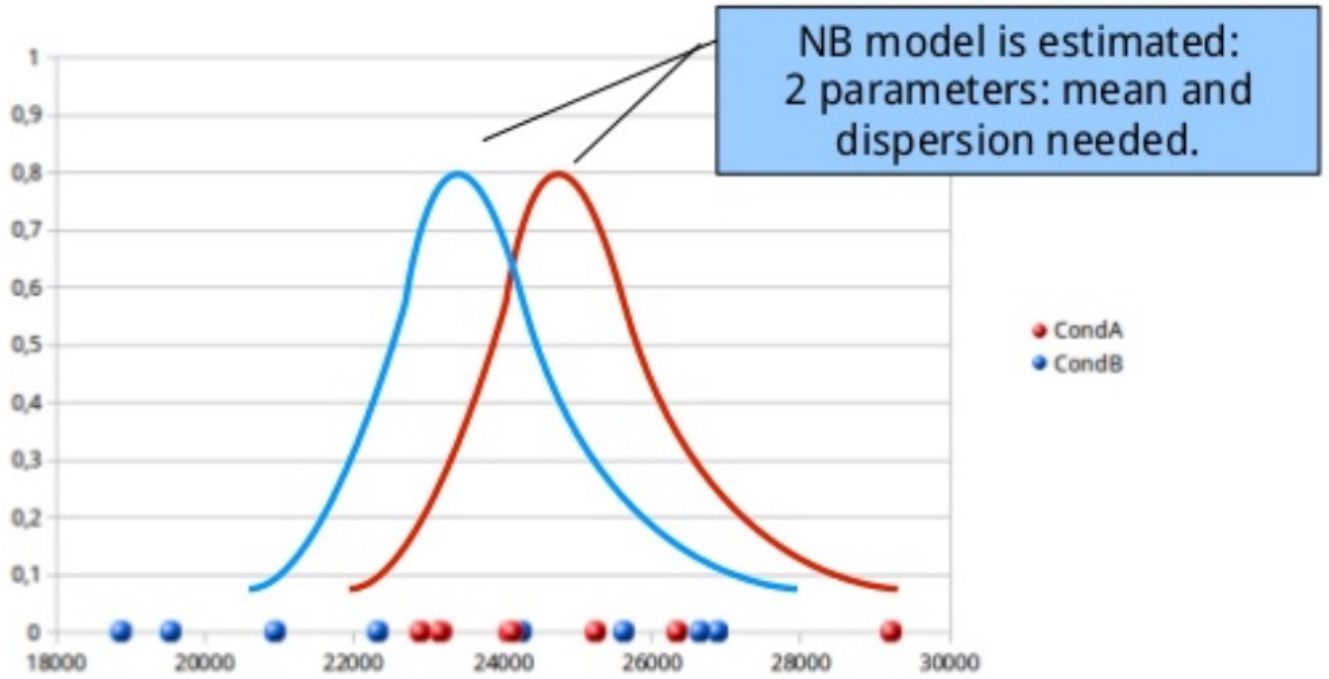


T**he NB model is a good approximation for data where the mean < variance**, as is the case with RNA-Seq count data.

DESeq2: Negative Binomial Model

gene_id CAF0006876

Condition A	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8
	23171	22903	29227	24072	23151	26336	25252	24122
Condition B	Sample9	sample10	sample11	sample12	sample13	sample14	sample15	sample16
	19527	26898	18880	24237	26640	22315	20952	25629



armandorp@ucm.es