RNA-seq
(RNA sequencing)

Beatriz Urda García

SGM Genomics Feb 2021

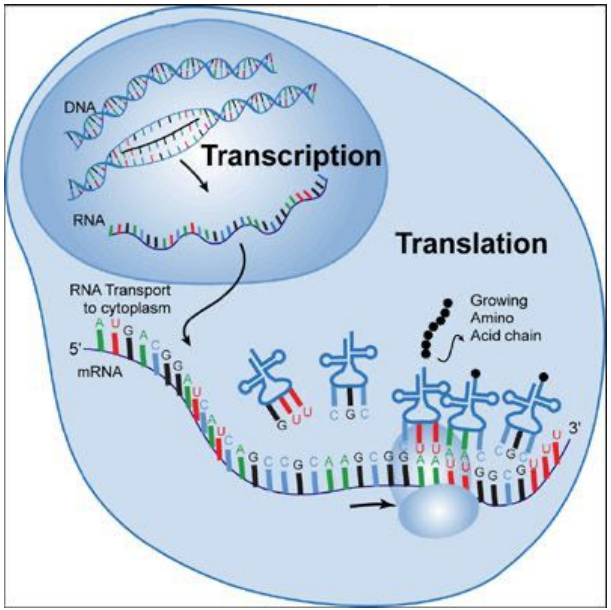
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Outline

- What is RNA-seq
- RNA sequencing steps and data analysis
  (focusing on Differential Expression Analysis)
- Applications

The transcriptome

The complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physical condition Wang et al., 2009.



RNA-seq

RNA-seq (RNA sequencing) is a technique destined to the **identification** and **quantification** of RNA in a biological sample in a given moment.

Steps:

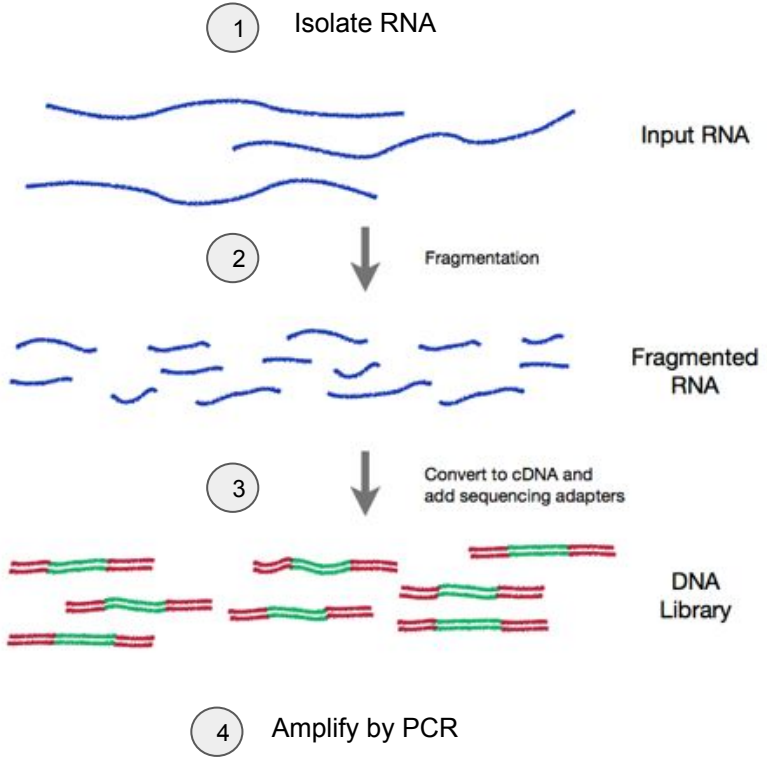
1. Preparing a sequencing library
2. Sequencing
3. Data Analysis

We are able to measure

gene expression

Preparing an RNA-seq library

Illumina protocol



Sequencing the library

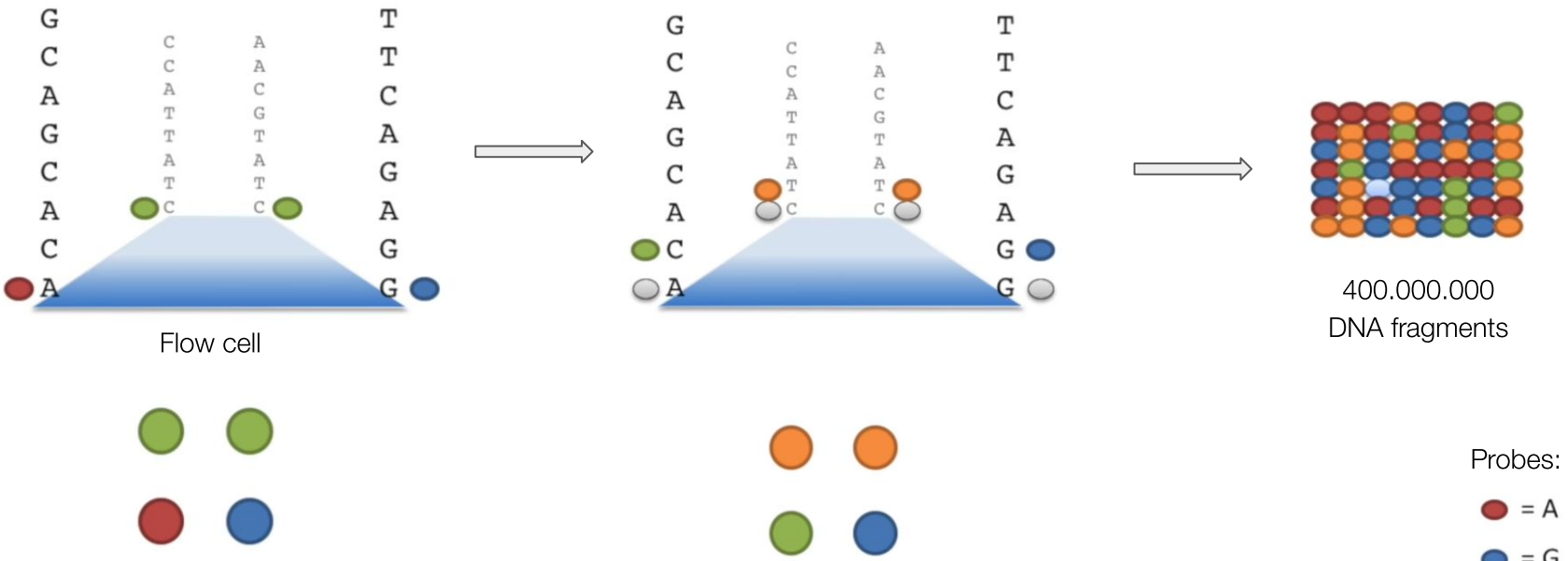


Figure adapted from StatQuest

Sequencing the library

The raw data:

```
@NS500177:196:HFTTTAFXX:1:11101:10916:1458 2:N:0:CGCGGCTG
ACACGACGATGAGGTGACAGTCACGGAGGATAAGATCAATGCCCTCATTAAAGCAGCCGGTGTAA
+
AAAAAEEEEEEEEEEE//AEEEAEEEEEEEEEEE/EE/<<EE/AAEEAEE///EEEEAEEEAEA<
```

Each sequencing read consists of 4 lines of data:

- The first line always starts with a @ and contains an unique ID for the sequence that follows.
- The second line contains the bases of the sequenced fragment.
- The third line is always ‘+’
- The fourth line contains quality scores for each base in the sequenced fragment.

RNA-seq Data Analysis - Quality control

Filter out:

- Low quality reads
- Poor-quality bases
- Adaptors

Software: FastQC, Trimmomatic

RNA-seq Data Analysis - Infer which transcripts are expressed

- If a reference genome is available, you can **map your reads to the genome.**
  - QC: 70-90% mapped reads to the genome.
  - You can also map your reads to a reference transcriptome (slightly lower mapping % are expected because unannotated transcripts will be lost).

- If not, RNA-seq reads can be **assembled de novo** into a transcriptome.
  - Reads are assembled into longer contigs. Then, these contigs are treated as the expressed transcriptome to which reads are mapped again for quantification.

Only genome mapping allows for **transcript discovery!**

RNA-seq Data Analysis - Infer which transcripts are expressed...

- If a reference genome is available, you can **map your reads to the genome.**
  - QC: 70-90% mapped reads to the genome. ★
  - You can also map your reads to a reference transcriptome (slightly lower mapping % are expected because unannotated transcripts will be lost).

- If not, RNA-seq reads can be **assembled de novo** into a transcriptome.
  - Reads are assembled into longet contigs. Then, these contigs are treated as the expressed transcriptome to which reads are mapped again for quantification.

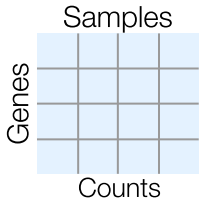
Only genome mapping allows for **transcript discovery!**

... and how much?

RNA-seq Data Analysis - Transcript quantification

The most common application of RNA-seq!

It consists of counting the number of reads per gene (**counts**).

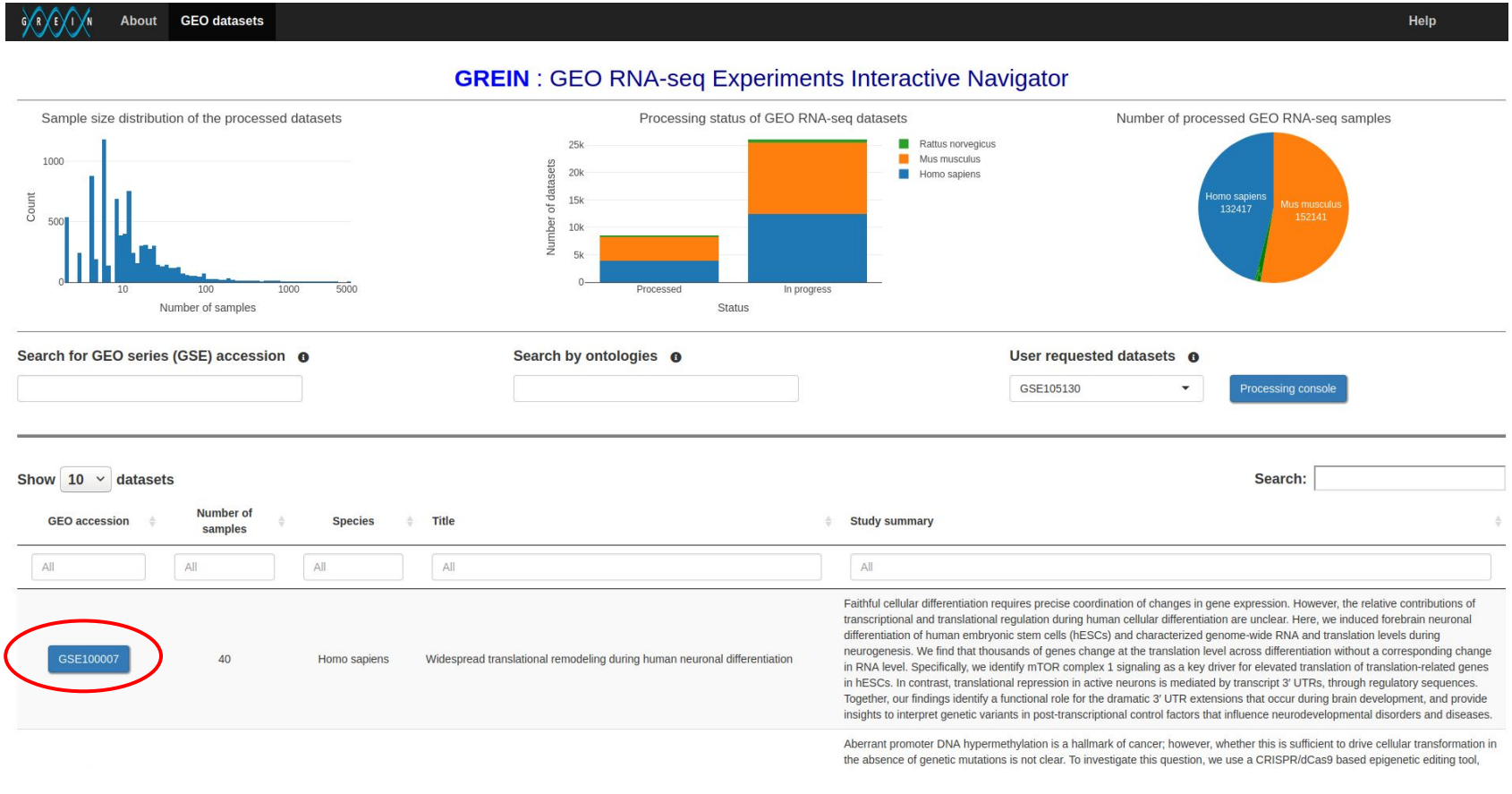


Gene	Sample1	Sample2	Sample3…
A1BG	30	5	13…
A1BG-AS1	24	10	18…
A1CF	0	0	0…
A2M	5	9	7…
A2M-AS1	3563	5771	4123…
A2ML1	13	8	7…
...	...	...	...

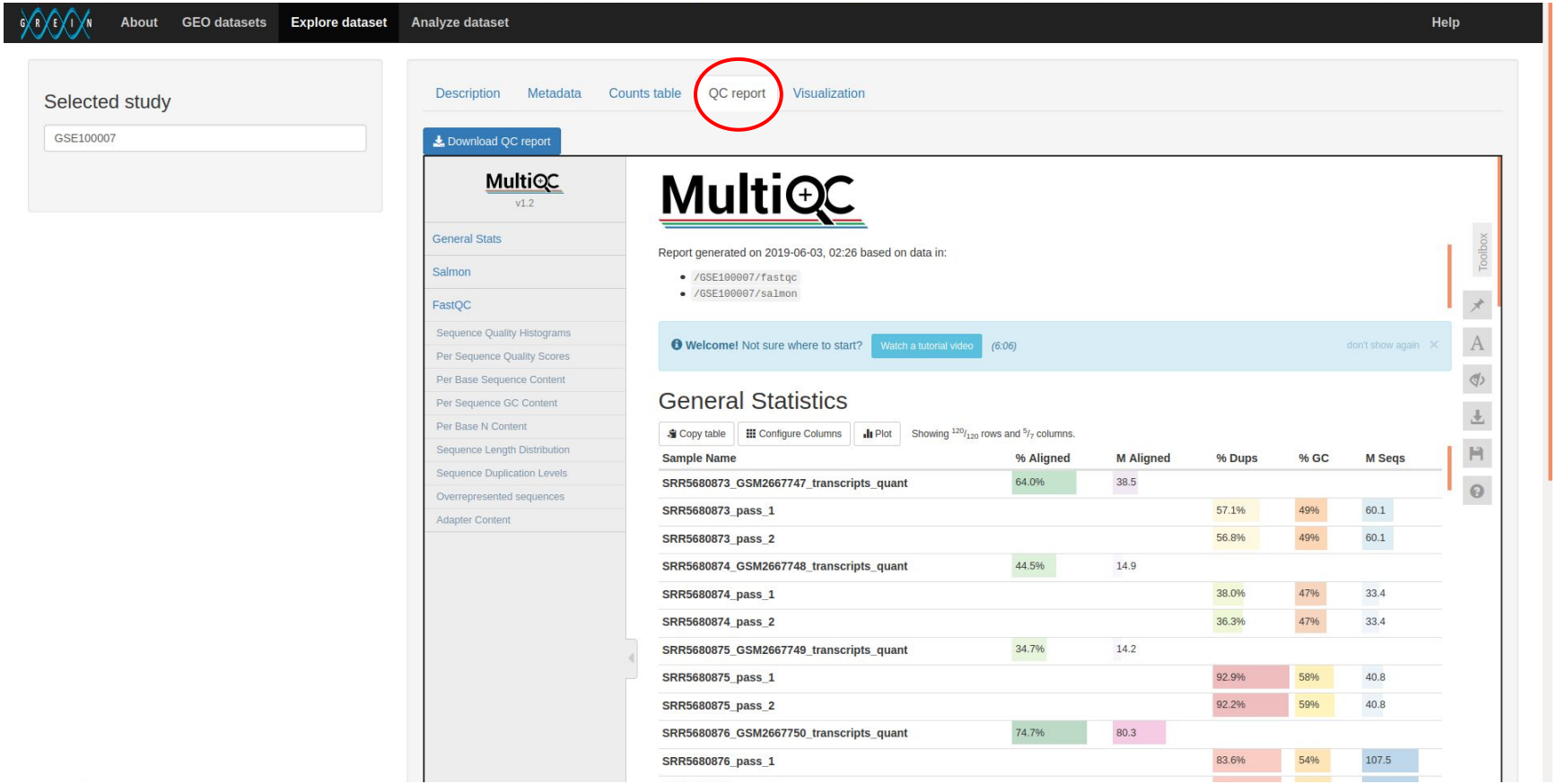
The human genome has ~20.000 **genes** (rows)
At least 3 control and 3 case **samples** (columns)

⟹ **Plot the data:**
PCA, t-SNEs

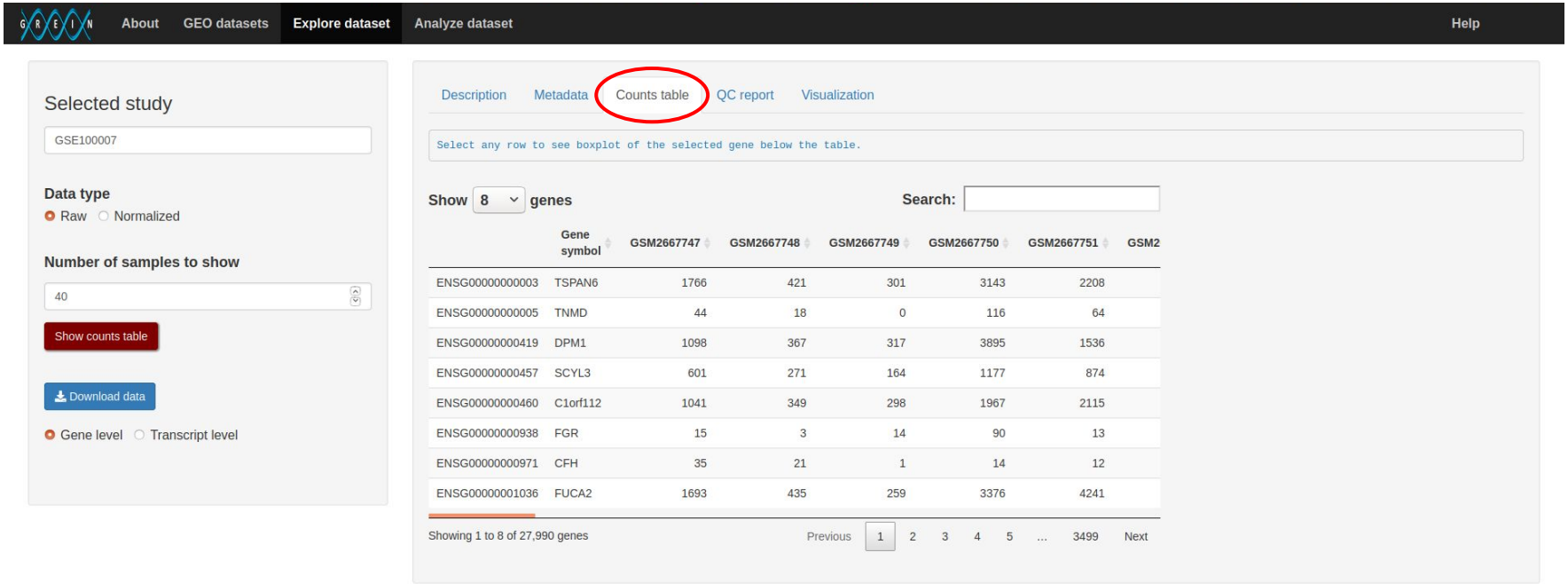
RNA-seq Data Analysis - Resource: GREIN



RNA-seq Data Analysis - Resource: GREIN

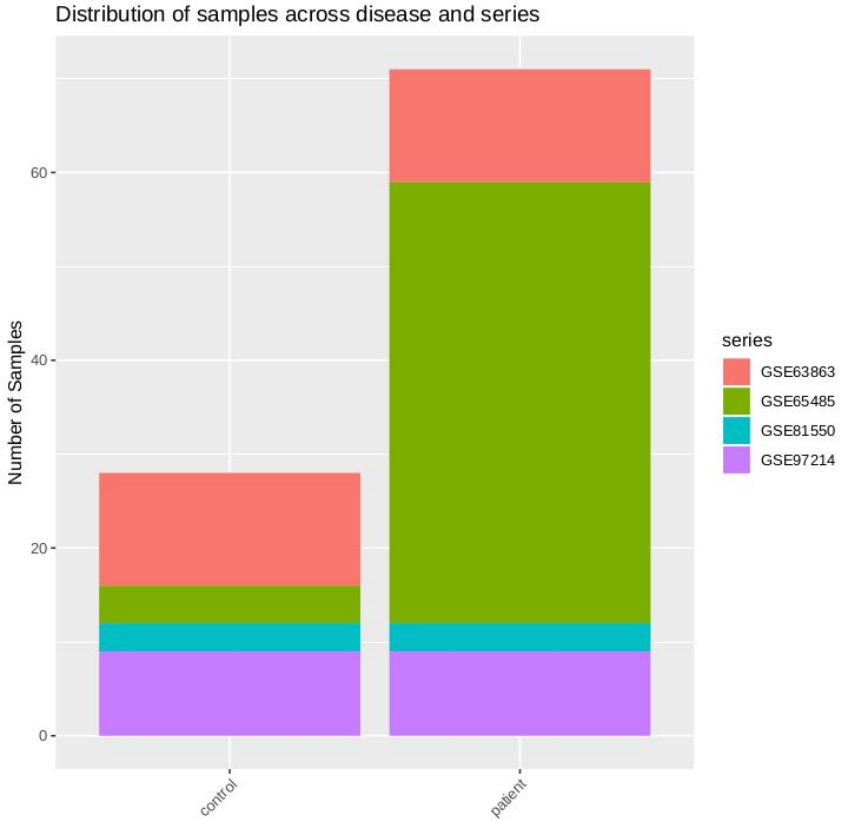


RNA-seq Data Analysis - Resource: GREIN



	Gene symbol	GSM2667747	GSM2667748	GSM2667749	GSM2667750	GSM2667751	GSM2
ENSG00000000003	TSPAN6	1766	421	301	3143	2208	
ENSG00000000005	TNMD	44	18	0	116	64	
ENSG00000000419	DPM1	1098	367	317	3895	1536	
ENSG00000000457	SCYL3	601	271	164	1177	874	
ENSG00000000460	C1orf112	1041	349	298	1967	2115	
ENSG00000000938	FGR	15	3	14	90	13	
ENSG00000000971	CFH	35	21	1	14	12	
ENSG00000001036	FUCA2	1693	435	259	3376	4241	

RNA-seq Data Analysis - Differential Expression Analysis



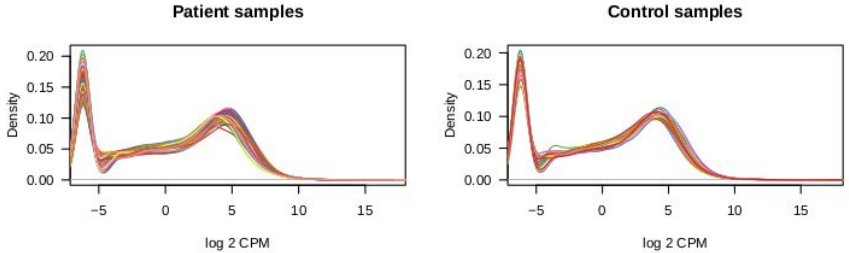
Hepatocellular carcinoma (liver)

Which genes have an altered expression in the patients with respect to the controls?

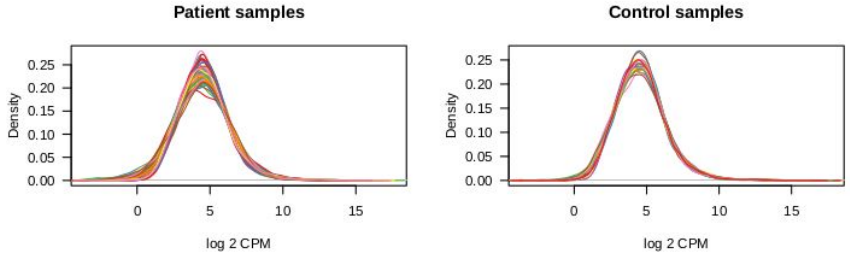
Significantly Differentially Expressed Genes (**sDEGs**)

RNA-seq Data Analysis - Filtering lowly expressed genes

Before filtering



After filtering lowly expressed genes



Lowly expressed genes (those with less than 1 $\log_2$CPM in more than 20% of the samples)

RNA-seq Data Analysis - Normalization

We need to normalize the data because each sample might have a different total number of reads assigned to it.

(One sample might have more low quality reads or slightly higher concentration on the flow cell).

Adjusting for differences in library sizes
(**within-sample normalization**)

Gene	Sample #1 635 reads	Sample #2 1,270 reads
A1BG	30	60
A1BG-AS1	24	48
A1CF	0	0
A2M	563	1126
A2M-AS1	5	10
A2ML1	13	26

Methods: logCPM, RPKM, FPK, RSEM

RNA-seq Data Analysis - Normalization

We also need to correct for the fact that different samples might have a different set of genes transcribed.

E.g., comparing tissues or in a knock-out experiment.

Adjusting for differences in library composition
(**between-sample normalization**)

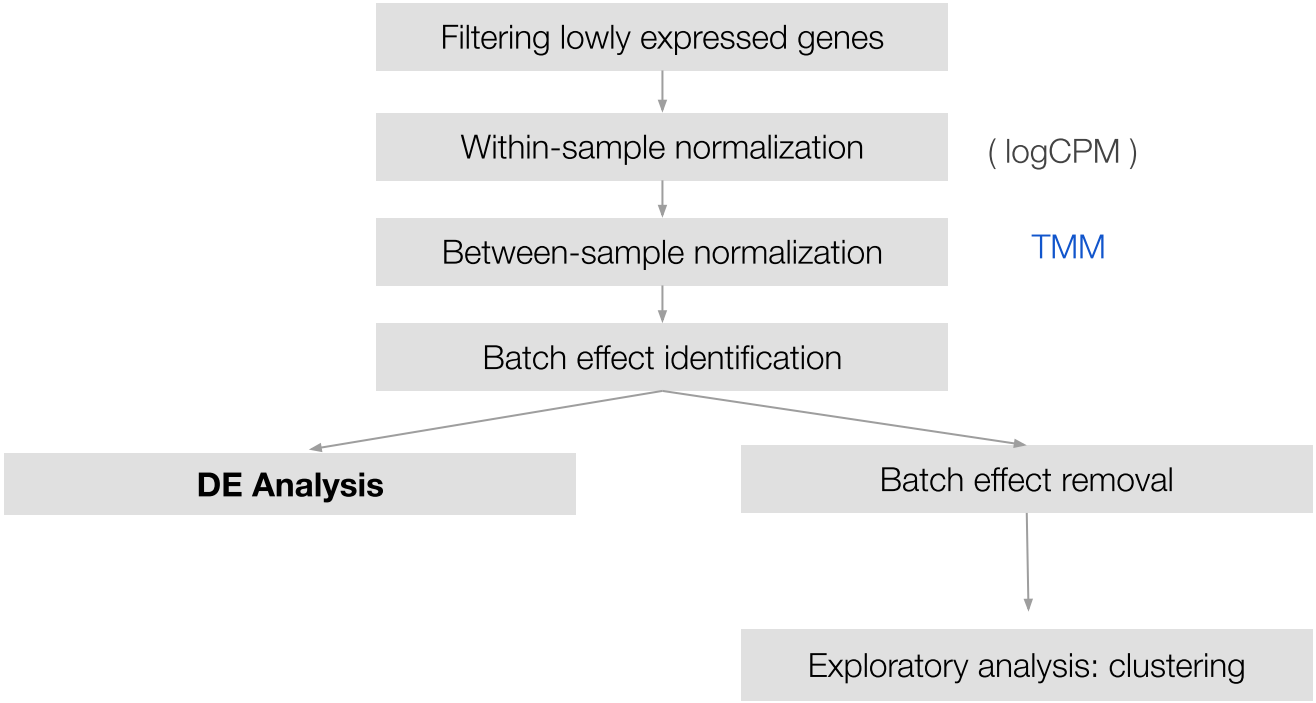
Gene	Sample #1 635 reads	Sample #2 635 reads
A1BG	30	235
A1BG-AS1	24	188
A1CF	0	0
A2M	563	0
A2M-AS1	5	39
A2ML1	13	102

Assume that only Sample 1 transcribed A2M

Then, the 563 reads used up by A2M in Sample 1 will be distributed to other genes in Sample 2

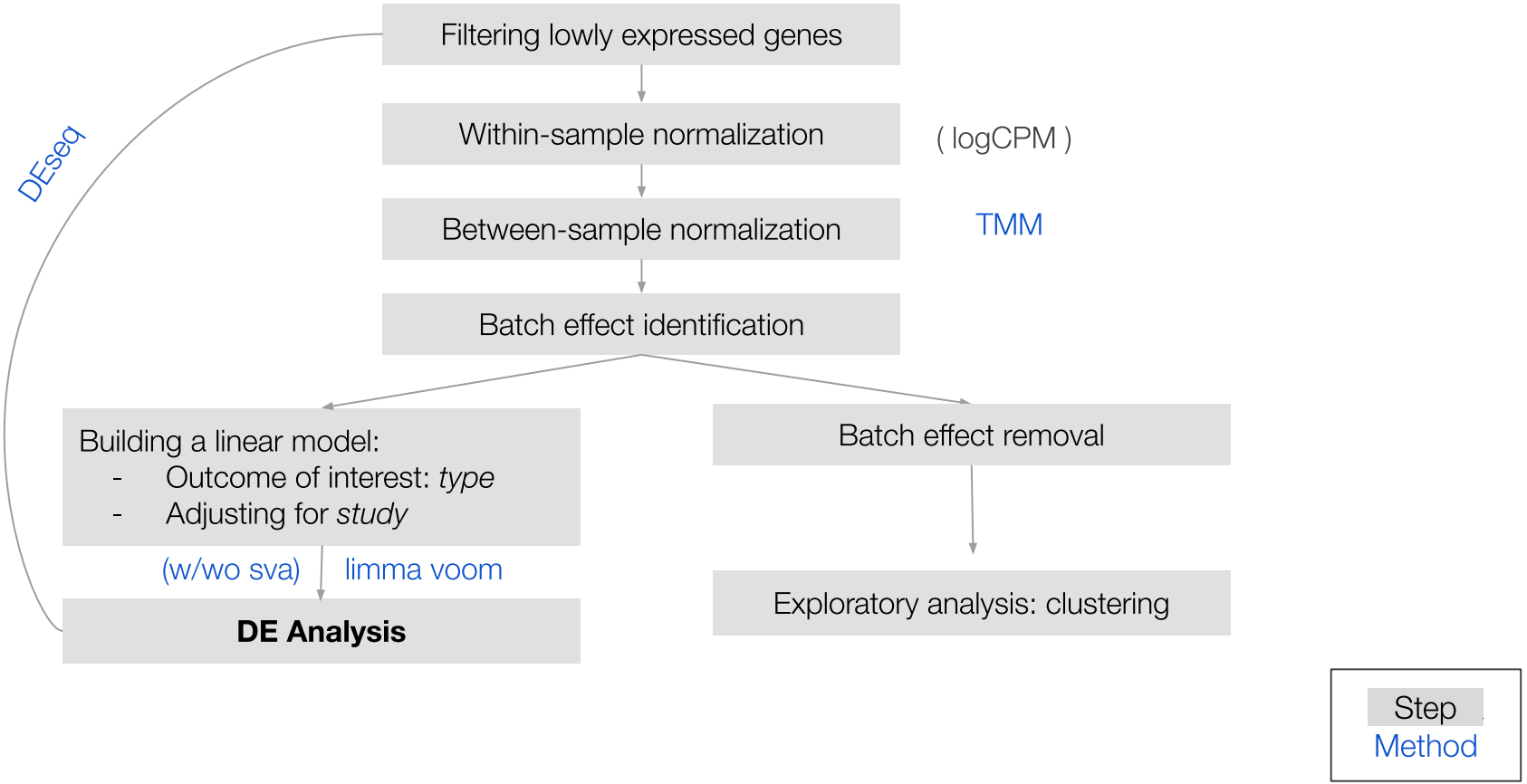
Methods: TMM

A real example - RNA-seq analysis for a human disease



Step
Method

A real example - RNA-seq analysis for a human disease

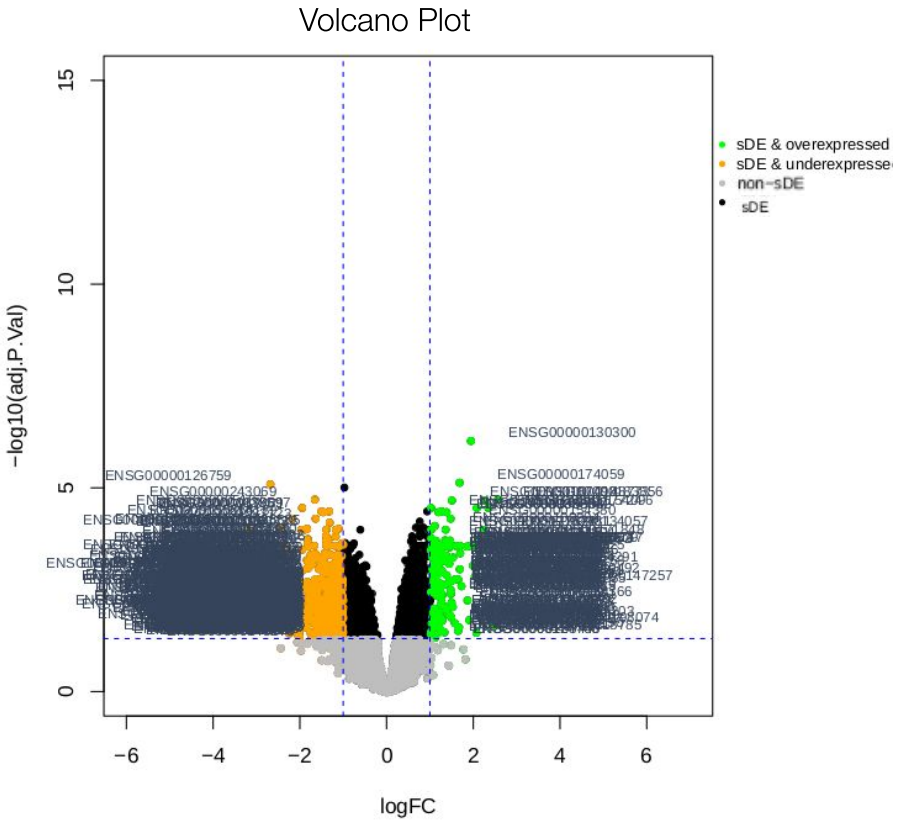


A real example - RNA-seq analysis for a human disease

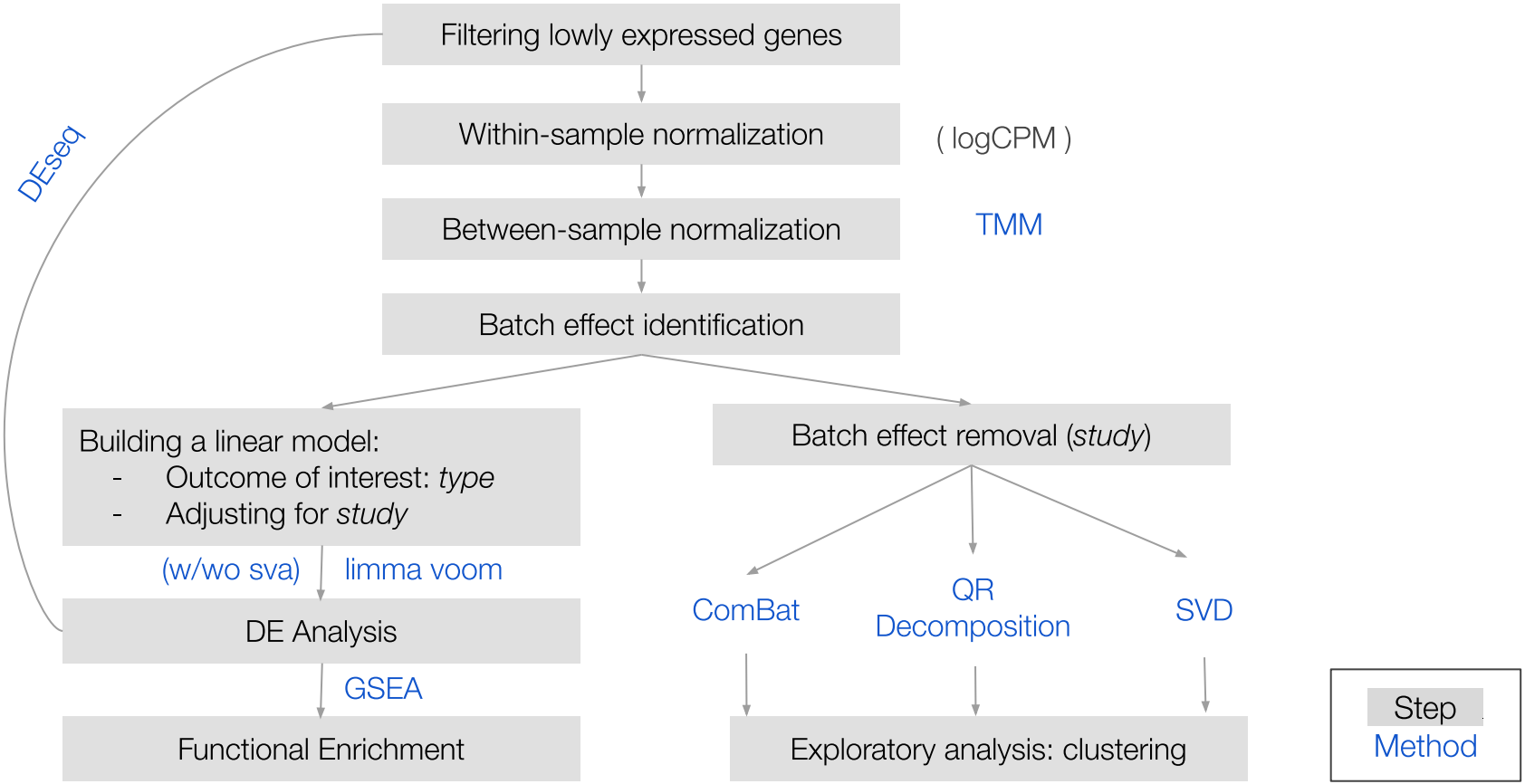
symbol	logFC	AveExpr	P.Value	adj.P.Val
ENSG00000136247	0.431193412624311	4.53206091658694	1.25515900002734E-07	0.001683168219037
ENSG00000123297	0.546242249989598	4.55757235090766	9.24594372649256E-07	0.006199405268613
ENSG00000196704	0.450928125588195	4.67783175684838	2.77249913109058E-06	0.012393071115975
ENSG00000128039	0.827160814514575	3.69123269239815	5.12379221449353E-06	0.014636695624259
ENSG00000142089	0.802968199645196	5.77745864548944	7.7331120050691E-06	0.014636695624259
ENSG00000180817	0.584449740405807	6.91369085334641	8.79547839444059E-06	0.014636695624259
ENSG00000124614	0.654441160333983	8.38732118508355	1.4658695900797E-05	0.014636695624259
ENSG00000204397	0.706090252140031	3.94699985101853	1.58683042963689E-05	0.014636695624259
ENSG00000105677	0.504256623697782	5.0595450008827	1.69579198106334E-05	0.014636695624259
ENSG00000197756	0.650467463308897	8.40492033517472	1.70609497565493E-05	0.014636695624259
ENSG00000274950	0.602595958739266	5.97886621269216	1.75006081959288E-05	0.014636695624259
ENSG00000124429	-0.65199875309962	7.48981448430548	1.75851231534655E-05	0.014636695624259
ENSG00000169740	0.574282207434585	3.51433694812706	1.71175774433926E-05	0.014636695624259
ENSG00000196233	-0.576867396671793	6.492375038996	1.99390489719467E-05	0.014636695624259
ENSG00000107862	-0.442479092291929	6.97126842861621	2.03873309358238E-05	0.014636695624259
ENSG00000131469	0.568446310713038	8.19176468461283	2.14630692750236E-05	0.014636695624259
ENSG00000116406	-0.382894010381702	6.86391180235511	2.24895368284029E-05	0.014636695624259
ENSG00000025772	0.500950529608195	4.29310370993621	2.27686749134424E-05	0.014636695624259
ENSG00000186472	-0.813912034209522	3.76771811263154	2.15002160523475E-05	0.014636695624259
ENSG00000233276	0.546532380109724	6.17485764851967	2.34621159475879E-05	0.014636695624259
ENSG00000065621	0.576288241226278	4.15666666426999	2.32765835072918E-05	0.014636695624259
ENSG00000109475	0.666424754154782	8.16983149693461	2.78764440862606E-05	0.014636695624259

...

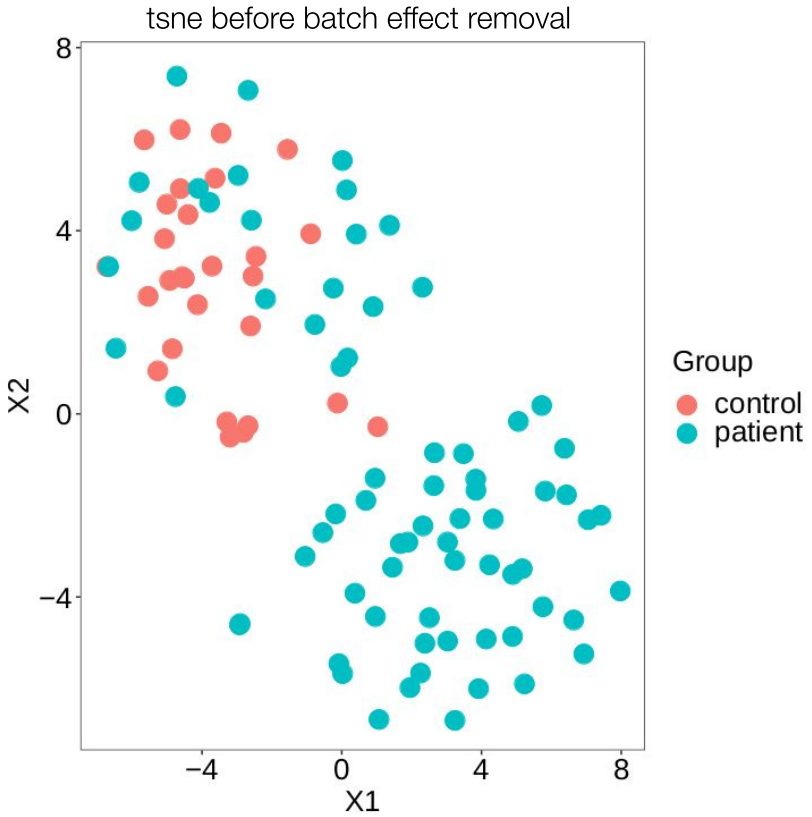
A real example - RNA-seq analysis for a human disease

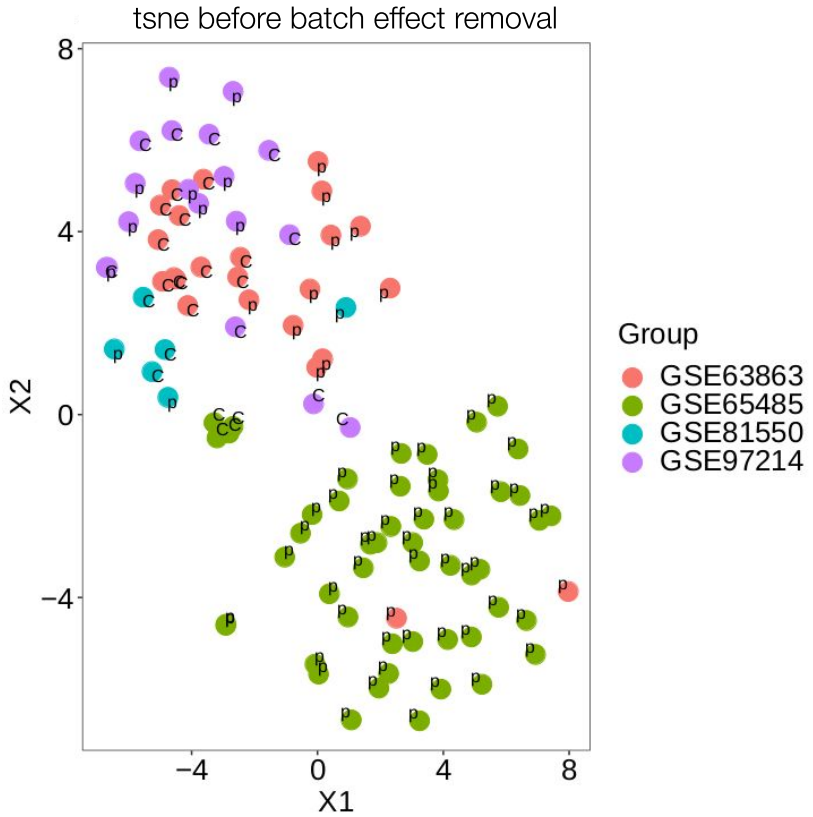


A real example - RNA-seq analysis for a human disease

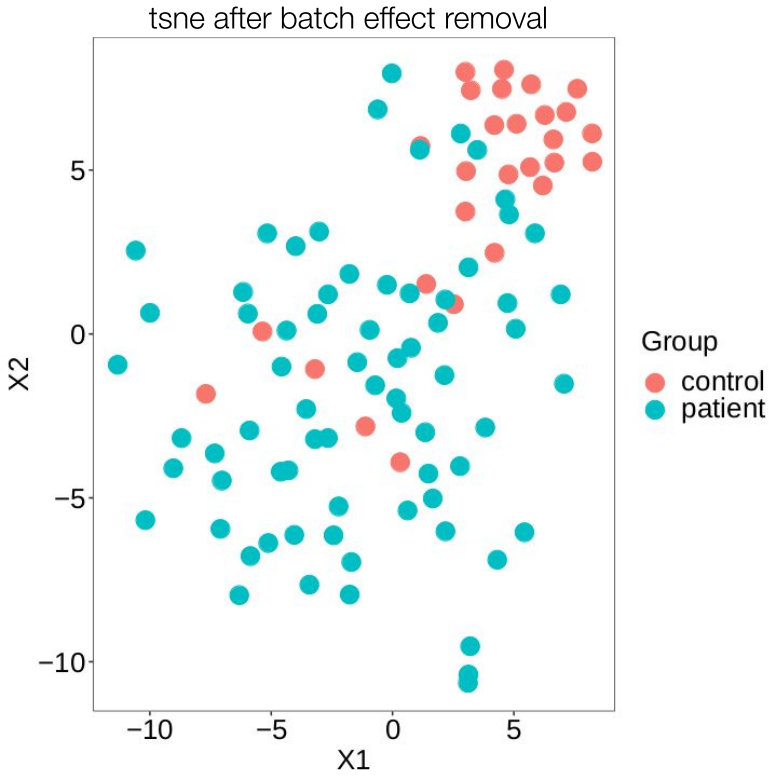


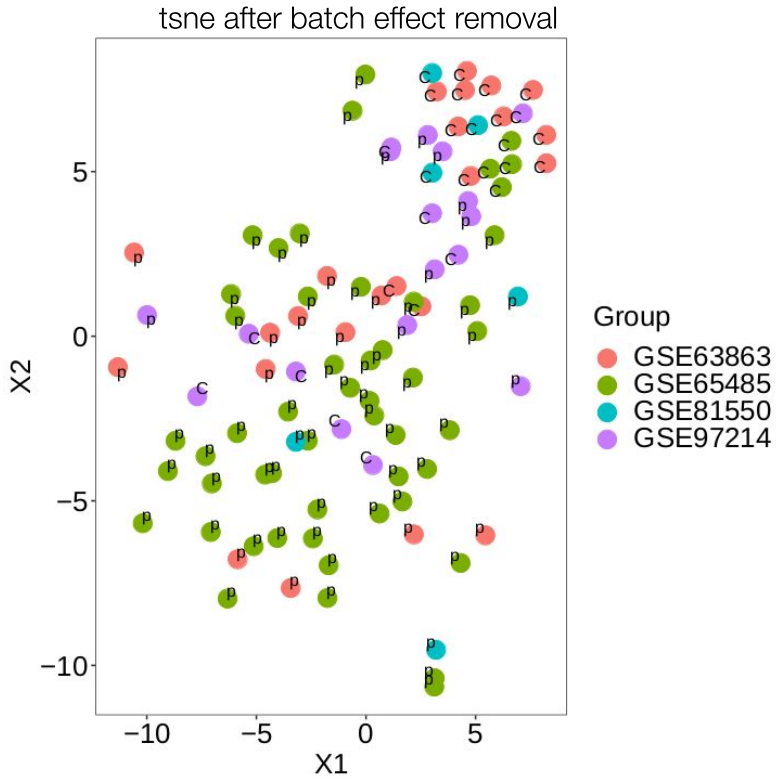
Pipeline results - Hepatocellular carcinoma (liver)





Pipeline results - Hepatocellular carcinoma (liver)

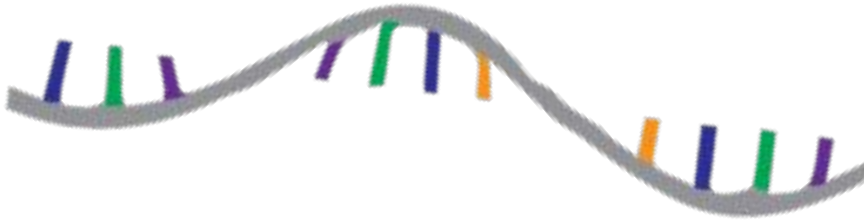




The study effect has been removed and controls and patients are better separated

RNA-seq applications

- Transcript identification and quantification
- Transcript discovery
- Characterization of alternative splicing patterns
- Gene fusion discovery
- Profiling small RNAs and long non-coding RNAs (lncRNAs)



Where to learn more

- Wang et al., 2009. RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet. (https://pubmed.ncbi.nlm.nih.gov/19015660/)

- Ozsolak et al., 2011. RNA sequencing: advances, challenges and opportunities, Nat Rev Genet. (https://pubmed.ncbi.nlm.nih.gov/21191423/)

- Conesa et al., 2016. **A survey of best practices for RNA-seq data analysis,** Genome Biology. (https://pubmed.ncbi.nlm.nih.gov/26813401/)


Thanks!