Statistical analysis of transcriptomics (miRNA) data

Giorgio Melloni

10/27/2022

TIMI

BROAD INSTITUTE

OUTLINE

- Overview of Transcriptomics data
- Input data
- Differentially Expressed Gene analysis (DEGs) using DESeq2
- Data representation and metrics
- Post DEG analysis
  - Multi-gene modeling
  - Pathway analysis

TRANSCRIPTOMICS

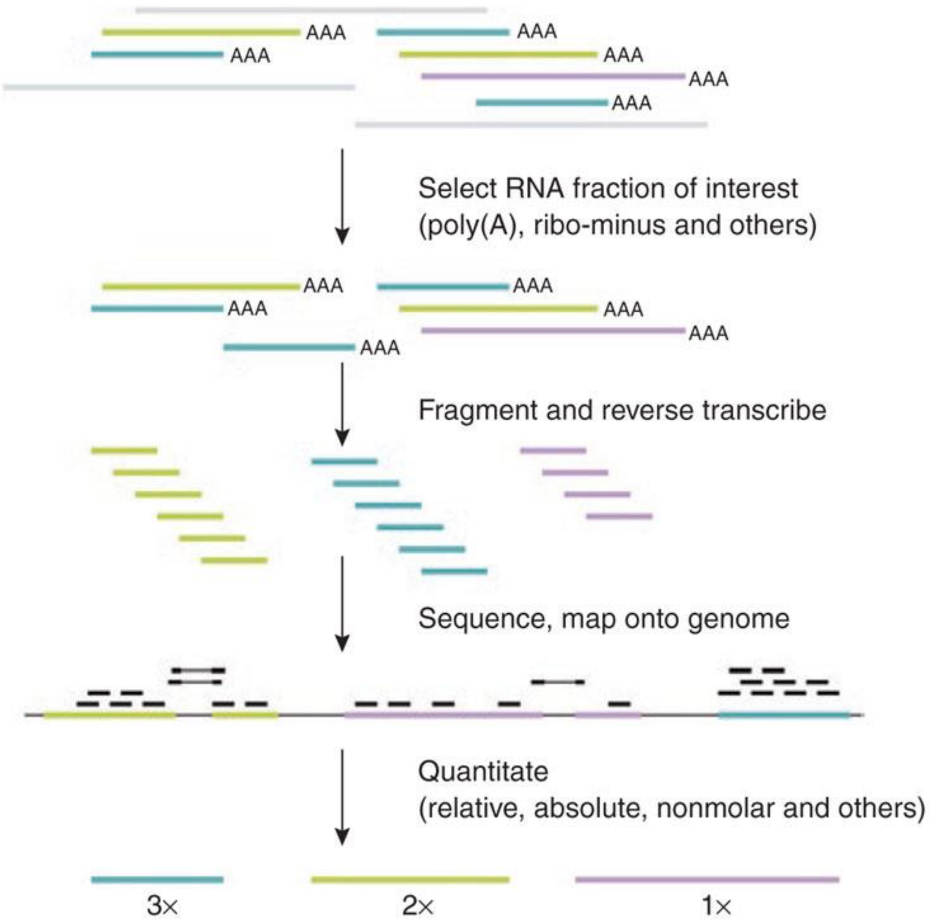
- There are many different types of transcriptomic data that use Next Generation Sequencing technology
    - miRNA, mRNA, ATACseq, Methylation Sequencing etc.

- Each technology and starting material provide different transcriptomic information…

- … but the underlying **analysis model is similar** and it’s based on **count data distributions**

- We’ll use **miRNA** data from HTG Edgeseq technology as an example

TRANSCRIPTOMICS


AAA
Select RNA fraction of interest
(poly(A), ribo-minus and others)
Fragment and reverse transcribe
Sequence, map onto genome
Quantitate
(relative, absolute, nonmolar and others)
3×
2×
1×

Similar to whole exome/genome sequencing, we align short sequences onto a reference genome but using a strand-aware aligner (e.g. STAR)

Aligned reads are then mapped onto a library of known targets (miRNA, mRNA, methylation sites etc.)

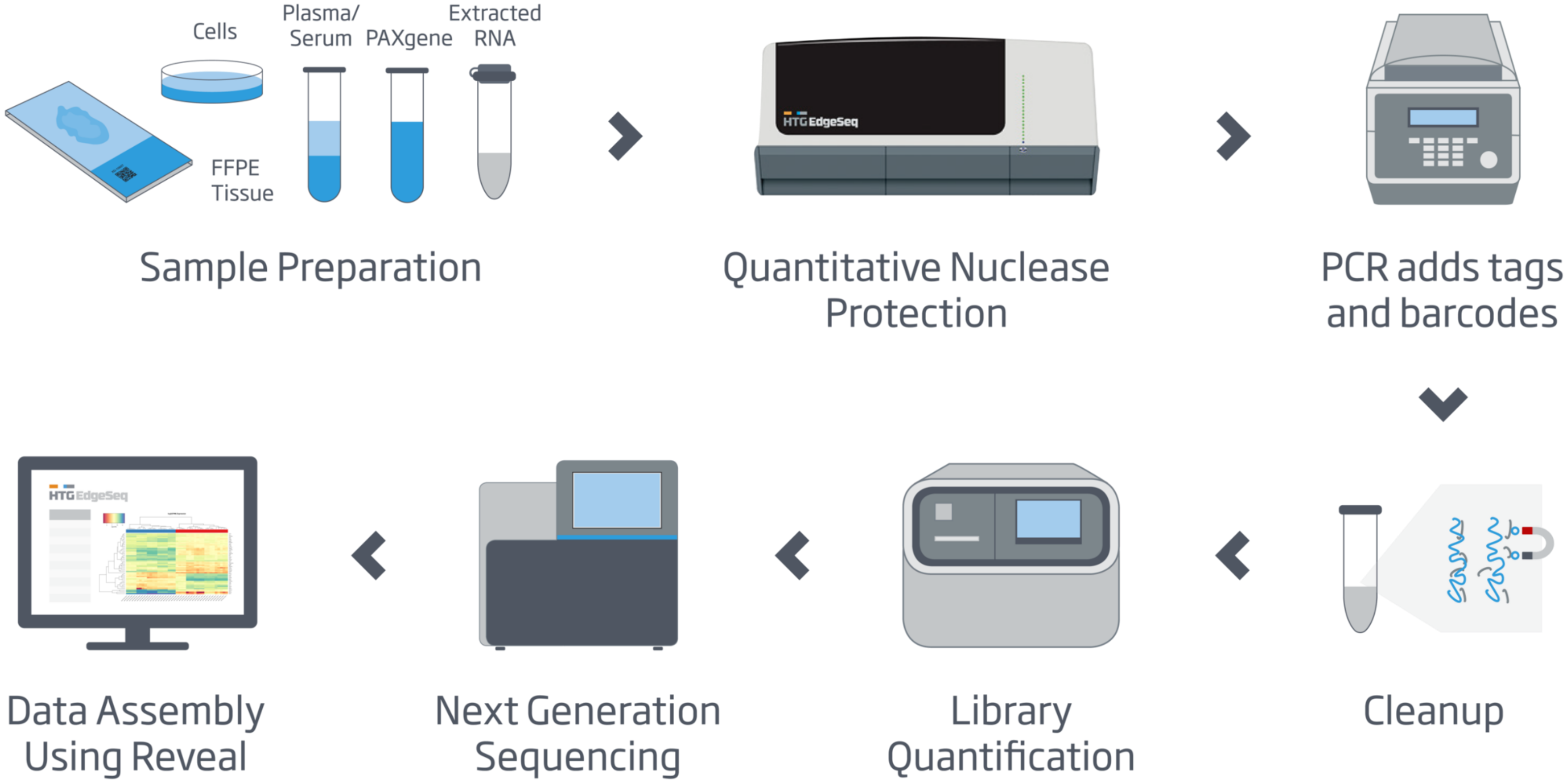
The measure of interest is the abundance of each transcript expressed in number of reads spanning a particular region (**read count**)

In our example, we'll use a library of ~2k miRNA targets

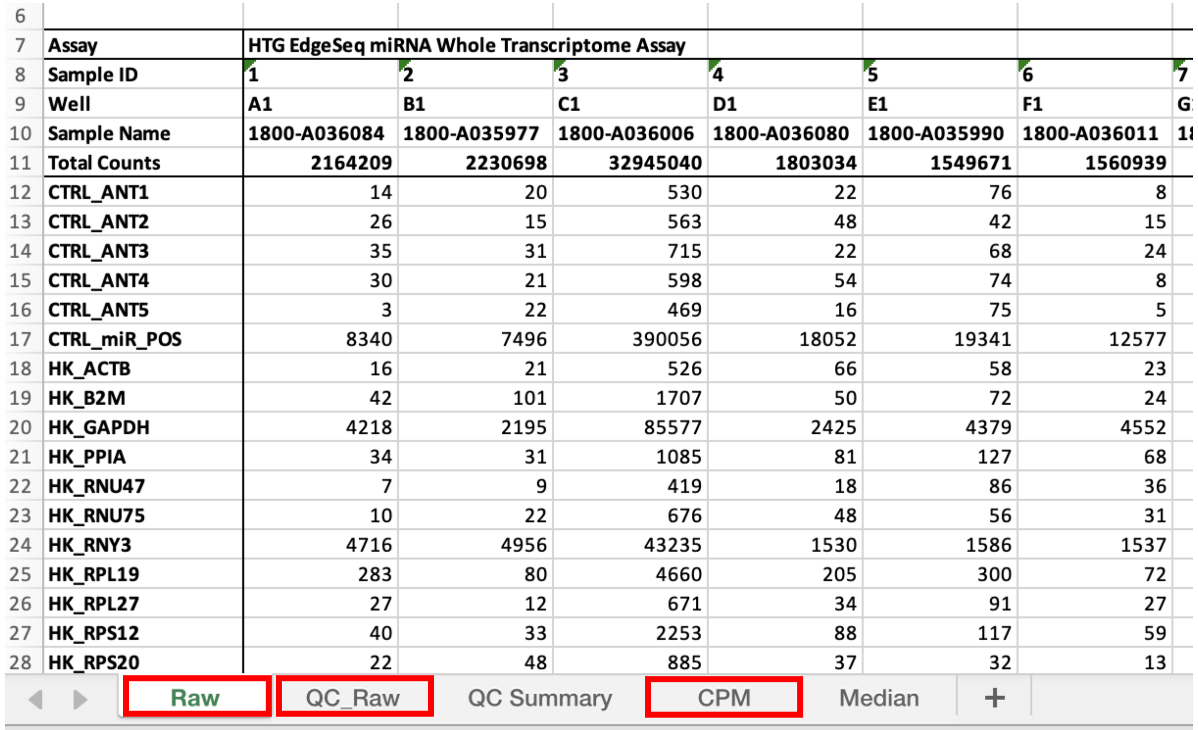
TRANSCRIPTOMICS

HTG EdgeSeq - miRNA Whole Transcriptome Assay

Measure the expression of 2,083 human miRNA transcripts using next generation sequencing (NGS)



INPUT DATA



6								
7	Assay	HTG EdgeSeq miRNA Whole Transcriptome Assay						
8	Sample ID	1	2	3	4	5	6	7
9	Well	A1	B1	C1	D1	E1	F1	G
10	Sample Name	1800-A036084	1800-A035977	1800-A036006	1800-A036080	1800-A035990	1800-A036011	18
11	Total Counts	2164209	2230698	32945040	1803034	1549671	1560939	
12	CTRL_ANT1	14	20	530	22	76	8	
13	CTRL_ANT2	26	15	563	48	42	15	
14	CTRL_ANT3	35	31	715	22	68	24	
15	CTRL_ANT4	30	21	598	54	74	8	
16	CTRL_ANT5	3	22	469	16	75	5	
17	CTRL_miR_POS	8340	7496	390056	18052	19341	12577	
18	HK_ACTB	16	21	526	66	58	23	
19	HK_B2M	42	101	1707	50	72	24	
20	HK_GAPDH	4218	2195	85577	2425	4379	4552	
21	HK_PPIA	34	31	1085	81	127	68	
22	HK_RNU47	7	9	419	18	86	36	
23	HK_RNU75	10	22	676	48	56	31	
24	HK_RNY3	4716	4956	43235	1530	1586	1537	
25	HK_RPL19	283	80	4660	205	300	72	
26	HK_RPL27	27	12	671	34	91	27	
27	HK_RPS12	40	33	2253	88	117	59	
28	HK_RPS20	22	48	885	37	32	13	

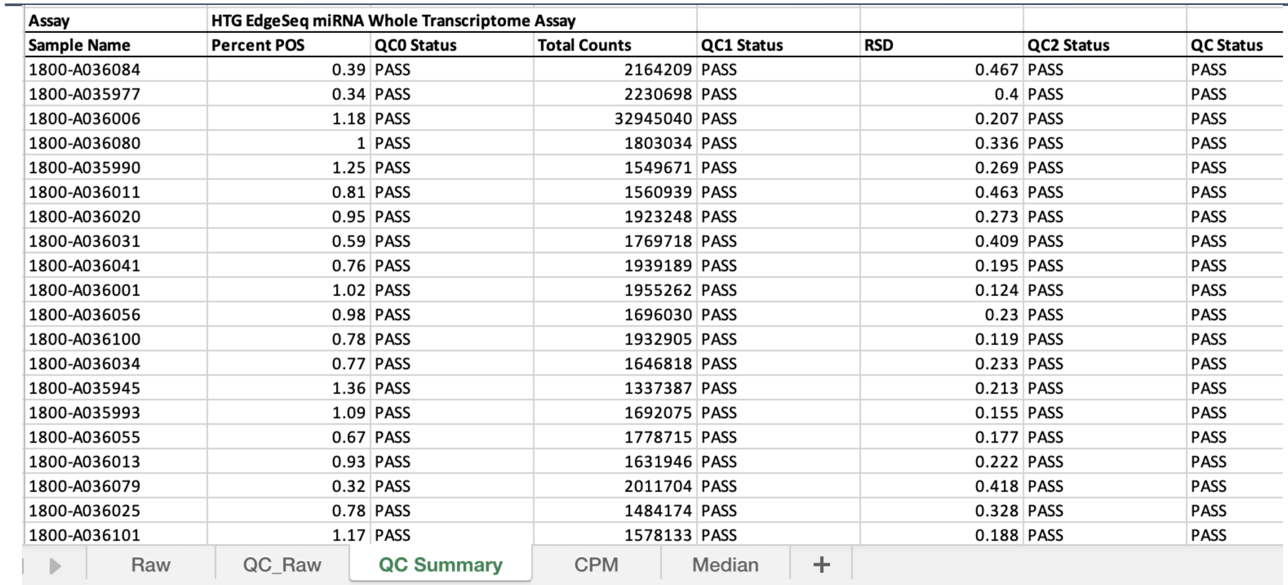
Raw Number of reads per target. This measure can't be directly compared across samples because each sample has different number of total output reads. **It requires normalization**

Same as Raw but excluding samples that did not pass **post-sequencing QC** and without control probes. **<u>This is what we will use for DEGs calling</u>**

Counts Per Million aligned reads is the most popular and simplest normalization procedure. Each count is divided by the total number of reads, multiplied by 1 million

$$\mathrm{CPM}_i = \frac{r_i}{\frac{R}{10^6}} = \frac{r_i}{R} \cdot 10^6$$

INPUT DATA



Assay	HTG EdgeSeq miRNA Whole Transcriptome Assay						
Sample Name	Percent POS	QC0 Status	Total Counts	QC1 Status	RSD	QC2 Status	QC Status
1800-A036084	0.39	PASS	2164209	PASS	0.467	PASS	PASS
1800-A035977	0.34	PASS	2230698	PASS	0.4	PASS	PASS
1800-A036006	1.18	PASS	32945040	PASS	0.207	PASS	PASS
1800-A036080	1	PASS	1803034	PASS	0.336	PASS	PASS
1800-A035990	1.25	PASS	1549671	PASS	0.269	PASS	PASS
1800-A036011	0.81	PASS	1560939	PASS	0.463	PASS	PASS
1800-A036020	0.95	PASS	1923248	PASS	0.273	PASS	PASS
1800-A036031	0.59	PASS	1769718	PASS	0.409	PASS	PASS
1800-A036041	0.76	PASS	1939189	PASS	0.195	PASS	PASS
1800-A036001	1.02	PASS	1955262	PASS	0.124	PASS	PASS
1800-A036056	0.98	PASS	1696030	PASS	0.23	PASS	PASS
1800-A036100	0.78	PASS	1932905	PASS	0.119	PASS	PASS
1800-A036034	0.77	PASS	1646818	PASS	0.233	PASS	PASS
1800-A035945	1.36	PASS	1337387	PASS	0.213	PASS	PASS
1800-A035993	1.09	PASS	1692075	PASS	0.155	PASS	PASS
1800-A036055	0.67	PASS	1778715	PASS	0.177	PASS	PASS
1800-A036013	0.93	PASS	1631946	PASS	0.222	PASS	PASS
1800-A036079	0.32	PASS	2011704	PASS	0.418	PASS	PASS
1800-A036025	0.78	PASS	1484174	PASS	0.328	PASS	PASS
1800-A036101	1.17	PASS	1578133	PASS	0.188	PASS	PASS

Metric	Corresponding Failure Mode	QC Failure by Cutoff
QC0	Insufficient RNA	60% or more reads allocated to POS
QC1	Insufficient Read Depth	100,000 or less
QC2	Insufficient Expression Variability	RSD equal to or lower than 0.082

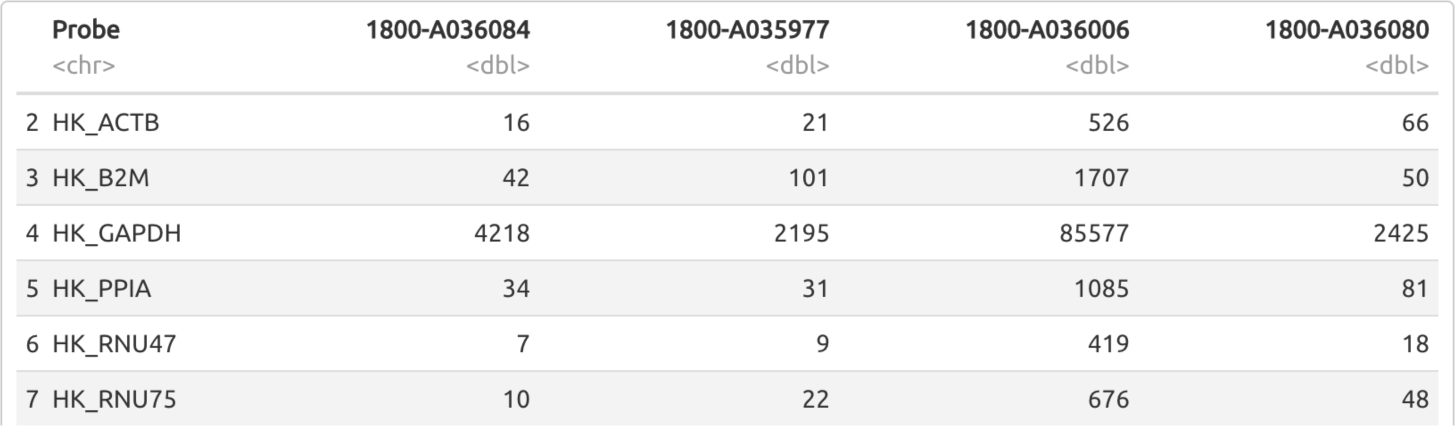
QC summary reports the reason why a sample should be excluded:

- **QC0** – excludes samples with **low RNA content**.
  - If positive control probes (POS) capture most of the reads, there’s not enough material for reliable calls
- **QC1** – excludes samples with **insufficient Read Depth**
  - If too many reads do not align to the reference panel, we’ll lack sensitivity for low expressed targets
  - These failures are typically caused by dilution or library pooling errors
- **QC2** – excludes samples with **low expression variability** across the targets
  - If the Relative Standard Deviation (RSD) is too low, the data do not reliably represent the true variability of a biological specimen
  - RSD is calculated as: SD(log2(counts+2)) / mean(log2(counts+2))
  - These failures are generally caused by a defective S1 nuclease activity

DIFFERENTIAL EXPRESSION

There are many tools for Differential Expression Analysis, edgeR, limma, DESeq2, NOISeq and others

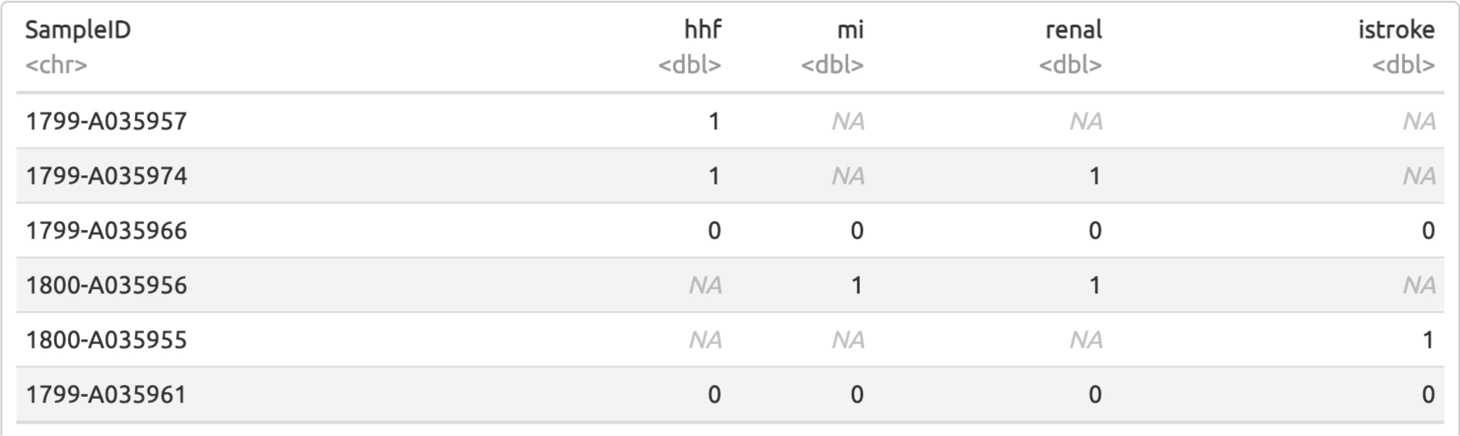
DESeq2 is probably the most popular and it comes as an R package



	Probe	1800-A036084	1800-A035977	1800-A036006	1800-A036080
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
2	HK_ACTB	16	21	526	66
3	HK_B2M	42	101	1707	50
4	HK_GAPDH	4218	2195	85577	2425
5	HK_PPIA	34	31	1085	81
6	HK_RNU47	7	9	419	18
7	HK_RNU75	10	22	676	48

QC_raw data is our input matrix

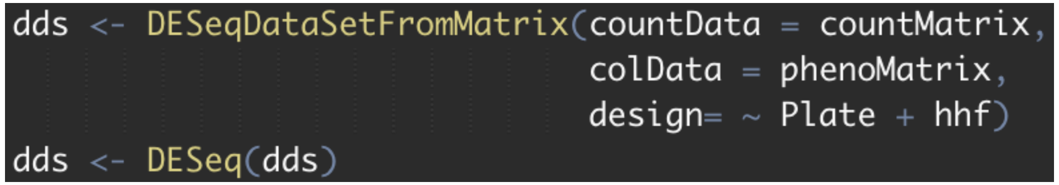
The normalization step is run by DESeq2



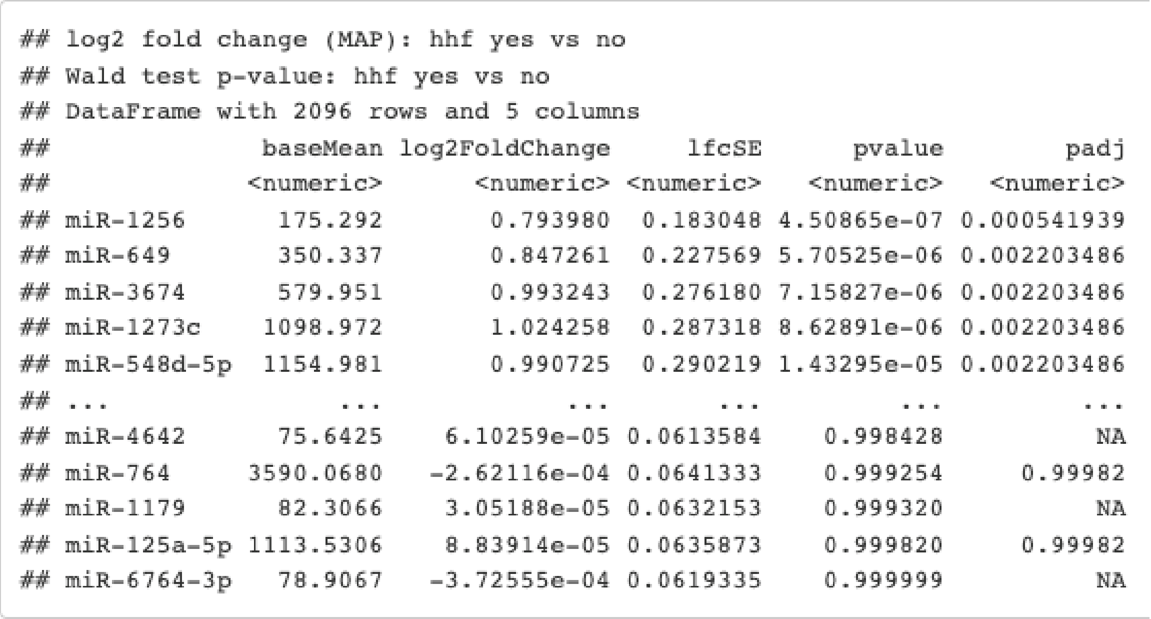
SampleID	hhf	mi	renal	istroke
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1799-A035957	1	NA	NA	NA
1799-A035974	1	NA	1	NA
1799-A035966	0	0	0	0
1800-A035956	NA	1	1	NA
1800-A035955	NA	NA	NA	1
1799-A035961	0	0	0	0

Phenotype matrix reports absence or presence of the disease of interest or treated vs untreated. Multiple levels are also accepted

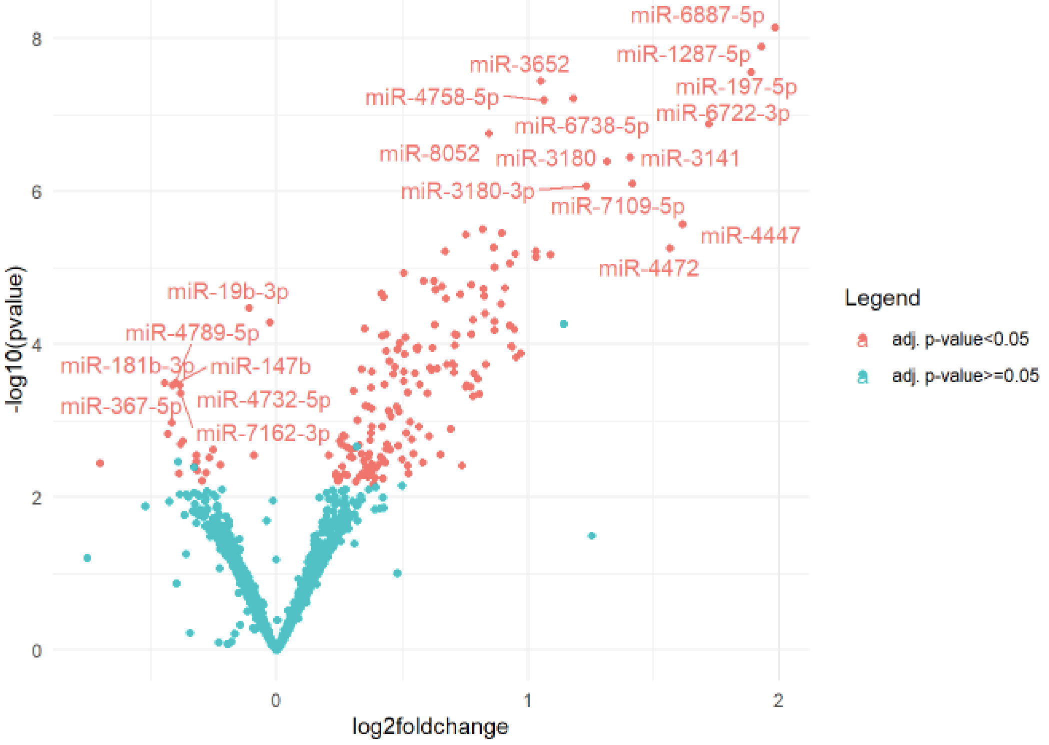
DEGs



```
dds <- DESeqDataSetFromMatrix(countData = countMatrix,
                              colData = phenoMatrix,
                              design= ~ Plate + hhf)
dds <- DESeq(dds)
```



```
## log2 fold change (MAP): hhf yes vs no
## Wald test p-value: hhf yes vs no
## DataFrame with 2096 rows and 5 columns
##              baseMean log2FoldChange     lfcSE      pvalue        padj
##             <numeric>      <numeric> <numeric>   <numeric>   <numeric>
## miR-1256      175.292       0.793980  0.183048 4.50865e-07 0.000541939
## miR-649       350.337       0.847261  0.227569 5.70525e-06 0.002203486
## miR-3674      579.951       0.993243  0.276180 7.15827e-06 0.002203486
## miR-1273c    1098.972       1.024258  0.287318 8.62891e-06 0.002203486
## miR-548d-5p  1154.981       0.990725  0.290219 1.43295e-05 0.002203486
## ...               ...            ...       ...         ...         ...
## miR-4642      75.6425    6.10259e-05 0.0613584    0.998428          NA
## miR-764     3590.0680   -2.62116e-04 0.0641333    0.999254     0.99982
## miR-1179      82.3066    3.05188e-05 0.0632153    0.999320          NA
## miR-125a-5p 1113.5306    8.83914e-05 0.0635873    0.999820     0.99982
## miR-6764-3p   78.9067   -3.72555e-04 0.0619335    0.999999          NA
```

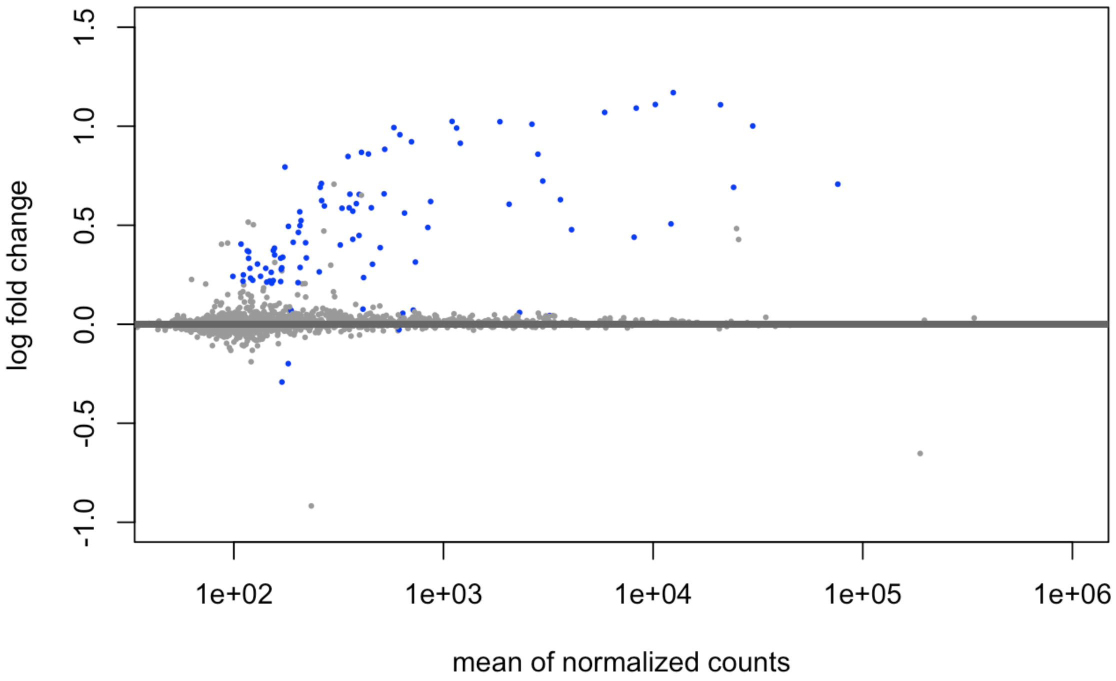


A negative binomial model is fitted for every marker comparing the mean abundance of cases vs controls

If the data come from different plates, we can add adjustments to our design

The strength of the association is expressed in log2 Fold Change and the p-value is adjusted via FDR

VISUALIZATION

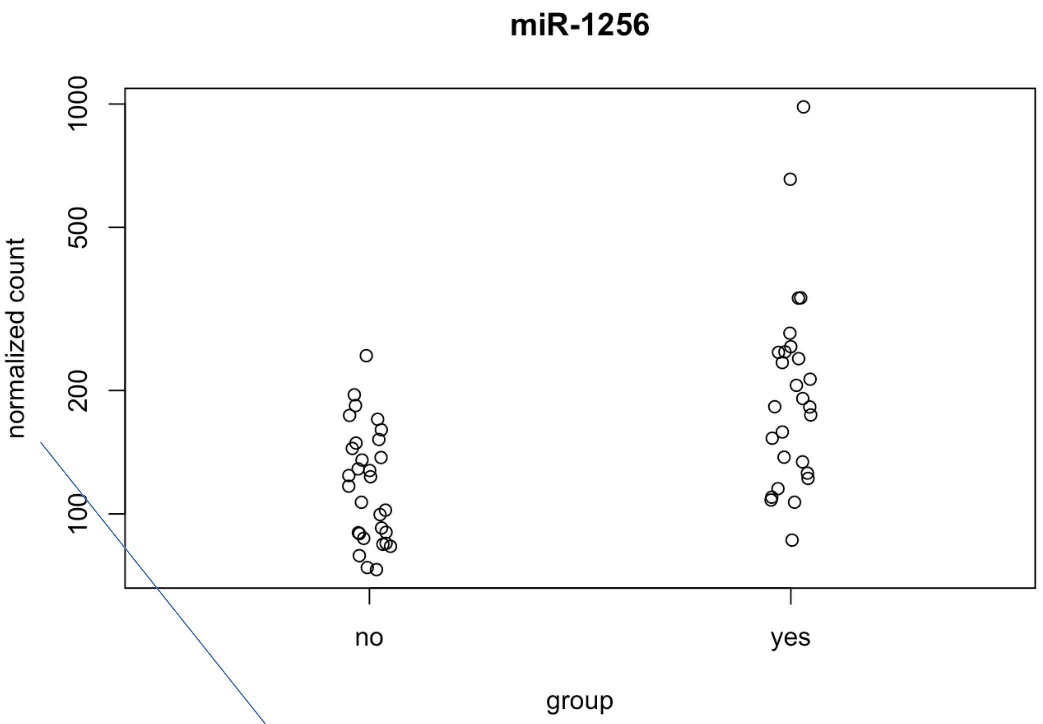


The **MA-plot** is a typical representation of transcriptomic data

It shows the log2 fold change by the mean of normalized counts and blue dots are targets with an adjusted p-value < 0.05

While the statistics of DEGs are based directly on the counts, **visualization and post-DEG analyses require normalized counts by sample and target**

VISUALIZATION



Counts have no absolute meaning. We always talk in terms of differential expression. More info on transcriptomics units at https://luisvalesilva.com/datasimple/rna-seq_units.html

- **Each sample has different total read counts**
  - ➢ Normalization by sample
- **Each target (miRNA in this case) has a different length. The larger the gene, the higher the count.**
  - ➢ Normalization by target

CPM takes care of the first point but not the second

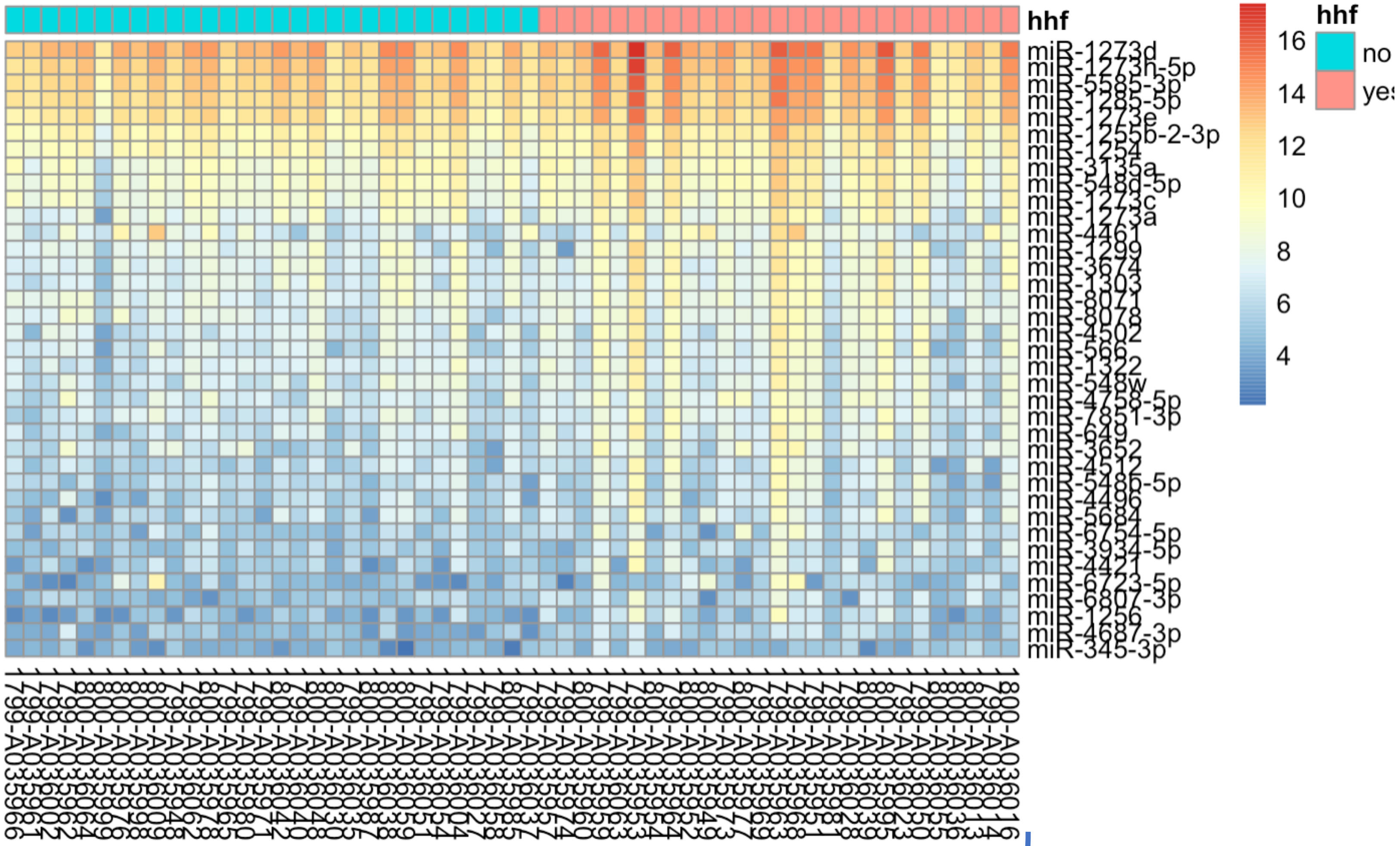
RPKM, FPKM, TPM are all different popular choices to export data and compare it directly across different datasets

DESeq2 has several internal normalizations that do not require feature length mapping and can be used for downstream analysis and normalization

Variance Stabilizing Transformation (VST, Tibshirani 1988; Huber et al. 2003; Anders and Huber 2010) is what the authors suggest

VISUALIZATION

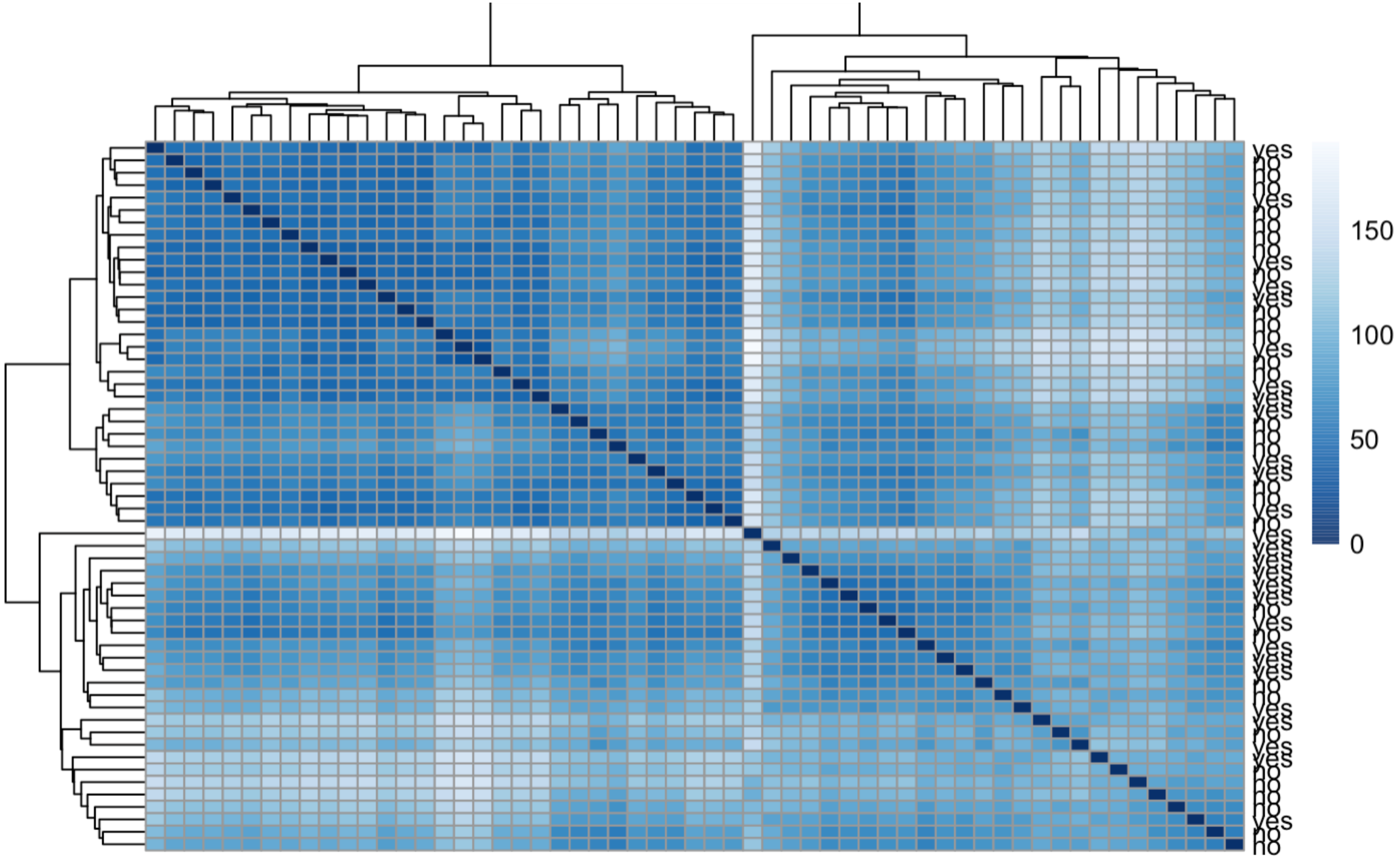
Since the advent of microarrays, heatmaps have become a classic way of visualizing expression data



Significant DEGs, ordered by mean counts

Samples are ordered by HHF status

MORE VISUALIZATION



Correlation Analysis between samples coupled with hierarchical clustering allows discovery of similarities among samples based on their transcriptional profile

A similar analysis can be run by miRNA

POST DEG

Why don’t we just fit a regular Cox model?

- □ Genomic data generally has too many targets (P) compared to samples (N)
  - In our example taken from SAVOR data, ~2000 miRNA are tested on ~100 samples
  - A good rule of thumb is to run a model that has less than log(N) predictors

…Then Why don’t we run a Cox LASSO?

- □ Even machine learning (ML) techniques will suffer from over fitting when N << P
- □ Discovery power with count data is maximized using Poisson-like distributions (like the negative binomial)

So how do I use my follow-up information?

Similar to what we do with proteomics data from O-link

1. Use count data to run a DEG analysis
2. Select targets that have an adjusted p-value < 0.05
3. Transform the count data with VST (or other normalizations)
4. Validate the selection with a longitudinal multivariable model for further feature selection (e.g. GMB, LASSO, Elastic Net and other ML techniques can support Cox modeling)

POST DEG

Pathway analysis is the most typical post-DEG step that can give insights to the following questions

1. What are the biological processes, cellular locations and molecular functions that are particularly over- or under-represented in my set of miRNAs/genes?

2. What are the pathways that are significantly impacted in the outcome of interest?

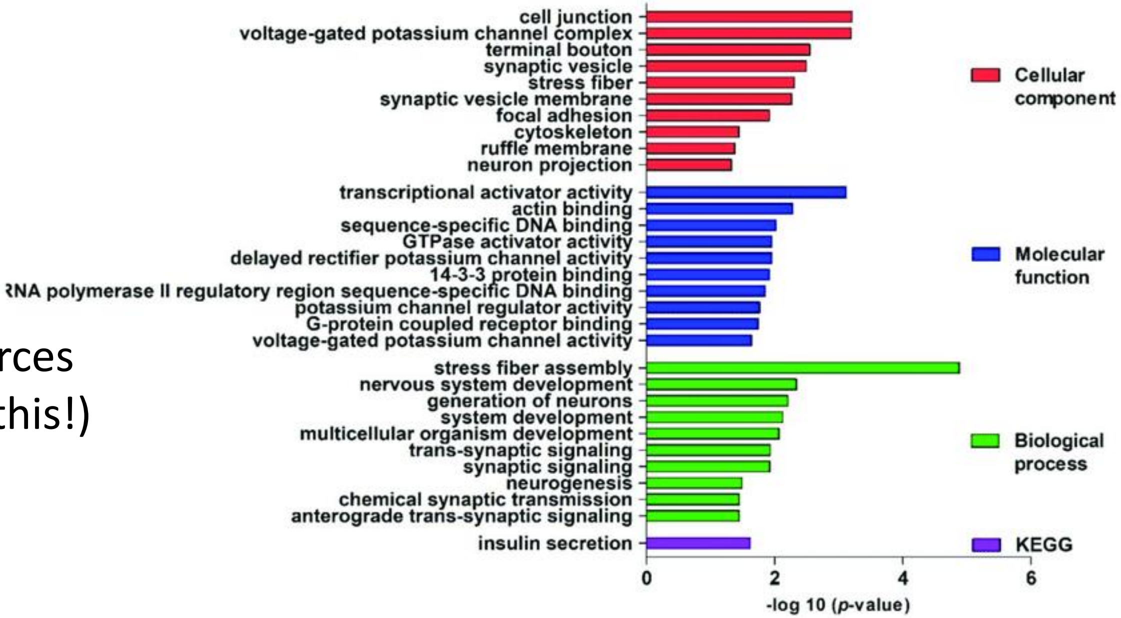
POST DEG

1. What are the biological processes, cellular locations and molecular functions that are particularly over- or under-represented in my set of miRNAs/genes?

Here the idea is to compare your list of DEGs to a database of Gene Ontologies (GO) whose choice largely depend on the transcriptomic data analyzed (miRNA targets, transcription factors, tissue specific gene expression etc.)

A few popular examples include:

- **KEGG** - https://www.genome.jp/kegg/
- **REACTOME** – https://reactome.org/
- **EnrichR** - https://maayanlab.cloud/Enrichr/
  - integrates several databases (including the two above)
- **Cytoscape** - https://cytoscape.org/
  - great for network visualization, allows plugins from other sources
- **Ingenuity** - from Qiagen, proprietary software (ask me or Fred for this!)



Specifically for miRNA and miRNA-gene interactions

- **miRDB** - https://mirdb.org/
- **TargetScan** - https://www.targetscan.org/
- **miRbase** - https://mirbase.org/

POST DEG

2. What are the pathways that are significantly impacted in the outcome of interest?

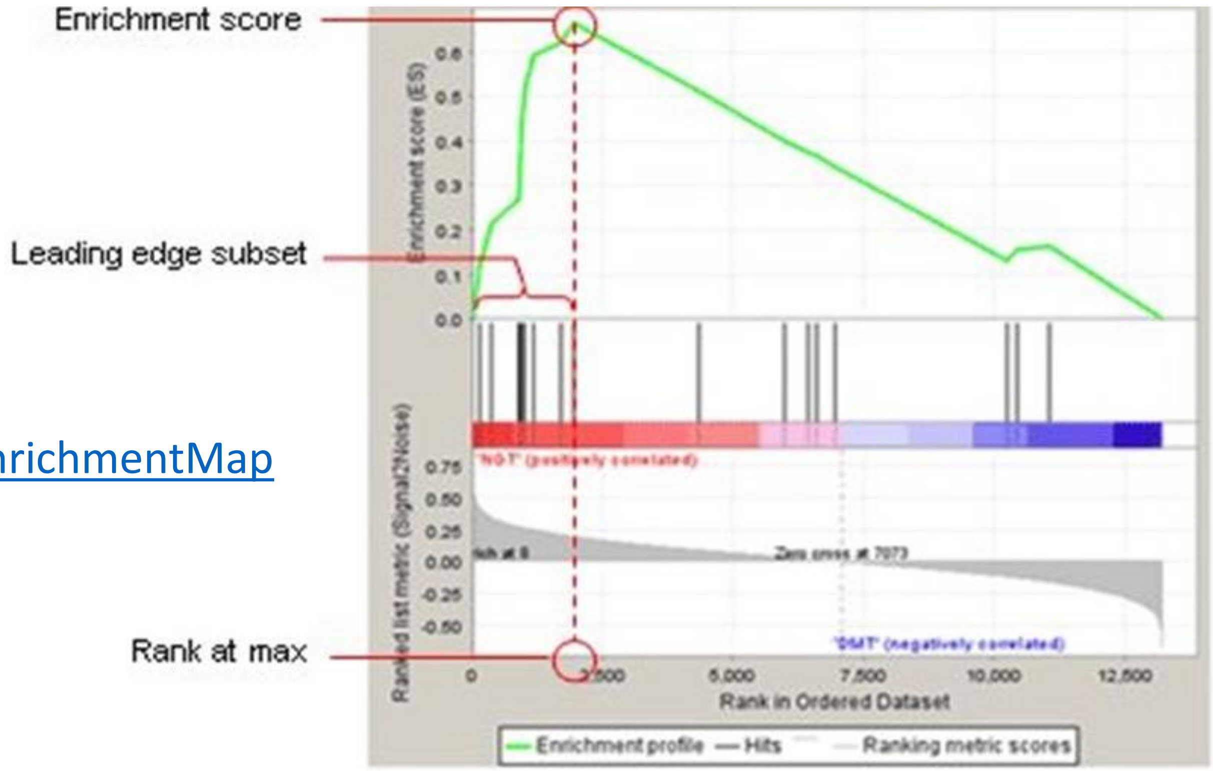
Gene Set Enrichment Analysis is a computational method that determines whether an a priori defined set of genes shows statistically significant concordant differences between two biological states (e.g. phenotypes). The main difference with a GO analysis is that **ranking and direction of the effect of each DEG are taken into consideration**

Popular choices are:

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

G:Profiler https://biit.cs.ut.ee/gprofiler/gost

EnrichmentMap - http://www.baderlab.org/Software/EnrichmentMap



Enrichment score

Leading edge subset

Rank at max

THANK YOU!

Questions?