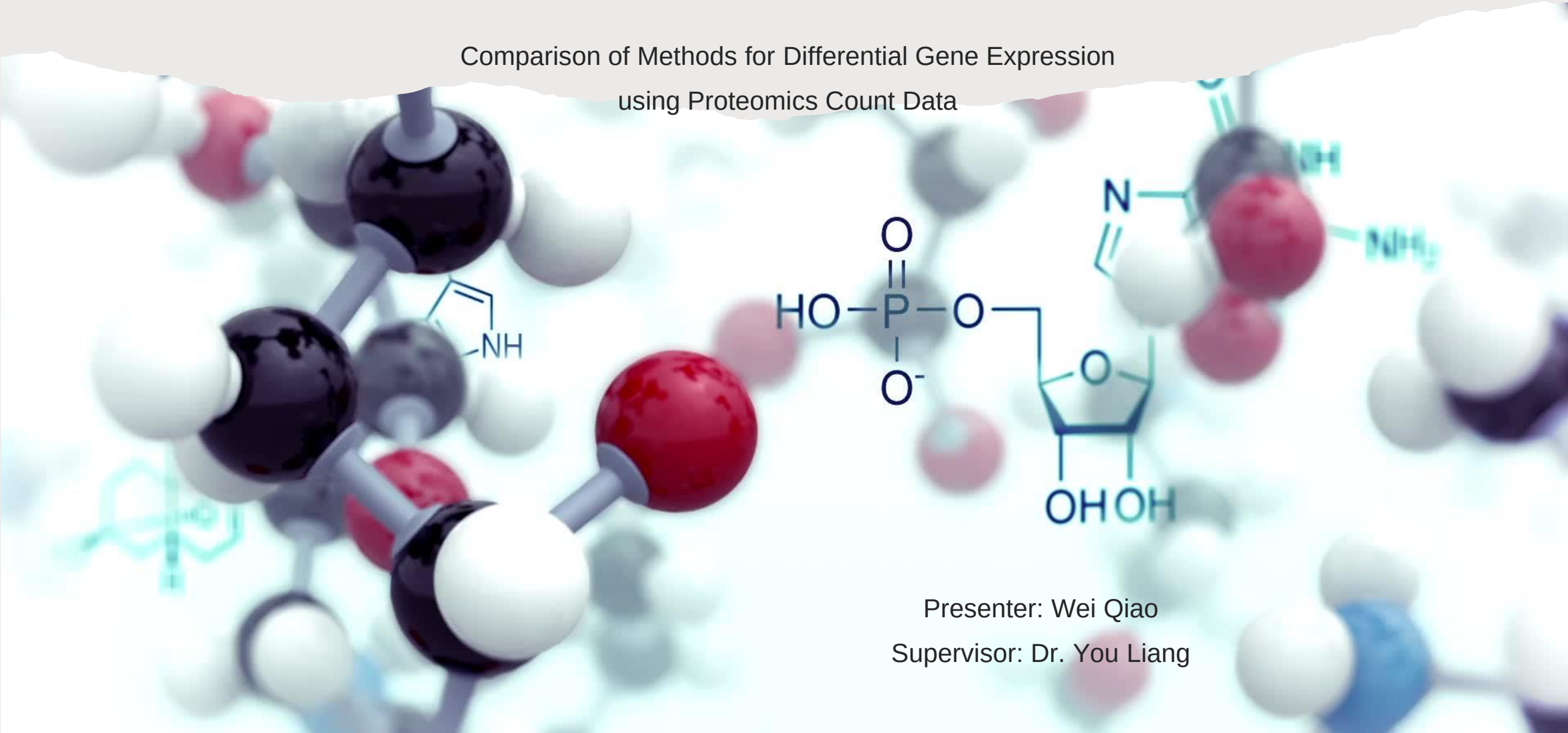
Covid Positive vs ICU Control

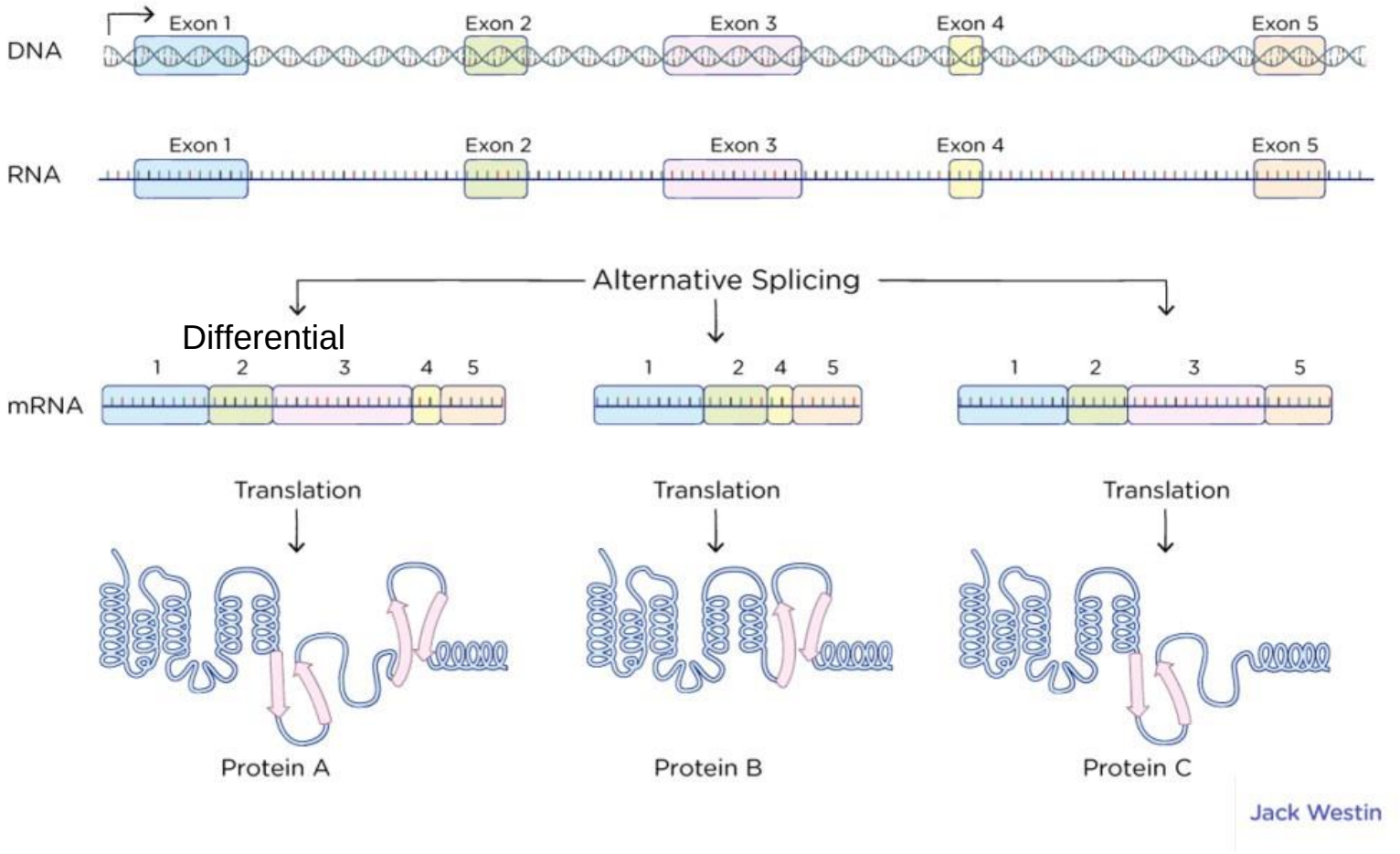
Comparison of Methods for Differential Gene Expression using Proteomics Count Data



Presenter: Wei Qiao

Supervisor: Dr. You Liang

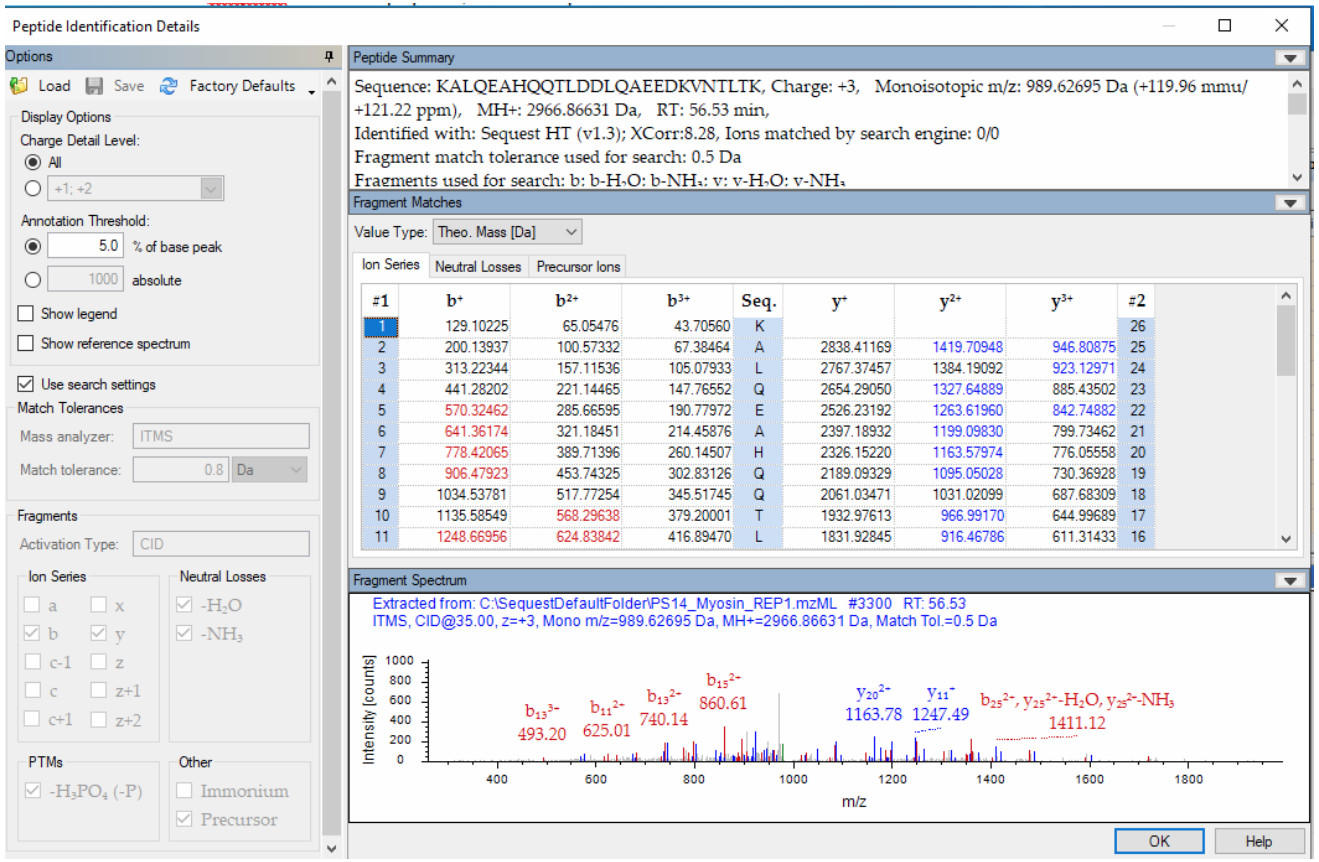
Transcription and Splicing



- https://jackwestin.com/resources/mcat-content/transcription/mrna-processing-in-eukaryotes-introns-exons

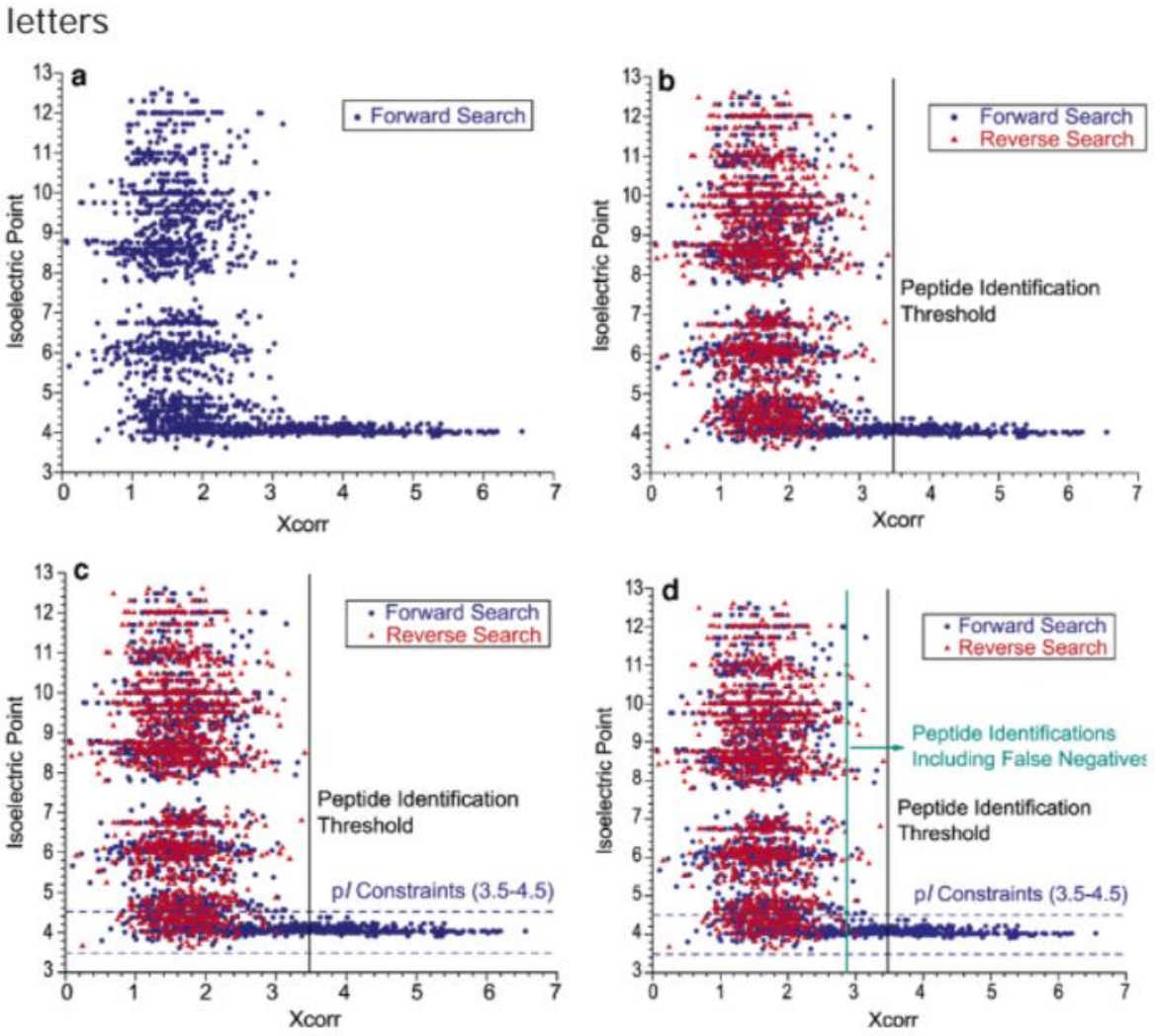
Process of Data Acquiring

1. Fractionation of blood
2. Tryptic digestion
3. Sample clean up on a C18 zip tip
4. Applied to a high pressure liquid chromatography (HPLC) C18 reverse phase column
5. Distinct peptides eluted sequentially over a gradient of increasingly organic solvent
6. Nano spray
   High voltage and heat applied to the particles to aerosolise them
7. Applied to LC-MSMS
8. Searched against a library and analyzed



Example spectra matching in SEQUEST

Decoy library searching



Most proteins have structural and evolutionary constraints on their amino acid sequences making them much more likely to be partially homologous.

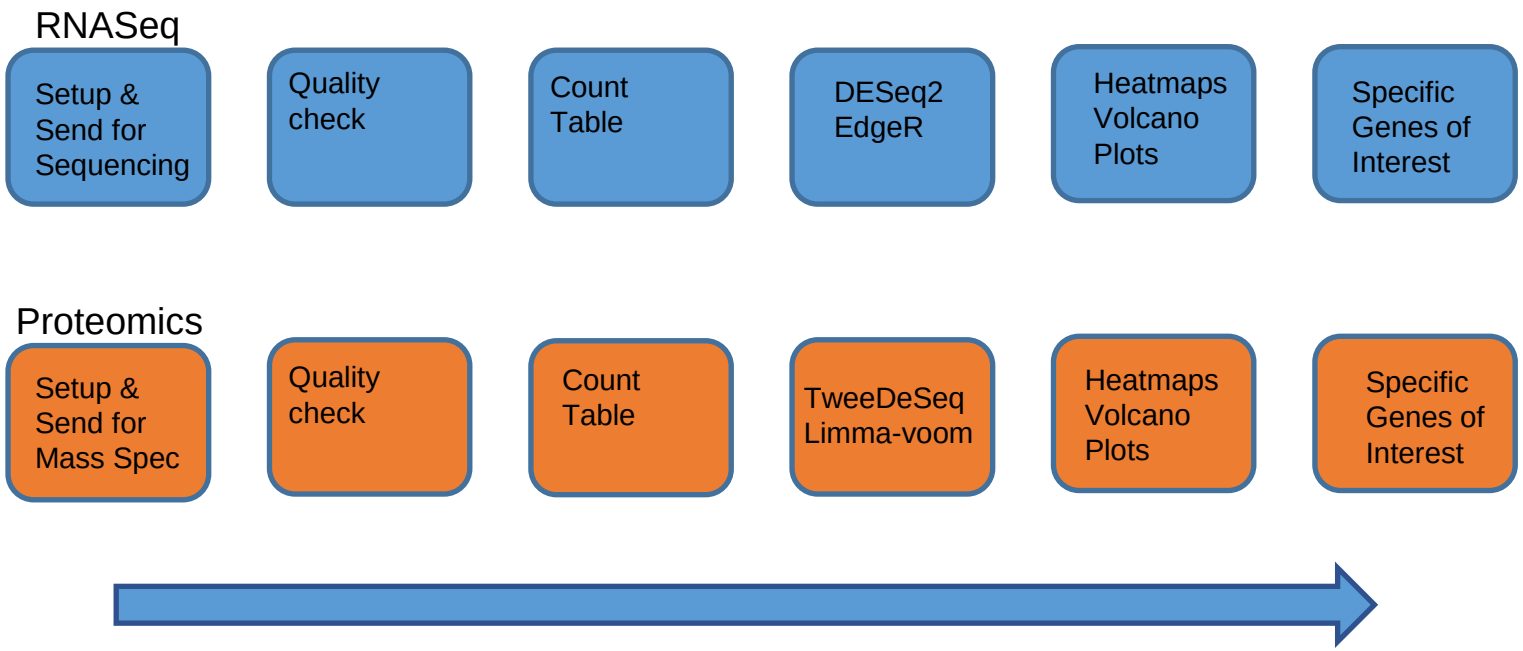
Samples were prepared from rat testis and digested with trypsin and searched against the rat library in forward and reverse direction (decoy library searching)

Samples were separated in a pH gradient gel followed by liquid chromatography tandem mass spectrometry. The “true” peptides are known to be between pH3.5 and 4.5

(Cargile et al., 2004)

RNA-seq vs Proteomics Workflow

Proteomics is a **high throughput** way to measure **all proteins in a cell** or tissue by using **mass spectrometry**.



https://www.youtube.com/watch?v=JPpKL1uzE0I&t=726s

Current Analysis Focus

1. Test the difference between two groups at a time

(Covid Positive vs ICU Ctrl)

2. Time Point Analysis

(Three time points for Covid Positive vs ICU Ctrl)

Forms of Two Obtained Data Table

SampleCode

Gene Symbol

	201	202	203	204	205	206	210	211	212	213	214	215	7D2	224	225	226	18D1	244	245	246	253	254	29D1	29D2	257	258
gene 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gene 2	0	0	0	0	0	18	0	11	0	12	20	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0
gene 3	14	10	0	0	10	0	0	16	10	0	0	0	12	23	0	0	12	11	0	12	0	17	10	0	0	0
gene 4	24	13	35	0	0	0	0	0	0	0	12	12	0	18	15	37	19	10	0	0	38	12	0	0	60	29
gene 5	0	12	17	0	27	0	18	0	0	13	0	0	0	0	11	0	15	0	0	0	23	32	0	0	0	0
gene 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	259	260	261	262	263	264	265	40D1	267	268	269	270	271	272	52D1	52D10	52D5	53D0	53D10	53D5	54D0	54D10	54D5
gene 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gene 2	12	0	0	0	17	10	0	0	10	10	14	24	42	0	0	0	0	0	0	0	0	0	0
gene 3	12	0	0	0	0	0	0	29	0	0	0	0	16	0	12	0	0	0	11	13	0	0	0
gene 4	49	18	30	23	107	31	70	0	97	71	154	191	206	30	0	0	0	0	0	0	0	12	0
gene 5	11	10	0	0	0	0	0	0	0	16	11	13	10	0	0	18	0	0	0	22	0	0	16
gene 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	56D1	56D3	56D7	59D0	59D10	59D5	207	208	209	218	219	220	221	222	223	227	228	229	230	231	232	233	234	235	273
gene 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0
gene 2	0	0	0	0	0	0	25	58	57	0	18	0	50	32	12	52	77	59	0	47	22	10	14	12	0
gene 3	0	0	0	11	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gene 4	0	0	0	15	30	0	62	16	106	0	71	0	137	56	0	41	76	36	10	20	0	303	502	218	142
gene 5	10	12	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	20
gene 6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- Count data table
  - How many times an ms/ms spectra was matched
  - Gene symbol

Treatments and Controls

Sample Code

Description: df [103 x 3]

	ProgramID <fctr>	TimePointOrdinal <fctr>	Pairs <fctr>
201	C	T1	C.T1
202	C	T2	C.T2
203	C	T3	C.T3
204	C	T1	C.T1
205	C	T2	C.T2
206	C	T3	C.T3
210	C	T1	C.T1
211	C	T2	C.T2
212	C	T3	C.T3
213	C	T1	C.T1

1-10 of 103 rows Previous 1 2 3 4 5 6 ... 11 Next

- Intensity data table
  - format of experiment is for discovery – we are looking to identify and characterize as much of the sample as possible)

Contingency Table Analysis of Count Data

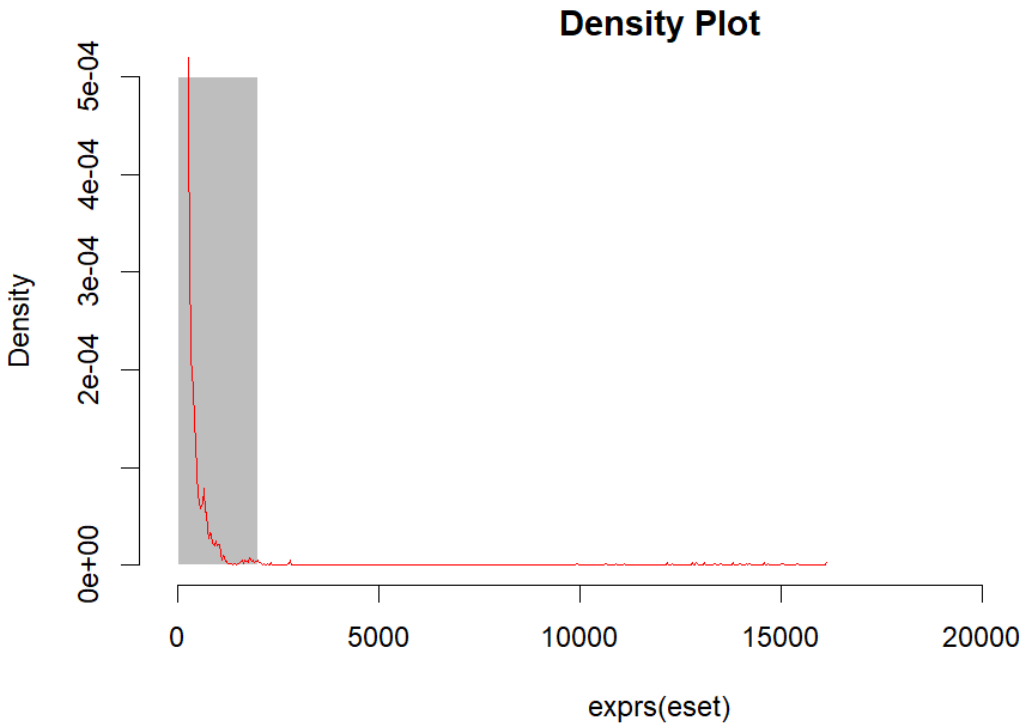
- Ultimately, these two tables will be built into different contingency tables depending on the contrast.
- Based on the right table, the **response variable** is the counts. The genes are **independent observations** from each independent samples.
- Two levels from the right table.

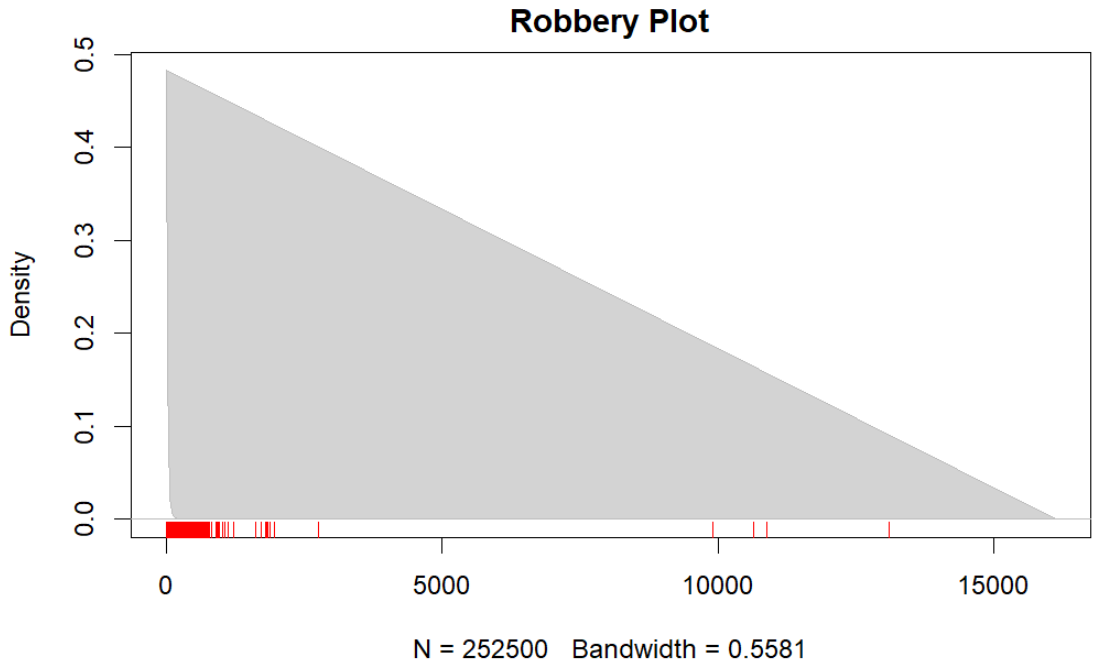
ProgramID	Covid				Icu			
Sample Code	**201**	**202**	**205**	**…**	**305**	**306**	**403**	**…**
gene1	20	10	1	…	33	0	20	…
gene2	100	131	23	…	5	12	51	…
gene3	55	132	34	…	67	34	33	…
gene4	56	133	6	…	77	55	51	…
gene5	57	134	0	…	0	0	33	…
….	….	….	….	….	….	….	….	….

GeneSymbols

Distribution

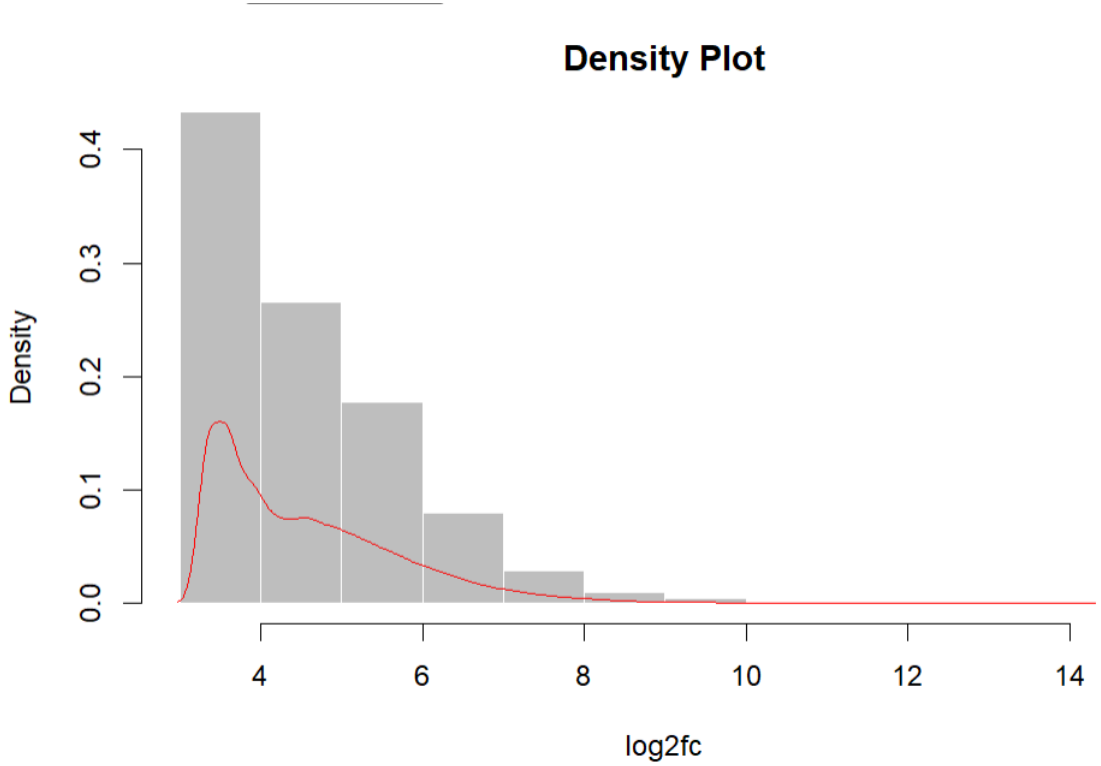
- Zero-inflated: Heavily skewed on the left (around 0)
- Long Right Tail
- Overdispersion: variance > mean

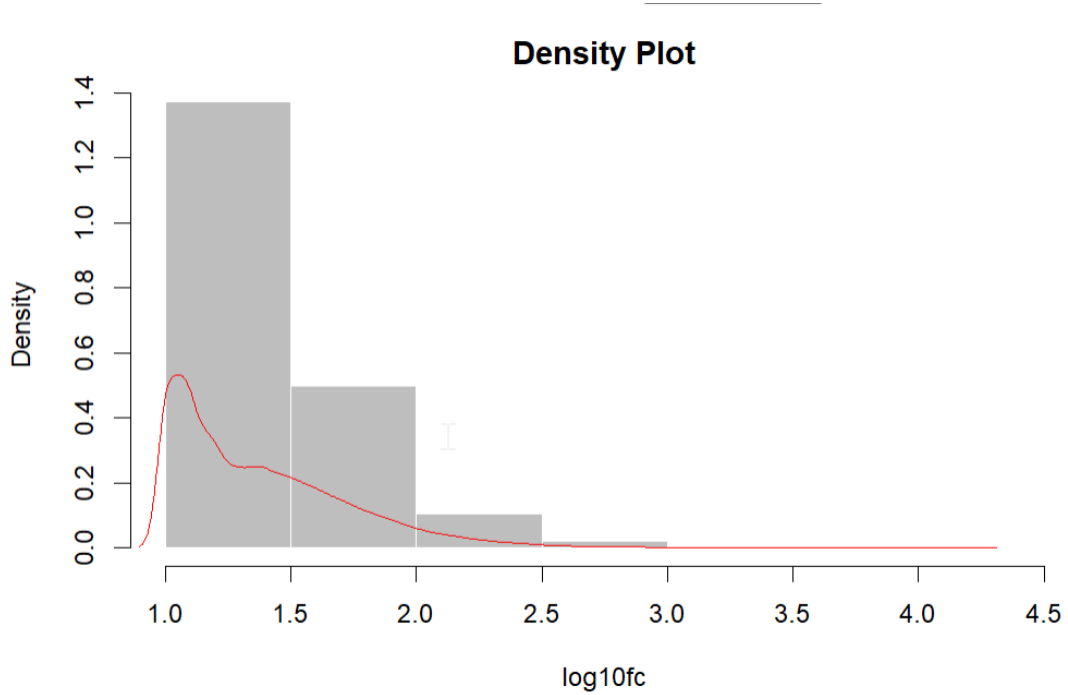




Distribution

- The Distribution after log2 or log10 transformation is still not normal.

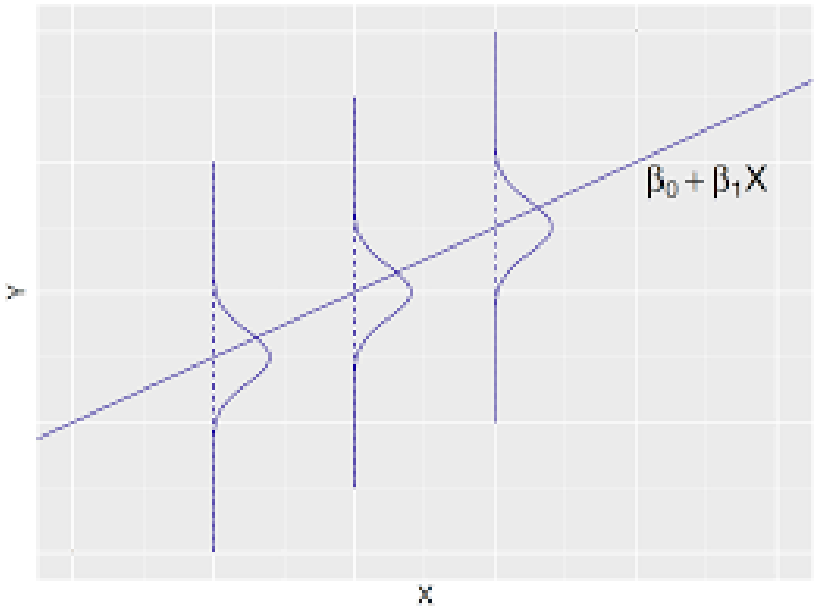




Linear Model ?

Problem is the random component of Y is not normally distributed.

$$Y = f(X) + \epsilon$$



$$1 \cdot y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

$$\Longrightarrow \quad Y = g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

GLMs for Count Data

- **Poisson** could not be a good model because of the overdispersion.

$$y_i \sim Pois(\lambda_i) \quad and \quad E[y_i] = Var[y_i] = \lambda_i$$

- **Negative Binomial** add a second layer of variability by allowing μ_i itself to be a random variable.

$$y_i|\lambda_i \sim Pois(\lambda_i) \quad and \quad \lambda_i \sim G(\mu_i, \psi)$$

$$Then \quad E[y_i] = \mu_i \quad and \quad Var[y_i] = \mu_i + \psi\mu_i^2$$

- **Tweedie EDMs** are distributions that generalize many of the EDBs

$$y_i \sim ED(\mu, \sigma^2, \xi)$$

$$Then \quad E[y_i] = \mu_i \quad and \quad Var[y_i] = \sigma^2\mu_i^{\xi}$$

Multiple Testing

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

- β_1 - Mean in group 1
- β_2 - Mean in group 2
- β_3 - Mean in group 3
- Tests:
  - $\beta_2 - \beta_1 = 0$
  - $\beta_3 - \beta_1 = 0$
  - $\beta_3 - \beta_2 = 0$

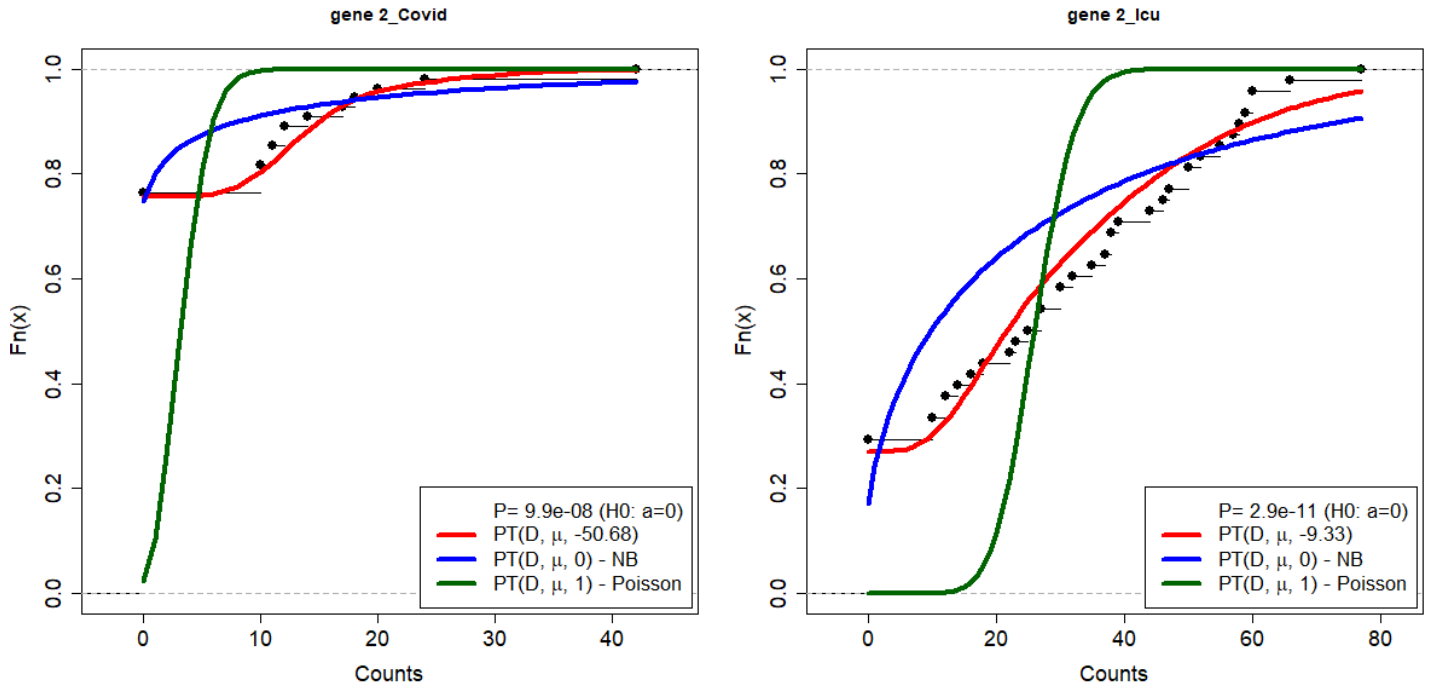
P values
Vs P Adj

The P-value indicates the probability that the observed difference of genes between groups.

P-value adjustments reduce the chance of making type I errors.

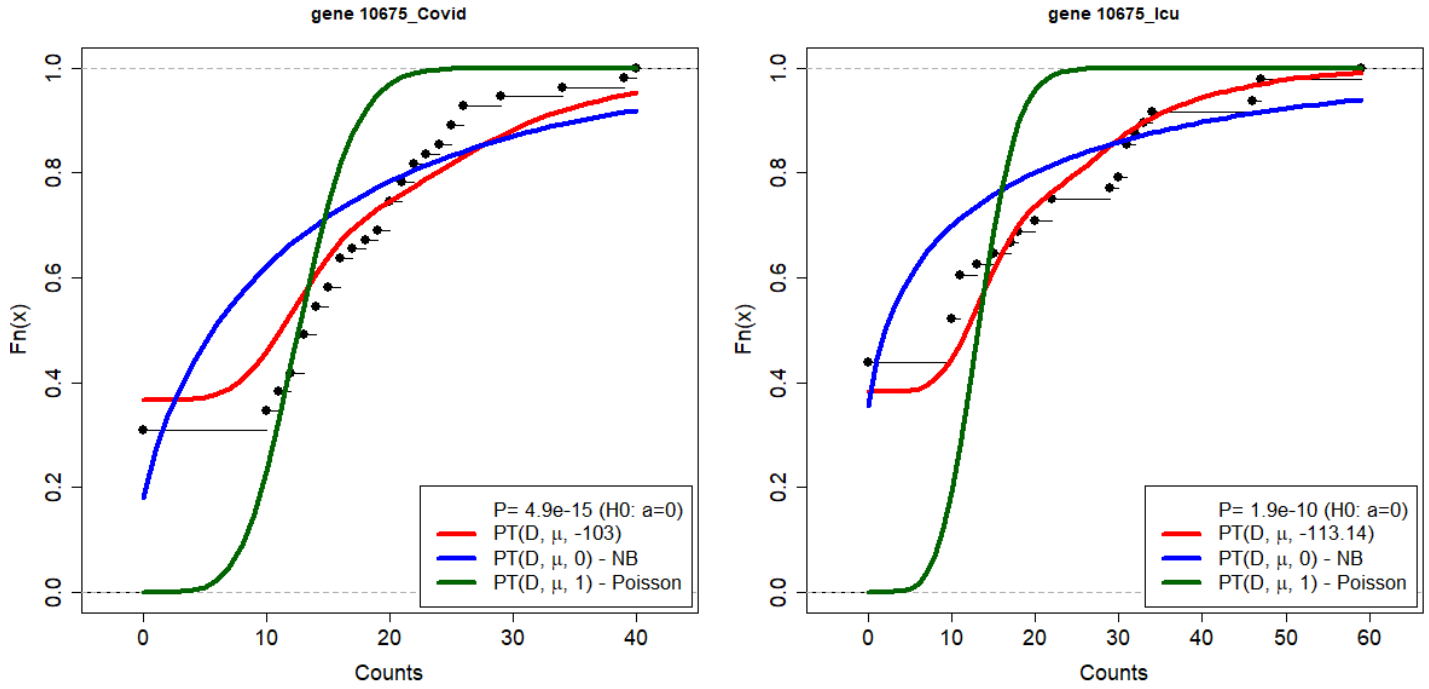
Poisson vs NG vs Tweedie

- Tweedie wins the other two models when **the sample size is small**.



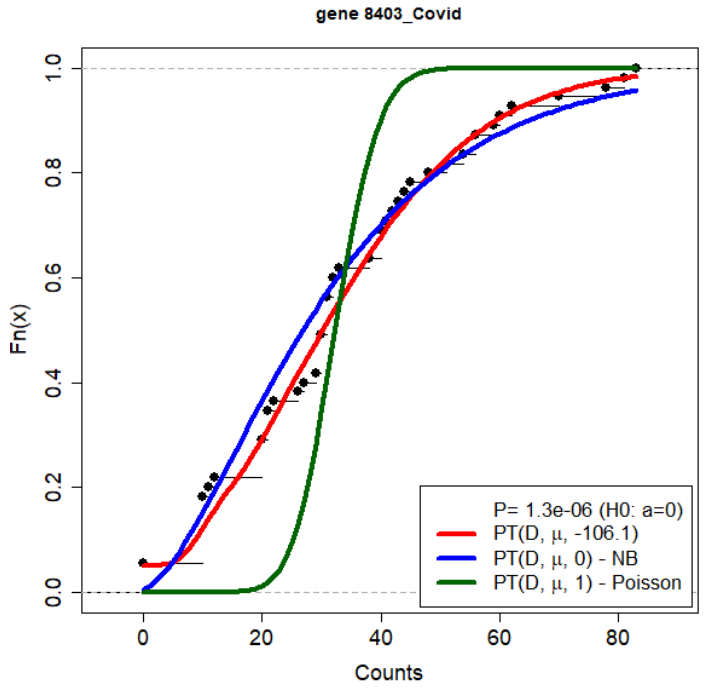
Poisson vs NG vs Tweedie

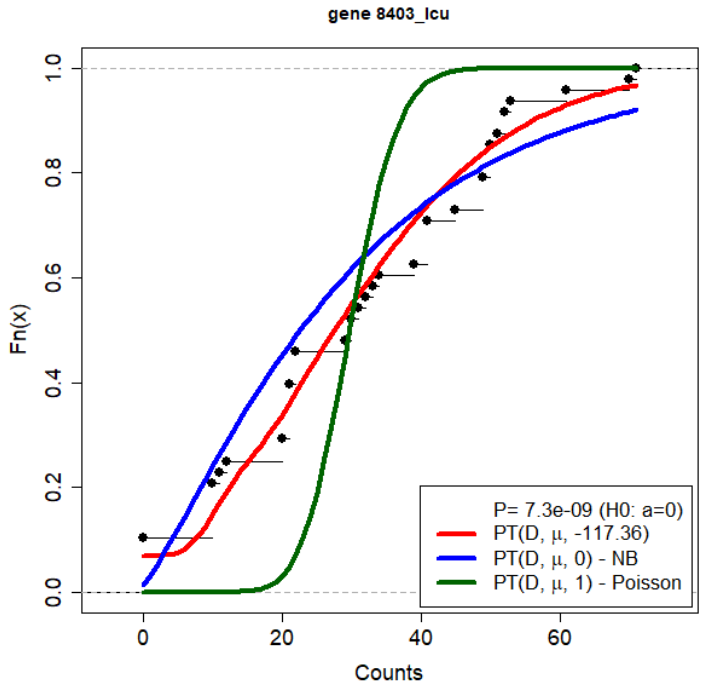
- Tweedie wins the other two models when **the sample size is small**.



Poisson vs NG vs Tweedie

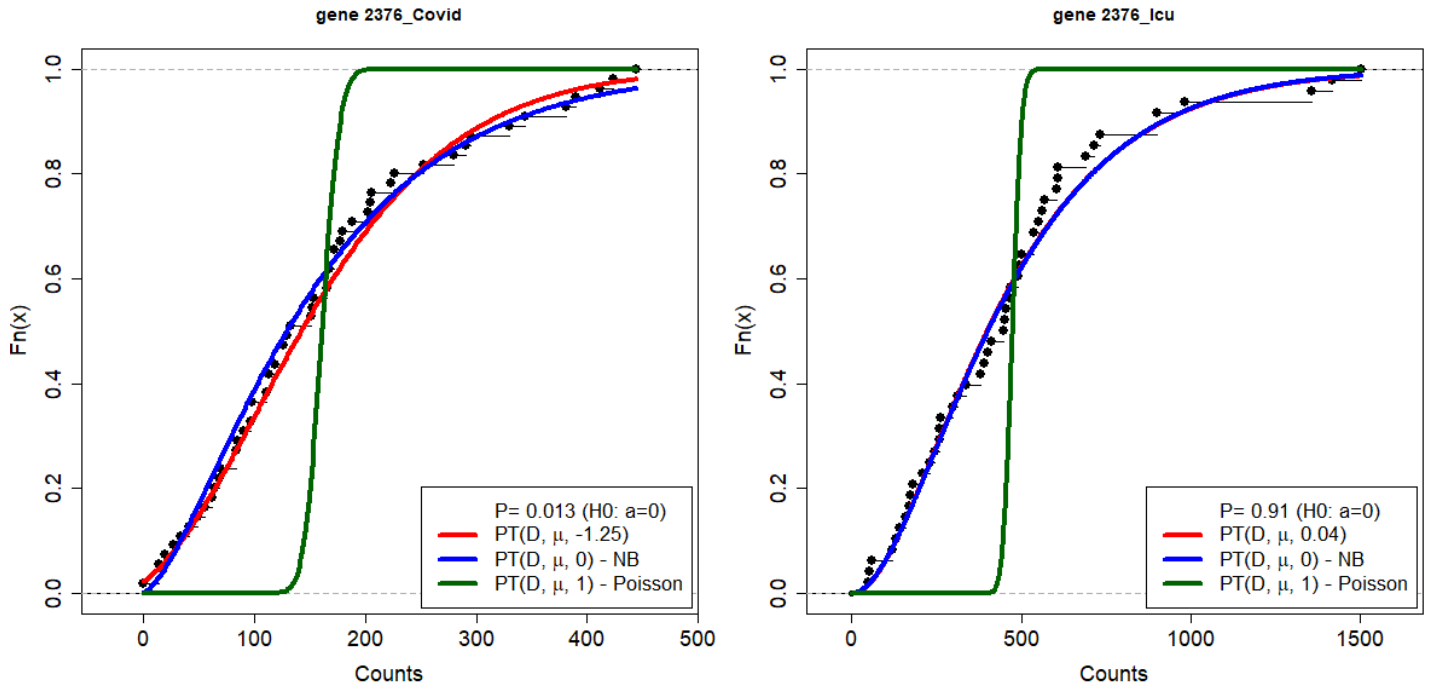
- Both of the **NG and Tweedie** are good and even aligned together when **the sample size is getting larger**. The Poisson seems too restrictive to fit the counts.





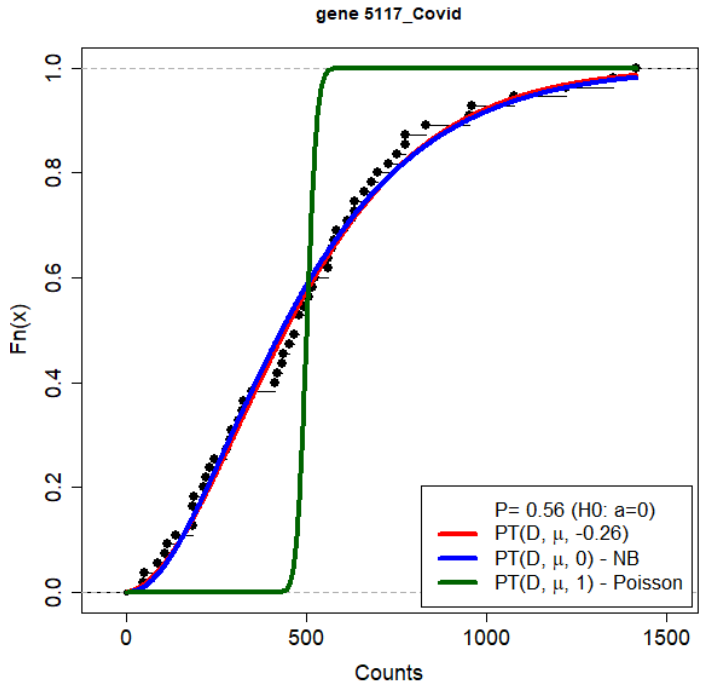
Poisson vs NG vs Tweedie

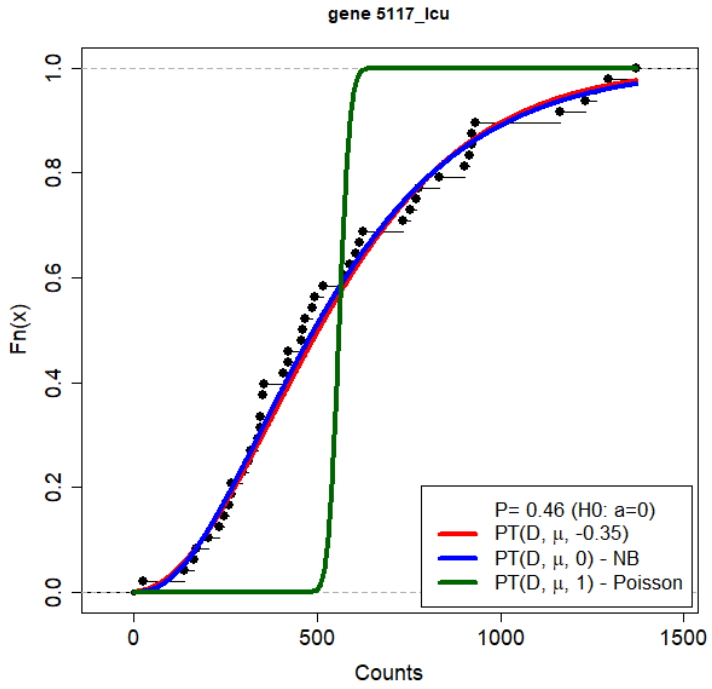
- Both of the **NG and Tweedie** are good and even aligned together when **the sample size is getting larger**. The Poisson seems too restrictive to fit the counts.



Poisson vs NG vs Tweedie

- Both of the **NG and Tweedie** are good and even aligned together when **the sample size is getting bigger**. The Poisson seems too restrictive to fit the counts.





Leading Packages

Method	Normalization	Distribution	Model Comparison Test
DESeq2 [1]	DESeq size Factors	Negative binomial distribution	Wald test
EdgeR [2]	Trimmed Mean of M-values	Negative binomial distribution	The empirical Bayes moderated t-statistics test
Limma-Voom [3]	Trimmed Mean of M-values	Negative binomial distribution	The empirical Bayes moderated t-statistics test
TweeDEseq [4]	Trimmed Mean of M-values	Tweedie distribution	ANOVA method

[1] Love, M.I., Huber, W., Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15:550. 10.1186/s13059-014-0550-8

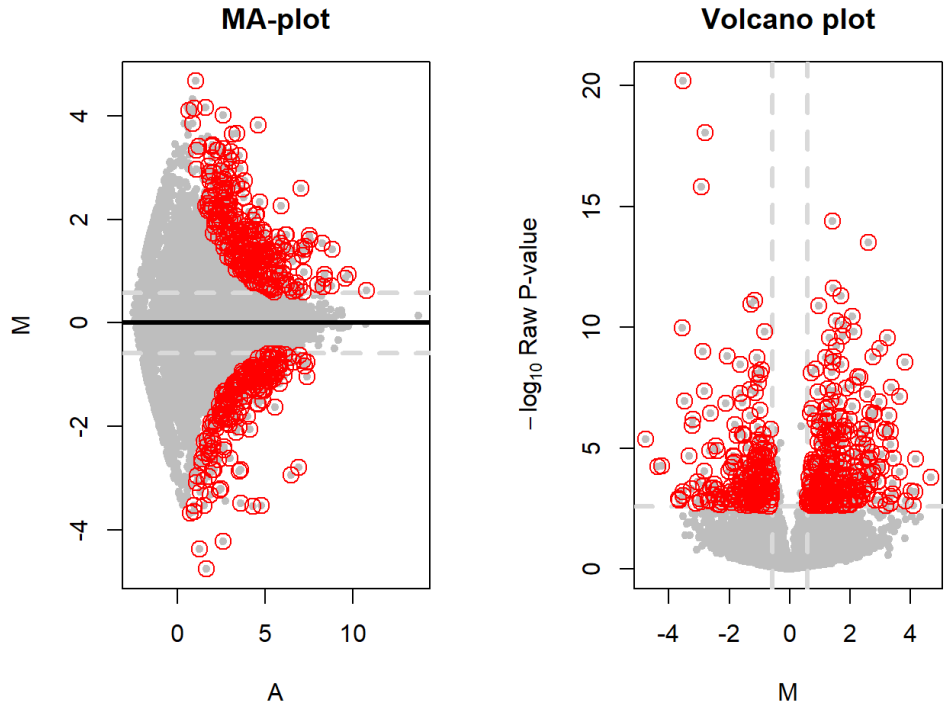
[2] Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F; French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013 Nov;14(6):671-83. doi: 10.1093/bib/bbs046. Epub 2012 Sep 17. PMID: 22988256.

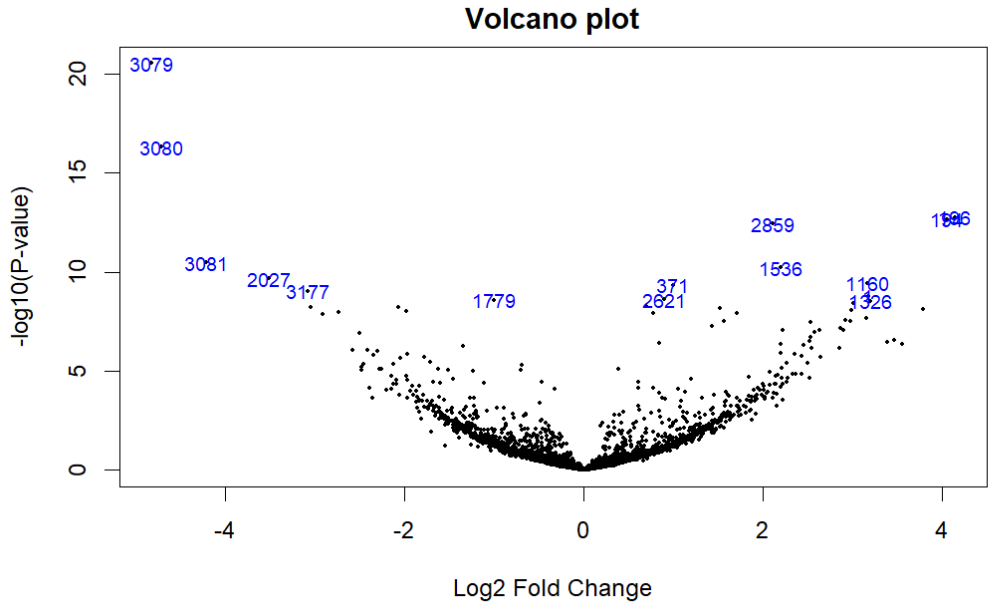
[3] Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F; French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013 Nov;14(6):671-83. doi: 10.1093/bib/bbs046. Epub 2012 Sep 17. PMID: 22988256.

[4] Mikel Esnaola1, Robert Castelo, Juan Ramon Gonzalez; tweeDEseq: analysis of RNA-seq data using the Poisson-Tweedie family of distributions; 2021 Oct.

Visulization

- Each package has their own plots functions, such as Heatmaps, MA-Plots, Volcano Plots.



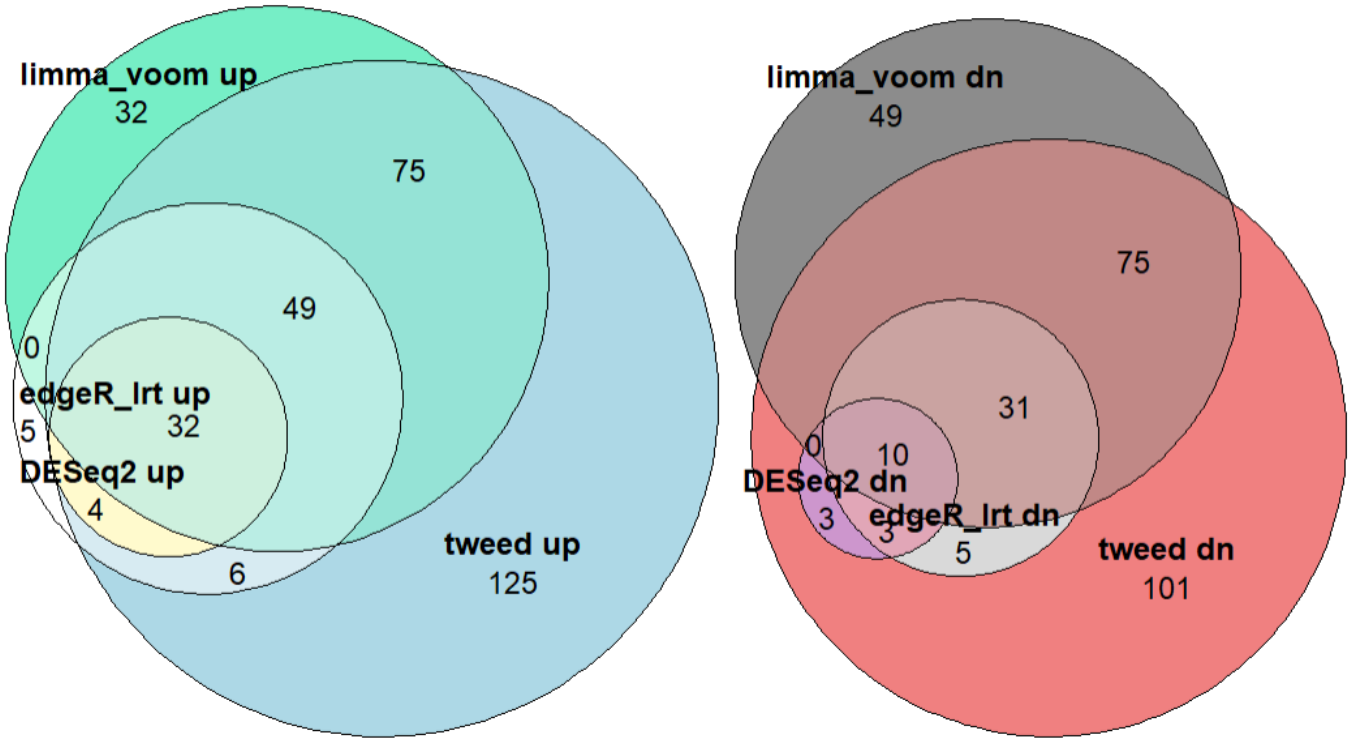


Gene Ranks Comparison

- After building each models , set the threshold as (P_ajusted< 0.05), and divided the selected genes as **UP**(positive influenced) and **DOWN** (negative influenced). We got different length of selected gene in the following:

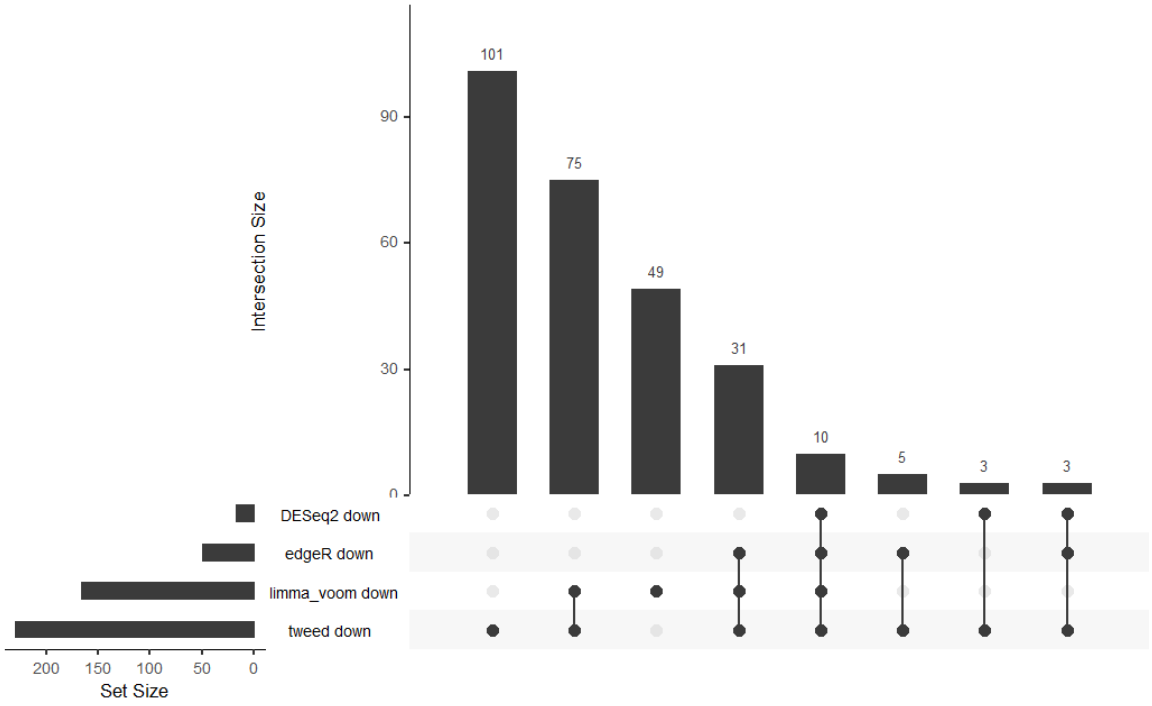
	UP	DOWN
DESeq2	43	16
EdgeR	98	49
Limma-Voom	190	165
TweeDEseq	296	228

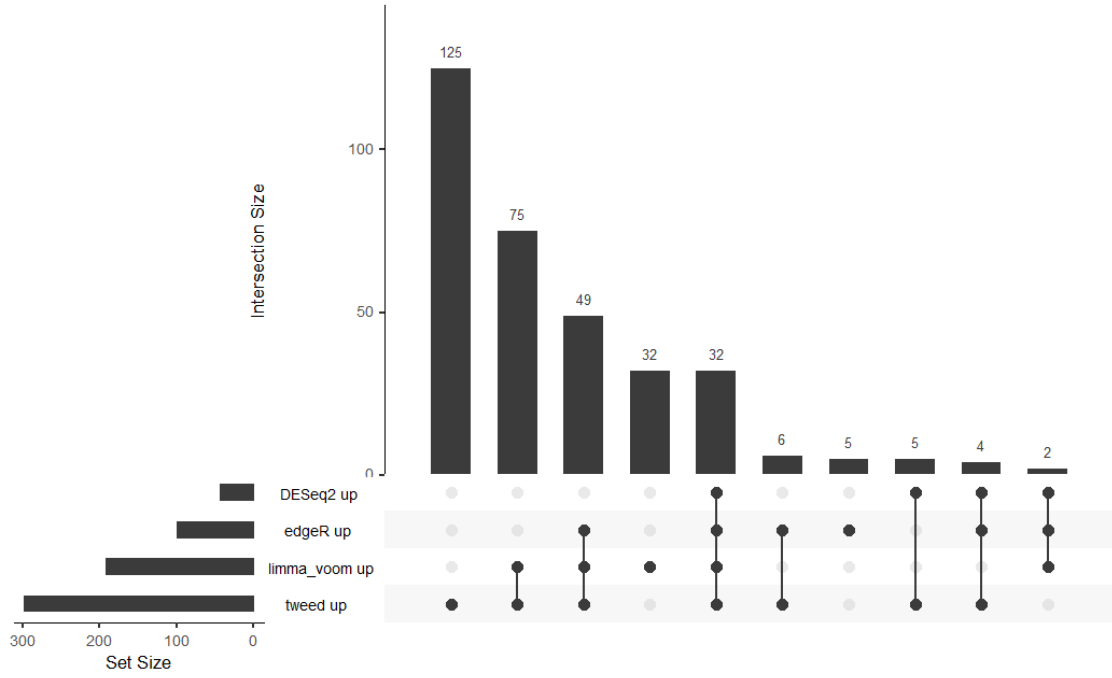
Gene Ranks Comparison



Gene Ranks Compariso

- In general, **TweedDEseq** and **Limma_Voom** have the **most selected genes**.
- **Big amount of** the genes from all four methods are **overlapped**.





References

- 1. Analyzing RNA-seq data with DESeq2 <https://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#model-matrix-not-full-rank> Michael I. Love, Simon Anders, and Wolfgang Huber 05/19/2021
- 2. Statistical models of Differential Expression <https://github.com/mistrm82/msu_ngs2015/blob/master/hands-on.Rmd> jessicalumian 08/15/2015
  3.edgeR: differential analysis of sequence read count data <https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf> Yunshun Chen, Davis McCarthy, Matthew Ritchie,Mark Robinson, and Gordon Smyth
- 4.limma:Linear Models for Microarray and RNA-Seq Data User's Guide <https://bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf> Gordon K. Smyth, Matthew Ritchie, Natalie Thorne,James Wettenhall, Wei Shi and Yifang Hu, Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia
- 5.Practical statistical analysis of RNA-Seq data - edgeR - tomato data<http://www.nathalievialaneix.eu/doc/html/solution_edgeR-tomato.html>,Annick Moisan, Ignacio Gonzales, Nathalie Villa-Vialaneix
- 6.tweeDEseq: analysis of RNA-seq data using the Poisson-Tweedie family of distributions, Mikel Esnaola, Robert Castelo, Juan Ramon Gonzalez
- 7. Generalized Linear Models With Examples in R, Mikel Esnaola, Peter K. Dunn, Gordon K.Smyth