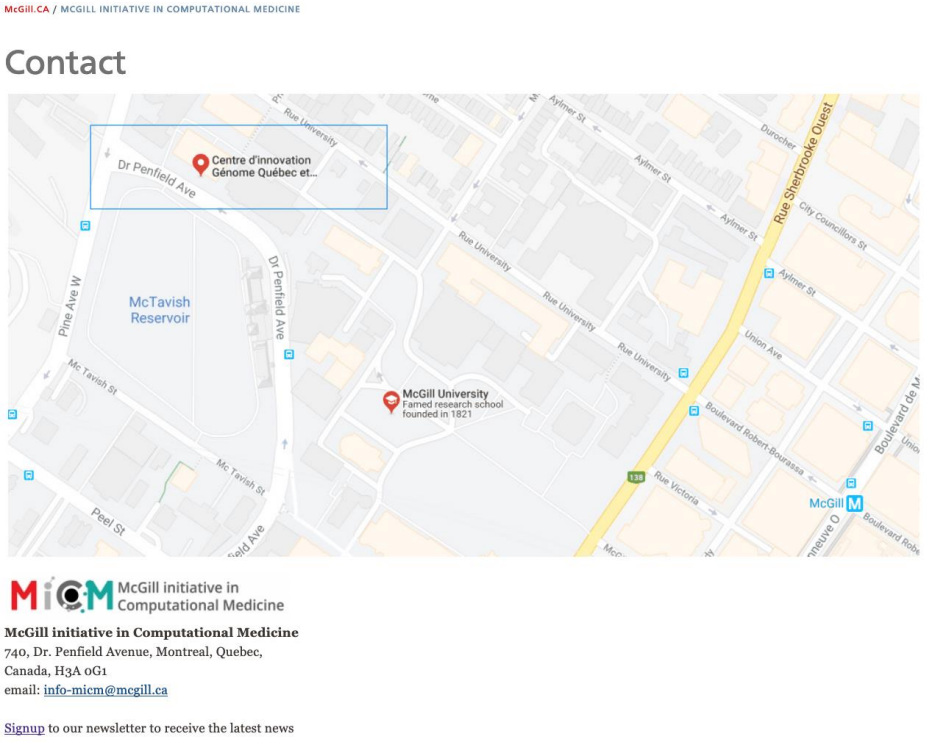
McGill initiative in Computational Medicine

RNA-seq analysis II

Instructor: Ariel Madrigal Aguirre
Facilitator: Zahra Takavol
October 23rd, 2025

McGill initiative in Computational Medicine

Mission : aims to deliver inter-disciplinary research programs and empower the use of data in health research and health care delivery



https://www.mcgill.ca/micm

Outline:

1. Differential expression using DESeq2
2. GSEA and ORA
3. Concluding remarks

Acknowledgements

Material:
Reinnier Padilla
Adrien Osakwe

Exercises:
DESeq2 vignette
fgsea vignette
Enrichr vignette

Data:
Link

This is an interactive workshop :)

Feel free to interrupt or raise your hand to ask questions

Part 1: Differential Expression using DESeq2

- Statistical modelling concepts used in DESeq2
- Log-fold shrinkage
- Introduction to design matrices for gene expression experiments
- Multi-factor designs
- Hands-on activity 1

Statistical modelling concepts used in DESeq2 and its differences with respect to edgeR

	edgeR	DESeq2
Dispersion	Negative Binomial dispersions capture global trend, quasi-dispersions account for gene-specific variability.	Gene-specific dispersion shrunken towards the trend using a MAP
Normalization	TMM	Median of ratios
Test	Quasi-likelihood F-tests	Wald-tests

Model used in DESeq2

DESeq2 models the counts using a negative binomial distribution

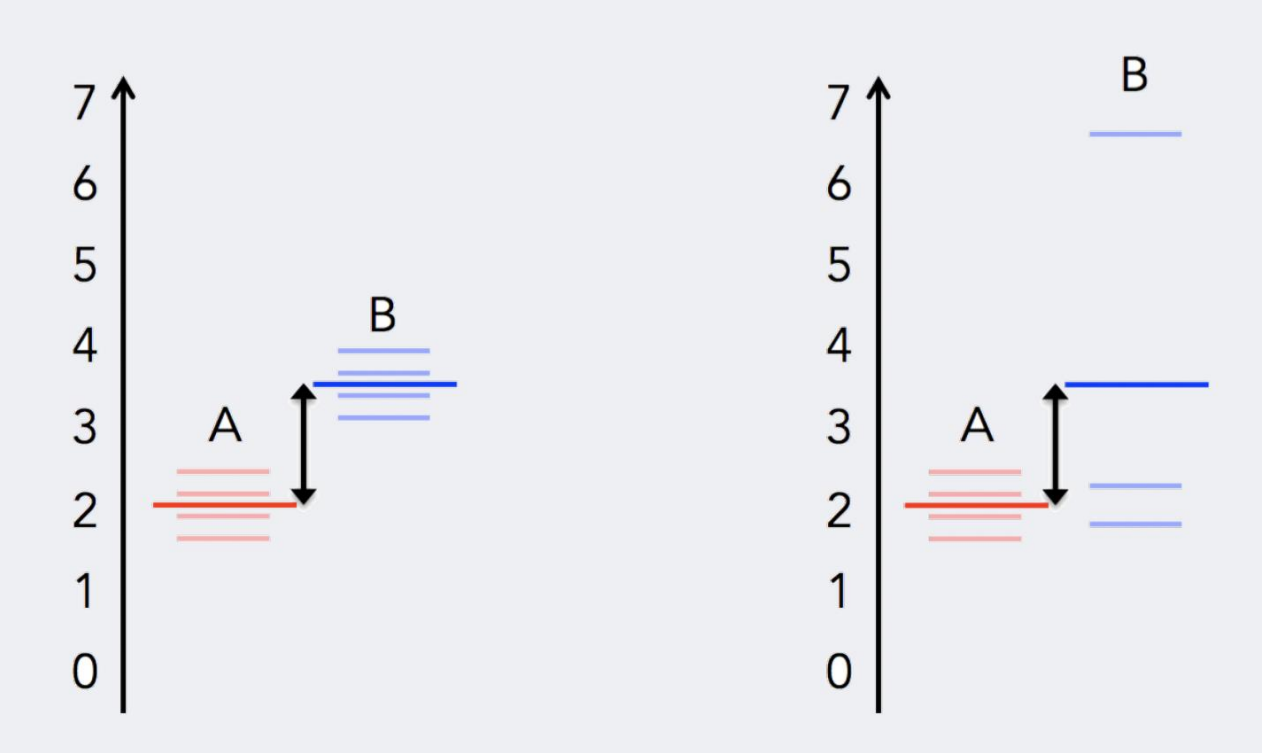
$$K_{ij} \sim \text{NB}(\text{mean} = \mu_{ij}, \text{dispersion} = \alpha_i)$$
$$\mu_{ij} = s_{ij} q_{ij} \quad (1)$$

$$\log q_{ij} = \sum_r x_{jr} \beta_{ir}. \quad (2)$$

K_{ij}	counts of reads for gene i, sample j
μ_{ij}	fitted mean
α_i	gene-specific dispersion
s_j	sample-specific size factor
s_{ij}	gene- and sample-specific normalization factor
q_{ij}	proportional to true concentration of fragments
x_{jr}	elements of the design matrix X
β_{ir}	the logarithmic fold change for gene i and covariate r

Love, M. *et al* (2014) *Genome Biology*

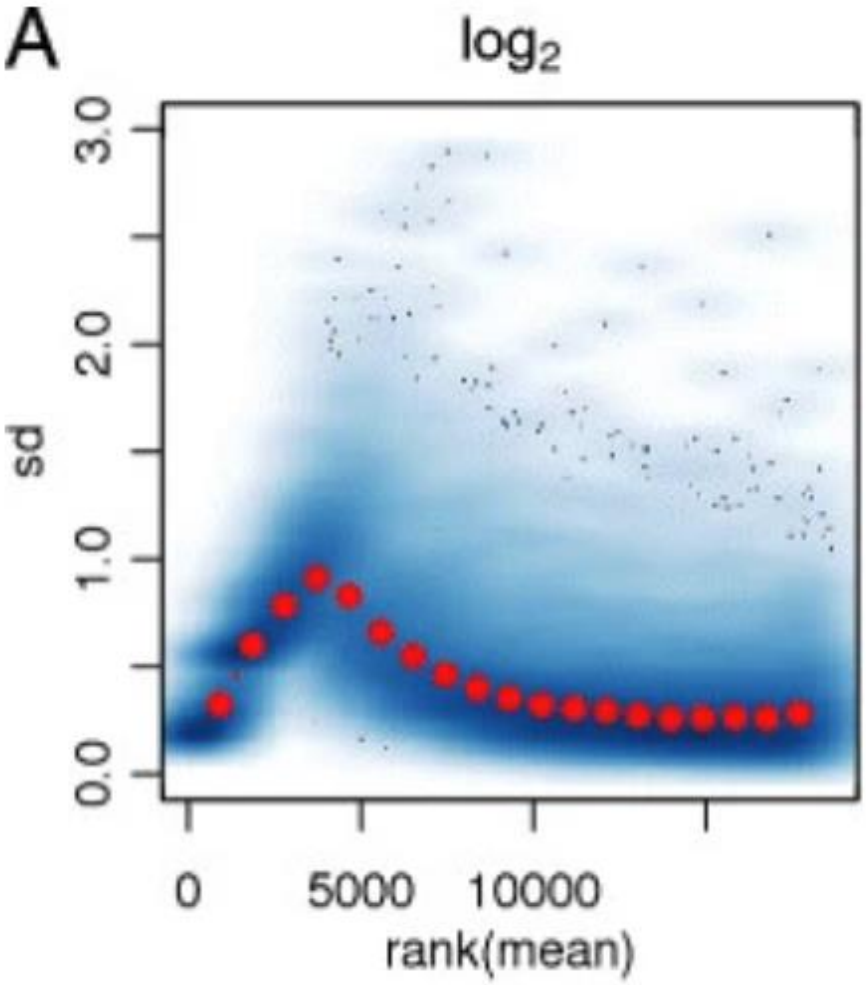
Uneven distribution of information



Replication introduces variance

Adapted from: https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to-lessons.html

Uneven distribution of information



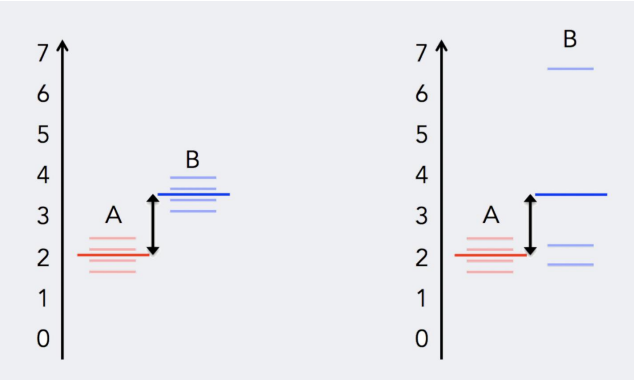
Love, M. *et al* (2014) *Genome Biology*

What is dispersion?

Within-group variability (the variability between replicates), is modeled by the dispersion parameter

$$K_{ij} \sim \mathrm{NB}(\mathrm{mean} = \mu_{ij}, \mathrm{dispersion} = \alpha_i)$$

$$\mathrm{Var}\, K_{ij} = \mu_{ij} + \alpha_i \mu_{ij}^2$$



K_{ij} counts of reads for gene i, sample j

μ_{ij} fitted mean

α_i gene-specific dispersion

Love, M. *et al* (2014) *Genome Biology*

Shrinkage estimation of dispersion

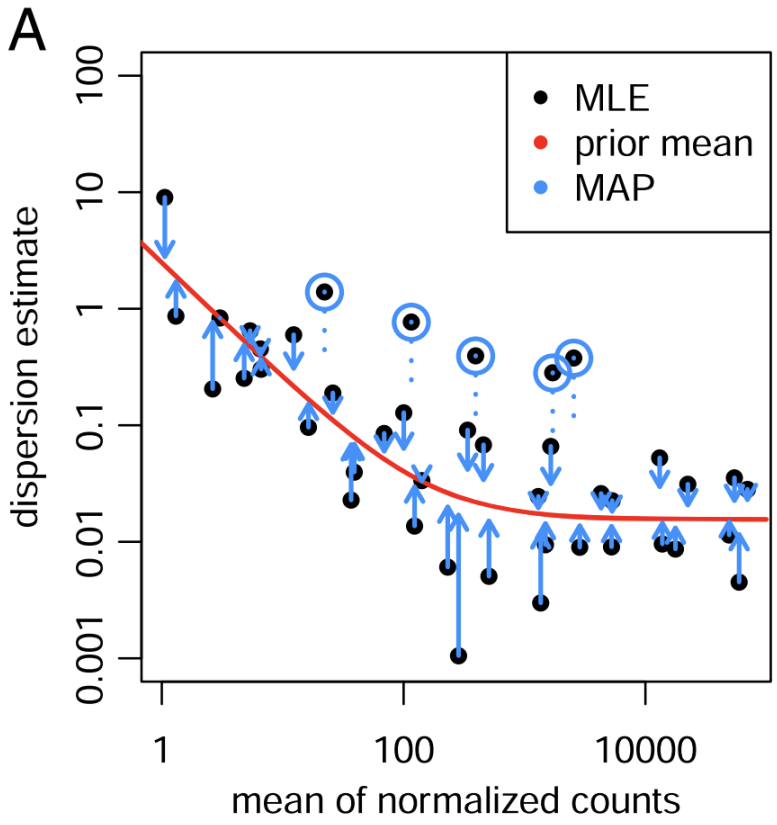


Fig 1A, adapted from: Love, M. *et al* (2014) *Genome Biology*

Wald tests

$$W = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

We can assume the log fold change under the null hypothesis to be distributed normally with mean zero and variance that depends on the standard error.

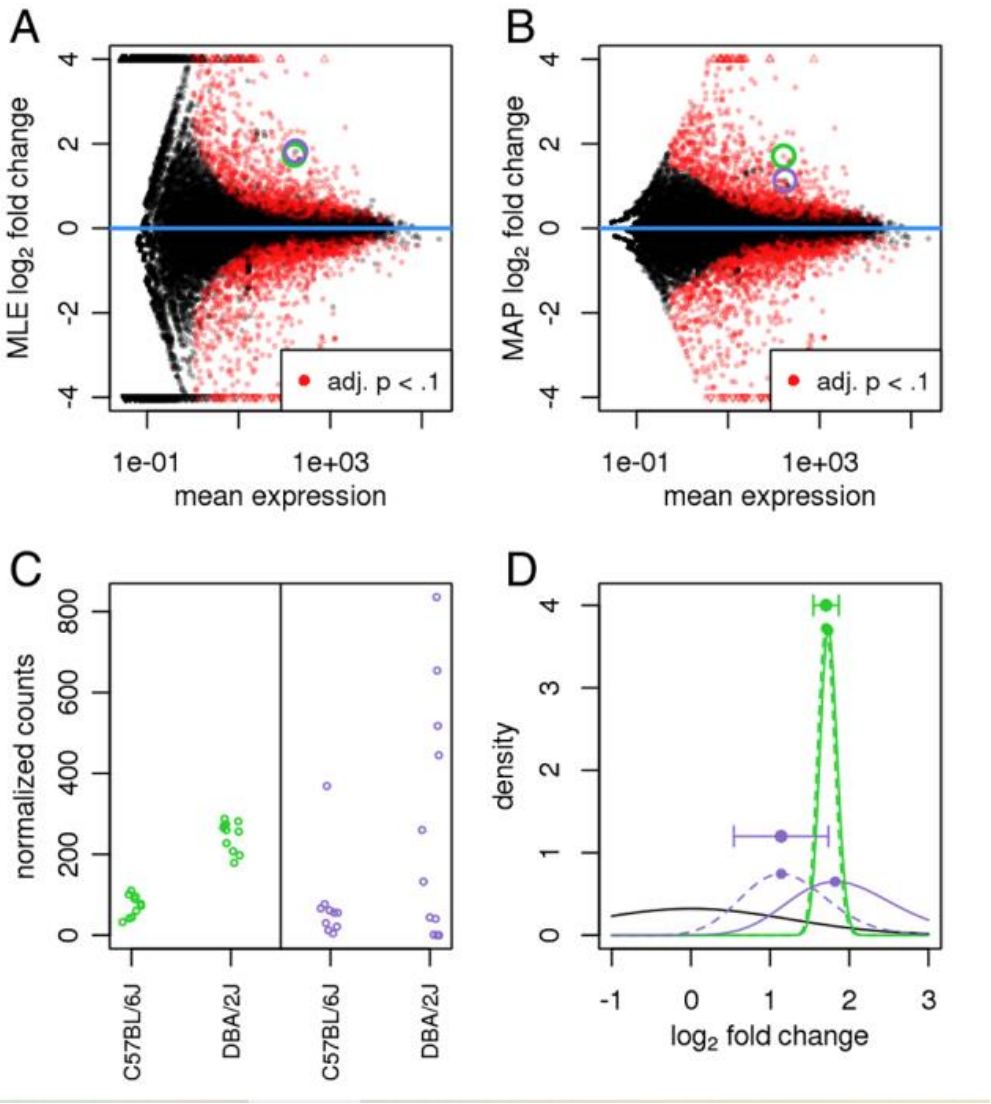
From there, we can compute a two-sided p-value

Love, M. *et al* (2014) *Genome Biology*

Part 1: Differential Expression using DESeq2

- Statistical modelling concepts used in DESeq2
- Log-fold shrinkage
- Introduction to design matrices for gene expression experiments
- Multi-factor designs
- Hands-on activity 1

Log-fold shrinkage



Shrinkage of FC addresses two issues:

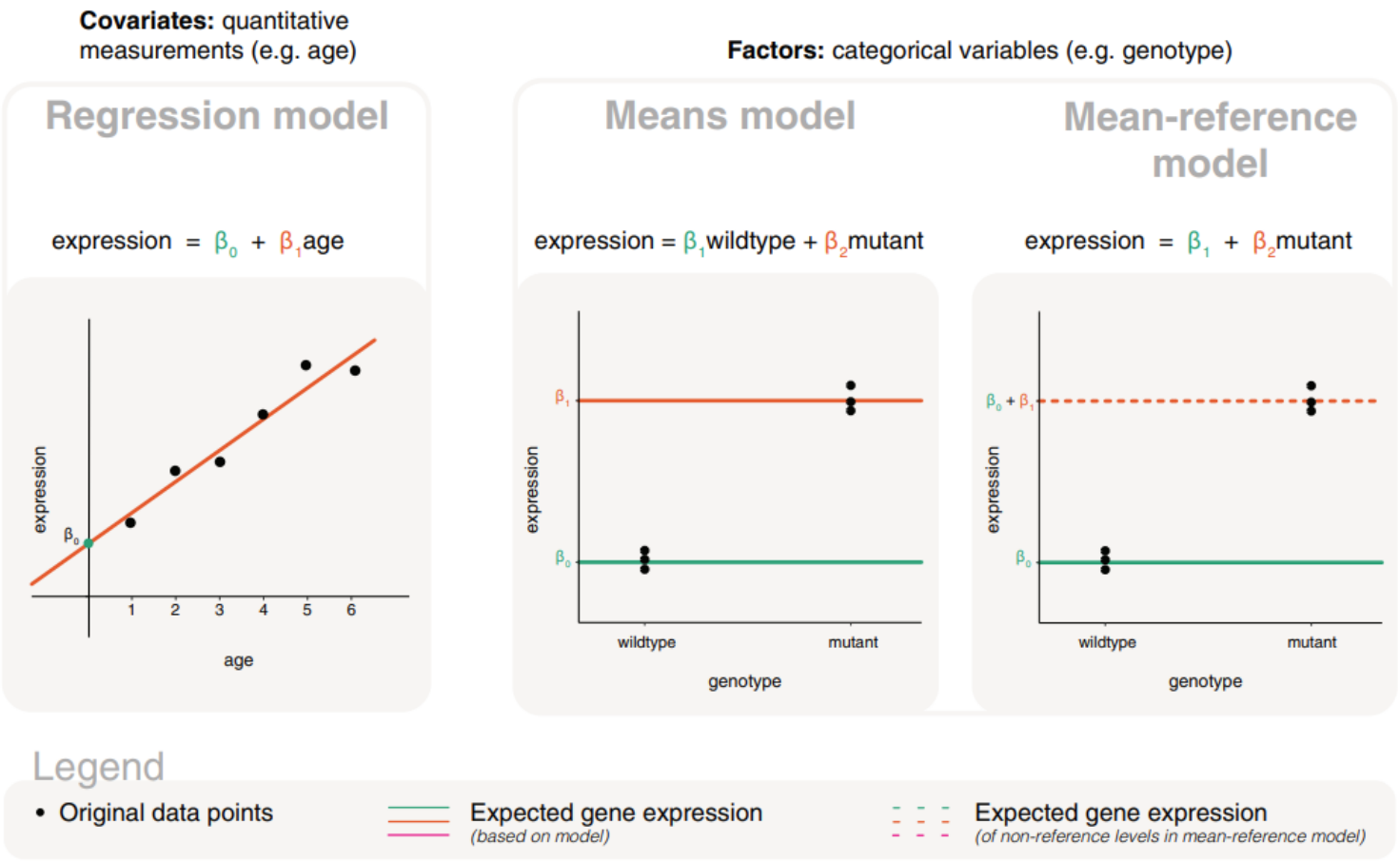
- Log-fold changes are noisier when counts are low
- Genes with high dispersion should be less reliable

Love, M. *et al* (2014) *Genome Biology*

Part 1: Differential Expression using DESeq2

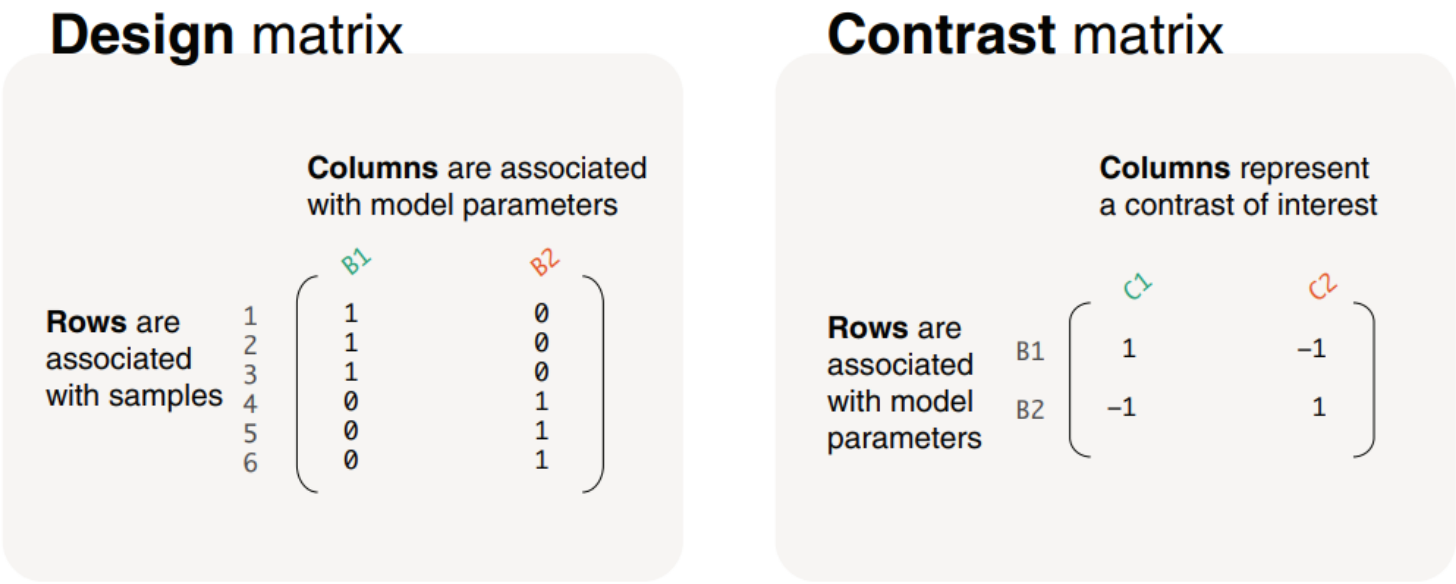
- Statistical modelling concepts used in DESeq2
- Log-fold shrinkage
- Introduction to design matrices for gene expression experiments
- Multi-factor designs
- Hands-on activity 1

Models for covariate and factor variables



Law, C. *et al* (2020) *F1000 Research*

Design and contrast matrices



Law, C. *et al* (2020) *F1000 Research*

Gene expression modelled by a group factor

~ group

Model

E(y) = 2.95 + 1.62x

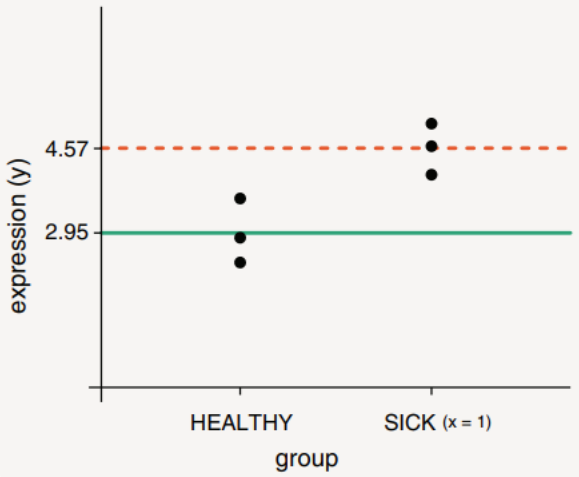
E(y) = 2.95	=	2.95	*(for healthy group)*
E(y) = 2.95 + 1.62	=	4.57	*(for sick group)*

Matrix

```
> model.matrix(~group)
```

	(Intercept)	groupSICK
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Plot



Law, C. *et al* (2020) *F1000 Research*

Gene expression modelled by a group factor (excluding intercept)

~ 0 + group

Model

$$E(y) = 2.95x_1 + 4.57x_2$$

E(y) = 2.95 = 2.95 *(for healthy group)*

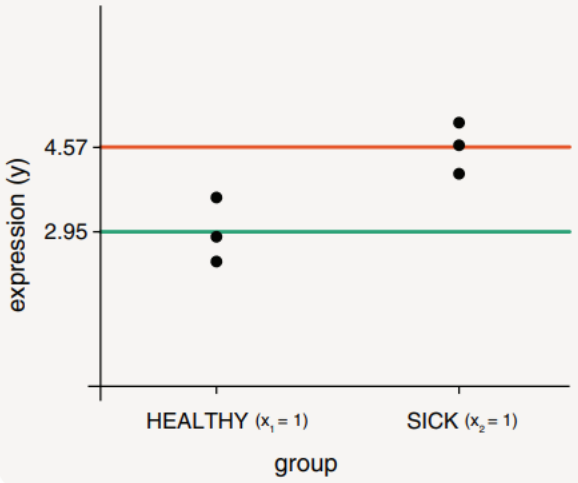
E(y) = 4.57 = 4.57 *(for sick group)*

Matrix

```
> model.matrix(~0 + group)
```

	groupHEALTHY	groupSICK
1	1	0
2	1	0
3	1	0
4	0	1
5	0	1
6	0	1

Plot



Law, C. *et al* (2020) *F1000 Research*

Part 1: Differential Expression using DESeq2

- Statistical modelling concepts used in DESeq2
- Log-fold shrinkage
- Introduction to design matrices for gene expression experiments
- Multi-factor designs
- Hands-on activity 1

An interaction model captures a synergistic effect

~ treat1 * treat2 equivalent to ~ treat1 + treat2 + treat1:treat2

Model

$$E(y) = 1.03 + 1.09x_1 + 1.97x_2 + 0.82x_1x_2$$

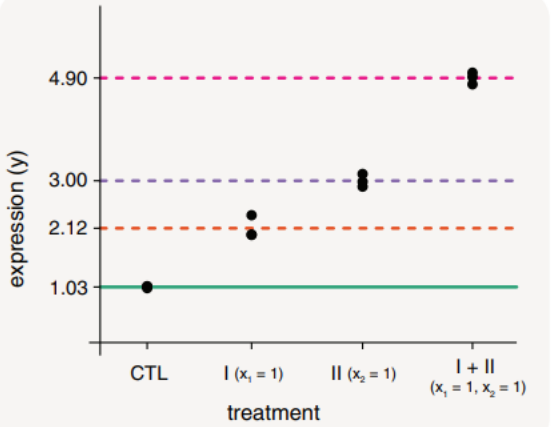
E(y) = 1.03	=	1.03	*(for control)*
E(y) = 1.03 + 1.09	=	2.12	*(for treatment I)*
E(y) = 1.03 + 1.97	=	3.00	*(for treatment II)*
E(y) = 1.03 + 1.09 + 1.97 + 0.82	=	4.90	*(for treatments I & II)*

Matrix

```
> model.matrix(~treat1 * treat2)
```

	(Intercept)	treat1YES	treat2YES	treat1YES:treat2YES
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	1	0	0
5	1	1	0	0
6	1	1	0	0
7	1	0	1	0
8	1	0	1	0
9	1	0	1	0
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1

Plot



One plus one is greater than two

Law, C. *et al* (2020) *F1000 Research*

A group factor can represent two factors (tissue sample and cell type)

```
##    expression      id tissue cells   group
## 1        1.01  MOUSE1   LUNG     B  LUNG_B
## 2        1.04  MOUSE2   LUNG     B  LUNG_B
## 3        1.04  MOUSE3   LUNG     B  LUNG_B
## 4        1.99  MOUSE4  BRAIN     B BRAIN_B
## 5        2.36  MOUSE5  BRAIN     B BRAIN_B
## 6        2.00  MOUSE6  BRAIN     B BRAIN_B
## 7        2.89  MOUSE7   LUNG     T  LUNG_T
## 8        3.12  MOUSE8   LUNG     T  LUNG_T
## 9        2.98  MOUSE9   LUNG     T  LUNG_T
## 10       5.00 MOUSE10  BRAIN     T BRAIN_T
## 11       4.92 MOUSE11  BRAIN     T BRAIN_T
## 12       4.78 MOUSE12  BRAIN     T BRAIN_T
```

Law, C. *et al* (2020) *F1000 Research*

A group factor can represent two factors (tissue sample and cell type)

~ 0 + group

Model

$$E(y) = 1.03x_1 + 2.12x_2 + 3.00x_3 + 4.90x_4$$

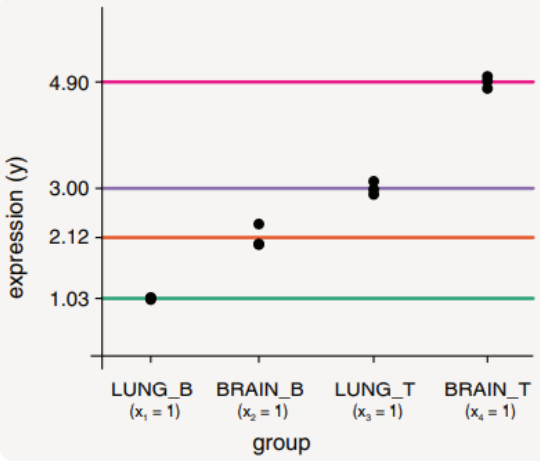
E(y) =	1.03				= 1.03	*(for lung B-cells)*
E(y) =		2.12			= 2.12	*(for brain B-cells)*
E(y) =			3.00		= 3.00	*(for lung T-cells)*
E(y) =				4.90	= 4.90	*(for brain T-cells)*

Matrix

```
> model.matrix(~0 + group)
```

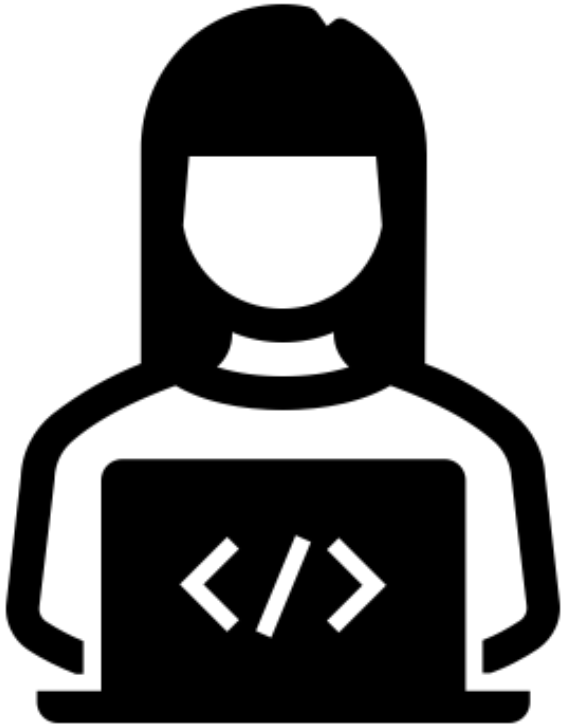
	groupLUNG_B	groupBRAIN_B	groupLUNG_T	groupBRAIN_T
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	0	1	0	0
5	0	1	0	0
6	0	1	0	0
7	0	0	1	0
8	0	0	1	0
9	0	0	1	0
10	0	0	0	1
11	0	0	0	1
12	0	0	0	1

Plot



Law, C. *et al* (2020) *F1000 Research*

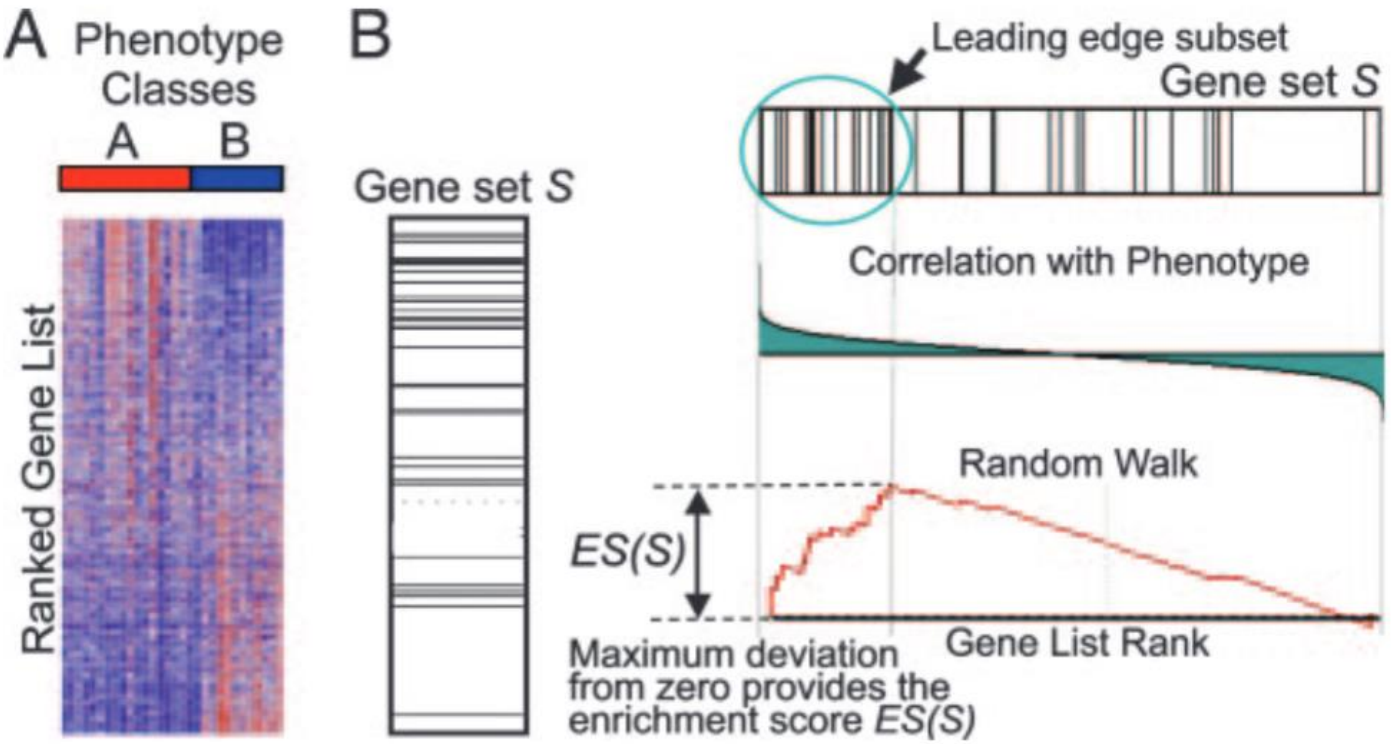
Hands-on 1



Part 2: GSEA and ORA

- Statistical concepts used in GSEA
- Statistical concepts used in ORA
- When to use GSEA or ORA?
- Hands-on activity 2

Gene Set Enrichment Analysis (GSEA)



Subramanian, A. *et al* (2005) *PNAS*

Part 2: GSEA and ORA

- Statistical concepts used in GSEA
- Statistical concepts used in ORA
- When to use GSEA or ORA?
- Hands-on activity 2

Enrichr

Enrichr

	In Pathway	Not in Pathway	Total
Selected genes	k	$n - k$	n
Not selected genes	$K - k$	$N - K - (n - k)$	$N - n$
Total	K	$N - K$	N

Where:

- N = total number of genes in the background (e.g. all detected genes).
- K = number of genes in the pathway.
- n = number of genes selected as "significant" (e.g. DE genes).
- k = overlap between selected genes and pathway genes.

P-value comes from a hypergeometric test

$$p = \sum_{i=k}^{\min(n,K)} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}$$

Chen E. *et al* (2013) *BMC Bioinformatics*

Part 2: GSEA and ORA

- Statistical concepts used in GSEA
- Statistical concepts used in ORA
- When to use GSEA or ORA?
- Hands-on activity 2

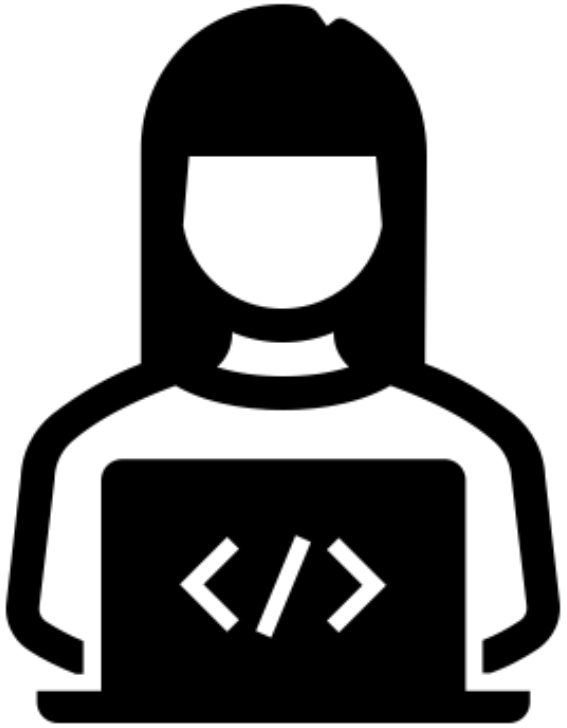
When to use GSEA or ORA?

	GSEA	ORA
Input	Ranked list of all genes	Significant gene list + background
Test type	Hypergeometric	Running-sum enrichment score + permutation
Uses cutoff?	Yes (DE)	No

ORA asks: Are my DE genes enriched in this pathway?

GSEA asks: Are genes from this pathway enriched towards the top or bottom of the ranked list?

Hands-on 2



Part 3: Concluding remarks

Now you are ready to:

- Perform a standard **DGE**, **GSEA** and **ORA** analyses
- Modify design matrices to address biological questions
- Choose between GSEA and ORA depending on the question at hand

Thanks for your attention!

McGill UNIVERSITY

Keep an eye for the workshops offered by the MiCM!

info-micm@mcgill.ca
https://www.mcgill.ca/micm/

References

- Love, M., Anders, S., & Huber, W. (2014). Differential analysis of count data–the DESeq2 package. *Genome Biol*, *15*(550), 10-1186.
- Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., & Mesirov, J. P. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics*, *23*(23), 3251-3253.
- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., ... & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, *14*(1), 128.
- Law, C. W., Zeglinski, K., Dong, X., Alhamdoosh, M., Smyth, G. K., & Ritchie, M. E. (2020). A guide to creating design matrices for gene expression experiments. *F1000Research*, *9*, 1444.

Normalization in DESeq2

Median of ratios

$$\mu_{ij} = s_{ij} q_{ij}$$

Size factor

$$s_j = \underset{i:K_i^{\mathrm{R}} \neq 0}{\text{median}} \frac{K_{ij}}{K_i^{\mathrm{R}}} \quad \text{with} \quad K_i^{\mathrm{R}} = \left(\prod_{j=1}^{m} K_{ij} \right)^{1/m}.$$

Geometric mean

K_{ij}	counts of reads for gene i, sample j
μ_{ij}	fitted mean
α_i	gene-specific dispersion
s_j	sample-specific size factor
s_{ij}	gene- and sample-specific normalization factor
q_{ij}	proportional to true concentration of fragments

Love, M. *et al* (2014) *Genome Biology*