McGill initiative in Computational Medicine

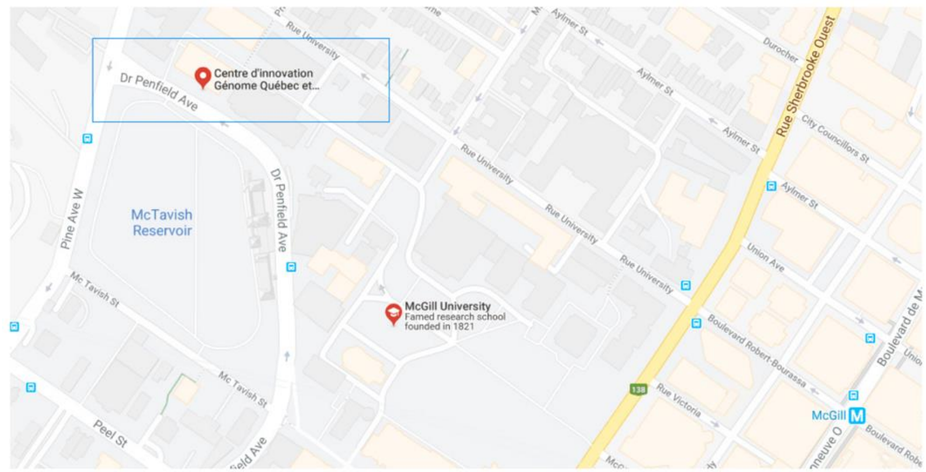
Statistical Foundation for RNA-Seq Analysis

Instructor: Ryan Huang
Oct. 22, 2025

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

McGill.CA / MCGILL INITIATIVE IN COMPUTATIONAL MEDICINE

Contact



McGill initiative in
Computational Medicine

McGill initiative in Computational Medicine
740, Dr. Penfield Avenue, Montreal, Quebec,
Canada, H3A 0G1
email: info-micm@mcgill.ca

Signup to our newsletter to receive the latest news

https://www.mcgill.ca/micm

Outline

1. Task of differential gene expression: pairwise vs multigroup comparisons
2. Preprocess for fair comparison: Filtering / Normalization
3. Statistical testing: Exact test for pairwise
4. Linear regression
5. Poisson and negative binomial distributions
6. Generalized linear model
7. Sample analysis using edgeR (R notebook)

Differential Gene Expression (DGE) analysis

Learning objectives:

- Understand the goal of differential gene expression analysis
- Understand how experimental designs affect choice of analysis methods
- Know the general pipeline of DGE analysis

Differential Gene Expression Analysis: What?

- **Goal:** identify differentially expressed genes (DEGs)
  - **Determine how treatment and genotype affect expression by identifying significantly upregulated and downregulated genes**
  - **We can then analyze these identified genes to infer underlying biological mechanisms causing the variance**
- Common packages to perform DGE analysis:
  - **edgeR**
  - DESeq2
  - limma

Differential Gene Expression: How?
edgeR

- Empirical analysis of Digital Gene Expression in R.
- One of the *Bioconductor* packages
  - *"The mission of the Bioconductor project is to develop, support, and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays."*

```r
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("edgeR")

#load libraries ----
library(edgeR)
```

McGill initiative in Computational Medicine

How to conduct DGE analysis with edgeR:

A - Pre-processing steps:

1. Convert genetic data into a DGE object → allows us to use analysis functions
2. Filter out low count genes → focusing on genes that are more likely to be involved in the biological processes of interest
3. Normalize data to enable comparison → consider each sample's library size and composition

B - Analysis steps:

1. Estimate the common and tag-wise dispersion
   - Common = overall variance
   - Tag-wise = gene specific variance
2. Perform exact test two sample groups of choice OR generalized linear model analysis

Preprocessing

Learning objectives:
- Understand why and how we filter counts
- Understand why and how we normalize counts

Filtering Genes: Why?

- Remove by functional category
  - Ribosomal RNA

- Remove genes with very low counts: genes with very low counts across all libraries provide little evidence for differential expression.
  - Reduces computational cost
  - Potentially reduces noise
  - Risk losing novel genes

Filtering Genes: How?

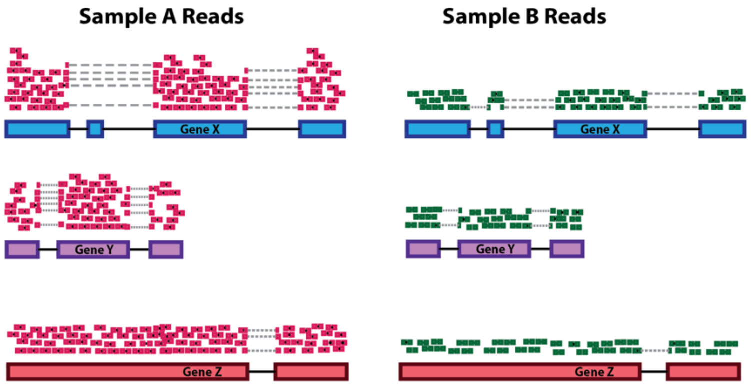
- As a rule of thumb, genes are dropped if they can't possibly be expressed in **all the samples for any of the conditions**.

- Threshold setting:
  - Usually, a gene is required to have a count of 5-10 in a library to be considered expressed in that library.
  - Multiple thresholds should be tried to find the appropriate one: account for experimental design and library size

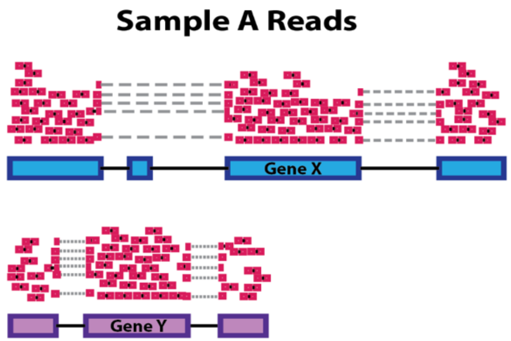
Use counts-per-million which normalizes with library size!!!

Normalization: Why?

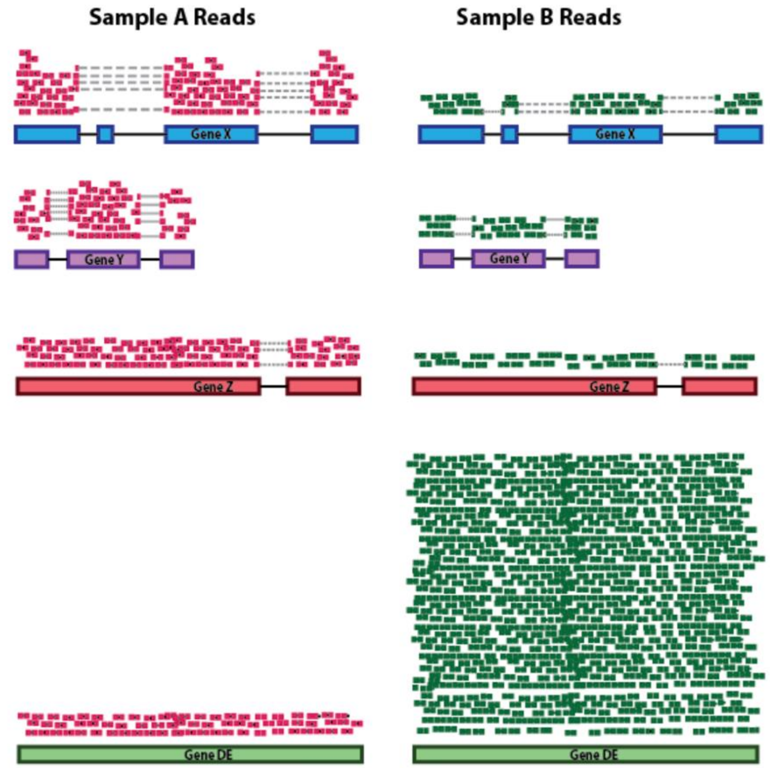
Sequencing Depth



Gene length



RNA composition



Normalization: How?

What are we assuming here by using CPM?

Same total expression

- Counts per million (CPM)
  - It normalizes RNA-seq data for sequencing depth but not gene length.
  - Raw counts are divided by the number of sequencing reads in your sample, multiplied by a million.
- Trimmed Mean of M-values (TMM)
  - "RNA-seq measures relative expression rather than absolute expression. This becomes important for differential expression analyses when a small number of genes are very highly expressed in some samples but not in others. If a small proportion of highly expressed genes consume a substantial proportion of the total library size for a particular sample, this will cause the remaining genes to be under-sampled for that sample. Unless this effect is adjusted for, the remaining genes may falsely appear to be down-regulated in that sample."
  - Original paper: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25

TMM: How it works?

How TMM works:

- **Scaling factors**: The key idea is to compute a scaling factor for each sample that normalizes the counts. A reference sample (usually the median sample or a pseudo-reference) is chosen, and the counts of all other samples are compared to it.
- **M-values**: For each gene, an M-value is calculated, which is the log ratio of the expression levels between the sample and the reference:

$$M = \log_2\left(\frac{\text{gene count in sample}}{\text{gene count in reference}}\right)$$

  - The M-values for highly expressed genes (that dominate the library) or genes with very low counts (often noisy) can distort the normalization.
  - To avoid this, TMM trims (excludes) genes that are too highly or too lowly expressed from the calculation of the scaling factor.
- **Final step**: After trimming, the mean of the remaining M-values is used to compute a scaling factor for each sample. These scaling factors are then used to normalize the raw counts, ensuring that differences in library size or sequencing depth do not skew the analysis.

From ChatGPT

Other normalization methods: Gene length

Do we need this?

Gene length is the same for all samples.

- Reads per kilobase per million reads (RPKM):
  - RPKM is a within-sample normalization method which removes transcript-length and library size effects
  - First normalize by library size and then normalize by transcript length
  - Designed for single-end reads
- Fragments per kilobase per million mapped reads (FPKM):
  - FPKM is an extension of RPKM designed for paired-end reads
  - Each pair of reads are treated as one fragment here (if they both were mapped). This avoids counting a fragment twice which cannot be done isby RPKM.
  - Renders the same output as RPKM on single-end reads.
- Transcripts per million (TPM)
  - TPM is an extension of RPKM where we first normalize by transcript length and then normalize by library size.
  - The sum of all TPMs is the same for each sample, making it easier to compare the proportion of reads mapped to a gene across samples.
  - Can also convert FPKM counts into TPM

Normalization Limitations

- Assume we have two identical samples
  - Knockout expression of Gene D in sample 2
- Fixed Library size means remaining counts are redistributed over the remaining genes

Gene	Sample 1	Sample 2
A	30	235
B	24	188
C	0	0
D	563	0
E	5	39
F	13	102
Total	635	635

In general, we don't expect the knocked-out genes to dominate the reads.

Statistical Testing

Learning objectives:

- Understand the formulation an assumption of standard statistical tests
- Understand the interpretation of a p-value

Hypothesis Tests

Goal

- Determine if two sets of data are **different**
- Can approach this with **test statistics**
  - Is the difference **significant** or not?

Many Types of Tests

- Parametric tests – **make assumptions on the data distribution**
  - Z-score vs. t-test
  - ANOVA
- Non-parametric – **does not make assumptions on data dist.**
  - Permutation test

Hypothesis Formulation

Null Hypothesis

- Effect size is 0
- Difference between conditions is 0

Alternative Hypothesis

- Effect size is NOT 0
- Difference between conditions is NOT 0

We **accept or reject** the null based on the generated **p-value (< 0.05)**

Understanding Significance

Statistical significance **DOES NOT** imply causality

It is a measurement of the likelihood of the observed data happening by chance.

A p-value tells us the probability of seeing the observation **if the null hypothesis is true.**

Dispersions and the exact test

Exact test definition:

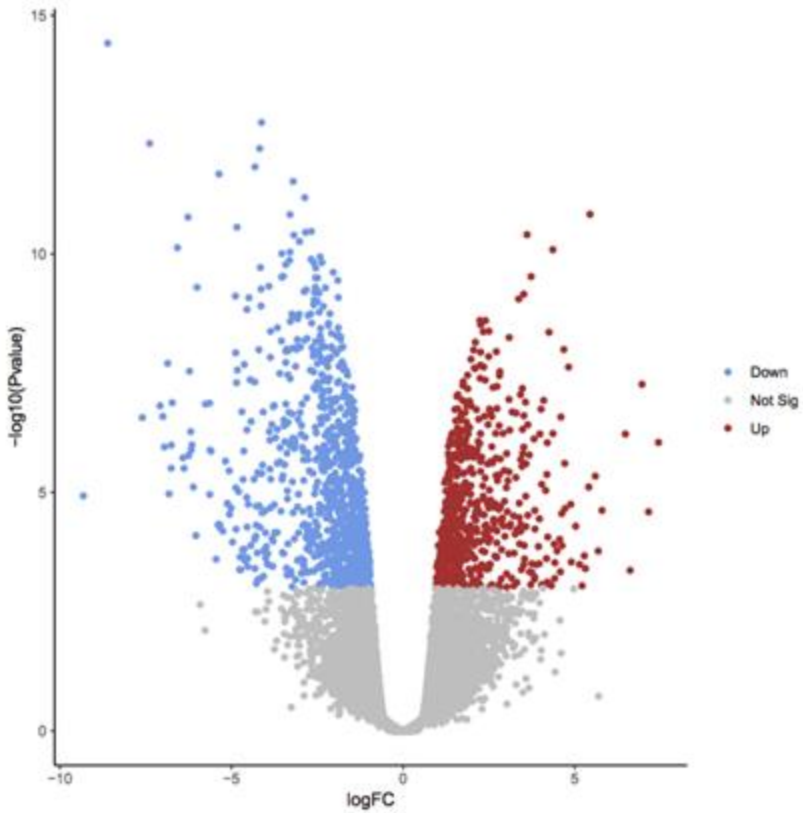
- Statistical method used to compare two groups (e.g., experimental conditions) when data doesn't meet the assumptions of traditional statistical tests (like the t-test).
- Identifies differentially expressed genes while accounting for the overdispersion of RNA-seq data

Dispersions to account for:

- Common dispersion = single value that represents the average level of variability across all genes
- Tag-wise dispersion = gene-specific variability based on its unique characteristics and the common dispersion

Volcano plot

- X axis: Fold change (log2)
- Y axis: pvalue/FDR (-log10)
- Goal: See what genes have a high FC AND high significance



What if we have multiple groups/dosages?

Example: test different siRNA knockdown conditions (25%, 50%, 100%) crossed with drug dosages (1 mM, 0.1 mM, 0.01 mM).

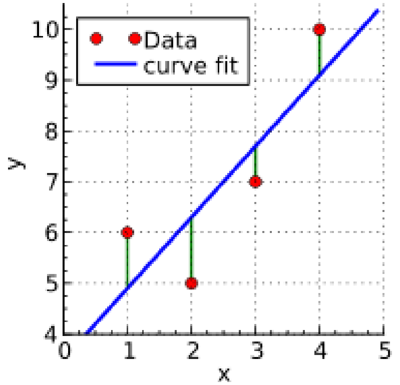
→ 9 groups, how many pair-wise?
→ **Solution: regression**

Linear Regression

Learning objectives:

- Understand the fundamental regression algorithms
- What is a coefficient/effect size

The Linear Model



Credit: Wikipedia

- The fundamental model for statistics & machine learning
- Follows **a Normal distribution**

Imagine a dataset with a dependent variable ***y*** and a set of descriptive features ***x***

We want to learn what features in *x* are good predictors of *y* (what is their **effect size/coefficient** β)

$$\boldsymbol{y} = \boldsymbol{\beta x} + \boldsymbol{\varepsilon} \qquad \varepsilon \sim Normal(0, \sigma^2)$$

$$y \sim Normal(\beta x, \sigma^2)$$

Model Assumptions

1. Outcome is **continuous** and a **linear combination of predictors**
2. Outcome is such that $y_i \sim Normal(\beta x_i, \sigma^2)$
3. Predictors must not be perfectly correlated (linear combination)
4. For every observation *i* the error is:
   1. **Normally distributed**
   2. **Mean zero**
   3. **Homoskedastic** (same variance as other observations)
   4. **Independent** (not correlated)

Model Fitting

Probabilistic Approach

- **Maximum Likelihood Estimation** $\underset{\beta}{\operatorname{argmax}}\ log(Normal(\beta x, \sigma^2))$

Machine Learning

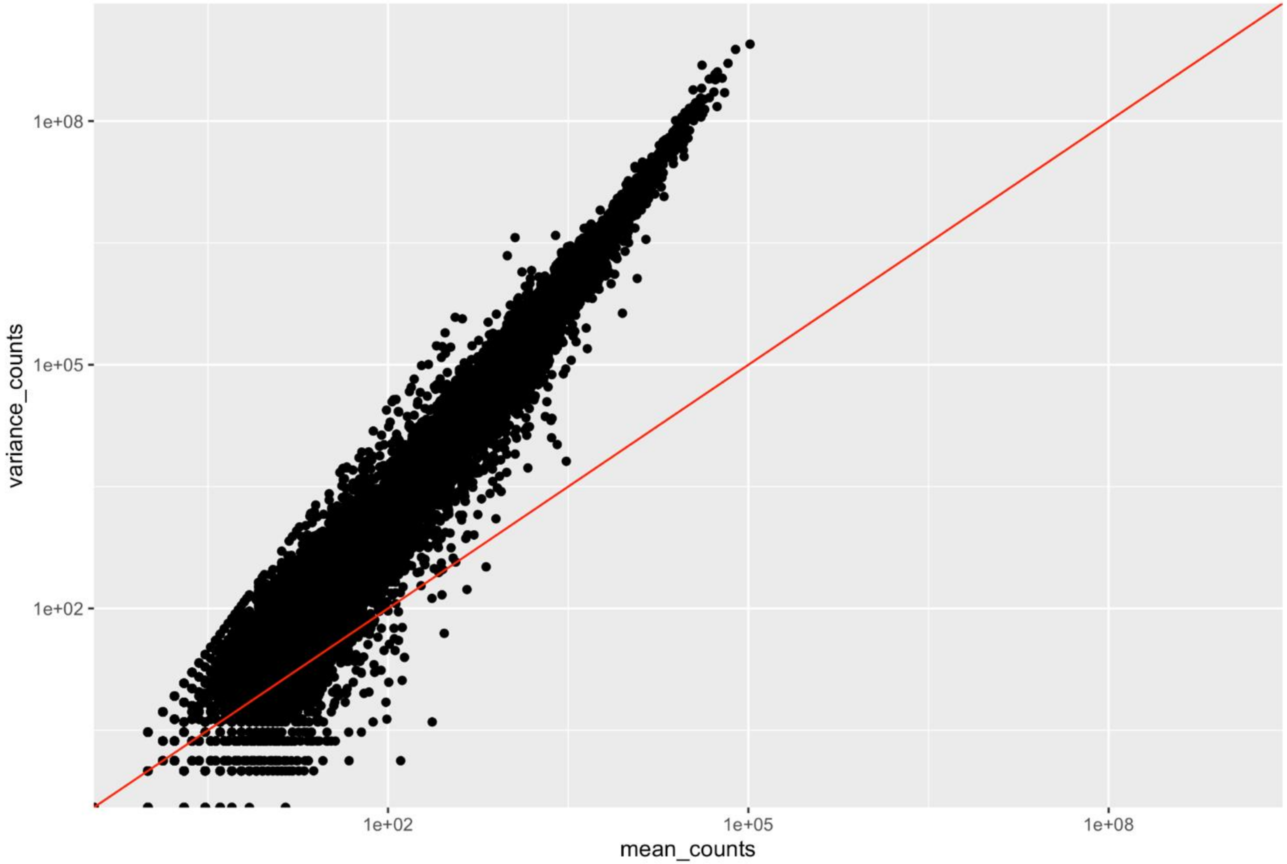
- **Mean Squared Error**

$$y = \beta x + \epsilon$$

$$\hat{y} = \hat{\beta} x$$

$$MSE = \frac{\sum_{i=1}^{N} (\hat{y_i} - y_i)^2}{N}$$

Are gene counts normally distributed? **NO**



Variance not fixed

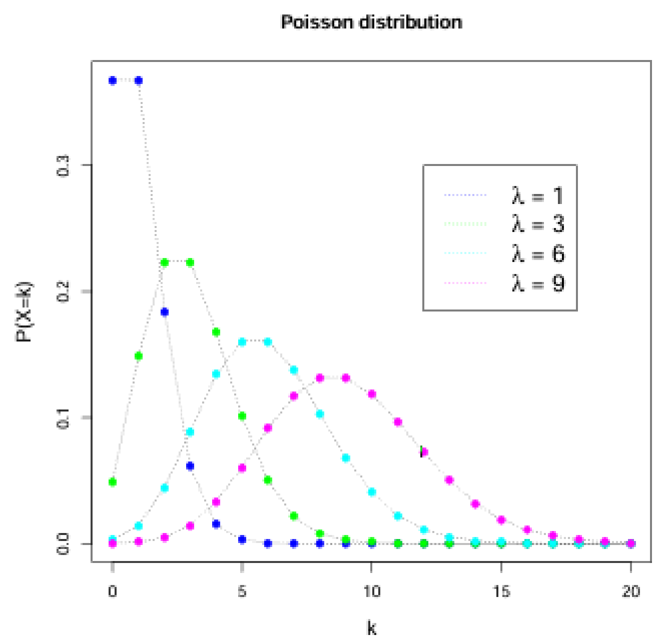
Adapted from: https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to-lessons.html

Poisson and Negative Binomial

Learning objectives:

- Understand properties of Poisson and negative binomial distributions
- Know why we use these to model RNA-seq count data

Poisson distribution

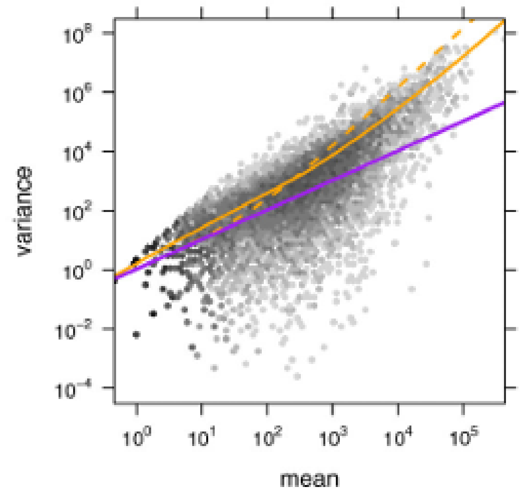


- For $X \sim \text{Poisson}(\lambda)$, both the mean and the variance are equal to λ

From Robinson

Negative binomial

Many studies have shown that the variance grows faster than the mean in RNAseq data. This is known as **overdispersion**.



- Mean count vs variance of RNA seq data. Orange line: the fitted observed curve. Purple: the variance implied by the Poisson distribution.

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11:R106.

From Robinson

What is Dispersion?

- A measure of spread or variability in the data

$$Var_{ij} = \mu_{ij} + \alpha_i \mu_{ij}^2$$

Rearranged:

$$\sqrt{\alpha}_i = \frac{Var_{ij} - \mu_{ij}}{\mu_{ij}}$$

Which is also the same as:

$$\sqrt{\alpha}_i = \frac{Var_{ij}}{\mu_{ij}} - 1$$

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

Generalized Linear Models

Learning objectives:

- Understand why and how to use generalized linear models to quantify differential gene expression

GLM for DEG

- Models Gene Expression using **Generalized Linear Models**
- Assumes counts are sampled from Negative-binomial distribution

Dispersion and Variability

The observed quadratic mean variance trend has motivated the use of the negative binomial distribution to model (bulk) RNA-seq gene expression data.

$$\begin{cases} Y_{gi} & \sim & NB(\mu_{gi}, \phi_g) \\ \log \mu_{gi} & = & \eta_{gi} \\ \eta_{gi} & = & \mathbf{X}_i^T \beta_g + log(O_{gi}) \end{cases}$$

coefficients β_g are log fold changes (with log link), tests on β_g tell us DE.

with

$$\text{var}[Y_{gi}] = \mu_{gi} + \phi_g \mu_{gi}^2$$

		Seq. technology		real expression
total variability	=	technical variability	+	biological variability
$\text{var}[Y_{gi}]$	=	μ_{gi}	+	$\phi_g \mu_{gi}^2$
total CV^2	=	$\frac{1}{\mu_{gi}}$	+	ϕ_g

From Berge and Clement

Coding
Notebook in R

Acknowledgements and References

- **Past MiCM slides: Intro to RNA-seq and Statistics in R (Adrien Osakwe)**
- **QLSC600 slides: myself and Megan Ng**
- **RNA-seq lecture by Peter N. Robinson**
- **Tutorial from Berge and Clement:** **https://statomics.github.io/SGA/sequencing_countData.html**

- **Sample data from**

Scheckel C, Drapeau E, Frias MA, Park CY et al. Regulatory consequences of neuronal ELAV-like protein binding to coding and non-coding RNAs in human brain. Elife 2016 Feb 19;5. PMID: 26894958