Differential Test with Generalized Linear Model

MCB 595A Genomics Journal Club

Jiacheng Ding

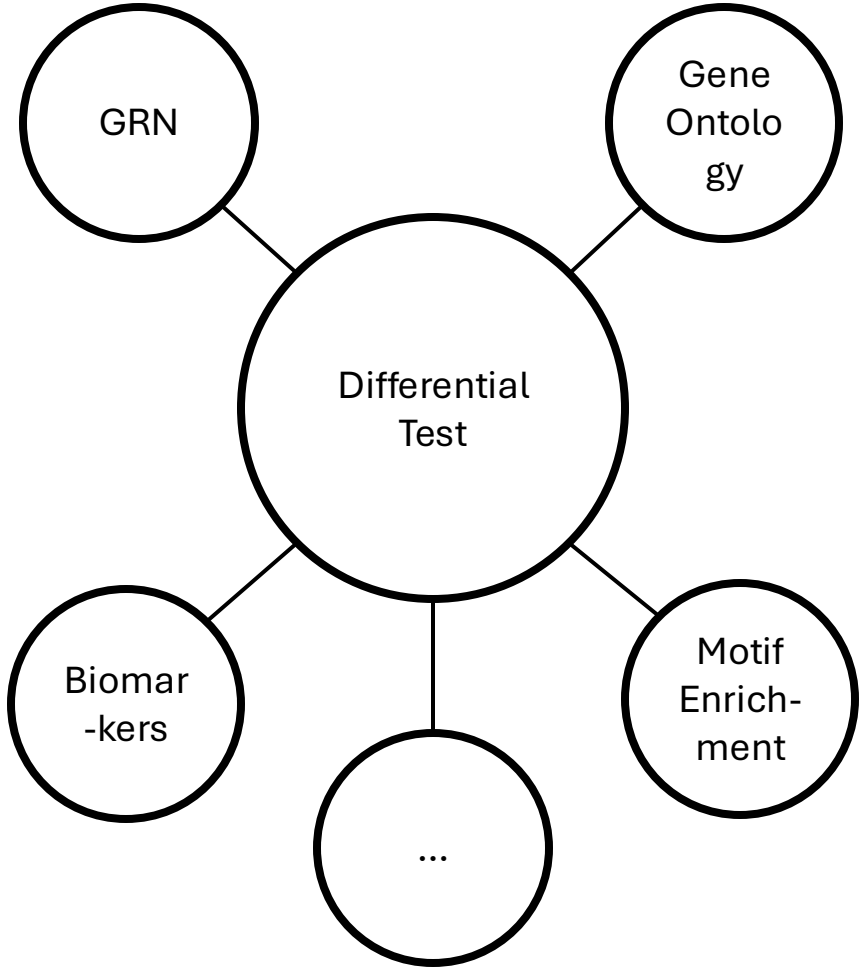
Cusanovich Lab

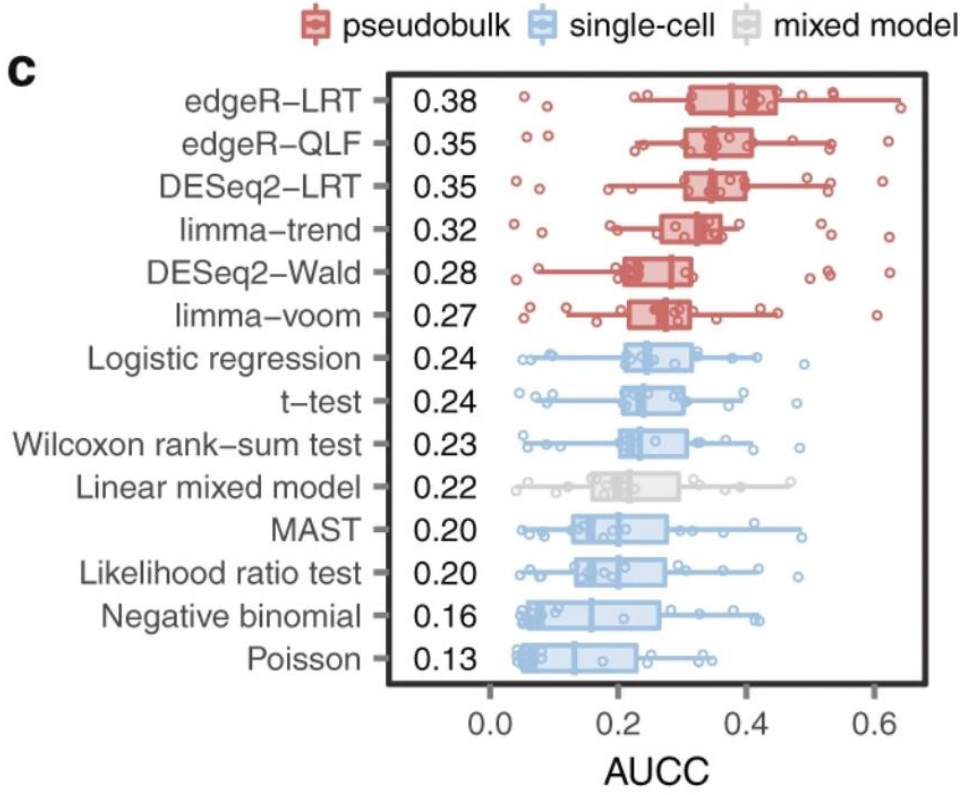
April 14th, 2025

An edgeR based discussion
edgeR tutorial - https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf
limma paper - lhttps://pmc.ncbi.nlm.nih.gov/articles/PMC4402510/
edgeRv4 paper - https://academic.oup.com/nar/article/53/2/gkaf018/7973897

Center of quantification study: differential test

Differential testing is at the heart of most quantification-based genomics studies, not only for traditional bulk genomics assay, but also for single-cell genomics assay..





Squair, J. W., et al. "Confronting false discoveries in single-cell differential expression. Nat. Commun. 12, 5692." 2021,

Murphy, Alan E., Nurun Fancy, and Nathan Skene. "Avoiding false discoveries in single-cell RNA-seq by revisiting the first Alzheimer's disease dataset." Elife 12 (2023): RP90214.

Murphy, A. E., and N. G. Skene. "A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. Nat Commun. 2022; 13: 7851."

What’s happening under the hood of differential test?

Have you encountered:

1) I observed grouping effects on PCA plots / heatmap, however only a handful of features tested significant.

2) I had one sample in a group, and I want do differential test.

3) I saw genes with very high log fold changes; however, p-values were not significant.

4) Differential test software asked me to plot a bunch of plots, however I don’t understand.

…

What’s going on under the hood of differential test?

Quantification test relies on underlying model / assumptions

Quantification test hinges on proper model predicting what is true, because:
1) Raw data is **noisy**.
2) Models give expectation and a way to measure **surprises** (variations).

In RNA-seq word, we now know gene expression counts are often model as negative binomial distribution, however…

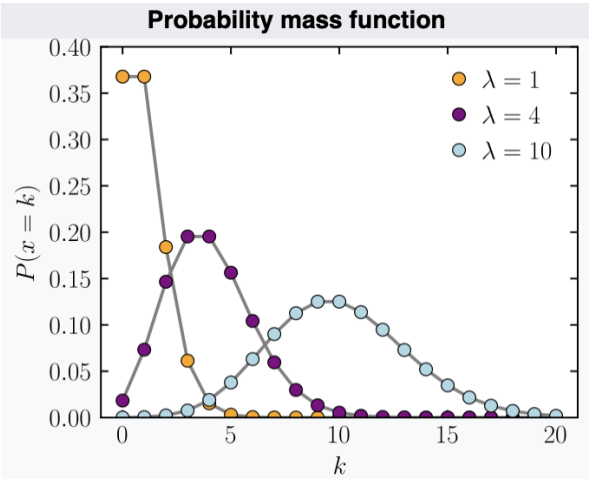
$$X \sim Poisson(\mu)$$

$$Var(X) = \mu$$

There is a **rationale** behind modeling RNA-seq counts with **Poisson** distribution (especially for experiments with only technical replicates).

If we consider sequencing is a process of sampling molecules from a mixed soup of RNA, the only thing contributing to sampling probability is the relative abundance of those RNA molecules.

Under this assumption, the only source of randomness is the sampling process itself, and the number of reads assigned to a gene follows a Poisson distribution.



Wikipedia: Poisson distribution

Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinform. 2010;11(1):1–13.

Real life is never perfect

Real life RNA-seq data is much nosier, because:

1) Amplification noise.
2) Sequencing bias (e.g., clustering efficiency).
3) Co-expression.
….

$X \sim Poisson(\mu)$

$Var(X) = \mu$

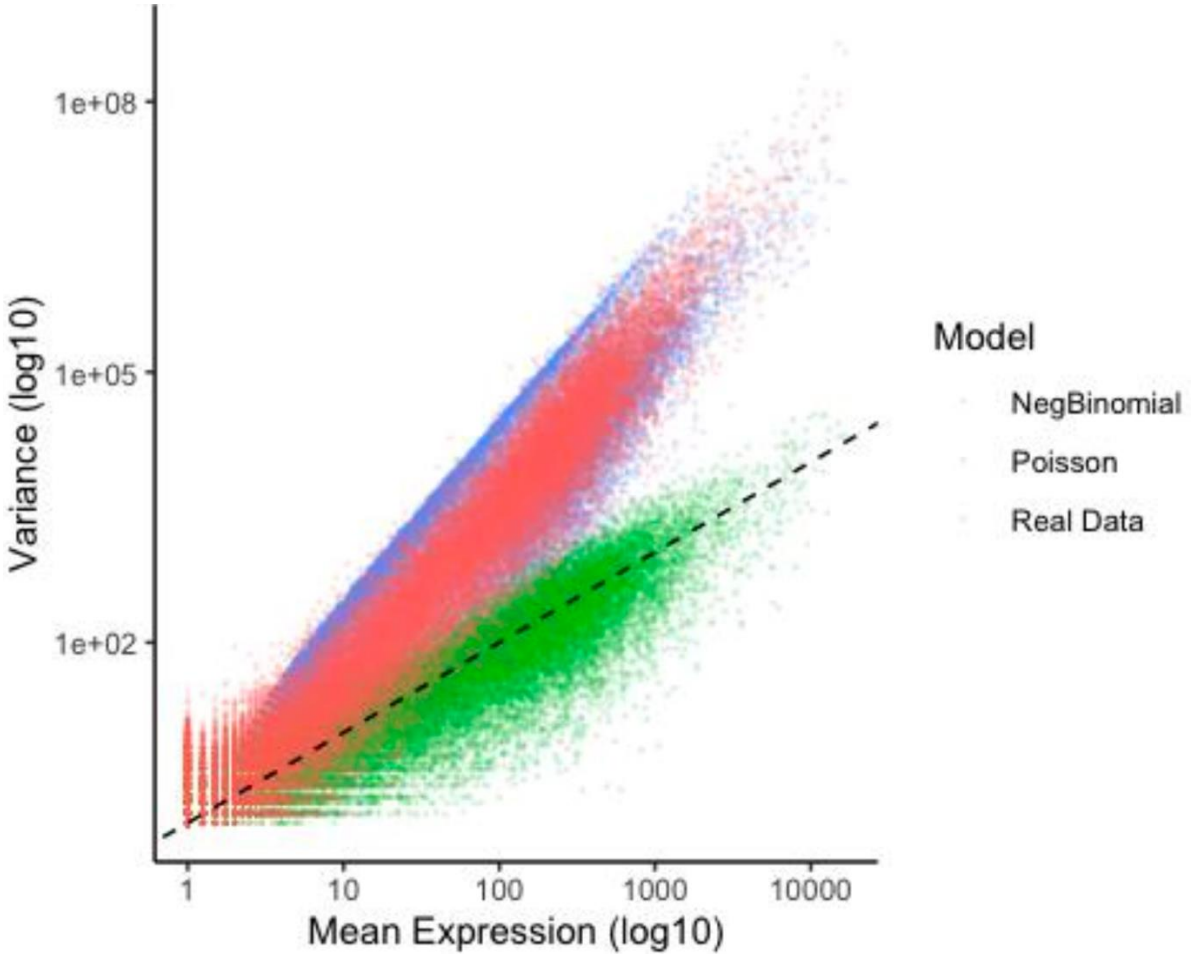
$Y \sim NB(r, p)$ *or equivalently* $Y \sim NB(\mu, \varphi)$

$\mu = r \cdot (\frac{1-p}{p})$

$Var(Y) = \mu + \varphi \cdot \mu^2$

overdispersion / mean-variance trend

$\varphi = \frac{1}{r}$, often referred as dispersion



Variance & dispersion

When people talk about a distribution, we often hear mean, median, **variance**, **variation**, **deviation** ,etc.…

$\boldsymbol{Variance} = \frac{1}{n} \cdot \sum (x_i - \bar{\boldsymbol{x}})\text{^}2$, while **variation** is an intuitive term descripting how **noisy** a data is.

$$Y = NB(r, p) \; or \; equivalently \; Y = NB(\mu, \varphi)$$

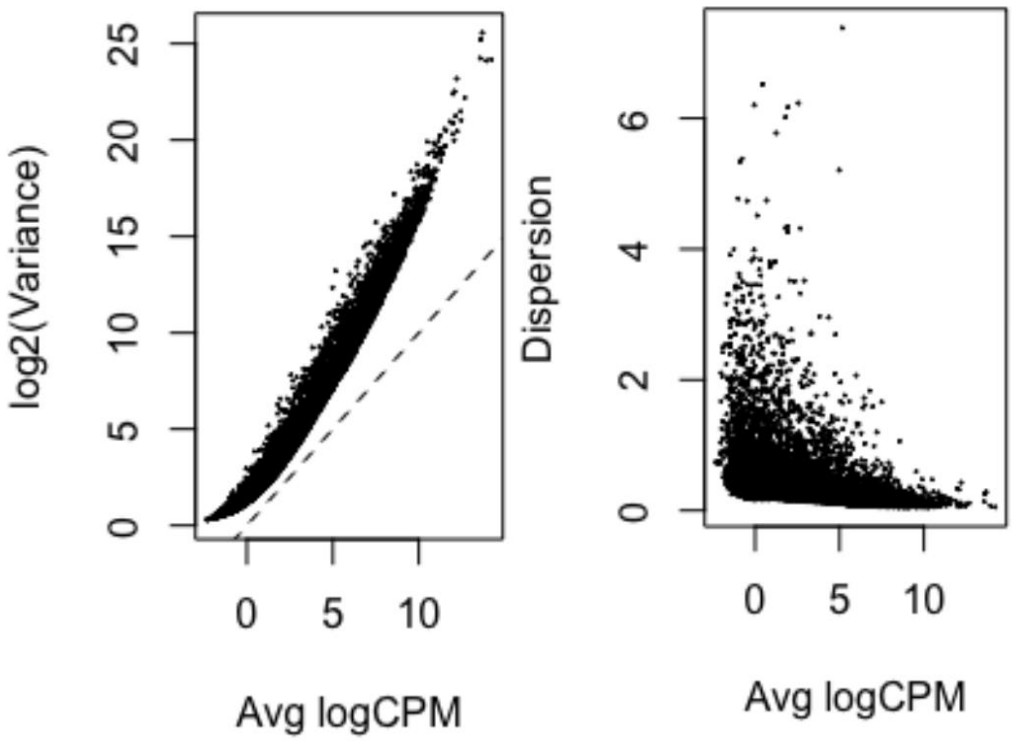
$$Var(Y) = \mu + \varphi \cdot \mu\text{^}2$$

Variance of unit of (counts)^2

variation, unitless

If we are really strict, quantitative form of variation should be:

$$CV\% = \left(\frac{\delta}{\mu}\right) \cdot 100, where \; \delta = \text{sqrt(Variance)}$$



Coefficient of variance

In negative binomial distribution, coefficient of variance is:

$$CV^2 = \frac{1}{\mu} + \varphi$$

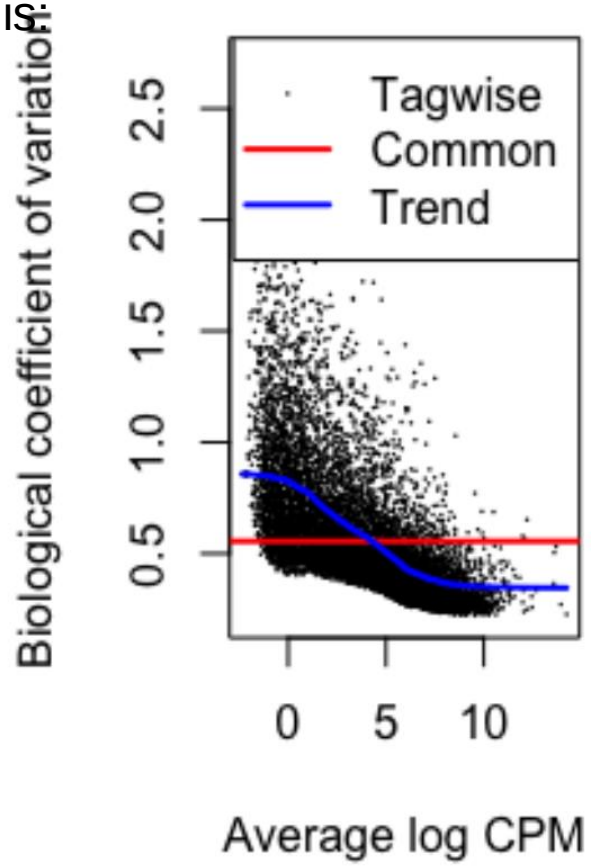
The above equation can be interpreted as:

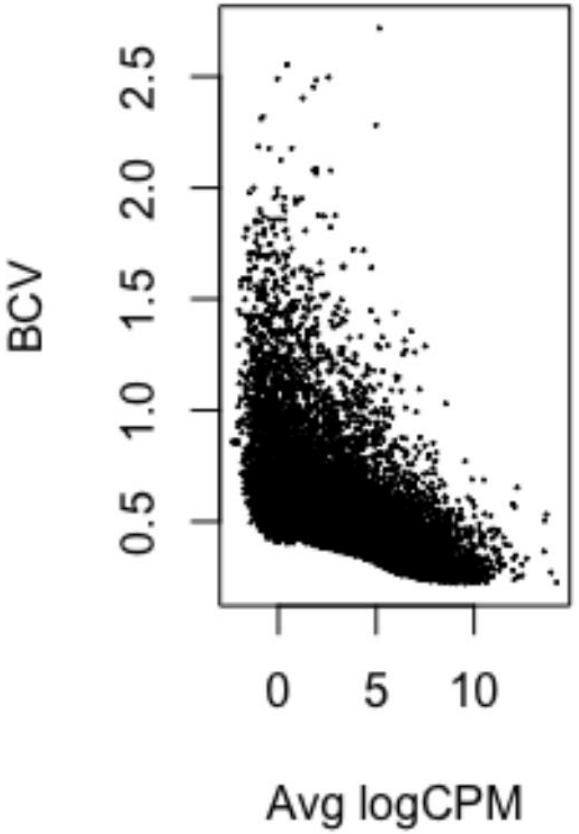
$$Total\ CV^2 = Technical\ CV^2 + Biological\ CV^2$$

Noise from Poisson model (ideal model)

Additional noise (real world correction)

$$Biological\ CV^2 = \varphi$$





Normalization

What's the goal of normalization?

Normalization: to make **non-differentially** expressed **genes** have **similar** expression **values** across samples, and **true** biological **differences stand out**.

$$log(\mu) = log(libSize \cdot normFactor) + \beta$$

Expression value adjustment
Often referred as normalization offset / effective library size

where expression adjustment happens

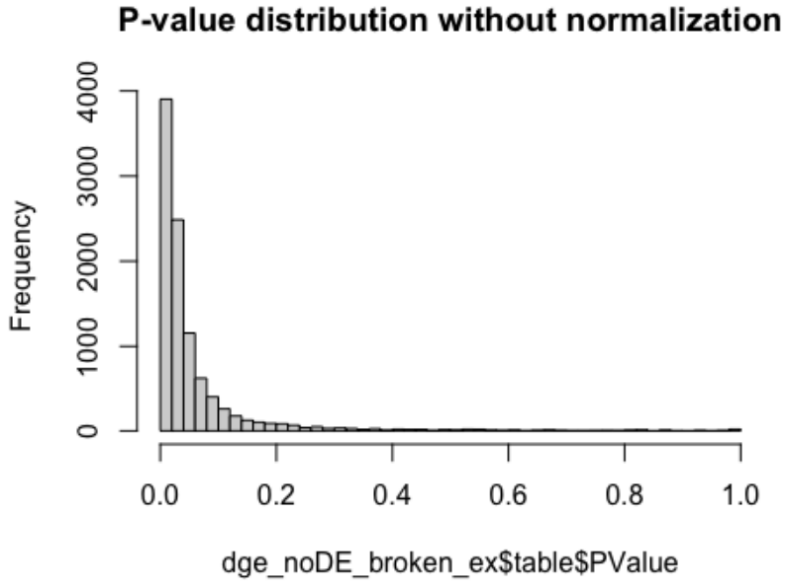
How is normalization factor $(normFactor)$ determined (TMM as an example):

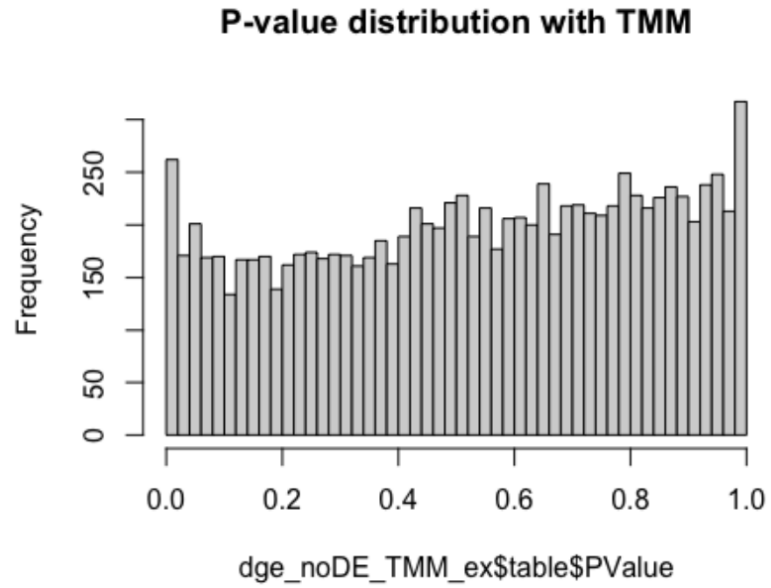
1) Filter lowly expressed genes.
2) Choose a reference sample (median if not assigned).
3) Calculate M-value (observed log fold change between test and ref) and A-value (log average expression value).
4) **Exclude genes with extreme M-value and A-value**.
6) Compute normFactor = 2^ (mean of M)

Reflects the assumption that most of genes are not differential!

What if I don’t normalize my data?

Matrices	Simulation_1	Simulation_2
N condition	2	2
N Samples	4	4
N genes	10,000	10,000
N true DEs	0	0
Lib Size control	1e7	1e7
Lib size treatment	1e8	1e8
Normalization	None	TMM
Detected Des (p-va < 5%)	7,053	538
Detected Des (FDR < 5%)	320	13





Please refer R script for an extension about **batch effects**

Classic linear model (CLM)

Classic linear model (CLM):

$$\begin{cases} Y = \beta_0 + \beta_1 \cdot X_i + r \\ Y \sim N(\beta_0 + \beta_1 \cdot X_i, \delta) \end{cases}$$

mean

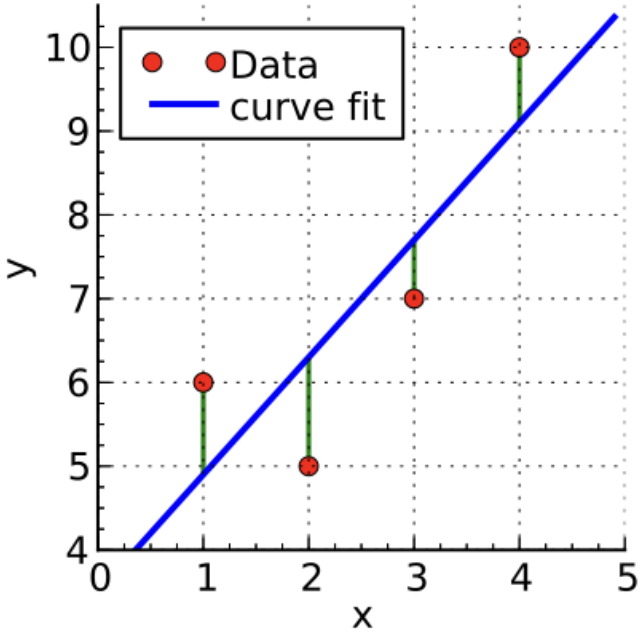
$Var(Y) = \delta^2$, *constant across genes*

$r = Y - Y_i$, with r $\sim$N$(0, \delta)$ *residual r is used to estimate* δ

How spread β_1 is (uncertainty)

Doesn't account for mean-variance relationship / overdispersion

Y *is epected expression value*
Y_i is observed expression value
β_0 is baseline expression (control)
β_1 is the effect of being covariate X_i
X_i is the i_{th}covariate
r is residual
δ is standard deviation



Wikipedia: linear regression

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

Generalized liner model (GLM)

Generalized linear model (GLM via NB):

$$\begin{cases} Y \sim Negative\ Binomial(\mu, \theta) \\ \log(\mu_i) = \beta_0 + \beta_1 \cdot X_i + \log(offset) \\ offset = libSize \cdot normFactor \end{cases}$$

link function

$$, with\ Var(Y) = \mu + \varphi \cdot \mu^2, varies\ across\ genes\ with\ different\ mean$$

$Y\ is\ epected\ expression\ value$

β_0 is baseline expression (control)

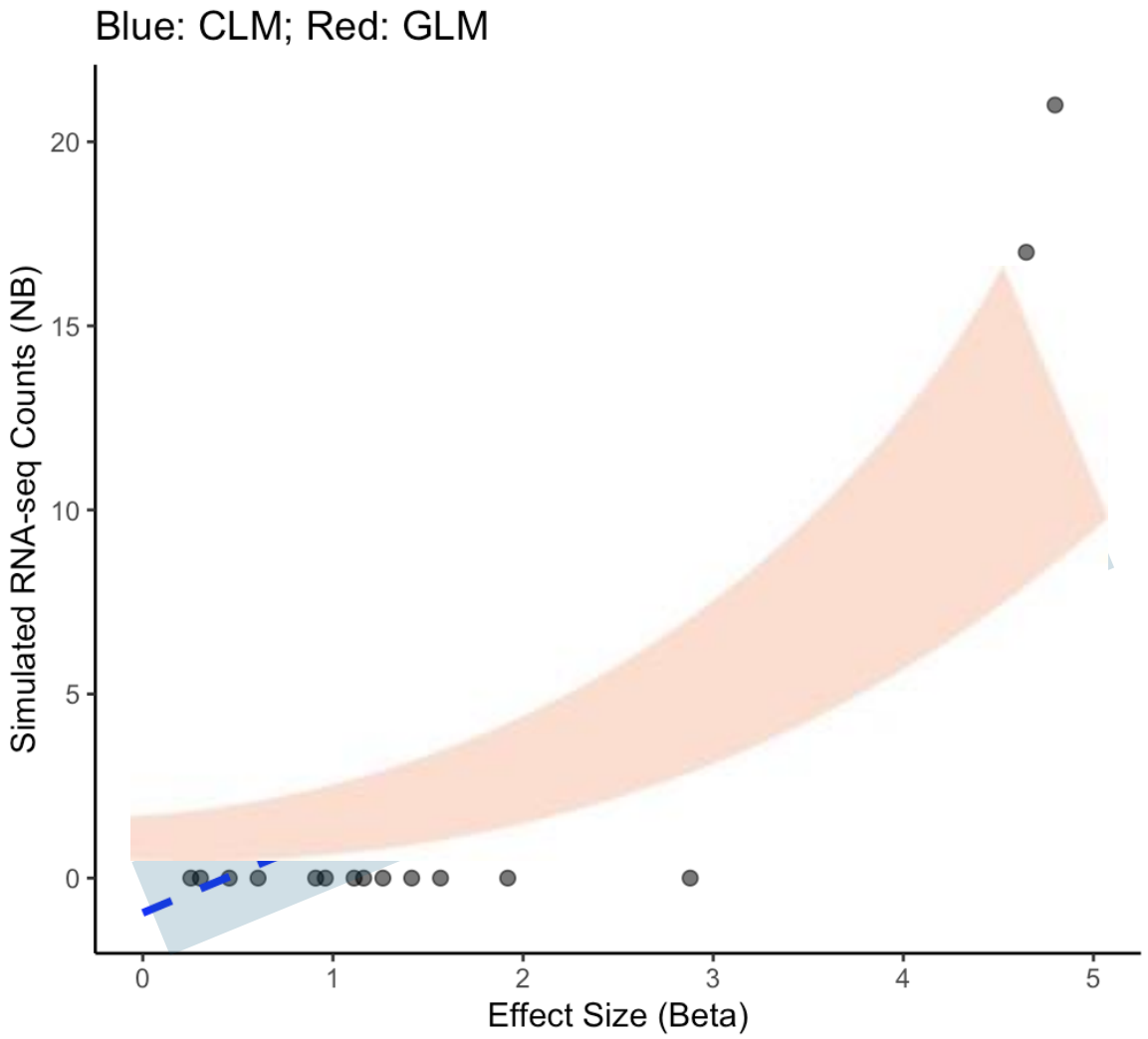
β_1 is the effect of being covariate X_i

X_i is the i_{th}covariate

You may notice there is no explicit residual in GLM, because variance is a function of mean, while CLM variance is an independent term.

However, it does NOT mean residual doesn't exist. It is implicit instead.

$$pearson\ rsidual = \frac{Yi\ - Y}{\sqrt{Variance}}$$



”A gene is differentially expressed”

Sampe1, Sample2, Sample3 – control – β_0
Sample4, Sample5, Smaple6 – treatment – β_1

	control	treatment
Sampe1	1	0
Sample 2	1	0
	1	0
Sample 3	0	1
	0	1
Sample 4	0	1
Sample 5		
Sample 6		

$$\times \begin{matrix}\beta_0\\ \beta_1\end{matrix} = \begin{matrix}\beta_0\\ \beta_0\\ \beta_0\\ \beta_1\\ \beta_1\\ \beta_1\end{matrix}$$

6 X 2 %*% 2 x 1 = 6 x 1

X = model.matrix(~0 + treatment) # no intercept
More flexibility – draw any comparisons

Is β_1 significantly different from 0?

	Intercept	treatment
Sampe1	1	0
Sample 2	1	0
	1	0
Sample 3	1	1
	1	1
Sample 4	1	1
Sample 5		
Sample 6		

$$\times \begin{matrix}\beta_0\\ \beta_1\end{matrix} = \begin{matrix}\beta_0\\ \beta_0\\ \beta_0\\ \beta_0 + \beta_1\\ \beta_0 + \beta_1\\ \beta_0 + \beta_1\end{matrix}$$

6 X 2 %*% 2 x 1 = 6 x 1

X = model.matrix(~treatment) # with intercept (common baseline)

fit.test <- glmQLFtest(fit, coef = 2)

Two Pillars of GLM via NB

Two pillars of GLM with NB model are: **dispersion** and **coefficients**

y <- calcNormFactors(y, method = 'TMM') → Estimate normalization factor
y <- estimateDisp(y, tagwise = T) → Estimate dispersion φ
fit <- glmQLFit(y) → Estimate coefficient β_1
fit.test <- glmQLFtest(fit) → Do statistic test

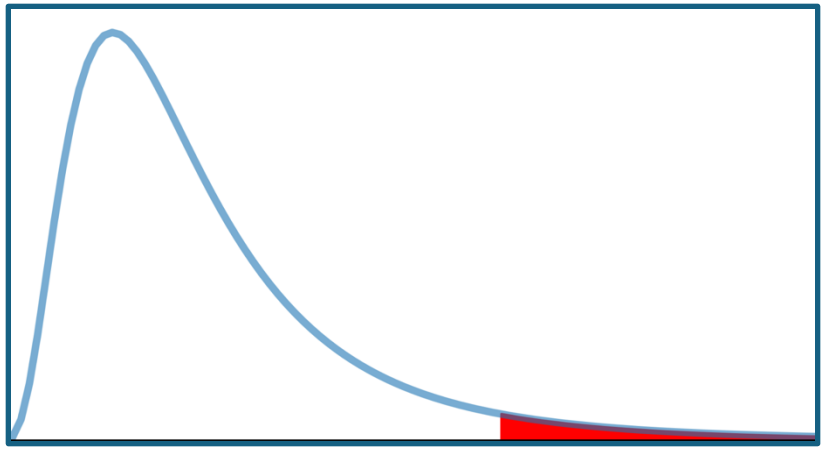
$$\begin{cases} Y \sim Negative\ Binomial(\mu, \theta) \\ \log(\mu_i) = \beta_0 + \beta_1 \cdot X_i + \log(offset) \\ offset = libSize \cdot normFactor \end{cases}$$

$Log(\mu) = \log(libSize \cdot normFactor) + \beta_1 \cdot X_i, \beta$ is effect size

$Var(\beta) = \mu + \theta \cdot \mu^2$, $hence\ SE(\beta) \propto sqrt(\varphi)\ is\ uncertainty$

$$test\ statistic = \frac{Effect\ Size}{Uncertainty}$$

FDR is enough for determining significance. LogFC > 0.5 is redundant!



Wikipedia: F test

It’s fun to reconstruct GLM with internalized knowledge

One function is all we need for reconstructing the generalized linear model:

$$Y_i \sim Y = 2\hat{}(\beta_0 + \beta_1 \cdot X_i + \log(libSize_i \cdot normFactor_i))$$