

# Differential Expression Analysis with RNA-SEQ: Review and Critique

David M. Rocke  
Division of Biostatistics and  
Department of Biomedical Engineering  
University of California, Davis

# Outline

- RNA-Seq for differential expression analysis
- Statistical methods for RNA-Seq: Structure and choices
- Criteria for comparison of statistical methods
- Results of three packages on a data set
- Null performance
- Source of differences in the results
  - Input data
  - Tests
  - Variance estimation
  - Normalization
- Conclusions

# RNA-Seq

- Gene expression is the transcription of the DNA in a gene into mRNA, which (in many cases) is later translated into a protein.
- We can measure expression of a single gene with PCR or other assays.
- Gene expression arrays measure expression of many genes simultaneously using spots each of which contains a matching sequence to the gene sequence to be detected.
- But this can only detect what we already suspected might be there.

# RNA-Seq

- For RNA-Seq, the RNA in the sample is reverse transcribed into the corresponding DNA sequence.
- Then the DNA fragments are sequenced (in an NGS sequencer, usually Illumina )
- Each fragment is mapped to the reference genome
- The data to be analyzed are the number of fragments mapping to each gene in a table where the columns are samples and the rows are genes.

# RNA-Seq

- This mapping can be complex
- We can choose to estimate isoforms or not
- We can choose how to handle ambiguous reads (omit or spread across genes)
- We can then use statistical analysis to determine when there is significantly more expression in one condition or another.
- This may or may not be better than an expression array depending on goals.

# Analysis of RNA-Seq Data

- For each gene/exon/isoform (we will say gene from now on), and for each sample, we have a count of fragments mapping to that gene.
- In principle, we need to test whether the counts from one group are significantly larger than another.
- Or we may have more than one factor or variable that could be associated.
- In practice, we may (probably) need to normalize the samples first, and may need to import some information across genes.

# Existing Methods

- Existing methods often contain complex combinations or filtering, normalization, transformation, and variance estimation before any statistical tests are performed.
- The methods are often poorly documented and change rapidly and often substantially between frequent versions, without any push notice.
- The results of different methods such as DESeq, edgeR, and voom/limma can differ substantially.
- It appears that most methods produce large numbers of false positives.

# Bottomly Data

- Two inbred mouse strains, C7BL6/6J (10 animals) and DBA/2J (11 animals)
- Gene expression in striatal neurons by RNA-Seq and expression arrays.
- 36,536 unique genes of which 11,870 had a count of at least 10 across the 21 samples.
- This data set has been used in other comparisons
- First, we compare three methods on this data set.
- Later we construct data sets where the null hypothesis is true.



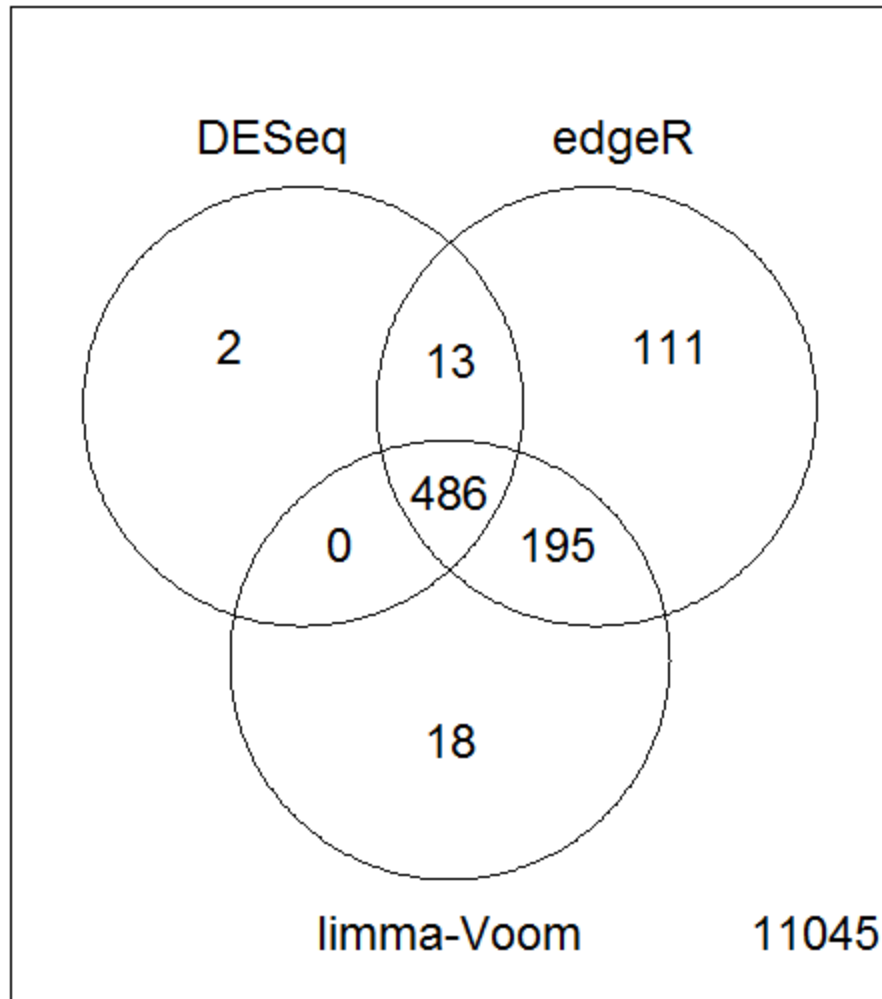
## Bottomly Data Significant Genes at $p = 0.001$

825 Genes Significant  
by at least one  
method

486 by all methods

More “significant”  
genes by edgeR

Is this greater power  
or is the test  
producing false  
positives?



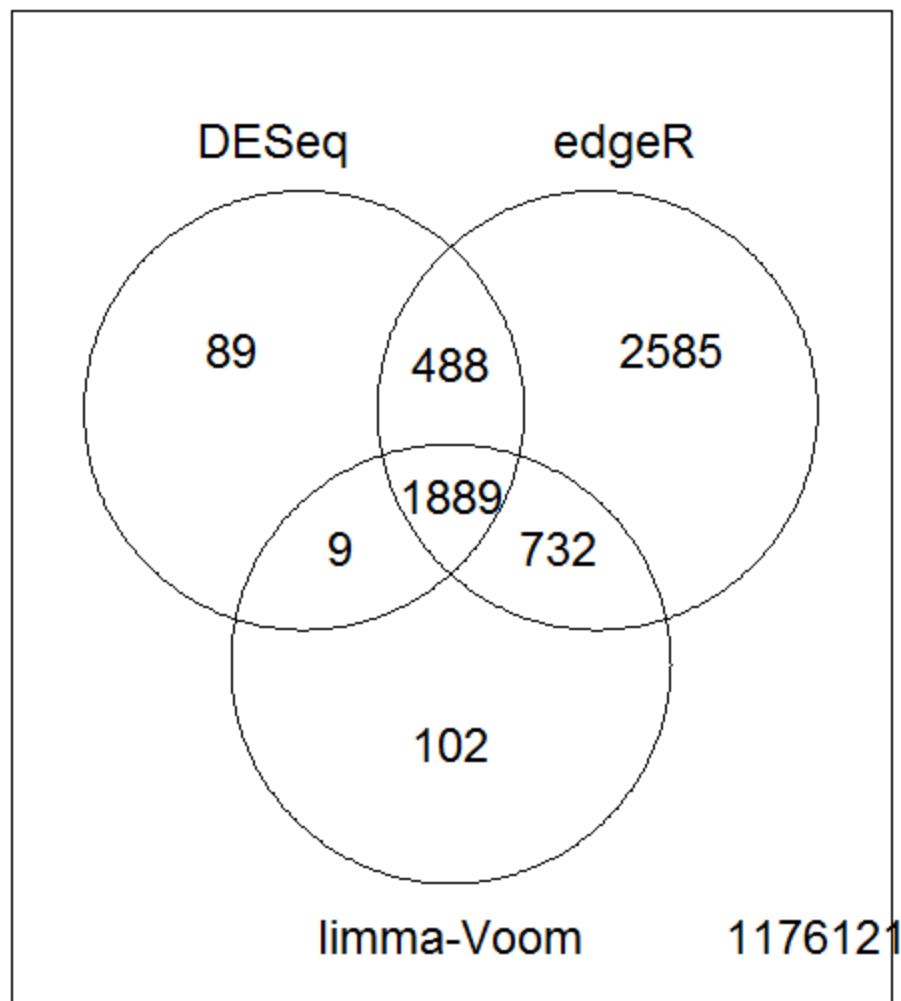
# Null Behavior

- 100 randomly selected subsets of size 6 out of the 10 C7BL6 mice (out of the 210 possible)
- In each subset, 3 assigned to treatment and 3 to control at random (out of 10 possible divisions).
- The null hypothesis is on the average true.
- Given the 11,870 genes with large enough counts, we would expect about 0.1% to be significant at the  $p = 0.001$  level.
- We have 1,187,000 tests, so there should be about 1187 rejections for each method

Significant at  $p = 0.001$   
Null Hypothesis is True

Estimator	Observed	Expected	Over-Rejection
DESeq	2475	1187	209%
edgeR	5694	1187	480%
limma-Voom	2732	1187	230%

## Null Test Significant Genes at $p = 0.001$



1187 Expected

# Important Factors

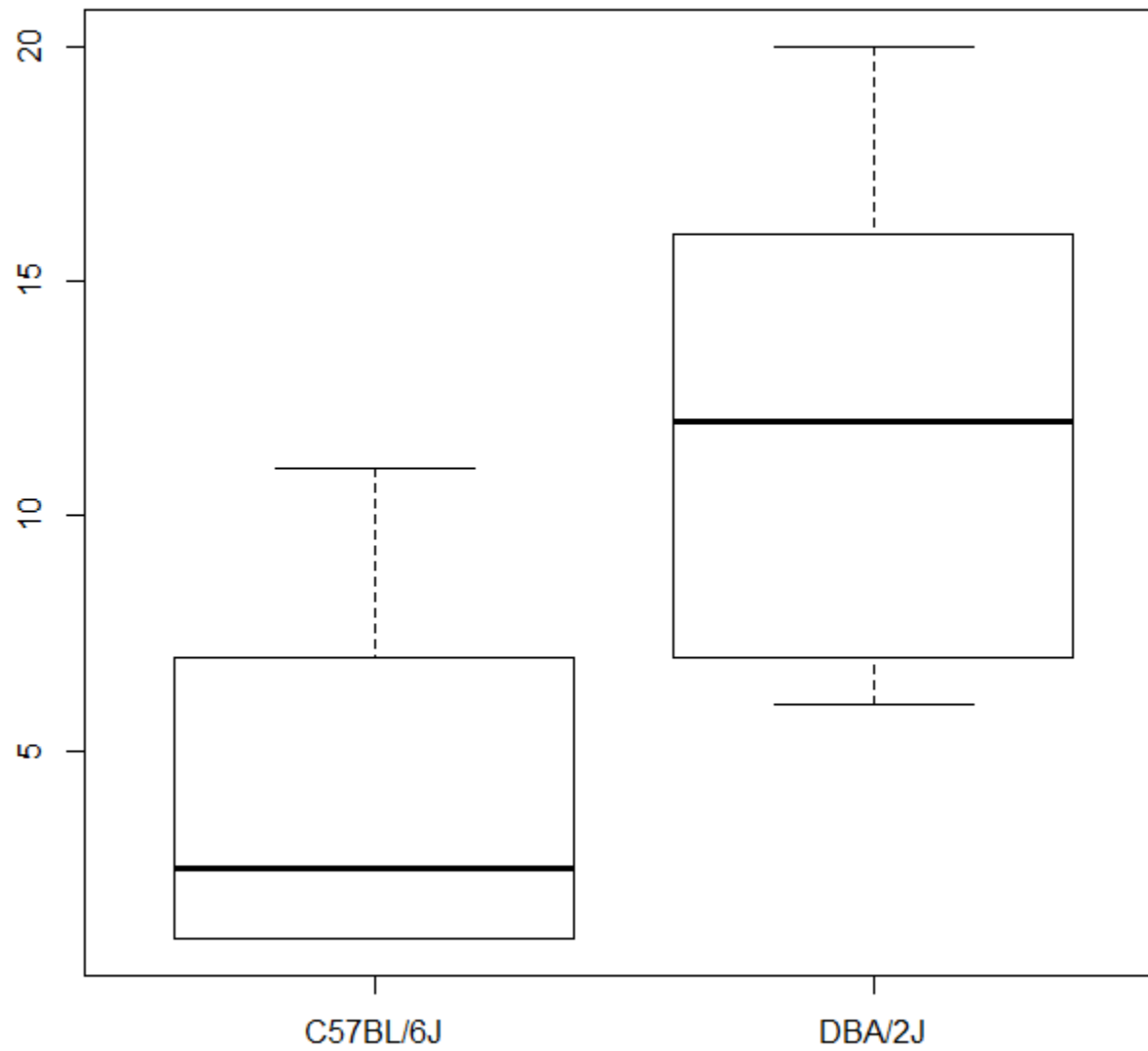
- These methods share some features but differ in their implementation.
- All of them treat the data as over-dispersed Poisson, or specifically negative binomial
- They may use a test based on the negative binomial distribution, but limma-Voom uses standard regression/anova with variance-based weights
- Since sample sizes tend to be small, they all have the possibility to replace the variance estimate from a given gene with a smoothed estimate based on all the genes.
- Normalization may be needed due to the differing total numbers of reads

# Factors That May Cause False Positives

- Normalization procedure
  - Normalization driven by high-abundance transcripts can cause spurious significance of low abundance ones.
- Variance estimation
  - Many methods “borrow strength” from other genes to get “better” variance estimates.
  - This may cause bias
- Statistical test used
  - Many choices
- Nature of the data
  - May require filtering

# Statistical Test

- Example gene from Bottomly data
  - ENSMUSG00000042638
  - C57BL/6J
  - 2, 7, 9, 11, 3, 7, 1, 1, 2, 1
  - Median 2.5, Mean 4.4, sd 3.75, var 14.04
  - DBA/2J
  - 8, 6, 12, 10, 19, 15, 17, 6, 20, 6, 14
  - Median 12, Mean 12.09, sd 5.28, var 27.89





# Poisson Analysis

- One of the first ideas was that the counts might be Poisson distributed and that we could use this for statistical tests.
- In this case, if the null hypothesis was true, then the 21 observations came from the same Poisson distribution with observed mean 8.429 and with variance also estimated at 8.429 (Poisson distributions have mean and variance that are equal).
- This implies that the variance of the mean of the 10 C7BL/6J counts should be  $8.429/10$  and the variance of the mean of the 11 DBA/2J counts should be  $8.429/11$
- So we can use  $z = (4.40 - 12.09)/\sqrt{(8.429/10 + 8.429/11)} = -6.06$  to test the difference

# Poisson Analysis

- This turns out to be a terrible idea.
- Frequently, the variance of the data is much larger than the mean. We can call this *overdispersion*.
- C57BL/6J
  - Mean 4.4, Variance 14.04
- DBA/2J
  - Mean 12.09, Variance 27.89
- Also, Poisson analysis can be done with one treatment and one control, and any belief that statistical evidence can be gathered with no biological replicates is evidence of terminal delusion!

# Over-dispersed Poisson Analysis

In this case, we estimate the means and the variances from the data using either the negative binomial distribution or simply an over-dispersed Poisson.

These yield similar results.

In either case, we model the data in such a way that, for a given gene  $g$  and /sample  $i$ , the count  $y_{ig}$  has

$$E(y_{ig}) = \mu_{ig}$$

$$\text{Var}(y_{ig}) = \mu_{ig} + c^2 \mu_{ig}^2$$

```
> summary(glm(y1~strain,family=poisson))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	1.4816	0.1508	9.828	< 2e-16	***
strainDBA/2J	1.0108	0.1739	5.812	<b>6.16e-09</b>	***

(Dispersion parameter for poisson family taken to be 1)

```
> anova(glm(y1~strain,family=poisson),test="Chisq")
```

	Df	Deviance	Resid.	Df	Resid. Dev	Pr(>Chi)
NULL			20		90.976	
strain 1	1	38.778	19		52.198	<b>4.749e-10</b> ***

**First test is Wald test, second is likelihood ratio test.  
But Poisson assumption is not valid.**

```
> summary(glm(y1~strain,family=quasipoisson))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	1.4816	0.2489	5.952	9.94e-06	***
strainDBA/2J	1.0108	0.2871	3.520	<b>0.00229</b>	**

(Dispersion parameter for quasipoisson family taken to be 2.72605)

```
> anova(glm(y1~strain,family=quasipoisson),test="Chisq")
```

	Df	Deviance	Resid.	Df	Resid. Dev	Pr(>Chi)
NULL			20	90.976		
strain	1	38.778	19	52.198	<b>0.0001622</b>	***

```
> summary(glm.nb(y1~strain))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	1.4816	0.2052	7.219	5.24e-13	***
strainDBA/2J	1.0108	0.2594	3.897	<b>9.73e-05</b>	***

(Dispersion parameter for Negative Binomial(5.1558) family taken to be 1)

```
> anova(glm.nb(y1~strain),test="Chisq")
```

	Df	Deviance	Resid.	Df	Resid. Dev	Pr(>Chi)
NULL			20	38.657		
strain	1	15.437	19	23.220	<b>8.53e-05</b>	***

# P-values ( $\times 10^5$ )

	Wald	LR
Quasi-Poisson	228.72	16.22
Negative Binomial	9.73	8.53

- Both the Wald test and the likelihood ratio test are valid asymptotically.
- The quasi-Poisson and negative binomial should be equivalent asymptotically
- But in finite samples there can be substantial differences

```
> t.test(y1~strain)
```

```
Welch Two Sample t-test
```

```
t = -3.8746, df = 18.007, p-value = 0.00111
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-11.860987 -3.520831
```

```
sample estimates:
```

```
mean in group C57BL/6J    mean in group DBA/2J
```

```
4.40000
```

```
12.09091
```

**Similar results to that using the negative binomial.**



# Is the Test Statistic Important?

## Not so much

- Simulated negative binomial count data
- Two groups of 10
- Mean 5 and variance 6
- 10,000 trials

Test	Reject at 5%	Reject at 1%	Reject at 0.1%
t-test	4.9%	0.90%	0.11%
Wilcoxon Test	4.6%	0.74%	0.03%
NB Likelihood Ratio Test	5.2%	0.95%	0.13%
NB Wald Test	5.7%	1.16%	0.22%
	4.8% – 5.2%	0.90% – 1.11%	0.07% – 0.12%

# Normalization

- Normalization is commonly used in differential expression analysis with microarrays.
- Normalization for RNA-Seq is often couched in terms like “library size” as if we should divide by the total (mapped) fragment count
- This can be problematic because it depends on only a few genes.
- Mostly, normalization in RNA-Seq is done with a single constant per sample, though this is unusual with expression arrays

# Library Size Normalization

- Using the total fragment count is problematic because highly expressed genes will provide most of the fragments
- Four genes with expression 10,000, 100, 150, 200 in condition A and 20,000, 100, 150, 200 in condition B.
- Normalized fragment counts use total fragment counts of 10,450 and 20,450 and can be normalized to 15,450
- Normalized fragment counts are 14,785, 148, 222, 296 in condition A and 15,110, 76, 113, 151 in condition B, so up-regulation of gene 1 has been turned into down-regulation of the other three.
- Fold changes are 2.0, 1.0, 1.0, 1.0 before “normalization” and 1.02, 0.51, 0.51, 0.51 after.

# Normalization Methods

- Total count
- Quantile Normalization of other signal based methods
- Geometric normalization (Cuffdiff2/DESeq version)
  - For each gene, compute the geometric mean of the total fragment count across libraries
  - Library “size” is the median across genes of the total fragment count divided by the geometric mean fragment count.
  - In our 4-gene example, the geometric means are 14,142, 100, 150, 200, the ratios for A are 0.707, 1, 1, 1 and for B are 1.414, 1, 1, 1, so the size factors are the medians, namely 1 and 1

$k_{ij}$  fragment count for gene  $i$  in library  $j$

$$g_i = \left( \prod_{v=1}^m k_{iv} \right)^{1/m} = \exp \left( \frac{1}{m} \sum_{v=1}^m \log(k_{iv}) \right)$$

$$s_j = \text{median}_i \frac{k_{ij}}{g_i}$$

- Cuffdiff2 first normalizes replicates under the same conditions giving an *internal* library size of  $s_j$
- Then the arithmetic mean of the scaled gene counts for each gene is used to compute an *external* library size of  $\eta_j$ .
- This is a possible source of problems, the scale of which is unknown

# Variance Estimation

- Many RNA-Seq experiments are small.
- Small studies have low power
- Very small p-values are needed to pass the false discovery filter
- So the default for small studies is no results
- Suppose two means differ by one standard deviation
- With 10,000 genes, the Bonferroni level is  $5 \times 10^{-7}$
- With two groups of 3, there are  $\sim 4$ df and  $5 \times 10^{-7}$  corresponds to almost 50 standard deviations

# Variance Estimation

- If we use tests that reference only the data from the specific gene, then usually variance estimation is not a problem.
- But with small sample sizes, the power is low, so there is a temptation to “improve” the variance estimates by smoothing or shrinkage.
- The variance of a negative binomial increases with the mean in a way controlled by a variance parameter.
- We can smooth the plot of the sample variance vs. the mean and use the smoothed estimate instead of the per-gene estimate.

A Poisson random variable with parameter  $\lambda$  has

$$\mu = \lambda$$

$$\sigma^2 = \lambda$$

A negative binomial random variable can be seen as a mixture of Poissons with varying  $\lambda$

$$\mu = \lambda$$

$$\sigma^2 = \lambda + c^2 \lambda^2$$

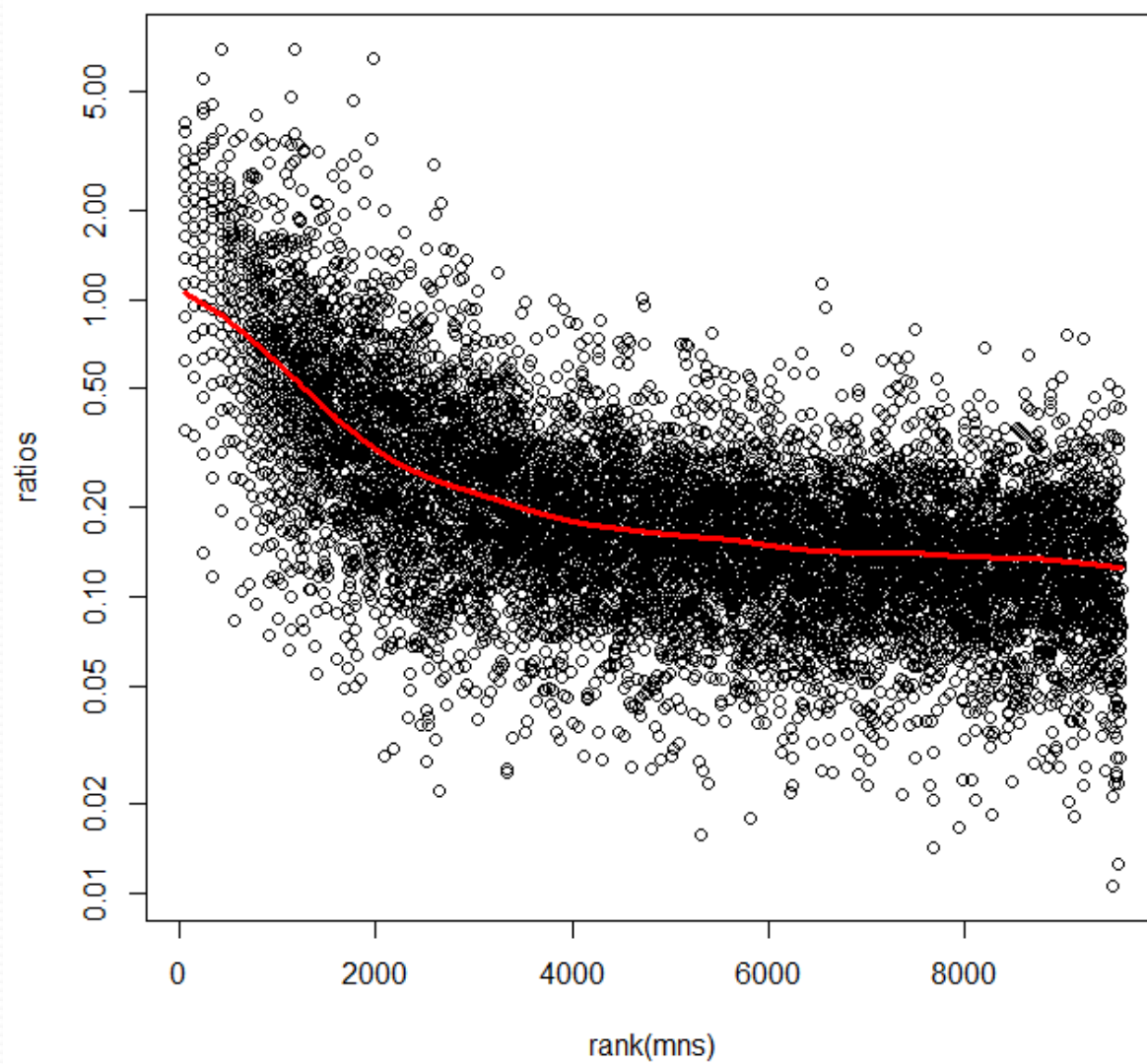
$$\sigma^2 / \mu^2 = \lambda^{-1} + c^2 \rightarrow c^2$$


$$\hat{c}^2 = \frac{s^2 - \bar{x}}{\bar{x}^2} \quad \text{if this is positive}$$

If we smooth a plot of the variance or the square CV or  $\hat{c}^2$  vs. the mean we obtain an average estimate of  $c$  for the given value of  $\mu$

We can use the individual variance, the smoothed variance, or a compromise






$$\mu = \lambda$$

$$\sigma^2 = \lambda + c^2 \lambda^2$$

$$c^2 = \frac{\sigma^2 - \lambda}{\lambda^2}$$

$$\hat{c}^2 = \frac{s^2 - \bar{x}}{\bar{x}^2} \quad \text{or 0 if negative}$$

$$\tilde{c}^2 = \alpha \bar{c}^2 + (1 - \alpha) \hat{c}^2 \quad \text{where } \bar{c}^2 \text{ is from the smoothed plot}$$

# Variance Estimation

- The sample variance is an unbiased estimate of the population variance
- A smoothed variance will be biased down or up depending on the data point
- While this can reduce the MSE of estimating the variance, it may increase false positives and false negatives for tests based on those variance estimates
- This is a possible source of the differences in results in various methods of analysis.

# Conclusions

- Analysis of RNA-Seq data is still in the early stages of development.
- Existing programs vary substantially in the results and it is unclear which is the most reliable
- More work is needed that focuses on the fundamental components of analysis, particularly variance estimation and sample normalization.