Differential analysis of gene regulation at transcript resolution with RNA-seq: Cuffdiff2

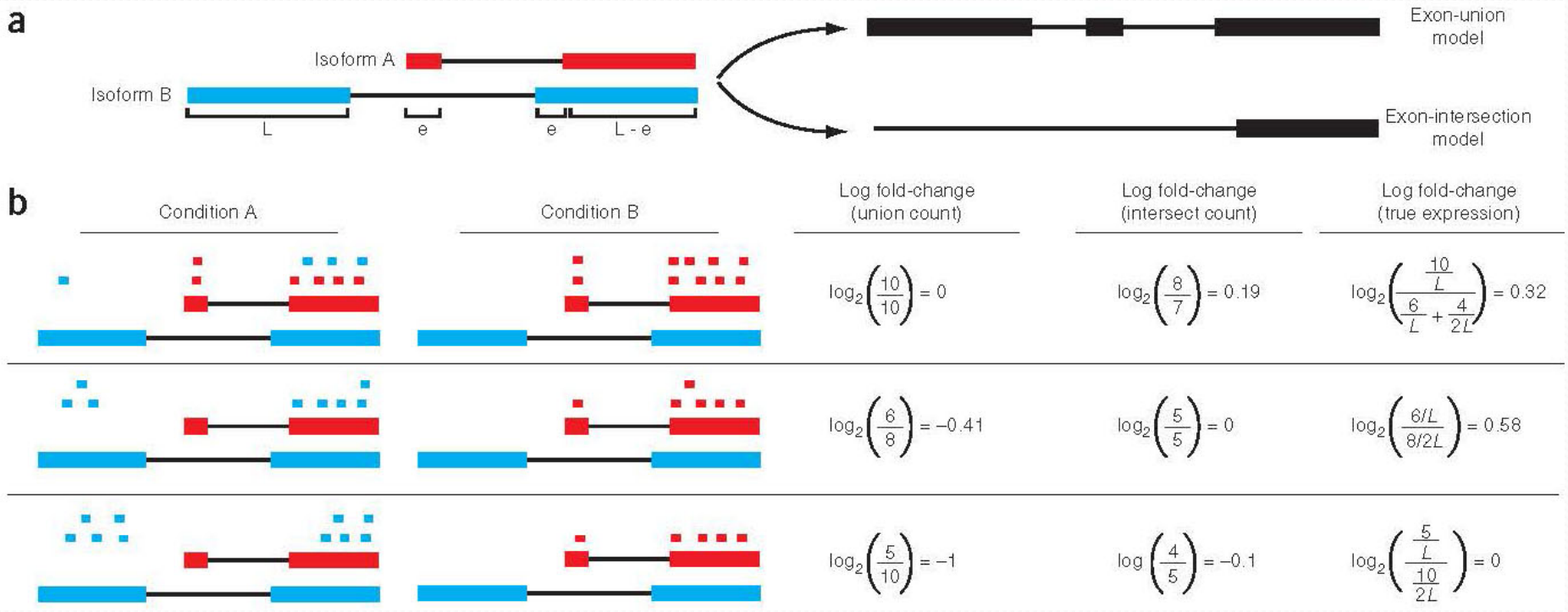
BST 226

Statistical Methods for Bioinformatics

David M. Rocke

Isoform Level Analysis

- DESeq and edgeR use raw counts mapped to genes as the element of analysis
- Cuffdiff tries to separate out expression of different isoforms, which they call the “true expression”
- Clearly, this has advantages (isoforms are potentially different in function) and disadvantages (isoform expression is not directly observable and has to be inferred from calculation.


a
Isoform A
Isoform B
L
e
e
L - e
Exon-union model
Exon-intersection model
b
Condition A
Condition B
Log fold-change (union count)
Log fold-change (intersect count)
Log fold-change (true expression)
$\log_2\left(\frac{10}{10}\right) = 0$
$\log_2\left(\frac{8}{7}\right) = 0.19$
$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L}+\frac{4}{2L}}\right) = 0.32$
$\log_2\left(\frac{6}{8}\right) = -0.41$
$\log_2\left(\frac{5}{5}\right) = 0$
$\log_2\left(\frac{6/L}{8/2L}\right) = 0.58$
$\log_2\left(\frac{5}{10}\right) = -1$
$\log\left(\frac{4}{5}\right) = -0.1$
$\log_2\left(\frac{\frac{5}{L}}{\frac{10}{2L}}\right) = 0$

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
- Geometric normalization (default)
  - 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 η_j.

Sources of Variability

- Transcript abundance in biological samples differs even if the conditions have been the same. This is *biological variability*.
- Library construction and sequencing adds *technical variability*, so that the relative fragment abundance is different from the relative transcript abundance in the sample.
- Two measurements of aliquots of the same RNA sample differ by technical variability. Replicates differ by the sum of biological and technical variability.

T set of all transcripts

$\{C_t\}_{t \in T}$ fragment count for transcript t

ρ_t relative abundance of transcript t

N total fragment count

$C_t \sim \text{Poisson}(\rho_t N)$

$\hat{\rho}_t = \frac{x_t}{N}$

- Fragment counts are not observed due to ambiguously mapped fragments, so are only estimated.
- Also, there may be variability in the rate of fragment production by location on the transcript
- Both lead to overdispersion
- So we can model the transcript count as a mixture of Poisson's, in particular as a negative binomial

Variance Estimation

- For each set of replicates of a given condition, we compute the mean and variance of the gene-level scaled fragment counts across replicates, so that there are as many pairs as there are genes.
- We fit a local regression to the variances as a function of the mean and use the fitted variance for the variance of the negative binomial distributions.
- This is done separately for each condition.
- If a condition has no replicates, then we use the estimates from the condition with the largest number of replicates.

- This is basically the Huber and Anders procedure.
- Alternatively, we could pool these across conditions as is usual in ANOVA
- This procedure will give biased estimates of the variances for two reasons:
  - The means are estimates as well as the variances, introducing biases into the regression function.
  - Two genes with the same mean fragment count can have different variances.
- Cuffdiff2 and DESeq account for the systematic change of variance with the mean.
- edgeR accounts for the variability around the mean.
- Probably, one should do both.

Assigning Counts to Isoforms

- Partition each gene into non-overlapping loci (one or more exons) so that all isoforms are accounted for
- We estimate the probabilities that a fragment maps to a given locus and the probability that a fragment mapping to a given locus comes from each of the transcripts containing that locus
- Incorporates fragment bias estimates
- Initializes probabilities that an ambiguously mapped fragment comes from each of the n loci it could come from as $1/n$.
- Iterates maximum likelihood estimates.

Modeling Transcript Abundance

- The transcript abundance ρ_t is modeled as a beta negative binomial; that is, a beta mixture of negative binomials.
- This has three parameters, which are chosen to fit three constraints:
  - The mean should equal the observed transcript count times the estimated chance that that count comes from transcript *t*, which is the mean of the estimated transcript count.
  - The variance should equal the variance of the estimated transcript count.
  - The variability of the mean of the negative binomial should be equal to the estimated uncertainty of the transcript assignment.

Poisson distribution with mean λ

$\mu = \lambda$

$\sigma^2 = \lambda$

The mean λ of the Poisson distribution is gamma distributed with parameters r and θ

This makes a negative binomial distribution with parameters p and r

$$\theta = \frac{p}{1-p}$$

$$p = \frac{\theta}{1+\theta}$$

$$\mu = \lambda = r\theta = \frac{rp}{1-p}$$

$$\sigma^2 = \frac{rp}{(1-p)^2}$$

The parameter p is then beta distributed with parameters α and β

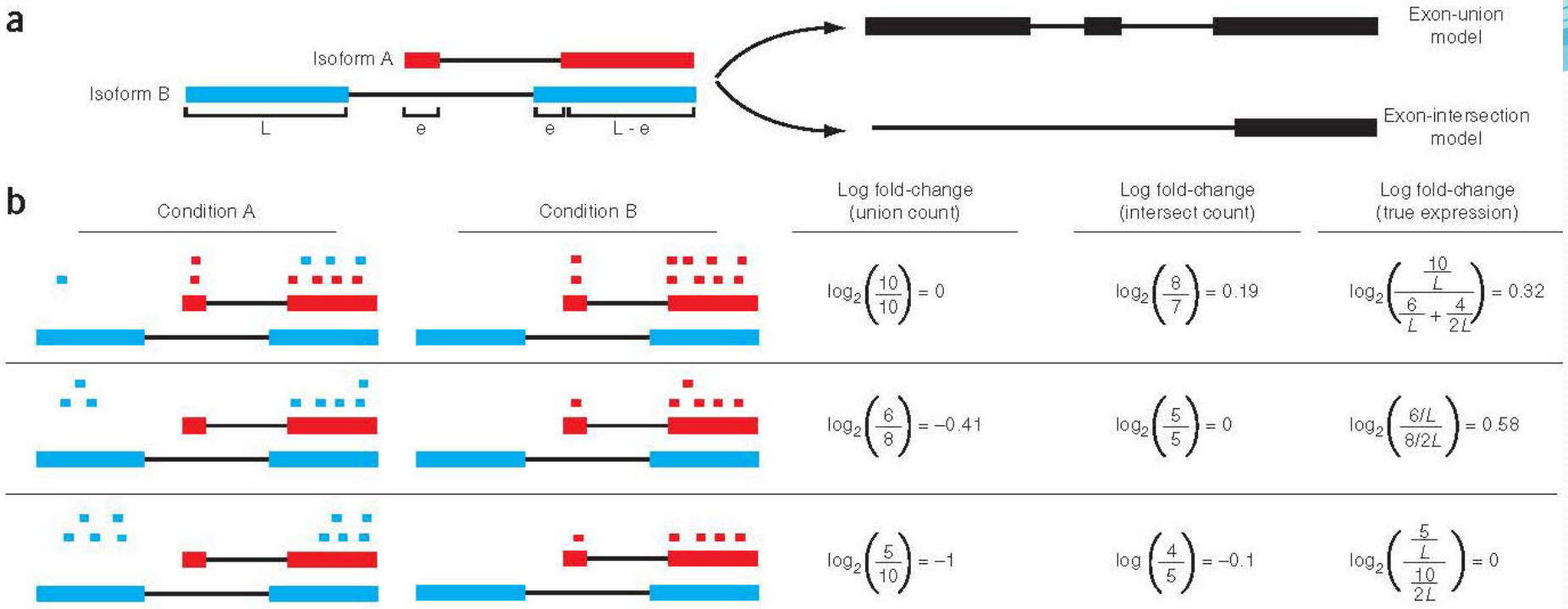
$$\mu = \frac{r\alpha}{\beta - 1}$$

Differential Expression

- Other methods of differential expression have one number per gene/sample that may be scaled, but is otherwise observed.
- Cuffdiff expression measurements for a given transcript/sample is actually a probability distribution, not a measurement.
- Expression for a gene is a sum of scaled transcript distributions.
- Uses resampling of assignment to produce variance estimates.
- This appears opaque at best.

Conclusion

- Although the point about isoforms is in a sense technically correct, it is not clear what advantage it confers on the statistical analysis.
- In particular, taking an observed count for a gene, dividing it into uncertain counts for transcripts and then adding them back together does not seem to help the clarity of the situation
- The variance function method is questionable.



In the top line, we have four exons, with observed counts (1, 2, 0, 7) and (0, 2, 0, 8). Counts in position 1 are isoform A and 2 are B. Counts in position 4 are either. Thus on the left, counts for the two isoforms are 8/2, 7/3, 6/4, 5/5, 4/6, 3/7, 2/8, or 1/9 with various probabilities. On the right, the counts for the two isoforms are 8/2, 7/3, … 2/8, also with various probabilities.

What does this get us that the original counts don't?

Poisson distribution with mean λ

$\mu = \lambda$

$\sigma^2 = \lambda$

The mean λ of the Poisson distribution is gamma distributed with parameters r and θ

This makes a negative binomial distribution with parameters p and r

$$\theta = \frac{p}{1-p}$$

$$p = \frac{\theta}{1+\theta}$$

$$\mu = \lambda = r\theta = \frac{rp}{1-p}$$

$$\sigma^2 = \frac{rp}{(1-p)^2}$$

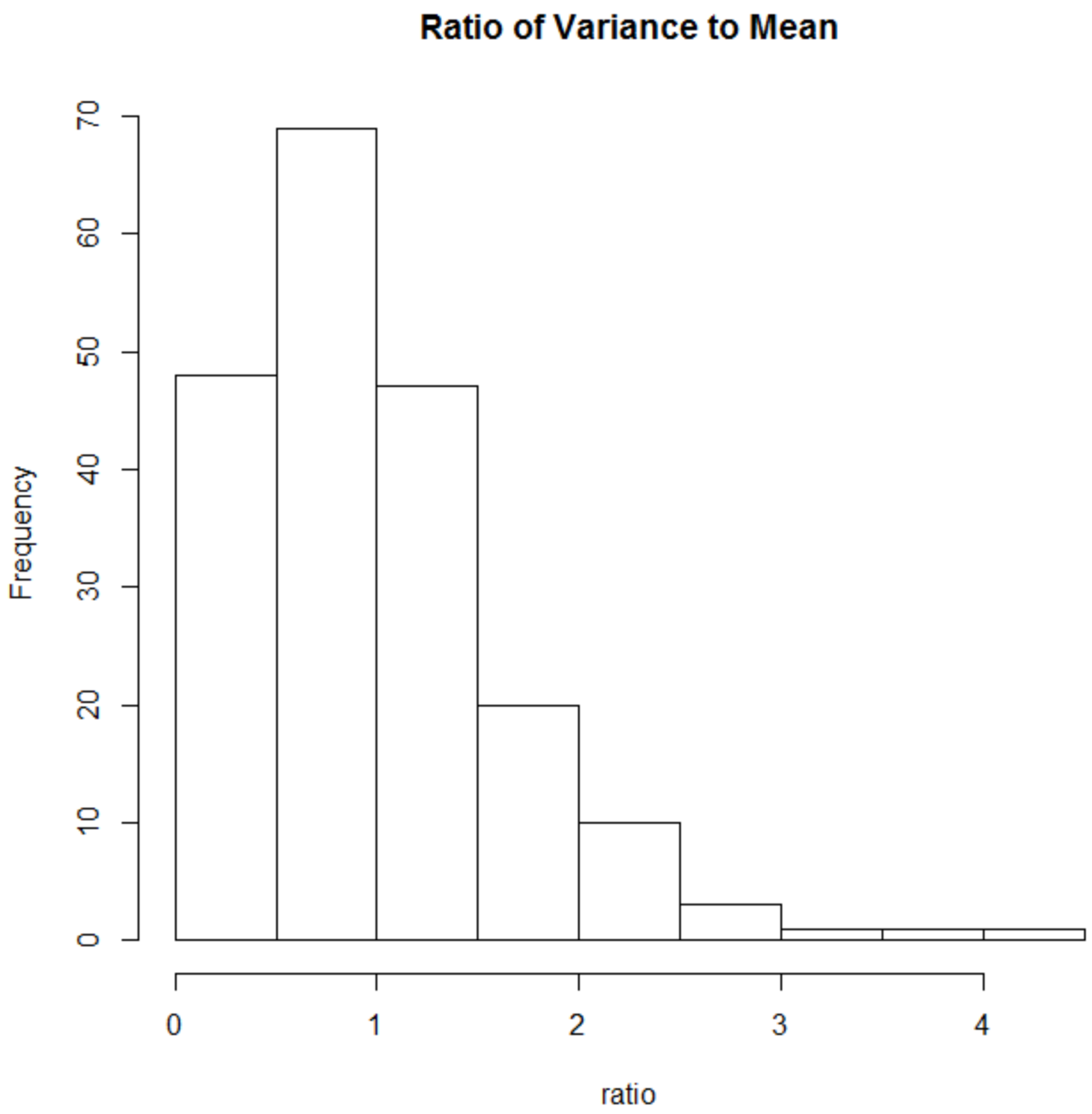
We can measure overdispersion with

$$\frac{\sigma^2}{\mu} = \frac{1}{1-p}$$

When $p = 0$ this is 1, so Poisson.

As $p \to 1$, the ratio goes to ∞

- I generated 200 samples of 5, Poisson(10).
- The observed ratio of the variance to the mean
- This varied from 0.02 to 4.3 with a mean very near 1.
- In this case, using the averages works (mean of means = 10.01, mean of variances = 10.11)
- But if the true variances vary, then this does not work well because it overestimates the variance of some and underestimates the variances of others.
- More work to be done


Ratio of Variance to Mean
Frequency
0
10
20
30
40
50
60
70
0
1
2
3
4
ratio