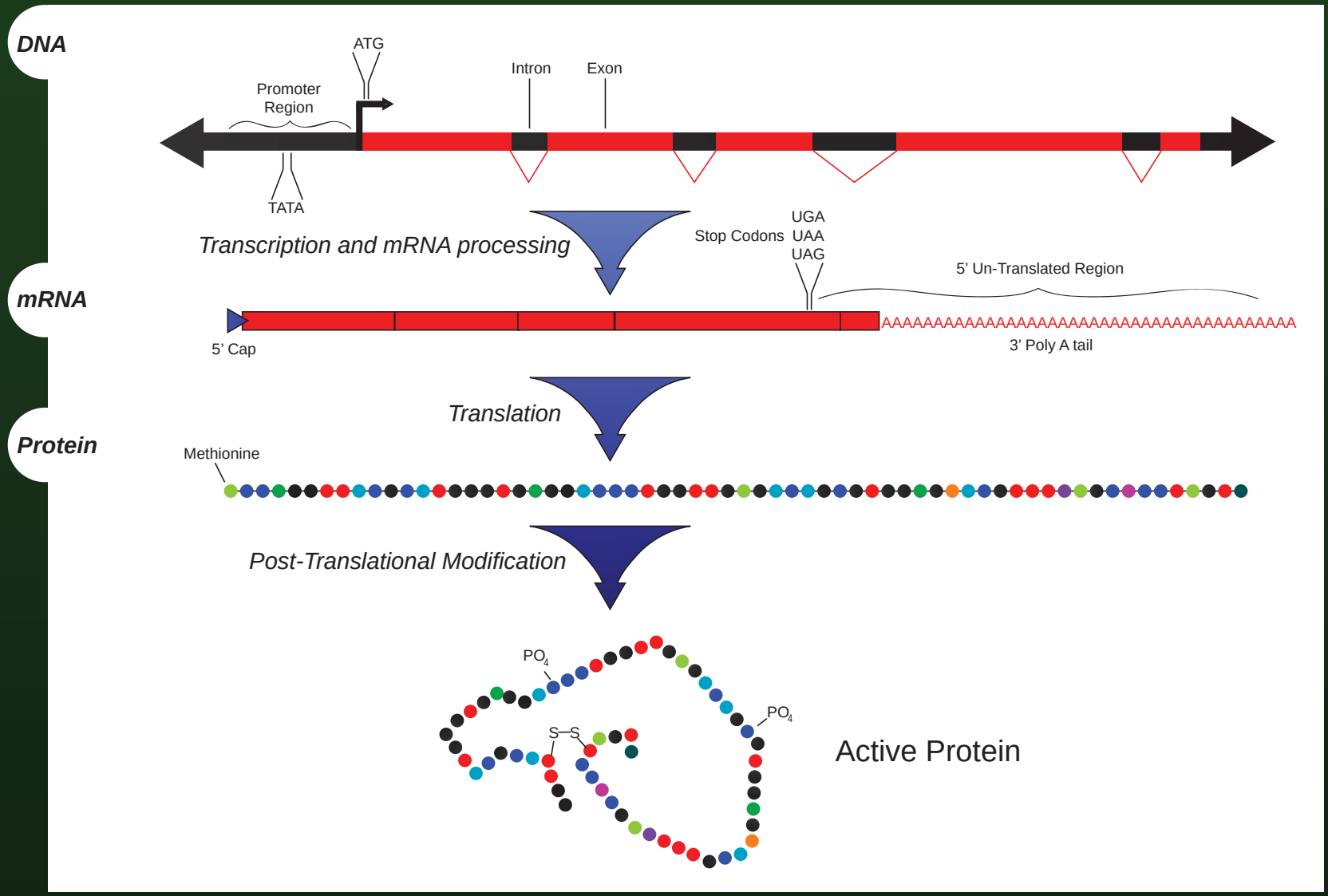
Excess False Positives in Negative-Binomial Based Analysis of Data from RNA-Seq Experiments

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

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.

Central Dogma of Molecular Biology : Eukaryotic Model



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/splice variants 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) choose to normalize the samples first, and may choose to import some information across genes.

Read Counts

- We assume that bioinformatic analysis has been conducted on the original RNA-Seq data for each sample and converted into reads K_{ij} per gene *i* per sample *j*.
- Since each K_{ij} is a count, we might first try the assumption that the underlying distribution is Poisson.
- Examination of RNA-Seq data sets shows that data are over-dispersed (see next slide) and so the Poisson assumption cannot hold.

Poisson Problems

- The Poisson distribution has a single parameter λ, which controls both the mean λ and the variance λ.
  - $E(X) = \lambda$
  - $V(X) = \lambda$
- If the sample standard deviation is not approximately equal to the mean, then one solution is the over-dispersed Poisson:
  - $E(X) = \lambda$
  - $V(X) = \theta\lambda$
- But empirically, the variance appears to rise quadratically with the mean, not linearly with the mean.
- This an happen via a mixture of Poissons.

Freezer

Cells grown to 50% confluence in 8 dishes

4 treated with TNF-α
4 controls

Actual transcript count for one gene in 8 dishes

RNA-Seq mapped read count for one gene in 8 dishes

Other (biological) variability. The Poisson parameter λ for a given gene varies from dish to dish.

Poisson variability (possibly)

Poisson Variability

- Is at best the technical variation
- Even then, the approximation is poor
- A better model for the overall experimental variability is a perhaps mixture of Poissons.
- Each sample under the same condition has a different actual transcript count before sampling, which means that λ_{ij} for gene *i* and sample *j* differs from sample to sample, even when these are replicates under the same conditions.
- So the actual observed mapped reads count for a given gene across samples in the same condition can better be considered as an approximate mixture of Poisson random variables.

The Negative Binomial

- Given a series of Bernoulli trials (coin flips) with chance of success p on each trial, if we let X be the number of successes at the time when r^{th} failure occurs, then X has a negative binomial distribution $NB(r;p)$.
- For example, if r is 3, and the sequence is SSFSFSSSSFSSF, then X is 7 and the length of the sequence is 10.
- The mean of X is $pr/(1-p)$ and the variance is $pr/(1-p)^2$.
- In this formulation, r is known and p is the only unknown parameter.

The Negative Binomial

- An more useful formulation for our purpose is that a negative binomial can be written as a gamma mixture of Poisson random variables.
- If an observed count is Poisson, with a random mean λ that differs from time to time and which is gamma distributed across sample replicates, then the count overall is negative binomial.

Gamma Mixtures

- The gamma distribution (same as the chi-squared) is an asymmetric distribution defined on $[0,\infty)$ with two parameters, the shape k and the scale θ. The mean is $k\theta$ and the variance is $k\theta^2$.
- $NB(r; p)$ is produced by a gamma mixture of Poissons with $k = r$ and $\theta = p/(1-p)$.
- Otherwise put, if we can estimate the mean μ and variance σ^2, then the parameters of the negative binomial can be determined from those.

Negative Binomial Parameters

- The negative binomial distribution can be parametrized in a number of ways.
- In each case, it requires two parameters.
- For the original definition in terms of waiting times, the parameter *r* is set and the parameter *p* is estimated.
- If we parametrize the negative binomial by the mean μ and the dispersion factor θ, then
  - $E(X) = \mu$
  - $V(X) = \mu + \theta \mu^2$

Negative Binomial Parametrizations

$$\mu = r\frac{p}{1-p}$$

$$\sigma^2 = r\frac{p}{(1-p)^2}$$ Mean and variance of a negative binomial

$$\sigma^2 = \mu + \theta\mu^2 \quad \text{where} \quad \theta = r^{-1}$$

Parametrize by μ and the dispersion factor θ

Gamma mixture that produces a negative binomial μ, θ

$$\text{shape} = r = \theta^{-1}$$

$$\text{scale} = \theta\mu$$

Existing Methods

- Existing methods often contain complex combinations or filtering, normalization, transformation, and variance estimation before any statistical tests are performed.
- The methods are sometimes poorly documented and can change rapidly and often substantially between versions.
- The results of different popular methods such as DESeq2, edgeR, and limma-voom can differ substantially.
- It appears that some methods produce large numbers of false positives.

DESeq2 and edgeR

- These methods share many similarities.
- Significance tests are based on the negative binomial likelihood.
- Estimation of the mean for each group is easy.
- The shared dispersion factor θ can be estimated from each gene separately, or from a smoothed curve relating the mean and variance, or from a compromise.
- DESeq2 is primarily due to Mike Love, Simon Anders, and Wolfgang Huber.
- edgeR is primarily due to Mark Robinson, Gordon Smyth and Davis McCarthy

limma

- limma is a package originally developed by Gordon Smyth for gene expression arrays.
- It conducts a linear model analysis for each gene, and by default uses empirical Bayes shrinkage of the MSE's.
- The voom function was developed for RNA-Seq data; it weights observations using the negative binomial variance function, but does not use the likelihood as such.
- Alternatively, one can transform the data to approximately stable variance.
- In either case, inference is based on standard linear models.

Variance Stabilizing Transformation

$$E(X) = \mu$$

$$V(X) = \mu + \theta\mu^2$$

$$Y = f(X)$$

$$V(Y) \approx c^2 \approx f'(\mu)^2 V(X) = f'(\mu)^2(\mu + \theta\mu^2)$$

$$f'(\mu) \approx c(\mu + \theta\mu^2)^{-1/2}$$

$$f(x) = \ln\left[x + \sqrt{x^2 + \theta^{-1}x} + \theta^{-1}/2 \right]$$

We choose θ so that the regression of variance on mean has zero slope.

This will use lots of means and variances or possibly means and MSE's.

We do not have to directly estimate θ.

MontPick Data

- RNA-Seq of RNA from lymphoblastoid cell lines from 129 individuals from the HapMap project.
- Montgomery data: 60 European heritage individuals from Utah (CEU), 33 females and 27 males.
- Pickrell data: 69 Yoruba heritage individuals of from Nigeria (YRI), 40 females and 29 males.
- 52,580 unique genes of which 8,124 had a count of at least 129 across the 129 samples.
- This data set has been used in other comparisons
- First, we compare four methods on this data set: DESeq2, edgeR-glm, limma-voom, and limma with a prior variance-stabilizing transformation.
- Later we construct data sets where the null hypothesis is true.

Results of Full Data Analysis

	Population (CEU/YRI)	Sex	Interaction
edgeR	4496	511	87
DESeq2	4495	503	66
limma-voom	4392	134	13
limma-trans	4536	193	3

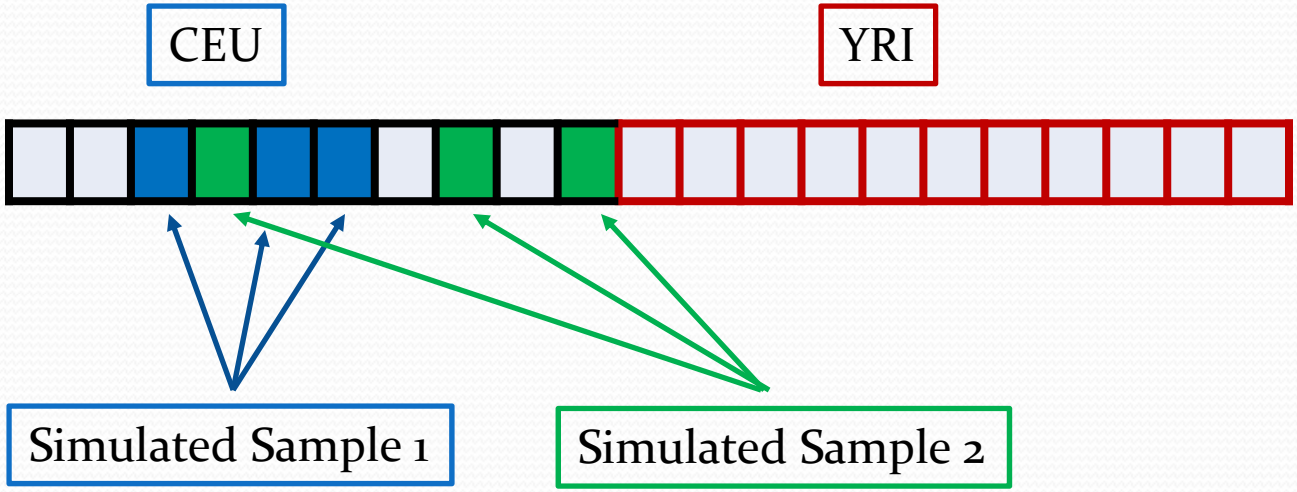
- Number of significant genes/8142 at 10% FDR.
- Total sample size is 129
- Largest difference is in the Sex factor and interaction.
- Only a small number of differences by sex in B-cell gene expression would be expected.
- Most RNA-Seq experiments have at most 2–5 replicates per group, not 60, and usually two groups.

Power to Detect or False Positives?

- Given the large sample size, the similarity in significant gene numbers by population is not surprising.
- But there are many more genes significant for sex in the two negative binomial based methods, and an even larger disparity for the interaction term.
- We may naively think that more genes detected is better.
- But this is only true if most of these are true positives.
- To evaluate the methods, we need to measure the performance when the null hypothesis is true.

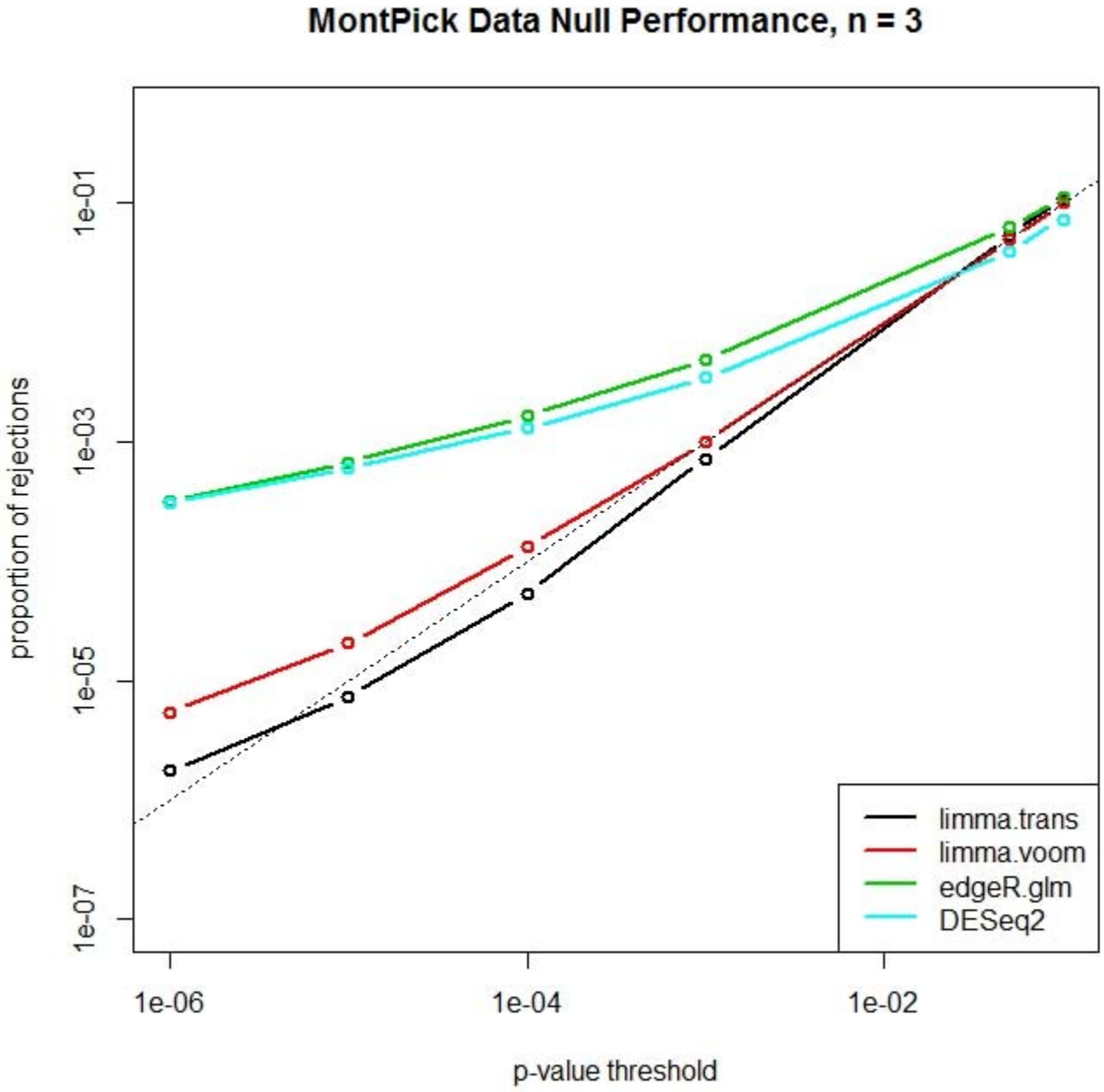
Null Behavior

- 10,000 randomly selected subsets of size 2*n* out of the 60 CEU samples.
- In each subset, *n* assigned to "treatment" and *n* to "control" at random.
- The null hypothesis is on the average true.
- Given the 8,124 genes with large enough counts, we would expect about 0.01% to be significant at the $p = 0.0001 = 10^{-4}$ level.
- We have 10,000 tests for each gene, so there should be about 1 rejection per gene for each method, and 8,124 rejections overall for each method
- We did this for n = 3, 5, 10, and 30.


CEU
YRI
Simulated Sample 1
Simulated Sample 2

Mean Number Significant at p = 0.0001 with n = 3 MontPick Data; Null Hypothesis is True

Estimator	Observed	Expected	Ratio
edgeR	13.36	0.81	16.44
DESeq2	10.52	0.81	12.95
limma-voom	1.06	0.81	1.31
limma-trans	0.43	0.81	0.53


MontPick Data Null Performance, n = 3
proportion of rejections
1e-01
1e-03
1e-05
1e-07
1e-06
1e-04
1e-02
p-value threshold
limma.trans
limma.voom
edgeR.glm
DESeq2

Anomalous Small p-values

- Many of the p-values for the negative binomial likelihood methods are very small. The numbers below are from edgeR with $n = 3$.
- Out of 8124 × 10,000 tests, there should be about 81 less than 10^{-6}, but there are actually 25,390.
- There should be around 0.8 less than 10^{-8}, but there are actually 7,325.
- There should be almost none less than 10^{-10}, but there are actually 2,843.
- The smallest p-value is 3×10^{-51}. For a t on 4df to reach this p-value would require separation by 7 trillion standard deviations.

Important Factors?

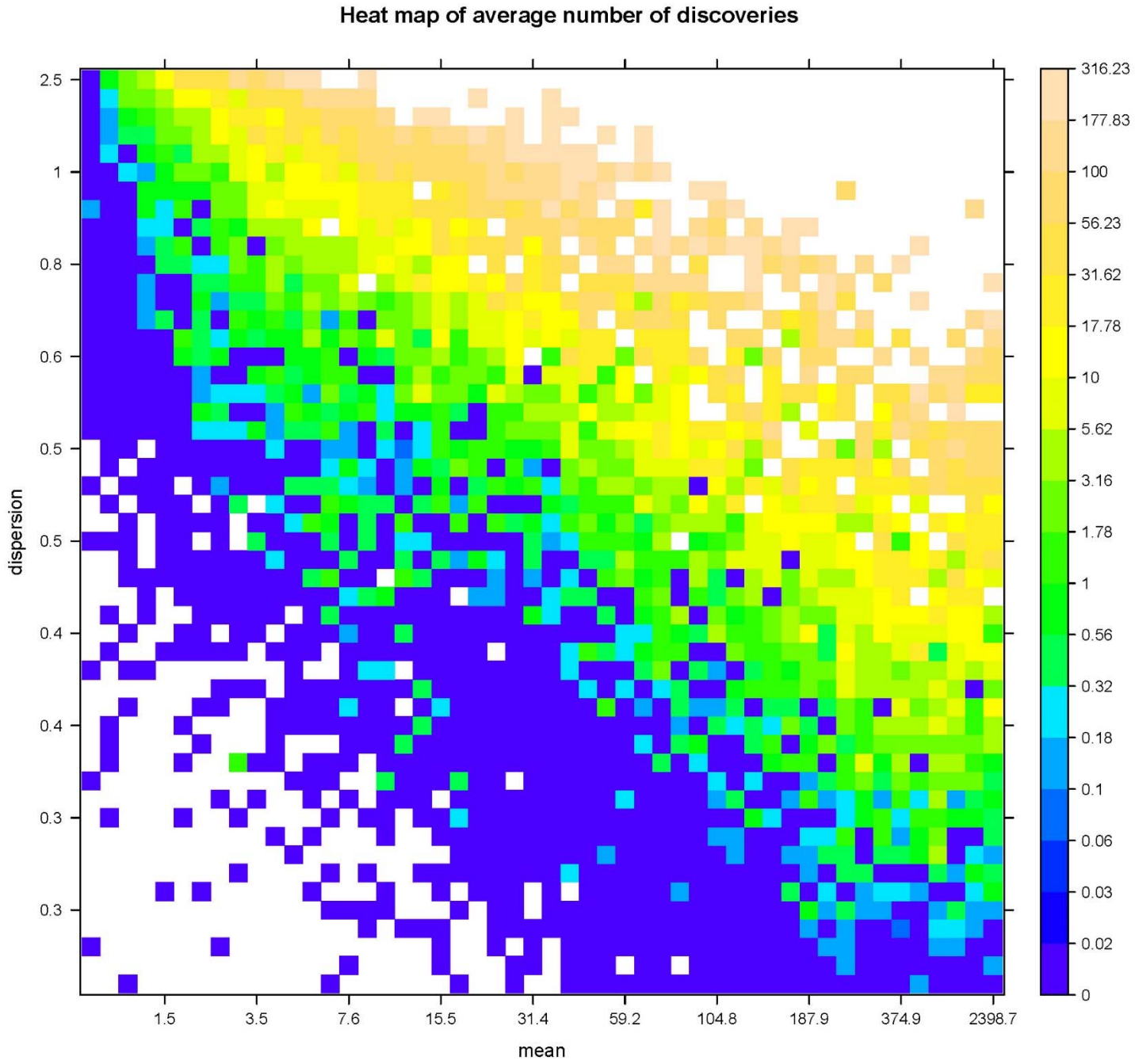
- These methods share some features but differ in their implementation.
- All of them treat the data as negative binomial, or at least use the variance-mean relationship.
- They may use a test based on the negative binomial distribution, but limma-voom and limma-trans uses standard regression/anova with 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 and other factors

Factors That May Cause False Positives

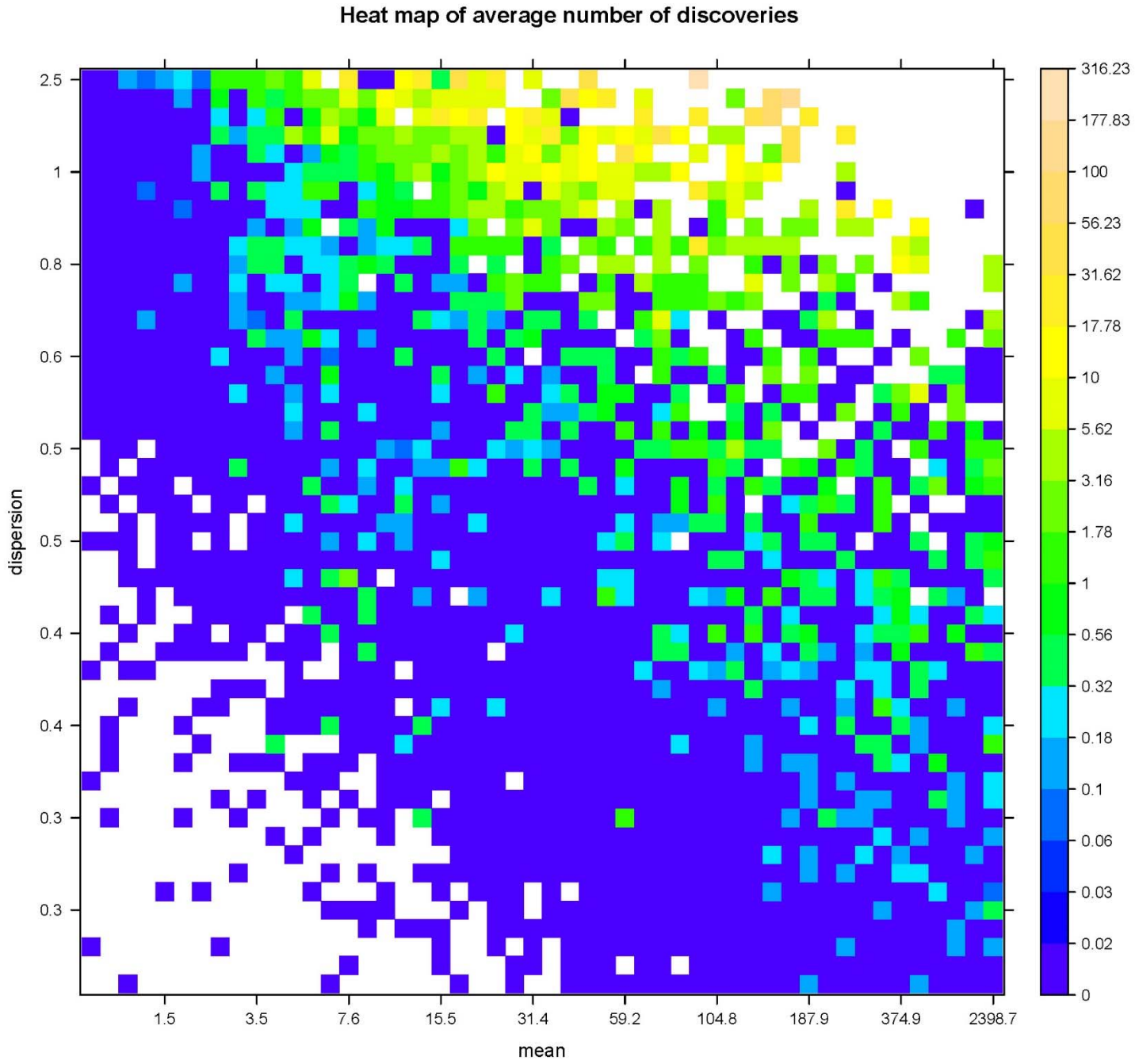
- Normalization procedure
  - Normalization driven by high-abundance transcripts can cause spurious significance of low abundance ones.
  - May need different normalization for high and low counts
- Variance estimation
  - Many methods “borrow strength” from other genes to get “better” variance estimates, which involves shrinking variances or dispersion factors towards a central tendency.
  - This may cause bias though increasing power
- Statistical test used
  - Many choices
- Nature of the data
  - May require filtering to avoid too many zero or low count situations.
  - May not be negative binomial.

Which Genes have High FP Rate?

- Some have conjectured that false positive rates will be high for genes with low counts, but this turns out not to be correct.
- All of the methods shown have rates of false positives that are low when the mean and dispersion factor are low, and high when the mean and dispersion factor are high.
- We show this, and later discuss the source of this phenomenon.



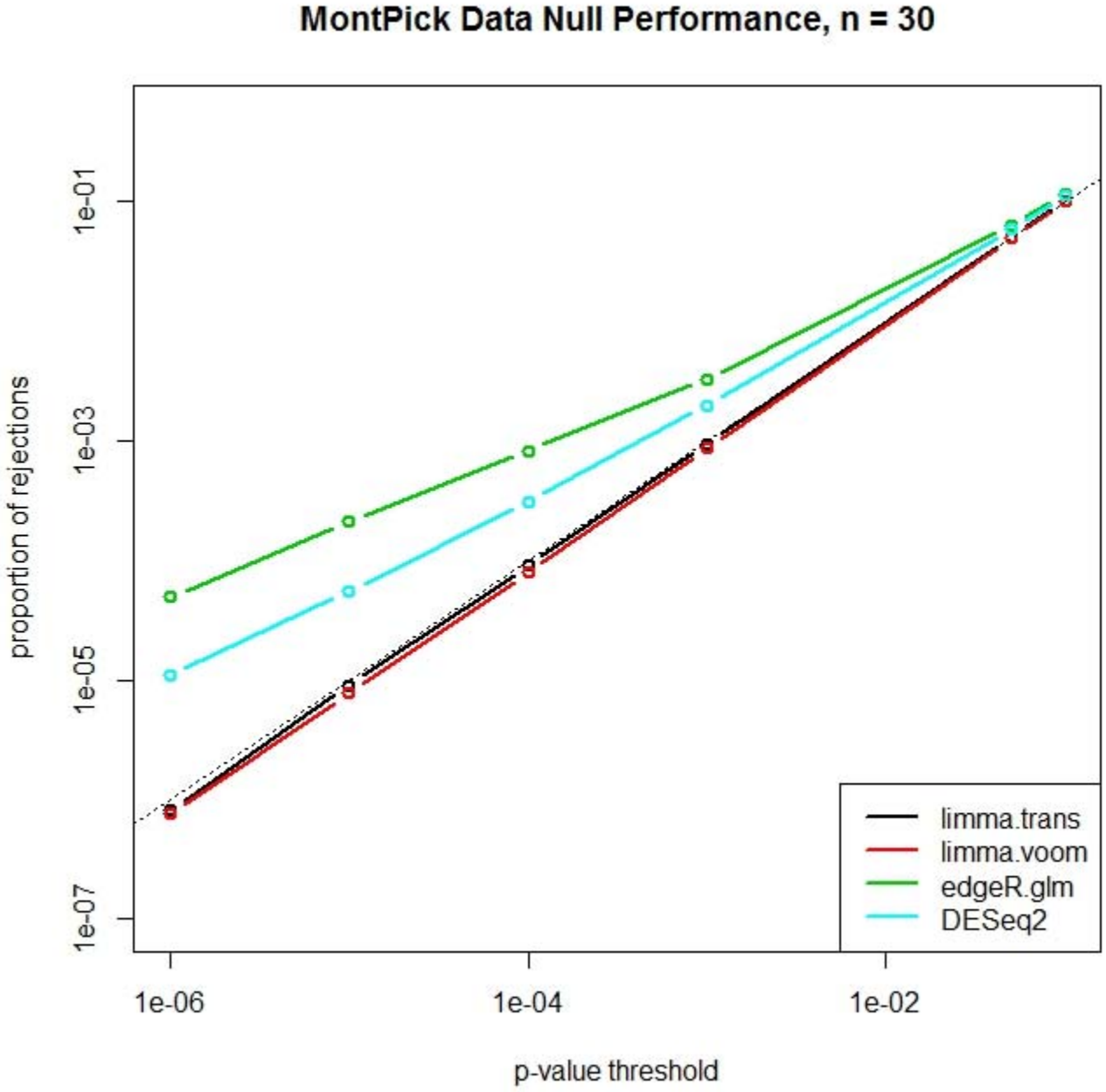
False Positives
edgeR
n = 3



False Positives
limma-trans
n = 3

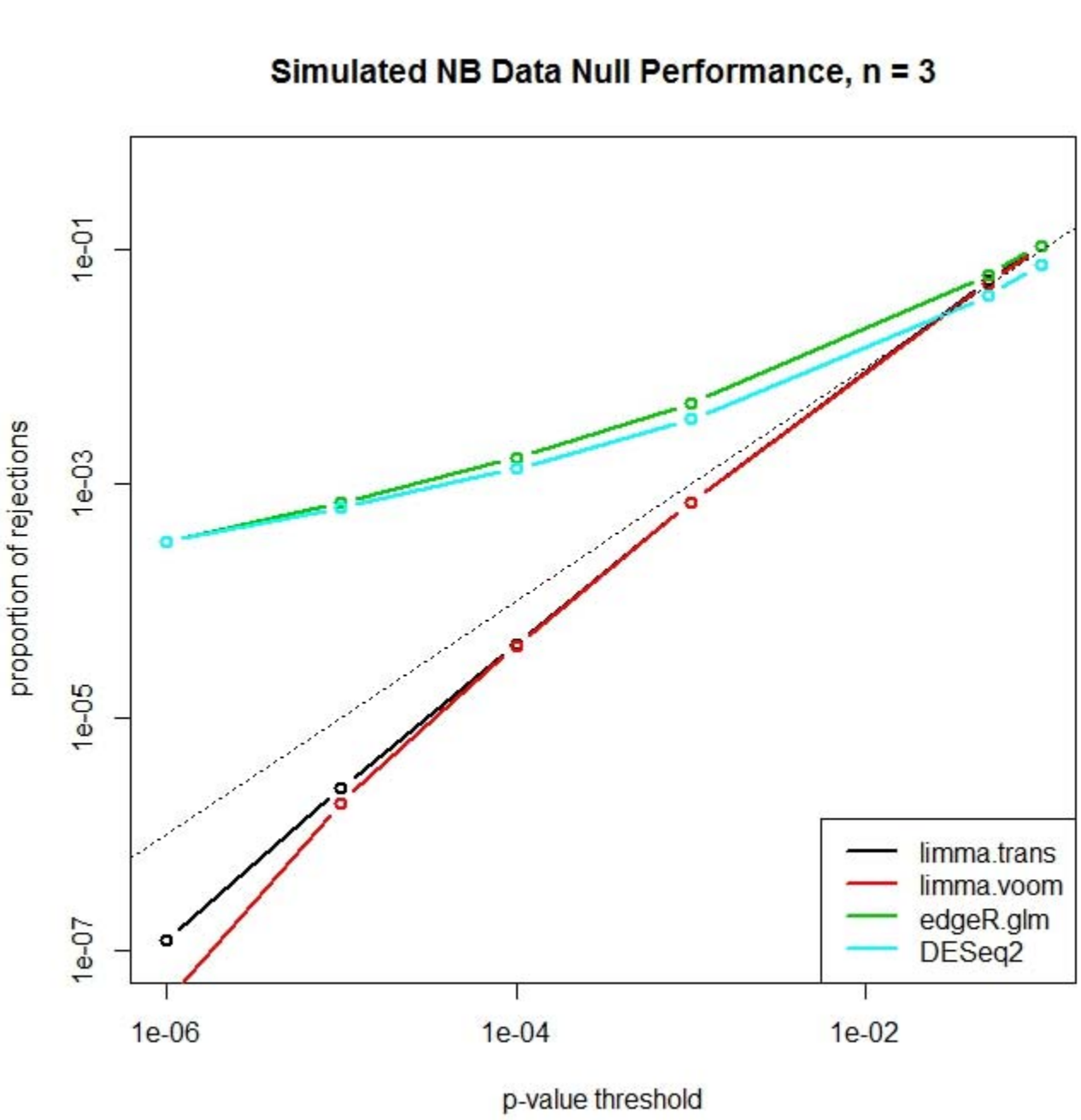
What Happens for Larger *n*?

- Although the false positives are fewer with larger sample sizes, they still occur often at a factor of ten more frequent than nominal


MontPick Data Null Performance, n = 30
proportion of rejections
1e-01
1e-03
1e-05
1e-07
1e-06
1e-04
1e-02
p-value threshold
limma.trans
limma.voom
edgeR.glm
DESeq2

Is the Real Distribution not Negative Binomial?

- We ran 10,000 trials at n = 3, 5, 10, and 30 using negative binomial random variables.
- For each of the 8124 "genes," we generated the data using values of μ and θ that matched those of the CEU data from the Montgomery half of the data set.
- Results differed very little from those of the resampled real data.
- The problem is not the distribution of the data.


Simulated NB Data Null Performance, n = 3
proportion of rejections
1e-01
1e-03
1e-05
1e-07
1e-06
1e-04
1e-02
p-value threshold
limma.trans
limma.voom
edgeR.glm
DESeq2

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
- Counts per Million in limma-voom for example
- 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 or 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

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

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

- 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.
- 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 5×10^{-7}
- With two groups of 3, there are ~4df and 5×10^{-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 λ has

$\mu = \lambda$

$\sigma^2 = \lambda$

A negative binomial random variable can be seen as a mixture of Poissons with varying λ

$\mu = \lambda$

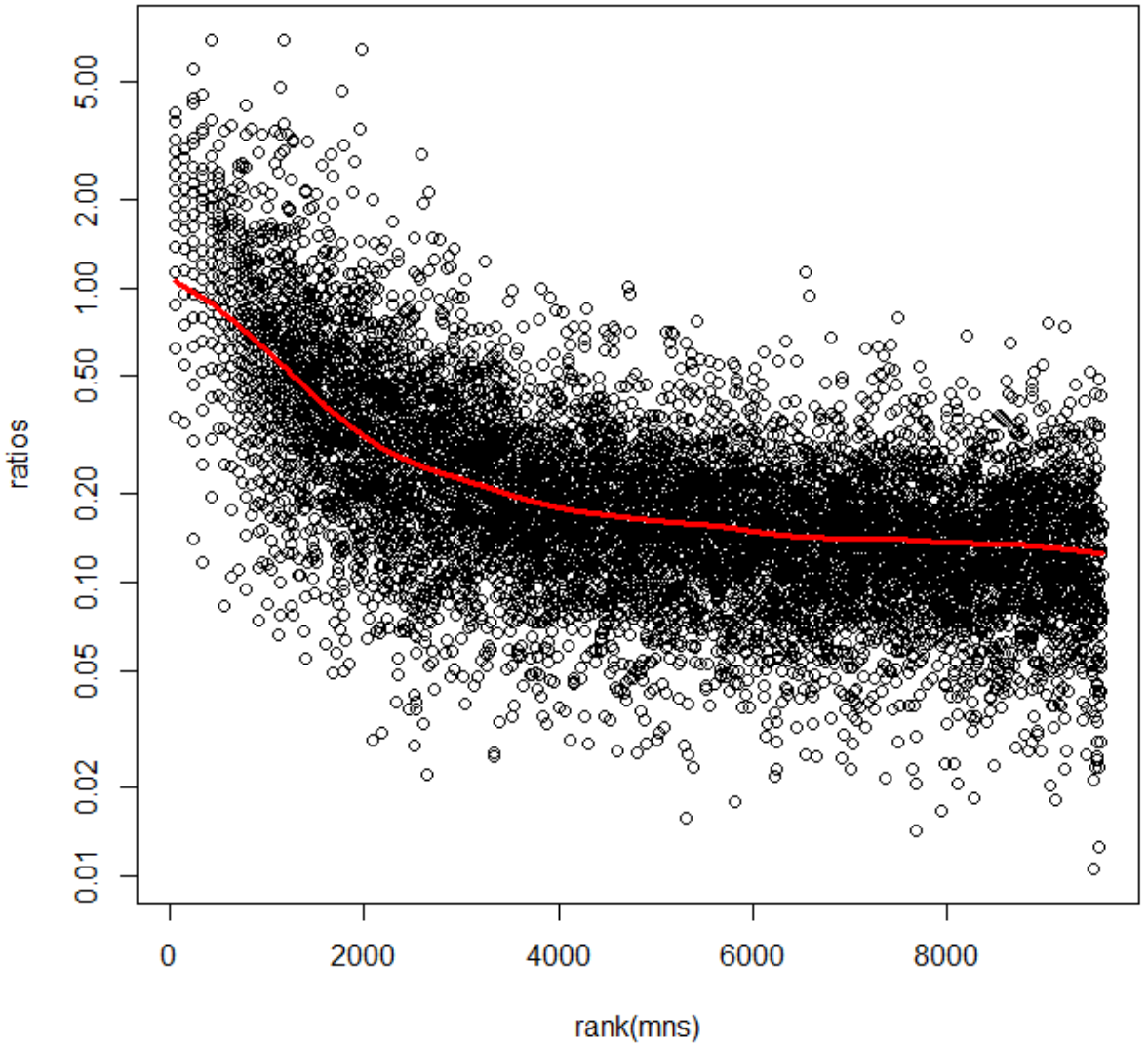
$\sigma^2 = \lambda + \theta\lambda^2$

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

$\hat{\theta} = \dfrac{s^2 - \overline{x}}{\overline{x}^2}$ if this is positive

If we smooth a plot of the variance or the square CV or $\hat{\theta}$ vs. the mean we obtain an average estimate of θ for the given value of μ

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


ratios
5.00
2.00
1.00
0.50
0.20
0.10
0.05
0.02
0.01
0
2000
4000
6000
8000
rank(mns)

$$\mu = \lambda$$

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

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

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

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

In practice, the Cox-Reid estimate is often used

instead of a method-of-moments estimate

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.
- It partially explains why large dispersion estimates have many false positives since they are biased down by the shrinkage.

Generalized Linear Models

- The basic approach in edgeR and DESeq2 is that of the generalized linear model.
- The base distribution is supposed to be of an exponential family, which means that the likelihood factors in a particular way.
- There is a canonical parameter that is a function of the mean and that can be modeled as a linear function of predictors.
- There is a variance function that depends on the mean only.
- Possible distributions include the normal, Poisson, binomial, gamma, and inverse Gaussian.

Poisson Distribution

$$f(x) = \frac{\lambda^x e^{-x}}{x!}$$

$$\ln(f(x)) = x\ln(\lambda) - x - \ln(x!)$$

$$= \frac{x\eta - b(\eta)}{a(\phi)} + c(x,\phi) \quad \text{where}$$

$$\eta = \ln(\lambda)$$

$$\phi = 1$$

$$a(\phi) = \phi$$

$$b(\eta) = 0$$

$$c(x,\phi) = -x - \ln(x!)$$

This is the definition of an exponential family distribution.

Negative Binomial Distribution

$$f(x;r,p)=\binom{x+r-1}{x}p^{x}(1-p)^{r}$$

$$\ln(f(x;r,p))=x\ln(p)+r\ln(1-p)+\ln\left[\binom{x+r-1}{x}\right]$$

$$=\frac{x\eta-b(\eta)}{a(\phi)}+c(x,\phi)\quad\text{where}$$

$$\mu=\frac{pr}{1-p}$$

$$\eta=\ln(p)=\ln\left(\frac{\mu}{r+\mu}\right)$$

$$\phi=r$$

$$a(\phi)=1$$

$$b(\eta)=-r\ln(1-e^{\eta})$$

$$c(x,\phi)=\ln\left[\binom{x+r-1}{x}\right]$$

This is an exponential family distribution only if

$r=1/\theta$ is known.

Otherwise, $b()$ is a function of θ as well as η and η is a function of θ as well as μ.

Consequences

- In practice, we are treating the dispersion factor θ as known, whereas it is estimated.
- For large enough samples, this will not matter, but it seems two groups of 30 are not large enough.
- It is as if we are treating a t on 4df as if it were normal.
- This does not happen if we use a standard (non-generalized) linear model, as in limma.
- Though some false positives can still occur, they are much rarer.

Quasi-Likelihood

- Quasi-likelihood is a generalization of these ideas. Instead of assuming a likelihood we assume only a variance function.
- We can add a multiplicative dispersion parameter without apparent harm (over-dispersed Poisson does not exhibit excess false positives).
- All of the usual distributions except the negative binomial have a quasi-likelihood formulation in which the variance function contains a dispersion parameter only multiplicatively (like σ^2 for the normal).
  - Poisson QL: $V(X) = \phi\lambda$
  - Negative Binomial: $V(X) = \mu+\theta\mu^2$
  - Negative Binomial QL?: $V(X) = \phi\ (\mu+\theta\mu^2)$ (not well identified)

Does it Matter?

- In some contexts in the aggregate the excess false positives may not seem to matter. If 8124 genes expect 1 false positive at $p = 10^{-4}$ (roughly FDR < 0.10), but there are 100 false positives instead, and there are 2000 positives in total, then this might increase the true FDR from 10% to 15%.
- If there are 500 total positives, then 100 extra false positives is an inferential disaster.
- And if one of the extra false positives is a gene you care about, then it is also a disaster.
- In any case, to argue that it does not matter when we get p-values less than 10^{-10} in a small-sample null case seems to show no respect for the concept of statistical testing.

Conclusions

- Methods for RNA-Seq data that depend strongly on a negative binomial assumption can generate large numbers of false positives.
- Use of standard linear models after a data transformation seems to work well. This depends only on the assumed variance function and not even strongly on that (we could use a started log instead).
- Shrinkage of variances or dispersions can also cause false positives for high mean/high dispersion genes, though not to the same extent.
- Two versions on bioarxiv—latest one will follow this month.

Acknowledgments

- The work presented here is an ongoing joint effort with Yilun Zhang, MS, Sharon Aviran, PhD, and Blythe Durbin-Johnson, PhD.
- We gratefully acknowledge support from the National Institutes of Health
  - NCATS UL1 TR000002 (UC Davis CTSA);
  - NIAID R33AI080604 (Rocke);
  - NHGRI R00 HG006860 (Aviran).