EBSeq: An empirical Bayes hierarchical model for inference in RNA-Seq experiments

Ning Leng[1]*, John A. Dawson[1], Anna Rissman[2], Bart Smits[2], Victor Ruotti[3], Ron M. Stewart[3], James A. Thomson[3], Michael Gould[2] and Christina Kendziorski[1,4]

[1]Department of Statistics, University of Wisconsin, Madison, WI 53706; [2]Department of Oncology, University of Wisconsin, Madison, WI 53706 ;
[3]Morgridge Institute for Research, Madison, WI 53707; [4]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53706

ABSTRACT

RNA-sequencing is a powerful approach providing estimates of both isoform and gene expression with unprecedented dynamic range and accuracy. A fundamental goal of RNA-seq experiments measuring expression in two or more biological conditions is the identification of differentially expressed isoforms and genes.

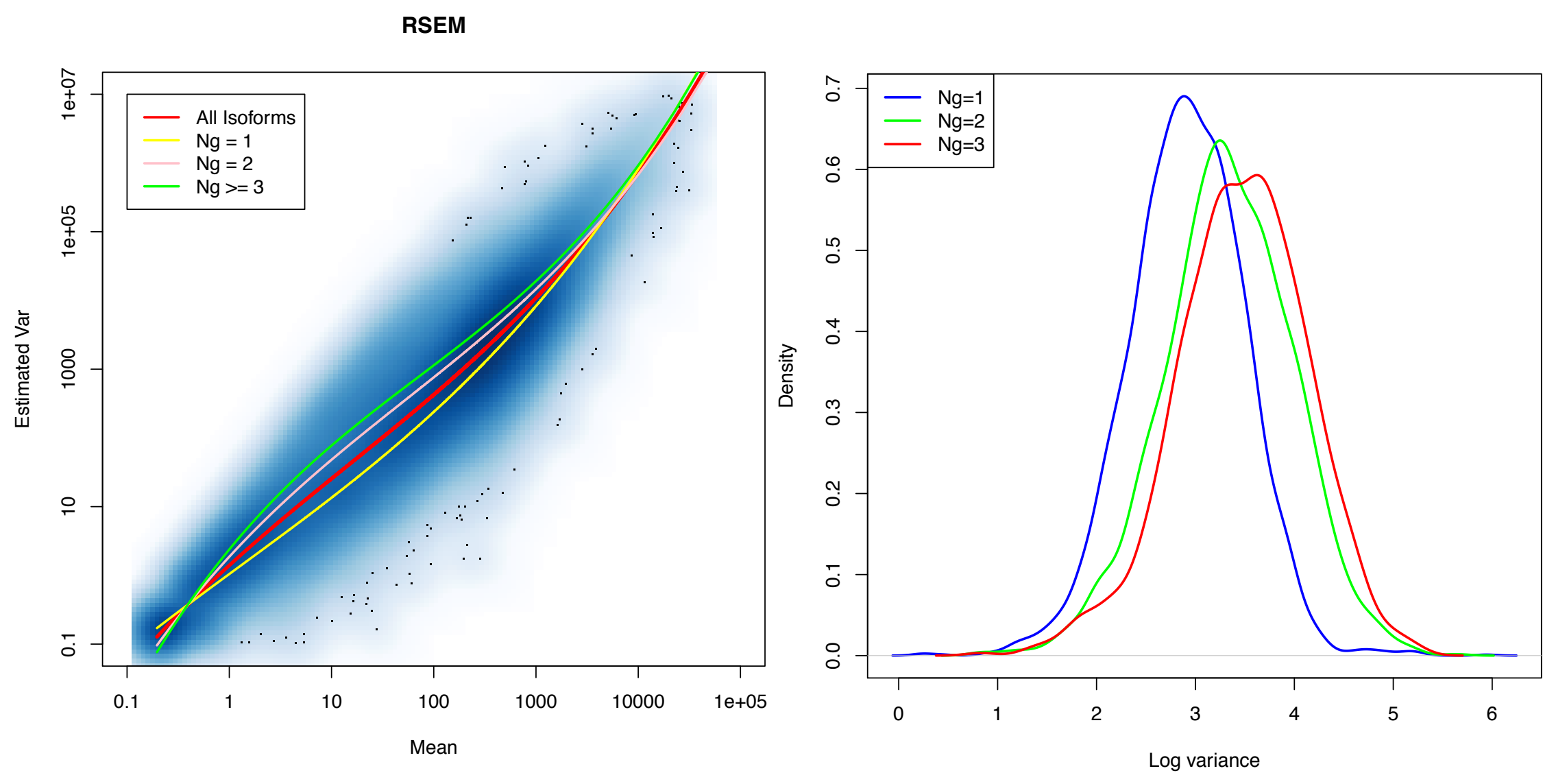
Most of the statistical methods developed to identify differentially expressed genes measured using microarrays do not directly apply, and the methods that have been developed specifically for RNA-seq measurements do not directly accommodate isoform level expression, dependence across isoforms, and mapping uncertainty.

We have developed an empirical Bayesian modeling approach that accounts for and capitalizes on these features. Motivation for and advantages of the approach are illustrated in simulations, in RNA-seq studies of human embryonic stem cells, and in an RNA-seq study of mammary carcinogenesis in a rat model for breast cancer.

INTRODUCTION

- RNA-seq is a revolutionary tool for transcriptomics. Unlike microarrays, it has a very wide dynamic range and low background noise; can be used to examine gene fusion events; doesn't require probe design before the experiment; allows for discovery of novel splicing events, exons, and even genes; and provides more precise expression levels with single base resolution. In addition, RNA-seq provides for quantification of allele-specific and isoform-specific expression levels. Given that alternative splicing is a major mechanism generating protein diversity and has been shown to be active in over 90% of human genes, the quantification of isoform-level expression is particularly important; and consequently so too are powerful and efficient methods for identifying DE isoforms.

- A number of methods exist for identifying DE at the gene level from RNA-seq data (DESeq, edgeR, baySeq, BBSeq, FDM). **However, simply applying the methods developed for RNA-seq gene DE analysis to isoform level data results in reduced power in some cases and significantly increased false discoveries in others.** This is due to mapping uncertainty, isoform structure and biases inherent to RNA-seq reads.

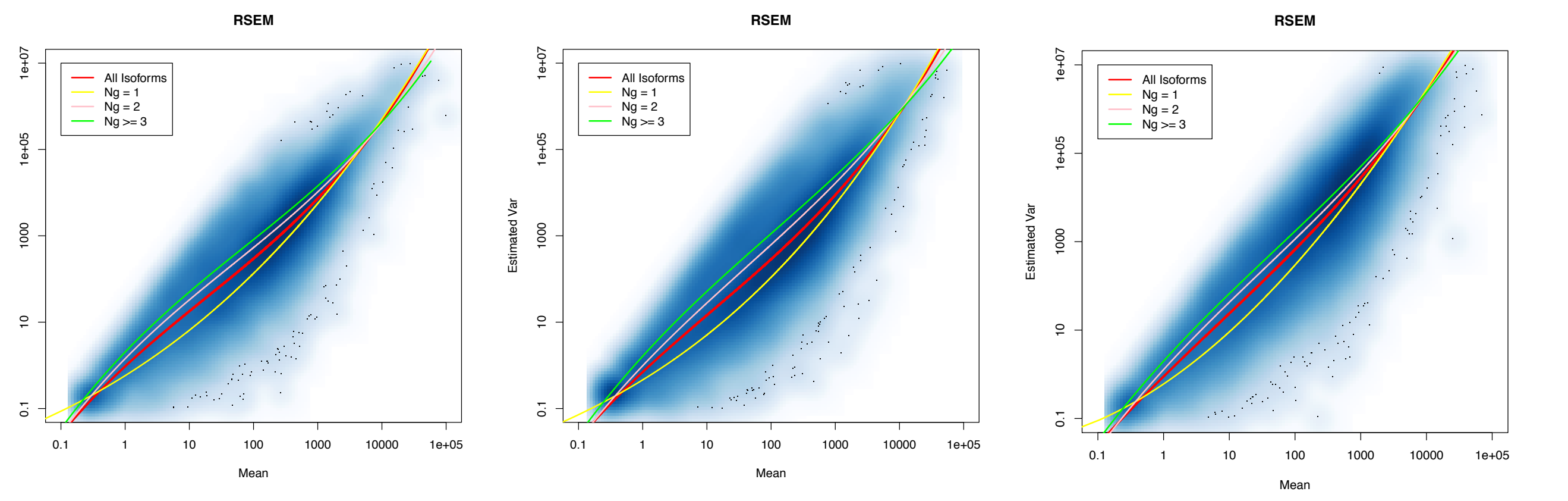


The figures above are isoform level expressions from Gould lab data (preprocessed by RSEM). We use Ng to denote the number of isoforms within a gene. The plots show that on the isoform level data, the mean-variance relationship depends on whether or not the isoform is coming from a gene with a single or multiple isoforms (Ng = 1 or Ng > 1).

If gene DE methods are applied to the isoform level data directly, the fit from all the data (red line) is used to estimate the common mean-variance relationship for each isoform.

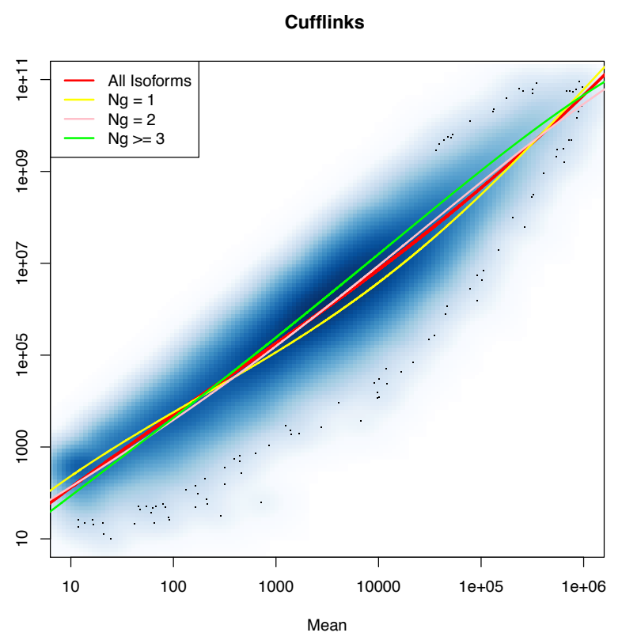
However, for isoforms with Ng = 1, the real relationship should follow the yellow line and the mean-variance relationship is overestimated by the full-data fit (red line) leading to low power.

When Ng > 1, the mean-variance relationship is underestimated leading to false discoveries.



MinAmp T7LA ILMN TrueSeq

The figures above show isoform expression from Thomson Lab data (preprocessed by RSEM). The data are from human H1 cells, each using different priming protocols. The Ng effect appears consistently.



The Ng effect is also present in the dataset shown on the left. This data set is a publicly available data set derived from human heart cells. (NCBI GEO GSM792454-GSM792461) The data are processed using TopHat and Cufflinks.

METHODS

The isoform level model:

s : Sample	$X_{gi,s}$: Isoform i expression in gene g and sample s
g : Gene	$r_{gi,0}$: Isoform specific parameter shared by all the samples
i : Isoform	p_0 : The prior probability of being EE
l_s : Library size parameter.	p_i : The prior probability of being DE

Assume:

$$X_{gi,s} \mid r_{gi,s}, q_{gi}^{C} \sim NB(r_{gi,s}, q_{gi}^{C}) \equiv NB\left(\mu_{gi,s} = \frac{r_{gi,s}(1-q_{gi}^{C})}{q_{gi}^{C}}, \sigma_{gi,s}^{2} = \frac{r_{gi,s}(1-q_{gi}^{C})}{(q_{gi}^{C})^{2}}\right)$$

$$q_{gi}^{C} \mid \alpha, \beta^{N_{gi}, b_{gi}} \sim Beta(\alpha, \beta^{N_{gi}, b_{gi}}) \text{ and } r_{gi,s} = l_s \bullet r_{gi,0}$$

The isoform is EE if $q_{gi}^{C1} = q_{gi}^{C2}$; is DE if $q_{gi}^{C1} \neq q_{gi}^{C2}$; then $X_{gi} \sim p_0 f_0(X_{gi}) + p_1 f_1(X_{gi})$ where

$$\text{EE:} \quad f_0(X_{gi}) = \int \prod_{X_{gi,s} \in X_{gi}} P(X_{gi,s} \mid r_{gi,s}, q) P(q \mid \alpha, \beta^{N_{gi}, b_{gi}}) dq$$

$$\text{DE:} \quad f_1(X_{gi}) = \int \prod_{X_{gi,s} \in X_{gi}^{C1}} P(X_{gi,s} \mid r_{gi,s}, q) P(q \mid \alpha, \beta^{N_{gi}, b_{gi}}) dq \bullet \int \prod_{X_{gi,s} \in X_{gi}^{C2}} P(X_{gi,s} \mid r_{gi,s}, q) P(q \mid \alpha, \beta^{N_{gi}, b_{gi}}) dq$$

Of primary interest is $P(DE \mid X_{gi}) = \dfrac{p_1 f_1(X_{gi})}{p_0 f_0(X_{gi}) + p_1 f_1(X_{gi})}$

The gene level model is similar but with a β shared by all the genes.

SIMULATION RESULTS

Isoform Simulation Without Artifacts

We follow a simulation set-up similar to that in the Robinson and Smith (2007) (EdgeR) paper with the assumption that isoform counts within condition are $X_{gi,s} \sim NB(\mu_{gi,s} = l_s \mu_{gi}^{C}, \sigma_{gi,s}^{2} = l_s \mu_{gi}^{C}(1 + \mu_{gi}^{C} \phi_{gi}))$. In which μ_{gi}^{C} is sampled from the empirical ones from all the data. ϕ_{gi} is sampled from the empirical ones within each Ng group. l_s is simulated from a uniform (0.8, 1.3). 10% of the isoforms are simulated as DE.

DESeq, edgeR, bayeSeq and BBSeq are applied to all of the isoforms at once, and then to each Ng group individually.

Results are averaged across 20 simulations, with thresholds from each method chosen to control an overall false discovery rate at 5%.

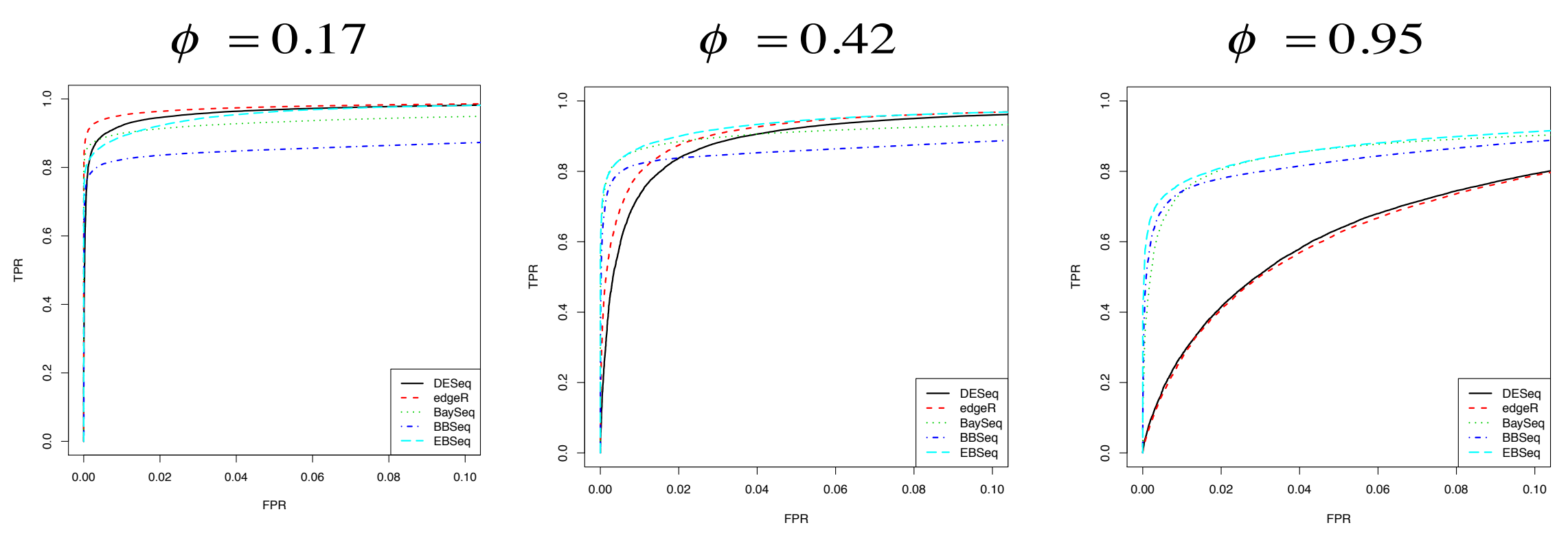
Table 1 shows that DESeq and edgeR are relatively underpowered for isoforms from Ng = 1 genes (compared to Ng = 2 and Ng = 3), with increased FDR for Ng = 2 and Ng = 3. EBseq shows improved power over baySeq and BBSeq, and although power is slightly reduced compared to DESeq and edgeR, the FDR is well controlled.

Table 1. Isoform level simulation without artifacts

	Ng=1	Ng=1	Ng=2	Ng=2	Ng=3	Ng=3
	Power	FDR	Power	FDR	Power	FDR
baySeq	0.64	0	0.62	0	0.55	0.01
baySeq Each	0.67	0	0.63	0	0.50	0.01
BBSeq	0.62	0.01	0.61	0.04	0.56	0.04
BBSeq Each	0.62	0.04	0.62	0.03	0.53	0.04
DESeq	0.78	0.02	0.86	0.24	0.89	0.29
DESeq Each	0.80	0.08	0.77	0.07	0.74	0.07
edgeR	0.79	0.02	0.86	0.18	0.88	0.24
edgeR Each	0.80	0.09	0.76	0.06	0.72	0.07
EBSeq	0.70	0.05	0.73	0.07	0.70	0.08

Gene Simulation Without Artifacts

The first gene level simulation set-up is similar to the isoform simulation, but with constant ϕ_g across all the genes. The figure below shows that the power of all methods decreases as the dispersion increases. However, with large dispersion parameters, EBSeq and baySeq perform better than edgeR and DESeq.



A more realistic simulation has been done by sampling the ϕ_g from the empirical ones calculated using Gould lab data. Table 2 shows results.

Table 2. Gene level simulation without artifacts

	Power	FDR
baySeq	0.71	0
BBSeq	0.7	0.02
DESeq	0.91	0.22
edgeR	0.89	0.15
EBSeq	0.79	0.05

Table 2 shows that EBSeq has increased power over baySeq and BBSeq. And although power is lower than DESeq and edgeR, FDR is well controlled.

Gene Simulation With Artifacts

Due to the PCR artifacts, priming bias or multi-read assignment variation, gene expression in a specific sample may be much higher than in others for the same gene within the same condition. To assess the potential impact of such outliers on our approach, we performed a simulation study. In addition to 10% DE genes, we include an additional 10% of genes that are EE, but contain an artifact (or outlier). For these latter genes, we define the expression of one of 10 samples to be 10 times its previous value. **Table 3** shows that DESeq and edgeR are affected by the artifact genes with decreasing power and increasing FDR. That is because their tests are both based on the summation of counts or pseudo counts within condition, by the assumption that the summation of independent NB distributed random variables with similar parameter *r* is still negative binomial distributed. EBseq is much more robust to outliers.

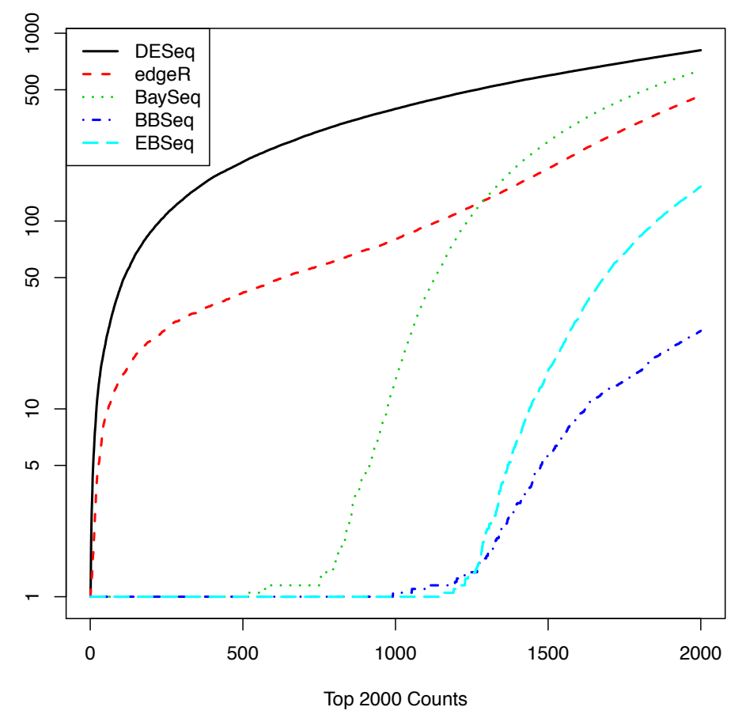


Table 3. Gene level simulation with artifacts

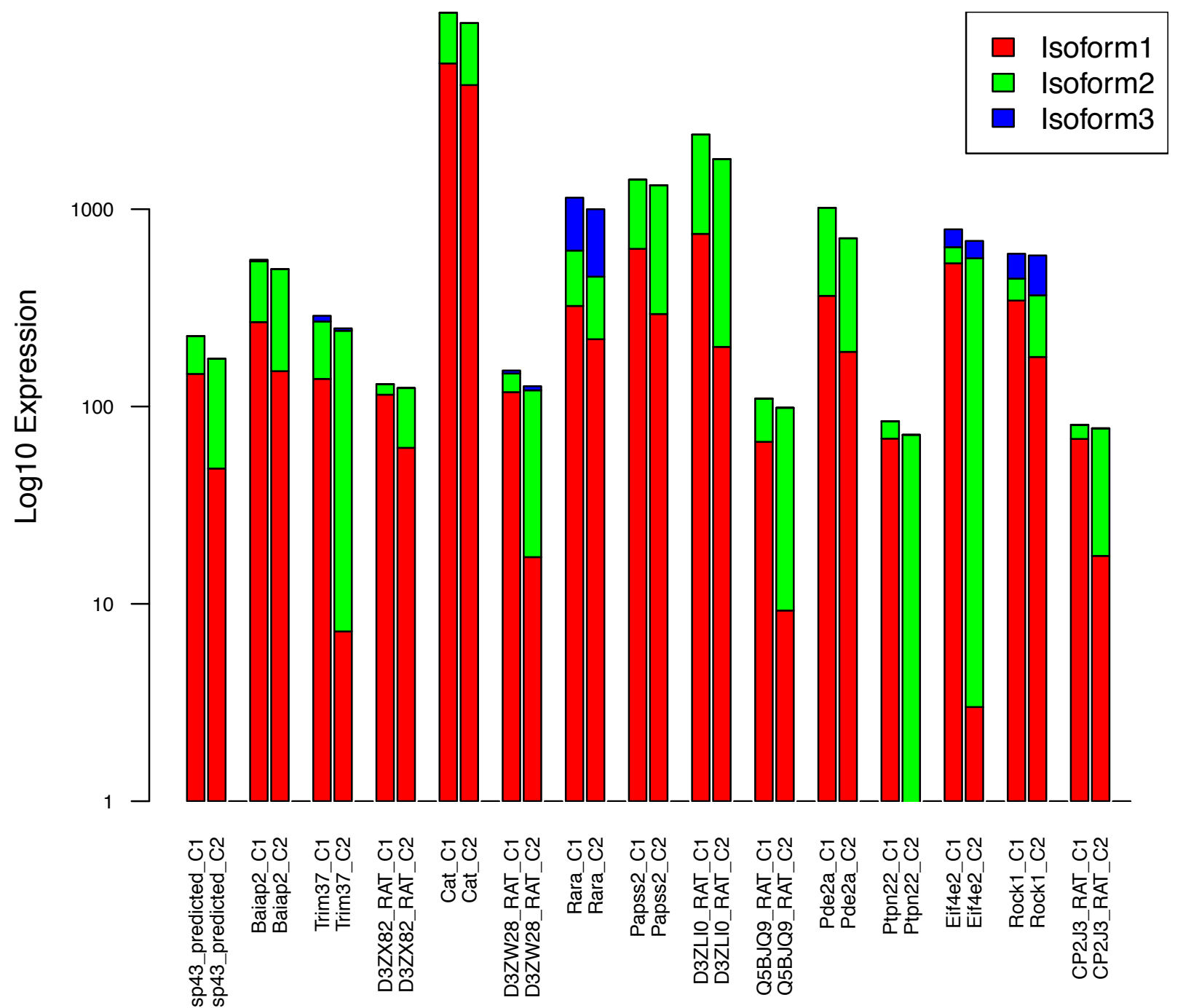
	Power	FDR
baySeq	0.63	0.04
BBSeq	0.69	0.01
DESeq	0.77	0.50
edgeR	0.85	0.36
EBSeq	0.79	0.05

CASE STUDY RESULTS

Of general interest in the Gould lab is the identification of the genetic factors underlying breast cancer. In this particular study, we consider Gould lab data where mRNA from 8 rats is obtained, 4 in each of two conditions (congenic rats carrying the resistant and susceptible Mcs1a allele). There are 20,267 expressed genes in total and 25,268 isoforms with group sizes 15315, 6908 and 3045 corresponding to Ng = 1, 2 and 3, respectively. EBSeq, baySeq, DESeq, and edgeR are each applied.

As a strength of EBSeq is the ability to identify both gene and isoform level DE, we first consider results where DE isoforms are identified in EE genes. Over 100 such DE isoforms were identified. The figure below shows 17 such isoforms.

To assess the effect of outliers, we identified genes whose maximum expression is greater than 10 times its minimum expression within condition. Out of 450 genes containing at least one outlier, EBSeq makes the fewest DE calls. In particular, baySeq, DESeq, edgeR and EBSeq identified 43, 133, 107 and 37 of these genes with outliers to be DE.



CONCLUSIONS

- Gene level DE methods are not optimal on isoform level data. Current methods cannot handle different mean-variance relationship between different isoform groups due to the mapping uncertainty, isoform sturucture and priming bias.

- We've developed an empirical bayes model for both isoform and gene level DE analysis. EBSeq accounts for and capitalizes on features in isoform level data and is more robust to outliers on both gene-level and isoform-level inference.

Contact Information:

Ning Leng
nleng@wisc.edu

Christina Kendziorski
kendzior@biostat.wisc.edu

Acknowledgement:

GM076274