Supplement to "EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments"

1 Package/annotation versions

Unless noted otherwise, calculations are carried out in R [14], version 2.14.1. The versions for each package considered in the manuscript are as follows: baySeq: 1.1.0 (default normalization method (upper quartile normalization) is used); DESeq: 1.8.2 (sharingMode="maximum"; default normalization method (median normalization) is used); edgeR: 2.6.3 (tag-wise overdispersion estimation and default normalization method (TMM) are used); BitSeq: 1.2.1 (default normalization method (total number of reads) is used);bowtie: 0.12.8; TopHat: 2.0.3; Cufflinks: 2.0.1 (default parameters and normalization method (upper quartile normalization) are used); HTSeq: 0.5.3p9; RSEM: 1.1.20; RSeq: 0.0.7. Splines were estimated using the polynomial fitting function in *lm*. Human genome annotation hg18 from RefSeq was used in read simulations. To obtain the ROC curves, genes/isoforms are sorted ascending by adjusted p-value (DESeq, edgeR, Cuffdiff2) or descending by posterior probability of being DE (EBSeq, baySeq). In BitSeq, to obtain the isoform ranking in favor of two-sided DE test via the one-sided statistic PPLR, we sorted the isoforms descending by max(PPLR, 1-PPLR).

2 Two additional data sets used in Figure 2c

2.1 Gould Lab Data

RNA-seq data was obtained from two groups of congenic rats (four samples in each condition) harboring the susceptible or resistance allele of the mammary carcinoma susceptibility locus (Mcs1a) ([5]). For these experiments, mammary glands are taken from 8 untreated, mammary cancer-free females per genotype. The tissue is disaggregated using physical shearing in a solution of Trireagent (Ambion). RNA is extracted using a total RNA extraction kit (Ambion). RNA integrity

is monitored using a 2100 Bioanalyzer (Agilent). Equal RNA (approximately 5 μg) from 2 rats is pooled to obtain a single sample for one RNA-seq lane. A total of four samples per genotype (*Mcs1a* susceptible or resistant) were processed by the University of Wisconsin Biotechnology Gene Expression Center using the Illumina Genome Analyzer IIx. Reads are post-processed to a length of 30 basepairs and aligned to the rat Ensemble RGS3.4 transcripts using Bowtie ([9]), allowing for up to 100 multiple matches and one mismatch with seed length 30. Expression is estimated using RSEM ([11, 10]).

2.2 Smith Lab Data

Tophat output files were downloaded from GEO GSM792454-61. Eight samples (4 in each of two conditions) are considered here. In short, RNA was extracted from atrial tissue samples and prepared using Illumina's mRNA protocol. The reads are single-end with read length 36-bp. Each sample was run on one lane of an Illumina Genome Analyzer IIx. Alignment was done using Bowtie and TopHat (without de novo transcript detection) with the hg19 RefSeq annotation. Isoform expression was estimated using Cufflinks ([18]).

3 Assessing the I_g effect in multiple data sets

We evaluate differences among I_g groups in multiple single-end and paired-end data sets processed under different priming protocols, in different labs, using different isoform expression estimation methods, using different definitions of isoform complexity, and for a wide range of sample sizes (from four to sixty-nine).

Supplementary Figure 1 shows data from James Thomson's lab at the Morgridge Institute for Research at UW-Madison. The data sets are distinct from those shown in the manuscript. For these experiments, RNA was extracted from human embryonic stem cell line H1 and prepared using the *Illumina TrueSeq*, T7LA[17], and the MinAmp (Thomson Lab internal) protocols, respectively. For each protocol, three samples were considered. Each sample was run on one lane of an Illumina Genome Analyzer IIx; the reads are single-end with read length 42-bp. Alignment was done using

Bowtie with the hg18 RefSeq annotation. Isoform expression was estimated using RSEM.

Supplementary Figures 2(a) and 2(e) show data from Michael Gould's lab at UW-Madison, detailed in Section 2.1 of this Supplement.

Supplementary Figures 2(b) and 2(f) show data from the Wold lab [12]. RNA was extracted from mouse brain tissue and two replicates were prepared using the Solexa protocol. For each replicate, random primers were used. The reads are single-end with read length 25-bp. Alignment was done using Bowtie and Tophat (without de novo transcript detection) with the UCSC mm9 annotation. Isoform expression was estimated using Cufflinks with multi-read correction.

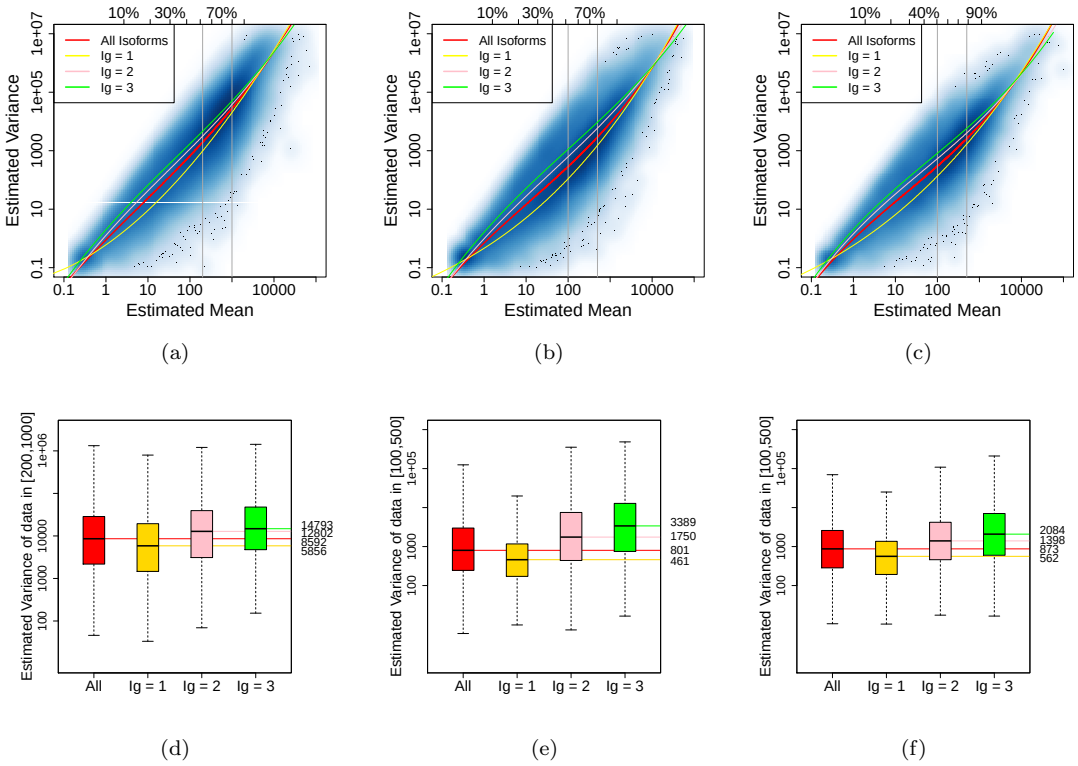
Supplementary Figures 2(c) and 2(g) show data from the MicroArray Quality Control (MAQC) experiment [3] that is distinct from what is shown in the manuscript. For these figures, raw read files (fasta format) were downloaded from GEO GSM475204-09. RNA was extracted from human brain tissue and 3 replicates were considered. For each replicate, random primers were used. The reads are paired-end with read length 50-bp. Each sample was run on one lane of an Illumina Genome Analyzer IIx. Alignment was done using SeqMap [7] with the hg18 RefSeq annotation. Isoform expression was estimated using RSeq [8].

Supplementary Figures 2(d) and 2(h) show data from Pickrell *et al.* [13]. RNA was extracted from Yoruba Hapmap cell lines and 69 samples were prepared using the Illumina Genome Analyzer II. For each replicate, random primers were used. The reads are single-end with read length 35-bp. Raw read files (fasta format) were downloaded from http://eqtl.uchicago.edu. Only Yale data are used and for the subjects assayed twice only the first replicate is used. Alignment was done using Bowtie with the hg18 annotation. Isoform expression was estimated using RSEM with multi-read correction.

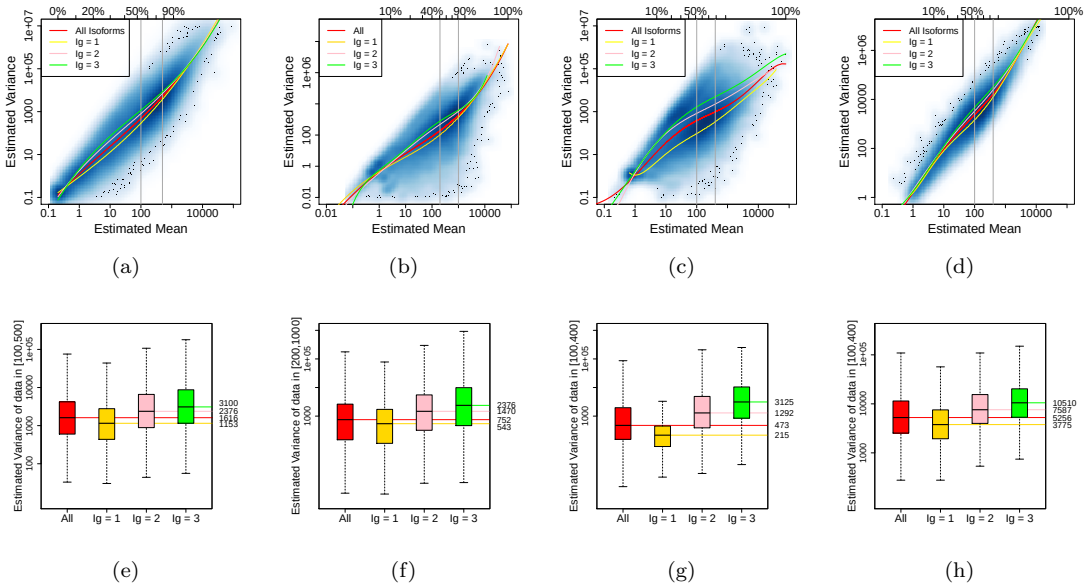
Supplementary Figure 3 shows results using an alternative method to define isoform complexity. Instead of I_g as defined in the manuscript, the unmappability score of each isoform is obtained from RSEM, and the unmappability scores are clustered to group isoforms. Panel (a) shows the results from K-means clustering with 3 centers; panel (b) shows the results from a Gaussian Mixture Model.

Recall that Figure 1(c) shows spline fits which are similar to the approaches used by DESeq and edgeR to estimate variance. Supplementary Figures 4(a), 4(b) and 4(c) show the exact estimators

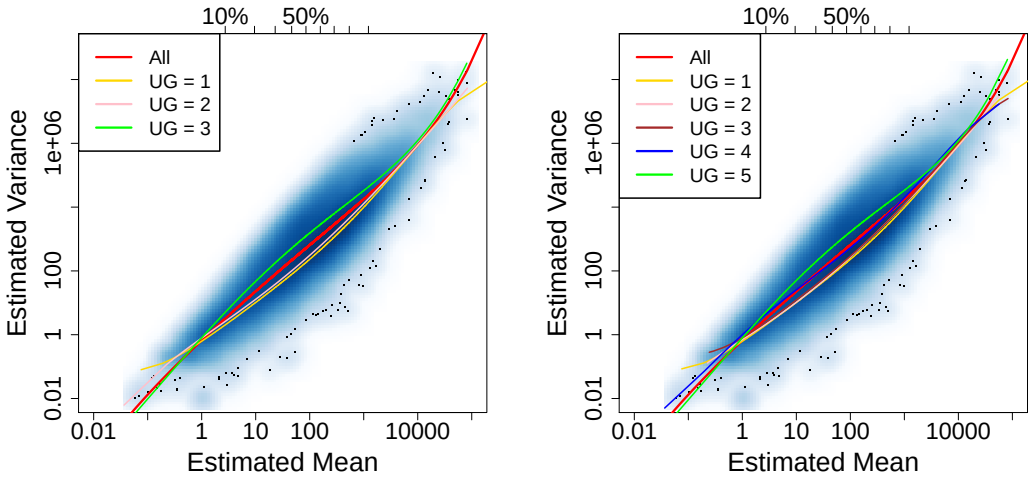
used in DESeq, and edgeR (both the common-dispersion model and the tag-wise-dispersion model) derived using the data from the ESC vs. iPSC experiment that is shown in Figure 1.



Supplementary Figure 1: Panel (a) shows the empirical variance vs. mean for each isoform profiled in the experiment comparing ESCs with iPSCs (TrueSeq Protocol); details of this experiment are given earlier in this Supplement. A spline fit to all isoforms is shown in red with splines fit within the $I_g = 1$, $I_g = 2$, and $I_g = 3$ isoform groups shown in yellow, pink, and green, respectively. Panels (b) and (c) are similar to (a), but for data processed under the T7LA and MinAmp protocols, respectively. The estimated variance of isoforms with average expression in 50^{th} and 80^{th} percentiles of expressions are shown in (d), (e), (f).

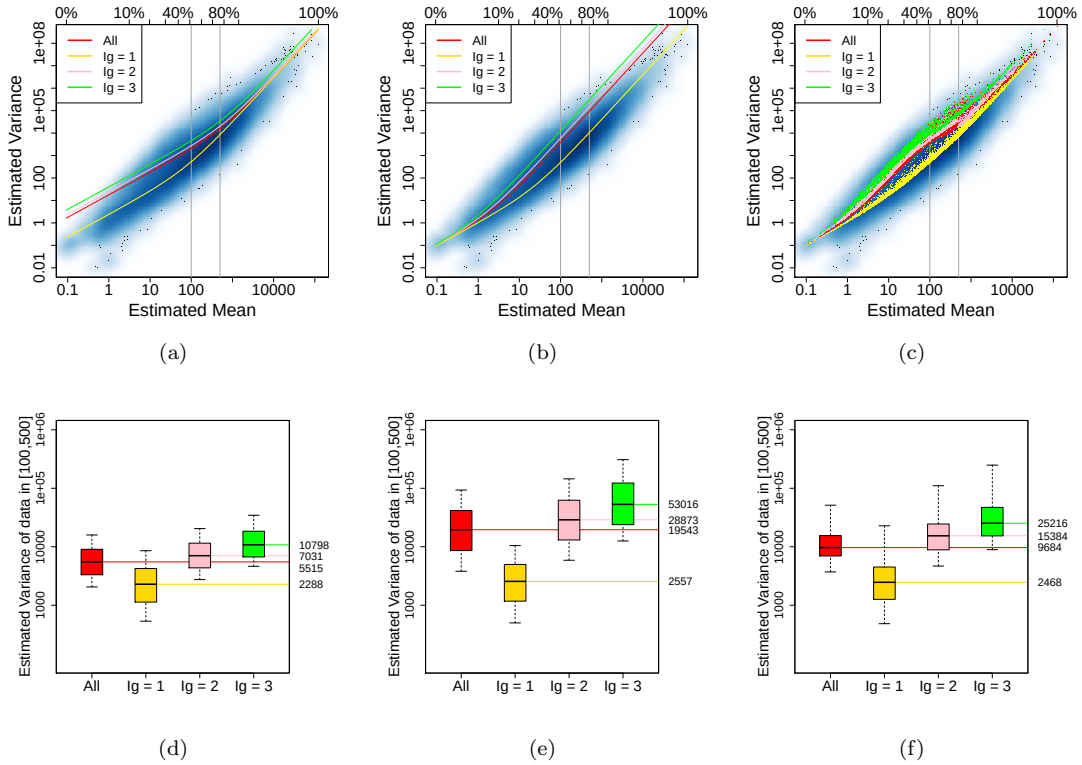


Supplementary Figure 2: Shown are plots similar to Supplementary Figure 1 generated using data from the Gould lab (panel (a)) processed by RSEM, data from Wold lab (panel (b)) processed by Cufflinks, , MAQC data from Wong lab (panel(c)) processed by RSeq and data from Pickrell *et al.* (panel (d)); details of these experiments are given earlier in this Supplement. The estimated variance of isoforms with average expression in the 50^{th} to 80^{th} percentiles of expressions are shown in (e), (f), (g), (h).



(a) (b)

Supplementary Figure 3: Shown are plots similar to Figure 1(c), but with uncertainty groups obtained by K-means clustering of unmappability scores instead of I_g groups. The unmappability score of each isoforms as well as the isoform expected counts are obtained from RSEM. Panel (a) shows the results using K-means clustering with 3 centers. Panel (b) shows the results using a Gaussian Mixture Model.



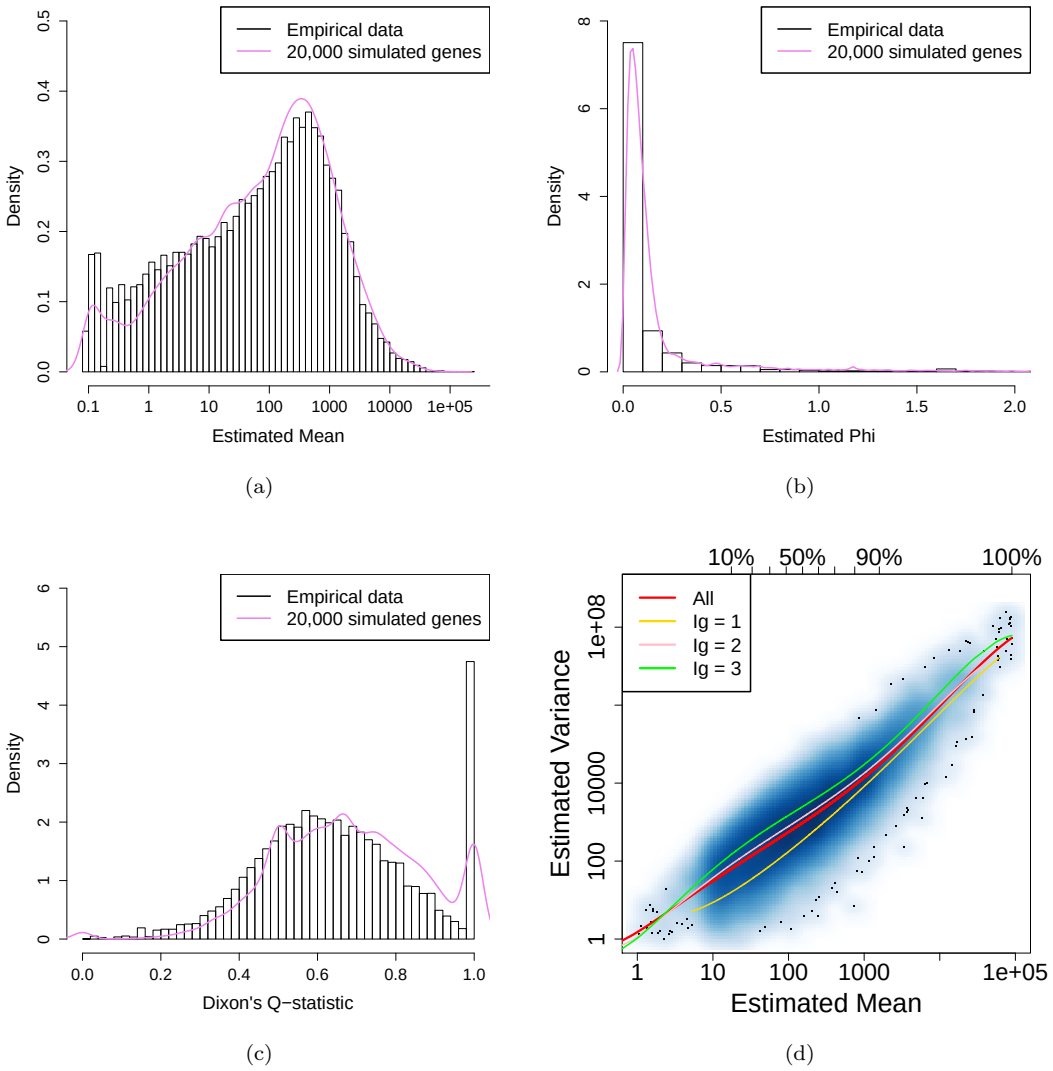
Supplementary Figure 4: Recall that Figure 1(c) shows spline fits which are similar to the approaches used by DESeq and edgeR to estimate variance. This figure shows the exact estimators used in DESeq, and edgeR (both the common-dispersion model and the tag-wise-dispersion model) derived using the data from the ESC vs. iPSC experiment that is shown in Figure 1(c). Specifically, panel (a) shows the fitted dispersion values provided by DESeq. The dispersion line is calculated across all isoforms (red) and within I_g group (shown in yellow, pink, and green, respectively). Panels (b) and (c) show similar plots from edgeR under their common dispersion (panel (b)) and tag-wise dispersion (panel (c)) models. Panels (d), (e) and (f) consider average expression in [100, 500]. The range was chosen as it approximates the 50^{th} and 80^{th} percentiles of expression across all isoforms. Shown are box-plots of the variances of these isoforms collectively, and within I_g group.

4 Simulations

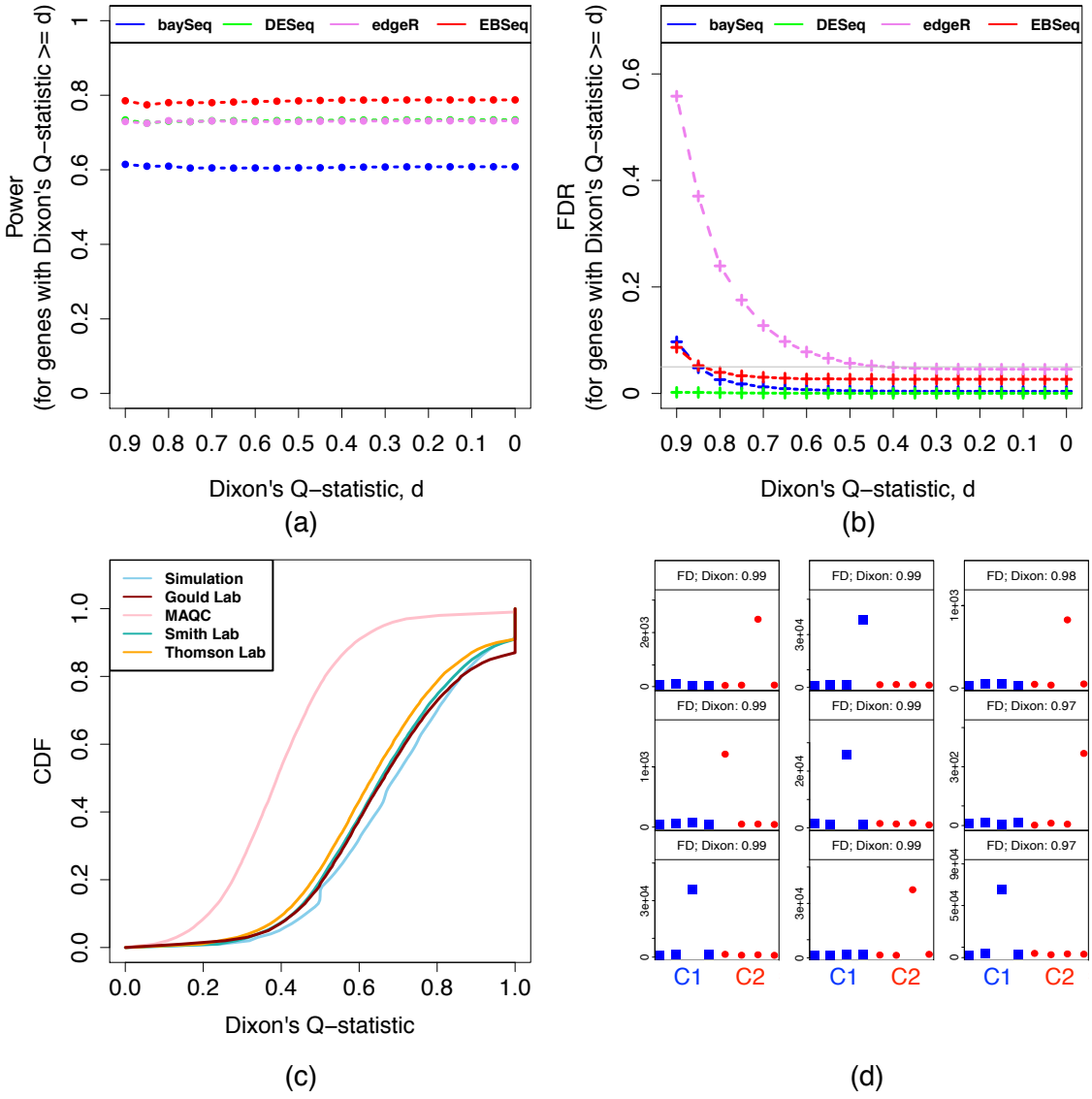
We followed the simulation setup of [16] by defining counts as Negative Binomial with isoform-specific mean in sample s and condition C given by $l_s\mu_{g_i}^C$ and variance $l_s\mu_{g_i}^C(1 + l_s\mu_{g_i}^C\phi_{g_i})$. For the isoform study, we simulated 30,802 isoforms, four lanes in each of two conditions. Sample sizes were taken to match those observed in the case study comparing ESCs with iPSCs; parameter values were sampled from empirical ones in that study. The percentages of DE isoforms were set at 2%, 4% and 5% in the $I_g = 1, 2$ and 3 groups, also to match the case study data. The empirical values from the isoforms belonging to same genes are sampled together to preserve dependence. For half of the DE isoforms, $\mu_{g_i}^{C1} = \delta_{g_i}\mu_{g_i}^{C2}$ with δ_{g_i} sampled from the 95%-97% quantile of the empirical isoform fold changes. For the other half, $\mu_{g_i}^{C2} = \delta_{g_i}\mu_{g_i}^{C1}$. The gene level simulation is similar, with 2% of the genes set to be DE. In addition to sampling μ and ϕ, for the gene-level simulation we also matched the simulated data to each gene's Dixon Q-statistic by adjusting one value when necessary (see the Methods section on Identification of Outliers for the definition of Dixon's Q-statistic). The library size factors for both the isoform and gene-level simulations were randomly simulated from Uniform $(0.8, 1.3)$. One hundred simulated data sets were generated for each scenario considered. Supplementary Figure 5 demonstrates that characteristics observed in the case study data are reproduced in the simulated data sets.

Table 2 in the manuscript reports that count-based methods have well-controlled FDR. Supplementary Figure 6 shows that the likelihood of a false call increases in the presence of outliers, especially for edgeR. In particular, Supplementary Figure 6, panels (a) and (b) evaluate the operating characteristics shown in Table 2 within subsets of genes grouped by their Dixon's Q-statistic ([4]). As detailed in Methods, a gene harboring an outlier will have a Dixon's Q-statistic near one. Panel (b) of Supplementary Figure 6 shows that FDR is relatively constant for most methods when outliers are present, with the exception of edgeR, where FDR increases substantially with increases in Dixon's Q-statistic. Panel (c) of Supplementary Figure 6 shows that values of the Dixon's Q-statistic considered in Sim III are consistent with those observed in many data sets (the MAQC data set has fewer outliers given it is comprised of technical, not biological, replicates). Panel (d) provides an example of the types of genes identified by edgeR having high Dixon's Q-statistic. Specifically, shown are the nine genes with highest Dixon's Q-statistics in those exclusively identified by edgeR. Although FDR is well-controlled for edgeR overall (detailed in Table 2), these figures suggest that the majority of false discoveries that are identified by edgeR are likely in genes harboring outliers.

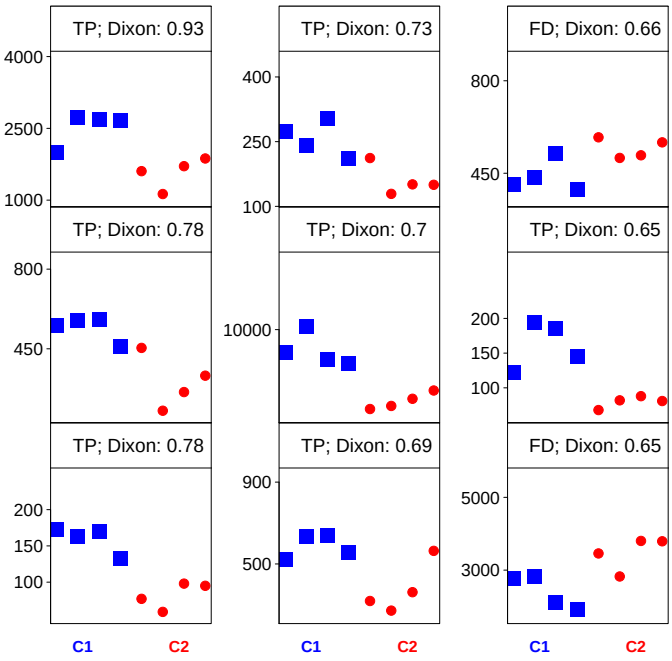
Genes identified exclusively by EBSeq having highest Dixon's Q-statistic are shown in Supplementary Figure 7.



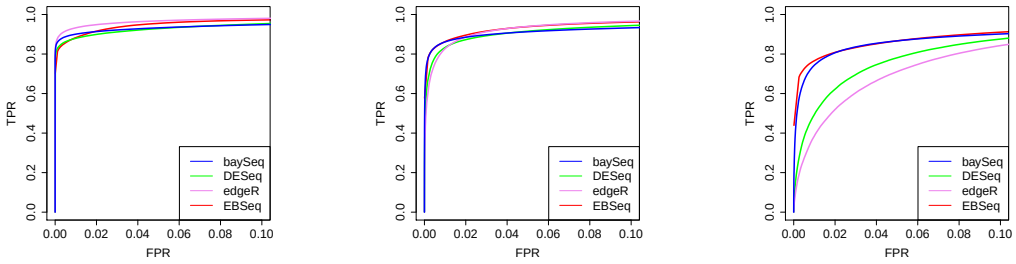
Supplementary Figure 5: Panels (a)-(c) show the distribution of μ, ϕ and Dixon's Q-statistic, d_g, comparing one simulated data set from Sim III (histogram) with the empirical data from the experiment comparing ESCs with iPSCs (density, pink line). Panel (d) shows the scatter plot shown in Figure 1, but from one of the simulated data sets from Sim I.



Supplementary Figure 6: Panel (a) and (b) show the operating characteristics of baySeq, DESeq, edgeR and EBSeq on subsets of genes averaged across 100 Sim III data sets for target FDR set at 5%. The subsets are defined as genes with Dixon's Q-statistic greater than the value given on the x-axis. Panel (c) shows the cumulative distribution function (CDF) of Dixon's Q-statistic in 4 empirical data sets as well as the CDF averaged across 100 simulations. Panel (d) shows 9 genes identified exclusively by edgeR having highest Dixon's Q-statistic for one simulated data set. The blue and red points correspond to two different conditions. The y-axis shows the normalized gene expression. The legend within each box shows whether the gene is a true positive (TP) or false discovery (FD) as well as the corresponding Dixon's Q-statistic value.



Supplementary Figure 7: Shown are nine genes with highest Dixon's Q-statistics of those exclusively identified by EBSeq on one simulation of Sim III (the same one as in Supplementary Figure 6 (d)). Blue and red points are samples from condition 1 and 2. The y-axis shows the gene expression counts. The legend within box shows whether the gene is a true positive (TP) or false discovery (FD) as well as the corresponding Dixon's Q-statistic.



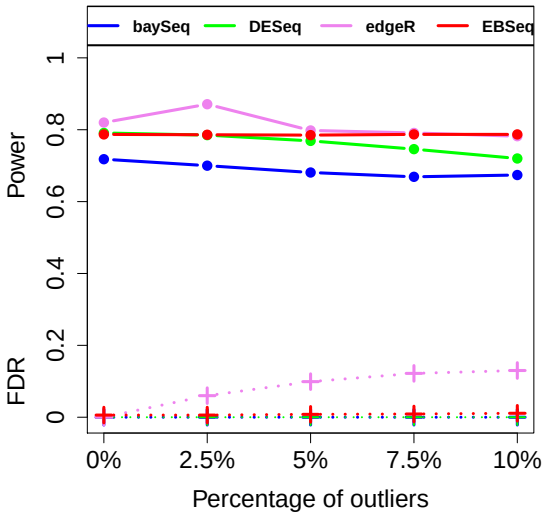
(a) $\phi_g = 0.17$ (b) $\phi_g = 0.42$ (c) $\phi_g = 0.95$

Supplementary Figure 8: ROC curves averaged over 100 simulations using the simulation set up shown in Figure 3 of Robinson and Smyth[16] and Figure 2 of Hardcastle *et al.*[6].

In addition to the simulations detailed in the manuscript, we reproduced the set-up shown in Figure 3 of Robinson and Smyth[16] and Figure 2 of Hardcastle *et al.* [6] by fixing $\phi_g = 0.17, 0.42$ and 0.95, as done in their work. For ten lanes (five in each of two conditions), we simulated data for 10,000 genes in which 50% were defined to be DE, also as in their work. For each value of ϕ, l_s's were sampled from Uniform(0.8, 1.3) and μ_g^C's were randomly sampled from the empirical means in the experiment comparing ESCs with iPSCs.

Supplementary Figures 8(a), 8(b) and 8(c) show the ROC curves for baySeq, DESeq, edgeR, and EBSeq, averaged over 100 simulations, for each of the three overdispersion parameters. As shown, baySeq and EBSeq perform consistently across the three overdispersion values.

To further evaluate how outliers affect the operating characteristics of each method, we investigate 5 simulations with outlier percentages 0%, 2.5%, 5%, 7.5% and 10%. Gene counts were simulated as in Sim III, but for a gene with an outlier, we modify the expression of one random sample to match a Dixon's Q-statistic randomly sampled from the 90^{th}-98^{th} percentile of Dixon Q-statistics (Dixon's Q-statistic is defined as in Methods). Supplementary Figure 9 shows average power and FDR across 100 simulations for each of the five sets.

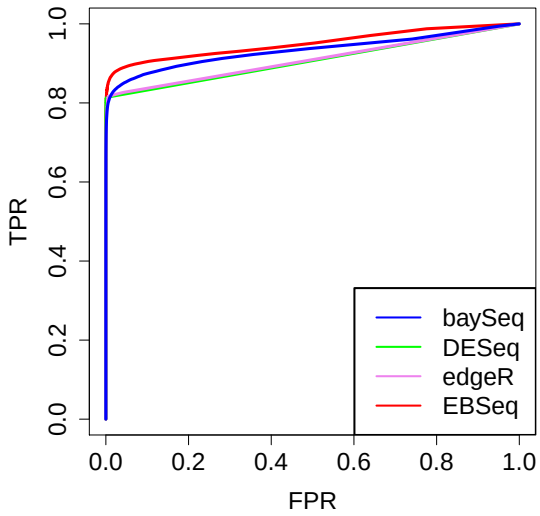


Supplementary Figure 9: Shown are Power and FDR averaged across 100 simulations for each of five sets of simulations, described earlier in this Supplement. Each set contains 0%, 2.5%, 5%, 7.5% or 10% outliers. Target FDR was set at 5%.

Supplement Table 1: Applying gene level (count-based) methods on simulated isoform data

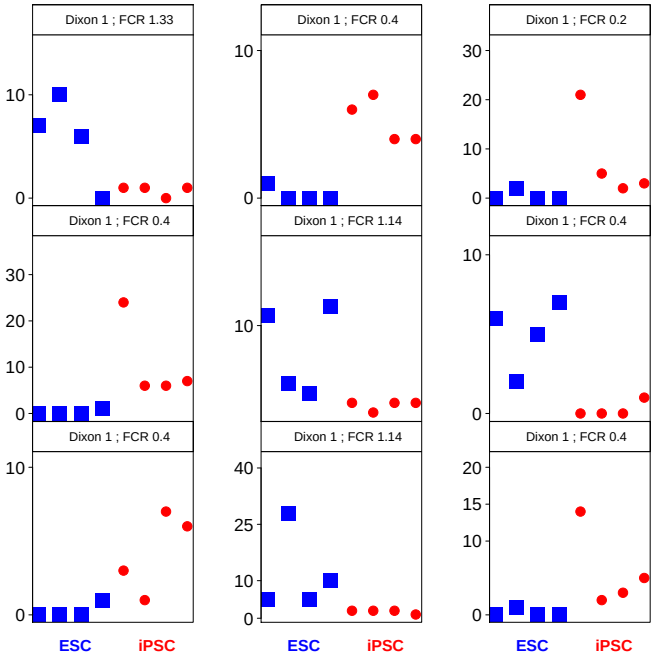
		baySeq	DESeq	edgeR	EBSeq
All Isoforms	Power	56.9%	72.3%	81.2%	81.4%
	FDR	0%	0.5%	13.1%	5.0%
$I_g = 1$ Isoforms	Power	**53.3%**	**56.8%**	**60.9%**	**78.2%**
	FDR	0%	0%	0%	0.5%
$I_g = 2$ Isoforms	Power	56.3%	75.2%	84.3%	83.1%
	FDR	0%	1.2%	4.6%	5.7%
$I_g = 3$ Isoforms	Power	59.4%	78.1%	90.3%	81.5%
	FDR	0.5%	0.9%	17.3%	7.0%

Power and FDR averaged across 100 isoform simulations. Thresholds were chosen to control FDR at 5% for each approach. Count-based DE methods are significantly underpowered in the $I_g = 1$ group when applied directly to estimates of isoform expression.

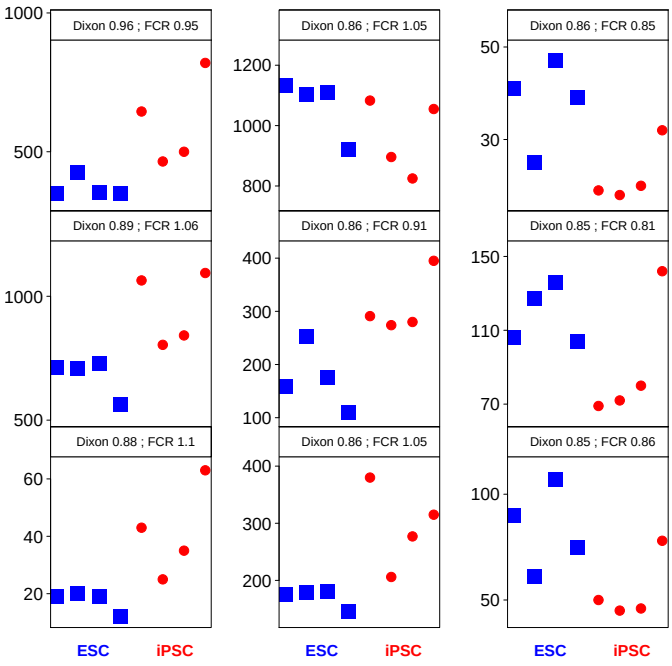


Supplementary Figure 10: Shown are ROC curves for baySeq, DESeq, edgeR and EBSeq averaging 100 Sim III data sets.

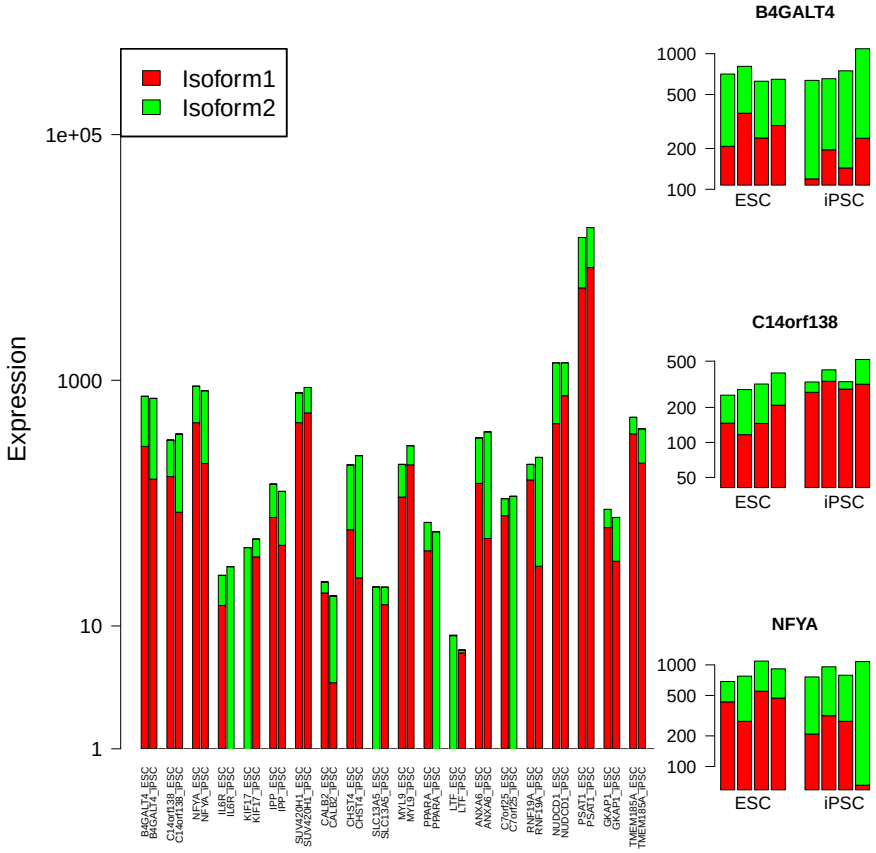
5 Case Study Results



Supplementary Figure 11: Shown are the top nine genes with highest Dixon's Q-statistics exclusively identified by edgeR in the experiment comparing ESCs with iPSCs. Blue and red points are samples from ES cell lines and iPS cell lines, respectively. The y-axis shows gene expression counts. Dixon's Q-statistic and fold change ratio (FCR) are shown in the legends.



Supplementary Figure 12: Shown are the top 9 genes with highest Dixon's Q-statistics exclusively identified by EBSeq in the experiment comparing ESCs with iPSCs. Blue and red points are samples from ES cell lines and iPS cell lines. The y-axis shows gene expression counts. Dixon's Q-statistic and fold change ratio (FCR) are shown in the legends.



Supplementary Figure 13: The left panel shows 20 genes in the $I_g = 2$ group identified as EE by EBSeq (gene level posterior probability of EE > 0.95) with DE isoforms (isoform level posterior probability of DE > 0.95) in the experiment comparing ESCs with iPSCs. Each bar shows the isoform expression in each condition; expression of the constituent isoforms is shown in different colors within each gene. The right panel shows 3 example EE genes with DE isoforms. Each bar shows the isoform expression in each sample; expression of the constituent isoforms is shown in different colors within each gene.

6 Parameter estimation and multiple group analysis

As in the text, we let $X_{g_i}^{C1} = X_{g_i,1}, X_{g_i,2}, ..., X_{g_i,S_1}$ denote data from Condition 1 and $X_{g_i}^{C2} = X_{g_i,(S_1+1)}, X_{g_i,(S_1+2)}, ..., X_{g_i,S}$ data from Condition 2. We assume that counts within condition C are distributed as Negative Binomial: $X_{g_i,s}^{C}|r_{g_i,s}, q_{g_i}^{C} \sim NB(r_{g_i,s}, q_{g_i}^{C})$ where

$$P(X_{g_i,s}|r_{g_i,s}, q_{g_i}^{C}) = \binom{X_{g_i,s} + r_{g_i,s} - 1}{X_{g_i,s}} (1 - q_{g_i}^{C})^{X_{g_i,s}} (q_{g_i}^{C})^{r_{g_i,s}} \tag{1}$$

and $\mu_{g_i,s}^{C} = r_{g_i,s}(1 - q_{g_i}^{C})/q_{g_i}^{C}$; $(\sigma_{g_i,s}^{C})^2 = r_{g_i,s}(1 - q_{g_i}^{C})/(q_{g_i}^{C})^2$.

We assume a prior distribution on $q_{g_i}^{C}$:$q_{g_i}^{C}|\alpha, \beta^{I_g} \sim Beta(\alpha, \beta^{I_g})$. The hyper parameter α is shared by all the isoforms and β^{I_g} is I_g specific. We further assume that $r_{g_i,s} = r_{g_i,0} l_s$. I.e., $r_{g_i,0}$ is an isoform specific parameter common across conditions. Of interest is distinguishing between EE and DE (two expression patterns) where EE refers to $q_{g_i}^{C1} = q_{g_i}^{C2}$ and DE refers to $q_{g_i}^{C1} \neq q_{g_i}^{C2}$.

On the null hypothesis (EE), the data $X_{g_i}^{C1,C2} = X_{g_i}^{C1}, X_{g_i}^{C2}$ arise from the prior predictive distribution $f_0^{I_g}(X_{g_i}^{C1,C2})$:

$$f_0^{I_g}(X_{g_i}^{C1,C2}) = \left[\prod_{s=1}^{S} \binom{X_{g_i,s} + r_{g_i,s} - 1}{X_{g_i,s}} \right] \frac{Beta(\alpha + \sum_{s=1}^{S} r_{g_i,s}, \beta^{I_g} + \sum_{s=1}^{S} X_{g_i,s})}{Beta(\alpha, \beta^{I_g})} \tag{2}$$

Under the alternative (DE), $X_{g_i}^{C1,C2}$ follows the prior predictive distribution $f_1^{I_g}(X_{g_i}^{C1,C2})$:

$$f_1^{I_g}(X_{g_i}^{C1,C2}) = f_0^{I_g}(X_{g_i}^{C1}) f_0^{I_g}(X_{g_i}^{C2}) \tag{3}$$

Denoting the latent variable Z_{g_i} where $Z_{g_i} = 1$ indicates that isoform g_i is DE and $Z_{g_i} = 0$ indicates isoform g_i is EE; $Z_{g_i} \sim Bernoulli(p)$. Thus, the marginal distribution of $X_{g_i}^{C1,C2}$ and Z_{g_i} is:

$$(1 - p) f_0^{I_g}(X_{g_i}^{C1,C2}) + p f_1^{I_g}(X_{g_i}^{C1,C2}) \tag{4}$$

The posterior probability of being DE at isoform g_i is obtained by Bayes' rule:

$$\frac{pf_1^{I_g}(X_{g_i}^{C1,C2})}{(1-p)f_0^{I_g}(X_{g_i}^{C1,C2})+pf_1^{I_g}(X_{g_i}^{C1,C2})} \tag{5}$$

6.1 Parameter estimation

With the assumption that $r_{g_i,s} = r_{g_i,0}l_s$, denote $\mu_{g_i,0}^C$ and $(\sigma_{g_i,0}^C)^2$ as the mean and variance of gene g isoform i under standard library size. Then $\mu_{g_i,0}^C = \frac{1}{l_s}\mu_{g_i,s}^C$ for any s within condition C, Assume there are S_C samples in condition C. We could obtain the unbiased estimator $\hat{\mu}_{g_i,0}^C = \frac{1}{S_C}\sum_{s\ in\ C}\frac{1}{l_s}\hat{\mu}_{g_i,s}^C$ where $\hat{\mu}_{g_i,s}^C = X_{g_i,s}^C$.

Since $(\sigma_{g_i,0}^C)^2 = \frac{1}{l_s}(\sigma_{g_i,s}^C)^2$ for any s within condition C, we could obtain the estimator $(\hat{\sigma}_{g_i,0}^C)^2 = \frac{1}{S_C}\sum_{s\ in\ C}\frac{1}{l_s}(\hat{\sigma}_{g_i,s}^C)^2$, which is unbiased conditioning on $\mu_{g_i,0} = \hat{\mu}_{g_i,0}$ where $(\hat{\sigma}_{g_i,s}^C)^2 = (X_{g_i,s}^C - l_s\hat{\mu}_{g_i,0}^C)^2$.

Denote $\hat{\mu}_{g_i,0} = \frac{\hat{\mu}_{g_i,0}^{C1}+\hat{\mu}_{g_i,0}^{C2}}{2}$ and $\hat{\sigma}_{g_i,0}^2 = \frac{(\hat{\sigma}_{g_i,0}^{C1})^2+(\hat{\sigma}_{g_i,0}^{C2})^2}{2}$ Then the estimator of $r_{g_i,0}$ is obtained by $\hat{r}_{g_i,0} = \frac{\hat{\mu}_{g_i,0}^2}{\hat{\sigma}_{g_i,0}^2-\hat{\mu}_{g_i,0}}$.

$\hat{l_s}$ could be obtained by the total number of reads, TMM [15], Median Normalization [1], or Quantile Normalization [2]. Since the total number of reads may be adversely affected by outliers from PCR or other artifacts, the latter 3 methods are more acceptable. We used Median Normalization.

The EM algorithm is used to estimate α, β^{I_g} and p via the **optim** function in **R**.

6.2 Multiple Condition Case

EBSeq naturally accommodates multiple condition comparisons. For example, in a study with 3 conditions, there are 5 possible patterns in which latent levels of expression may vary across conditions: $q_{g_i}^{C1} = q_{g_i}^{C2} = q_{g_i}^{C3}$; $q_{g_i}^{C1} = q_{g_i}^{C2} \neq q_{g_i}^{C3}$; $q_{g_i}^{C1} = q_{g_i}^{C3} \neq q_{g_i}^{C2}$; $q_{g_i}^{C1} \neq q_{g_i}^{C2} = q_{g_i}^{C3}$; and $q_{g_i}^{C1} \neq q_{g_i}^{C2} \neq q_{g_i}^{C3}$.

The prior predictive distributions for these are given, respectively, by:

$g_1^{I_g}(X_{g_i}^{C1,C2,C3}) = f_0^{I_g}(X_{g_i}^{C1,C2,C3})$; $g_2^{I_g}(X_{g_i}^{C1,C2,C3}) = f_0^{I_g}(X_{g_i}^{C1,C2})f_0^{I_g}(X_{g_i}^{C3})$; $g_3^{I_g}(X_{g_i}^{C1,C2,C3}) = f_0^{I_g}(X_{g_i}^{C1,C3})f_0^{I_g}(X_{g_i}^{C2})$; $g_4^{I_g}(X_{g_i}^{C1,C2,C3}) = f_0^{I_g}(X_{g_i}^{C1})f_0^{I_g}(X_{g_i}^{C2,C3})$; and $g_5^{I_g}(X_{g_i}^{C1,C2,C3}) = f_0^{I_g}(X_{g_i}^{C1})f_0^{I_g}(X_{g_i}^{C2})f_0^{I_g}(X_{g_i}^{C3})$

in which $f_0^{I_g}$ is the same as in equation 2. Then the marginal distribution in equation 4 becomes:

$$\sum_{j=1}^{5} p_j g_j^{I_g}(X_{g_i}^{C1,C2,C3}) \tag{6}$$

in which $\sum_{j=1}^{5} p_j = 1$.

Thus, the posterior probability that isoform g_i is in pattern P_J is readily obtained by:

$$\frac{p_J g_J^{I_g}(X_{g_i}^{C1,C2,C3})}{\sum_{j=1}^{5} p_j g_j^{I_g}(X_{g_j}^{C1,C2,C3})} \tag{7}$$

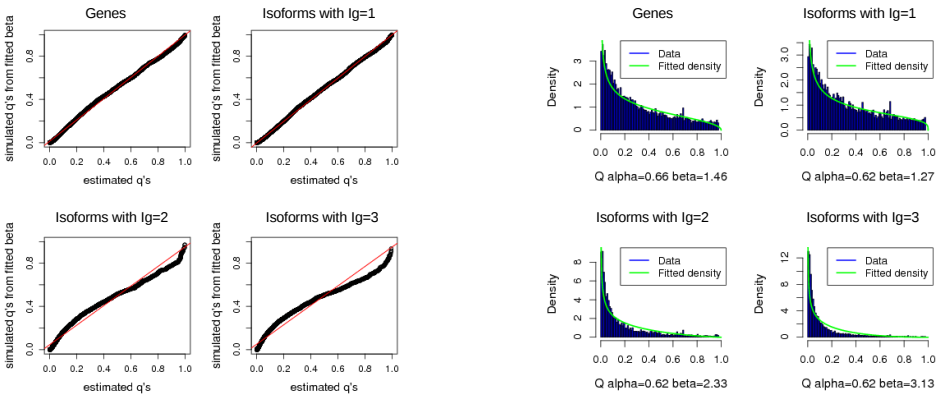
Parameter estimation closely follows that given in the previous section. For the 3 condition case presented here, $\hat{\mu}_{g_i,0} = \frac{\hat{\mu}_{g_i,0}^{C1} + \hat{\mu}_{g_i,0}^{C2} + \hat{\mu}_{g_i,0}^{C3}}{3}$ and $\hat{\sigma}_{g_i,0}^2 = \frac{(\hat{\sigma}_{g_i,0}^{C1})^2 + (\hat{\sigma}_{g_i,0}^{C2})^2 + (\hat{\sigma}_{g_i,0}^{C3})^2}{3}$.

6.3 EBSeq may also be used to identify equivalently expressed isoforms and genes

Unlike most approaches which classify non-DE genes as EE, as we detail above, EBSeq is based on a mixture model which facilitates evaluation of the posterior probabilities associated with DE, as well as EE. With these posterior probabilities, a user may identify an FDR controlled list of EE genes. This may be of particular interest for genes with more than one isoform, since compensatory mechanisms may give rise to DE isoforms in EE genes; and consequently subtle, yet important, differences may be missed if focus is placed exclusively on DE genes alone. Using the EE posterior probabilities from the case study, EBSeq identified 64 EE genes with DE isoforms contributing at least 30% of the gene expression (20 are shown in Supplementary Figure 13).

7 Model Diagnostics

Supplementary Figure 14(a) shows the estimated $q_{g_i}^{C1}$'s (q_g^{C1}'s) and the same number of points simulated from the prior assumed in EBSeq, namely a Beta distribution with hyperparameters estimated as described in Section 6 of this Supplement using data from the experiment comparing



(a) (b)

Supplementary Figure 14: Panel (a) shows a QQ-plot comparing the estimated $q_{g_i}^{C1}$'s (q_g^{C1}'s) and the same number of points simulated from a Beta distribution with parameters estimated via EBSeq. The data is from the experiment comparing ESCs with iPSCs. Panel (b) shows a histogram of the estimated $q_{g_i}^{C1}$'s (q_g^{C1}'s) and the corresponding Beta densities.

ESCs with iPSCs. Supplementary Figure 14(b) shows the histogram of estimated $q_{g_i}^{C1}$'s (q_g^{C1}'s) and the fitted Beta density using that same data. These figures indicate that the prior assumed by EBSeq is reasonable for the experiments considered here.

References

[1] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.

[2] Bullard, J. H., Purdom, E. A., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, **11**, 94.

[3] Consortium, M. (2006). The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, **24**, 1151–1161.

[4] Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics*, **21**, 488.

[5] Haag, J. D., Shepel, L. A., Kolman, B. D., Monson, D. M., Benton, M. E., Watts, K. T., Waller, J. L., Lopez-Guajardo, C. C., Samuelson, D. J., and Gould, M. N. (2003). Congenic rats reveal three independent copenhagen alleles within the mcs1 quantitative trait locus that confer resistance to mammary cancer. *Cancer Res*, **63**, 5808.

[6] Hardcastle, T. J. and Kelly, K. A. (2010). bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

[7] Jiang, H. and Wing, W. H. (2008). Seqmap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, **24(20)**, 2395–2396.

[8] Jiang, H. and Wing, W. H. (2009). Statistical inferences for isoform expression in rna-seq. *Bioinformatics*, **25(8)**, 1026–1032.

[9] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2010). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, page R25.

[10] Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

[11] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010). Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26(4)**, 493–500.

[12] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, **5(1)**, 621–628.

[13] Pickrell, J., Marioni, J., Pai, A., Degner, J., Engelhardt, B., Nkadori, E., Veyrieras, J., Stephens, M., Gilad, Y., and Pritchard, J. (2010). Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, **464**(7289), 768–772.

[14] R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

[15] Robinson, M. D. and A, O. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, **11**, R25.

[16] Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23(21)**, 2881–2887.

[17] Sengupta, S., Ruotti, V., Bolin, J., Elwell, A., Hernandez, A., Thomson, J., and Stewart, R. (2010). Highly consistent, fully representative mrna-seq libraries from ten nanograms of total rna. *Biotechniques*, **49**, 898–904.

[18] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28(5)**, 211–215.