UNIVERSITY OF WISCONSIN

DEPARTMENT OF BIOSTATISTICS AND MEDICAL INFORMATICS

Technical Report **March 2012**
 # 226

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

Ning Leng, John Dawson , James Thomson, Victor Ruotti, Anna Rissman, Bart Smits, Jill Haag, Michael Gould, Ron Stewart and Christina Kendziorski

UNIVERSITY OF WISCONSIN
DEPARTMENT OF BIOSTATISTICS
AND MEDICAL INFORMATICS
K6/446 Clinical Science Center
600 Highland Avenue
Madison, Wisconsin 53792-4675
(608) 263-1706

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

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

[1] *Department of Statistics, University of Wisconsin, Madison, WI*

[2] *Morgridge Institute for Research, Madison, WI*

[3] *McArdle Laboratory for Cancer Research, Department of Oncology, University of Wisconsin, Madison, WI*

[4] *Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI*

Messenger RNA expression is important in normal development and differentiation, as well as in manifestation of disease. High-throughput cDNA sequencing (RNA-seq) experiments allow for the identification of differentially expressed (DE) genes and their corresponding isoforms on a genome-wide scale. However, statistical methods are required to ensure that accurate identifications are made. A number of methods have been developed for identifying DE genes in an RNA-seq experiment, but they are deficient for identifying DE isoforms. Because uncertainty in estimated isoform expression varies directly with isoform complexity, applications of gene-centric approaches to isoform inference results in reduced power for some classes of isoforms and increased false discoveries for others. In addition, the most popular gene-centric

DE methods are not robust to outliers, which further increases the potential for false discoveries in both gene and isoform-level inference. We have developed an empirical Bayesian modeling approach for identifying differential expression in an RNA-seq experiment (EBSeq) comparing two or more biological conditions. Evaluation via simulation and case studies demonstrates that EBSeq is a powerful and robust approach that outperforms existing methods. Application of EBSeq to a study of human embryonic and induced pluripotent stem cells provides novel insights into genomic differences underlying these cell types and illustrates the importance of appropriate statistical analyses.

1 Introduction

Appropriate expression of a gene's isoforms via alternative splicing is fundamental to normal development and maintenance in eukaryotes; and aberrations in alternative splicing are common in disease[1,2,3]. Consequently, there is much interest in identifying isoforms with expression that varies across biological conditions. High-throughput cDNA sequencing (RNA-seq) experiments provide the potential to identify such differentially expressed (DE) isoforms on a genome-wide scale, but statistical methods are required to ensure that accurate identifications are made.

The statistical methods available for isoform inference in an RNA-seq experiment are focused largely on changes in the proportion of gene-specific reads assigned to an isoform[4,5], so-called differential transcription or differential splicing. These methods do not consider changes in overall expression levels and are therefore not appropriate for identify-

ing DE isoforms. When a precise quantification of isoform-level DE is required, methods developed for identifying DE genes are often applied directly to the isoform expression estimates[6,7]. Doing so unduly inflates both false negative and false positive error rates.

Because uncertainty in estimated isoform expression varies directly with isoform complexity[8], applications of gene-centric approaches to isoform level inference results in critically reduced power for some classes of isoforms and increased false discoveries for others. This is due to the general structure of test statistics used by most methods for DE gene identification, which calibrate a difference in expression levels between conditions by a variance. For valid inference, it is essential that this variance be well estimated. A common approach for variance estimation is to use a spline obtained empirically from the mean-variance relationship observed in data. Figure 1 shows that this relationship varies dramatically for different groups of isoforms, where groups are defined by isoform complexity as quantified by the number of constituent isoforms of the parent gene.

Specifically, an isoform of gene g is assigned to the $N_g = m$ group, for example, where $m = 1, 2$ or 3, if the total number of isoforms from gene g is m (the $N_g = 3$ group contains all isoforms from genes having 3 or more isoforms). As shown in Figure 1, there is decreased variability in the $N_g = 1$ group, but increased variability in the others, due to the relative increase in uncertainty inherent in estimating isoform expression when multiple isoforms of a given gene are present. This observation is not specific to the data set and/or the method used for isoform expression estimation (see Supplementary Figures 1-5 for additional examples).

If isoforms are analyzed collectively, there is reduced power for identifying isoforms in the $N_g = 1$ group (since the true variances in that group are lower, on average, than that derived from the full collection of isoforms) and increased false discoveries in the $N_g = 2$ and $N_g = 3$ groups (since the true variances are higher, on average, than those derived from the full collection). In addition to this limitation, we find that the most popular statistical methods for identifying differential expression in an RNA-seq experiment are not robust to outliers, which further increases the number of false discoveries made not only when identifying DE isoforms, but also when identifying DE genes.

Taking advantage of the merits of empirical Bayesian approaches, we developed a specific formalism suited to RNA-seq inference called EBSeq. We use EBSeq to enable improved identification of the expression differences between human embryonic stem (ES) and induced pluripotent stem (iPS) cells. Results from this case study as well as a series of simulations show that EBSeq is a powerful and robust approach for identifying differential expression in an RNA-seq experiment that outperforms the most commonly used approaches. The case study further demonstrates the importance of using appropriate statistical tests when characterizing differences between ES and iPS cell types.

2 Results

Simulation Studies Simulation studies to investigate the operating characteristics of both isoform and gene-level inference revealed that DESeq[9] and edgeR[7] have an inflated false discovery rate (FDR) and are sensitive to outliers. Specifically, Table 1 shows the power and FDR averaged across 100 simulations where target FDR was set at 5%. In the absence

of outliers, DESeq and edgeR had the highest power for identifying DE genes. However, the FDR was increased almost three-fold. EBSeq was robust, showing the highest power among the methods with well controlled FDR.

When outliers are present, the FDR for baySeq, DESeq and edgeR was increased almost ten-fold. Figure 2 shows that this poor performance is not a function of the FDR target rate. There we show the number of equivalently expressed (EE) genes containing outliers identified as DE for lists ranging in size from 1 to 2,000. At 5% FDR, baySeq[10], DESeq, and edgeR identified an average of 1,987, 2,561, and 2,624 genes as DE, respectively. Of these, over 30% were false discoveries due to outliers in otherwise EE genes. At 10% FDR each method identified 2,152, 2,691, and 2,777, respectively, of which over 35% were due to outliers. At a 5% (10%) FDR level, EBSeq identified 1,434 (1,456) genes, none of which contained outliers.

Each of these methods was also applied to simulated isoforms. It is important to note that with the exception of EBSeq, the methods considered here were not developed specifically to identify DE isoforms, and consequently do not accommodate estimation uncertainty at the isoform level. As a result, some decrease in performance was expected. In particular, DESeq and edgeR assume that a gene's variance may be represented as the sum of its mean and extra biological variance which is estimated in part by pooling genes with common means. When isoforms are being considered, isoform-specific variability also depends on exon assignment uncertainty induced in isoform expression estimation. EBSeq quantifies this uncertainty via N_g; and so a naive remedy when considering other approaches developed for gene inference would be to apply each approach within each

N_g group. Table 1 shows results derived from applying these approaches to all isoforms at once as well as within N_g group. As shown, the results are similar to those from the gene-level simulation study, both with and without outliers: power is highest for DESeq and edgeR, but FDR is increased. Evaluating the operating characteristics within N_g group shows that the increase in FDR is due to false calls in the $N_g = 2$ and $N_g = 3$ groups, as expected given the differential variability observed among these groups (see Figure 1). Applying each of the approaches within N_g group did slightly decrease the FDR within group, but overall the FDR remains high. As in the gene-level simulation study, EBSeq showed the highest power among methods for a given level of well-controlled FDR.

Case study Induced pluripotent stem cells hold great promise for therapeutic and research purposes as they are nearly indistinguishable from ES cells with respect to morphology, capacity to self-renew, and developmental potential. However, the extent to which the promise may be realized depends critically on a precise characterization of the molecular differences between the cell types, a question which remains controversial. An early study using Student's t-test identified numerous genes DE between ES and iPS cells[11], but subsequent studies noted that the vast majority of identifications were false ones induced by not properly accounting for multiplicities in the statistical analysis[12]. The most recent work, which also used the t-test, found far fewer differences between ES and iPS cells when only one replicate per cell line was used and pointed out the importance of utilizing replicates to increase statistical power[13]. The small number of genes identified with just one replicate per cell line is likely the result of the fact that the t-test is underpowered for high throughput studies[14]. It is also likely due to an exclusive focus on DE at the gene,

not isoform, level. As compensatory mechanisms may give rise to DE isoforms in EE genes, subtle, yet important, differences were missed. Both considerations are addressed by applying EBSeq.

Unlike the t-test, EBSeq directly models read counts and allows for information sharing across genes and isoforms, leading to increased power for identifying DE. Analyzing the data from Phanstiel *et al.*[13], EBSeq identified 90 genes to be consistently DE across three experiments comparing ES and iPS cells, whereas the Benjamini-Hochberg adjusted t-test found just one[13]. Many of genes identified by EBSeq (detailed in Supplementary Table 1) are involved in developmental processes including mesodermal and neural differentiation. Amongst the most highly iPS-enriched genes, several are known to be involved in mesoendodermal differentiation including EOMES, SOX17, GATA6, TBX15, and SP5. This may be a result of epigenetic memory, as the parental cells are mesodermal in origin, and may be one explanation for the reduced neural differentiation capability of these iPS cells (when compared to differentiation of ES cells). It may also be possible that this epigenetic memory gives the iPS cells a proclivity to establish some transcriptional programs of the early mesendodermal lineage. In addition to these genes, a number of the ES- and iPS-enriched genes identified by EBSeq have demonstrated differences in other studies that considered expression from microarrays in these same strains of ES and iPS cells[13] as well as others[15,16]. Of the 90 genes identified by EBSeq, 27 were not identified by DEseq and edgeR (Figure 4). The gene DNMT1 was included in the list of 27 showing significantly higher expression in iPS cells. This is particularly interesting given that inappropriate DNMT1 levels are known to cause differentiation defects in other model

systems[17,18,19].

Supplementary Table 2 shows the total number of genes and isoforms identified as consistently DE across the three experiments by all approaches. As shown, CuffDiff identified only 6 genes whereas DESeq and edgeR identified over 250 genes as consistently DE. A consideration of the overlap among gene lists from EBSeq, DESeq and edgeR provides insight into these approaches. As shown in Figure 4(b), DESeq and edgeR identified 144 genes not found by EBSeq. Supplementary Figure 9 shows a heatmap of the 144 along with the 27 genes identified exclusively by EBSeq, as well as the 63 identified by all three approaches. The figure suggests that, unlike EBSeq, many of the genes identified exclusively by DESeq and edgeR contain outliers.

To more precisely identify genes containing outliers, for each gene we evaluated Dixon's Q-statistic[20] as well as the gene's fold change with and without its most extreme value (FCRatio). A gene harboring an outlier will have a large Dixon's Q-statistic as well as FCRatio far from 1 (see Methods). Figure 3(a) shows the FCRatio and Dixon's Q-statistics for the genes identified by each method at 5% FDR for one of the three experiments comparing ESCs with iPSCs (results from the other two experiments were similar). As shown, DESeq and edgeR favored the genes with extreme FCRatio and large Dixon's Q-statistic. Figure 3(b) shows that this holds for varying levels of Dixon's Q-statistic as well as varying FDR thresholds. Specifically, Figure 3(b) considers the lists of genes identified at 5% FDR for each method and reports (solid lines) the proportion of genes identified as DE that contain outliers, where outlier is defined for varying Dixon's Q-statistics. As ordered lists of genes may be preferred to FDR controlled lists, we also

evaluated performance for the top 500 genes rank ordered by each method. Shown in Figure 3(b) (dashed lines) are the proportion of outliers in these top 500 genes. Note that DESeq and edgeR identified 600 and 599 DE genes, respectively, and so these top 500 were on their 5% FDR controlled lists and consequently there is little difference in the proportions identified on the FDR controlled vs. the rank ordered lists. On the other hand, CuffDiff identified only 23 genes at 5% FDR, and so larger differences are observed.

As shown in Figure 3(b), approximately 10-20% of the genes identified as DE by DESeq or edgeR contain outliers. For a Dixon's Q-statistic of 0.76, for example, approximately 20% of the genes identified by DESeq or edgeR contain outliers, whereas only 2.6% of genes in the full dataset do. It is important to note that not all of these identifications were necessarily false ones, as truly DE genes may contain outliers (see Supplementary Figure 8). However, Figure 4 and Supplementary Figure 9 show that many of the calls were due to the presence of a single outlier in an otherwise EE gene.

To further investigate the relationship between outliers and DE identifications, we used the multiple condition EBSeq model on the combined data set with 24 samples to evaluate the posterior probability of each of 18 patterns of expression. The patterns included EE and DE, as in the two condition comparison, as well as EE with an outlier cell line (EEO; 8 possibilities since the outlier could be in any one of the 8 cell lines) and DE with an outlier cell line (DEO; another 8 possibilities); see Figure 4. Delineating these expression patterns allows one to disentangle identifications of DE genes due to outliers in otherwise EE genes (EEO pattern) vs. truly DE genes that harbor an outlier (DEO pattern). The model identified 17,390 genes as being most likely EE or EEO (posterior probability

of EE or EEO greater than 0.5). None of these were in the list of 90 genes identified as consistently DE by EBSeq (two condition analysis), whereas 26 were in the list of 144 genes identified as consistently DE by DESeq and edgeR but not EBSeq (see Figure 4 and Supplementary Figure 9); 21 of the 26 showed posterior probability of EE or EEO greater than 0.95. The 9 of the 26 showing strongest evidence of being EE or EEO are shown in Figure 4(e), from which it is clear that many of the false identifications are in genes that are very lowly expressed (near zero) in all but one sample. For comparison, although none were significant, in Figure 4(e), we also show the 9 genes identified exclusively by EBSeq that have the highest evidence of being EE or EEO; the average intensities within condition are higher and it is clear that identifications are not due to a single outlier in an otherwise EE gene.

A comparison of the genes identified as DE or DEO provides further insight into the differences among these approaches. In particular, the left panel of Figure 4(d) shows the median expression for the 118 DE or DEO genes identified exclusively by DESeq and edgeR but not EBSeq. The median intensity of this collection is comparable to that of the 63 genes found by all approaches, but the boxplots show that the 118 contain many genes with very low counts in both conditions. The right panel shows specifically, for example, that approximately 20% of the 118 genes are ones where 75% of the samples have counts less than 20. On the other hand, identifications exclusive to EBSeq tend to be those with higher expression.

Supplementary Table 2 also shows the number of DE isoforms identified by each approach. While the total number identified by DESeq, edgeR and EBSeq is comparable at

157, 161, and 135, respectively, the table shows that DESeq and edgeR found a disproportionately low number of calls in the $N_g = 1$ group (suggesting low power for this group as demonstrated in the simulation studies) and a disproportionately high number in the $N_g = 2$ and $N_g = 3$ groups (suggesting increased FDR). In particular, DESeq calls 0.5% of all isoforms DE, but 0.2% of all $N_g = 1$ isoforms DE (and 0.8% of all $N_g = 3$ isoforms DE); similarly, edgeR calls 0.5% of all isoforms DE with 0.1% in the $N_g = 1$ group (and 0.9% in the $N_g = 3$ group). The proportion identified across N_g groups was stable for EBSeq. In addition, EBSeq identified a number of DE isoforms in non-DE genes, including CFLAR, CYB5A, DLL3, LRRTM4 and TAP2, which are thought to be important in mesodermal differentiation (see Supplementary Figure 10). An analysis of data from the mammary cancer susceptibility experiment shows that these results are not specific to an individual dataset (see Figures 1 and 3, Supplementary Table 3, Supplementary Figures 11 and 12).

3 Discussion

The identification of isoforms DE across two or more biological conditions is a common and important problem in RNA-seq experiments that is often addressed by applying statistical methods for identifying DE genes directly to isoforms. As gene-based methods do not account for differential uncertainty inherent in isoform expression estimation, their application leads to underpowered inference for some classes of isoforms and inflated false discoveries for others. To address this, we have developed an empirical Bayesian hierarchical modeling approach, EBSeq, that enables accurate, efficient, and robust inference in

RNA-seq experiments.

The main difference between EBSeq and the other approaches considered here is that EBSeq models isoform expression directly, as opposed to gene expression, and in so doing accommodates isoform expression estimation uncertainty. In particular, estimation uncertainty is partitioned into three groups defined by isoform complexity ($N_g = 1, 2,$ or 3), following our empirical observation that uncertainty is increased on average in isoforms that share a parent gene. EBseq is not restricted to three groups and for some genomes, additional N_g groups may be warranted. An alternative approach would be to model a more direct measure of uncertainty, such as the number of discriminatory reads between isoforms, or credibility interval length provided by an isoform expression estimation method such as RSEM[21] or Cufflinks[22]. Although we continue to explore the potential advantages of doing so, we note that credibility intervals are estimated slightly differently among these methods, and consequently incorporating credibility interval length into EBSeq requires tailoring the model to each method. Modeling N_g, on the other hand, allows for direct application of EBSeq to expression estimates obtained from a variety of methods. At the same time, applicability is limited to genomes for which N_g is known, or at least well approximated. When dealing with de novo transcripts or reconstructed transcriptomes, grouping information could be defined by the percentage of overlapping exons for each transcript as provided by de novo assembly algorithms[23]. Evaluation of this type of approach is underway.

Another difference is that EBseq is based on a mixture model which facilitates evaluation of the posterior probabilities associated with various expression patterns. A com-

parison of two biological conditions provides for the evaluation of two patterns, EE and DE. Unlike other approaches that classify non-DE genes as EE, EBSeq allows for a precise quantification of the evidence in favor of both DE and EE and in so doing enables a more precise identification of EE genes harboring DE isoforms. The mixture model framework also enables comparisons of more than two patterns which is required when more than two biological conditions are being compared, or when more specific questions related to differences among subgroups of samples are being addressed, such as those considered here related to outliers (see Figure 4).

Simulations demonstrate the advantages of EBSeq for both isoform and gene-level inference. When identification of DE isoforms is of interest, methods previously developed to identify DE genes should not be used as they are underpowered for $N_g = 1$ isoforms and they yield false identifications for $N_g = 2$ or 3 isoforms. On the other hand, EBSeq has well controlled FDR and appreciable power for isoform inference. Although developed to facilitate isoform inference, the model underlying EBSeq is also useful for identifying DE genes, with simulations demonstrating that EBSeq outperforms the most popular DE gene identification methods when outliers are not present, and greatly outperforms them when they are.

It is important to note that the simulations assume read counts are distributed as Negative Binomial, an assumption common to most methods considered here (baySeq, DESeq, edgeR and EBSeq). The form of the variance in the simulation was not one we assumed in EBSeq. Rather, it was defined to match that of Robinson and Smyth[7]. As a result, data were not simulated directly from the model underlying EBSeq, and so some

violation of model assumptions is expected. In addition, when outliers are included, the Negative Binomial assumption is violated as well. In spite of this, EBSeq demonstrated substantial power with well controlled FDR for most simulation scenarios. At the same time, major departures from model assumptions could lead to poor performance, and diagnostics should always be checked to assess model fit (see the Supplement).

The case study results were consistent with what we observed in simulations; and they also demonstrate the importance of using appropriate statistical methods. In particular, prior comparisons of ESCs and iPSCs used a t-test to identify DE genes and found very few, largely due to insensitivity of the test. Model-based methods such as those evaluated here provide increased power, but as demonstrated, they can be prone to false discoveries. In practice, an accurate characterization of the differences underlying ESCs and iPSCs requires a more detailed consideration of isoform expression as well as further refinement of statistical models. The analysis of differences between ES and iPS cells by EBSeq uncovers a core set of genes that differ between these cell types and identifies several iPS-enriched genes that are key regulators of mesendodermal differentiation that may represent epigenetic memory in the iPS cells or a proclivity for the iPS cells to take on some transcriptional characteristics of the early mesendodermal lineage. Given that we found many DE isoforms in EE genes (see Supplementary Figures 10 and 12), we expect that the expression based differences previously reported are likely underestimates as those prior studies focused exclusively on gene, not isoform, expression[13]. Furthermore, the analysis considered three replicate experiments comparing ESCs and iPSCs. Since the statistical methods currently available do not accommodate multiple experiments, each

was analyzed separately, and we reported genes and isoforms that were consistently DE (DE in each experiment). We are extending EBSeq to accommodate not just biological, but also technical, replicates within a study which will enable a unified analysis of these types of experiments. Additional extensions are also underway to accommodate correlation among isoforms sharing exons, which should further improve the power and accuracy for identifying DE genes and isoforms in an RNA-seq experiment.

Acknowledgments The authors would like to thank Bo Li, Michael Newton, Justin Brumbaugh, and Colin Dewey for comments that helped improve the manuscript. This work was supported in part by NIH CA28954 and NIEHS ES17400.

4 Methods

Statistical Model We model read counts from isoform i in gene g and sample s as Negative Binomial, $X_{g_i,s}$, where $g = 1, 2, ..., G$, $s = 1, 2, ..., S$, and $i = 1, 2, ..., N_g$. Specifically, we assume that within condition C, $X^C_{g_i,s}|r_{g_i,0}, l_s, q^C_{g_i} \sim NB(r_{g_i,0}l_s, q^C_{g_i})$ where l_s represents the library size in sample s. A prior distribution describes fluctuations in technical and biological variation: $q^C_{g_i}|\alpha, \beta^{N_g} \sim Beta(\alpha, \beta^{N_g})$. The hyperparameter α is shared across isoforms while β depends on N_g, accommodating the systematic differences in variability among the N_g groups.

Within this framework, the latent mean level of expression, $\mu^C_{g_i}$, is given by $r_{g_i,0}(1 - q^C_{g_i})/q^C_{g_i}$. When RNA-seq reads in two biological conditions are available, identifying DE isoforms corresponds to identifying those isoforms for which $\mu^{C1}_{g_i} \neq \mu^{C2}_{g_i}$. Letting

p denote the prior probability of DE, counts are modeled by the mixture distribution $(1-p)f_0^{N_g}(X_{g_i}^{C1,C2}) + pf_1^{N_g}(X_{g_i}^{C1,C2})$ where $X_{g_i}^{C1,C2}$ represents g_i's read counts across the two conditions; f_0 and f_1 are the predictive distributions under EE and DE, respectively, defined in the Supplement. Once parameters are estimated via method-of-moments and the expectation-maximization (EM) algorithm[24], the posterior probability of DE is readily obtained via Bayes' Rule. A mixture model with additional components may be used when data from more than two conditions are available (see the Supplement).

Simulated Data We followed the simulation set-up of Robinson and Smyth[7] 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 27,468 isoforms, five lanes in each of two conditions. Sample sizes were taken to match those observed in the mammary carcinoma susceptibility case study (as that study had the most samples); parameter values were sampled from empirical ones in that study (see the Supplement). Ten percent of the isoforms were simulated as DE. For half of the DE isoforms, $\mu_{g_i}^{C1} = \delta_{g_i}\mu_{g_i}^{C2}$ with δ_{g_i} sampled from the 97%-98% 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 10% of the genes set to be DE. The library size factors for both the isoform and gene-level simulations were randomly simulated from Uniform $(0.8, 1.3)$. Simulations similar to those described above were also conducted to evaluate performance in the presence of outliers. For these, in addition to the 10% DE genes, another 10% were simulated to contain outliers. An outlier gene was simulated initially as EE, then a randomly selected sample from one of the groups was multiplied by 4, 6, 8, or 10 (see Supplement for further details). Two

hundred simulated datasets were generated, one hundred without outliers and one hundred with. CuffDiff (`http://cufflinks.cbcb.umd.edu/manual.html`) was not evaluated in the simulation study as it does not apply to expression estimates, but rather requires raw reads. As shown in Results, it was evaluated in the case studies.

Identification of genes and isoforms DE across two biological conditions To quantify evidence in favor of DE, EBSeq and baySeq provide posterior probabilities whereas DESeq, edgeR, and CuffDiff provide p-values which are adjusted for multiplicities using Benjamini-Hochberg (DESeq, edgeR) or by converting to q-values (CuffDiff). To construct a list of DE genes with target FDR α, we considered those genes for which the posterior probability of DE was greater than or equal to $1 - \alpha$ (baySeq and EBseq) or those genes for which adjusted p-values were less than α (DESeq, edgeR, CuffDiff). For the case study comparing ESCs with iPSCs, the experiment was repeated three times. A gene is called "consistently DE" if it is DE in each of the three experiments.

Identification of outliers For simulated data, the exact genes containing outliers are known (see Simulated Data section within Methods for the definition). To identify putative outliers in the case studies, for each gene we evaluated Dixon's Q-statistic[20] as well as the fold change ratio (FCRatio). A Dixon's Q-statistic for a collection of values is defined as the gap over range, where gap is the absolute difference between an outlier in question and the number closest to it; the range is the max minus min. For each gene in each condition, we calculated the Dixon's Q-statistics for the smallest and the largest value. The sample with the largest Dixon's Q-statistic was defined as the potential outlier for that gene; and the largest Dixon's Q-statistic was taken as the Dixon's Q-statistic for

the gene. The FCRatio is the ratio of the fold change without the outlier over the fold change with the outlier. A gene containing an outlier will have a large Dixon's Q-statistic and FCRatio far from 1. Here, a gene was deemed to contain an outlier if its FCRatio was greater than the 95^{th} quantile or less than the 5^{th} quantile and its Dixon's Q-statistic was large, where large was quantified by d_q. Results are reported for varying d_q, as shown in Figure 3.

Experimental Data ES vs. iPS: We considered the RNA-seq data from Phanstiel *et al.*[13]. Briefly, Phanstiel *et al.*[13] evaluated RNA-seq reads from embryonic stem cell lines H1, H7, H9 and H14 and induced pluripotent stem cell lines DF4.7, DF6.9, DF19.7 and DF19.11. Experiments were run in triplicate. For data processing, we filtered 42-base-pair reads to remove adapters in each lane. Reads were aligned to the human RefSeq Hg18 transcripts using Bowtie[25], allowing for up to 200 multiple matches and two mismatches with seed length 42. Expression was estimated using RSEM[21,26] for 19,784 genes and 30,563 isoforms with expression greater than 0 in at least one sample on average across the three experiments.

Mammary carcinoma susceptibility: We consider two groups of congenic rats (4 samples in each condition) harboring the susceptible or resistance allele of the mammary carcinoma susceptibility 1a locus (Mcs1a). The resistance allele derived from the resistant Copenhagen (COP) inbred rat strain when introgressed in the susceptible genetic background of the Wistar-Furth (WF) inbred rat strain reduces mammary carcinoma multiplicity by 50%[27]. The allele is non-protein coding as it is located entirely within a gene desert on rat chr 2. For these experiments, abdominal and adjacent inguinal mammary glands

were taken from 8 untreated, mammary cancer-free females per genotype. The tissue was disaggregated using physical shearing in a solution of Tri-reagent (Ambion). RNA was extracted using a total RNA extraction kit (Ambion). RNA integrity was monitored using a 2100 Bioanalyzer (Agilent). Equal RNA (approximately 5 μg) from 2 rats were pooled to obtain a single sample for one RNA-seq lane. A total of 4 samples per genotype (*Mcs1a* susceptible or resistant) were submitted to the University of Wisconsin Biotechnology Gene Expression Center for RNA-seq sample preparation and next-generation sequencing using the Illumina Genome Analyzer IIx. Reads were retrieved and post-processed to a length of 30 basepairs. Reads were aligned to the rat Ensemble RGS3.4 transcripts using Bowtie[25], allowing for up to 100 multiple matches and one mismatch with seed length 30. Expression was estimated using RSEM[21,26] for 20,267 genes and 27,468 isoforms.

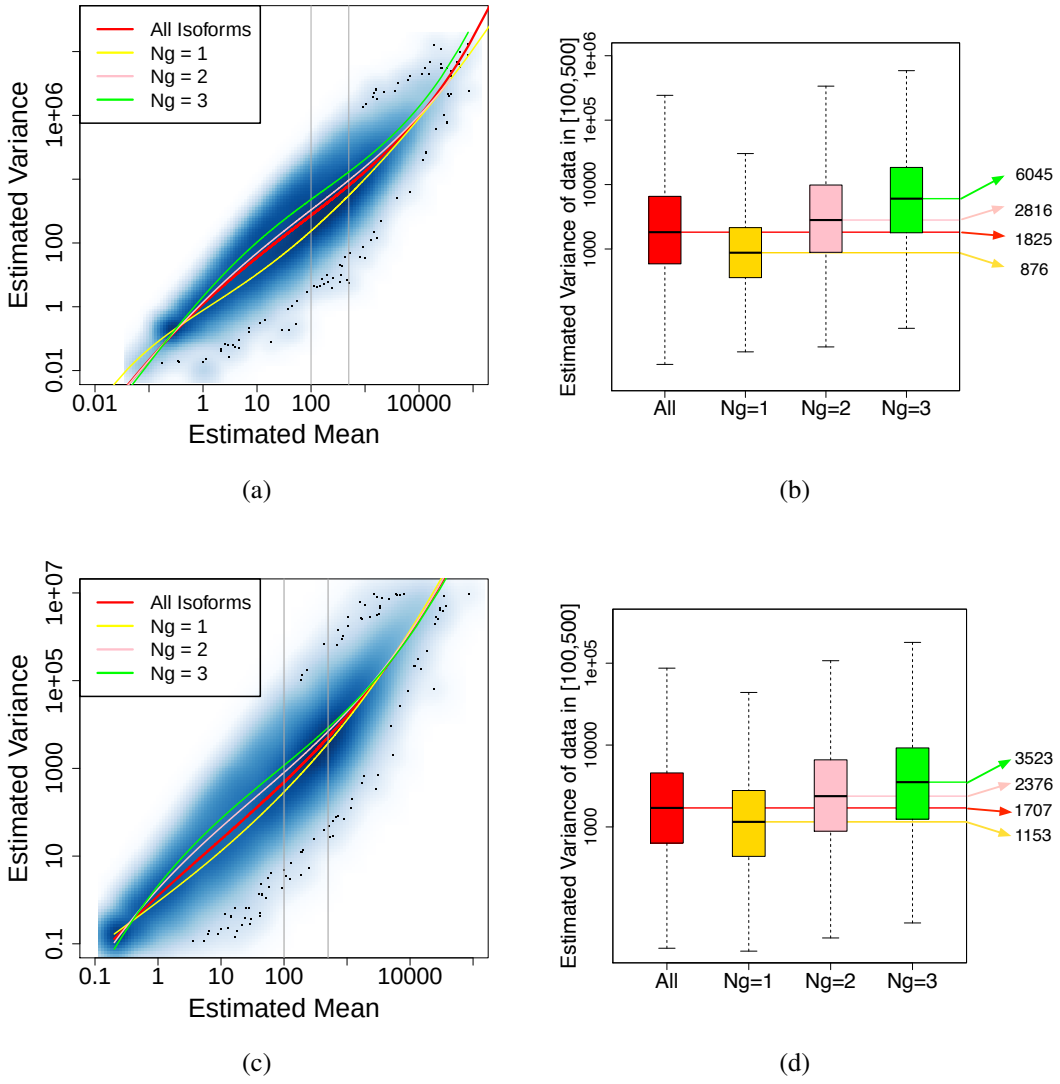


Figure 1: Panel (a) shows the empirical variance vs. mean for each isoform profiled in one of the three experiments comparing ESCs with iPSCs (results for the other two experiments were similar). A spline fit to all isoforms is shown in red with splines fit within the $N_g = 1$, $N_g = 2$, and $N_g = 3$ isoform groups shown in yellow, pink, and green, respectively. Panel (b) considers isoforms with average expression (expected count) in [100, 500], delineated by the grey lines in panel (a). The range was chosen as it approximates the 60^{th} and 80^{th} percentiles of expression across all isoforms. Shown in panel (b) are box-plots of the variances of these isoforms collectively, and within N_g group. Median variance within each group is shown right. Panels (c) and (d) are similar using data from the mammary carcinoma susceptibility experiment.

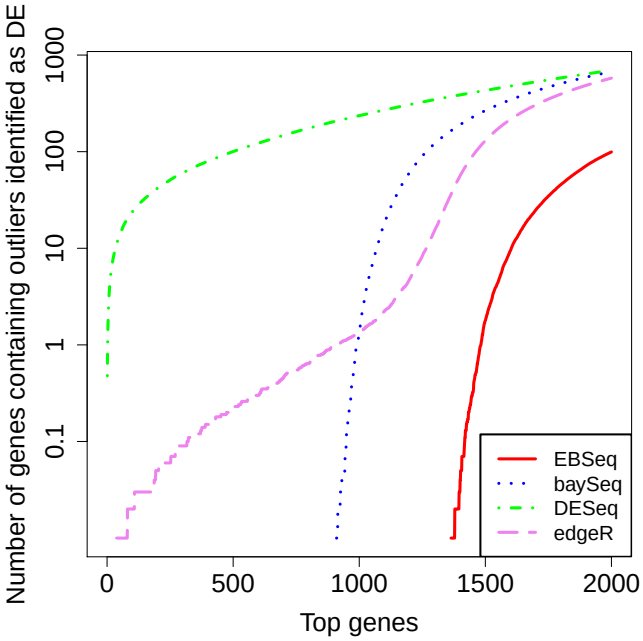


Figure 2: Shown are the average number of EE genes containing outliers identified as DE for gene lists of varying size, where the average is taken over 100 simulated datasets. At 5% FDR, baySeq, DESeq, edgeR and EBSeq yield (on average) a list size of 1,987, 2,561, 2,624 and 1,434, respectively. Of these, an average of 664, 1,085, 1,123 and 0 are identified due to outliers. At 10% FDR each method identifies 2,152, 2,691, 2,777, and 1,456 genes, respectively; of these 801, 1,182, 1,239 and 0, respectively, are due to outliers.

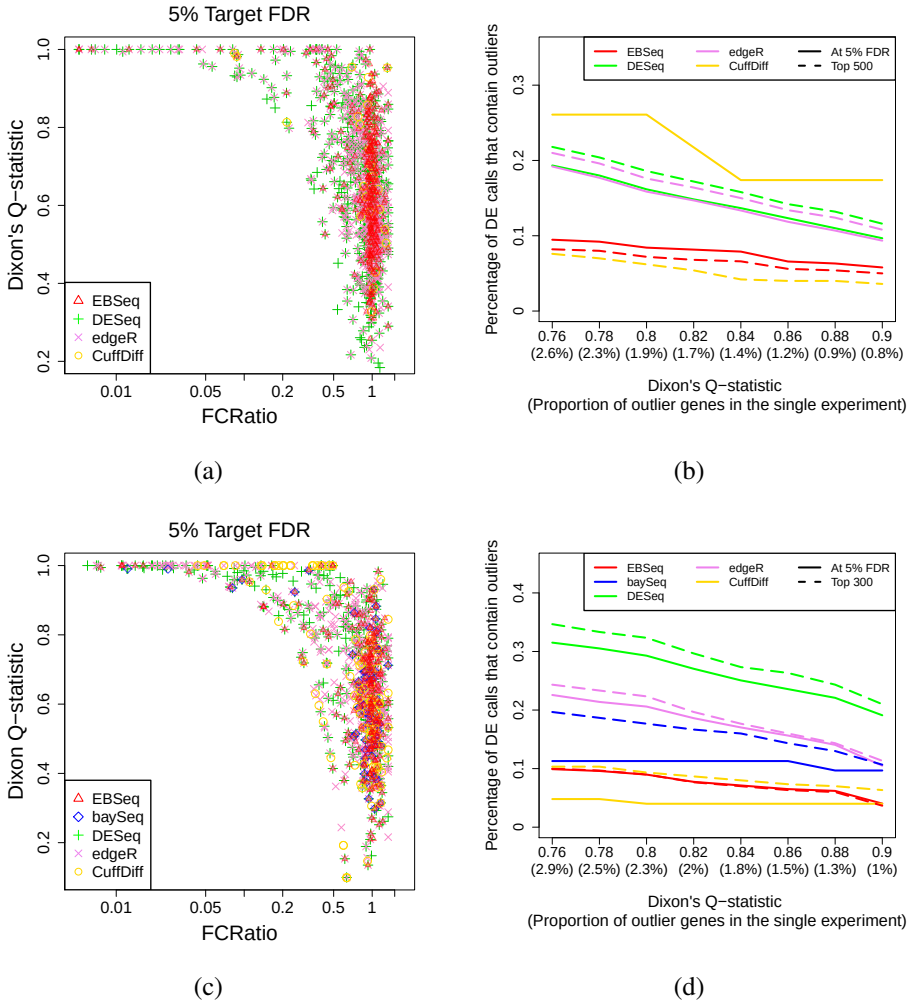


Figure 3: (a) Shown are the FCRatio and Dixon's Q-statistic of the genes identified by each method at 5% FDR in one of the three experiments comparing ESCs with iPSCs (results for the other two experiments were similar). Note that CuffDiff, DESeq, edgeR and EBseq identified 23, 600, 599, and 380 DE genes, respectively, at 5% FDR; baySeq did not converge for over 50% of the genes and consequently results are not shown. Panel (b) shows the proportion of genes identified as DE at 5% FDR (solid line) that contain outliers and also the proportion on the top 500 genes ranked by each method (dashed line) that contain outliers. Since we cannot know with certainty which genes contain outliers, a gene was deemed to contain an outlier if its FCRatio was greater than the 95^{th} quantile or less than the 5^{th} quantile and its Dixon's Q-statistic was greater than the value given on the x-axis in at least one condition. The proportion of genes satisfying these criteria is shown in parentheses for each Dixon's Q-statistic considered. Panels (c) and (d) are similar, with results shown from the mammary carcinoma susceptibility experiment. For this experiment, baySeq, CuffDiff, DESeq edgeR, and EBSeq identified 62, 117, 403, 505, and 323 DE genes, respectively, at 5% FDR. Given the lower numbers identified by DESeq, edgeR and EBSeq, results are shown for the top 300, not 500, genes.

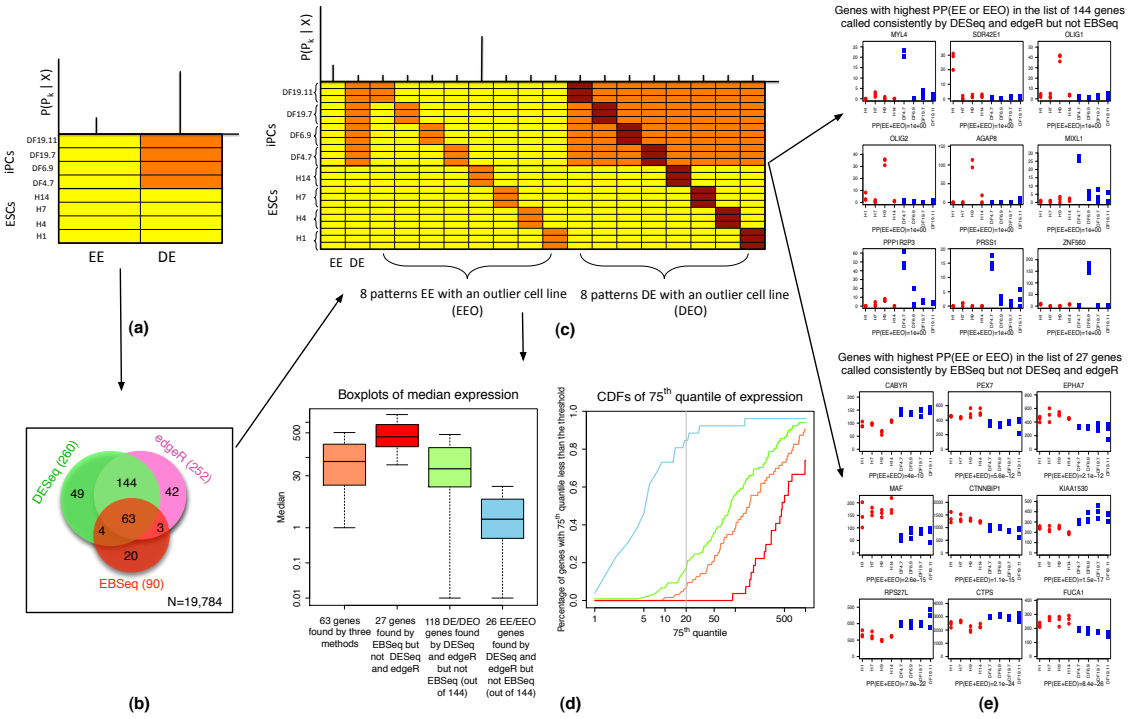


Figure 4: Panel (a) shows a schematic of the EE and DE expression patterns considered in a two condition EBSeq model for the experiment comparing ESCs with iPSCs. The upper part shows the posterior probability (PP) of each pattern for a hypothetical gene; these probabilities classify the gene into one of the patterns. Panel (b) shows a Venn diagram of the genes called consistently DE (DE in each of the three replicate experiments comparing ESCs with iPSCs) by DESeq, edgeR or EBSeq. Panel (c) shows the patterns used in the multiple condition EBSeq model on the combined data set with 24 samples. The patterns included EE and DE as well as 8 patterns for EE with an outlier cell line (EEO) and another 8 for DE with an outlier cell line (DEO). Panel (d) considers the 63 genes identified as consistently DE by all three methods, the 27 identified exclusively by EBSeq, and the 144 identified exclusively by DESeq and edgeR, but not EBSeq (a heatmap of these genes is shown in Supplementary Figure 9). Within these groups, 0,0, and 26 genes, respectively, were identified as significantly EE or EEO by the multiple condition EBSeq model. Shown in the left panel are boxplots of the median expression across the 24 samples within each group. The right panel considers each gene's 75^{th} percentile of expression across the 24 samples, and shows the percentage of genes in each of the 4 groups shown in the left panel (with the same color) having their 75^{th} percentile less than the value on the x-axis. Both panels indicate that EBSeq tends to identify more highly expressed genes. (e) The upper panel shows the 9 genes with highest posterior probability of EE or EEO out of the 144 identified as consistently DE by both DESeq and edgeR, but not EBSeq (all 9 had posterior probabilities exceeding 0.95). The x-axis indicates the 8 cell lines (4 in each of the two cell types denoted by red and blue, respectively); the y-axis shows normalized expression for each of the three replicate experiments. For comparison, the lower panel shows the 9 genes with highest posterior probability of EE or EEO out of 27 identified as consistently DE by EBSeq, but not DESeq and edgeR (note that none of these has posterior probability of EE or EEO exceeding 0.5).

Table 1: Simulation Results

			baySeq	DESeq	edgeR	baySeqEach	DESeqEach	edgeREach	EBSeq
Without outliers	All Genes	Power	0.90	0.98	0.98	-	-	-	0.94
		FDR	0.00	0.13	0.13	-	-	-	0.03
	All Isoforms	Power	0.62	0.73	0.76	0.63	0.72	0.75	0.68
		FDR	0.00	0.18	0.19	0.00	0.15	0.18	0.05
	$N_g = 1$ Isoforms	Power	0.62	0.67	0.70	0.63	0.70	0.71	0.66
		FDR	0.00	0.05	0.04	0.00	0.13	0.14	0.03
	$N_g = 2$ Isoforms	Power	0.63	0.80	0.84	0.63	0.77	0.83	0.72
		FDR	0.00	0.27	0.29	0.01	0.18	0.21	0.08
	$N_g = 3$ Isoforms	Power	0.60	0.82	0.87	0.59	0.73	0.81	0.71
		FDR	0.01	0.37	0.41	0.01	0.19	0.23	0.10
With outliers	All Genes	Power	0.89	0.97	0.98	-	-	-	0.94
		FDR	0.33	0.43	0.44	-	-	-	0.03
	All Isoforms	Power	0.62	0.72	0.76	0.62	0.72	0.74	0.69
		FDR	0.29	0.44	0.44	0.30	0.45	0.44	0.05
	$N_g = 1$ Isoforms	Power	0.62	0.67	0.70	0.63	0.70	0.70	0.66
		FDR	0.34	0.44	0.42	0.37	0.47	0.49	0.03
	$N_g = 2$ Isoforms	Power	0.63	0.79	0.83	0.62	0.76	0.81	0.72
		FDR	0.23	0.44	0.44	0.19	0.40	0.37	0.08
	$N_g = 3$ Isoforms	Power	0.60	0.82	0.86	0.59	0.73	0.78	0.72
		FDR	0.18	0.47	0.49	0.14	0.39	0.33	0.10

Power and FDR averaged across 100 simulations with (lower half) and without (upper half) outliers. For the isoform simulations, DESeq, edgeR and baySeq were applied to all isoforms collectively, and also applied separately within each subgroup of isoforms, defined by N_g. "Each" is used to denote the corresponding method applied separately within each subgroup. Thresholds were chosen to control FDR at 5% for each approach.

1. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).

2. Stamm, S. *et al.* Function of alternative splicing. *Gene* **344**, 1–20 (2005).

3. Smith, C. W., Patton, J. G. & Nadal-Ginard, B. Alternative splicing in the control of gene expression. *Annu Rev Genet.* **23**, 527–77 (1989).

4. Katz, Y., Wang, E., Airoldi, E. & Burge, C. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature Methods* **7**, 1009–1015 (2010).

5. Singh, D. *et al.* Fdm: a graph-based statistical method to detect differential transcription using rna-seq data. *Bioinformatics* **27**, 2633–2640 (2011).

6. Chang, P., Dunham, J., Nuzhdin, S. & Arbeitman, M. Somatic sex-specific transcriptome differences in drosophila revealed by whole transcriptome sequencing. *BMC Genomics* **12**, 364 (2011).

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

8. Trapnell, C. *et al.* Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature Protocols* **7(3)**, 562–578 (2012).

9. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106 (2010).

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

11. Chin, M. *et al.* Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell stem Cell* **5(1)**, 111–123 (2009).

12. Guenther, G. M., M Gand Frampton *et al.* Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell stem Cell* **7(2)**, 249–257 (2010).

13. Phanstiel, H. P. *et al.* Proteomic and phosphoproteomic comparison of human es and ips cells. *Nature Methods* **8**, 821–827 (2011).

14. Yang, H. & Churchill, G. Estimating p-values in small microarray experiments. *Bioinformatics* **23(1)**, 38–43 (2007).

15. Ohi, Y. *et al.* Incomplete dna methylation underlies a transcriptional memory of somatic cells in human ips cells. *Nat Cell Biol.* **13(5)**, 541–9 (2011).

16. Bock, C. *et al.* Reference maps of human es and ips cell variation enable high-throughput characterization of pluripotent cell lines. *Cell.* **144(3)**, 439–52 (2011).

17. Sen, G., Reuter, J., Webster, D., Zhu, L. & Khavari, P. Dnmt1 maintains progenitor function in self-renewing somatic tissue. *Nature* **463(7280)**, 563–7 (2010).

18. Rai, K. *et al.* Zebra fish dnmt1 and suv39h1 regulate organ-specific terminal differentiation during development. *Mol Cell Biol.* **26(19)**, 7077–85 (2006).

19. D'Aiuto, L. *et al.* Mouse es cells overexpressing dnmt1 produce abnormal neurons with upregulated nmda/nr1 subunit. *Differentiation.* **82(1)**, 9–17 (2011).

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

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

22. Trapnell, C. *et al.* Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28(5)**, 211–215 (2010).

23. Grabherr, M. *et al.* Full-length transcriptome assembly from rna-seq data without a reference genome. *Nat Biotechnol.* **29**, 644–652 (2011).

24. Dempster, A., Laird, N. & Rubin, D. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society* **39**, 1–38 (1977).

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

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

27. Haag, J. D. *et al.* Congenic rats reveal three independent copenhagen alleles within the mcs1 quantitative trait locus that confer resistance to mammary cancer. *Cancer Res* **63**, 5808 (2003).

28. Sengupta, S. *et al.* Highly consistent, fully representative mrna-seq libraries from ten nanograms of total rna. *Biotechniques* **49**, 898–904 (2010).

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

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

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

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

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

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

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

1 Details on the 90 genes identified by EBSeq in the experiment comparing ESCs with iPSCs

Supplementary Table 1 is included as EBSeq90.xls. Each row represents one gene. The columns contain normalized expression estimated via RSEM for each of the 12 ES and 12 iPS cell lines along with the fold change and posterior probability of DE as reported by EBSeq.

2 The N_g effect in multiple datasets

We check the N_g effect in multiple single-end and paired-end datasets processed under different priming protocols, in different labs, using different isoform expression estimation methods.

Supplementary Figures 1(a), 1(c) and 1(d) show data from James Thomson's lab at the Morgridge Institute for Research. For these experiments, which are distinct from the experiments comparing ESCs vs. iPSCs considered in the manuscript, RNA was extracted from human embryonic stem cells and prepared using the *Illumina TrueSeq*, T7LA[28], 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 An-

alyzer 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.

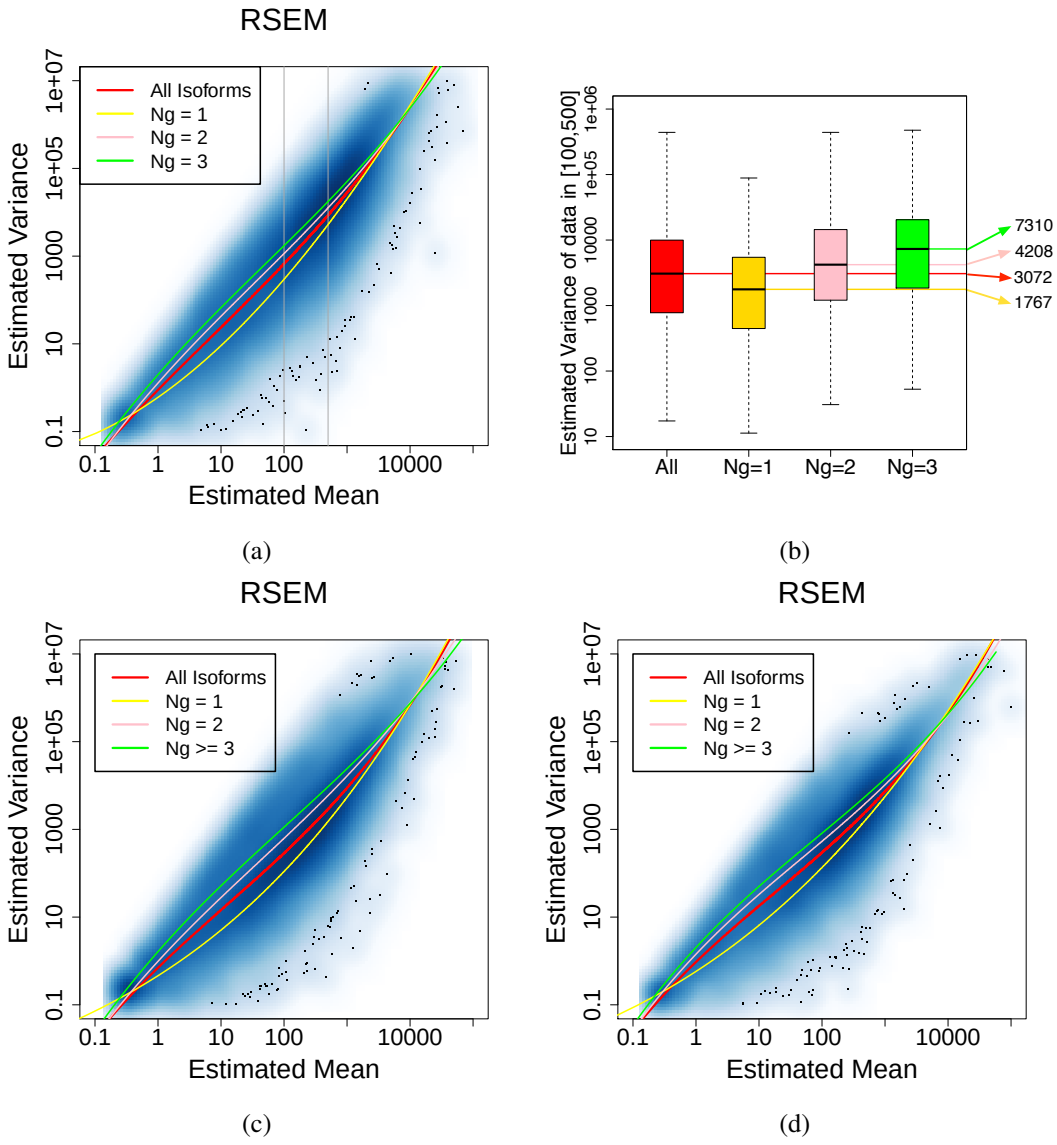
Supplementary Figures 2(a) and 2(b) show data from Michael Gould's lab at UW-Madison. This data is identical to that considered in the mammary carcinoma susceptibility experiment considered in the manuscript, but unlike there (where RSEM was used), here isoform expression was estimated using Cufflinks, with (Supplementary Figure 2(a)) and without (Supplementary Figure 2(b)) multi-read correction.

Supplementary Figures 3(a), 3(b) and 3(c) show three publicly available data sets. Supplementary Figure 3(a) shows data from the Smith lab. Tophat output files were downloaded from GEO GSM792454-61. Eight samples (4 in each of two conditions) were considered here. 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 with multi-read correction.

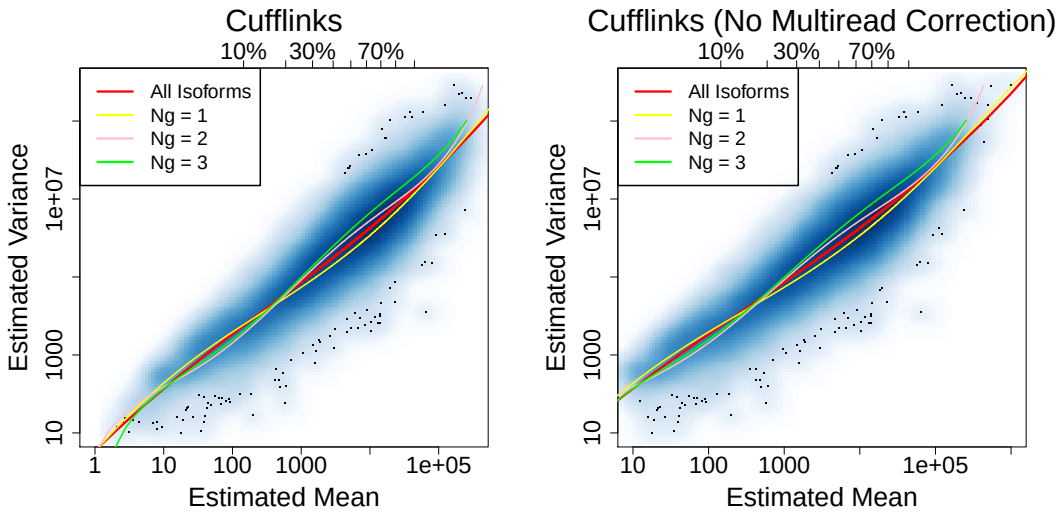
Supplementary Figure 3(b) shows data from the Wold lab[29]. Two samples were considered here. For these, RNA was extracted from mouse brain tissue and prepared using the Solexa protocol. For each dataset, random primers were used. The reads are single-end with read length 25-bp. Alignment was done using Bowtie and Tophat (without de novo transcript detection) with the UCSC mm9 annotation. Isoform expression was estimated using Cufflinks with multi-read correction.

Supplementary Figure 3(c) shows data from the MicroArray Quality Control (MAQC) experiment[30]. The raw read files (fasta format) were downloaded from GEO GSM475204-09. Three samples were considered here. For these, RNA was extracted from human brain tissue. For each dataset, 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[31] with the hg18 RefSeq annotation. Isoform expression was estimated using RSeq[32].

Supplementary Figures 1(a) - 3(c) show 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 mammary carcinoma susceptibility experiment that is shown in Figure 1(c).

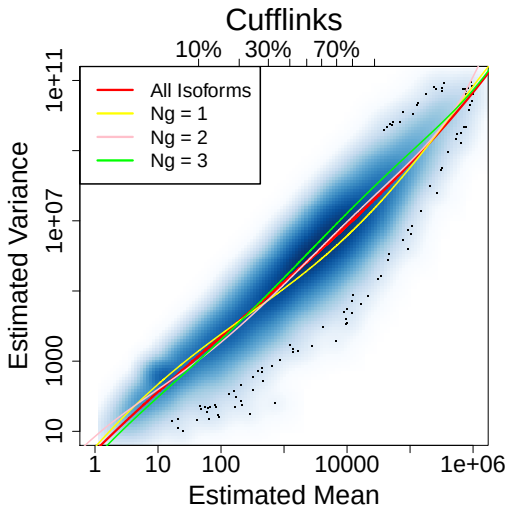


Supplementary Figure 1: Panel (a) shows the empirical variance vs. mean for each isoform profiled in the experiment comparing ESCs with iPSCs (TrueSeq Protocol); details of this experiment are given earlier in this Supplement. A spline fit to all isoforms is shown in red with splines fit within the $N_g = 1$, $N_g = 2$, and $N_g = 3$ isoform groups shown in yellow, pink, and green, respectively. Panel (b) considers isoforms with average expression in [100,500], delineated by the grey lines in panel (a). The range was chosen as it approximates the 60^{th} and 80^{th} percentiles of expression across all isoforms. Shown in panel (b) are box-plots of the variances of these isoforms collectively, and within N_g group. Median variance within each group is shown right. Panels (c) and (d) are similar to (a), but for data processed under the T7LA and MinAmp protocols, respectively.

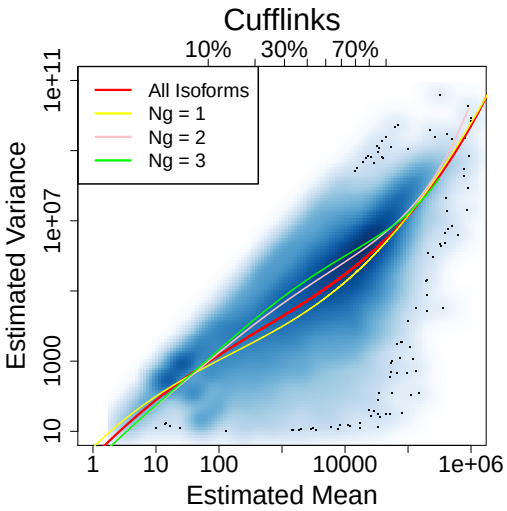


(a) (b)

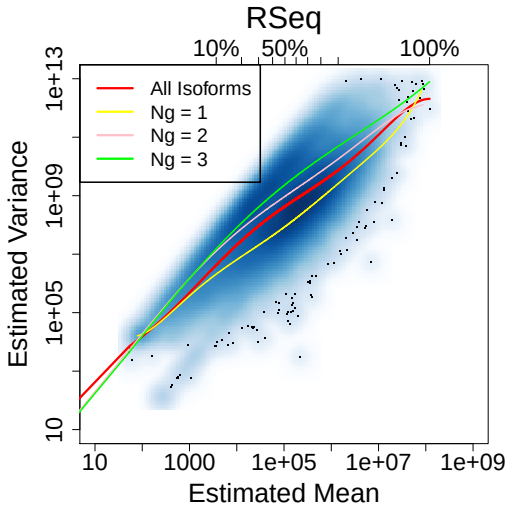
Supplementary Figure 2: Shown are plots similar to Supplementary Figure 1 generated using data from the mammary carcinoma susceptibility experiment; details of this experiment are given in Methods. Isoform expression is estimated via Cufflinks with (panel (a)) and without (panel (b)) the multi-read correction.



(a) Smith Data preprocessed by Cufflinks

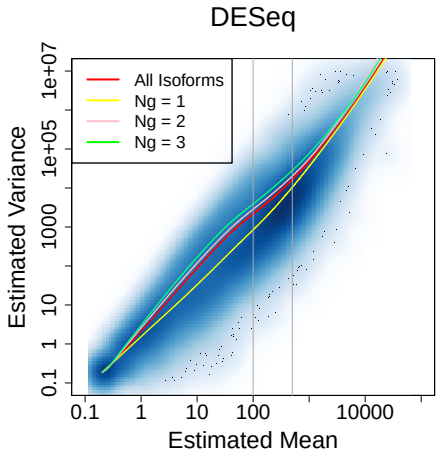


(b) Wold Data preprocessed by Cufflinks

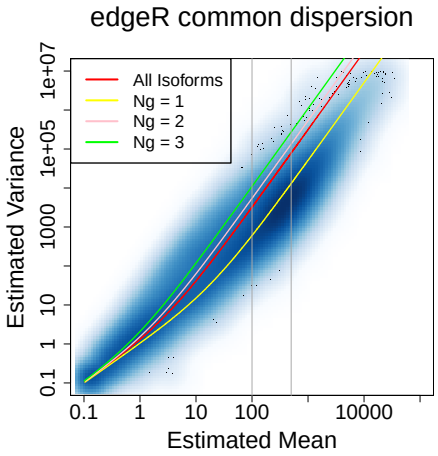


(c) MAQC Data preprocessed by RSeq

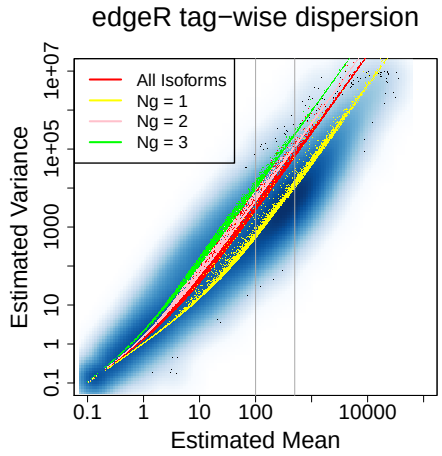
Supplementary Figure 3: Shown are plots similar to Supplementary Figure 1 generated using data from the Smith lab (panel (a)), the Wold lab (panel (b)), and MAQC (panel(c)); details of these experiments are given earlier in this Supplement.



(a) DESeq derived dispersion lines

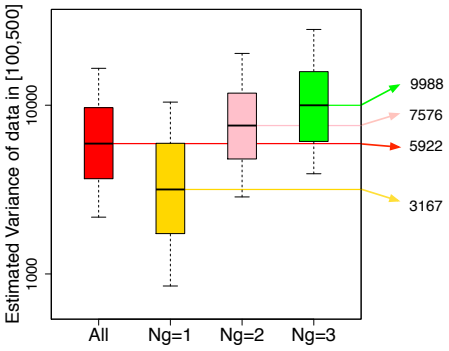


(b) edgeR derived dispersion lines (common dispersion)

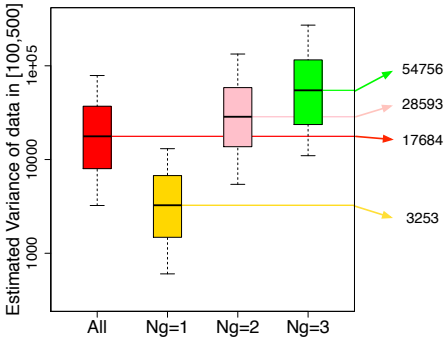


(c) edgeR derived dispersion lines (tag-wise dispersion)

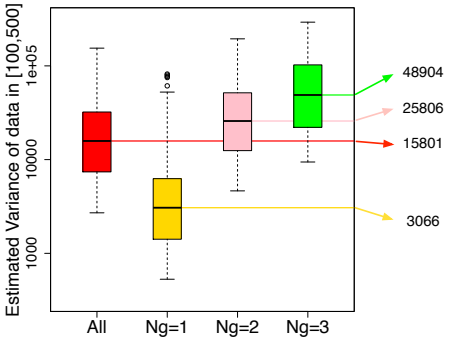
Supplementary Figure 4: Panel (a) shows the fitted dispersion values provided by DESeq applied to the mammary carcinoma susceptibility experiment data shown in Figure 1(c). The dispersion line is calculated across all isoforms (red) and within N_g group (shown in yellow, pink, and green, respectively). Panels (b) and (c) show similar plots from edgeR under their common dispersion (panel (b)) and tag-wise dispersion (panel (c)) models. Supplementary Figures 5(a), 5(b) and 5(c) consider average expression in [100, 500], delineated by the grey lines in panel (a). The range was chosen as it approximates the 60^{th} and 80^{th} percentiles of expression across all isoforms. Shown in Supplementary Figures 5(a), 5(b) and 5(c) are box-plots of the variances of these isoforms collectively, and within N_g group.



(a) Variance estimated using DESeq



(b) Variance estimated using edgeR (common dispersion model)



(c) Variance estimated using edgeR (tag-wise dispersion model)

Supplementary Figure 5: Considered here are isoforms with mean expression in the range [100, 500], delineated by the grey lines in Supplementary Figure 4(a). Shown in panel (a) are boxplots of the variance estimates obtained from DESeq applied to these isoforms collectively, as well as within N_g group. Panels (b) and (c) are similar, with variance estimated via edgeR using the common dispersion (panel (b)) and tag-wise dispersion (panel (c)) model.

3 The Simulation Study

Following Robinson and Smyth[7] and Hardcastle *et al.*[10], counts are assumed to be distributed as Negative Binomial with gene-specific mean in sample s and condition C given by $l_s\mu_g^C$ and variance $l_s\mu_g^C(1 + l_s\mu_g^C\phi_g)$. To reproduce the set-up shown in Figure 3 of Robinson and Smyth[7] and Figure 2 of Hardcastle *et al.*[10], we fixed $\phi_g = 0.17, 0.42$ and 0.95, as done in their work. For each value of ϕ, the 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 mammary carcinoma susceptibility experiment data. 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 half of the DE genes, $\mu_g^{C1} = \delta_g\mu_g^{C2}$ in which δ_g was sampled from the 97%-98% quantile of the empirical gene fold changes (in the range $[2.69 - 3.27]$). For the other half, $\mu_g^{C2} = \delta_g\mu_g^{C1}$. One hundred simulated datasets were considered. BaySeq, DESeq, edgeR and EBSeq were applied to each simulated dataset. We defined thresholds within each method to control FDR at 5%. For baySeq and EBSeq, this implies identifying a gene as DE if its posterior probability of DE exceeds 0.95. For DESeq and edgeR, a gene was defined to be DE if the Benjamini-Hochberg adjusted p-value < 0.05. Default functions were used for each method for library size estimation.

Supplementary Figures 6(a), 6(b) and 6(c) show the ROC curve for each method, averaged over 100 simulations, for each of three overdispersion parameters as considered in Figure 3 of Robinson and Smyth[7] and Figure 2 of Hardcastle *et al.*[10]. As shown there, DESeq and edgeR perform well when overdispersion is minimal but decline dramatically with the increase of overdispersion. EBSeq and baySeq perform consistently across the

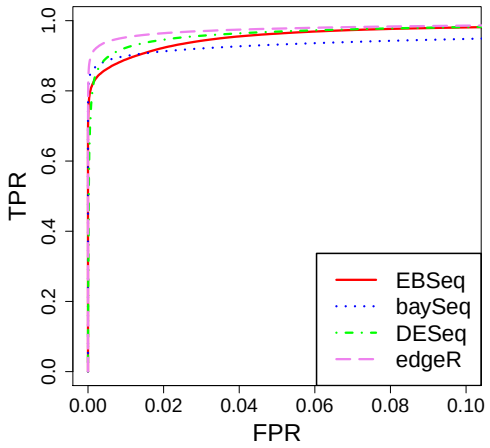
three overdispersion values.

A second set of (more realistic) simulations was conducted by using empirical estimates of overdispersion (results shown in the manuscript). In the isoform level simulations, the $\mu_{g_i}^C$, ϕ_{g_i}'s were sampled as a pair from the 10%-90% quantile of the mammary carcinoma experiment data's empirical ϕ_{g_i}'s within N_g group. Ten percent of the isoforms were simulated as DE. Half of the DE isoforms were defined as $\mu_{g_i}^{C1} = \delta_{g_i}\mu_{g_i}^{C2}$ while the others were $\mu_{g_i}^{C2} = \delta_{g_i}\mu_{g_i}^{C1}$. The δ_{g_i}'s were randomly sampled from the 97%-98% quantile of the empirical isoform fold changes (in the range of $[3.27-4.04]$). Supplementary Figure 7(b) shows that the variance differences between the N_g groups in the isoform simulation without outliers reproduces that observed in the empirical data from which the simulation was generated.

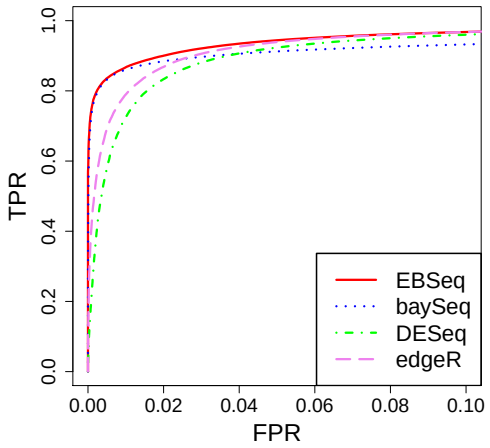
In the simulation with outliers, an extra 10% of the isoforms were simulated to contain a single outlier sample. For these, we first simulated EE isoforms, then multiplied a sample selected at random by 4, 6, 8, or 10. The gene level simulations were conducted similarly. Each simulation was repeated 100 times. The numbers of genes and isoforms simulated were chosen to match the mammary carcinoma experiment data. In particular, 20,267 genes and 27,468 isoforms were simulated (15,324, 6,902 and 5,242 within each N_g group). We only consider the non-zero genes and isoforms.

BaySeq, DESeq, edgeR, and EBseq were applied to each dataset. For the isoform simulations, baySeq, DESeq, and edgeR were applied to all isoforms and also to subsets of isoforms defined by N_g group. As in the gene-level simulations, target FDR was set

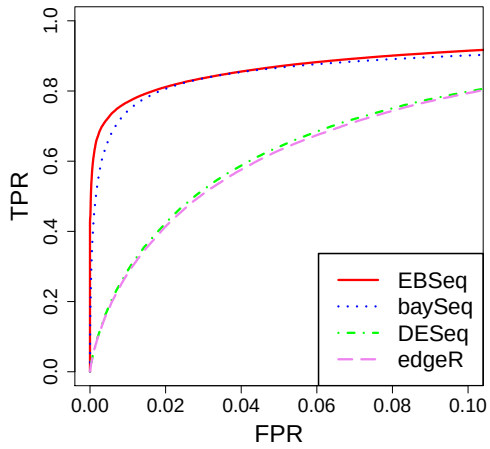
at 5%. For baySeq and EBSeq, this corresponds to identifying an isoform as DE if the posterior probability of DE exceeds 0.95. For DESeq and edgeR we identify an isoform as DE if the Benjamini Hochberg adjusted p-value is less than 0.05. Table 1 shows the power and false discovery rate (FDR) averaged across 100 simulations.



(a) $\phi_g = 0.17$

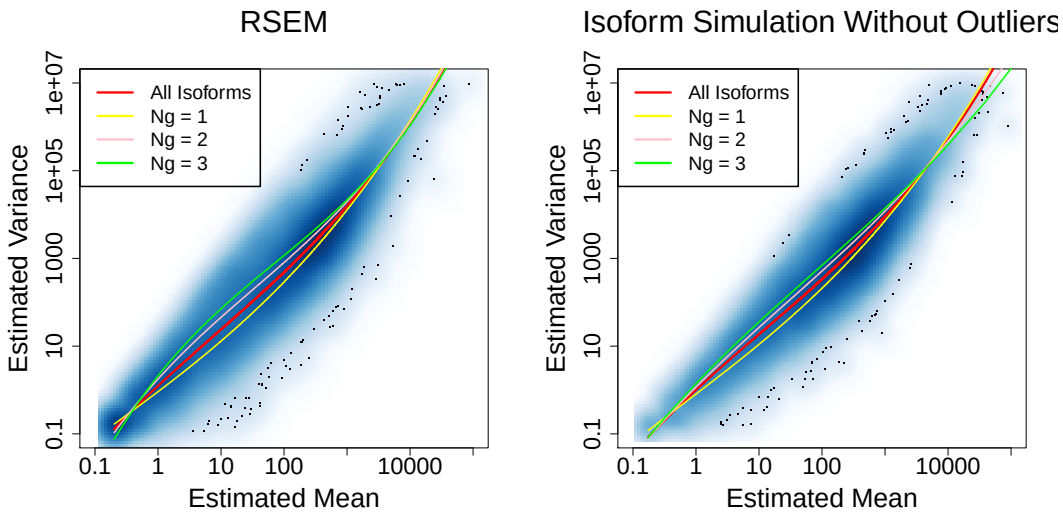


(b) $\phi_g = 0.42$



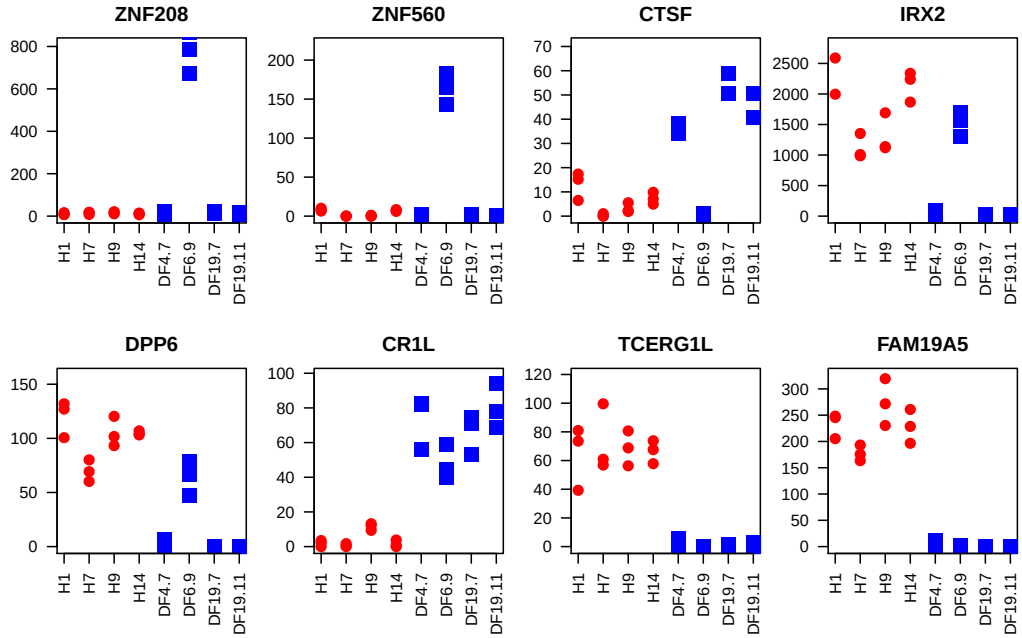
(c) $\phi_g = 0.95$

Supplementary Figure 6: The ROC curve averaged over 100 simulations, using the simulation set up shown in Figure 3 of Robinson and Smyth[7] and Figure 2 of Hardcastle *et al.*[10].



(a) (b)

Supplementary Figure 7: Panel (a) shows empirical variance vs. mean for isoforms profiled in the mammary carcinoma susceptibility experiment (this figure is identical to Figure 1(c)). Panel (b) shows the same, but using one of the simulated datasets (without outliers). A spline fit to all isoforms is shown in red with splines fit within the $N_g = 1$, $N_g = 2$, and $N_g = 3$ isoform groups shown in yellow, pink, and green, respectively.



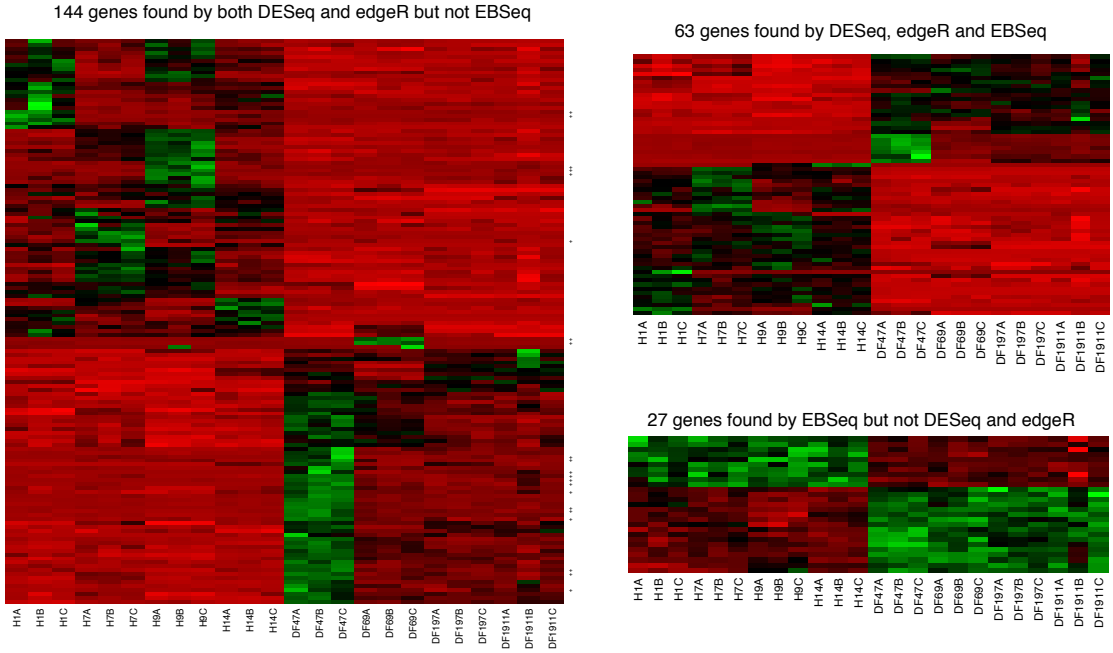
Supplementary Figure 8: Shown are 8 genes identified as DE by Phanstiel *et al.*[13] (see their Supplementary Table 9). Four (upper panel) were identified as consistently DE by DESeq and edgeR, but not EBSeq (the two condition model) in three experiments; and four (lower panel) were identified consistently by all three methods. Using the 18 pattern EBSeq model, the genes are classified as EE but containing one outlier (ZNF208, ZNF560), DE but containing one outlier (CTSF, IRX2, DPP6 and CR1L), and DE without outliers (TCERG1L and FAM19A5) using a posterior probability cutoff of 0.95.

4 Case Study Results

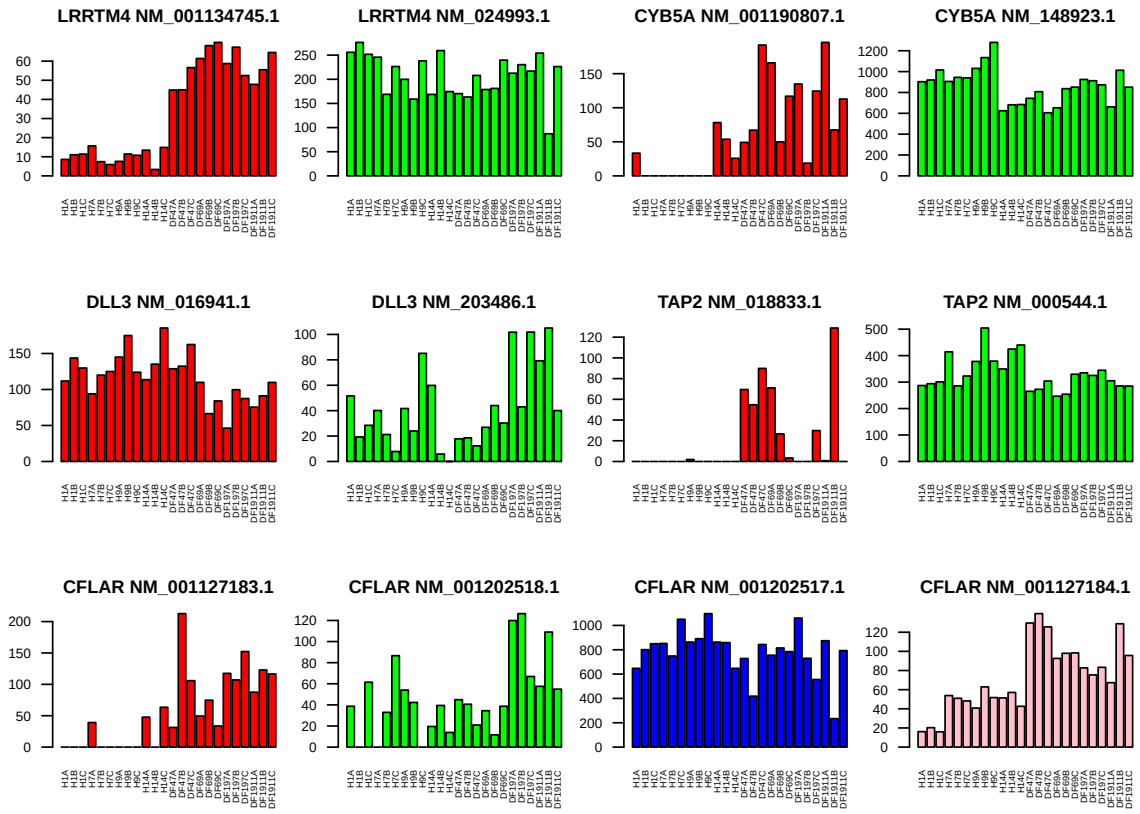
Supplementary Table 2: ESCs vs. iPSCs

		CuffDiff	DESeq	edgeR	EBSeq
Number identified as DE	All Genes	6	260	252	90
	All Isoforms	4	157	161	135
	$N_g = 1$ Isoforms	4	29	20	68
	$N_g = 2$ Isoforms	0	58	55	37
	$N_g = 3$ Isoforms	0	70	86	30
Proportion identified as DE	All Genes	0%	1.3%	1.3%	0.5%
	All Isoforms	0%	0.5%	0.5%	0.4%
	$N_g = 1$ Isoforms	0%	0.2%	0.1%	0.5%
	$N_g = 2$ Isoforms	0%	0.8%	0.7%	0.5%
	$N_g = 3$ Isoforms	0%	0.8%	0.9%	0.3%

The table shows the number of genes and isoforms identified as DE in the experiment comparing ESCs with iPSCs. There are 19,784 genes and 30,563 isoforms with non-zero expression on average across the three experiments. Isoform group sizes are 13,493, 7,721, and 9,349 corresponding to $N_g = 1, 2$ and 3, respectively. CuffDiff, DESeq, edgeR and EBSeq were applied to all genes and isoforms. Shown in the top half of the table are the number of DE genes and isoforms consistently identified by each method (identified in each of the 3 replicate experiments); the percent identified is shown in the bottom half. BaySeq did not converge for over 50% of the genes and consequently results are not shown. The target FDR was set at 5%.



Supplementary Figure 9: The left panel shows a heat map of the 144 genes consistently called as DE by DESeq and edgeR, but not EBSeq, in the experiment comparing ESCs with iPSCs; green (red) shows high (low) expression. The upper right panel shows the 63 genes consistently called as DE by all three methods and the bottom right panel shows the 27 genes consistently called as DE by EBSeq but not DESeq and edgeR. Genes identified as EE or EE with an outlier cell line in the EBSeq multiple condition analysis are marked "+"; 26 of the 144, 0 of the 27, and 0 of the 63 were identified as EE or EE with an outlier cell line.

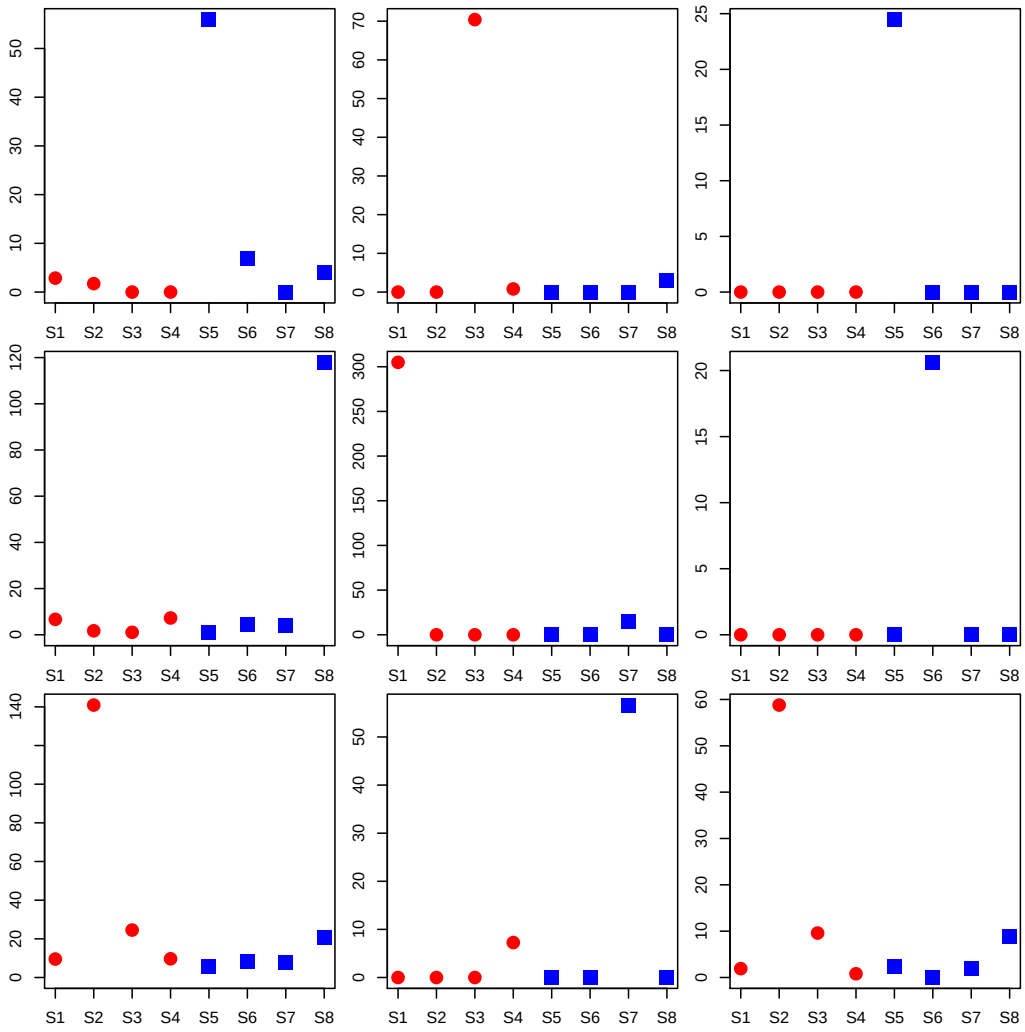


Supplementary Figure 10: Shown are 5 out of 308 genes identified as non-DE by EBSeq's multiple condition model (gene level posterior probability of DE and DEO < 0.95) with DE isoforms (isoform level posterior probability of DE or DEO > 0.95) in the experiment comparing ESCs with iPSCs. Each bar shows the isoform expression in each sample; expression of the constituent isoforms is shown in different colors within each gene.

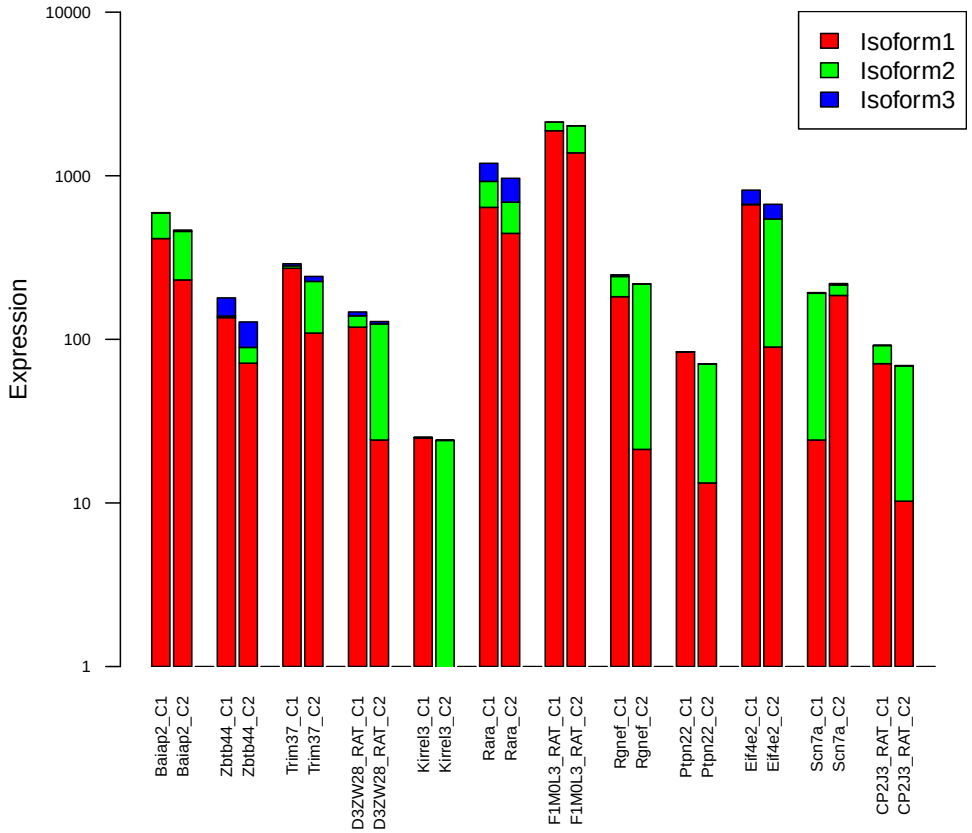
Supplementary Table 3: Mammary carcinoma susceptibility

		baySeq	CuffDiff	DESeq	edgeR	EBSeq
Number identified as DE	All Genes	62	117	403	505	323
	All Isoforms	66	138	786	941	849
	$N_g = 1$ Isoforms	39	95	135	129	488
	$N_g = 2$ Isoforms	19	22	303	370	185
	$N_g = 3$ Isoforms	8	21	348	442	176
Percent identified as DE	All Genes	0.3%	0.6%	2.0%	2.5%	1.6%
	All Isoforms	0.2%	0.5%	2.9%	3.4%	3.1%
	$N_g = 1$ Isoforms	0.3%	0.6%	0.9%	0.9%	3.1%
	$N_g = 2$ Isoforms	0.3%	0.3%	4.4%	5.4%	2.7%
	$N_g = 3$ Isoforms	0.2%	0.4%	6.6%	8.4%	3.4%

The table shows the number of genes and isoforms identified as DE in the mammary carcinoma susceptibility experiment. There are 20,267 non-zero expressed genes in total and 27,468 isoforms with group sizes 15,324, 6,902 and 5,242 corresponding to $N_g = 1, 2$ and 3, respectively. baySeq, CuffDiff, DESeq, edgeR and EBSeq were applied to all genes and isoforms. Shown in the top half of the table are the number of DE genes and isoforms identified by each method; the percent identified is shown in the bottom half. The target FDR was set at 5%.



Supplementary Figure 11: Shown are 9 genes (out of 311) in the mammary carcinoma susceptibility experiment identified as DE by DESeq and edgeR, but not EBSeq. The x-axis indicates the 8 samples (4 in each of two conditions denoted by red and blue, respectively); the y-axis shows normalized expression.



Supplementary Figure 12: Shown are 12 genes identified as EE by EBSeq (posterior probability of DE < 0.005) with DE isoforms (posterior probability of DE > 0.95) in the mammary carcinoma susceptibility experiment. Each pair of bars shows the average gene expression in each condition; constituent isoform expression is shown in color within each bar.

5 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 $q_{g_i}^{C}$:$q_{g_i}^{C}|\alpha, \beta^{N_g} \sim Beta(\alpha, \beta^{N_g})$. The hyper parameter α is shared by all the isoforms and β^{N_g} is N_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

$$H_0 \text{ (EE)} : q_{g_i}^{C1} = q_{g_i}^{C2} \text{ vs } H_1 \text{ (DE)} : q_{g_i}^{C1} \neq q_{g_i}^{C2}.$$

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

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

Alternatively (DE), $X_{g_i}^{C1,C2}$ follows the prior predictive distribution $f_1^{N_g}(X_{g_i}^{C1,C2})$:

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

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:

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

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

$$\frac{p f_1^{N_g}(X_{g_i}^{C1,C2})}{(1-p) f_0^{N_g}(X_{g_i}^{C1,C2}) + p f_1^{N_g}(X_{g_i}^{C1,C2})} \tag{5}$$

Parameter estimation With the assumption that $r_{g_i,s} = r_{g_i,0} l_s$, denote $\mu_{g_i,0}^C$ and $(\sigma_{g_i,0}^C)^2$ are 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 $\hat{\mu}_{g_i,0} = \frac{\hat{\mu}^{C1}_{g_i,0}+\hat{\mu}^{C2}_{g_i,0}}{2}$ and $\hat{\sigma}^2_{g_i,0} = \frac{(\hat{\sigma}^{C1}_{g_i,0})^2+(\hat{\sigma}^{C2}_{g_i,0})^2}{2}$ Then the estimator of $r_{g_i,0}$ is obtained by $\hat{r}_{g_i,0} = \frac{\hat{\mu}^2_{g_i,0}}{\hat{\sigma}^2_{g_i,0}-\hat{\mu}_{g_i,0}}$.

$\hat{l_s}$ could be obtained by RPKM[29], TMM[33], Median Normalization[9], or Quantile Normalization[34]. Since RPKM 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 the α, β^{N_g} and p via the **optim** function in **R**.

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: $q^{C1}_{g_i} = q^{C2}_{g_i} = q^{C3}_{g_i}$; $q^{C1}_{g_i} = q^{C2}_{g_i} \neq q^{C3}_{g_i}$; $q^{C1}_{g_i} = q^{C3}_{g_i} \neq q^{C2}_{g_i}$; $q^{C1}_{g_i} \neq q^{C2}_{g_i} = q^{C3}_{g_i}$; and $q^{C1}_{g_i} \neq q^{C2}_{g_i} \neq q^{C3}_{g_i}$.

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

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

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

$$\sum_{k=1}^{5} p_k g_k^{N_g}(X_{g_i}^{C1,C2,C3}) \tag{6}$$

in which $\sum_{k=1}^{5} p_k = 1$.

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

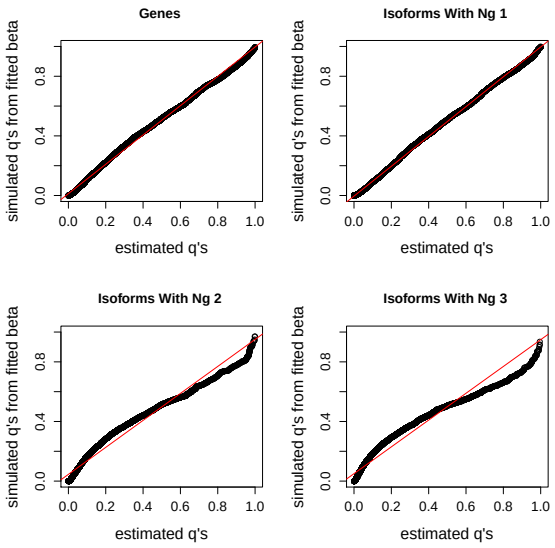
$$\frac{p_K g_K^{N_g}(X_{g_i}^{C1,C2,C3})}{\sum_{k=1}^{5} p_k g_k^{N_g}(X_{g_i}^{C1,C2,C3})} \tag{7}$$

In the evaluation of outliers, we considered 18 patterns: EE and DE between ESCs and iPSCs (2 patterns), EE with an outlier in one of the 8 cell lines (8 patterns; one for each cell line), and DE with an outlier in one of the 8 cell lines (8 patterns; one for each cell line). A gene is classified into a pattern with FDR α if the posterior probability of that pattern exceeds $1 - \alpha$.

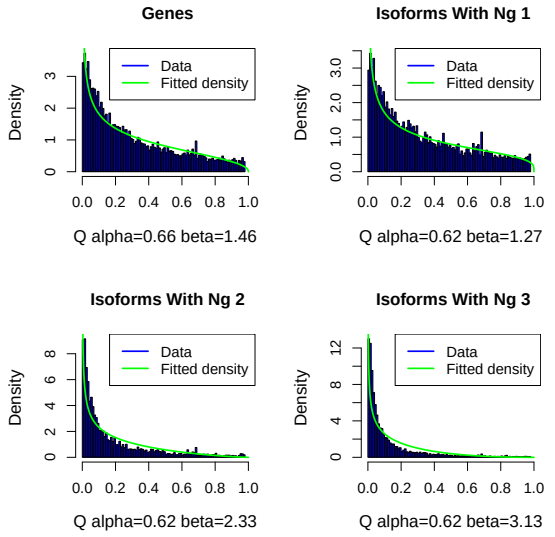
6 Model Diagnostics

Supplementary Figure 13(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 5 of this Supplement using data from one of the three experiments for the comparison of ESCs with iPSCs. Supplementary Figure 13(b) shows the histogram of estimated $q_{g_i}^{C1}$'s (q_g^{C1}'s) and the fitted Beta density using that

same data. Figures 14(a) and 14(b) show the same, but using data from the mammary carcninoma susceptibility experiment. These figures indicate that the prior assumed by EBSeq is reasonable for the experiments considered here.

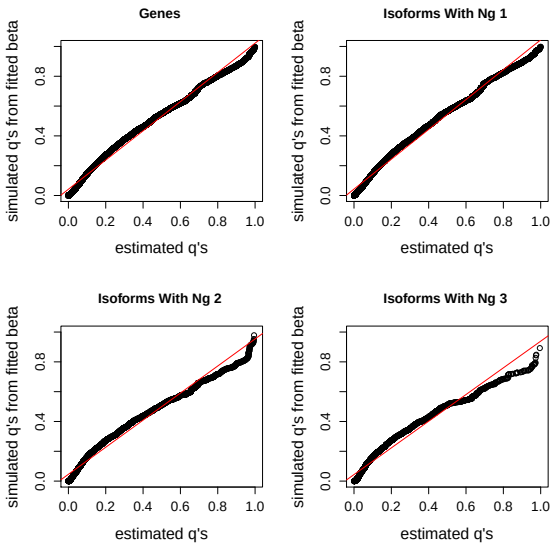


(a) The QQ plot for model diagnostics in the ESCs vs. iPSCs experiment

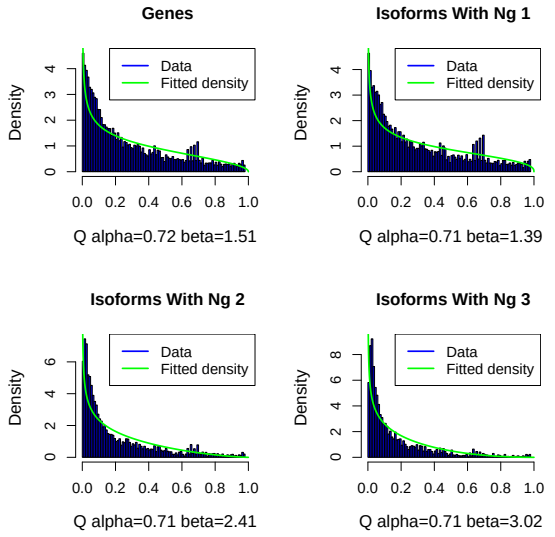


(b) The density plot for model diagnostics in the ESCs vs. iPSCs experiment

Supplementary Figure 13: A QQ-plot (a) 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. Panel (b) shows a histogram of the estimated $q_{g_i}^{C1}$'s (q_g^{C1}'s) and the corresponding Beta densities.



(a) The QQ plot for model diagnostics in the mammary carcinoma susceptibility experiment



(b) The density plot for model diagnostics in the mammary carcinoma susceptibility experiment

Supplementary Figure 14: A QQ-plot (a) 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. Panel (b) shows a histogram of the estimated $q_{g_i}^{C1}$'s (q_g^{C1}'s) and the corresponding Beta densities.