PLOS | ONE

RESEARCH ARTICLE

Modeling Exon-Specific Bias Distribution Improves the Analysis of RNA-Seq Data

Xuejun Liu*, Li Zhang, Songcan Chen

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China

* xuejun.liu@nuaa.edu.cn

Abstract

RNA-seq technology has become an important tool for quantifying the gene and transcript expression in transcriptome study. The two major difficulties for the gene and transcript expression quantification are the read mapping ambiguity and the overdispersion of the read distribution along reference sequence. Many approaches have been proposed to deal with these difficulties. A number of existing methods use Poisson distribution to model the read counts and this easily splits the counts into the contributions from multiple transcripts. Meanwhile, various solutions were put forward to account for the overdispersion in the Poisson models. By checking the similarities among the variation patterns of read counts for individual genes, we found that the count variation is exon-specific and has the conserved pattern across the samples for each individual gene. We introduce Gamma-distributed latent variables to model the read sequencing preference for each exon. These variables are embedded to the rate parameter of a Poisson model to account for the overdispersion of read distribution. The model is tractable since the Gamma priors can be integrated out in the maximum likelihood estimation. We evaluate the proposed approach, PGseq, using four real datasets and one simulated dataset, and compare its performance with other popular methods. Results show that PGseq presents competitive performance compared to other alternatives in terms of accuracy in the gene and transcript expression calculation and in the downstream differential expression analysis. Especially, we show the advantage of our method in the analysis of low expression.

Citation: Liu X, Zhang L, Chen S (2015) Modeling Exon-Specific Bias Distribution Improves the Analysis of RNA-Seq Data. PLoS ONE 10(10): e0140032. doi:10.1371/journal.pone.0140032

Editor: Yan W. Asmann, Mayo Clinic, UNITED STATES

Received: January 14, 2015

Accepted: September 21, 2015

Published: October 8, 2015

Data Availability Statement: The four datasets used in the manuscript are freely available in public repositories. Please see the references for further information. Our simulation data is available via GitHub (https://github.com/PUGEA/PGSeq).

Funding: XL is supported by the National Science Foundation of China (61170152), LZ is supported by the Fundamental Research Funds for the Central University (CXZZ11_0217), and SC is supported by Qing Lan project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Introduction

Alternative splicing (AS) is a common phenomenon observed in eukaryotes. In this process, the multiple exons for a gene are connected in multiple ways, leading to various protein isoforms. It has been found that AS exists for more than 95% human genes [1]. Unexpected variation in AS is often associated to many diseases [2]. Therefore, the study on the variation of AS has received more and more interest in the area of biomedicine in recent years. The analysis of gene and isoform expression provides an important approach to study the variation of AS. RNA-Seq technology offers a vital tool to quantify transcript expression by generating millions

Competing Interests: The authors have declared that no competing interests exist.

of short transcript reads from an RNA population of biological samples [3]. The processing of RNA-Seq data typically involves three aspects [4]. First, reads are aligned to a reference genome or transcriptome. Second, the expressed genes and isoforms are assembled by using mapped reads. Third, given a transcriptome assembly gene and isoform expression can be calculated by counting the reads mapped to a gene and the associated isoforms. Naturally, differential expression of transcripts can also be analyzed using the obtained expression quantification. However, this paper just focuses on the third aspect in the processing of RNA-Seq data.

The expression quantification from short reads ($25 \sim 300$ base pairs) is challenging. First, many reads are mapped to multiple isoforms, which belong to the same gene, since they share exons. For the paralogous genes with close sequences, it is possible to map reads to multiple genes. This read assignment uncertainty makes accurate expression quantification difficult since it is unclear which isoform a read originates if it comes from a shared exon across multiple isoforms. As read counts are proportional to the abundance of the fragments originating from a gene, the RPKM (reads per kilobase of transcript per million mapped reads) was proposed to represent the expression level of genes [5]. However, this method cannot be directly used to calculate the expression for isoforms due to read assignment uncertainty. In order to solve this problem, various Poisson-based approaches are proposed to model reads that are mapped ambiguously to multiple transcripts [6–8]. These methods utilize the additive property of the Poisson distribution to deal with the ambiguities of read mappings. Alternatively, other sophisticated approaches use probabilistic graphic models to simulate the stochastic process of generating read sequences, such as RSEM [9, 10], BitSeq [11] and the approach proposed in [12].

Second, many expression calculation approaches assume the uniform distribution of the read counts along the reference sequence, such as the RPKM representation and the Poisson-based method [6]. However, read counts follow obviously non-uniform distribution in reality due to the positional, sequencing and mappability biases [13]. This violates the uniform assumption of Poisson models. In order to obtain accurate expression estimates, these biases should be modeled and removed. Therefore, a number of bias correction strategies have been introduced to the Poisson-based models to account for the non-uniformity in read sequencing rates [8, 14, 15]. [14] used a linear model to explicitly estimate the sequencing preference of the Poisson rate at each nucleotide position based on the local surrounding sequences. [15] proposed a two-parameter generalized Poisson model for gene and exon expression computation. One parameter was used to represent expression and the other was used to model the average sequencing bias. [8] used bias curves to characterize the non-uniformity of read distributions, and incorporated these curves into a Poisson model to relieve the effects of sequencing bias. RSEM allowed the use of an empirical read start position distribution to account for the non-uniformity of the read distribution in a generative graphical model [9]. Cufflinks used a variable length Markov model to learn sequence-specific bias and positional bias based on the surrounding sequences [16]. [17] proposed a Bayesian network to predict bias at each position within a locus. The prediction can be used to adjust the biased read counts. CEM used a quasi-multinomial distribution model to capture various types of RNA-Seq biases [13]. [18] made use of the multi-sample information to jointly estimate the isoform expression and the isoform-specific read bias factor. By correcting for fragment bias, all these methods have been proved to obtain improved expression estimates showing that the bias correction is vital for accurate expression quantification. However, most of these methods explicitly estimate the average bias based on the sequence contents and position information in empirical data and do not consider the diversity of bias pattern among genes. Some approaches can model gene- or isoform-specific biases, but they use only a single point estimate to account for the average of bias properties. They therefore ignore the variability of biases across samples.

By checking the similarities among the variation patterns of read counts for individual genes, we found that the count variation is exon-specific and has the conserved pattern across different conditions for each individual gene (see the section of Materials and Methods). The same phenomenon was also observed in [14, 18]. The read sequencing preference in RNA-seq data is analogous to the probe affinities in microarray data [20–22]. In microarrays, probes have different sensitivities to the specific hybridization signals. The pattern of probe intensities varies in a gene-specific way and is also conserved for the same gene across various conditions. This characteristic has been intensively studied in the area of microarray analysis and a number of approaches have been proposed to model the probe affinity [20, 22–24]. [23] introduced the Gamma distributed latent variables to model the probe-specific sensitivity for the Affymetrix oligonuleotide arrays. This strategy has been proven to be effective in terms of both computational accuracy and efficiency.

In this paper, we propose a statistical model, PGseq (Poisson-Gamma model for RNA-Seq data), to estimate accurate gene and isoform expression by accounting for the exon-specific bias for each gene. Our approach uses Poisson distribution to model the read counts and borrows the idea in [23] of using Gamma distributed latent variables to capture the overall exon-specific read bias for each gene. An important feature of our new model is that it accounts for the distribution of the read bias instead of using a single point estimate to represent the average bias properties or explicitly depicting individual types of biases by taking into account the sequence content and position information. The bias modeled in our method is exon-specific and shared across all conditions for each individual gene, it can therefore automatically capture all the intrinsic exon-specific effects, including the sequence-specific and positional effects. Another advantage of this strategy is that the exon-specific variables representing the overall bias can be integrated out of the likelihood and this leads to an efficient maximum likelihood (ML) solution of the model. Finally, in addition to calculating gene/isoform expression our method also provides a level of uncertainty associated with these measurements. This level of uncertainty can be used to improve the downstream analysis, such as the detection of differential expression (DE).

Materials and Methods

Modeling distribution of exon-specific bias

In RNA-Seq data analysis, it is natural to use Poisson distribution to model read counts. However, the assumption of the constant rate is violated by the serious overdispersion of read counts. Fig 1 shows the counts of a randomly selected gene ENSG00000197746 (PSAP) of four tissues from the Human Body Map project (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/). It can be seen that the counts are highly non-uniformly distributed and the variation patterns are almost consistent across the tissues. The same characteristic also holds for other genes and for other species [14]. We also plot the normalized read counts (by exon length) for each exon along this gene in different tissues as shown in Fig 2. The variation patterns are highly conserved across tissues for the normalized counts of exons. Analogously, there is high variability for the probe intensities within the same probe-set in microarray data and the variation due to probe effects is larger than the variation across technical replicates [20, 22–24]. Like the analogy with microarrrays, some exons will be preferentially sequenced due to the technical biases (e.g. GC content, secondary structure, distance from the 3' end). There is also a more obvious reason for the low counts of some exons which is that these exons are not incorporated into transcripts as frequently. As we can see from Figs 1 and 2, the two exons with the lowest read counts in the middle part of gene ENSG00000197746 are used by a single transcript respectively, while other exons with the high read counts are included in at least two

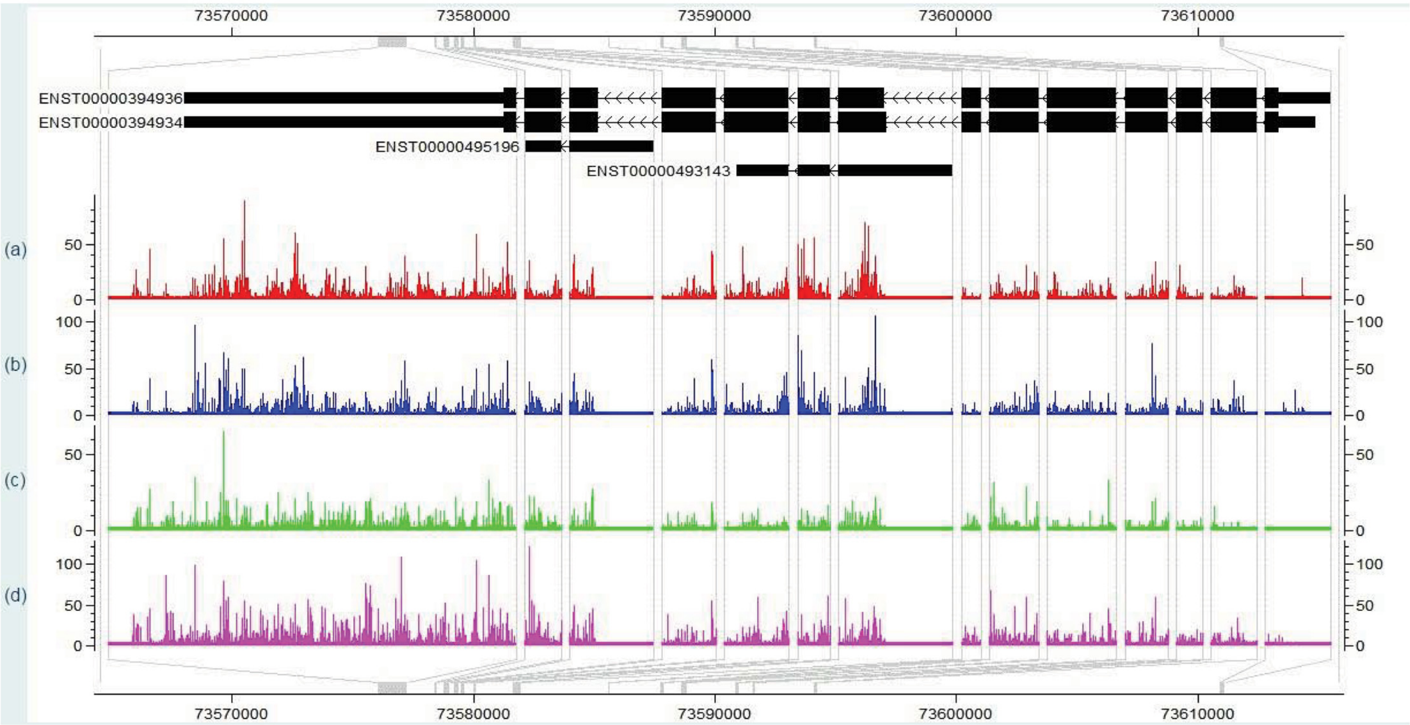

Fig 1. Read counts for each exonic nucleotide (nt) position in CisGenome Browser [19] along gene ENSG00000197746 (PSAP) in different tissues of the Human Body Map dataset (a) brain, (b) breast, (c) liver and (d) lung. Counts for reads starting at each exonic nucleotide position are shown.

doi:10.1371/journal.pone.0140032.g001

transcripts. Considering all these bias sources, we model the exon-specific sequencing preference in our approach.

We borrow the strategy of modeling the probe effects in microarray data in [23] to deal with the exon-specific fluctuation of read counts in RNA-Seq data. We believe the variability in the sequencing preference can explain the consistent overdispersion pattern across samples in the read counts. Accordingly, we introduce a latent variable β_i for exon i to describe the associated sequencing preference and share β_i across multiple samples. We assume that $\{\beta_i\}$ are independent and identically distributed random variables and are drawn from a gene-specific Gamma distribution (we ignore the subscript of gene for notation simplicity),

$$\beta_i \sim \mathrm{Ga}(a, b), \tag{1}$$

with a shape parameter a and a scale parameter b. As we can see from Fig 2 that although the overall count variation patterns are highly consistent across tissues, we can still observe the visible variability of this pattern among samples. For this reason, unlike using a single point value [13, 15, 18] to summarize all possible biases which were revealed in the previous studies [9, 14, 16, 17], we use the Gamma distribution in Eq (1) to describe the stochastic property of the gene-specific sequencing preference for each exon. It thus covers all possible types of the intrinsic sequencing biases and considers the variability of bias distribution across samples.

The Poisson-Gamma model

For each gene, let y_{icr} represent the observed read count mapped to the ith exon for the rth technical or biological replicate under the cth condition. Allowing any number of isoform contributions to y_{icr}, y_{icr} can be decomposed as the sum of the normalized count contributions

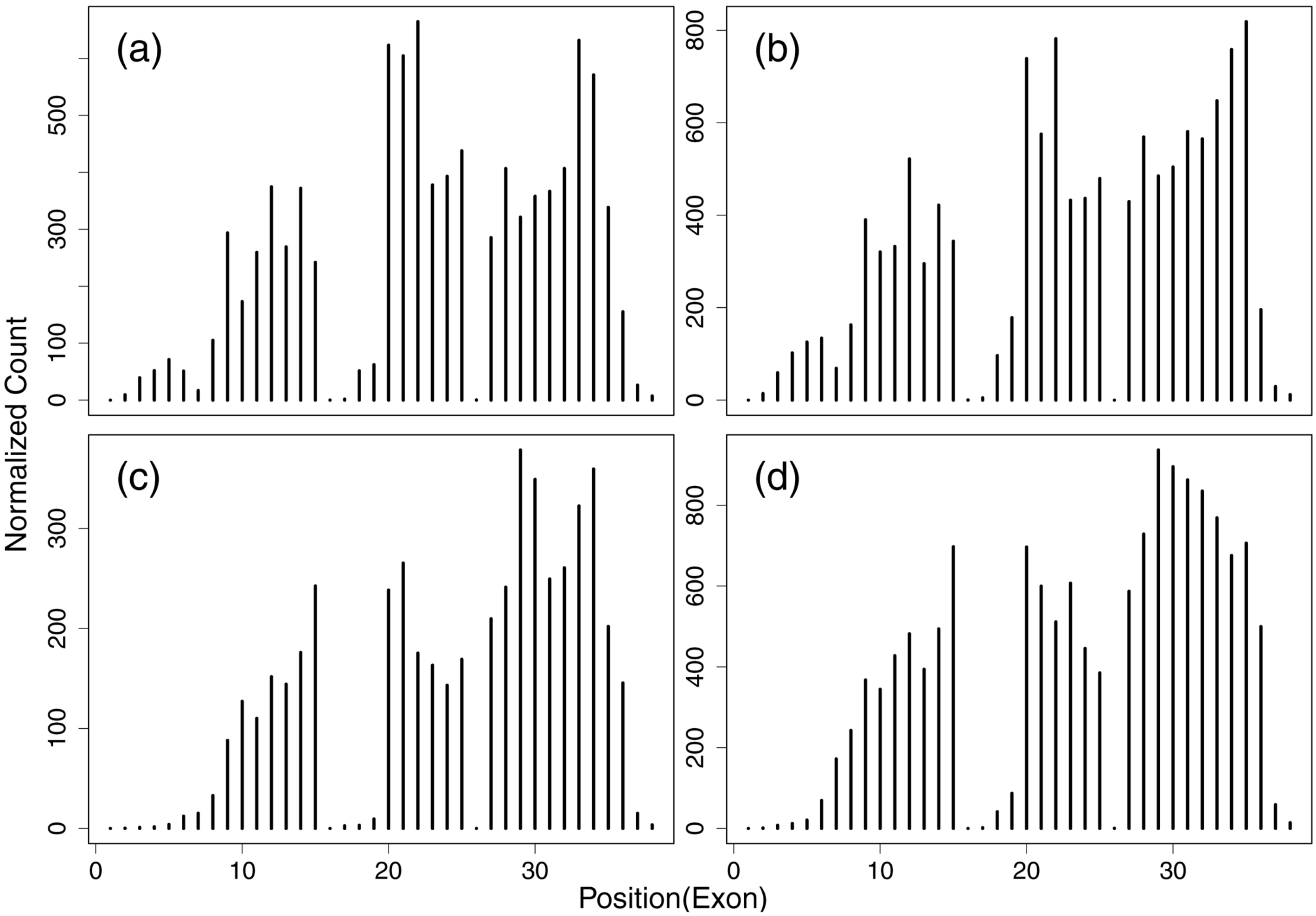

Fig 2. Normalized read counts for each exon along gene ENSG00000197746 (PSAP) in different tissues of the Human Body Map dataset (a) brain, (b) breast, (c) liver and (d) lung. Reads mapped to each exon are counted and normalized according to the exon length. Overlapping exons are segmented into multiple non-overlapping ones.

doi:10.1371/journal.pone.0140032.g002

from isoforms t_{icrk}, in which exon i is included. We assume $y_{icr} = w_{cr}\ l_i \Sigma_k M_{ik}\ t_{icrk}$, where w_{crl} and l_i are the scaling factors related to the sequencing depth and the exon length, respectively, and M_{ik} is defined as the indicator functions $M_{ik} = 1$ if exon i belongs to transcript k. Here, we assume t_{icrk} follow a Poisson distribution, $t_{icrk} \sim \text{Pois}(\beta_i \alpha_{crk})$, where β_i is the sequencing preference of exon i and follows the distribution in Eq (1). We share β_i across all replicates and conditions in order to capture the consistent sequencing preference for each exon. The parameter α_{crk} is a quantity proportional to the abundance of transcript k for replicate r under condition c. Under this assumption, y_{icr} also follows a Poisson distribution,

$$y_{icr} \sim \text{Pois}\left(w_{cr} l_i \beta_i \sum_k M_{ik} \alpha_{crk} \right). \tag{2}$$

The reads mapped to the exon-junction region proportionally contribute to the counts for each adjacent exon according to the fraction of the length of the mapping sequence on each exon. For paired-end data, we count the fragment between paired reads. If fragments cover multiple exons, we add the fraction of the mapping sequence length to the count of each involved exon.

Also, overlapping exons are divided into multiple exons to avoid the redundant read counting. Note that MMSEQ [7] also uses a Poisson-Gamma model to fit read counts, whereas the Gamma prior is put on transcript expression and it does not consider the variability of the sequencing preference for each fragment region.

The likelihood of observed counts for a specific gene is

$$\begin{aligned} &L(\{y_{icr}\}|\{\alpha_{crk}\}, a, b) \\ &= \prod_i \int \mathrm{d}\beta_i P(\beta_i|a, b) \prod_{cr} P\left(y_{icr}|w_{cr} l_i \beta_i \sum_k M_{ik}\alpha_{crk}\right). \end{aligned} \tag{3}$$

The conjugacy of the Poisson-Gamma model makes the integral tractable. But there are no closed-form solution for the parameters of this model, $\{\alpha_{crk}\}$, a and b. By using the efficient optimization tool, such as donlp2 [25], the ML estimates of these parameters, $\{\hat{\alpha}_{crk}\}$, $\hat{a}$ and $\hat{b}$, can be easily obtained. Practically, we find that the large number of $c \times r$ can help with the estimation of the sequencing preference distribution. For the experiments where biological and technical replicates are not available, data from the multiple lanes or runs of a single library can be taken as "technical" replicates to aid the estimation of model parameters.

Expression inference

In our model, α_{crk} represents the relative transcript abundance across conditions and replicates for the same transcript k. However, it cannot represent the absolute expression across transcripts belonging to different genes, as the gene-specific random variable β_i also accounts for the overall read count of the transcript across conditions and replicates. For this consideration, we choose to use the normalized read count t_{icrk} to represent transcript expression since it considers the effect of β_i. This representation is different from other Poisson-based approaches for the reason that they do not adopt a gene-specific parameter to the Poisson rate as β_i in our model.

We assume that the normalized read counts of transcript k on all exons, $t_{\cdot crk}$, are independent and identically distributed random variables. With the estimated model parameters, we can obtain the distribution of β_i, $P(\beta_i|\hat{a}, \hat{b})$, which represents the distribution of the gene-specific sequencing preference for all exons. Considering all possible values of latent random variable β_i, each of $t_{\cdot crk}$ follows the same distribution

$$\begin{aligned} P(t_{icrk}) &= \int \mathrm{d}\beta_i P(t_{icrk}|\hat{\alpha}_{crk}\beta_i) P(\beta_i|\hat{a}, \hat{b}) \\ &= \mathrm{NB}\left(\hat{a}, \frac{\hat{\alpha}_{crk}}{\hat{b} + \hat{\alpha}_{crk}}\right), \end{aligned} \tag{4}$$

where NB denotes the negative binomial distribution. The expectation and variance of t_{icrk} are then

$$\langle t_{icrk} \rangle = \frac{\hat{a}}{\hat{b}} \hat{\alpha}_{crk} \text{ and } \mathrm{Var}[t_{icrk}] = \frac{\hat{a}}{\hat{b}^2} \hat{\alpha}_{crk} \left(\hat{b} + \hat{\alpha}_{crk}\right). \tag{5}$$

Assume that the normalized gene expression, s_{icr}, can be calculated as the sum of the transcript contributions, i.e. $s_{icr} = \sum_k t_{icrk}$. Similarly, the gene expression on all exons $s_{\cdot cr}$ are also assumed to be independent and identically distributed random variables, each of which follows the same distribution

$$\begin{aligned} P(s_{icr}) &= \int \mathrm{d}\beta_i P\left(s_{icrk} \middle| \sum_k \hat{\alpha}_{crk}\beta_i\right) P\left(\beta_i | \hat{a}, \hat{b}\right) \\ &= \mathrm{NB}\left(\hat{a}, \frac{\sum_k \hat{\alpha}_{crk}}{\hat{b} + \sum_k \hat{\alpha}_{crk}}\right). \end{aligned} \tag{6}$$

The expectation and variance of s_{icr} are therefore

$$\langle s_{icr} \rangle = \frac{\hat{a}}{\hat{b}} \sum_k \hat{\alpha}_{crk} \text{ and } \mathrm{Var}[s_{icr}] = \frac{\hat{a}}{\hat{b}^2} \sum_k \hat{\alpha}_{crk} \left(\hat{b} + \sum_k \hat{\alpha}_{crk}\right). \tag{7}$$

Using sampling from the negative binomial distributions in Eqs (4) and (6), the expectation and variance of the logarithmic transcript/gene expression can be obtained. This expression representation is useful for propagating the measurement error in the subsequent downstream analyses of both gene and transcript expression.

Note that the transcript and gene expression are both expressed as negative binomial distributions here. The two-parameter NB distribution has been thought to be advantageous for modeling the read counts in the differential expression analysis of RNA-Seq data due to its ability to model the overdispersion in read distributions [26–28]. The expression in Eq (7) has the similar parametrization as the NB model, sSeq, proposed in [28]. The expected expression μ and the dispersion parameter ϕ in sSeq are analogous to $\frac{a}{b}\alpha$ and $\frac{1}{a}$, respectively, in our method. However, our approach is different from sSeq and other NB-based approaches. First, instead of modeling the distribution of the whole count for each gene we model the variability of the count for each individual exon. Consequently, it is possible to estimate the gene-specific bias distribution and to apply the ML method for parameter estimation. Second, we decompose the total count for each exon into a sum of the contributions from related transcripts and thus obtain the expression for each transcript which can be useful for downstream transcript level analyses.

Software

The proposed PGseq method is implemented in Python and C. After aligning the primary reads to the reference transcriptome by Bowtie 2 [29], the alignment files are then processed using our Python scripts to obtain the read counts for each exon. We employ the fast optimization toolkit, donlp2 [25], to estimate the model parameters. Both parameter optimization and expression inference are implemented in fast C codes. We also make use of parallel computing to improve the computation efficiency. The software and documentation are freely available online from the website https://github.com/PUGEA/PGSeq.

Datasets

We evaluate the proposed approach, PGseq, on the estimation of gene and isoform expression using three real datasets and one simulated dataset, and considering both single-end and paired-end data.

We use the well studied Microarray Quality Control (MAQC) dataset [30] to validate the expression estimation from PGseq at gene level. MAQC project measured gene expression from high-quality RNA samples to assess the comparability across multiple platforms. This dataset has been widely used as the benchmark to verify various analysis methods [31–33]. We select two RNA samples, the universal human reference (UHR) RNA and the human brain reference (HBR) RNA, from the Illumina platform. The Short Read Archive accession number is SRA010153 for single-end data and SRA012427 for paired-end data. Around one thousand genes have been measured by the qRT-PCR experiments and can be served as the gold standard to benchmark the gene expression estimation obtained from other platforms. We used the Ensembl annotation data (NCBI37/hg19) and obtained 841 matching qRT-PCR validated genes. Among these qRT-PCR validated genes, we use the method in [32] to filter 217 DE genes and 88 non-DE genes with high confidence according to the qRT-PCR measurements. Data of these 305 qRT-PCR validated genes is used as a gold standard to evaluate the sensitivity and the specificity of various DE analysis approaches.

A real human colorectal cancer (HCC) dataset [34] is also used to further validate the gene expression estimation of PGseq. In this dataset, the fluorouracil-resistant (MIP101) and -non-resistant (MIP/5-FU) human colorectal cancer cell lines were investigated using paired-end RNA-seq experiments. Since there are no biological or technical replicates in this dataset, we select seven lanes for each condition and take them as seven "technical" replicates in order to obtain better estimation of model parameters. For each replicate we use about 9 million reads. Reads are aligned using Ensembl annotation data (NCBI36/hg18). There are 192 genes which were quantified by the qPCR experiments in this dataset. The number of qPCR validated genes is reduced to 101 by merging redundancy and being successfully mapped to the reference. The qPCR measurements of these 101 genes are used to validate the gene expression estimates of our method. Among these genes, we use the similar selection method in [28] to choose 21 DE genes and 14 non-DE genes with high confidence. Data of these 35 genes is also used as a gold standard to compare methods in DE analysis.

A publicly available human breast cancer (HBC) dataset [35] is used to validate the estimation of transcript expression of PGseq. This dataset contains single-end data and includes two conditions, the human breast cancer cell line (MCF-7) and the normal cell line (HME). Four genes (TRAP1, ZNF581, WISP2 and HIST1H2BD) which contain multiple transcripts were validated using the qRT-PCR experiments [36]. Two transcripts for each gene have been interrogated for both cell lines. We used the UCSC knownGene transcriptome annotation (NCBI36/hg18) for obtaining all annotation information for the eight qRT-PCR validated transcripts.

Since the true expression for a large number of transcripts are not available, we generated simulated data using our model based on the calculated parameters from the qRT-PCR validated genes of HBR sample in MAQC dataset. This dataset is mainly used for sanity checking of our method. Around 100 million reads are generated individually for each of the seven "technical" replicates. Since we simulate data of a single condition, we omit the subscript c in the following mathematical symbols. For each gene, we first sample β_i from $\mathrm{Ga}(\hat{a}, \hat{b})$ and α_{rk} from $\mathrm{N}(\hat{\alpha}_{1k}, \hat{\alpha}_{1k}/20)$, where $\hat{\alpha}_{1k}$ is the estimated parameters for the first replicate. The count for each exon, y_{ir}, is then drawn from $\mathrm{Pois}(w_l l_i \hat{\beta}_i \sum_k M_{ik} \hat{\alpha}_{rk})$. Reads with count y_{ir} are then sequenced from the reference sequence by considering the start position along the reference. If $y_{ir}/l_i > 0.1$, reads are sequenced according to the true histogram in the real dataset. Otherwise, reads are uniformly sequenced. The length of the sampled fragments is 35 base pairs for the sing-end dataset, and sampled from N(206, 19.6) for the paired-end dataset. The sequenced read length for both ends of paired-end data is 50. The constants used in the above distributions are chosen based on empirical data. The sampled reads hold the consistent realistic non-uniformity in the distribution across technical replicates.

Finally, we use the human brain dataset (HBD) downloaded from DDBJ [37] with accession number SRA009447 to show the use of our method for datasets with biological replicates. We also use this dataset to verify the performance of our approach for lowly expressed genes. HBD dataset includes two conditions, the adult and fetal human brains, each of which contains three biological replicates. For each biological replicate, two or three technical replicates were used. We pool reads from technical replicates for each biological replicate and make expression estimation from the six pooled sets.

Results

We compare PGseq with other three popular alternatives, Cufflinks v2.2.1, RSEM v1.2.19 and MMSEQ v1.0.7, for gene and transcript expression estimation. All these softwares are used with default parameters. We use the MAQC, HCC and simulation datasets to evaluate the performance of PGseq on gene expression estimation. The HBC and simulation datasets are used to verify the accuracy of our method for isoform expression estimation. Finally, we apply PGseq to DE analysis and compare it with other competitive approaches.

Transcript expression deconvolution

The major advantage of PGseq is that it is able to deconvolute the read counts for a gene to obtain the individual NB distributed isoform expression by considering the gene-specific read bias distribution. Before evaluating the accuracy of expression measurements estimated from our method, we randomly select two examples, genes ENSG00000130816 and ENSG00000152291, from the MAQC SRA010153 data to show this advantage as presented in Figs 3 and 4, respectively. As we can see that these two genes present very different count variation patterns. Consequently, we obtain the different bias distributions as shown in subplots (c) in both figures. We believe the obtained bias distributions are able to capture the gene-specific count variation patterns for individual genes. We note that gene ENSG00000130816 is up-regulated in sample UHR while gene ENSG00000152291 is invariant across the two samples by comparing the observed total read counts for both samples. From Fig 3 we can find that even though the gene is obviously differentially expressed, isoform ENST00000340748 is low expressed and largely invariant across the two samples while the other three are up-regulated in sample UHR. Fig 4 indicates that the gene expression is unchanged while many isoforms are differentially expressed across the two samples. The examples here show that a reasonable approach, which is able to accurately deconvolute transcript expression, is vital for the investigation of AS variation.

Validation of gene expression estimation

The SRA010153 data in the MAQC dataset contains two RNA samples, HBR and UHR. Each sample includes seven lanes which can be seen as seven “technical” replicates. The SRA012427 data contains three technical replicates for the single sample UHR. We apply all approaches to each replicate and compute the average gene expression for each sample. All the methods are run with the bias correction mode turned on. The squared Pearson correlation coefficient (R^2) of the logarithmic average gene expression estimates with the logarithmic qRT-PCR measurements for the 841 mapping qRT-PCR validated genes are calculated as shown in the first three rows in Table 1. We can see that PGseq outperforms the other three alternatives for the comparisons on all three samples. Notice that Cufflinks, RSEM and MMSEQ all obtain less consistent results with the qRT-PCR results for sample UHR using the paired-end data than the single-end data. This contradicts the common sense that the paired-end protocol is more advantageous than the single-end protocol and thus should lead to more accurate expression measurements. We find that only PGseq produces more accurate expression estimates for the paired-end data than the single-end data. The fourth and fifth rows show the calculation accuracy of various methods at gene level against the qPCR measurements for the two samples in the HCC dataset. For this dataset, PGseq performs as almost equally well as other approaches for sample MIP101, but significantly better for sample MIP/5-FU. Similarly, the last two rows

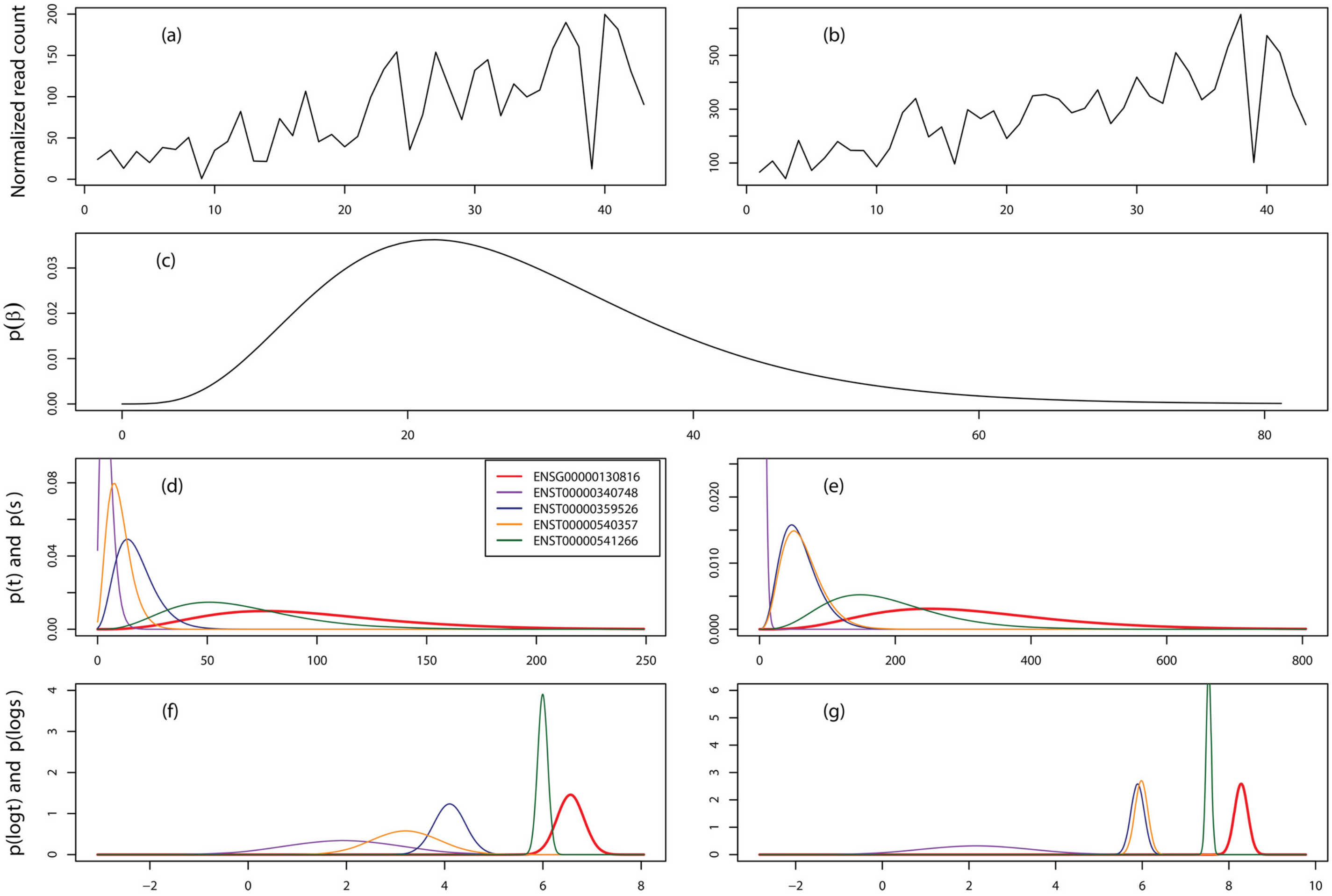

Fig 3. Transcript expression deconvolution by PGseq for gene ENSG00000130816. (a) and (b) show the normalized read counts for samples HBR and UHR, respectively, for each exon. (c) shows the estimated gene-specific read distribution. The 3rd panel shows the obtained NB distributions for both gene and isoforms for the two samples. The last panel presents the approximated Gaussian distributions of logged gene and isoform expression. This gene contains four isoforms and the total normalized read counts for both samples are 601 and 2341, respectively.

doi:10.1371/journal.pone.0140032.g003

show the comparison results of expression estimation accuracy against the ground truth at gene level using the simulated datasets. Since the data is generated from our model, it is not surprising that PGseq obtains the most accurate results than other approaches. This shows the consistency of our model. Note that although the data is biased to PGseq, the other three methods also present the high accuracy on gene expression calculation showing good performance of these approaches on gene expression calculation.

We have shown that PGseq provides better correlation with the qRT-PCR findings. Next, we divide the 841 qRT-PCR validated genes in the MAQC data set into three groups with low, medium and high expression. The genes with qRT-PCR measurement below 0.02 are assigned to the "low" group, 0.02 and 0.2 to the "medium" group and above 0.2 to the "high" group. For each group we examine the correlation between the calculated gene expression with the qRT-PCR measurements to reveal the performance of our approach for individual group with low, medium and high expression, respectively. Table 2 shows the correlation of each method for each group. We can see that PGseq presents the outstanding performance for the low group, while obtains moderate results for the medium and high groups. This testifies that the

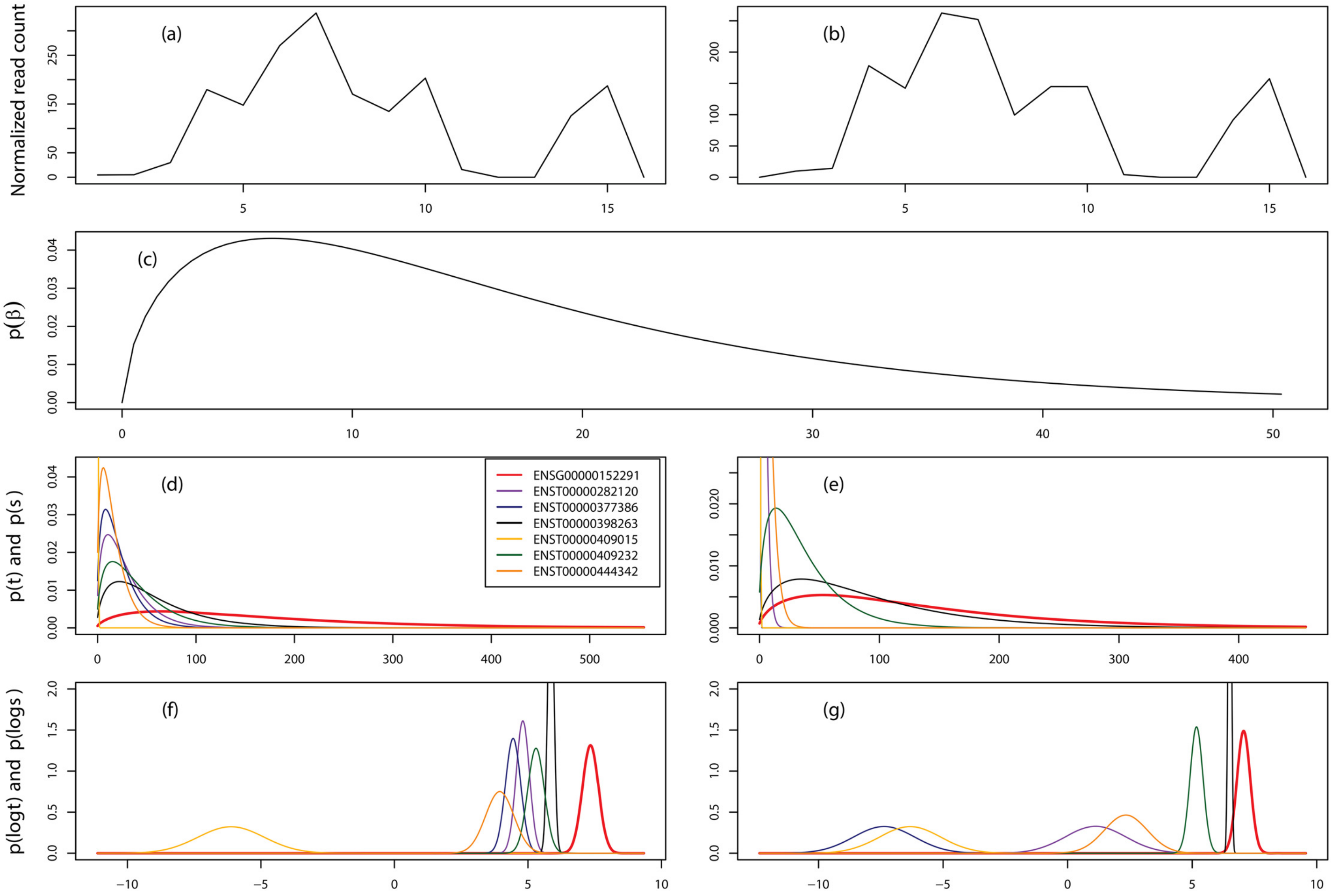

Fig 4. Transcript expression deconvolution by PGseq for gene ENSG00000152291. (a) and (b) show the normalized read counts for samples HBR and UHR, respectively, for each exon. (c) shows the estimated gene-specific bias distribution. The 3rd panel shows the obtained NB distributions for both gene and isoforms for the two samples. The last panel presents the approximated Gaussian distributions of logged gene and isoform expression. This gene contains six isoforms and the total normalized read counts for the two samples are both 1321.

doi:10.1371/journal.pone.0140032.g004

Table 1. Comparison of expression estimation accuracy at gene level using various datasets.

Dataset	Cufflinks	RSEM	MMSEQ	PGseq
MAQC.HBR.SE	0.812	0.808	0.800	**0.845**
MAQC.UHR.SE	0.837	0.840	0.832	**0.854**
MAQC.UHR.PE	0.723	0.800	0.805	**0.860**
HCC.MIP101.PE	0.770	**0.785**	0.779	0.776
HCC.MIP/5-FU.PE	0.844	0.853	0.852	**0.881**
Simulation.SE	0.930	0.974	0.974	**0.979**
Simulation.PE	0.950	0.981	0.980	**0.984**

For the MAQC and HCC datasets, the R^2 correlation coefficients of the logarithmic average expression for the matching PCR-validated genes with the logarithmic qRT-PCR or qPCR results are calculated. Two samples (HBR and UHR) in single-end (SE) data (SRA010153) and one sample (UHR) in paired-end (PE) data (SRA012427) are used for the MAQC dataset. Seven lanes with 9 million paired-end reads for each lane are used for the HCC dataset. For the simulated dataset, the R^2 correlation coefficients of the estimated gene expression with the ground truth are calculated. Both single-end and paired-end simulated data are used. The best result for each comparison is highlighted in bold.

doi:10.1371/journal.pone.0140032.t001

Table 2. Comparison of gene expression estimation accuracy for groups with different level of expression.

Dataset		Cufflinks	RSEM	MMSEQ	PGseq
Low	MAQC.HBR.SE(282)	0.333	0.326	0.303	**0.492**
	MAQC.UHR.SE(258)	0.257	0.268	0.250	**0.382**
	MAQC.UHR.PE(282)	0.203	0.328	0.357	**0.530**
Medium	MAQC.HBR.SE(246)	0.452	0.499	0.472	**0.520**
	MAQC.UHR.SE(246)	**0.480**	0.473	0.446	0.406
	MAQC.UHR.PE(246)	0.399	0.480	0.489	**0.533**
High	MAQC.HBR.SE(313)	**0.707**	0.694	0.673	0.663
	MAQC.UHR.SE(337)	0.690	**0.719**	0.693	0.693
	MAQC.UHR.PE(313)	0.658	**0.682**	0.678	0.649

For the MAQC dataset, the R^2 correlation coefficients of the logarithmic average expression measurements with the logarithmic qRT-PCR results. Data are divided into the "low", "medium" and "high" groups according to the level of the qRT-PCR measurements. The number after each dataset shows the number of genes belonging to this group. The best result for each comparison is highlighted in bold.

doi:10.1371/journal.pone.0140032.t002

overall outperformance of PGseq is mainly due to its superiority in low expression. Normally, it is difficult to measure expression of genes with low read counts because there is usually high level of noise associated to these genes [38].

Validation of transcript expression estimation

We use a real human breast cancer dataset with qRT-PCR validation to verify the transcript expression measured from PGseq. The qRT-PCR measurements for eight transcripts under the two conditions are taken as the gold standard to compare the performance of various methods. We calculate R^2 correlation coefficients between the obtained logarithmic transcript expression with the logarithmic qRT-PCR measurements. The consistency of results between various approaches with qRT-PCR experiments is shown in the first row in Table 3. We can see that PGseq obtains the most consistent results with the qRT-PCR measurements as compared with other alternatives.

We then use the simulated datasets to verify the transcript expression estimation of our method for a larger number of transcripts. The last two rows in Table 3 show the comparison results of expression estimation accuracy against the ground truth at transcript level. It can be found that all the four methods obtain more consistent results with the ground truth using the paired-end data than the single-end data. However, the obtained R^2 values are much lower than those for gene expression calculation in Table 1. This shows that the computation of transcript expression is much more difficult than that of gene expression. We find that PGseq outputs the highest accuracy for both comparisons and the superiority for the single-end data is

Table 3. Comparison of expression estimation accuracy at transcript level.

Dataset	Cufflinks	RSEM	MMSEQ	PGseq
HBC	0.558	0.615	0.578	**0.657**
Simulation.SE	0.724	0.768	0.738	**0.853**
Simulation.PE	0.831	0.862	0.836	**0.900**

For the HBC dataset, the R^2 correlation coefficients of the logarithmic average expression measurements with the logarithmic qRT-PCR results for the eight validated transcripts under the two conditions are calculated. For the simulated dataset, the R^2 correlation coefficients with the ground truth are calculated. Both single-end and paired-end simulated data are simulated. The best result for each comparison is highlighted in bold.

doi:10.1371/journal.pone.0140032.t003

especially significant. Note that both PGseq and MMSEQ are Poisson-Gamma models, and the difference between these two models is that PGseq considers the variability of the gene-specific read sequencing preference for each exon while MMSEQ does not. The comparison results demonstrate that properly modeling the distribution of the sequencing preference contributes to the estimation of transcript expression.

Propagating measurement error in DE analysis

Our method is also advantageous for providing measurement error in expression estimates. Propagating measurement uncertainty in the downstream analyses has been approved to obtain biologically more relevant results for microarray data analyses [39–41]. The recent study in [42] has also shown that accounting for posterior uncertainty in expression measurements can improve the power of DE analysis of RNA-seq data. We make use of the proposed DE analysis method, MMDiff, in [42] to show the usefulness of the measurement error obtained by PGseq in DE analysis.

The left column in Fig 5 shows the scatter plots of the standard deviation vs. the logarithmic gene expression calculated from PGseq for the PCR-validated genes in the MAQC and HCC datasets. It can be seen that as the expression increases the measurement error decreases. For the genes with logged expression below -5, the measurement error does not increase accordingly. By investigating the raw read data, we find that the read counts related to these genes for all the samples are close to zero, the obtained low expression estimates are then associated with relatively high certainty. In order to show the usefulness of the calculated measurement error of PGseq, we make use of MMDiff, which is able to propagate expression measurement error to the DE analysis, and combine PGseq and MMDiff to produce receiver operator characteristic (ROC) curves for the true DE genes and the non-DE genes by considering the measurement error and ignoring this error. The comparison results are shown in the right column in Fig 5. By considering the measurement error calculated from PGseq, we obtain the better ROC curves for both datasets than ignoring measurement error (i.e. setting zero measurement error). The area under the ROC curve (AUC) for the MAQC dataset is 0.959 if ignoring the measurement error, while 0.977 if considering the measurement error. For the HCC dataset, AUC values for ignoring and considering the measurement error are 0.801 and 0.912, respectively. This demonstrates that the obtained measurement error from PGseq significantly helps with the downstream DE analysis.

We produce the ROC curves for various combination of the expression estimation methods and the DE analysis approaches as shown in Fig 6. The corresponding AUCs for the MAQC and HCC datasets are shown in Table 4. Cufflinks and MMSEQ are combined with their own embedded DE methods, Cuffdiff [43] and MMDiff, respectively. The embedded DE method of RSEM is EBSeq [44]. We find that DESeq [26] combined with RSEM obtains better results than EBSeq (data not shown) and we thus choose DESeq as the DE method for RSEM in the following comparisons. The combination of PGseq and MMDiff obtains higher accuracy than other combinations for both datasets. Note that even though using the same DE analysis method, PGseq still outperforms MMSEQ. Since PGseq and MMSEQ are both Poisson-Gamma models, we believe that the difference in the performance is due to the fact that PGseq models the distribution of exon-specific sequencing bias while MMSEQ does not take this into consideration. Comparisons in this section show that modeling bias distribution in expression estimation can lead to improved DE analysis results for RNA-seq data.

Finding DE for lowly expressed genes

Finally, we consider a real HBD dataset which includes biological replicates. In this dataset, we pool the technical replicates and consider only biological replicates. We apply the above four

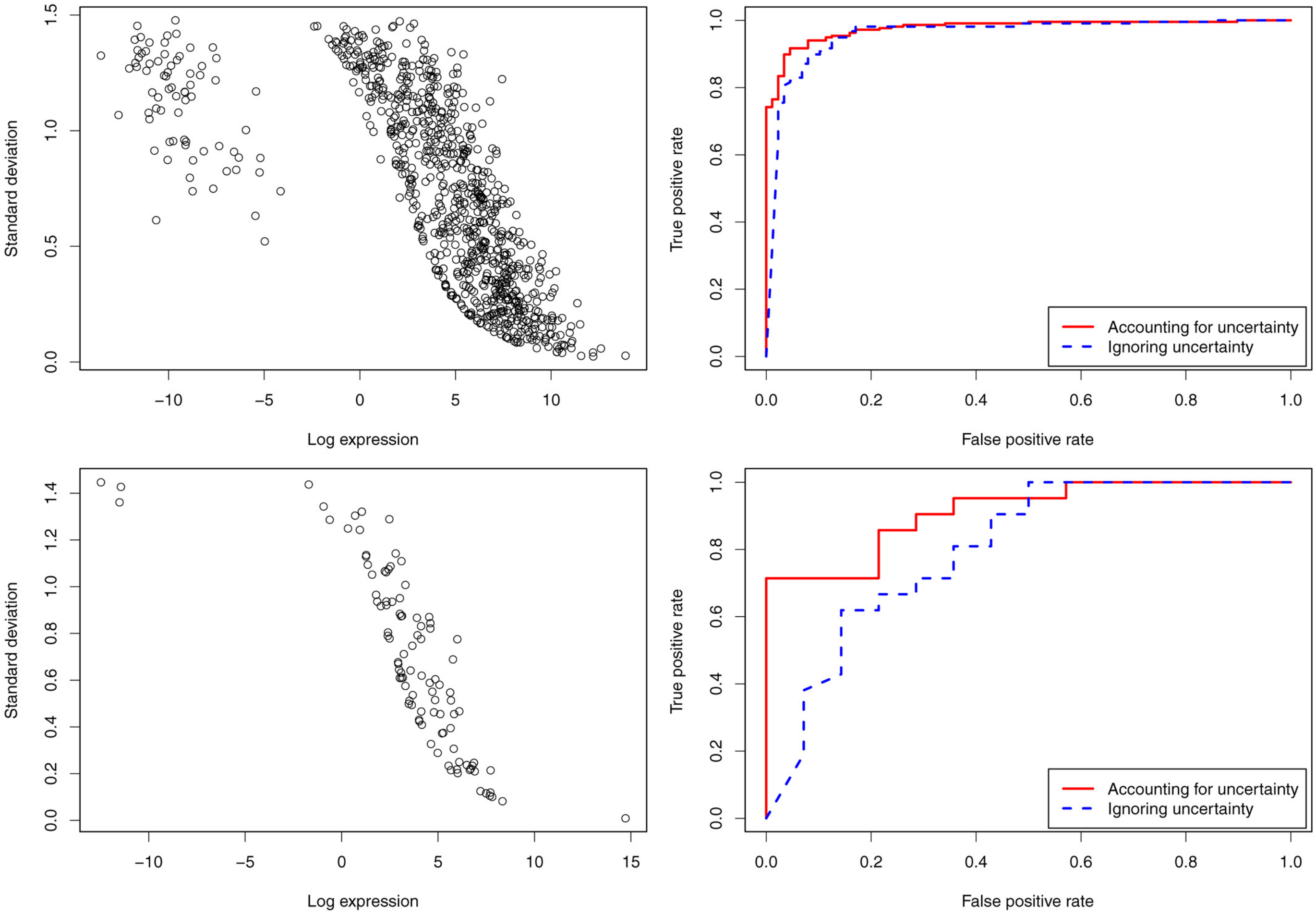

Fig 5. Usefulness of measurement error obtained by PGseq. The scatter plots (left column) show the standard deviation vs. the logarithm of gene expression for the PCR-validated genes in the MAQC (upper panel) and HCC (lower panel) datasets. As the expression level increases the associated measurement error decreases. The ROC curves (right column) indicate the difference between accounting for and ignoring measurement uncertainty in the DE analysis. The DE analysis employs MMDiff which considers expression measurement error. The solid curves show the performance of the DE analysis considering expression measurement error, and the dashed lines ignore measurement error by setting zero measurement error.

doi:10.1371/journal.pone.0140032.g005

combined methods to this dataset for DE detection. Since these methods use the different statistics for significance test, the significant levels are not comparable. Hence, for each method we select the top 2,000 genes in the significance ranking of differential expression. Fig 7 shows the Venn diagram of the significant DE genes found by the four approaches. It can be seen that there are quite a number of common DE genes found by any pair of these methods. We found 729 genes which are declared DE by all the four methods. We then plot the scatter plot of the average logged RPKM estimation as shown in Fig 8. We find that the majority of the 729 common DE genes distribute over the medium and high expression areas and few are found in the lower end. It shows the obvious difficulty of detecting DE for lowly expressed genes.

We have demonstrated in previous section that PGseq obtains more accurate expression estimation for lowly expressed genes. Here, we use the HBD dataset to show the power of our approach in the DE detection for the low expression. We filter the lowly expressed genes with expression between 0.01 and 2.0 (measured in RPKM) to obtain 10,157 low expression genes. We apply the four combined methods to these genes for the DE detection. For each method we find a set of 2,000 most significant DE genes. The union of these four sets contains 4,373 genes,

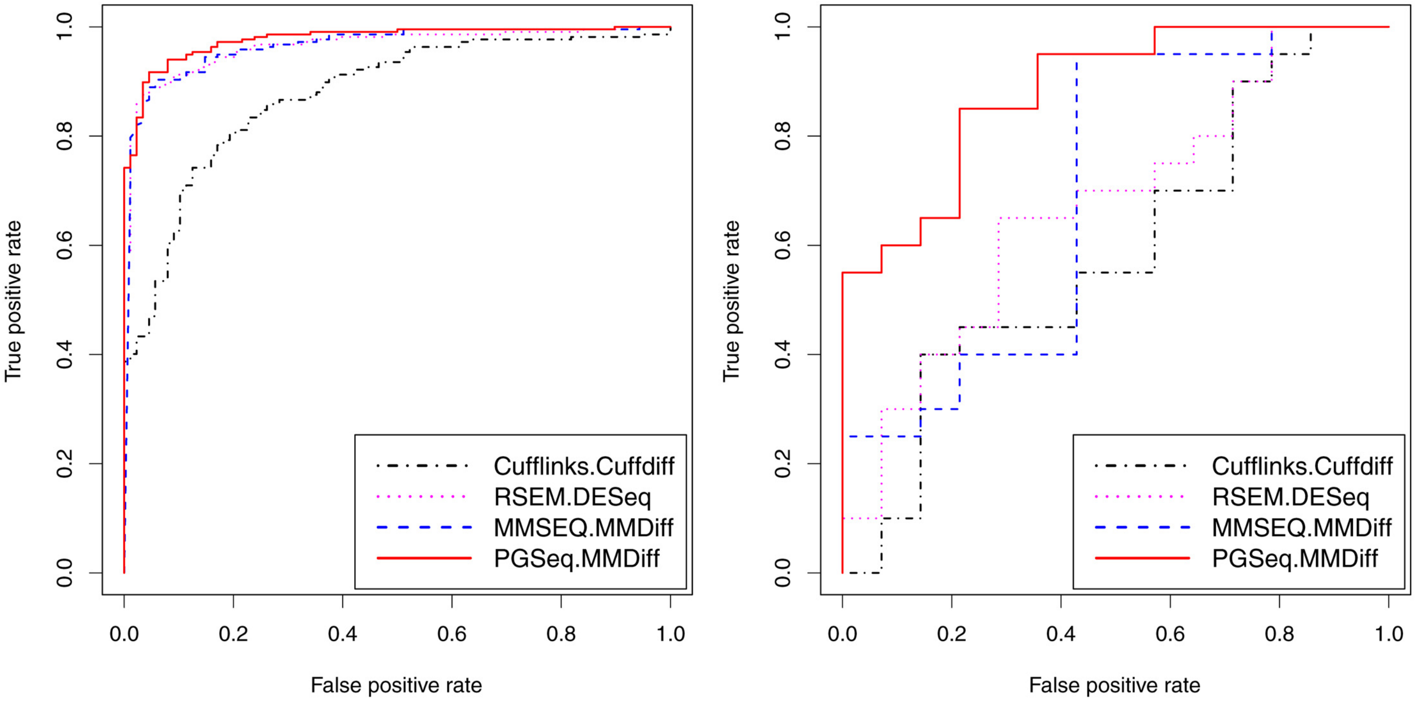

Fig 6. ROC curves of DE analysis for the selected PCR-validated genes in the MAQC (left) and HCC (right) datasets. Cufflinks and MMSEQ are combined with the corresponding embedded DE analysis methods, Cuffdiff and MMDiff, respectively. RSEM is combined with DESeq. PGseq is combined with MMDiff for propagating measurement error in the DE analysis.

doi:10.1371/journal.pone.0140032.g006

each of which appear at least once in the four sets. Correspondingly, we find 348 genes in the intersection which is the overlap of these four sets. We take these 348 genes as the significant DE genes since all of the four approaches find them as significant DE genes, and the rest among the 4,373 genes as the non-DE genes. We use this data to draw ROC curve for each method to show its power in finding "true" lowly expressed DE genes. The higher the curve, the more powerful the related method in finding DE genes in low expression area. Fig 9 shows the ROC curves for the four DE approaches. We can see that PGseq combined with MMDiff presents the highest power in the DE detection. This example shows again that our approach has the advantage in the analysis of low expression genes over other alternatives.

In addition, we leave one method out and find the "true DE set", validated by three methods each time, among the 4,373 genes. For the four sets of the "true DE genes", we perform the comparisons for all the four approaches respectively and show the results in S1 Fig. It can be seen that our approach shows competitive strength in plots (a) $\sim$ (c). For plot (d) where the DE genes are agreed by the three alternative methods except PGseq, our method fails to find many "true positives". When we examine the distribution of the "true DE genes" for each plot

Table 4. Area under ROC curves for detection of DE genes.

Dataset	Cufflinks (Cuffdiff)	RSEM (DESeq)	MMSEQ (MMDiff)	PGseq (MMDiff)
MAQC	0.876	0.965	0.965	**0.977**
HCC	0.422	0.725	0.757	**0.912**

Cufflinks, RSEM and MMSEQ are combined with the corresponding embedded DE analysis methods, Cuffdiff, DESeq and MMDiff, respectively. PGseq is combined with MMDiff for propagating measurement error in the DE analysis.

doi:10.1371/journal.pone.0140032.t004

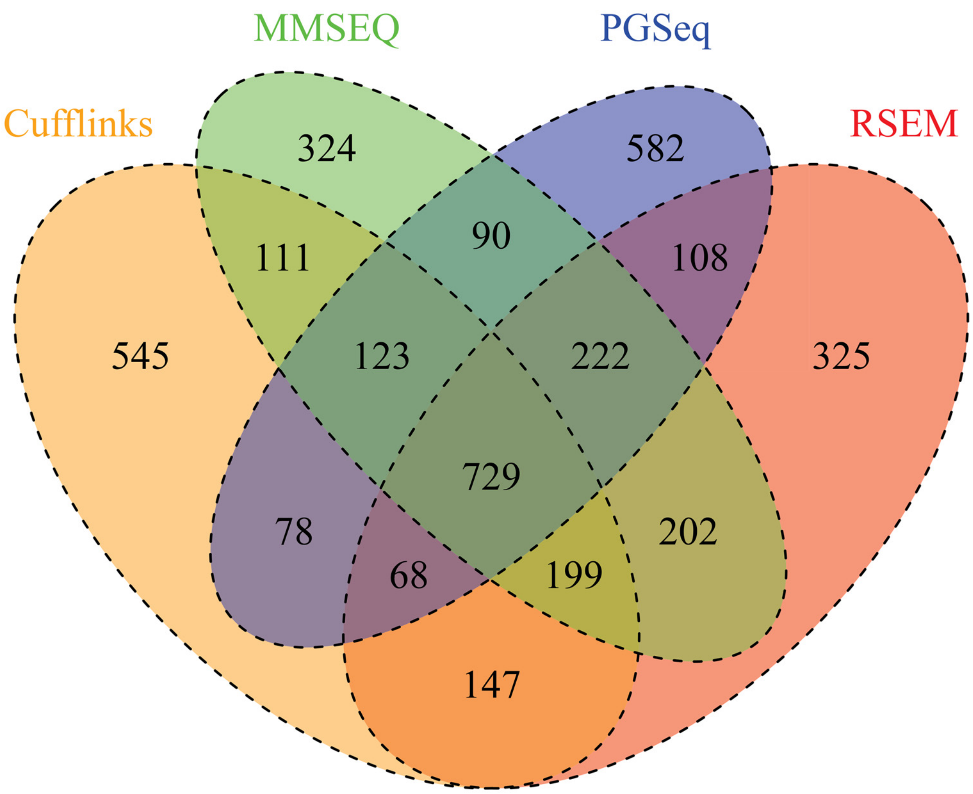

Fig 7. Venn diagram of the significant DE genes for the HDB dataset. The big ovals represent the number of the significant DE genes found by the four methods: Cufflinks, MMSEQ, PGseq and RSEM, which combined with CuffDiff, MMDiff, MMDiff and DESeq, respectively. The overlap of the four ovals in the middle of the diagram is 729, which is the number of the DE genes found by all of the four approaches.

doi:10.1371/journal.pone.0140032.g007

as shown in S2 Fig, we find that although plot (d) contains more "DE genes" than other plots but there are many genes locating close to the diagonal and these genes are likely false positives. In contrast, there are few false positives on the other three plots.

Discussion

In this manuscript, we proposed a Poisson model to fit the read counts for each gene and use the gene-specific Gamma-distributed latent variables to capture the variability of the read sequencing preference for every exon. The bias property modeled in our method is shared across all conditions for each individual gene, and automatically captures all the intrinsic exon-specific effects. We used four real datasets and one simualted dataset to verify the performance of our method and compared it with other popular alternatives, Cufflinks, RSEM and MMSEQ. For the real datasets, we calculated the R^2 correlation coefficients of the estimated gene and transcript expression with the PCR measurements, and performed DE analysis to show the advantages of our approaches. For the simulated dataset, the consistence with the ground truth was also compared. The comparison results have shown that the proposed PGseq approach obtains competitive results for most comparison cases and performs especially better for lowly expressed genes.

Our work indicates that the non-uniformity of read distribution is one of the most important characteristics of RNA-seq data and appropriately modeling the sequencing bias can remarkably improve the accuracy of the expression calculation. We merged all possible biases

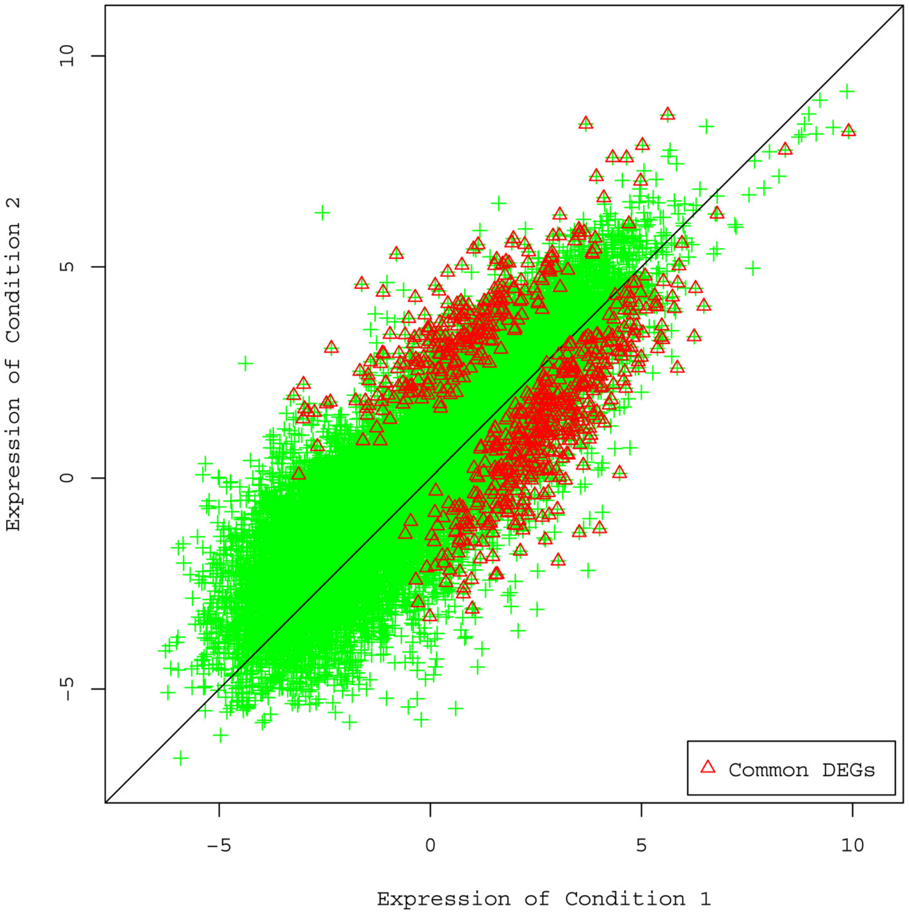

Fig 8. Scatter plot of the average logged RPKM estimation for the HDB dataset. There are 23,402 expressed genes (RPKM>0 for both conditions), among which the 729 common DE genes found by all the four DE methods are represented by red triangles and others by green crosses.

doi:10.1371/journal.pone.0140032.g008

in the read sequencing into an exon-specific random variable and did not look into any specific sequence content around any specific position as many methods did. A Gamma prior was put on this variable and it was integrated out in the ML estimation. Therefore, all possible values of this variable were considered in the bias correction. This is distinct from many other methods which explicitly calculate the point estimate for each of gene- or isoform-specific biases. Our approach seems desirable based on the comparison results in this manuscript.

Another advantage of our method is that PGseq is able to provide a level of uncertainty associated with the gene and transcript expression estimates. This level of uncertainty can be propagated to the downstream analysis and obtain improved analysis results. We combined our method with a recently proposed DE analysis method, MMDiff, which incorporates measurement error of expression estimates to improve DE analysis. We evaluated this approach using two real PCR-validated datasets for DE analysis. The obtained ROC curves showed that our method significantly outperforms other popular combinations for finding differentially expressed genes. This demonstrates the usefulness of the measurement error provided by our method in downstream analysis.

We assumed that the exon-specific read variation pattern is conserved across multiple samples. We therefore shared the bias distribution across all samples to capture this pattern. In the

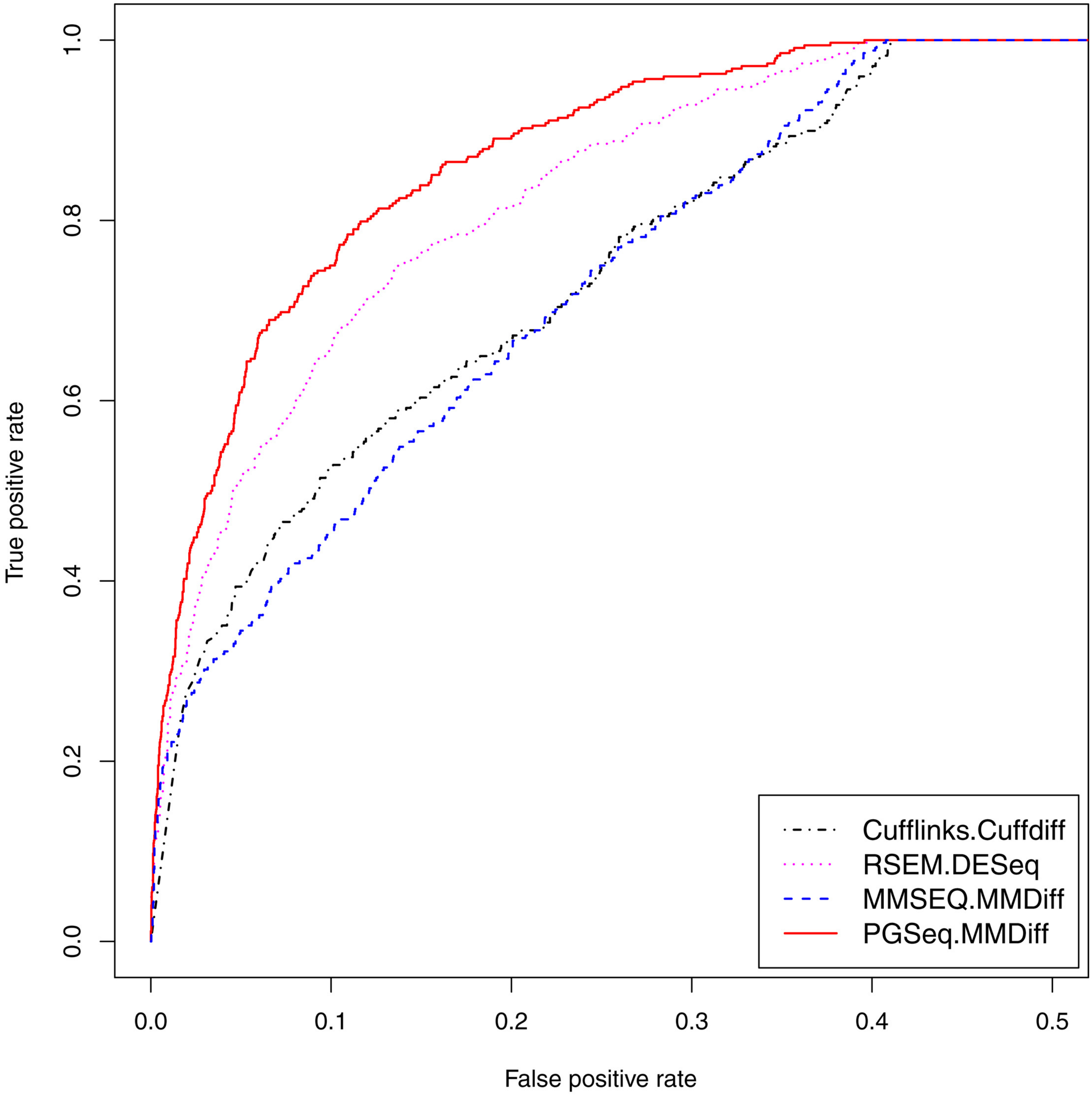

Fig 9. ROC curves of DE analysis for the lowly expressed genes in the HBD datasets. The 348 DE genes found by all the four methods are taken as the "true" DE genes and the rest as non-DE genes.

doi:10.1371/journal.pone.0140032.g009

application to the datasets in this manuscript, our method processed all samples in a single run. In case significant biological variance violates this assumption, the model can be applied sample by sample to estimate the sample-specific bias distribution. For each run, it would be helpful to model estimation if considering as much replicate information as possible.

Practically, biological replicates are preferred to be considered. If biological replicates are not available or the individual sample-specific bias is of interest, the multiple lane information for a single library can also be considered. Finally, we mainly applied our method to data from human genome which have a relatively large number of splicing, we thus modeled the bias variation for each exonic position in order to have an appropriate population size for estimating bias distribution. For simpler genome, such as yeast, where many genes do not have lots of exons, the positions of interest along the reference sequence are not necessary exonic. They can be any sub-sequences of short length. In that case, a proper segmentation of the reference sequence would be needed.

Supporting Information

S1 Fig. ROC curves from the four approaches for datasets validated by three methods. The datasets are validated by the three methods except (a) Cufflinks, (b) RSEM, (c) MMSEQ and (d) PGSeq.
(PDF)

S2 Fig. Scatter plots of the averaged logged RPKM estimation for lowly expressed genes in HDB dataset. The common DE genes (represented by red triangles) are agreed by three methods except (a) Cufflinks, (b) RSEM, (c) MMSEQ and (d) PGSeq, respectively. Others are represented by green crosses.
(PDF)

Acknowledgments

The authors thank M. Rattray for helpful discussions and Michele Busby for valuable suggestions, and support from NSFC (61170152), Qing Lan Project and the Fundamental Research Funds for the Central University (CXZZ11_0217).

Author Contributions

Conceived and designed the experiments: XL. Performed the experiments: LZ. Analyzed the data: XL LZ SC. Contributed reagents/materials/analysis tools: XL LZ. Wrote the paper: XL. Developed the software: LZ.

References

1. Pan Q., Shai O., Lee L.J., Frey B.J., Blencowe B.J (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40, 1414–1415.
2. Matlin A.J., Clark F., Smith C.W.J (2005) Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6, 386–398. doi: 10.1038/nrm1645 PMID: 15956978
3. Wang Z., Gerstein M., Snyder M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10, 57–63. doi: 10.1038/nrg2484 PMID: 19015660
4. Garber M., Grabherr M.G., Guttman M., Trapnell C. (2011) Computational methods for transcriptome annotation and quantification using rna-seq. *Nature Methods*, 8, 469–477. doi: 10.1038/nmeth.1613 PMID: 21623353
5. Mortazavi A., Williams B.A., McCue K., Schaeffer L., Wold B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5, 621–628. doi: 10.1038/nmeth.1226 PMID: 18516045
6. Jiang H., Wong W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25, 1026–1032. doi: 10.1093/bioinformatics/btp113 PMID: 19244387
7. Turro E., Su S.Y., Gonçalves Â., Coin L.J., Richardson S., Lewin A. (2011) Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology*, 12, R13. doi: 10.1186/gb-2011-12-2-r13 PMID: 21310039

8. Wu Z., Wang Xi., Zhang X. (2011) Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics*, 27, 502–508. doi: 10.1093/bioinformatics/btq696 PMID: 21169371

9. Li B., Ruotti V., Stewart R.M., Thomson J.A., Dewey C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26, 493–500. doi: 10.1093/bioinformatics/btp692 PMID: 20022975

10. Li B., Dewey C.N. (2010) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323. doi: 10.1186/1471-2105-12-323

11. Glaus P., Honkela A., Rattray M. (2012) Identifying differentially expressed transcripts from RNA-Seq data with biological variation. *Bioinformatics*, 28, 1721–1728. doi: 10.1093/bioinformatics/bts260 PMID: 22563066

12. Katz Y., Wang E.T., Airoldi E.M., Burge C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 12, 1009–1015. doi: 10.1038/nmeth.1528

13. Li W., Jiang T. (2012) Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. *Bioinformtics*, 28, 2914–2921. doi: 10.1093/bioinformatics/bts559

14. Li J., Jiang H., Wong W.H. (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology*, 11, R50. doi: 10.1186/gb-2010-11-5-r50 PMID: 20459815

15. Srivastava S., Chen L. (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research*, 38, e170. doi: 10.1093/nar/gkq670 PMID: 20671027

16. Roberts A., Trapnell C., Donaghey J., Rinn J.L., Pachter L. (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, 12, R22. doi: 10.1186/gb-2011-12-3-r22 PMID: 21410973

17. Jones D.C., Ruzzo W.L., Peng X., Katze M.G. (2012) A new approach to bias correction in RNA-Seq. *Bioinformtics*, 28, 921–928. doi: 10.1093/bioinformatics/bts055

18. Suo C., Calza S., Salim A., Pawitan Y. (2014) Joint estimation of isoform expression and isoform-specific read distribution using multisample RNA-Seq data. *Bioinformtics*, 30, 506–513. doi: 10.1093/bioinformatics/btt704

19. Ji H., Jiang H., Ma W., et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data, *Nature Biotechnology*, 26:1293–1300. doi: 10.1038/nbt.1505 PMID: 18978777

20. Li C., Wong W. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science USA*, 98, 31–36. doi: 10.1073/pnas.98.1.31

21. Naef F., Magnasco M.O. (2005) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Physical Review—E. Statistical, Nonlinear, and Soft Matter Physics*, 68, 011906. doi: 10.1103/PhysRevE.68.011906

22. Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U., Speed T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4, 249–264. doi: 10.1093/biostatistics/4.2.249

23. Liu X., Milo M., Lawrence N.D., Rattray M. (2005) A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. *Bioinformtics*, 21, 3637–3644. doi: 10.1093/bioinformatics/bti583

24. Hein A.M., Richardson S., Causton H.C., Ambler G.K., Green P.J. (2005) BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data. *Biostatistics*, 6, 349–373. doi: 10.1093/biostatistics/kxi016

25. Spellucci P. (1998) An SQP method for general nonlinear programs using only equality constrained subproblems. *Mathematical programming*, 82: 413–448. doi: 10.1007/BF01580078

26. Anders S., Huber W. (2010) Differential expression analysis for sequence count data. *Genome Biology*, 11: R106. doi: 10.1186/gb-2010-11-10-r106 PMID: 20979621

27. Robinson M., Oshlack A. (2010) A scaling normaliztion method for differential expression analysis of RNA-seq data. *Genome Biology*, 11: R25. doi: 10.1186/gb-2010-11-3-r25 PMID: 20196867

28. Yu D., Huber W., Vitek O. (2013) Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics*, 29: 1275–1282. doi: 10.1093/bioinformatics/btt143 PMID: 23589650

29. Langmead B., Salzberg S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9: 357–359. doi: 10.1038/nmeth.1923 PMID: 22388286

30. Shi L., et al. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24: 1151–1161. doi: 10.1038/nbt1239 PMID: 16964229

31. Bemmo A., Benovoy D., Kwan T., Gaffney D.J., Jensen R.V., Majewski J. (2008) Gene expression and isoform variation analysis using Affymetrix Exon Arrays. *BMC Genomics*, 9: 529. doi: 10.1186/1471-2164-9-529 PMID: 18990248

32. Bullard J., Purdom E., Hansen K., Dudoit S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11: 94. doi: 10.1186/1471-2105-11-94 PMID: 20167110

33. Rapaport F., Khanin R., Liang Y., Pirun M., Krek A., Zumbo P., Mason C.E., Socci N.D., Betel D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14: R95. doi: 10.1186/gb-2013-14-9-r95 PMID: 24020486

34. Griffith M., Griffith O.L., Mwenifumbo J., et al. (2010) Alternative expression analysis by RNA sequencing, *Nature Methods*, 7:843–847. doi: 10.1038/nmeth.1503 PMID: 20835245

35. Wang E.T., Sandberg R., Luo S., Khrebtukova I., Zhang L., Mayr C., Kingsmore S.F., Schroth G.P., Burge C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456: 470–476. doi: 10.1038/nature07509 PMID: 18978772

36. Kim H., Bi Y., Pal S., Gupta R., Davuluri R.V. (2011) IsoformEx: isoform level gene expression estimation using weighted non-negative least squares from mRNA-Seq data. *BMC Bioinformatics*, 12, 305. doi: 10.1186/1471-2105-12-305 PMID: 21794104

37. Kaminuma E., Kosuge T., Kodama Y., et al. (2011) DDBJ progress report, *Nucleic Acids Research*, 39 (Database issue):D22–D27. doi: 10.1093/nar/gkq1041 PMID: 21062814

38. Busby M.A., Stewart C., Miller C.A., et al. (2013) Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression, *Bioinformatics*, 29:656–657. doi: 10.1093/bioinformatics/btt015 PMID: 23314327

39. Liu X., Milo M., Lawrence N.D., Rattray M. (2006) Probe-level measurement error improves accuracy in detecting differential gene expression, *Bioinformatics*, 22: 2107–2113. doi: 10.1093/bioinformatics/btl361 PMID: 16820429

40. Hein A.M.K., Richardson S. (2006) A powerful method for detecting differentially expressed genes from GeneChip arrays that does not require replicates. *BMC bioinformatics*, 7: 353. doi: 10.1186/1471-2105-7-353 PMID: 16857053

41. Liu X., Lin K.K., Andersen B., Rattray M. (2007) Including probe-level uncertainty in model-based gene expression clustering, *BMC Bioinformatics*, 8:98. doi: 10.1186/1471-2105-8-98 PMID: 17376221

42. Turro E., Astle W.J., Tavaré S. (2013) Flexible analysis of RNA-seq data using mixed effects models. *Bioinformatics*, Epub ahead of print. PMID: 24281695

43. Trapnell C., Hendrickson D.G., Sauvageau M., Goff L., Rinn J.L., Pachter L. (2012) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31: 46–53. doi: 10.1038/nbt.2450 PMID: 23222703

44. Leng N., Dawson J.A., Thomson J.A., Ruotti V., Rissman A.I., Smits B.M.G., Haag J.D., Gould M.N., Stewart R.M., Kendziorski C. (2013) EBSeq: an empirical bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29: 1035–1043. doi: 10.1093/bioinformatics/btt087 PMID: 23428641