Published online 29 July 2010

Nucleic Acids Research, 2010, Vol. 38, No. 17 ***e170***
doi:10.1093/nar/gkq670

A two-parameter generalized Poisson model to improve the analysis of RNA-seq data

Sudeep Srivastava and Liang Chen*

Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA

Received May 3, 2010; Accepted July 15, 2010

ABSTRACT

Deep sequencing of RNAs (RNA-seq) has been a useful tool to characterize and quantify transcriptomes. However, there are significant challenges in the analysis of RNA-seq data, such as how to separate signals from sequencing bias and how to perform reasonable normalization. Here, we focus on a fundamental question in RNA-seq analysis: the distribution of the position-level read counts. Specifically, we propose a two-parameter generalized Poisson (GP) model to the position-level read counts. We show that the GP model fits the data much better than the traditional Poisson model. Based on the GP model, we can better estimate gene or exon expression, perform a more reasonable normalization across different samples, and improve the identification of differentially expressed genes and the identification of differentially spliced exons. The usefulness of the GP model is demonstrated by applications to multiple RNA-seq data sets.

INTRODUCTION

With the advance of high-throughput sequencing technologies, transcriptomes can be characterized and quantified at an unprecedented resolution. Deep sequencing of RNAs (RNA-seq) has been successfully applied to many organisms (1–5). However, there are still many challenges in analyzing RNA-seq data. In this work, we focus on a basic question in RNA-seq analysis: the distribution of the position-level read count (i.e. the number of sequence reads starting from each position of a gene or an exon). It is usually assumed that the position-level read count follows a Poisson distribution with rate θ. The gene length-normalized read count, which is a popular gene expression estimate, is then the maximum likelihood estimator (MLE) of θ. In addition, Jiang *et al.* (6) modeled the read count as a Poisson variable to estimate isoform expression. However, as we show in this work, a Poisson distribution with rate θ cannot explain the non-uniform distribution of the reads across the same gene or the same exon. A different distribution is in need to better characterize the randomness of the sequence reads.

We propose using a two-parameter generalized Poisson (GP) model for the gene and exon expression estimation. Specifically, we fit a GP model with parameters θ and λ to the position-level read counts across all of the positions of a gene (or an exon). The estimated parameter θ reflects the transcript amount for the gene (or exon) and λ represents the average bias during the sample preparation and sequencing process. Or the estimated θ can be treated as a shrunk value of the mean with the shrinkage factor λ. We found that the GP model can better estimate gene (or exon) expression by separating true signals from sequencing bias.

It has been shown that normalization continues to be a critical component of RNA-seq analysis (7). A few highly expressed genes specific to one sample can make the library-size based normalization approach inappropriate. Our proposed GP model shows that for some highly expressed genes, many of the reads might be caused by sequencing bias. By removing the sequencing bias captured by λ, we can better perform the normalization across different samples. It has also been observed that in differentially expressed gene studies, although the gene-level read counts across replicate lanes can be fitted by a Poisson distribution, the fit is inappropriate when applied to biological replicates or different biological samples (1). In addition, the parameter estimation for the biological effect may be unreliable given the small number of replicate lanes. On the other hand, based on our GP model, we can better identify differentially expressed genes and differentially spliced exons through log-likelihood ratio approaches. These conclusions are demonstrated by applying our model to multiple RNA-seq data sets in multiple organisms.

*To whom correspondence should be addressed. Tel: +1 213 740 2143; Fax: +1 213 740 8631; Email: liang.chen@usc.edu

MATERIALS AND METHODS

Two-parameter generalized Poisson model for RNA-seq data

For each gene, let X represent the number of mapped reads starting from an exonic position of the gene. The observed counts are $\{x_1, \ldots, x_n\}$ where n is the total number of non-redundant exonic positions (or gene length). The sum of x_i's is equal to the total number of reads mapped to this gene. We assume that X follows a GP distribution with parameters θ and λ:

$$\Pr(X = x) = \begin{cases} \theta(\theta+x\lambda)^{x-1}e^{-\theta-x\lambda}/x!, & x = 0,\ 1,\ 2, ..., \\ 0 & \text{for } x > q \text{ if } \lambda < 0, \end{cases}$$

where $\theta > 0$, $\max(-1, -\theta/q) \le \lambda \le 1$, and q (≥ 4) is the largest positive integer for which $\theta + q\lambda > 0$ when $\lambda < 0$. The lower limits on λ and $q \ge 4$ are imposed to ensure that there are at least five classes with non-zero probabilities and the truncation errors [i.e. $\sum_{x=0}^{\infty} \Pr(X = x)$ is a little <1] do not affect practical applications. When $\lambda = 0$, the GP model reduces to the Poisson model. From our results on the real datasets, more than 99% of the λ estimates were >0. The mean of X is: $\mu = \theta(1-\lambda)^{-1}$, and the variance of X is: $\sigma^2 = \theta(1-\lambda)^{-3}$. The GP distribution was defined and studied by Consul and Jain (8,9). In RNA-seq experiments, the parameter θ can be treated as the transcript amount for the gene and λ represents the bias during the sample preparation and sequencing process. The underlying mechanisms for the sequencing bias remain unknown and need further investigation.

The MLE $\hat{\lambda}$ of λ can be obtained by solving the following equation using the Newton–Raphson method:

$$\sum_{i=1}^{n} \frac{x_i(1-x_i)}{\bar{x}+(x_i-\bar{x})\lambda} - n\bar{x} = 0, \quad \text{where } \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}.$$

The MLE $\hat{\theta}$ of θ can be obtained from: $\hat{\theta} = \bar{x}(1-\hat{\lambda})$. Thus, $\hat{\theta}$ is a shrunk value of the sample mean if $\hat{\lambda} > 0$. This relationship can also be inferred by the equation that $\theta = \mu(1-\lambda)$. Note that there is no MLE if all of the x's are equal to 0 or 1. The same model can be specified for an exon. The observed read counts at the positions are $\{z_1, \ldots, z_m\}$ and m is the exon length.

Normalization issue

To identify differentially expressed genes, we need to perform normalization. The total amount of sequenced RNAs in sample 1 can be estimated by $s_1 = \sum_{g=1}^{G} \hat{\theta}_{1,g} l_g$, where $\hat{\theta}_{1,g}$ is the MLE of θ in the GP model for gene g in sample 1, l_g is the gene length, and G is the total number of genes. Similarly, the total amount of sequenced RNAs in sample 2 can be estimated by $s_2 = \sum_{g=1}^{G} \hat{\theta}_{2,g} l_g$, where $\hat{\theta}_{2,g}$ is the MLE of θ for gene g in sample 2. To perform normalization, we assume that the total amount of RNAs in sample 1 is equal to the total amount of RNAs in sample 2. Therefore, the scaling factor for the comparison between the two samples can be estimated as:

$$w = \frac{s_2}{s_1}.$$

when $\lambda_g = 0$, $\hat{\theta}_g = \bar{x}_g$ which is the MLE for the Poisson model. When $\lambda_g = 0$ for all g's, the scaling factor is the same as that in the normalization procedure using the total number of mapped reads (here only reads mapped to the transcriptome were considered). When $\lambda_g > 0$, $\hat{\theta}_g < \bar{x}_g$ we remove extra reads that are due to the biased sequencing for this gene. When $\lambda_g < 0$, $\hat{\theta}_g > \bar{x}_g$ we compensate read counts for this gene because some of reads cannot be sequenced successfully. In our applications, we found that the majority of genes had a positive λ_g. The ratio of the gene expression between two samples can be estimated as $w\hat{\theta}_{1,g}/\hat{\theta}_{2,g}$.

Methods to identify differentially expressed genes

We used the likelihood ratio test to identify differentially expressed genes. Let X represents the position-level read count in sample 1. Similarly, Y is the random variable for the gene in sample 2. To estimate the unrestricted MLEs, we have:

$$\Pr(X = x) = \theta_1(\theta_1+x\lambda_1)^{x-1}e^{-\theta_1-x\lambda_1}/x!,$$
$$\Pr(Y = y) = \theta_2(\theta_2+y\lambda_2)^{y-1}e^{-\theta_2-y\lambda_2}/y!,$$

where (θ_1, λ_1) and (θ_2, λ_2) can be estimated by the maximum likelihood approach mentioned above independently. When $\lambda_1 < 0$ or $\lambda_2 < 0$ and some x's or y's are larger than the corresponding q values (see the probability mass function of the GP distribution for the meaning of q), the likelihood is zero and the parameter estimation fails. Under the null hypothesis (restricted parameter space), in which the gene is not differentially expressed, we have

$$\Pr(X = x) = \theta(\theta+x\lambda_1)^{x-1}e^{-\theta-x\lambda_1}/x!,$$
$$\Pr(Y = y) = w\theta(w\theta+y\lambda_2)^{y-1}e^{-w\theta-y\lambda_2}/y!,$$

where w is a normalization constant associated with the different sequencing depths for the two samples. We can choose $w = s_2/s_1$, and s_1 and s_2 were calculated based on the unrestricted maximum likelihood model. Through the parameter specification, we preserved the original counts. λ_1 and λ_2 were taken as the same values of the MLEs from the unrestricted maximum likelihood model. Thus, we assumed that the MLE of λ from the unrestricted maximum likelihood model was close to the true value. Then the restricted profile MLE $\hat{\theta}$ can be obtained by solving the equation using the Newton–Raphson method:

$$\frac{2n}{\theta} - (n+nw) + \sum_{i=1}^{n} \frac{(x_i-1)}{\theta+x_i\hat{\lambda}_1} + \sum_{i=1}^{n} \frac{(y_i-1)w}{w\theta+y_i\hat{\lambda}_2} = 0.$$

The log-likelihood ratio test statistic can be calculated as:

$$T = -2\ln\left(\frac{L\left(\hat{\theta}, \hat{\lambda}_1, \hat{\lambda}_2 | \mathbf{x}, \mathbf{y}\right)}{L\left(\hat{\theta}_1, \hat{\theta}_2, \hat{\lambda}_1, \hat{\lambda}_2 | \mathbf{x}, \mathbf{y}\right)}\right).$$

If the null model is true, T is approximately chi-square distributed with one degree-of-freedom.

To perform the comparison, we also used the Poisson model and the log-likelihood ratio approach to identify differentially expressed genes. For the unrestricted Poisson model:

$$\Pr(X=x)=\frac{\theta_1^x e^{-\theta_1}}{x!} \quad \text{and} \quad \Pr(Y=y)=\frac{\theta_2^y e^{-\theta_2}}{y!}.$$

The MLEs are $\hat{\theta}_1=\bar{x}$ and $\hat{\theta}_2=\bar{y}$. For the restricted null model:

$$\Pr(X=x)=\frac{\theta^x e^{-\theta}}{x!} \quad \text{and} \quad \Pr(Y=y)=\frac{(w\theta)^y e^{-w\theta}}{y!},$$

where w can be chosen as $\sum_{g=1}^{G}\bar{y}_g l_g/\sum_{g=1}^{G}\bar{x}_g l_g$. The profile MLE under the null is

$$\hat{\theta}=\frac{\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}y_i}{n+nw}.$$

The log-likelihood ratio test statistic can be calculated as:

$$T=-2\ln\left(\frac{L\left(\hat{\theta}|\mathbf{x},\mathbf{y}\right)}{L\left(\hat{\theta}_1,\hat{\theta}_2|\mathbf{x},\mathbf{y}\right)}\right),$$

and it follows a chi-square distribution with one degree of freedom if the null model is true.

We also used the generalized linear model (GLM) proposed in ref. (1) for the ROC curve study. The log/Poisson link function was used for the linear model:

$$\log\left(E\left(R_{i,j}|d_i\right)\right)=\log d_i+\tau_{a(i),j}+\delta_{i,j},$$

where $R_{i,j}$ is the number of reads mapped to gene j in lane i, $\tau_{a(i),j}$ is the biological effect for gene j in the biological group $a(i)$, $\delta_{i,j}$ is other technical effect (i.e. flow cell effect), d_i is the total number of uniquely mapped reads for lane i and the regression coefficient of $\log d_i$ is set to 1. For the MAQC-2 data, there are two biological groups: UHR and brain samples distributed among 14 lanes. Therefore i ranges from 1 to 14. $a(i)=1$(UHR) for the first seven lanes, and $a(i)=0$ (brain) for the last seven lanes. These 14 lanes were distributed among two flow cells so that $\delta=(1,0,1,0,1,0,1,0,1,0,1,0,1,0)$. The log-likelihood ratio statistic was calculated based on the full and the null model in which the biological effect is zero. Besides the log/Poisson link function, we also applied the negative binomial and the quasi-Poisson link functions in the GLM.

Methods to identify differentially spliced exons

For the differentially spliced exon study, we focused on 'skipped exon' events without considering other alternative splicing events such as alternative 5′ splice sites or alternative 3′ splice sites. In sample 1, the read count for the considered middle exon is Z, and the read count for the corresponding gene is X. In sample 2, the variable for the exon is V, and the variable for the gene is Y. Under the unrestricted parameter space, we can estimate the related parameters for Z, X, V and Y using the method mentioned before. The resultant estimators are: $\hat{\theta}_Z$, $\hat{\lambda}_Z$, $\hat{\theta}_X$, $\hat{\lambda}_X$, $\hat{\theta}_V$, $\hat{\lambda}_V$, $\hat{\theta}_Y$, $\hat{\lambda}_Y$. Under the restricted null model, in which the exon is not differentially skipped or the splicing ratio between the exon expression and the gene expression (b) is the same between the two samples, we have:

$$\Pr(Z=z)=\frac{b\theta_1(b\theta_1+z\lambda_1)^{z-1}e^{-b\theta_1-z\lambda_1}}{z!},$$

$$\Pr(X=x)=\frac{\theta_1(\theta_1+x\lambda_2)^{x-1}e^{-\theta_1-x\lambda_2}}{x!},$$

$$\Pr(V=v)=\frac{b\theta_2(b\theta_2+v\lambda_3)^{v-1}e^{-b\theta_2-v\lambda_3}}{v!},$$

$$\Pr(Y=y)=\frac{\theta_2(\theta_2+y\lambda_4)^{y-1}e^{-\theta_2-y\lambda_4}}{y!},$$

where $b>0$. The λ_1, λ_2, λ_3 and λ_4 were taken as $\hat{\lambda}_Z$, $\hat{\lambda}_X$, $\hat{\lambda}_V$ and $\hat{\lambda}_Y$, respectively. The profile MLEs of θ_1, θ_2 and b can be obtained by solving:

$$\frac{m+n}{\theta_1}+\sum_{i=1}^{m}(z_i-1)\frac{b}{b\theta_1+z_i\hat{\lambda}_Z}+\sum_{i=1}^{n}(x_i-1)\frac{1}{\theta_1+x_i\hat{\lambda}_X}-(mb+n)=0,$$

$$\frac{m+n}{\theta_2}+\sum_{i=1}^{m}(v_i-1)\frac{b}{b\theta_2+v_i\hat{\lambda}_V}+\sum_{i=1}^{n}(y_i-1)\frac{1}{\theta_2+y_i\hat{\lambda}_Y}-(mb+n)=0,$$

$$\sum_{i=1}^{m}(z_i-1)\frac{\theta_1}{b\theta_1+z_i\hat{\lambda}_Z}+\sum_{i=1}^{m}(v_i-1)\frac{\theta_2}{b\theta_2+v_i\hat{\lambda}_V}-m(\theta_1+\theta_2)+\frac{2m}{b}=0.$$

The log-likelihood ratio test-statistic can be calculated as:

$$T=-2\ln\left(\frac{L\left(\hat{b},\hat{\theta}_1,\hat{\theta}_2,\hat{\lambda}_Z,\hat{\lambda}_X,\hat{\lambda}_V,\hat{\lambda}_Y|\mathbf{z},\mathbf{x},\mathbf{v},\mathbf{y}\right)}{L\left(\hat{\theta}_Z,\hat{\lambda}_Z,\hat{\theta}_X,\hat{\lambda}_X,\hat{\theta}_V,\hat{\lambda}_V,\hat{\theta}_Y,\hat{\lambda}_Y|\mathbf{z},\mathbf{x},\mathbf{v},\mathbf{y}\right)}\right).$$

It follows a chi-square distribution with the degree of freedom = 1 if the null model is true.

For the Poisson model, the MLEs of the Poisson parameters are those sample means: $\hat{\theta}_Z=\bar{z}$, $\hat{\theta}_X=\bar{x}$, $\hat{\theta}_V=\bar{v}$, $\hat{\theta}_Y=\bar{y}$. Under the null model:

$$\Pr(Z=z)=\frac{(b\theta_1)^z e^{-b\theta_1}}{z!},\ \Pr(X=x)=\frac{\theta_1^x e^{-\theta_1}}{x!},$$

$$\Pr(V=v)=\frac{(b\theta_2)^v e^{-b\theta_2}}{v!},\ \Pr(Y=y)=\frac{\theta_2^y e^{-\theta_2}}{y!}.$$

The restricted profile MLEs are:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{m} z_i + \sum_{i=1}^{n} x_i}{n+mb}, \hat{\theta}_2 = \frac{\sum_{i=1}^{m} v_i + \sum_{i=1}^{n} y_i}{n+mb},$$
$$\hat{b} = \frac{n\left(\sum_{i=1}^{m} z_i + \sum_{i=1}^{m} v_i\right)}{m\left(\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i\right)}.$$

$T = -2\ln(L(\hat{b}, \hat{\theta}_1, \hat{\theta}_2 | \mathbf{z}, \mathbf{x}, \mathbf{v}, \mathbf{y}) / L(\hat{\theta}_Z, \hat{\theta}_X, \hat{\theta}_V, \hat{\theta}_Y | \mathbf{z}, \mathbf{x}, \mathbf{v}, \mathbf{y}))$ follows a chi-square distribution with one degree of freedom if the null model is true.

Simulation strategy to calculate *P*-values

In the likelihood ratio tests to identify differentially expressed genes or differentially spliced exons, we treated the λ estimates from the unrestricted model as true values and put them into the profile likelihood calculation for simplicity. The derived test statistics T may not exactly follow a chi-square distribution with one degree of freedom under the null. To better estimate the *P*-values, we adopted a simulation strategy. For each nucleotide position, we randomly assigned the mapped reads to the two considered conditions. The simulated sample pair should have no differentially expressed genes or differentially spliced exons. We repeated the simulations 1000 times. For each simulated sample pair i, we calculated the test statistics T_{gi} for each gene g. The distribution of T_{gi}'s ($I = 1, \ldots, 1000$) is the null distribution of the test statistic for gene g. To obtain an accurate *P*-value for the observed test statistic, we fitted the distribution of T_{gi}'s by a Gamma distribution and calculated the *P*-value based on the Gamma distribution. Remember that if Q follows a chi-square distribution with degrees of freedom υ and c is a positive constant, then $c \times Q$ follows a gamma distribution with parameter $\upsilon/2$ and $2c$. In our considered datasets, the null distribution can be well fitted by a Gamma distribution for more than 84% of genes (the *P*-value based on the Kolmogorov–Smirnov test >0.05).

Real data sets

The real data sets we used included: (i) The MAQC data including the MAQC-2 and the MAQC-3 data from ref. (1). The MAQC-2 data were for two biological samples (human UHR and brain samples) and each sample had seven lanes distributed across two flow-cells. The MAQC-3 data were for the UHR sample with four different library preparations. (ii) The human data were for nine human tissues and five breast cancer cell lines (5). (iii) The mouse data were for three mouse tissues, each with two replicates (3). The controlled hydrolysis of RNA samples was performed before cDNA synthesis. (iv) The yeast data were for oligo(dT)-primed and random-hexamer-primed cDNA samples, each with original, biological and technical replicates (4). All of these data were obtained from the Illumina sequencing platforms. These samples had multiple lanes distributed in one flow-cell or across multiple flow-cells.

For the mouse data, the uniquely mapped reads including body reads and junction reads were downloaded from http://woldlab.caltech.edu/rnaseq/. Other data were downloaded from the Sequence Read Archive http://www.ncbi.nlm.nih.gov/sra/ as the .fastq files: SRA010153 for the MAQC data, SRP000727 for the human data (the two low-coverage MAQC samples were excluded), SRX000559-SRX000564 for the yeast data. The sequence reads were mapped to the human genome (NCBI 37.1 or hg19, downloaded from the NCBI website) or the yeast genome (downloaded from the Saccharomyces Genome Database http://www.yeastgenome.org/). The unmapped reads were further mapped against the human refseq RNA sequences (downloaded from the NCBI website) or the yeast ORF sequences (downloaded from the Saccharomyces Genome Database). The resultant reads were treated as junction reads. Note that we only used the uniquely mapped reads and removed the multi-reads. The mapping was performed by using Bowtie, version 0.12.1 and the default settings (10). The Refseq gene annotations for human and mouse were downloaded from the NCBI website and the ORF gene annotations for yeast were downloaded from the Saccharomyces Genome Database. We only considered genes on autosomes or X chromosomes. The position-level read count was the number of body or junction reads starting from an exonic position of a gene (or an exon) without considering the strand information because the sample preparation did not consider strands. A non-redundant exon list was assembled from the Refseq gene annotations or the ORF gene annotations. If two exons were overlapped, they were broken into three 'exons': the region specific to exon 1, the shared region, and the region specific to exon 2. Note that both alternative exons and constitutive exons were included.

To demonstrate that $\hat{\theta}$ of the GP model can better represent gene expression levels, we utilized the QuantiGene data for the MAQC-2 samples (11). The QuantiGene system detects RNA directly without reverse transcription and PCR amplification. The background-corrected signals were averaged across the replicates with detectable expression. If more than two-thirds of the replicates had detectable expression, the gene was included in the regression analysis in Table 2. For each regression, we further required that the gene had an available expression estimate from the considered statistical model. The sample size for these regression models was around 170. To validate our differentially expressed genes, we also utilized the quantitative real-time PCR (qRT-PCR) data for the MAQC-2 samples (12). The DCt values with respect to POLR2A (endogenous control gene) were averaged across the replicates with detectable expression. If $\geq$75% of the replicates had detectable expression for both samples, the gene was claimed to have a reliable log-ratio value in the qRT-PCR data set. To validate our identified differentially spliced exons, we used an enlarged junction list downloaded from http://genes.mit.edu/burgelab/mRNA-Seq/ which included a large number of known and predicted junctions. The liftOver tool from the UCSC genome browser was used to convert the coordinates of the junctions between version hg18 and hg19. We remapped the reads to the enlarged junction lists using BLAST (13), requiring at most two mismatches and at least four matched bases on each side of the junction.

The mapped junction reads can better represent the inclusion rate of a middle exon. The inclusive junction reads were those starting at the 3′ end of the exon (5′ splice site) and those ending at the 5′ end of the exon (3′ splice site). The exclusive junction reads were those skipping the considered middle exon.

Computation efficiency and software package

The codes for the GP model were written in C and the program was computationally efficient. For example, for the MAQC data with a total of six samples, starting from the position-level read counts, it took ~1 min to estimate the gene and exon expression, 2 min to identify differentially expressed genes, and 5 min to identify differentially spliced exons. The memory used was ~800 Mb on a single CPU and the memory can be further decreased if the six samples were run separately. But it took more time if we used the simulation strategy to calculate the *P*-values for the identification of differentially expressed genes and differentially spliced exons. We also prepared an R-package 'GPseq' to implement the methods proposed here. These can be downloaded at: http://www-rcf.usc.edu/~liangche/software.html.

RESULTS

Gene and exon expression estimation

We fitted a GP model to the position-level counts of a gene or an exon. The parameters θ and λ were estimated using the maximum likelihood approach. We first examined the lane and flow-cell effect on the bias parameter λ using the MAQC-2 data from ref. (1). In the MAQC-2 data, each of the two biological samples (UHR and brain) was sequenced in seven lanes distributed across two flow-cells. We estimated λ_g ($g = 1, \ldots, G$; G is the total number of genes) for each lane separately (the goodness-of-fit will be further discussed in Table 1). The correlation of the λ values between the seven lanes for the same sample was around 0.96–0.98. Because the seven lanes were distributed across two flow-cells, the high correlations also indicated that the flow-cell effect on λ was negligible. The λ values were plotted in Supplementary Figure S1 to further show that they were similar across replicate lanes. However, the correlation between the lanes for different biological samples was only around 0.61–0.63, which indicated that the sequencing bias was biological sample dependent. Note that if two independent variables X and Y follow the GP distribution with parameters (θ_1, λ) and (θ_2, λ), the sum of X and Y is also a GP variable with parameters ($\theta_1 + \theta_2$, λ) (14). This justifies the pooling of read counts across lanes that are believed to have similar sequencing bias. We therefore pooled the read counts across lanes for the same sample in the following studies. In addition, we should notice that the λ estimate was not robust if a small number of lanes were used for genes with low expression levels (Supplementary Figure S2). We further examined the library preparation effect on the sequencing bias λ. In the MAQC-3 data (1), four different library preparations were used for the UHR sample. The correlation of the λ values between the library preparations was around 0.97–0.98, indicating that the library preparation effect was also negligible. In the yeast data, cDNAs were prepared by either random hexamers or oligo(dT) primers. The correlation of the λs between the two library preparations was 0.92.

Table 1. The percentages of genes and exons with the position-level count data fitted well by the GP model or the Poisson model

	Gene level		Exon level	
	GP (%)	Poisson (%)	GP (%)	Poisson (%)
MAQC data	85.72	1.57	89.62	19.71
Human data	77.28	3.22	88.78	28.35
Mouse data	88.57	7.88	91.73	39.67
Yeast data	93.24	20.49	93.21	23.73
MAQC-2_sep	92.93	10.18	92.90	41.51

If the *P*-value for the chi-square test >0.05, we declared that the model fitted the data well. The MAQC data contained both the MAQC-2 and the MAQC-3 data. The MAQC-2 data were for two biological samples (human UHR and brain samples). The MAQC-3 data were for the UHR sample with four different library preparations. The human data were for nine human tissues and five breast cancer cell lines. The mouse data were for three mouse tissues, each with two replicates. The yeast data were for oligo (dT)-primed and random-hexamer-primed cDNA samples, each with original, biological and technical replicates. Note that the majority of yeast genes were annotated based on ORFs and only a few of genes had introns. The reads from multiple replicate lanes were pooled together. But for the MAQC-2_sep data, reads from the seven replicate lanes for each of the MAQC-2 samples were fitted separately.

The goodness-of-fit of the GP model was examined by the chi-square test. If the *P*-value for the chi-square test-statistic was >0.05, we declared that the model fitted the data well. Table 1 lists the percentages of genes (or exons) with the position-level count data fitted well by the GP model or by the Poisson model. The position-level counts can be fitted well by the GP model for the majority of genes or exons (77–93%). However, the Poisson model only fitted well for 2–42% of the genes or exons. The *P*-values of the genes that were declared to be well fitted by the GP model were uniformly distributed between 0.05 and 1 (Supplementary Figure S3A). The uniformity was slightly worse for exons, but still much better than that of the Poisson model (Supplementary Figure S3B). For the genes or exons that cannot be fitted by the GP model, 93–99% of them cannot be fitted by the Poisson model either, and the fit of the Poisson model was even worse than the fit of the GP model. The expression levels of the un-fitted genes and exons were generally higher than those fitted well by the GP model (Supplementary Figure S4). It indicated that there was additional bias not captured by λ for the highly-expressed genes.

Based on the definition of the GP distribution, the expectation is: $\mu = \theta(1-\lambda)^{-1}$ and the variance is: $\sigma^2 = \theta(1-\lambda)^{-3}$. The parameter λ is a measure of the departure from Poissonicity. The variance of the GP distribution is greater than, equal to or less than the expectation according to whether $\lambda > 0$, $\lambda = 0$ or $\lambda < 0$, respectively. Consul (14) stated that the parameter θ is the average rate for the natural Poisson process. The parameter λ is

Table 2. Assessment of gene expression estimation based on the QuantiGene gold standards

	θ estimate from the GP model		Mean estimate from the Poisson model	
	Regression line	R^2	Regression line	R^2
UHR sample	$\log(y) = 4.20 + 0.94 \times \log(x)$	0.62	$\log(y) = 3.38 + 0.72 \times \log(x)$	0.68
Brain sample	$\log(y) = 3.92 + 0.90 \times \log(x)$	0.54	$\log(y) = 3.23 + 0.71 \times \log(x)$	0.57

The QuantiGene system detects RNA directly without reverse transcription and PCR amplification. The bias introduced in the reverse transcription and the PCR amplification steps is therefore avoided. Y represents the averaged detectable signal from the replicates of the QuantiGene data. X represents the θ estimate from the GP model or the mean estimate from the Poisson model. A slope of 1.0 for the regression line indicates that the estimated gene expression is accurate and the estimated fold change between two different genes is the same as that from the QuantiGene data. Thus, $\log(y_1/y_2) = 1.0 \times \log(x_1/x_2)$. The slopes for the θ estimates from the GP model were closer to 1.0 compared with those for the mean estimates from the Poisson model.

the average rate of the effort that the subjects are making to deviate from the process. A positive value of λ indicates that the subjects are making an effort to accelerate the natural process and a negative value of λ denotes an effort to retard the process. The GP distribution can also be interpreted through the quasi-binomial distribution. Considering a random variable X following a quasi-binomial distribution I (QBD-I):

$$\Pr(X = x) = \binom{m}{x} p\,(p+x\varphi)^{x-1}(1-p-x\varphi)^{m-x}.$$

The X represents the number of successes in m trials. The probability of success in any one trial is p and in all other trials is $p + x\varphi$. The probability of success increases or decreases depending on the positive or negative φ value. And the change of the probability is proportional to the number of successes. When $m \to \infty$, $p \to 0$, and $\varphi \to 0$ such that $mp = \theta$ and $m\varphi = \lambda$, the QBD-I approaches the GP model with parameter (θ, λ) (15). For the RNA-seq data, the actual transcript amount can be approximated by a Poisson distribution with parameter θ. The bias introduced in the sample preparation and sequencing process is represented by λ. Therefore, the observed position-level read count follows the GP distribution with parameters (θ, λ). Many other models can lead to the GP distribution. For example, under certain limits, the generalized negative binomial distribution and the generalized Markov–Pólya distribution approaches the GP distribution (14). The GP model has been applied to many biological problems such as the frequency of chemically induced chromosome aberrations in human leukocytes (16).

Furthermore, we used the QuantiGene arrays as gold standards to show that θ can better represent the 'true signal' for gene expression. The QuantiGene system detects RNA directly without reverse transcription and PCR amplification. The bias introduced in the reverse transcription and the PCR amplification steps can be avoided. Therefore, the QuantiGene system provides an accurate gene expression measurement. The QuantiGene data were obtained from ref. (11). A linear regression model was fitted between the log of the QuantiGene signal and the log of the estimated θ value. The slope was 0.94 and 0.90 for the two MAQC-2 samples (Table 2). However, the slope was 0.72 and 0.71 if the Poisson estimate was used. The R-square values were similar for the regression models. A slope different from 1.0 indicates the compression or expansion effects between the estimated expression levels and the gold standard signals. If a slope is equal to 1.0, the estimated fold change between two different genes of the same sample is equal to the gold-standard fold change. As we can see, the slope was closer to 1.0 for the GP θ estimates than the Poisson estimates. It indicates that θ can better represent the 'true signal' and it makes the direct comparison between two different genes reliable.

It has been observed that a few highly expressed genes contributed a large fraction of sequence reads in certain tissue samples (17). Figure 1 shows the fraction of mRNA sample contributed by the highly expressed genes in the human tissues. The mRNA amount derived from a gene was equal to the product of the estimated gene expression multiplied by the gene length: $\bar{x}_g l_g$ for the Poisson model, and $\theta_g l_g$ for the GP model. According to the Poisson model, the top 10 genes contributed to ~15–31% of the total mRNAs in the human heart, muscle and liver tissues (Figure 1A). If we used the GP model, the fraction contributed by highly expressed genes was smaller (Figure 1B). We examined the top four genes with the largest contributions to the total RNAs in the liver tissue based on the Poisson model. The four genes were also the top genes based on the GP model (ranks 8, 1, 12 and 3). Figure 2 shows the observed frequency of the read count equal to k ($k = 0, 1, 2, \ldots$) (blue bars) and the expected frequencies based on the GP model (magenta bars) or the Poisson model (brown bars). We only plotted the 'k' values with at least one frequency among the three types of frequencies ≥ 5 for visual convenience. However, all of the k's had been used in the parameter estimation. The GP model fitted much better than the Poisson model. Note that for these genes, there were ~1.4–3.8% of positions with thousands of reads starting exactly from these positions (i.e. $k \geq 1000$). For each of these 'k' values, the frequency of positions with exactly k number of reads was small (<5) so that they were not plotted in Figure 2. However, these outliers greatly affected the MLE of the Poisson model, which resulted in a peak at a high 'k' value for the expected frequencies (peak of the brown bars), but the observed frequencies

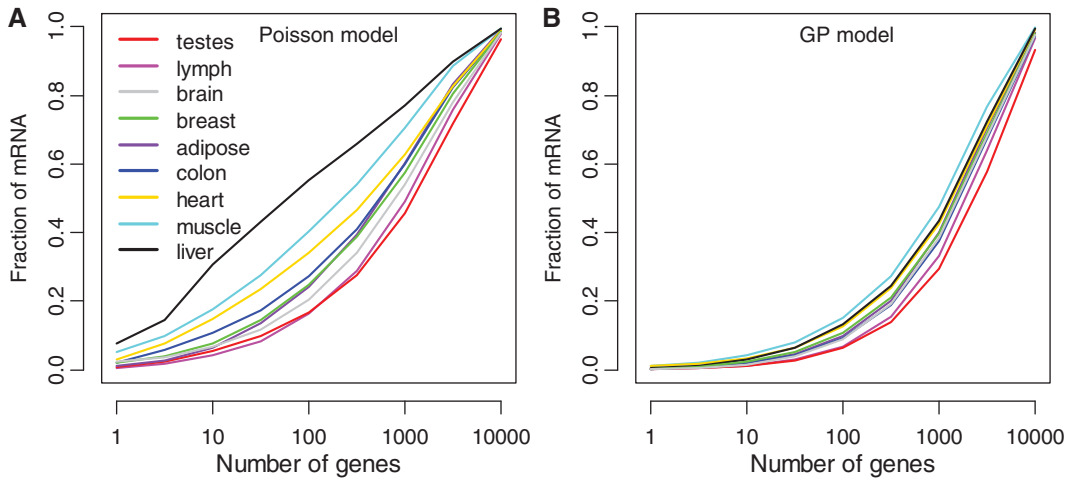

Figure 1. The fraction of total mRNA amount derived from highly expressed genes for the human tissue samples. Genes were ranked based on the product of its estimated expression level and the gene length: $\bar{x}_g l_g$ for the Poisson model (**A**) or $\hat{\theta}_g l_g$ for the GP model (**B**). Then the percentage of mRNA amount contributed from the top 1, 10,..., 10 000 genes was calculated and plotted.

(blue bars) were low in this region. Although the effect of outliers on the GP model was relatively smaller, they still caused the *P*-values of the chi-square tests to be <0.05. It was unrealistic to assume that these outlier reads reflected the gene expression correctly because the majority of the gene positions did not exhibit such high read counts. It was more likely that these reads contained sequencing bias. Therefore, if we counted all the reads mapped to the gene to estimate the gene expression, we counted extra reads due to sequencing bias. On the other hand, the GP model can better separate the bias from the true signal. The frequencies of the k values beyond the range of Figure 2 were plotted in Supplementary Figure S5. Note that all the frequencies were <5 and the GP model still fitted better than the Poisson model. These highly expressed genes would affect the normalization across different samples if we used the Poisson estimates, as argued in ref. (7). We will further discuss about it in the 'Normalization' section.

Factors associated with bias λ

As we mentioned above, the bias parameter λ was independent of lanes, flow-cells, and library preparations, but dependent on biological samples. We further explored possible factors related to λ. We first considered the average number of reads mapped to a gene ($\bar{x}_g$). Although the MLE of λ does not have a closed-form expression, the method of moments estimator of λ is: $\tilde{\lambda}_g = 1 - \sqrt{\bar{x}_g / s_g^2}$, where s_g^2 is the sample variance. Therefore we expect a negative correlation between $\hat{\lambda}_g$ and $\sqrt{\bar{x}_g / s_g^2}$. In our real datasets, we found a negative correlation between $\sqrt{\bar{x}_g / s_g^2}$ and $\sqrt{\bar{x}_g}$, the correlation was from around −0.61 to −0.87. Thus, the increase of the variance was faster than the increase of the mean. We therefore expect a positive correlation between $\hat{\lambda}_g$ and $\sqrt{\bar{x}_g}$. Specifically, the Pearson correlation between $\hat{\lambda}_g$ and $\sqrt{\bar{x}_g}$ was around 0.66–0.84 for the human tissue and cell line data, 0.83–0.85 for the yeast data and 0.54–0.75 for the mouse tissue data. To remove the possible effect of different robustness of $\hat{\lambda}$'s for different genes, we focused on genes with converged $\hat{\lambda}_g$ from the MAQC-2 data. We pooled the reads from the seven lanes gradually by adding one lane each time. Once we added one lane, $\hat{\lambda}_g$ was calculated again. If $\hat{\lambda}_g$ from the first six lanes was within ±5% of the final $\hat{\lambda}_g$ from the total seven lanes, the gene was said to have a converged $\hat{\lambda}_g$. The correlation between $\hat{\lambda}_g$ and $\sqrt{\bar{x}_g}$ for these genes was 0.78 for the UHR sample and 0.81 for the brain sample. Because the increase of the sample variance was faster than the increase of the sample mean, there was also a negative correlation between $\sqrt{\bar{x}_g / s_g^2}$ and s_g (the correlation was from around −0.19 to −0.71 for the real datasets). This indicates a positive correlation between s_g and $\hat{\lambda}_g$. Specifically, we found that the Pearson correlation between $\hat{\lambda}_g$ and s_g was around 0.22–0.71 for all the considered datasets. The results about $\bar{x}_g$ and s_g also explained why λ was dependent on biological samples because the expression levels were different.

We next studied the nucleotide composition of each gene. We counted the occurrence of every possible oligonucleotide with length ≤6. Because the library preparations did not consider the strand information, we pooled the oligonucleotides with their reverse complementary counterparts. Table 3 lists the top 10 oligonucleotides with the largest absolute correlations with $\hat{\lambda}_g$. The mean value of $\hat{\lambda}_g$ was used if there were multiple biological samples. For the human tissue and yeast data (oligo(dT) primed or random-hexamer primed), if a gene was A/T enriched, the bias parameter λ generally tended to be smaller. However, the pattern was different for the mouse data in which the hydrolysis of RNA was used before cDNA priming. It has been shown that the controlled RNA hydrolysis step significantly improves the uniformity of sequence coverage (3). The nature of the bias λ could be different for the mouse data. In summary, the bias parameter λ is a combined effect of the expression level, nucleotide composition and experimental procedures such as RNA hydrolysis before cDNA reverse transcription.

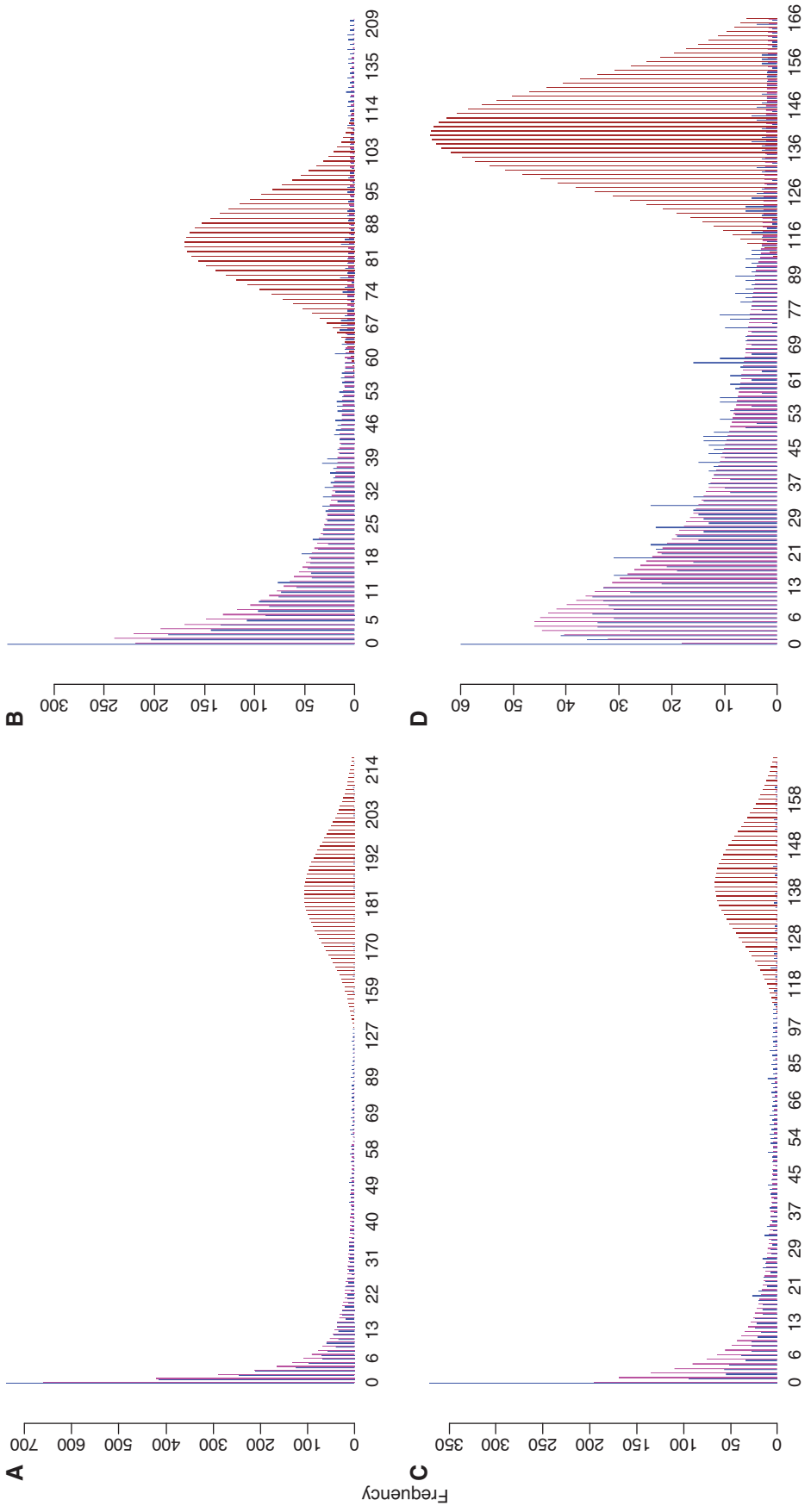

Figure 2. The observed and the expected frequencies of the read count equal to k. (**A–D**) are for the top four highly expressed genes in the liver tissue. Blue bars are for the observed frequencies, brown bars are for the expected frequencies based on the Poisson model, and magenta bars are for the expected frequencies based on the GP model. We only plotted the 'k's with at least one frequency among the three types of frequencies ≥ 5. The total number of reads mapped to the genes ($\tilde{x}_g l_g$) were about 677 476, 329 818, 277 529 and 272 551. And the total number of estimated reads from the GP model ($\theta_g l_g$) were about 6340.2, 11297.0, 4589.7 and 9138.9 for these four genes.

Table 3. Top 10 oligonucleotides with the largest absolute correlations with $\hat{\lambda}$

Human tissue		Yeast_random_hexamer		Yeast_oligo(dT)		Mouse tissue	
Motif	Corr	Motif	Corr	Motif	Corr	Motif	Corr
ACT or AGT	−0.312	ATA or TAT	−0.550	ATA or TAT	−0.543	AG or CT	−0.350
A or T	−0.311	ATAA or TTAT	−0.507	ATAA or TTAT	−0.495	AGAG or CTCT	−0.349
TA	−0.310	AATA or TATT	−0.495	AATA or TATT	−0.488	CAG or CTG	−0.347
TTTC or GAAA	−0.308	TAAA or TTTA	−0.495	TAAA or TTTA	−0.484	CTC or GAG	−0.346
ATTC or GAAT	−0.308	AT	−0.486	AT	−0.474	TC or GA	−0.342
TCA or TGA	−0.308	AAAA or TTTT	−0.481	AAAA or TTTT	−0.469	TCTG or CAGA	−0.342
TAG or CTA	−0.308	ATG or CAT	−0.478	AAAT or ATTT	−0.462	AGC or GCT	−0.341
AT	−0.307	AAAT or ATTT	−0.473	ATG or CAT	−0.462	C or G	−0.340
TTCA or TGAA	−0.306	ATTA or TAAT	−0.471	ATTA or TAAT	−0.458	TCC or GGA	−0.339
ACA or TGT	−0.306	TA	−0.468	TA	−0.458	AGA or TCT	−0.339

Every possible oligonucleotide with length ≤6 was considered. The occurrences of the oligonucleotides and their reverse complementary counterparts were pooled together for each gene. The correlation between the log of the oligonucleotide frequency and $\hat{\lambda}$ was calculated and ranked based on the absolute value. For the human tissue data, the average $\hat{\lambda}$ across the nine tissues was used. For the yeast data, the samples were prepared using random hexamer or oligo(dT) primers. For the mouse data, the average $\hat{\lambda}$ across three tissues were used. The hydrolysis of RNA was performed before cDNA priming for the mouse data.

Normalization

To perform the normalization across different samples, the total RNA amount was always assumed to be equal. Denote the total RNA amount for the kth sample as $s_k = \sum_{g=1}^{G} \hat{\mu}_{k,g} l_g$, where $\hat{\mu}_{k,g}$ is the estimated expression level of gene g in sample k, and l_g is the gene length. To compare sample 1 with 2, the expression level in sample 2 can be adjusted by multiplying s_1/s_2. If we assume that the position-level count across different positions of the gene follows a Poisson distribution, the MLE $\hat{\mu}_{k,g}$ is the average number of reads mapped to gene $(\bar{x}_{k,g})$, and then s_k is the total number of reads mapped to the transcriptome. Thus, this is the standard normalization to the total number of reads. If we assume that the position-level count follows a two-parameter GP distribution, $\hat{\mu}_{k,g} = \hat{\theta}_{k,g}$. We remove the reads due to sequencing bias when we calculate the total RNA amount. Robinson and Oshlack argued that some highly expressed genes specific to one sample will lead to more falsely declared differentially expressed genes if the standard normalization is applied (7). They proposed a weighted trimmed mean approach to remove extreme values when calculate the normalization factor. Our MLE $\hat{\theta}$ for the GP model is equal to $\bar{x}(1-\hat{\lambda})$. It can be treated as a shrunk value of $\bar{x}$, which trims the extreme RNA amount contributed by a single gene in nature. We should note that more than 99% of the $\hat{\lambda}$'s in our applications were >0. In addition, we have $\theta = \mu(1-\lambda)$. Therefore $1-\lambda$ represents how much μ needs to shrink to approach the true gene expression rate θ. Therefore, we treated θ as the parameter representing gene expression and λ as the shrinkage parameter. As shown in Figure 1, the GP expression estimates of highly expressed genes were shrunk toward smaller values. Therefore the effect of a few outlier genes on the normalization factor was relatively small.

Identification of differentially expressed genes

Based on the GP model, the log-likelihood ratio approaches were proposed to identify differentially expressed genes. The scaling normalization factor (s_2/s_1) was used directly in the parameter estimation, while the data itself were not modified and the sampling properties were preserved. To evaluate our model, we used the technical or biological replicates. There should be no differentially expressed genes between technical replicates if the sequencing depth is similar. If the sequencing depth is unbalanced, certain genes are detected in one sample but they are not detected or poorly detected in the other sample, which will still lead to 'differentially expressed genes'. Table 4 lists the number of differentially expressed genes declared from the comparison of these replicates. The false discovery rate was controlled at 0.0001. In general, the GP model declared less differentially expressed genes. For the MAQC-3 data, they were technical replicates of the UHR sample with different library preparations, and the sequencing depth was comparable between the replicates. The GP model identified no differentially expressed genes. However, the Poisson model still identified as many as 120 differentially expressed genes, which can be treated as false positives. For the mouse data, the sequencing depth was much deeper for the second replicates. Although we performed normalization, both models identified many false positives (left panel). If we randomly selected a subset of the uniquely mapped reads from the original larger sample to make the sequencing depth equal, the GP model identified much less false positives (right panel). However, the Poisson model still identified as many as 1007 false positives. The results indicated that we should design the same number lanes for the two-sample comparison to keep the sequencing depth similar. The results also indicated the need for a more complicated normalization method for two samples with very different sequencing depth (e.g. the ratio of sequencing depth = 0.55). For the technical replicates of the yeast data, although the sequencing depth was very different, the GP model identified no false positives while the Poisson model identified a few false positives. The difference between biological replicates was larger than the difference between technical replicates,

Table 4. The number of differentially expressed genes declared from the comparison of technical or biological replicates

Original data					
	GP	Poisson		GP	Poisson
MAQC-3: library 1 versus library 2 (1.11)	0	47	MAQC-3: library 1 versus library 3 (0.98)	0	6
MAQC-3: library 1 versus library 4 (1.02)	0	120	MAQC-3: library 2 versus library 3 (0.89)	0	43
MAQC-3: library 2 versus library 4 (0.92)	0	7	MAQC-3: library 3 versus library 4 (1.04)	0	118

Original data			Random subset from the original data		
	GP	Poisson		GP	Poisson
Mouse_brain: original versus technical replicates (0.55)	422	390	Mouse_brain: original versus technical replicates (1.0)	2	265
Mouse_liver: original versus technical replicates (0.75)	67	823	Mouse_liver: original versus technical replicates (1.0)	27	720
Mouse_muscle: original versus technical replicates (0.86)	171	1085	Mouse_muscle: original versus technical replicates (1.0)	119	1007
Yeast_ random_hexamers: original versus technical replicates (0.66)	0	8	Yeast_ random_hexamers: original versus technical replicates (1.0)	0	8
Yeast_ oligo_dT: original versus technical replicates (0.64)	0	2	Yeast_ oligo_dT: original versus technical replicates (1.0)	0	1
Yeast_ random_hexamers: original versus biological replicates (0.96)	280	590	Yeast_ random_hexamers: original versus biological replicates (1.0)	275	579
Yeast_ oligo_dT: original versus biological replicates (1.38)	203	571	Yeast_ oligo_dT: original versus biological replicates (1.0)	163	487

The numbers in the brackets were the sequencing depth ratio between the two samples. The sequencing depth was defined as the total number of reads that can be uniquely mapped to the genome or junctions. For the mouse and yeast data, besides the studies on the original data sets, we also randomly selected a subset of the uniquely mapped reads from the original larger sample so that the sequencing depth was the same for these two samples. The false discovery rate was controlled at 0.0001. Only genes with existing MLEs of the parameters for both the GP and the Poisson models in the two samples were considered.

even after we made the sequencing depth the same. In the above analysis, we obtained the *P*-values for the likelihood ratio tests based on the chi-square distribution with one degree of freedom. We also calculated the *P*-values based on the simulation strategy mentioned in 'Materials and Methods' section. The results are listed in Supplementary Table S1. The GP model still identified less number of differentially expressed genes between technical replicates with similar sequencing depth, compared with the Poisson model.

To further validate our methods, we utilized the real-time PCR (qRT-PCR) data for the MAQC UHR and brain samples (12). Genes with qRT-PCR absolute log-ratio <0.2 were treated as negatives and genes with qRT-PCR absolute log-ratio >2.0 were treated as positives. The same criteria were used in ref. (1). Then the log-likelihood ratio tests based on the GP model and the Poisson model were performed for these gold-standard genes, respectively. In addition, we considered the GLM proposed in ref. (1) with the log/Poisson link function, the negative binomial link function, and the quasi-Poisson link function. In the GLM, the replicate lanes were considered separately. The true positive rate and the false positive rate were calculated for each given threshold on the test statistics. Figure 3 shows the ROC curves. It clearly shows that the GP model performs better than the Poisson and the generalized linear models. The ROC curves ignored the genes with absolute log-ratio between 0.2 and 2.0 ('no-call' genes). Ideally, these 'no-call' genes should have the *P*-values between the *P*-values for the

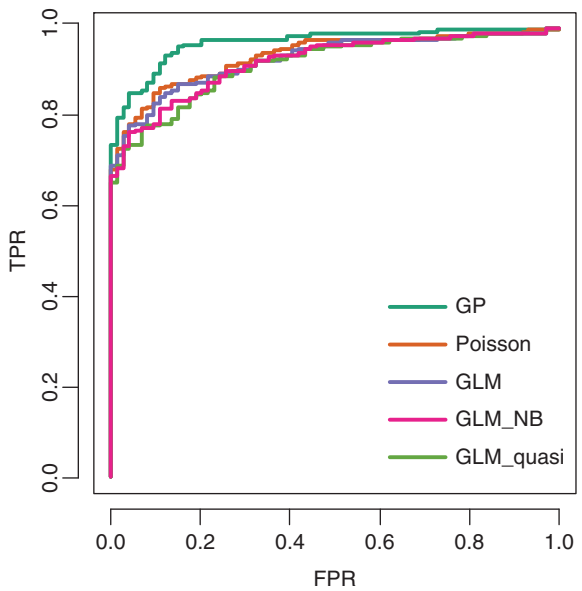

Figure 3. ROC curves for the GP, the Poisson and the GLM (Poisson, negative binomial and quasi-Poisson links) in the identification of differentially expressed genes. Genes with the estimates in the six models and a reliable log-ratio value in the qRT-PCR experiments were considered. We further limited our studies on the standard positives (qRT-PCR absolute log-ratio >2.0, $n = 218$) and the standard negatives (qRT-PCR absolute log-ratio <0.2, $n = 74$). A true positive was required to be differentially expressed in the same direction according to both RNA-seq and qRT-PCR.

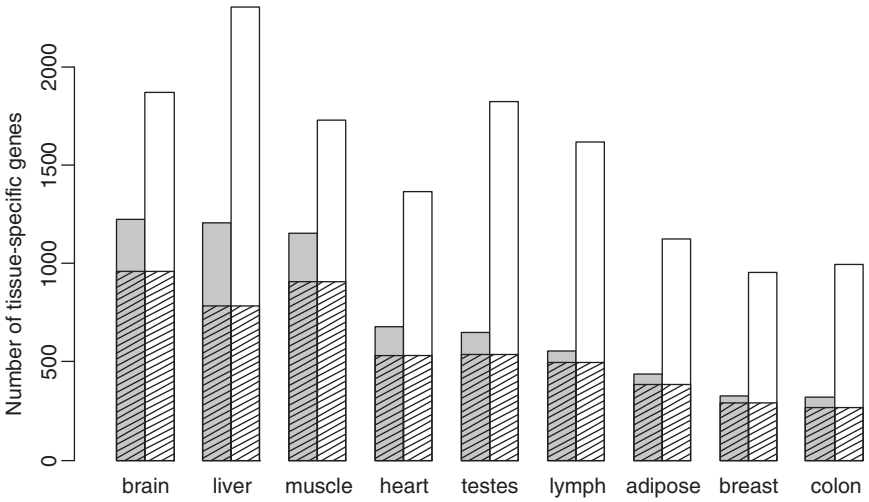

Figure 4. Number of tissue-specific genes. For each human tissue, we counted the number of genes differentially expressed between this tissue and all the other eight tissues. Grey bars are for the GP model and the white bars are for the Poisson model. The shaded regions represent the shared genes between the GP and Poisson models. Only genes with existing MLEs of the parameters for both the GP and the Poisson models in the two compared samples were considered.

positive standards and the *P*-values for the negative standards. For the GP model, 10 'no-call' genes had a *P*-value $\leq$ the median *P*-value of the positive standards (i.e. false positives), and 95 'no-call' genes had a *P*-value $\geq$ the median *P*-value of the negative standards (i.e. false negatives). However, if we used the Poisson model, the number increased to 56 for the false positives and 117 for the false negatives. The number further increased to 58 and 125 for the GLM model. The results were similar if the QuantiGene data were used as gold standards (Supplementary Figure S6). Note that for the MAQC-2 data, there were no extremely high gene expression levels even based on the Poisson model (see Supplementary Figure S7). But we still observed a significant improvement of the GP model over other methods. Thus, the advantages of the GP model are not limited to the normalization issue.

We applied the GP model to the human tissue data and identified the tissue-specific genes. Specifically, for each tissue, we counted the number of genes differentially expressed between this tissue and all the other eight tissues. The differentially expressed genes were determined with FDR threshold 0.0001. Figure 4 shows that the GP model identified much fewer tissue-specific genes than the Poisson model. The brain tissue had the largest number of tissue-specific genes followed by liver and muscle (grey bars). However, if we used the Poisson model, liver had much more tissue-specific genes than brain (2301 versus 1871). Based on the GP model, we identified 119 housekeeping genes with constant expression levels across the nine different tissues (thus, not differentially expressed between any two tissues). Enriched gene annotations included 'ribonucleoprotein complex', 'ribosome biogenesis', 'rRNA processing', 'RNA processing' and so on. These were analyzed by the David gene annotation tool (18). The number of housekeeping genes dropped to 19 if the Poisson model was specified. The pattern was similar if we used the simulation strategy to calculate the *P*-values (Supplementary Figure S8).

Identification of differentially spliced exons

Based on the GP model, we can also identify differentially spliced exons using the log-likelihood ratio tests. We focused on the middle exons and compared the splicing ratio of θ_e/θ_g between two samples. Figure 5A plots the number of differentially spliced exons for each pair of human tissues. It clearly shows that brain was the tissue that had the greatest number of the differentially spliced exons, which was consistent with previous reports using microarrays (19) or EST-based approaches (20). However, if we assumed a Poisson model, many other tissues had a larger number of differentially spliced exons (Figure 5B). We used a more stringent FDR criterion for the Poisson model to make the total number of discoveries approximately equal to that of the GP model. Still, the pattern was different from that of the GP model (Figure 5C). Note that Pan *et al*. (21) used the similar way to present their results by plotting the boxes. In our data preprocessing, we mapped the reads that cannot be mapped to the genome to the refseq RNA sequences, and treated them as junction reads. Both junction reads and body reads were used for the GP model. Here, we remapped the reads to an enlarged junction list from ref. (5) to get a more complete junction read counts. Then we calculated the inclusion rate of a middle exon as: the number of inclusive junction reads/(the number of inclusive junction reads + the number of exclusive junction reads). The absolute differences of the inclusion rates between each pair of tissues were calculated and compared between the differentially spliced exons identified by the GP model and those identified by the Poisson model (Figure 5D). Apparently, the differentially spliced exons identified by the GP model had a larger inclusion rate difference between the two compared

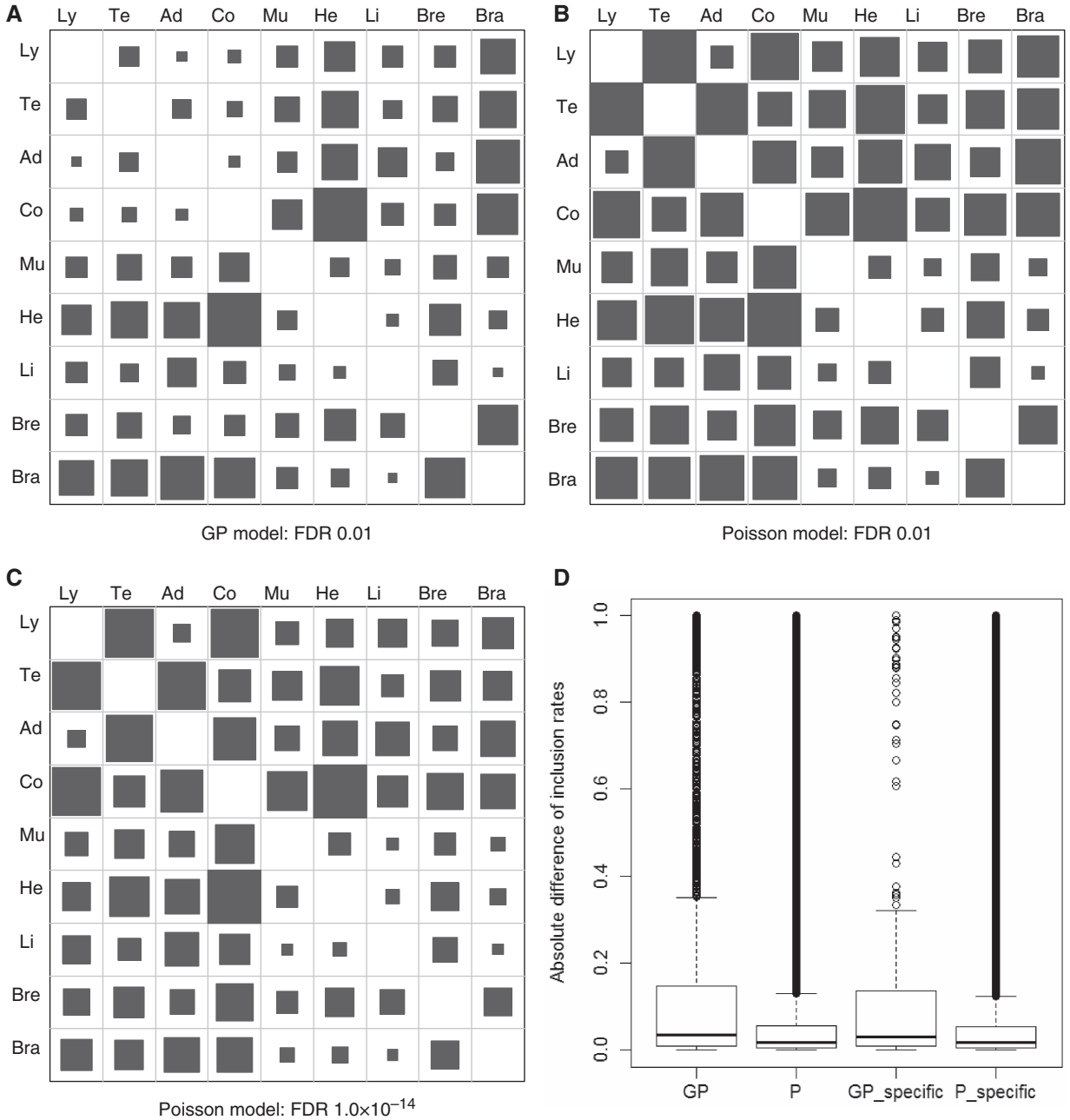

Figure 5. Identification of differentially spliced exons. (**A–C**) The number of differentially spliced exons for each pair of human tissues. Only middle exons with existing MLEs of the parameters for both the GP and the Poisson models in each pair of tissues were considered. (A) is for the GP model with FDR 0.01 for each two-tissue comparison, (B) is for the Poisson model with FDR 0.01, (C) is for the Poisson model with FDR 1.0×10^{-14}. The size of each black box indicates the number of differentially spliced exons. However the size of the boxes between different panels (A–C) is not comparable. The total discoveries for (A–C) are: 5099, 94849 and 5719. Ly, lymph node; Te, testes; Ad, adipose; Co, colon; Mu, muscle; He, heart; Li, liver; Bre, breast; Bra, brain. (**D**) The difference of the inclusion rates calculated from the junction reads for the differentially spliced exons identified by the GP model (GP) or the Poisson model (P). The differentially spliced exons specific to the GP (GP_specific) or the Poisson (P_specific) models are also shown. The FDR was controlled at 0.01.

tissues. For the exons identified only by the GP model, they also had a larger inclusion rate difference than those identified only by the Poisson model. The results indicated that the GP-based approach can better identify differentially spliced exons. The pattern was similar if we used the simulation strategy to calculate the *P*-values (Supplementary Figure S9).

Comparison with other models

The negative binomial distribution has been used to model the additional variation from biological replicates for the digital gene expression (DGE) data (22). In the review article (23), the authors also mentioned that the background tag distribution for the chip-seq data can be

modeled with a negative binomial distribution. In our study, the ROC in Figure 3 shows that the negative binomial performs even worse than the Poisson model. Actually we found that the variance of the gene-level read count was not always larger than the mean across replicates. For example, for the MAQC-2 UHR sample, considering the gene-level read counts across seven replicate lanes, 37% of the genes had a smaller sample variance than the sample mean and 60% of the genes had a larger sample variance than the sample mean. For the MAQC-2 brain sample, 32% of the genes had a smaller sample variance than the sample mean and 64% of the genes had a larger variance. These indicate that for the gene-level read counts, there is both overdispersion and underdispersion. The negative binomial which can only handle overdispersion is inappropriate to handle the gene-level read counts. If we fitted the negative binomial distribution to the position-level read counts, the goodness-of-fit was much better than that for the Poisson model. But it was still worse than the GP model (Supplementary Table S2). The estimated mean from the negative binomial model was treated as the gene expression estimate. The QuantiGene data were used as gold standards to test the performance again. The slope for the linear regression between the 'true' signal and the estimate on the log scale was 0.72 ($R^2 = 0.62$) and 0.63 ($R^2 = 0.49$) for the two MAQC-2 samples. They were different from 1.0 and worse than the GP model. In terms of computational efficiency, although the calculation of MLEs for both the GP model and the negative binomial model utilized a numerical optimization method, the latter was slower because of the involvement of the Gamma function. For example, it took the GP model ~1 min to estimate the expression levels for the MAQC data, but it took the negative binomial model ~12 min. All the results indicate that the GP model was better than the negative binomial model.

Recently, Li *et al*. (24) used the properties of local sequences to model the varying Poisson rates for different positions and provided an R package 'mseq'. The Poisson rate of each position was modeled as dependent on the gene expression level and the nucleotide sequence surrounding this nucleotide. To compare the GP model with the mseq model, we simulated a bootstrap sample for the position-level read counts based on the fitted GP distribution or the fitted Poisson rates from the mseq model. Then the Kolgomorov–Smirnov test was performed to compare the bootstrap sample from the fitted model and the observed sample. Ideally the two samples should have similar distributions. For the GP model, ~50–98% genes with a P-value >0.05, which indicates that the bootstrap sample approximated to the observed sample. However, the percentage was only ~0–2.04% for the mseq model. Details can be found in Supplementary Table S3. The QuantiGene data were used as gold standards to test the expression estimates of the mseq model. The slope for the regression lines was 0.61 ($R^2 = 0.62$) and 0.58 ($R^2 = 0.54$). It indicates that the expression estimates from the mseq model were worse than those from the GP model. We should also note that the mseq model provides a read count estimate for each position, but the GP model only provides the overall gene expression estimate without specifying the expression level of each position.

DISCUSSION

In this work, we focused on the position-level read count (i.e. the number of reads starting from each position). We found that a two-parameter GP model can fit the position-level read counts more appropriately than a traditional Poisson model. The parameter θ reflects the transcript amount for a gene (or an exon) and the parameter λ represents the average bias during the sample preparation and sequencing process for this gene (or exon). The goodness-of-fit studies showed that the GP model fits the position-level read count much better than the Poisson model. Through the GP model, we can better estimate gene and exon expression levels and perform the normalization across different samples. The GP model improves the identification of differentially expressed genes and the identification of differentially spliced exons.

The estimated gene expression $\hat{\theta}$ can be treated as a shrunk value of the sample mean, because $\hat{\theta} = \bar{x}\left(1 - \hat{\lambda}\right)$ and $\hat{\lambda}$ is a positive value (but <1) for more than 99% of the exons and genes in our applications. In this sense, $\hat{\lambda}$ can be treated as a shrinkage factor for the gene expression estimation. This relationship can also be inferred by the equation that $\theta = \mu(1-\lambda)$. Because $\hat{\lambda}$ had a positive correlation with $\sqrt{\bar{x}}$, we shrunk more for the highly expressed genes or we removed more reads due to sequencing bias. In contrast to microarray data in which the signal is saturated at a certain threshold, RNA-seq data have some extremely large values. However, when we examined the details of the highly expressed genes, we found some suspicious large number of reads starting from several positions of the genes. These outliers substantially affected the sample mean estimation (Figure 2). On the contrary, our GP model can remove these suspicious reads and obtain a more reasonable expression estimate. From our real data analysis, we also found a positive correlation between $\hat{\lambda}$ and the sample standard deviation s. In addition, $\hat{\lambda}$ also had high correlations with certain A/T enriched oligonucleotides for the commonly used library preparation procedures such as random hexamer priming or oligo(dT) priming in human and yeast. However, if the RNA hydrolysis was used before cDNA priming such as in the mouse data, the associated oligonucleotides were changed. More RNA-seq data are needed to draw a conclusion here. Further investigation is needed to study the underlying mechanisms of λ. Hansen *et al*. (25) found that the random hexamer priming induces bias in the nucleotide composition at the beginning of the sequence reads and they proposed a reweighting scheme for the read count. We can easily incorporate it into our data preprocessing to further remove such bias.

The shrinkage property of the GP model has an immediate impact on the normalization across different samples, because a few highly expressed genes specific to one sample makes the library-size based normalization

problematic. For example, in the human liver tissue, the top 10 genes contributed ~31% of the total RNA amount if the $\bar{x}$ was used to estimate the gene expression. If these 10 genes are not expressed in the second tissue and we assume that the total RNA amounts are equal for the normalization, the RNA amounts of the remaining genes are actually very different between the two tissues. Let's insert hypothetical values for this example. The RNA amount from the 10 genes in the liver tissue is 31 and the RNA amount from the remaining thousands of genes is 69. If the 10 genes are not expressed in the second tissue (i.e. 0 RNA amount), the RNA amount for the remaining genes will be scaled to 100 to make the total RNA amount equal. However, the identification of differentially expressed genes among the remaining genes will be problematic because their total RNA amount is 69 versus 100. On the contrary, in our GP model, the contribution from the top 10 genes was shrunk to 1–4% for all of the considered tissues. It therefore improves the identification of differentially expressed genes and the identification of differentially spliced exons (see Table 4, Figures 4 and 5). In the future, we can improve the normalization by further removing extreme values from the GP model and adding the weights about the robustness of the estimates, similar to the method proposed in ref. (7).

To evaluate the performance of the GP model on the identification of differentially expressed genes, we used technical or biological replicates as negative controls. For the technical replicates, ideally, there should be no differentially expressed genes. We found that if the sequencing depth was comparable, the GP model identified no false positives in general. But the Poisson model identified many false positives. As we expected, the difference between biological replicates was larger than the difference between technical replicates. For the very unbalanced design (i.e. the sequencing depth was very different between the two samples), the GP model still performed better than the Poisson model. If the sequencing depth was the same, the performance of the GP model was much better than that of the Poisson model. The results also demonstrate that the advantages of the GP model over the Poisson model are not limited to the normalization issue. The ROC curve studies also concluded that the GP model can better identify differentially expressed genes (Figure 3).

If the sequencing depth is very different for the two samples, besides the difficulty of normalization, the different robustness of the parameter estimation is another issue. As reported in ref. (3), the RPKM estimation was sensitive to the total number of mapping reads for genes with low expression levels. We also showed that the parameter estimation for the GP model was sensitive to the sequencing depth for genes with low expression levels (Supplementary Figure S2). Therefore, we suggest designing a comparable sequencing depth to perform sample comparison. In addition, a deeper sequencing depth will provide us more stable expression estimates. The deeper sequencing yields more sequence reads. A longer read can help to map to the genome more accurately. However it does not contain any addition information about the sampling property of the RNA-seq data.

We used the maximum likelihood approach to estimate the parameters in the GP model. If a gene expression level is too low and there is no position with more than one read, the MLE does not exist. For the identification of differentially expressed genes and differentially spliced exons, we used the log-likelihood ratio tests. Under the null hypothesis, we calculated the profile likelihood. Our parameter of interest is $\Psi = \theta_1/\theta_2$ (or $\Psi = (\theta_Z/\theta_X)/(\theta_V/\theta_Y) = b_1/b_2$ for differentially spliced exons study) and under the null $\Psi = 1$. There are nuisance parameters θ, b and λs. For the first two parameters, we estimated them in a usual way. However, we found that λs are very dependent on Ψ and they cannot be treated as orthogonal parameters (results not shown). For simplicity, we treated the λ estimates from the unrestricted model as true values and put them into the profile likelihood calculation. A more sophisticated method can be specified to better calculate the profile likelihood, but it will increase the computational complexity. Here we also provided a simulation strategy to estimate the null distribution of the test statistics instead of using the chi-square distribution with one degree of freedom.

In the GP model, we made an implicit assumption that the observed position level counts are independent for different positions of a gene. If the dependence of different positions is high, the error rate for the expression estimate will be high (~10-fold increase for correlation around 0.5 compared with the independent case). Details about the simulation study can be found in Supplementary Table S4. However, in the real situation, the positions close to each other may be dependent on each other, but the positions far away should be independent. Thus, there is a block-wise correlation structure. The error rate should be smaller than the case where all of the positions are dependent on each other with a correlation 0.5. More sophisticated model can be developed to better address the dependence issue in the future.

The current studies of alternative splicing using RNA-seq data focus on the junction reads and calculate the 'inclusion rate'. The 'inclusion rate' was represented by the ratio between inclusive junction reads and the sum of inclusive and exclusive junction reads. In exon array studies, the differentially spliced exons were examined in terms of the comparison between the 'splicing ratios', and the 'splicing ratio' was represented by the ratio between exon expression and gene expression. The 'inclusion rate' is always ≤ 1, but the 'splicing ratio' can be >1 because the 'gene expression' is an overall estimate of the expression levels of multiple isoforms. Therefore, the value of the 'splicing ratio' in a single sample alone cannot give us much information, but the comparison of the two 'splicing ratios' is meaningful to identify differentially spliced exons. On the contrary, the 'inclusion rate' in a single sample contains the information about whether this exon is alternatively spliced. However, the sequencing depth is usually too low to obtain enough junction reads and the positional bias is strong. In this work, we showed that the GP model can better identify differentially spliced exons using the 'splicing ratio' approach. More

importantly, the fold change of two different genes of the same sample can be accurately estimated by the GP model. Therefore, in the future, we will compare an exon with a constitutive exon in the same sample to identify alternatively spliced exons based on the GP model.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Lei Li for the discussion about the profile likelihood.

FUNDING

National Institutes of Health (P50 HG 002790, partial); start-up grant from the University of Southern California. Funding for open access charge: a start-up grant from the University of Southern California.

Conflict of interest statement. None declared.

REFERENCES

1. Bullard,J.H., Purdom,E., Hansen,K.D. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.
2. Cloonan,N., Forrest,A.R., Kolle,G., Gardiner,B.B., Faulkner,G.J., Brown,M.K., Taylor,D.F., Steptoe,A.L., Wani,S., Bethel,G. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.
3. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
4. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
5. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
6. Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.
7. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
8. Consul,P.C. and Jain,G.C. (1973) Generalization of Poisson distribution. *Technometrics*, **15**, 791–799.
9. Consul,P.C. and Jain,G.C. (1973) Some interesting properties of generalized Poisson distribution. *Biometr. Z*, **15**, 495–500.
10. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
11. Canales,R.D., Luo,Y., Willey,J.C., Austermiller,B., Barbacioru,C.C., Boysen,C., Hunkapiller,K., Jensen,R.V., Knight,C.R., Lee,K.Y. *et al.* (2006) Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.*, **24**, 1115–1122.
12. Shi,L., Reid,L.H., Jones,W.D., Shippy,R., Warrington,J.A., Baker,S.C., Collins,P.J., de Longueville,F., Kawasaki,E.S., Lee,K.Y. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.
13. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
14. Consul,P.C. (1989) *Generalized Poisson Distributions: Properties and Applications*. Marcel Dekker Inc, New York.
15. Consul,P.C. (1974) A simple urn model dependent upon predetermined strategy *Sankhyā. Indian J. Stat., Ser. B*, **36**, 391–399.
16. Janardan,K.G. and Schaeffer,D.J. (1977) Models for the analysis of chromosomal aberrations in human leukocytes. *Biometrical J.*, **19**, 599–612.
17. Ramskold,D., Wang,E.T., Burge,C.B. and Sandberg,R. (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, **5**, e1000598.
18. Dennis,G. Jr. Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.
19. Castle,J.C., Zhang,C., Shah,J.K., Kulkarni,A.V., Kalsotra,A., Cooper,T.A. and Johnson,J.M. (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat. Genet.*, **40**, 1416–1425.
20. Xu,Q., Modrek,B. and Lee,C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.*, **30**, 3754–3766.
21. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
22. Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.
23. Pepke,S., Wold,B. and Mortazavi,A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6**, S22–S32.
24. Li,J., Jiang,H. and Wong,W.H. (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.*, **11**, R50.
25. Hansen,K.D., Brenner,S.E. and Dudoit,S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, **38**, e131.