Published online 5 May 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 10* ***e75***
doi:10.1093/nar/gkp282

A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level

Sika Zheng[1] and Liang Chen[2,*]

[1]Howard Hughes Medical Institute, University of California, Los Angeles, Los Angeles, CA 90095 and [2]Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA

Received December 15, 2008; Revised April 13, 2009; Accepted April 14, 2009

ABSTRACT

The complexity of mammalian transcriptomes is compounded by alternative splicing which allows one gene to produce multiple transcript isoforms. However, transcriptome comparison has been limited to differential analysis at the gene level instead of the individual transcript isoform level. High-throughput sequencing technologies and high-resolution tiling arrays provide an unprecedented opportunity to compare transcriptomes at the level of individual splice variants. However, sequence read coverage or probe intensity at each position may represent a family of splice variants instead of one single isoform. Here we propose a hierarchical Bayesian model, BASIS (Bayesian Analysis of Splicing IsoformS), to infer the differential expression level of each transcript isoform in response to two conditions. A latent variable was introduced to perform direct statistical selection of differentially expressed isoforms. Model parameters were inferred based on an ergodic Markov chain generated by our Gibbs sampler. BASIS has the ability to borrow information across different probes (or positions) from the same genes and different genes. BASIS can handle the heteroskedasticity of probe intensity or sequence read coverage. We applied BASIS to a human tiling-array data set and a mouse RNA-seq data set. Some of the predictions were validated by quantitative real-time RT–PCR experiments.

INTRODUCTION

It has been estimated that more than 90% of human genes are alternatively spliced (1,2). Multiple transcript isoforms produced from a single gene can lead to protein isoforms with distinct functions (3). Alternative splicing (AS) is widely involved in different physiological and pathological processes. Different tissues exhibit different AS patterns, and malfunctions in AS regulatory factors result in various developmental defects (4–8). Abnormal mRNA splicing contributes to many human diseases (9–11). Identifying differentially expressed distinct transcript isoforms is crucial to understanding transcriptional and post-transcriptional regulation of various processes. Thus, there is an urgent need to study the differences between transcriptomes at the individual transcript isoform level. Although full-length cDNA sequencing is a potential approach, it is expensive and labor-intensive, making transcriptome comparison an elusive goal.

High-resolution tiling arrays and high-throughput sequencing technologies (e.g. RNA-seq) provide an unprecedented opportunity to compare transcriptomes at the individual splice variant level. Probes of tiling array are fixed, whereas RNA-seq provides a collection of randomly distributed reads. In the two platforms, each transcript is represented by many probes or covered by a large number of sequence reads. However, a short probe (~25-mer for Affymetrix chips or ~60-mer for NimbleGen chips) matches only a small portion of the transcript sequence. Thus, the probe intensity may not represent the expression level of a single transcript, but rather a family of splice variants. RNA-seq faces the same challenges due to short sequence reads (~35-mer for Illumina Solexa and Applied Biosystems SOLiD, ~200-mer for Roche 454 Life Sciences). Although junction reads are useful for identifying AS events, the low coverage hampers their statistical power, which is more obvious for low-abundance transcripts. In addition, some junction reads are not specific to one transcript isoform, but to a group of transcript isoforms. Novel data analysis methods are needed to fully utilize these high-throughput techniques for inferring transcriptome differences at the individual transcript isoform level, as AS is one of the major means of expanding genome information.

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

Transcriptome comparison at the individual transcript isoform level must jointly consider probes or sequence reads belonging to the same gene, because many genes have multiple alternatively spliced regions and many transcript isoforms do not contain any sequence positions or exon–exon junctions which exclusively appear in these isoforms. Instead, the uniqueness of these transcript isoforms is reflected by the uniqueness of exon combinations. If a nucleotide position (or exon junction) exclusively appear in one isoform but not in others, we call this position an isoform-specific position (or isoform-specific exon junction). Through the analysis (see details in Supplementary Data S1), we found that about 42% of human transcripts exhibit no isoform-specific positions, and about 57% have ≤ 50 base pair (bp) of isoform-specific positions. Approximately 66% of human multi-exon transcripts have no isoform-specific exon junctions. Among mouse transcripts, roughly 39% have no isoform-specific sequence positions, and about 57% have ≤ 50 bp of isoform-specific positions. Approximately 70% of the mouse multi-exon transcripts exhibit no isoform-specific exon junctions. The distribution of the number of isoform-specific positions and isoform-specific exon junctions in a transcript is shown in Supplementary Figure S1. Another complication confronted during the analysis of differential isoform expression is that contiguous splicing choices cannot be directly obtained from either the high-resolution microarray data or the high-throughput sequencing data.

Here we introduce an approach to comparing transcriptomes at the individual transcript isoform level. This was achieved by developing a hierarchical Bayesian model based on transcript splicing patterns assembled from public databases and high-resolution tiling-array or high-throughput sequencing data (specifically RNA-seq). We call this model BASIS (Bayesian Analysis of Splicing IsoformS). BASIS has the ability to borrow information across different probes (or positions) from the same and different genes when making statistical inferences. Differentially expressed transcript isoforms can be directly inferred from the model by introducing a latent variable and accounting for the heteroskedasticity of probe intensity or sequence read coverage. The usefulness of BASIS is illustrated by its application to a human tiling-array data set to compare HeLa and HepG2 cell lines (12) and to a mouse RNA-seq data set to compare brain, liver and muscle tissues (13).

MATERIALS AND METHODS

Hierarchical Bayesian model (BASIS)

For each probe i that appears in at least one transcript isoform of gene g, consider the linear model:

$$\Delta y_{gi} = \sum \Delta\beta_{gj} x_{gij} + \Delta\varepsilon_{gi},$$

where Δy_{gi} is the intensity difference between two conditions for probe i of gene $g(\Delta y_{gi} = y^1_{gi} - y^2_{gi})$, $\Delta\beta_{gj}$ is the expression difference between two conditions for the j-th transcript isoform of gene g, x_{gij} is the binary indicator of whether probe i belongs to isoform j's exon region, and $\Delta\varepsilon_{gi}$ is the error term for probe i of gene g. Within one data set, g ranges from 1 to G, where G is the total number of genes. i ranges from 1 to n_g where n_g is the total number of probes for gene g. And j ranges from 1 to s_g where s_g is the total number of transcript isoforms for gene g. The total $\Delta\varepsilon_{gi}$'s ($g = 1, \ldots, G$ and $i = 1, \ldots, n_g$) are divided into 100 bins. Each bin contains thousands of probes with similar $y^1_{gi} + y^2_{gi}$ values. Because probe intensity variance is dependent on probe intensity mean (see Results), probes in the same bin exhibit similar variances. The same model can be specified for RNA-seq data with y representing the read coverage over each position. Thus, Δy_{gi} is the coverage difference between two conditions for position i of gene g, $\Delta\beta_{gj}$ is the expression difference between two conditions for the j-th transcript isoform of gene g, x_{gij} is the binary indicator of whether position i belongs to isoform j's exon region, and $\Delta\varepsilon_{gi}$ is the error term for position i of gene g. And we only consider nucleotide positions that appear in at least one transcript isoform.

A hierarchical Bayesian model is constructed as:

$$\Delta\mathbf{Y}_g|\boldsymbol{\Delta\beta}_g,\boldsymbol{\Sigma}_g \ \sim N_{n_g}(\mathbf{X}_g\boldsymbol{\Delta\beta}_g,\boldsymbol{\Sigma}_g),\ g = 1,\ldots,G;$$

$$\boldsymbol{\Sigma}_g \equiv diag\ (\pi_{g1,\ldots,}\pi_{gn_g}),\ \pi_{gi} =$$

$$\quad \delta_m \text{ if probe (or position) i of gene g} \in \text{bin m};$$

$$\delta_m \ \sim IG(\nu/2,\nu\lambda/2),\ m = 1,\ldots,100;$$

$$\boldsymbol{\Delta\beta}_g|\gamma_g \ \sim N_{s_g}(\mathbf{0},\mathbf{R}_g);$$

$$\mathbf{R}_g \equiv diag\ (\kappa_{g1},\ldots,\kappa_{gs_g}),\ \kappa_{gj} = \tau_{gj} \text{ if } \gamma_{gj} = 0 \text{ and } \kappa_{gj} =$$

$$\quad \psi_{gj} \text{ if } \gamma_{gj} = 1;$$

$$f(\boldsymbol{\gamma}_g) = \prod_{j=1}^{s_g} p^{\gamma_{gj}}(1-p)^{1-\gamma_{gj}};$$

where $\boldsymbol{\Delta}\mathbf{Y}_g$, $\boldsymbol{\Delta\beta}_\mathrm{g}$ and $\mathbf{X}_g$ are matrixes with elements described before, $\boldsymbol{\gamma}_g$ is a latent variable, N_{n_g} and N_{s_g} stand for multivariate normal distributions, and IG stands for the inverse gamma distribution. Given the isoform amount differences ($\boldsymbol{\Delta\beta}_\mathrm{g}$) and the probe arrangements ($\mathbf{X}_g$), the probe intensity (or read coverage) differences ($\boldsymbol{\Delta}\mathbf{Y}_g$) follow a multivariate normal distribution with mean $\mathbf{X}_g\boldsymbol{\Delta\beta}_\mathrm{g}$ and variance $\boldsymbol{\Sigma}_g$. For the variance $\boldsymbol{\Sigma}_g$, specifically, if a probe (or position) is assigned to bin m, the variance of the intensity (or coverage) difference is δ_m. δ_m itself is a random variable following an inverse gamma distribution. γ_{gj} is an indicator whether the j-th isoform is differentially expressed. When $\gamma_{gj} = 0$, the isoform difference $\Delta\beta_{gj} \ \sim N(0,\tau_{gj})$ and when $\gamma_{gj} = 1, \Delta\beta_{gj} \ \sim N(0,\psi_{gj})$. Here N stands for normal distribution. τ_{gj} was set as a small value so that when $\gamma_{gj} = 0$, $\Delta\beta_{gj}$ is small enough to be estimated as 0. ψ_{gj} was set as a large value so that when $\gamma_{gj} = 1$, $\Delta\beta_{gj}$ is large enough to be included in the final model. Therefore, the latent variable γ can perform variable selection for the linear model. The errors for probes belonging to the same gene can be heteroskedastic and assigned to different bins. In our prior distributions for parameters $(\Delta\beta, \delta, \gamma)$, there are hyperparameters $(\tau,\psi,\nu,\lambda,p)$. The hierarchical structure of BASIS can be found in Figure 1. Through BASIS, we can identify transcript isoforms that are differentially expressed between two conditions.

The Gibbs sampler was used to generate a Markov chain and the posterior probabilities of $\Delta\beta, \delta$ and γ were estimated from the chain. The variance parameter $\delta_m^{[0]}$ was initialized to be the mean of intensity sum (y^1+y^2) for probes or positions in bin m. $\gamma^{[0]}$ was initialized as $(1,\ldots,1)^T$. The Gibbs sampler at the k-th iteration proceeds as follows:

(i) Sample the isoform amount differences $\boldsymbol{\Delta\beta}_g^{[k]}$ $(g = 1,\ldots, G)$ from the conditional posterior distribution:

$$\boldsymbol{\Delta\beta}_g^{[k]} \sim f(\boldsymbol{\Delta\beta}_g^{[k]}|\Delta\mathbf{Y}_g,\boldsymbol{\delta}^{[k-1]},\gamma_g^{[k-1]}) =$$
$$N_{s_g}\left(\mathbf{A}\times\mathbf{X}_g^T\left(\boldsymbol{\Sigma}_g^{[k-1]}\right)^{-1}\Delta\mathbf{Y}_g,\mathbf{A}\right),$$
$$\text{where } \mathbf{A} = \left(\mathbf{X}_g^T\left(\boldsymbol{\Sigma}_g^{[k-1]}\right)^{-1}\mathbf{X}_g + \left(\mathbf{R}_g^{[k-1]}\right)^{-1}\right)^{-1}.$$

(ii) Sample $\delta_m^{[k]}$ $(m = 1,\ldots, 100)$, the variance for probes (or positions) in bin m, from the conditional posterior distribution:

$$\delta_m^{[k]} \sim f(\delta_m^{[k]}|\Delta\mathbf{Y}_m,\boldsymbol{\Delta\beta}_m^{[k]},\gamma_m^{[k-1]}) =$$
$$IG\left(\frac{\nu+q_m}{2},\frac{\nu\lambda+(\Delta\mathbf{Y}_m-\mathbf{X}_m\boldsymbol{\Delta\beta}_m^{[k]})^T(\Delta\mathbf{Y}_m-\mathbf{X}_m\boldsymbol{\Delta\beta}_m^{[k]})}{2}\right)$$

where $\Delta\mathbf{Y}_m$, $\mathbf{X}_m$, $\boldsymbol{\Delta\beta}_{\mathrm{m}}$ are for probes (or positions) falling in bin m, q_m is the number of probes (or positions) in bin m. The probes (or positions) in bin m may be from different genes.

(iii) Sample $\gamma_{gj}^{[k]}$ $(g = 1,\ldots, G$ and $j = 1,\ldots, s_g)$, the indictor of whether the j-th isoform should be declared as differentially expressed, from the conditional posterior distribution:

$$\gamma_{gj}^{[\mathrm{k}]} \sim f(\gamma_{gj}^{[\mathrm{k}]}|\Delta\mathbf{Y},\boldsymbol{\Delta\beta}_g^{[k]},\boldsymbol{\delta}^{[\mathrm{k}]},\boldsymbol{\gamma}_{(gj)}^{[\mathrm{k}]}),$$
$$Pr(\gamma_{gj}^{[\mathrm{k}]} = 1|\Delta\mathbf{Y},\boldsymbol{\Delta\beta}_g^{[k]},\boldsymbol{\delta}^{[\mathrm{k}]},\boldsymbol{\gamma}_{(gj)}^{[\mathrm{k}]}) =$$
$$\frac{f(\boldsymbol{\Delta\beta}_g^{[k]}|\boldsymbol{\gamma}_{(gj)}^{[\mathrm{k}]},\gamma_{gj}^{[\mathrm{k}]}=1)p}{f(\boldsymbol{\Delta\beta}_g^{[k]}|\boldsymbol{\gamma}_{(gj)}^{[\mathrm{k}]},\gamma_{gj}^{[\mathrm{k}]}=1)p+f(\boldsymbol{\Delta\beta}_g^{[k]}|\boldsymbol{\gamma}_{(gj)}^{[\mathrm{k}]},\gamma_{gj}^{[\mathrm{k}]}=0)(1-p)}$$
$$\text{where } \boldsymbol{\gamma}_{(gj)}^{[\mathrm{k}]} = (\gamma_1^{[\mathrm{k}]},\ldots,\gamma_{j-1}^{[\mathrm{k}]},\gamma_{j+1}^{[\mathrm{k-1}]},\ldots,\gamma_{s_g}^{[\mathrm{k-1}]})^T.$$

For the choice of hyperparameters τ and ψ, we adopt a semi-automatic approach proposed by George *et al.* (14). In this approach, τ_{gj} and ψ_{gj} were selected by considering the prior odds of excluding an isoform from the model and a t-statistic threshold of including an isoform in the model. $\sqrt{\psi_{gj}/\tau_{gj}}$ is the ratio of the heights of $N(0,\tau_{gj})$ and $N(0,\psi_{gj})$ at 0. Therefore $\sqrt{\psi_{gj}/\tau_{gj}}$ can be interpreted as the prior odds that transcript isoform j is declared as a non-differentially expressed transcript when $\Delta\beta_{gj}$ is very close to zero. We also consider the marginal densities $(\Delta\hat{\beta}_{gj}|\sigma_{\Delta\beta_{gj}},\gamma_{gj}=0) \sim N(0,\sigma_{\Delta\beta_{gj}}^2+\tau_{gj})$ and $(\Delta\hat{\beta}_{gj}|\sigma_{\Delta\beta_{gj}},\gamma_{gj}=1) \sim N(0,\sigma_{\Delta\beta_{gj}}^2+\psi_{gj})$, where $\Delta\hat{\beta}_{gj}$ is the least squares estimator and $\sigma_{\Delta\beta_{gj}}^2$ is the variance of $\Delta\hat{\beta}_{gj}$. The intersection point of these two marginal densities

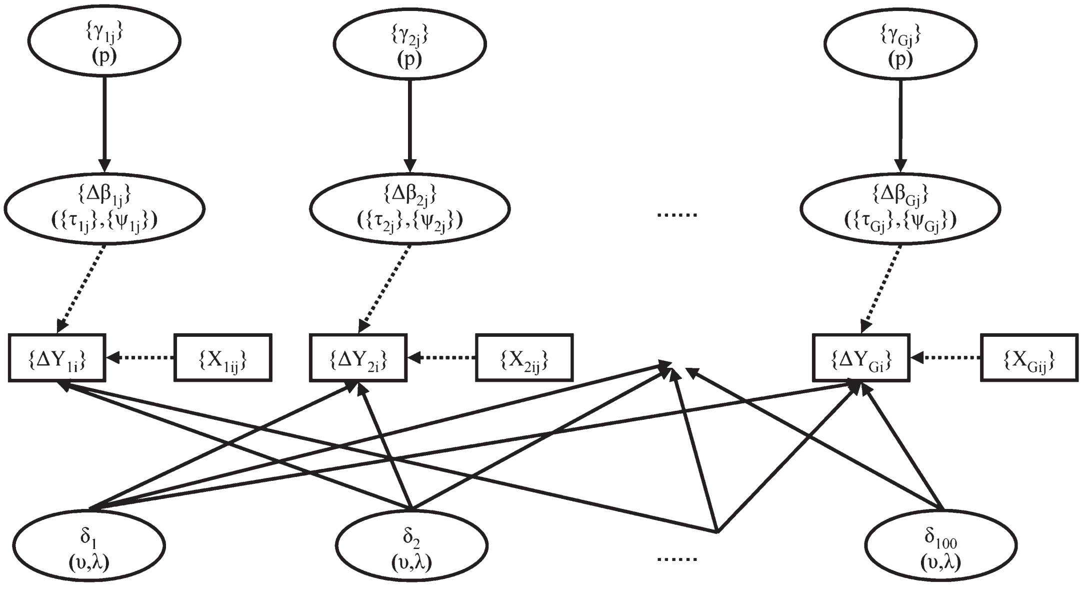


Figure 1. Hierarchical structure of BASIS. The observed data are denoted as rectangles. The random variables besides $\mathbf{Y}_g$ are denoted as ovals. The hyperparameters are listed in brackets. A solid arrow indicates a stochastic dependence while a dashed arrow indicates a logical function. The details of BASIS can be found in Materials and Methods section.

is denoted as $t_{gj}\sigma_{\Delta\beta_{gj}}$ so that the density of $N(0,\sigma^2_{\Delta\beta_{gj}}+\psi_{gj})$ will be larger than the density of $N(0,\sigma^2_{\Delta\beta_{gj}}+\tau_{gj})$ if and only if $\Delta\hat{\beta}_{gj}/\sigma_{\Delta\beta_{gj}} > t_{gj}$. Therefore t_{gj} can be interpreted as a *t*-statistic threshold of whether transcript isoform *j* should be declared as a differentially expressed transcript. Through simple calculation, it can be shown that t_{gj} is a function of $\sigma_{\Delta\beta_{gj}}/\sqrt{\tau_{gj}}$ and $\sqrt{\psi_{gj}/\tau_{gj}}$. Specifically, we chose $(\hat{\sigma}_{\Delta\beta_{gj}}/\sqrt{\tau_{gj}},\sqrt{\psi_{gj}/\tau_{gj}}) = (10, 100)$ where $\hat{\sigma}_{\Delta\hat{\beta}_{gj}}$ is the standard error of the least squares estimator $\Delta\hat{\beta}_{gj}$. This setting was suggested by George *et al.* (14). It indicates that the prior odds that transcript isoform *j* is declared as a non-differentially expressed transcript when $\Delta\beta_{gj}$ is very close to zero is 100, and the t-statistic threshold t_{gj} for the marginal density of $\Delta\hat{\beta}_{gj}$ is about 2.17. Hyperparameter $v = 0$ (and any λ) and $p = 0.5$ were used to represent ignorance as suggested (14–16).

To study the robustness of BASIS to initial values and bin size, in the real data analysis, four Markov chains were generated according to four different settings: (I) Hyperparameters were chosen as described above. We divided the probes (or positions) into 100 bins and δ_m was initialized as the mean of intensity sum (y^1+y^2) for probes or positions in bin *m*; (II) the same as (I) except that we used 20 bins; (III) the same as (I) except that we used 500 bins; (IV) the same as (I) except that we used 100 as the initial value for each δ_m. A total of 10 000 burn-in iterations followed by 40 000 iterations were generated to estimate the posterior probabilities.

To identify differentially expressed transcript isoforms, we used the median model decision rule (17) that includes variables with posterior probability $\Pr(\gamma = 1|\text{data})$ larger than 0.5. Thus, transcript isoforms with posterior mean of γ larger than 0.5 were declared as differentially expressed. If the posterior mean of $\Delta\beta_{gj}$ for the differentially expressed transcript is positive, the isoform was declared to be up-regulated in HeLa for the comparison between HeLa and HepG2, or to be up-regulated in brain for the comparison between brain and liver or the comparison between brain and muscle. Otherwise, the isoform was declared to be down-regulated in HeLa or brain. The lists of genes and their isoforms analyzed can be found in Supplementary Tables S1–3. The differentially expressed isoforms can be found in Supplementary Tables S4–6. BASIS can be downloaded at http://www-rcf.usc.edu/~liangche/software.html.

Simulations

A total of 100 genes were simulated. Nine of them were simulated to have five transcript isoforms and some transcript isoforms were simulated to be differentially expressed. The other 91 genes were created by randomly drawn from the real data and simulated to have no differentially expressed isoforms. The probe arrangements of the five isoforms for the nine differentially expressed genes were simulated as:

$$E = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ \vdots & & \vdots & & \vdots \\ 1 & 0 & 1 & 1 & 1 \\ \vdots & & \vdots & & \vdots \\ 1 & 1 & 1 & 1 & 0 \\ \vdots & & \vdots & & \vdots \\ 1 & 1 & 0 & 0 & 1 \\ \vdots & & \vdots & & \vdots \\ 1 & 1 & 1 & 0 & 1 \\ \vdots & & \vdots & & \vdots \\ 1 & 1 & 1 & 1 & 1 \\ \vdots & & \vdots & & \vdots \end{pmatrix}_{300\times 5} \qquad (^*)$$

Probes 1–50 appear in isoforms 2–5; probes 51–100 appear in isoforms 1 and 3–5; and so on. The matrix $\mathbf{E} = \{e_{ij}\}$ was used as matrix $\mathbf{X}_g$ ($g = 1,\ldots,9$) in BASIS for the nine genes. ΔY_{gi} was simulated as $\sum \Delta\beta_{gj}e_{ij} + \Delta\varepsilon_{gi}$ where $\Delta\varepsilon_{gi}$ follows a normal distribution with mean 0 and variance δ_m which is determined by the bin number for the probe. The choices $\mathbf{\Delta\beta}_g$ and δ_m are discussed as follows.

(i) All five isoform annotations are known for genes 1–3. Isoform 1 and isoform 2 are differentially expressed. The coefficients of the three genes are: $\boldsymbol{\Delta\beta}_1 = (-1.8,1.8,0,0,0)^T$, $\boldsymbol{\Delta\beta}_2 = (-1.8,2.4,0,0,0)^T$ and $\boldsymbol{\Delta\beta}_3 = (-2.4,2.4,0,0,0)^T$.

(ii) For genes 4–6, the annotations of isoform 5 are missing, although it is differentially expressed together with isoform 1 and isoform 2. The coefficients are:
$\boldsymbol{\Delta\beta}_4 = (-1.8,2.4,0,0,1.2)^T$, $\boldsymbol{\Delta\beta}_5 = (-1.8,2.4,0,0,1.8)^T$ and $\boldsymbol{\Delta\beta}_6 = (-1.8,2.4,0,0,2.4)^T$. The correlations between the probe arrangements of isoform 5 and those of isoforms 1, 2, 3 and 4 are −0.2, −0.2, −0.2 and −0.32, respectively.

(iii) For genes 7–9, the annotations of isoform 4 are missing, although it is differentially expressed together with isoform 1 and isoform 2. The coefficients are:
$\boldsymbol{\Delta\beta}_7 = (-1.8,2.4,0,1.2,0)^T$, $\boldsymbol{\Delta\beta}_8 = (-1.8,2.4,0,1.8,0)^T$, and $\boldsymbol{\Delta\beta}_9 = (-1.8,2.4,0,2.4,0)^T$. The correlations between isoform 4 and isoforms 1, 2, 3 and 5 are −0.32, −0.32, 0.63 and −0.32, respectively.

For the other 91 genes, we randomly selected the **X** matrix from the human data. They were simulated to have no differentially expressed isoforms (i.e. ΔY_{gi} was simulated as $\Delta\varepsilon_{gi}$ because $\boldsymbol{\Delta\beta}_g = \mathbf{0}$ for $g = 10,\ldots,100$). In total, there were 28 132 probes and 368 transcript isoforms. These probes were randomly assigned to 100 bins. For the *m*-th bin, the variance δ_m was simulated as m. About 1000 simulations were performed. For each simulation, we used a burn-in of 1000 iterations, followed by 4000 iterations. Hyperparameters were chosen as described before. Different thresholds for the posterior mean of γ were used to declare transcript isoforms as differentially expressed. The power and the false-positive rates were calculated as average values from those 1000 simulations.

Instead of using the purely simulated matrix **E** [shown in (*)] for the nine differentially expressed genes, we also randomly selected a probe arrangement matrix from genes with five isoforms in the human data and used the matrix as **E** to simulate $\Delta\mathbf{Y}$'s for the differentially expressed genes as described. The other 91 genes without any differentially expressed isoforms were the same as described above. We performed 1000 simulations to calculate the average power when the average false positive rate is 0.005. Then, we repeated these procedures 100 times. Thus we tested 100 different matrix **E**'s for the nine genes with differentially expressed isoforms.

Tiling array data preprocessing

Non-redundant transcript isoform information of human genes was downloaded from the AS and Transcript Diversity database (18) (http://www.ebi.ac.uk/astd/, release 1.1, names begin with 'TRAN') and the Ensembl Genome Browser (http://www.ensembl.org/index.html, release 50, names begin with 'ENST'). Expression levels of these transcripts were from the whole-genome tiling arrays in which the human genome is split into 91 chips at 5-bp resolution (as measured from the central positions of adjacent 25-mer oligonucleotides) (12). We considered the expression data for HeLa and HepG2 cell lines in the cytosol. For each cell line, there were about three replicates. Those RNAs were polyadenylated and longer than 200 nt. The probe coordinates were from the NCBI version 35. The UCSC liftover tool (http://genome.ucsc.edu/) was used to convert the coordinates between version 35 and version 36. The probes mapped to intergenic regions were used as background probes to train the sequence-specific model that considers the composition of the nucleotides at each position of a 25-mer probe (19,20). Thus:

$$\log_2 \mathrm{Bkg}(PM_i) = \alpha n_{iT} + \sum_{j=1}^{25} \sum_{k\in\{A,C,G\}} \beta_{jk} I_{ijk} + \sum_{k\in\{A,C,G,T\}} \gamma_k n_{ik}^2 + \varepsilon_i,$$

where PM_i is the intensity of the perfect match probe i, n_{ik} is the number of nucleotide k in probe i, I_{ijk} is the indicator of nucleotide k in position j of probe i, α, β_{jk}, γ_k are effect parameters and ε_i is the probe-specific error term. The parameters of the model were estimated by using about 300 000 intergenic probes for each chip separately. For eight chips with less than 300 000 intergenic probes, we used all of the intergenic probes. Then the probe intensity on each chip was background corrected according to the estimated parameters and their nucleotide contents. Thus:

$$Corrected(PM_i) = \max(PM_i - Bkg(PM_i), 0)$$

The background-corrected probe intensities were further quantile-normalized and averaged across replicates as the final expression level. Note that those probe intensities were not in the log-scale.

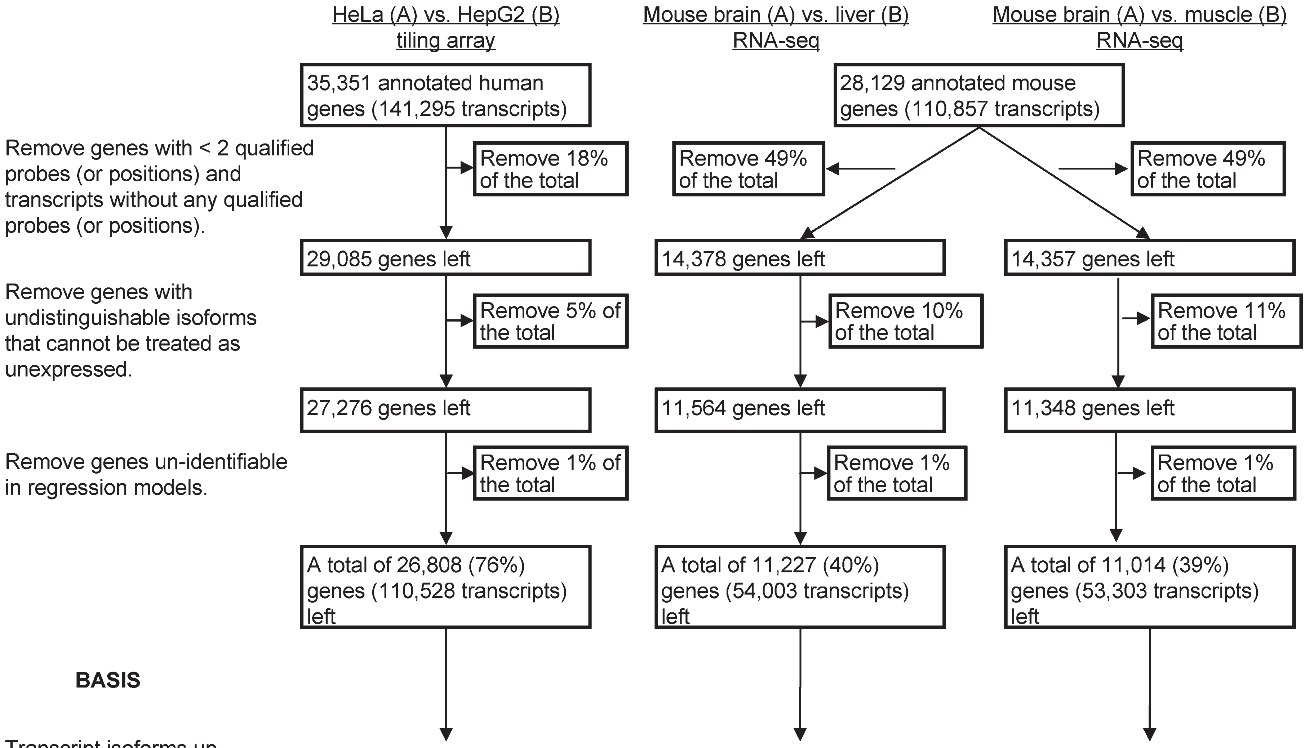
Probes with intensity level larger than 4 in at least one cell line were counted as qualified probes. We removed transcripts without any qualified probe. In other words, if a column of **X** is equal to vector **0**, the corresponding transcript isoform was considered not expressed and removed thereby. Recall that **X** is the probe arrangement matrix and each row represents a qualified probe and each column represents an isoform. Among the 35 351 annotated genes (141 295 transcripts), 29 085 genes have at least two qualified probes. Among the 29 085 genes, 3197 do not have enough qualified probes to distinguish different transcript isoforms. The case of not having enough qualified probes results in identical columns in matrix **X** (e.g. $\mathbf{X}_j = \mathbf{X}_k$ where $\mathbf{X}_j$ is the j-th column and $\mathbf{X}_k$ is the k-th column). For example, considering gene A and gene B each of which have three isoforms, their exon arrangements are:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \text{or } \mathbf{B} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

When the fourth exon (or the fourth row) has no designed probe or qualified probe, the matrix **X** will become:

$$\mathbf{X}_a = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \text{or } \mathbf{X}_b = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then the first two isoforms have the same probe arrangements and they cannot be distinguished (or they are undistinguishable isoforms). There are two possible reasons for undistinguishable isoforms: (I) The isoforms are not expressed and the signal of qualified probes (e.g. the non-zero elements in $\mathbf{X}_j$ or $\mathbf{X}_k$) is from other expressed transcript isoforms that share these probes. (II) The designed probes are not dense enough to have unique probe combinations to represent these isoforms. For case (I), we removed isoform j and k and retained the gene with other isoforms (1388 genes). For example, for matrix $\mathbf{X}_a$, the first two isoforms can be treated as unexpressed isoforms and the signal of the first probe (the non-zero element in $\mathbf{X}_1$ and $\mathbf{X}_2$) is from isoform 3. We therefore retain this gene with isoform 3. On the other hand, we removed genes with undistinguishable isoforms that cannot be treated as unexpressed in case (II). For example, for matrix $\mathbf{X}_\mathrm{b}$, the first two isoforms cannot be treated as un-expressed because probe 1 is not shared by isoform 3 and the signal of probe 1 must come


HeLa (A) vs. HepG2 (B)
tiling array
Mouse brain (A) vs. liver (B)
RNA-seq
Mouse brain (A) vs. muscle (B)
RNA-seq
35,351 annotated human genes (141,295 transcripts)
28,129 annotated mouse genes (110,857 transcripts)
Remove genes with < 2 qualified probes (or positions) and transcripts without any qualified probes (or positions).
Remove 18% of the total
Remove 49% of the total
Remove 49% of the total
29,085 genes left
14,378 genes left
14,357 genes left
Remove genes with undistinguishable isoforms that cannot be treated as unexpressed.
Remove 5% of the total
Remove 10% of the total
Remove 11% of the total
27,276 genes left
11,564 genes left
11,348 genes left
Remove genes un-identifiable in regression models.
Remove 1% of the total
Remove 1% of the total
Remove 1% of the total
A total of 26,808 (76%) genes (110,528 transcripts) left
A total of 11,227 (40%) genes (54,003 transcripts) left
A total of 11,014 (39%) genes (53,303 transcripts) left
BASIS

Transcript isoforms up-regulated in (A)	8,851 transcripts	21,188 transcripts	19,851 transcripts
Transcript isoforms up-regulated in (B)	3,003 transcripts	14,527 transcripts	14,275 transcripts
Genes only with transcript isoforms up-regulated in (A)	6,597 genes	3,343 genes	3,240 genes
Genes with transcript isoforms up-regulated in (A) and (B)	789 genes	5,711 genes	5,498 genes
Genes only with transcript isoforms up-regulated in (B)	1,805 genes	1,717 genes	1,816 genes

Figure 2. Workflow of the prescreening steps and BASIS results.

from isoform 1 and (or) isoform 2. We have to remove the whole gene. After the above procedures, we have 27276 genes left. Among them, 468 genes are un-identifiable because the columns of $\mathbf{X}$ are perfectly collinear (i.e. $\mathbf{X}_j$ is a linear combination of the other columns). The un-identifiability is detected by considering the rank of $\mathbf{X}$ which is determined by the singular value decomposition. The prescreening procedures are summarized in Figure 2. Finally, a total of 26 808 human genes (110 528 transcripts) were considered in BASIS. The average probe number for each gene is 323 and the median is 206.

RNA-seq data preprocessing

Non-redundant transcript isoform information of mouse genes was downloaded from the ASTD and Ensembl databases. Expression levels of these transcripts were from the high-throughput RNA-seq data for adult mouse brain, liver and muscle (13). For each tissue, there were two replicates. Uniquely mapped sequence reads from two replicates were pooled together and mapped to genes. The number of reads mapped to a position was treated as the read coverage over that position. The read coverage was multiplied by a constant to make the total number of reads equivalent for the three tissues (28 million). We compared brain with liver and brain with muscle.

Positions with read coverage larger than 4 in at least one tissue were counted as qualified positions. For the comparison between brain and liver, among the 28 129 annotated genes (110 857 transcripts), 49% of them have less than two qualified positions. About 10% of them do not have enough qualified positions to distinguish different transcript isoforms and this cannot be explained by the lack of expression for these isoforms. In addition, 1% of them are un-identifiable. For the comparison between brain and muscle, 49% of them have less than two qualified positions. About 11% of them were removed because some isoforms cannot be distinguished and this cannot be explained by the fact that these isoforms are not expressed. Another 1% of them are un-identifiable. The details of the prescreening procedures can be found in Figure 2. Finally, a total of 11 227 genes (54 003 transcripts) and 11 014 genes (53 303 transcripts) were considered in BASIS for the comparison between mouse brain and liver and the comparison between brain and muscle respectively. For the brain versus liver comparison, the average position number for each gene is 1652 and the median is 1304. For the brain versus muscle comparison, the average position number is 1640 and the median is 1248.

RNA preparation and qRT–PCR

Adult C57BL mouse brain, liver and muscle tissues were dissected and quickly submerged in Trizol (Invitrogen, CA) followed by immediate tissue homogenization. Total RNA samples were prepared according to manufacturer's protocol (Invitrogen, CA). Cytosolic RNA of HeLa and HepG2 cells were generous gifts from Gingeras's group, original authors of the tilling array data (12). RNA were treated with RQ1 RNase-free DNase I (Roche Applied Science) at 1 U/μg RNA and reverse transcription was done as described previously (21). Real-time RT–PCR was performed as previously described (22) using SYBR Green Supermix on a Bio-Rad iQ5 thermocycler for 40 cycles at 60°C annealing temperature. Primers are listed in Supplementary Table S7. Each primer pair amplifies only one amplicon and the identity of RT–PCR product was confirmed by direct sequencing. Relative mRNA levels between brain and liver (brain/liver ratio) or between brain and muscle (brain/muscle ratio) were first normalized by geometric averaging of multiple internal control genes (including Gapdh, Sdha and mRps18a) (23) and then quantified using ΔΔCt method. Relative mRNA levels between HeLa and HepG2 cells were first normalized by geometric average of three internal control genes (HPRT1, RPLP0 and SDHA) and then quantified using ΔΔCt method.

RESULTS

In BASIS, for gene g, the probe intensity (or read coverage over each position) is modeled as the sum of the intensity of transcript isoforms containing this probe (or position):

$$y_{gi} = \sum \beta_{gj} x_{gij} + \varepsilon_{gi}$$

where y_{gi} is the intensity value of probe i (or read coverage at position i) of gene g, β_{gj} is the abundance of the j-th transcript isoform, x_{gij} is the binary indicator of whether probe i (or position i) belongs to isoform j's exon region, and ε_{gi} is the error term for probe i (or position i). The difference in probe intensity (or read coverage) under two conditions (Δy_{gi}) is modeled as the combination of transcript isoforms' differences ($\Delta\beta_{gj}$'s):

$$\Delta y_{gi} = \sum \Delta\beta_{gj} x_{gij} + \Delta\varepsilon_{gi}.$$

To infer the differentially expressed transcript isoforms ($\Delta\beta_{gj} \neq 0$), we introduced a latent variable $\gamma_g = (\gamma_{g1}, \ldots, \gamma_{gs_g})^T$, where $\gamma_{gj} = 1$ means that the j-th isoform is differentially expressed and $\gamma_{gj} = 0$ means that it is not differentially expressed. A homogeneous ergodic Markov chain was generated by our Gibbs sampler. The empirical distribution of γ based on the Markov chain will converge to the actual posterior probability of γ (14). The hierarchical Bayesian model also accounts for the heteroskedastic errors associated with different probes (or positions), as discussed below. The hierarchical structure of BASIS is represented in Figure 1 as we mentioned in Materials and Methods.

Heteroskedasticity of probe intensity and sequence read coverage

Microarray noise has been shown to be scale dependent (24). Similarly for RNA-seq data, the noise associated with read coverage over each position is proportional to the mean. Figure 3A and B illustrates the relationship between the mean and the standard deviation across replicates. Because of the scale-dependent noise, log-transformed intensity was always used in the microarray study to minimize the effect of such heteroskedastic errors with different variances. However, the log-transformed intensity cannot be modeled simply as the sum of the intensity of individual transcript isoforms (i.e. $\log(y_{gi}) \neq \sum \log(\beta_{gj}) x_{gij} + \varepsilon_{gi}$). In addition, the log ratio of probe intensity under two conditions cannot be modeled as the sum of isoform differences at the log scale (i.e. $\log(\Delta y_{gi}) \neq \sum \log(\Delta\beta_{gj}) x_{gij} + \Delta\varepsilon_{gi}$). To handle the heteroskedasticity and concomitantly maintain the valid linear isoform combination assumption, we divided all of the probes (or positions) across the whole genome into bins according to their intensity values (or read coverage) under two conditions. Probes (or positions) with similar intensity values have similar variances. And different variance parameters were specified for different bins. Thus, $\Delta\varepsilon_{gi} \sim N(0,\delta_m)$ if probe i (or position i) falls into bin m. Large number of probes (or positions) in each bin provided a more stable variance estimate for $\Delta\varepsilon_{gi}$ than that estimated from very few experimental replicates (e.g. two or three replicates) of single probe. We therefore borrowed strength across probes from different genes. Figure 3C and D shows the histograms of the number of different bins for a gene in the tiling array data or the RNA-seq data. For most genes, probes (or positions) fall into multiple bins, further showing their heteroskedasticity.

Power analysis

We studied the statistical power of BASIS, particularly when the isoform information was incomplete. A total of 100 genes were simulated each time. Nine genes were simulated to have five potential transcript isoforms, and some transcript isoforms were differentially expressed. The other 91 genes were simulated to have no differentially

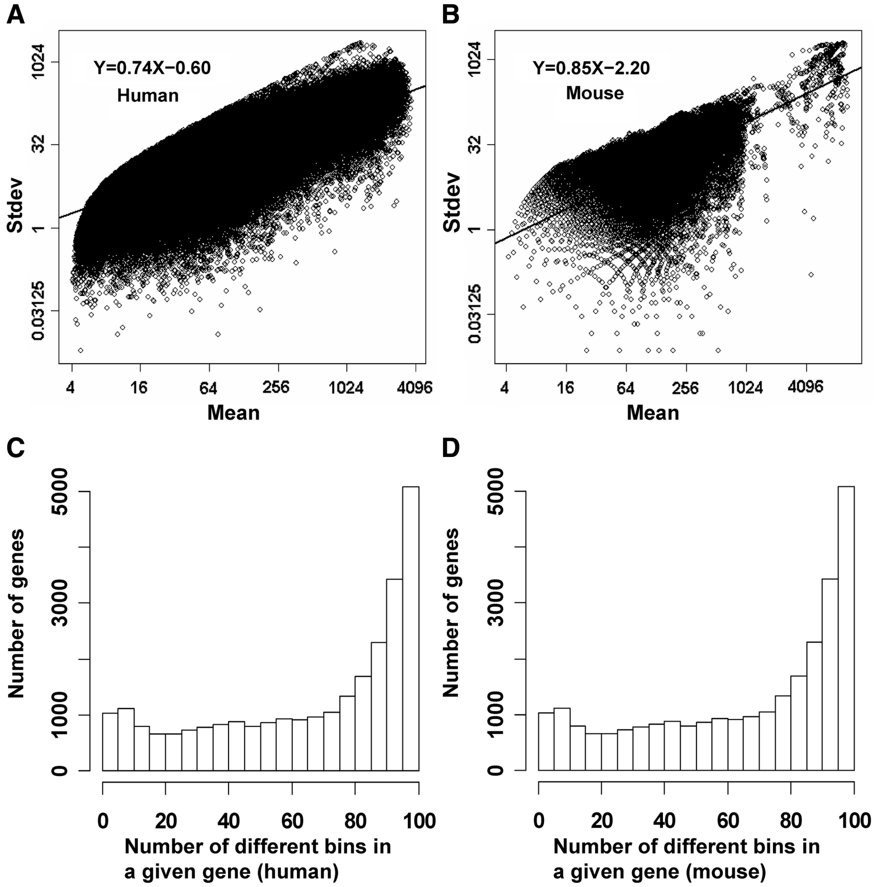


Figure 3. Heteroskedasticity of probe intensity and sequence read coverage. (**A** and **B**) Multiplicative error of probe intensity (A) and sequence read coverage (B). The *x*-axis represents the mean probe intensity across three replicates (A) or the mean sequence read coverage across two replicates (B). The *y*-axis represents the corresponding standard deviation of probe intensity or sequence read coverage. In (A), 500 000 normalized probe intensity data obtained from the human HeLa tiling array were used. In (B), normalized sequence read coverage at 500 000 nucleotide positions from the mouse liver RNA-seq data were used. *x* and *y* are plotted in a $\log_2$ scale for visual convenience. (**C** and **D**) Histograms demonstrating the number of different bins in a given gene for the tiling array data (C) and the RNA-seq data (D). Probes or positions were divided into 100 bins according to their intensity or sequence read coverage. For each gene, we counted the number of different bins that its probes or positions belong to. The number of genes with a specific number of bins was shown in the histograms.

expressed transcript isoforms. Different scenarios in terms of $\Delta\beta_{gj}$ and the completeness of the transcript isoform information were examined. The details of the simulation settings can be found in 'Materials and Methods' section.

Table 1 compares the power of BASIS and the least squares fit when the total false-positive rate was controlled at 0.005. A particular probe arrangement matrix **E** [shown in (*)] was used for this study. The power of BASIS is 0.76 when the false positive rate is 0.005. It demonstrates that BASIS can correctly identify most of the differentially expressed isoforms (13.68 out of 18) and also correctly declare non-differentially expressed isoforms (348.25 out of 350). Note that we have additional 91 genes with no differentially expressed isoforms and there are a total of 350 non-differentially expressed isoforms. BASIS has a much larger statistical power than the least squares fit (0.76 versus 0.31). This is due to the fact that errors for probes of the same genes are heteroskedastic, and BASIS takes this into account. We also separately calculated the power and the false-positive rate for genes 1–9. For example, gene 1 has five transcripts isoforms, two of which are differentially expressed. Thus, the total number of positive instances is 2 and the total number of negative instances is 3 when we calculate the power and false-positive rate for gene 1. The settings for genes 2 and 3 are similar to those for gene 1, except for the differential signals. When the differential signal increases ($\Delta\beta_{gj}$ from 1.8 to 2.4), the performance of the model improves (power from 0.74 to 0.96). When the information for one differentially expressed isoform is lacking (i.e. there is no annotation about the transcript isoforms, but it exists in cells and is differentially expressed), the inferences for other isoforms are still reliable (genes 4–6). The worst situation is when the differential signal for the missing isoform is very high ($\Delta\beta_{gj}$ = 2.4 for isoform 4 of gene 9)

Table 1. Performance of BASIS and the least squares fit

	BASIS		Least squares fit	
	Power	False-positive rate	Power	False-positive rate
Total	0.76	0.005	0.31	0.005
Gene 1	0.74	0.002	0.16	0.002
Gene 2	0.88	0.002	0.29	0.005
Gene 3	0.96	0.0003	0.43	0.004
Gene 4	0.65	0.001	0.33	0.006
Gene 5	0.56	0.001	0.35	0.007
Gene 6	0.59	0.001	0.38	0.008
Gene 7	0.83	0.1	0.27	0.02
Gene 8	0.89	0.3	0.28	0.08
Gene 9	0.72	0.4	0.29	0.2

The total false-positive rate was controlled at 0.005. The power and the false–positive rate were the average values across 1000 simulations. They were also calculated for genes 1–9 separately. For genes 1–3, the annotations for all five isoforms are known and $\Delta\beta_1 = (-1.8,1.8,0,0,0)^T$, $\Delta\beta_2 = (-1.8,2.4,0,0,0)^T$, and $\Delta\beta_3 = (-2.4,2.4,0,0,0)^T$. For genes 4–6, the annotations of isoform 5 are missing but isoform 5 is differentially expressed together with isoform 1 and isoform 2:$\Delta\beta_4 = (-1.8,2.4,0,0,1.2)^T$, $\Delta\beta_5 = (-1.8,2.4,0,0,1.8)^T$, and $\Delta\beta_6 = (-1.8,2.4,0,0,2.4)^T$ The correlations between the exon arrangements of isoform 5 and those of isoforms 1, 2, 3 and 4 are –0.2, –0.2, –0.2 and –0.32. For genes 7–9, the annotations of isoform 4 are missing. But isoform 4 is differentially expressed together with isoform 1 and isoform 2: $\Delta\beta_7 = (-1.8,2.4,0,1.2,0)^T$, $\Delta\beta_8 = (-1.8,2.4,0,1.8,0)^T$, and $\Delta\beta_9 = (-1.8,2.4,0,2.4,0)^T$. The correlations between isoform 4 and isoforms 1, 2, 3 and 5 are –0.32, –0.32, 0.63 and –0.32.

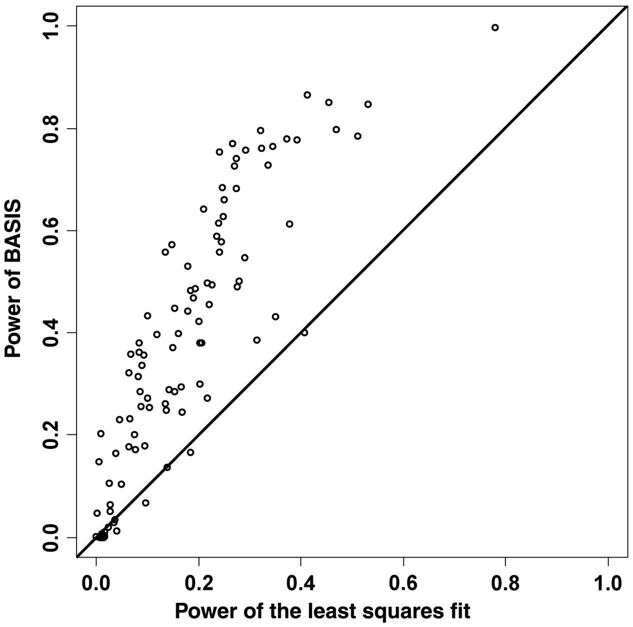


Figure 4. Power of BASIS and the least squares fit for 100 different matrix **E**s. The powers were calculated based on 1000 simulations on 100 genes. The total false-positive rate was controlled at 0.005.

and this isoform demonstrates a high correlation with other known isoforms (the correlation between isoforms 4 and 3 is about 0.63). The false-positive rate can be as high as 0.4. The results demonstrate that when the AS information is incomplete and the missing transcript isoform that has not been annotated is actually differentially expressed, the model still performs well if the missing transcript isoform has a reasonably low correlation with other known transcripts or the differential signal is low.

Besides the purely simulated probe arrangement matrix **E** [shown in (*)] for genes with differentially expressed isoforms, we also tested another 100 different probe arrangement matrix **E**'s randomly drawn from the real data (genes in the human data and with five isoforms). For each matrix **E**, the same simulation settings as mentioned in 'Materials and Methods' section were preformed: nine genes with differentially expressed isoform were simulated and there were another 91 non-differentially expressed genes. The overall power of BASIS and the least squares fit for the 100 genes were calculated based on 1000 simulations for each **E**. As shown in Figure 4, BASIS consistently performs better than the least squares fit. There is about 2-fold increase in the power of BASIS most of time. The results also indicate that the gene annotation structure (**E**) will affect the power of BASIS. Specifically, if a gene has more probes (or positions), thus the number of rows of **E**(n) is larger; the power of BASIS is larger. The Pearson correlation between n and the power is 0.34 which is significant with a P-value of 0.0005. The correlation was calculated based on the 100 different **E**'s. In addition, if the difference among isoforms is larger, the power of BASIS is larger. Here the difference among isoforms was measured as the average Manhattan distances among isoforms (i.e. among columns of **E**) divided by n. The Pearson correlation between the difference measure and the power of BASIS is 0.38 with a P-value of 0.0001. Finally, BASIS does not rely on the percentage of isoform-specific positions of a gene. For each **E**, we calculated the percentage of positions which appear in only one isoform. The power of BASIS is not related to the percentage of isoform-specific positions with a P-value of 0.29. This is because BASIS considers the joint behavior of probes targeting on the same gene.

HeLa and HepG2 tiling-array data analysis

The array data was obtained from Kapranov *et al.* (12), who profiled the cytosolic polyadenylated [poly(A)$^+$] RNAs in HeLa and HepG2 cell lines using whole-genome 5-bp resolution tiling arrays. The known or predicted human transcript isoform splicing patterns were obtained from the ASTD and Ensembl databases. After the preprocessing to remove unexpressed genes etc., a total of 110 528 transcripts (26 808 genes) were considered in BASIS. Overall, 11 854 transcripts were differentially expressed between HeLa and HepG2 cells. About 8851 transcripts were up-regulated in HeLa cells, and the remaining 3003 transcripts were up-regulated in HepG2 cells. These differentially expressed transcripts belong to 9191 genes, indicating that some genes have more than one differentially expressed transcript isoform. Specifically, 1892 genes have more than one differentially expressed transcript isoform. More interestingly, 789 exhibited at least one up-regulated isoform and at least one down-regulated isoform in HeLa compared to HepG2 cells. These have been summarized in the workflow Figure 2. The list of differentially expressed transcripts can be found in Supplementary Table S4.

The convergence of the chain was evaluated by tracing the variance estimate δ_m for each bin m. All of the variance estimate δ_m passed the Geweke's diagnostic, the Raftery and Lewis's diagnostic, and the Heidelberger and Welch's convergence diagnostic implemented in the R package 'coda' (25). Using the posterior mean of $\Delta\beta$, we calculated the residual of each probe ($\Delta\varepsilon$). Residuals falling in the same bin should be approximately normally distributed, with a mean of 0 and variance equal to the estimated variance δ_m for this bin. The residual Q–Q plots (Supplementary Figure S2) show that the distributions of residuals were similar to those expected.

Mouse brain, liver and muscle RNA-seq data analysis

The RNA-seq data was obtained from Mortazavi *et al.* (13), who used Solexa high-throughput sequencing to quantify the poly(A)$^+$ RNA in adult mouse brain, liver and muscle. The sequence read coverage at nucleotide resolution was normalized across different tissues such that the total number of reads was equivalent. Similarly, the known or predicted mouse transcript isoform splicing patterns were obtained from the ASTD and Ensembl databases. For the comparison between brain and liver, 35 715 transcripts were differentially expressed. About 21 188 transcripts were up-regulated in brain, and the others were up-regulated in liver. These transcripts correspond to 10 771 genes. About 7699 genes have more than one differentially expressed transcript isoform. Among them, 5711 exhibited at least one up-regulated isoform and at least one down-regulated isoform in brain compared to liver. For the comparison between brain and muscle, 34 126 transcripts belonging to 10 554 genes were differentially expressed. About 19 851 of the transcripts were up-regulated in brain and the others were up-regulated in muscle. Among these differentially expressed genes, 7392 have more than one differentially expressed transcript isoforms and 5498 of them exhibited at least one up-regulated isoform and at least one down-regulated isoform in brain compared to muscle. The above results have also been summarized in the workflow Figure 2.

The convergence of the chain was further evaluated by tracing the variance estimate δ_m for each bin. All of the variance estimate δ_m passed the Geweke's diagnostic, the Raftery and Lewis's diagnostic, and the Heidelberger and Welch's convergence diagnostic. The residual Q-Q plots (Supplementary Figures S3 and S4) show that the residuals in the same bin were approximately normally distributed, with a mean of 0 and variance equal to the variance estimated for this bin.

Using the junction reads as an independent data resource, we evaluated the performance of BASIS. We mapped the splice-spanning reads to the transcript isoforms considered in BASIS. About 3732 transcript isoforms have at least one sequence read over their isoform-specific splice junctions in brain and (or) liver. For the comparison between brain and muscle, the number is 3679. Isoform-specific splice junction means that no other transcript isoforms contain the same junction. Such transcripts were designated as 'present' in tissues. This is a stringent criterion, as many of the truly present transcripts may not have any isoform-specific splice junctions (Supplementary Figure S1) or may not have any reads over their isoform-specific splice junctions owing to the low abundance. We declared the transcripts with junction read difference larger than four as differentially expressed. As shown in Figure 5, for the comparison between brain and liver, among the differentially expressed transcripts declared by junction read difference, about 83% or 81% of them were also predicted as up-regulated in brain or liver by BASIS. For the comparison between brain and muscle, about 80% or 83% of them were also predicted as up-regulated in brain or muscle by BASIS. The results indicate that BASIS has a statistical power of 80–83%.

Robustness of BASIS to bin size and initial value specifications

The hyperparameters of BASIS were chosen by a semi-automatic approach or chosen as non-informative values to represent ignorance as described in Materials and Methods section. We also studied the robustness of BASIS for different bin sizes and initial values. Four Markov chains were generated according to different bin sizes or different initial values (see details in Materials and Methods section). For the tiling-array data, among the declared differentially expressed transcripts, 95% of them can be detected by all of the chains. For the RNA-seq data, 91% and 88% of them can be detected by all of the chains for the brain-liver comparison and the brain-muscle comparison respectively. Specifically, for different bin sizes, the overlap among the three scenarios (20, 100, 500 bins) is about 89–95%. The results are not exactly the same because of the different strength borrowed from probes (or positions) due to different bin sizes. For different initial values, the overlap among results is about 98–99%. The above results suggest the robustness of the inference results for different bin sizes and initial values.

Experimental validation

To further examine the prediction power of BASIS, we subsequently performed real time RT–PCR experiments to assay transcript isoforms' relative expression levels between adult mouse brain and liver, between adult mouse brain and muscle, and between HeLa and HepG2 cells. We were particularly interested in genes whose isoforms show distinct differential expression patterns between the two conditions. For example, one transcript isoform is up-regulated in brain than in liver, whereas

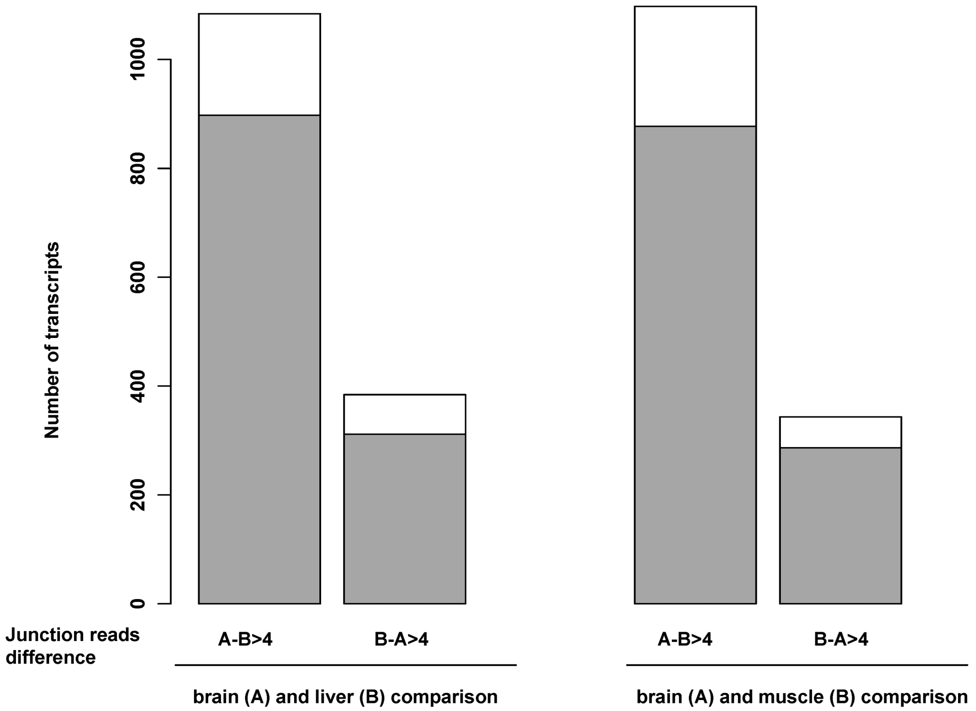


Figure 5. Differentially expressed transcript isoforms predicted by isoform-specific junction reads and BASIS. For each comparison (A versus B), we declared a transcript isoform as up-regulated in A or B if the junction read difference A – B > 4 or B – A > 4. The junction read has to be isoform-specific to the transcript. Grey areas represent the proportions of transcripts declared as differentially expressed by both junction read difference and BASIS. White areas represent the proportions of transcripts predicted as differentially expressed by junction read difference but not by BASIS.

anther transcript isoform of the same gene is down-regulated or is not differentially expressed. For each tested transcript isoform, we designed one of the two PCR primers from the isoform-specific exonic region or exon junction that exclusively represents the isoform. For the RNA-seq data, we randomly tested the relative expression levels of 14 transcript isoforms between mouse brain and liver (Figure 6A), but the transcript isoforms were required to have an isoform-specific exonic region or exon junction and the selection was biased toward genes with isoforms showing distinct expression patterns. Transcripts TRAN00000157032 (Slc25a25), ENSMUST 00000115599 (Pcdh1), TRAN00000139600 (Mrps12), TRAN00000123912 (M6prbp1) and TRAN00000143381 (Clu) were predicted to be up-regulated in brain than in liver by BASIS (black bars in Figure 6A). Transcripts TRAN00000157033 (Slc25a25), ENSMUST00000057185 (Pcdh1), ENSMUST00000019726 (M6prbp1), TRAN00000161590 (Esd), TRAN00000143382 (Clu) and ENSMUST00000000335 (Comt) were predicted to be down-regulated in brain than in liver (white bars). Transcripts TRAN00000139599 (Mrps12), TRAN 00000161592 (Esd) and ENSMUST00000115609 (Comt) were predicted not to be differentially expressed between the two tissues (grey bars). As shown in Figure 6A, all of the transcripts except TRAN00000143381 (Clu) and ENSMUST00000115609 (Comt) show the predicted differential expression patterns. We also tested these transcripts' relative expression ratios between mouse brain and muscle (Figure 6B). All transcripts except Transcripts TRAN00000157033 (Slc25a25), TRAN00000 161592 (Esd), ENSMUST00000115609 (Comt) and ENSMUST00000000335 (Comt) show the predicted differential expression patterns. More importantly, most of genes (except Clu in Figure 6A and B; Pcdh1 and Esd in Figure 6B) have their two transcript isoforms showing significantly different relative expression ratios (*P*-values based on Student's *t*-test ≤ 0.05). It shows that transcript isoforms of the same gene can have distinct expression patterns. However, the standard differentially expressed gene analysis cannot detect such subtle differences.

For the tilling array data, we randomly tested 12 transcript isoforms in HeLa and HepG2 cells (Figure 6C), but the transcript isoforms were required to have an isoform-specific exonic region or exon junction and the selection was biased toward genes with isoforms showing distinct expression patterns. Transcripts ENST00000226225 (TNFAIP1), TRAN00000076466 (PTDSS2), TRAN0000 0076464 (PTDSS2) and ENST00000368680 (NPR1) show the predicted patterns of being up-regulated in HeLa cells compared to HepG2 cells. Transcripts TRAN00000094700 (TNFAIP1), TRAN00000112564 (WDR39), ENST00000394196 (CHD2) and ENST00000361900 (SCAMP5) show the predicted patterns of being down-regulated in HeLa cells compared to HepG2 cells. Although the other four transcripts

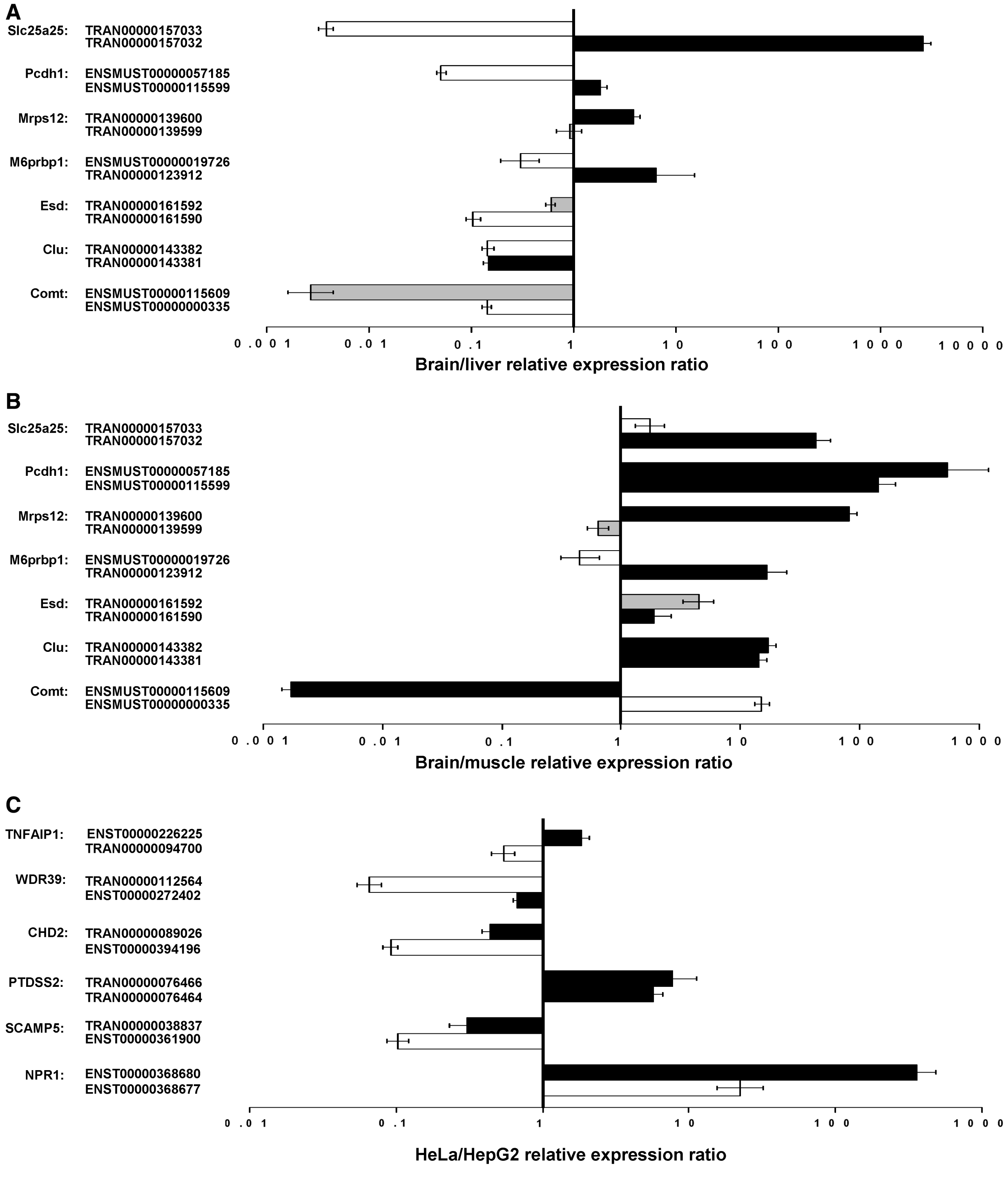
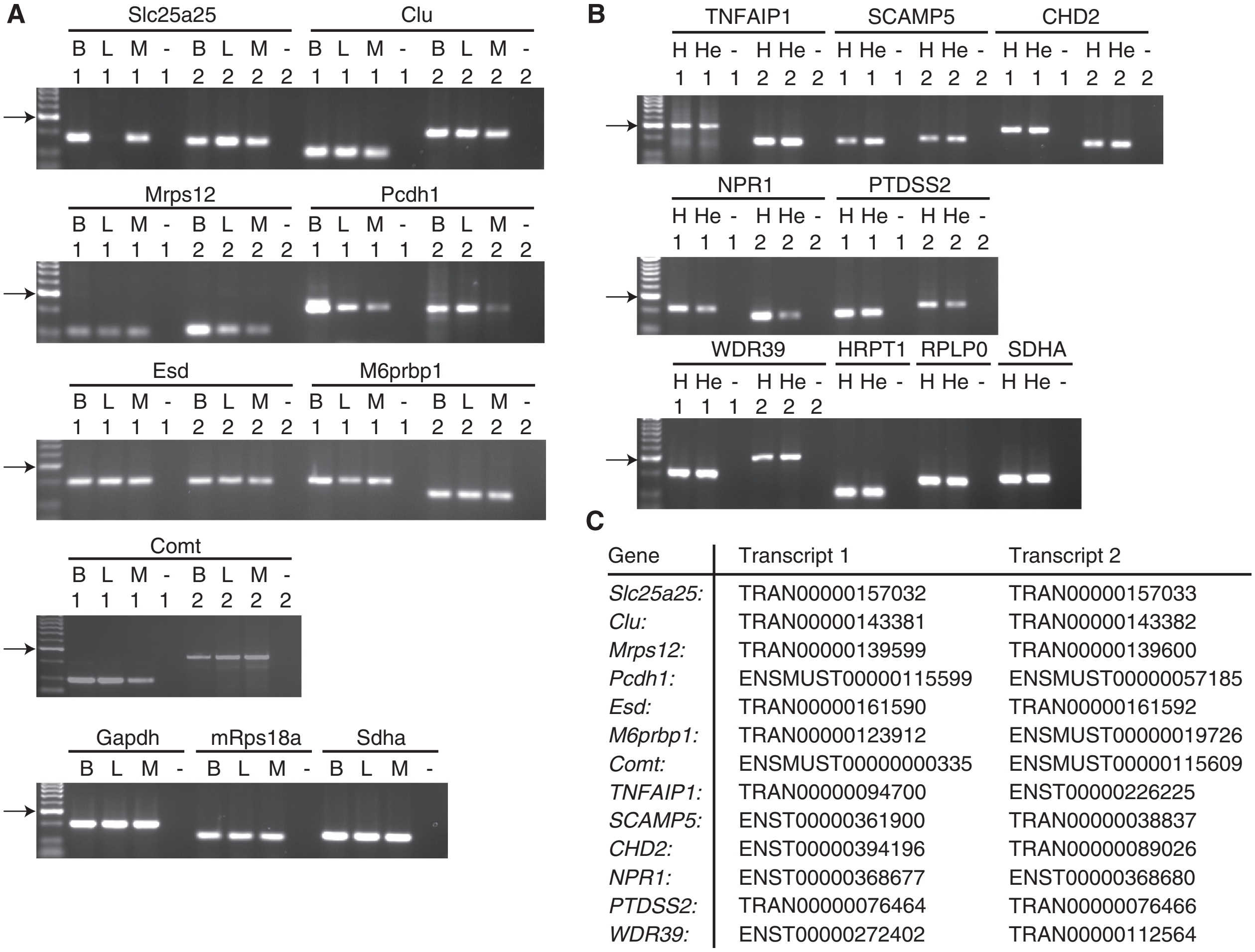


Figure 6. Experimental validation of BASIS prediction. Real time RT–PCR barplots of tested transcripts' relative expression levels between mouse brain and liver (**A**), between mouse brain and muscle (**B**), and between HeLa and HepG2 cells (**C**). Relative expression ratio (condition 1/condition 2) = 1 means no differential expression between two conditions. Relative expression ratio >1 means higher expression in condition 1. Relative expression ratio <1 means higher expression in condition 2. Black bars are transcripts predicted to have higher expression levels in condition 1 by BASIS and white bars are transcripts predicted to have higher expression levels in condition 2. Gray bars are those predicted not to be differentially expressed between two conditions. Value represents mean ± SEM, $N = 3$.



Gene	Transcript 1	Transcript 2
Slc25a25:	TRAN00000157032	TRAN00000157033
Clu:	TRAN00000143381	TRAN00000143382
Mrps12:	TRAN00000139599	TRAN00000139600
Pcdh1:	ENSMUST00000115599	ENSMUST00000057185
Esd:	TRAN00000161590	TRAN00000161592
M6prbp1:	TRAN00000123912	ENSMUST00000019726
Comt:	ENSMUST00000000335	ENSMUST00000115609
TNFAIP1:	TRAN00000094700	ENST00000226225
SCAMP5:	ENST00000361900	TRAN00000038837
CHD2:	ENST00000394196	TRAN00000089026
NPR1:	ENST00000368677	ENST00000368680
PTDSS2:	TRAN00000076464	TRAN00000076466
WDR39:	ENST00000272402	TRAN00000112564

Figure 7. Agarose gel electrophoresis of RT–PCR products. (**A**). Tested transcripts were separated on agarose gels in lanes denoted by B (brain), L (liver), M (muscle) and −(RT negative control). One and two are two different isoforms whose transcript IDs can be referred in (**C**). Gapdh, mRps18a and Sdha are internal control genes. (**B**) Tested transcripts were separated on agarose gels in lanes denoted by H (HeLa cell), He (HepG2 cell) or −(RT negative control). HRPT1, RPLP0 and SDHA are internal control genes. The length of each product is identical to the PCR target region. Arrow points to the brightest DNA size marker 300 bp. Below it, they are 200 bp, 100 bp and 50 bp. Above it, they are 400 bp, 500 bp, 600 bp, 700 bp, 800 bp and 1 kb.

show opposite differential expression patterns compared to the BASIS predictions, their differential ratios are significantly smaller than their counterpart transcript isoform of the same gene. Note that all of the genes except PTDSS2 have their two transcript isoforms showing significantly different relative expression ratios (*P*-values based on Student's *t*-test ≤0.05). The above results suggest that BASIS has a better power when the differential signal is stronger. In addition, BASIS can distinguish the different expression patterns among transcript isoforms of the same genes.

The RT–PCR products of the tested transcripts in different tissues/cells were examined by agarose gel electrophoresis (Figure 7). It shows that the RT–PCR experiments generated the correct amplicons for each transcript under different conditions. In summary, BASIS correctly predicts 22 out of 28 times for the RNA-Seq data and 8 out of 12 times for the tilling array data. BASIS has a relatively better prediction power for RNA-Seq data than tilling array data.

DISCUSSION

Because AS dramatically increases the complexity of eukaryotic transcriptomes, two transcriptomes can be precisely compared only through the expression level of each isoform, but not individual probes or exons, to more accurately deduce gene expression regulation. In this article, we proposed a hierarchical Bayesian model (BASIS) to identify splicing isoforms that are differentially expressed between two conditions. BASIS integrates known splicing information to fully utilize high-density tiling-array or high-throughput RNA-Seq data.

BASIS jointly considers all probes (or positions) targeting the same gene to infer the differential expression level. Sequence read coverage or probe intensity at each position may represent a family of splice variants instead of one single isoform. As shown in Supplementary Figure S1, many transcript isoforms do not contain any isoform-specific sequence positions or isoform-specific exon–exon junctions. Individual probe intensities or sequence reads may not provide direct evidence to distinguish differentially expressed transcript isoforms. BASIS tackles this problem by allocating the intensity of each probe (or sequence read coverage) to multiple transcript isoforms and integrating multiple probe intensities (or sequence read coverage values) for the same gene. Another advantage of jointly considering probes is that a superior signal-to-noise estimate can be achieved by utilizing information from every probe (or sequence read). If the expression intensity is compared probe by probe between two conditions, the high noise level of an individual probe would make the comparison less reliable. However, if we consider the joint behavior of all probes targeting the same gene, the results become much more reliable. In addition, inferences at the transcript isoform level instead of the probe level deliver a more biologically interpretable result.

Second, BASIS accounts for the heteroskedasticity of probe intensity or sequence read coverage and has much higher statistical power than the least squares fit. We gathered together all of the probes (or read coverage over positions) from different genes and divided them into 100 bins. Probes (or positions) within the same bin share the same variance. Therefore, strength could be borrowed across genes in estimating the variance in probe intensity (or the variance in read coverage). This is particularly crucial when there are only a few replicates for each tiling-array or RNA-seq experiment. The approach to binning probes to calculate stable estimates of variances has also been used by Johnson *et al.* (19). In addition, BASIS can be extended to handle the 'large p and small *n*' issue. When the number of potential transcript isoforms is larger than the number of data points available, BASIS maintains the flexibility in statistical inference, whereas the traditional least squares fit requires the number of potential transcript isoforms to be smaller than the number of probes (or positions). Empirical and hierarchical Bayesian approaches have been applied to gene-level microarray analyses in which each gene is represented by one probe and information across different genes are borrowed from each other (26,27).

Third, the latent variable γ was introduced into BASIS in order to perform variable selection. In many biological conditions, only a portion of the transcript isoforms is expressed. The latent variable can directly identify the transcript isoforms of interest and leads to an interpretable model.

Simulation studies show that BASIS has a about 2-fold increase in power compared to the least squares fit (Table 1 and Figure 4). And the power of BASIS is related to gene structure. Specifically, if a gene has more probes (or positions), the power of BASIS is larger (*P*-value for the correlation is 0.0005). If the difference among isoforms is larger, the power of BASIS is larger ($P = 0.0001$). BASIS does not rely on the percentage of isoform-specific positions ($P = 0.29$), and it considers the joint behavior of positions. The model also depends on the completeness of the known splicing patterns of each gene. However the incompleteness does not jeopardize the legitimacy of our model as shown in simulation studies (Table 1). In the real data analysis, using the junction reads as an independent data source, we showed that BASIS has a statistical power of 80–83%. The real time RT–PCR experiments validated 22 out of 28 predictions by BASIS for the RNA-Seq data, and 8 out of 12 predictions for the tilling array data. As information accumulates and novel transcript isoforms are discovered through experiments or isoform reconstruction algorithms (28), a more accurate and complete AS annotation database will further improve results derived from our model. In addition, junction reads from RNA-seq experiments can provide prior information on AS patterns and further improve the signal-to-noise ratio. In the post-genomic era, there is an increasing demand for the complete identification of alternative transcripts and thorough genome annotations. cDNA cloning and/or longer read sequencing (e.g. pair-end sequencing) remain necessary experimental tools for identifying long-range contiguous splice choices.

The predicted tissue-specific transcript isoforms have functional significance. For example, *Slc25a25* is a type of calcium binding mitochondrial ATP-Mg/Pi transporter (29,30). Interestingly, its rat ortholog was found to be expressed much more highly in liver than in brain (31), whereas its human homolog was shown to be expressed much more highly in brain than in liver (29). Such a discrepancy may not be due to species variation. It is likely due to the tissue-specific expression of alternatively spliced variants, as Mashima *et al.* (31) used a probe specific to the liver isoform (TRAN00000157033), consistent with our real time RT–PCR results (Figure 6A). The difference between *Slc25a25* transcripts TRAN00000157032 and TRAN00000157033 occurs at their *N*-termini, which encode calcium-sensitive EF-hand binding motifs. This indicates that the tissue-specific expression of these two transcripts may be related to their differential physiological functions responding to Ca^{2+} signals in different tissues. Another example is represented by the two isoforms of *Mrps12*. Although the *Mrps12* 5′-UTRs are not well conserved between human and mouse in terms of sequence identity, their AS patterns are conserved, indicating the functional importance of such splicing regulation. Indeed, the two human ortholog transcripts of *Mrps12* TRAN00000139599 and TRAN00000139600 are subject to different translational regulation (32), and this could be functionally related to their tissue-specific expression.

BASIS focuses on the direct detection of differentially expressed transcript isoforms between two conditions. Several groups have developed algorithms to estimate transcript abundance, but not difference. The difference in isoform abundances can be conveniently modeled as a normal distributed variable. This is based on the fact that the difference between two normal distributions remains normal (for tiling-array data) and the difference between two Poisson distributions is approximately normal (for sequencing data). The Q–Q plots in Supplementary Figures S2–S4 confirm that the normal distribution is a valid assumption. Shai *et al.* (33) developed the GenASAP algorithm to infer the expression levels of transcript isoforms including or excluding a cassette exon. This was designed specifically for a custom microarray in which an exon-skipping event are represented by three exon body probes and three junction probes (33). If a gene has more than one alternative exon and more than two transcript isoforms consequently, GenASAP cannot distinguish isoforms which all include the tested cassette exon, neither can it further distinguish isoforms which all exclude the tested cassette exon. On the contrary, BASIS can deal with genes with more than two transcript isoforms. Shai *et al.* (33) used a truncated normal distribution ($\beta \geq 0$) to satisfy the non-negative constraint on isoform abundance and maximized the lower bound of the log likelihood instead of the log likelihood itself during their variational EM learning because the exact posterior cannot be computed. Such normal distribution approximation may be inappropriate for RNA-seq data. BASIS focuses on the difference of isoform abundances and the normal approximation for the difference is valid for both the tiling-array data and the RNA-seq data. In GenASAP, the calculation based on the lower bound of the log likelihood may introduce bias in the estimation of isoform abundance. In addition, there was no direct statistical inference for the differential expression patterns and genes are tested separately. However, BASIS performs direct inference on the differentially expressed isoforms and it borrows information from different genes. Anton *et al.* (34) proposed the SPACE algorithm to predict the structures and the abundances of transcript isoforms from microarray data (34). A ‘non-negative matrix factorization’ method was applied to handle the non-negative constraints. The numerical approximation involving non-negative constraints is a computation-intensive task, especially when thousands of genes are considered in a single study. In addition, SPACE has no direct statistical inference for the comparison of two transcriptomes. Anton *et al.* (34) provided the MATLAB code for SPACE. We therefore used SPACE to predict transcript isoform abundances and carried out differential studies by comparing the isoform abundances between two conditions. We performed simulation studies to compare BASIS and SPACE. BASIS has a much higher statistical power than SPACE given the same false positive rate (e.g. 0.87 versus 0.04 when false positive rate is 0.06. See details in Supplementary Data S1 and Supplementary Table S8). The low power of SPACE may be due to the fact that SPACE assumes that the gene structure is un-known and only two experiments (or conditions) were considered. Anton *et al.* (34) reported that the estimation of isoform structure and abundance depends on the number of experiments (34). When there are only a few experiments, the estimation error tends to be high. On the contrary, BASIS utilizes the known isoform structure and borrows information across different genes. It works well even there are only two experiments (or conditions). BASIS focuses on the direct inference of differentially expressed transcript isoforms. The Markov chains generated by the Gibbs sampler converged very quickly and, in theory, the empirical distributions of the hidden variables based on those homogeneous ergodic Markov chains will converge to the actual posterior probabilities (14). BASIS can handle both microarray data and RNA-seq data.

For the RNA-seq data, we used the uniquely mapped reads for each gene and ignored the multireads that can be mapped to multiple positions in the mouse genome. Inclusion and proportionate allocation of multireads have been reported to impact RNA quantification (13). In the present study, we focused on the differential expression patterns of transcript isoforms. Either exclusion or inclusion of the multireads under both conditions has only a small effect on the final results.

An isoform-specific exonic region is needed to accurately assay the expression level of a transcript isoform by real time RT–PCR. Because many transcripts are unique in their exon combinations rather than in isoform-specific exon positions (Supplementary Figure S1), the number of transcripts one can directly test is significantly reduced. Novel experimental techniques are needed in the future to solve this problem. However, through simulation studies, we found that the power of BASIS is not related to the percentage of isoform-specific positions (P-value for the correlation = 0.29). Therefore, the real time RT–PCR validation results on transcripts with isoform-specific positions can still be treated as a fair evaluation of BASIS. We also noted that about 1% genes in our data are un-identifiable because the columns of $\mathbf{X}$ are perfectly collinear (i.e. $\mathbf{X}_j$ is a linear combination of the other columns). For those genes, additional information from other types of experiments is required to infer the differentially expressed isoforms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Michael Waterman (USC) for critical reading of this manuscript and Hongyu Zhao (Yale) for the scientific critiques during the manuscript preparation. We would also like to acknowledge Thomas Gingeras' group (Affymetrix) for providing the cytosolic RNA of HeLa and HepG2 cell line. We would like to specially thank Doug Black (UCLA) for his scientific advice on the study and generous support for the validation experiments.

FUNDING

National Institutes of Health [P50 HG 002790]; the American Federation for Aging Research Grant. Funding for open access charge: University of Southern California.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. 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.
3. Black,D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367–370.
4. Ding,J.H., Xu,X., Yang,D., Chu,P.H., Dalton,N.D., Ye,Z., Yeakley,J.M., Cheng,H., Xiao,R.P., Ross,J. *et al.* (2004) Dilated cardiomyopathy caused by tissue-specific ablation of SC35 in the heart. *EMBO J*., **23**, 885–896.
5. Jumaa,H., Wei,G. and Nielsen,P.J. (1999) Blastocyst formation is blocked in mouse embryos lacking the splicing factor SRp20. *Curr. Biol*., **9**, 899–902.
6. Xu,X., Yang,D., Ding,J.H., Wang,W., Chu,P.H., Dalton,N.D., Wang,H.Y., Bermingham,J.R. Jr, Ye,Z., Liu,F. *et al.* (2005) ASF/SF2-regulated CaMKIIdelta alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle. *Cell*, **120**, 59–72.
7. Jensen,K.B., Dredge,B.K., Stefani,G., Zhong,R., Buckanovich,R.J., Okano,H.J., Yang,Y.Y. and Darnell,R.B. (2000) Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*, **25**, 359–371.
8. Kanadia,R.N., Johnstone,K.A., Mankodi,A., Lungu,C., Thornton,C.A., Esson,D., Timmers,A.M., Hauswirth,W.W. and Swanson,M.S. (2003) A muscleblind knockout model for myotonic dystrophy. *Science*, **302**, 1978–1980.
9. Faustino,N.A. and Cooper,T.A. (2003) Pre-mRNA splicing and human disease. *Genes Dev*., **17**, 419–437.
10. Garcia-Blanco,M.A., Baraniak,A.P. and Lasda,E.L. (2004) Alternative splicing in disease and therapy. *Nat. Biotechnol*., **22**, 535–546.
11. Blencowe,B.J. (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci*., **25**, 106–110.
12. Kapranov,P., Cheng,J., Dike,S., Nix,D.A., Duttagupta,R., Willingham,A.T., Stadler,P.F., Hertel,J., Hackermuller,J., Hofacker,I.L. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.
13. 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.
14. George,E.I. and Mcculloch,R.E. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc*., **88**, 881–889.
15. Chipman,H., George,E.I. and Mcculloch,R.E. (2001) The practical implementation of Bayesian model selection. *IMS Lect. Notes Monogr. Ser*., **38**, 67–131.
16. George,E.I. and McCulloch,R.E. (1997) Approaches for Bayesian variable selection. *Stat. Sinica*., **7**, 339–373.
17. Barbieri,M.M. and Berger,J.O. (2004) Optimal predictive model selection. *Ann. Stat*., **32**, 870–897.
18. Stamm,S., Riethoven,J.J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res*., **34**, D46–D55.
19. Johnson,W.E., Li,W., Meyer,C.A., Gottardo,R., Carroll,J.S., Brown,M. and Liu,X.S. (2006) Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl Acad. Sci. USA*, **103**, 12457–12462.
20. Kapur,K., Xing,Y., Ouyang,Z. and Wong,W.H. (2007) Exon arrays provide accurate assessments of gene expression. *Genome Biol*., **8**, R82.
21. Chen,L. and Zheng,S. (2008) Identify alternative splicing events based on position-specific evolutionary conservation. *PLoS ONE*, **3**, e2806.
22. Boutz,P.L., Chawla,G., Stoilov,P. and Black,D.L. (2007) MicroRNAs regulate the expression of the alternative splicing factor nPTB during muscle development. *Genes Dev*., **21**, 71–84.
23. Vandesompele,J., De Preter,K., Pattyn,F., Poppe,B., Van Roy,N., De Paepe,A. and Speleman,F. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*., **3**, RESEARCH0034.
24. Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol*., **8**, 557–569.
25. Plummer,M., Best,N., Cowles,K. and Vines,K. (2006) CODA: convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11.
26. Lonnstedt,I. and Speed,T. (2002) Replicated microarray data. *Stat. Sinica*., **12**, 31–46.
27. Nott,D.J., Yu,Z.M., Chan,E., Cotsapas,C., Cowley,M.J., Pulvers,J., Williams,R. and Little,P. (2007) Hierarchical Bayes variable selection and microarray experiments. *J. Multivariate Anal*., **98**, 852–872.
28. Xing,Y., Yu,T., Wu,Y.N., Roy,M., Kim,J. and Lee,C. (2006) An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res*., **34**, 3150–3160.
29. Fiermonte,G., De Leonardis,F., Todisco,S., Palmieri,L., Lasorsa,F.M. and Palmieri,F. (2004) Identification of the human mitochondrial ATP-Mg/Pi transporter. *Bba-Bioenergetics*, **1658**, 191.
30. del Arco,A. and Satrustegui,J. (2004) Identification of a novel human subfamily of mitochondrial carriers with calcium-binding domains. *J. Biol. Chem*., **279**, 24701–24713.
31. Mashima,H., Ueda,N., Ohno,H., Suzuki,J. and Omata,M. (2003) A novel mitochondrial Ca2+-dependent solute carrier in the liver identified by mRNA differential display. *Gastroenterology*, **124**, A127.
32. Mariottini,P., Shah,Z.H., Toivonen,J.M., Bagni,C., Spelbrink,J.N., Amaldi,F. and Jacobs,H.T. (1999) Expression of the gene for mitoribosomal protein S12 is controlled in human cells at the levels of transcription, RNA splicing, and translation. *J. Biol. Chem*., **274**, 31853–31862.
33. Shai,O., Morris,Q.D., Blencowe,B.J. and Frey,B.J. (2006) Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics*, **22**, 606–613.
34. Anton,M.A., Gorostiaga,D., Guruceaga,E., Segura,V., Carmona-Saez,P., Pascual-Montano,A., Pio,R., Montuenga,L.M. and Rubio,A. (2008) SPACE: an algorithm to predict and quantify alternatively spliced isoforms using microarrays. *Genome Biol*., **9**, R46.