Published online 20 January 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 8* ***e61***
doi:10.1093/nar/gkr1291

MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data

Shihao Shen[1], Juw Won Park[2], Jian Huang[3], Kimberly A. Dittmar[4], Zhi-xiang Lu[2], Qing Zhou[5], Russ P. Carstens[4,6] and Yi Xing[2,7,*]

[1]Department of Biostatistics, [2]Department of Internal Medicine, [3]Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, [4]Renal Division, Department of Medicine, University of Pennsylvania, School of Medicine, Philadelphia, PA 19104, [5]Department of Statistics, University of California, Los Angeles, CA 90095, USA, [6]Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104 and [7]Department of Biomedical Engineering, University of Iowa, Iowa City, IA 52242

Received August 6, 2011; Revised December 10, 2011; Accepted December 15, 2011

ABSTRACT

Ultra-deep RNA sequencing has become a powerful approach for genome-wide analysis of pre-mRNA alternative splicing. We develop MATS (multivariate analysis of transcript splicing), a Bayesian statistical framework for flexible hypothesis testing of differential alternative splicing patterns on RNA-Seq data. MATS uses a multivariate uniform prior to model the between-sample correlation in exon splicing patterns, and a Markov chain Monte Carlo (MCMC) method coupled with a simulation-based adaptive sampling procedure to calculate the *P*-value and false discovery rate (FDR) of differential alternative splicing. Importantly, the MATS approach is applicable to almost any type of null hypotheses of interest, providing the flexibility to identify differential alternative splicing events that match a given user-defined pattern. We evaluated the performance of MATS using simulated and real RNA-Seq data sets. In the RNA-Seq analysis of alternative splicing events regulated by the epithelial-specific splicing factor ESRP1, we obtained a high RT–PCR validation rate of 86% for differential exon skipping events with a MATS FDR of $<$10%. Additionally, over the full list of RT–PCR tested exons, the MATS FDR estimates matched well with the experimental validation rate. Our results demonstrate that MATS is an effective and flexible approach for detecting differential alternative splicing from RNA-Seq data.

INTRODUCTION

Alternative splicing plays critical roles in regulating gene function and activity in higher eukaryotes (1). By alternative selection of exons and splice sites during splicing, a single gene locus can produce multiple mRNA and protein isoforms with divergent structural and functional properties (2). Over 90% of multi-exon human genes are alternatively spliced and many genes undergo changes in alternative splicing during development, cell differentiation and disease (3–6). Changes in alternative splicing patterns can be manifested as all-or-none switches between distinct mRNA isoforms, but more frequently as shifts in the relative abundance of multiple mRNA isoforms of a gene. In some disease genes, a slight shift (by as few as several percent) in the relative isoform proportions can trigger disease pathogenesis (7,8). Due to the prevalent role of alternative splicing in gene regulation and disease, there is a pressing need for genomic tools that can accurately quantify isoform ratios and reliably detect changes in isoform ratios (i.e. differential alternative splicing) among different biological conditions.

Direct sequencing of full-length mRNAs and mRNA fragments has been a popular and effective approach for discovery and quantification of alternative splicing events (3,4,9). Since mRNAs represent the end products of splicing, by aligning mRNA sequences to the genome one can delineate exon–intron structures and identify alternative splicing events. If sequencing reaches the sufficient depth, such that the relative abundance of distinct mRNA isoforms can be confidently estimated by the numbers of RNA sequences matching to specific isoforms, we can compare the mRNA sequence counts across biological conditions to identify differential alternative

*To whom correspondence should be addressed. Tel: +1 319 384 3099; Fax: +1 319 384 3150; Email: yi-xing@uiowa.edu splicing events. This approach was first adopted for the discovery of tissue-specific exons from expressed sequence tags (ESTs) (10). Xu *et al*. categorized the ESTs of human genes according to their tissue origins. The exon inclusion level of an alternatively spliced cassette exon in any given tissue was estimated from the counts of ESTs mapped uniquely to the exon inclusion or skipping exon-exon junctions (10) (for a formal definition, see 'Materials and Methods' section). Using a Bayesian approach, Xu *et al.* compared the EST counts across different tissues to detect exons with tissue-specific shifts in exon inclusion levels. Specifically, the prior distribution of an exon's exon inclusion levels in two tissues was modeled as two independent uniform (0, 1) distributions. The EST read count of the exon inclusion isoform in each tissue was assumed to follow a binomial distribution. A Bayesian posterior probability was calculated to assess whether the exon inclusion levels of an exon differed between two tissues. Similar approaches were later used to identify cancer-specific alternative splicing events (11,12). These studies pioneered the use of RNA sequence count data for quantitative splicing analysis. However, due to the low throughput of EST sequencing, EST-based analysis of differential alternative splicing has limited capacity and high noise (13).

Recently, with the advent of the high-throughput RNA sequencing technology (RNA-Seq), it has become feasible to generate hundreds of millions of short RNA-Seq reads on any RNA sample of interest (14). This technology enables genome-wide quantitative analyses of RNA alternative splicing (3,4). Pan and colleagues demonstrated that if the RNA-Seq coverage of an alternatively spliced cassette exon reaches at least 20 reads for one of its exon–exon junctions, the exon inclusion levels estimated by RNA-Seq strongly correlate with splicing microarray measurements (4). Other studies comparing RNA-Seq data to RT–PCR data reached similar conclusions (15–17). Thus, by analyzing and comparing deep RNA-Seq data from different biological conditions, one can identify exons with changes in exon inclusion levels on a genome scale. From the RNA sequence count data, differential alternative splicing events are commonly identified by testing the equality of transcript isoform ratios between samples (11,15,16,18–20). Various methods have been used to assess the statistical significance of such differential alternative splicing events, including Fisher exact test of isoform-specific read counts (15,19,20), and Bayesian approaches that model read counts as sampled from a probabilistic mixture of distinct isoforms (11,16,21).

In this article, we introduce MATS (multivariate analysis of transcript splicing), a Bayesian statistical framework for flexible hypothesis testing of differential alternative splicing patterns on RNA-Seq data. Compared to previous computational methods for detecting differential alternative splicing events from RNA sequence count data, MATS has several novel features. First and most importantly, MATS offers the flexibility to identify differential alternative splicing events that match a given user-defined pattern. For example, MATS can calculate the statistical significance that the absolute difference in the exon inclusion levels of an exon between two conditions exceeds a given threshold (e.g. 10%). This allows biologists to identify alternative splicing changes that reach any specified magnitude. MATS can also be used to detect exons with the extreme 'switch-like' differential alternative splicing pattern, i.e. exons predominantly included in the transcripts in one condition but predominantly skipped in another condition. This switch-like pattern is of considerable biological interest, because it is a strong indicator of functional alternative splicing events (3,22). Second, MATS uses a multivariate uniform distribution as the joint prior for the exon inclusion levels in two conditions. Compared to the independent uniform priors commonly used by previous methods, the multivariate uniform prior is more general and better captures the genome-wide similarity in exon splicing patterns between biological conditions. Of note, this prior distribution is motivated by the observation that between any two conditions there is generally an overall positive correlation in exon splicing patterns, and only a small percentage of alternatively spliced exons undergo differential splicing (see 'Results' section). Finally, MATS employs a Markov chain Monte Carlo (MCMC) method coupled with a simulation-based adaptive sampling procedure to calculate the *P*-value and false discovery rate (FDR) of differential alternative splicing, by comparing the posterior probability of the observed data to a set of posterior probabilities simulated from the null hypothesis. This approach is applicable to almost any type of null hypotheses of interest. To evaluate the performance of MATS, we analyzed a deep RNA-Seq data set with 256 million reads on a human breast cancer cell line (MDA-MB-231) with ectopic expression of the epithelial-specific splicing factor ESRP1 or an empty vector (EV) control. In this experimental system, the splicing factor ESRP1 induces large-scale changes in alternative splicing towards an epithelial-specific splicing signature (23). Based on the MATS result, we selected 164 exons that covered a broad range of FDR values for RT–PCR validation. For exons with a MATS FDR of <10%, we obtained a high overall RT–PCR validation rate of 86%, demonstrating that MATS can reliably detect differential alternative splicing events. Additionally, over the full list of RT–PCR tested exons, we observed a progressive decrease in the RT–PCR validation rate with increasing FDR values, suggesting MATS can yield experimentally meaningful FDR estimates to help biologists interpret RNA-Seq predictions and design follow-up experiments.

MATERIALS AND METHODS

Overview of MATS

MATS is a Bayesian statistical framework to detect differential alternative splicing events from RNA-Seq data. For every alternatively spliced cassette exon, MATS assesses the statistical significance of differential alternative splicing based on a user-defined hypothesis. The default analysis of MATS is to test whether the difference in the exon inclusion levels between two samples exceeds a given user-defined threshold (e.g. 10%). Compared to existing RNA-Seq analysis methods that test the *equality* of exon inclusion levels between samples, the MATS test of splicing *difference* has three advantages. First, it provides a rigorous statistical framework for biologists to identify alternative splicing changes that reach any specified magnitude. Second, it improves robustness against random sampling noise in the RNA-Seq data, which could cause a minor shift in the estimated isoform ratio between samples. For exons with high RNA-Seq read counts such as those from highly expressed genes, such random noise might introduce false positives to a test of equality on exon inclusion levels. Third, the flexible hypothesis formulation also allows testing of other types of differential alternative splicing behavior such as the 'switch-like' pattern, in which an exon is predominantly included in the transcripts in one condition but predominantly skipped in another condition.

The major steps of MATS are illustrated schematically in Figure 1. First, for each exon MATS uses the counts of RNA-Seq reads mapped to the exon-exon junctions of its inclusion or skipping isoform to estimate the exon inclusion levels in two samples (Figure 1A). Second, the exon inclusion levels of all alternatively spliced cassette exons are used to construct a multivariate uniform prior that models the overall similarity in alternative splicing profiles between the two samples (Figure 1B). Third, based on the multivariate uniform prior and a binomial likelihood model for the RNA-Seq read counts of the exon inclusion/skipping isoforms, MATS uses a MCMC method to calculate the Bayesian posterior probability for splicing difference. Under the default setting, MATS calculates the posterior probability that the change in the exon inclusion level of a given exon exceeds a given user-defined threshold (e.g. 10%; Figure 1C). Finally, MATS calculates a *P*-value for each exon by comparing the observed posterior probability (from Step C) with a set of simulated posterior probabilities from the null hypothesis (Figure 1D). These *P*-values are then transformed to FDR values by the Benjamini–Hochberg FDR method (24). Below we describe the details of the MATS algorithm.

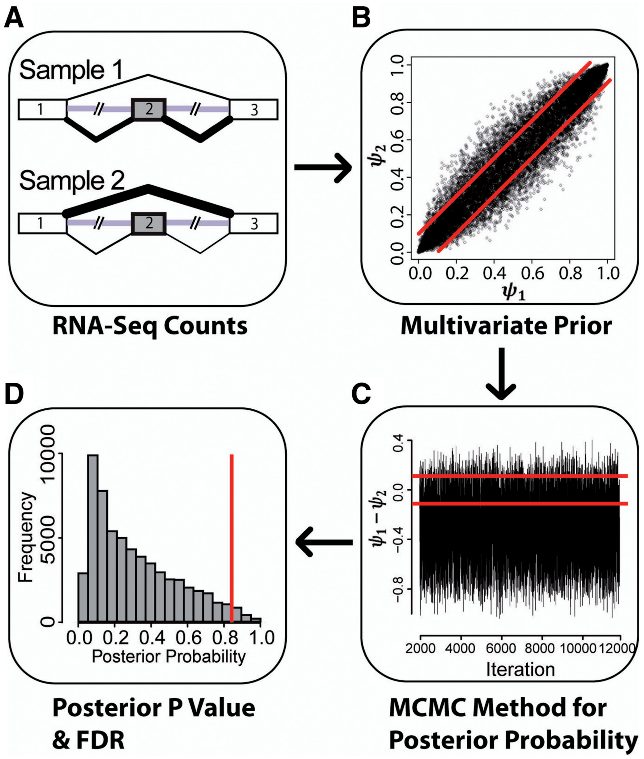

Figure 1. Basic steps of MATS. (**A**) For each exon MATS uses the counts of RNA-Seq reads mapped to the exon–exon junctions of its inclusion or skipping isoform to estimate the exon inclusion levels in two samples. (**B**) The exon inclusion levels of all alternatively spliced cassette exons are used to construct a multivariate uniform prior that models the overall similarity in alternative splicing profiles between the two samples. (**C**) Based on the multivariate uniform prior and a binomial likelihood model for the RNA-Seq read counts of the exon inclusion/skipping isoforms, MATS uses a Markov chain Monte Carlo (MCMC) method to calculate the Bayesian posterior probability for splicing difference. (**D**) MATS calculates a *P*-value for each exon by comparing the observed posterior probability with a set of simulated posterior probabilities from the null hypothesis, followed by adjustment for multiple testing to obtain the FDR value.

Estimating exon inclusion levels

We define the exon inclusion level (ψ) of an alternatively spliced cassette exon as the percentage of 'exon inclusion' transcripts that splice from its upstream exon into the cassette exon then into its downstream exon among all such 'exon inclusion' transcripts plus 'exon skipping' transcripts that splice from its upstream exon directly into its downstream exon (16,17,25). In a RNA-Seq study, for each exon in a given sample we count the number of RNA-Seq reads uniquely mapped to its upstream, downstream or skipping exon–exon junctions (Figure 1A). The upstream junction count (UJC) and the downstream junction count (DJC) reflect the abundance of the exon inclusion isoform, while the skipping junction count (SJC) reflects the abundance of the exon skipping isoform. Let *I* and *S* represent the counts of exon inclusion and skipping isoforms respectively. Assuming that the read counts follow a binomial distribution, the maximum likelihood estimate (MLE) of the exon inclusion level (ψ) of an exon in a given sample can be calculated as:

$$\hat{\psi} = I/(I{+}S) = \frac{(UJC{+}DJC)/2}{(UJC{+}DJC)/2{+}SJC}$$

Calculating the Bayesian posterior probability for differential alternative splicing

To compare alternative splicing patterns between two samples, for each exon we define ψ_1 and ψ_2 as its exon inclusion levels in sample 1 and 2. Under the default setting, MATS tests the hypothesis that the difference in the exon inclusion levels of a given exon between sample 1 and 2 is above a user-defined cutoff *c*, i.e. $|\psi_1 - \psi_2| > c$. The cutoff *c* is a user-defined parameter that represents the extent of splicing change one wishes to identify. For example, if a researcher is interested in identifying exons with at least 10% change in exon inclusion levels, the cutoff *c* should be set as 10%. The values of ψ_1 and ψ_2 under the null hypothesis (H_0) and the alternative hypothesis (H_1) of this test are illustrated graphically in

Figure 2A. In this section, we describe how MATS calculates the posterior probability of $|\psi_1 - \psi_2| > c$ from the RNA-Seq counts, i.e. $P(|\psi_1 - \psi_2| > c|\text{Data})$. The *P*-value and FDR calculation is described in the next section.

To calculate the posterior probability of $|\psi_1 - \psi_2| > c$, we need to define the prior probability and the likelihood model. In MATS, the joint prior distribution of ψ_1 and ψ_2 is modeled as a multivariate uniform distribution (Figure 1B), with marginal distributions as uniform (0, 1) and correlation $\rho \sim$ uniform (0, 1):

$$\text{Prior}: \begin{array}{l}(\psi_1,\psi_2) \sim MultiVarUniform\left(0,1,cor=\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)\\ \rho \sim Uniform(0,1)\end{array}$$

The multivariate uniform distribution was obtained by applying cumulative standard normal distribution functions to a random vector that follows a multivariate normal distribution. Specifically, $(\psi_1,\psi_2) = (\Phi(Z_1), \Phi(Z_2))$, where Z_1 and Z_2 are standard normal random variables $N(0,1)$ with correlation ρ and Φ is the cumulative distribution function of the standard normal distribution. The obtained multivariate uniform distribution is equivalent to a bivariate distribution with uniform marginals.

We note that our choice of prior distribution in MATS differs from previous methods which model the priors of ψ_1 and ψ_2 as two independent uniform distributions (11,16). This multivariate uniform prior distribution of ψ_1 and ψ_2 is motivated by the observation that between any two biological conditions, there is generally an overall similarity in the genome-wide alternative splicing profiles, and only a small percentage of alternatively spliced exons undergo differential splicing. Indeed, our analysis of several RNA-Seq data sets suggests that this multivariate uniform prior provides a good fit with empirical data (see 'Results' section). In contrast, the commonly used independent uniform prior distributions assume that the splicing activities of the same exon in two different samples are independent, even if these two samples have the identical cell type and tissue origin. This lacks biological justification and fits empirical data poorly.

In each sample, the exon inclusion isoform count *I* follows a binomial distribution with $n = I{+}S$ and $p = \psi$, where *S* is the skipping isoform count.

Assume we have a total of *N* alternatively spliced cassette exons, for each exon $i = 1,...,N$, we denote:

ψ_{i1},ψ_{i2}: exon inclusion levels of exon *i* in sample 1 and 2;

I_{i1},I_{i2}: counts of the exon inclusion isoform of exon *i* in sample 1 and 2;

S_{i1},S_{i2}: counts of the exon skipping isoform of exon *i* in sample 1 and 2.

$$\begin{array}{rl}\text{Likelihood}: I_{i1}|\psi_{i1} & \sim \text{Binomial}\ (n_{i1} = I_{i1}+S_{i1}, p_{i1} = \psi_{i1})\\ I_{i2}|\psi_{i2} & \sim \text{Binomial}\ (n_{i2} = I_{i2}+S_{i2}, p_{i2} = \psi_{i2})\end{array}$$

The posterior probability of differential alternative splicing for exon *i* can be calculated as $P_i = P(|\psi_{i1} - \psi_{i2}| > c|I_{i1},I_{i2},S_{i1},S_{i2},I_{[-i]1},I_{[-i]2},S_{[-i]1},S_{[-i]2})$, where the counts of the exon inclusion/skipping isoforms of exon *i* and all other alternatively spliced cassette exons (indexed by $[-i]$) are used to infer the parameter ρ in the multivariate uniform prior as well as ψ_{i1} and ψ_{i2}, and *c* represents the user-defined threshold for splicing change.

As this posterior probability cannot be calculated with an analytic solution in closed-form, we adopt a numeric solution based on the MCMC method. Specifically, the posterior probability is calculated by the JAGS program (Just Another Gibbs Sampler; http://sourceforge.net/projects/mcmc-jags/). This program estimates the posterior probability $P(|\psi_{i1} - \psi_{i2}| > c|I_{i1},I_{i2},S_{i1},S_{i2},I_{[-i]1},I_{[-i]2},S_{[-i]1},S_{[-i]2})$ for all exons and the parameter ρ of the multivariate uniform prior. The parameter ρ is a global parameter for all exons, which describes the overall correlation of the exon inclusion levels of all alternatively spliced exons between two samples. Therefore, for *N* exons there are a total of $2N{+}1$ parameters, including $2N$ parameters for exon inclusion levels in two samples and the

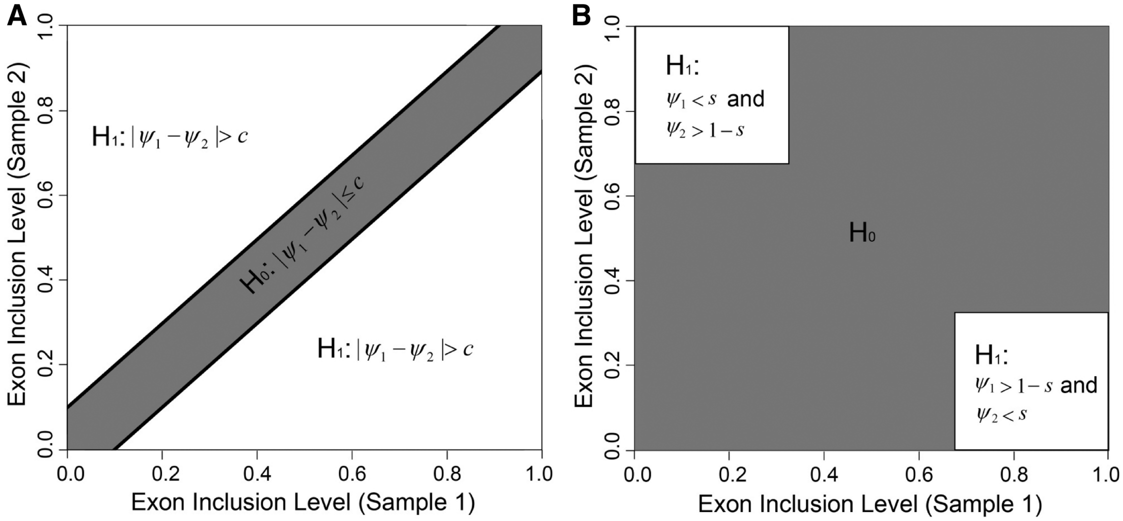

Figure 2. Null hypotheses in MATS. (**A**) Under the default setting of MATS, the H_1 alternative hypothesis is that the difference in the exon inclusion levels between two samples is above the user-defined cutoff *c* (the white area). The H_0 null hypothesis is that the difference is below the user-defined cutoff *c* (the gray area). (**B**) MATS can test the extreme 'switch-like' differential alternative splicing pattern with a different hypothesis. The H_1 alternative hypothesis is that the exon inclusion level is below a user-defined threshold *s* in sample 1 and above 1-*s* in sample 2, or vice versa (the white area). The H_0 null hypothesis is outside the alternative hypothesis region (the gray area).

global parameter ρ. To estimate the global parameter ρ along with all the ψ values, the data of all N exons are used as the input for the MCMC sampler. We burn in the MCMC sampler for 2000 iterations, followed by another 10 000 iterations to calculate the posterior probabilities. The posterior probability of a given exon i (denoted as P_i^{obs}) is estimated by the fraction of iterations with $|\psi_{i1} - \psi_{i2}| \geq c$ among all 10 000 iterations (Figure 1C).

Calculating the *P*-value and FDR for differential alternative splicing

We use a simulation-based adaptive sampling procedure to calculate the P-value and FDR for differential alternative splicing. In theory, P-value comes from the comparison of the observed test statistics with statistics from the null hypothesis. In MATS, when we test $|\psi_1 - \psi_2| > c$, we consider the null hypothesis that $|\psi_1 - \psi_2| \leq c$. We calculate the P-value of each exon by comparing the posterior probability of the observed data (P_i^{obs}) to a set of posterior probabilities simulated from the null hypothesis. For each exon, we find the maximum likelihood estimate (MLE) of the constrained ψ_1 and ψ_2 (denoted as $\hat{\psi}_1^{\mathrm{c}}$ and $\hat{\psi}_2^{\mathrm{c}}$) from the binomial distributions with the counts I_1, I_2, S_1 and S_2 in two conditions, subject to the constraint that $|\psi_1 - \psi_2| \leq c$. Specifically,

$$(\hat{\psi}_1^c, \hat{\psi}_2^c) = \underset{|\psi_1 - \psi_2| \leq c}{\arg\max}(I_1 \log \psi_1 + S_1 \log(1 - \psi_1) + I_2 \log \psi_2 + S_2 \log(1 - \psi_2))$$

The limited-memory Broyden–Fletcher–Goldfarb–Shanno box-constraints (L-BFGS-B) algorithm is used to search for the constrained MLE (26,27). Then we simulate RNA-Seq count data from the constrained MLE of $\hat{\psi}_1^c$ and $\hat{\psi}_2^c$, and calculate the posterior probability of $|\psi_1 - \psi_2| > c$ given the simulated data. The P-value of each exon is calculated by comparing the posterior probability of differential splicing based on the observed data to a set of simulated posterior probabilities. The details of this calculation for a given exon i are summarized below:

(1) Retrieve the estimated global parameter $\hat{\rho}$ from the MCMC calculation of posterior probabilities of all alternatively spliced exons. The value of $\hat{\rho}$ is fixed in the P-value calculation.

(2) For exon i, retrieve the observed posterior probability from the MCMC calculation $P_i^{\mathrm{obs}} = P(|\psi_{i1} - \psi_{i2}| > c | I_{i1}, I_{i2}, S_{i1}, S_{i2}, I_{[-i]1}, I_{[-i]2}, S_{[-i]1}, S_{[-i]2})$.

(3) For exon i, simulate M posterior probabilities from the constrained MLE of $\hat{\psi}_{i1}^{\mathrm{c}}$ and $\hat{\psi}_{i2}^{\mathrm{c}}$. For $j = 1,\ldots,M$:

i) Simulate data

$$I_{i1j} | \hat{\psi}_{i1}^{\mathrm{c}} \sim \text{Binomial } (n_{i1} = I_{i1} + S_{i1}, p_{i1} = \hat{\psi}_{i1}^{\mathrm{c}}),$$
$$S_{i1j} = n_{i1} - I_{i1j}$$
$$I_{i2j} | \hat{\psi}_{i2}^{\mathrm{c}} \sim \text{Binomial } (n_{i2} = I_{i2} + S_{i2}, p_{i2} = \hat{\psi}_{i2}^{\mathrm{c}}),$$
$$S_{i2j} = n_{i2} - I_{i2j}$$

ii) Calculate the posterior probability from the simulated data $I_{i1j}, I_{i2j}, S_{i1j}, S_{i2j}$ using the MCMC method as $P_{ij}^{\mathrm{sim}} = P(|\psi_{i1j} - \psi_{i2j}| > c | I_{i1j}, I_{i2j}, S_{i1j}, S_{i2j}, \hat{\rho})$.

iii) Calculate the P-value for exon i by comparing P_i^{obs} with the simulated $\{P_{ij}^{\mathrm{sim}}\}$ as $\frac{1}{M}\sum_{j=1}^{M} I(P_i^{\mathrm{obs}} \leq P_{ij}^{\mathrm{sim}})$.

For each exon, the number of M is determined by an adaptive sampling procedure (see below).

The number of simulations (M) in calculating the simulated posterior probabilities is determined by an adaptive sampling procedure. Initially, we aim to reach a P-value precision of 0.01 by setting $M = 100$. One hundred simulated posterior probabilities are calculated for each exon, and the exon's P-value is generated by comparing the observed posterior probability to the simulated ones. A zero or close-to-zero P-value for any exon indicates that the number of simulations is insufficient for generating a reliable P-value estimate. For all exons with a P-value of smaller than three times the precision, the number of simulations is increased by 10-fold in a new round of simulation, which increases the precision of the P-value estimate from 0.01 to 0.001. This adaptive sampling procedure is repeated for multiple rounds. The default setting of MATS is to repeat this procedure for at most six rounds to reach the highest P-value precision of 10^{-7}, but this parameter can be adjusted by users. This adaptive sampling procedure enables us to selectively increase the precision and running time for exons with significant (i.e. small) P-values, thus reducing the overall running time needed for all exons.

After we obtain the P-values of all exons, we apply the classic Benjamini–Hochberg method (24) on these P-values to obtain the FDR values.

Detecting switch-like differential alternative splicing

MATS offers the flexibility for testing different types of hypotheses on the differential alternative splicing pattern. An analysis of considerable biological interest is to identify exons with the extreme 'switch-like' differential alternative splicing pattern, i.e. ($\psi_1 < s$ and $\psi_2 > 1 - s$) or ($\psi_1 > 1 - s$ and $\psi_2 < s$) where s is a user-defined parameter between 0 and 1/2. For example, if we set s as 1/3, we test if the exon inclusion level of an exon switches from less than 1/3 in one sample to more than 2/3 in the other sample. Such an extreme switch of exon inclusion levels between conditions is a strong indicator of functional alternative splicing events (3,22). In the 'switch-like' test, MATS considers the null hypothesis that the ψ_1 and ψ_2 are outside of the region defined by ($\psi_1 < s$ and $\psi_2 > 1 - s$) or ($\psi_1 > 1 - s$ and $\psi_2 < s$). The values of ψ_1 and ψ_2 under the null hypothesis (H_0) and the alternative hypothesis (H_1) for the 'switch-like' test are illustrated graphically in Figure 2B. The same MCMC and simulation procedures for testing $|\psi_1 - \psi_2| > c$ can be used to calculate the Bayesian posterior probability, P-value, and FDR for 'switch-like' differential alternative splicing.

Exon–exon junction database of human genes

We constructed a database of exon–exon junctions in human genes using the Ensembl transcript annotations (release 57) (28). The database includes all known exon–exon junctions observed in Ensembl transcripts, as well as hypothetical exon-exon junctions obtained by all possible pairwise fusions of exons within genes. In total, the database contains ~3.5 million exon–exon junctions. Each exon–exon junction sequence is 84 bp long with 42 bp from the 3′-end of the upstream exon and 42 bp from the 5′-end of the downstream exon. This exon–exon junction database is available for download from the MATS website http://intron.healthcare.uiowa.edu/MATS/.

RNA-Seq analysis of ESRP1 regulated differential alternative splicing events in the MDA-MB-231 breast cancer cell line

MDA-MB-231 cells were maintained and retrovirally transduced by a cDNA encoding the epithelial-specific splicing factor ESRP1 or the empty vector (EV) control as described previously (23,29). Sequencing libraries were prepared using the mRNA-Seq Sample Prep Kits (Illumina) according to the manufacturer's instructions. Ten micrograms of total RNA was used to prepare polyA RNA for fragmentation followed by cDNA synthesis with random hexamers and ligation to Illumina adaptor sequences. The samples underwent an RNA quality assurance check and were quantified using an Agilent 2100 Bioanalyzer, loaded onto flow-cells for cluster generation, and sequenced on an Illumina Genome Analyzer II using single-end protocol to generate 76 bp reads (Illumina). The resulting RNA-Seq dataset consisted of 256 million single-end reads, including 136 million reads for the ESRP1 sample and 120 million reads for the EV sample.

During the quality assessment of our 76 bp single-end RNA-Seq data, we found that the first two 25 bp segments of these reads had a high mapping rate to the human genome, while the 3rd 25 bp segment had a much lower mapping rate (data not shown). This is likely due to the increased sequencing error rate near the 3′-end of the Illumina RNA-Seq reads. Thus, we decided to use the first 50 bp of each read for mapping and subsequent analysis. We mapped RNA-Seq reads to the human genome (hg19) and the exon–exon junction database, using the software Bowtie (30) allowing up to 3 bp mismatches. Each mapped exon–exon junction read required at least 8 bp from each side of the exon–exon junction. We further removed exon–exon junction reads that mapped to either the human genome (hg19) or multiple exon–exon junctions. For all identified alternatively spliced cassette exons, the exon–exon junction counts were used as the input for MATS.

Illumina Human Body Map 2.0 data on 16 human tissues

We obtained a paired-end RNA-Seq data set from Illumina, with ~80 million read pairs per tissue for 16 human tissues (adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid and white blood cells). This data set was referred to as the 'Human Body Map 2.0' by Illumina. For each paired-end read (50 bp × 2), we mapped each end to the human genome (hg19) and the exon–exon junction database, using the software Bowtie (30) allowing up to 3 bp mismatches. The final mapping location of the paired-end read was determined by requiring that the two ends be mapped in the opposite orientation to the same genomic region, with a maximum genomic distance of 50 kb between the two ends (to allow introns between the two mapped ends).

RT–PCR validation

Quantitation of alternative splicing was performed using standard RT–PCR incorporating radiolabeled dCTP or fluorescently labeled primers, or high-throughput RT–PCR at the Université de Sherbrooke as described (23,31). Since we used MATS to test $|\psi_1 - \psi_2| > 0.1$, we defined a differential alternative splicing event as validated if the RT–PCR-based exon inclusion levels differed by at least 10% between the two samples with the direction of change matching the RNA-Seq prediction.

RESULTS

Multivariate uniform prior in MATS

MATS uses a multivariate uniform distribution to model the joint prior of exon inclusion levels of alternatively spliced cassette exons in two samples. This is different from and more general than the independent uniform priors used by previous methods (11,16). Note that the multivariate uniform prior includes the independent uniform prior model as a special case ($\rho = 0$). These two types of prior distributions have distinct underlying assumptions on the alternative splicing patterns of different biological conditions. Intuitively, the multivariate uniform prior allows the modeling of similarity in alternative splicing patterns between two samples (using the correlation parameter ρ). In contrast, the independent uniform priors assume that the global splicing pattern of one sample is independent of the other sample. To determine whether the multivariate uniform prior is appropriate and able to capture the correlation pattern in the data, we analyzed two RNA-Seq data sets covering diverse tissues and cell types.

We first compared the alternative splicing profiles of a single cell line subject to two different treatments. The data set came from our deep single-end RNA sequencing of a human breast cancer cell line (MDA-MB-231) with ectopic expression of the epithelial-specific splicing factor ESRP1 or an empty vector (EV) control (see 'Materials and Methods' section). ESRP1 encodes a master cell-type specific regulator of alternative splicing that controls a global epithelial-specific splicing network (23,29). In the MDA-MB-231 cell line, the ectopic expression of ESRP1 drives coordinated switches of ESRP1-regulated exons towards an epithelial splicing signature (23). This provides an excellent system for testing algorithms of alternative splicing analysis. We generated 136 million single-end reads on the ESRP1 sample and 120 million single-end reads on the EV sample. We identified a total of 18 859 alternatively spliced cassette exons in this data set. Pan and colleagues previously demonstrated that RNA-Seq can reliably estimate the exon inclusion levels of alternatively spliced exons, when the sequencing coverage reaches at least 20 reads for one of the three exon-exon junctions (4). Therefore, to assess the global correlation in alternative splicing patterns between these two samples (ESRP1 and EV), we restricted our analysis to 12 890 alternatively spliced cassette exons with at least 20 reads mapped to one of the three exon-exon junctions in both samples. We observed a high correlation in the exon inclusion levels of these exons between the ESRP1 and EV samples (Pearson correlation $r = 0.95$, $P < 2.2e\text{-}16$; Figure 3A). In contrast, the exon inclusion levels simulated from two independent uniform priors had no correlation between the two samples (Pearson correlation $r = 0$; Figure 3B), contradicting with the real data.

To assess if the multivariate uniform prior is capable of modeling the correlation pattern in the data, we used our MCMC procedure to obtain the estimate of the correlation parameter ρ on the real ESRP data set. As the control, we analyzed 10 000 simulated data sets using exon inclusion levels simulated from two independent uniform priors, in which no correlation existed between the two samples. For the ESRP data, our MCMC procedure obtained an estimate of ρ of 0.93, consistent with the high overall positive correlation observed in the data. In contrast, the estimates of ρ on the 10 000 simulated data sets were close to 0, indicating no correlation between the two samples (Figure 3C). These results suggest that the multivariate uniform prior model is flexible enough for both situations, and that the MCMC procedure is capable of obtaining an estimate of the parameter ρ that reflects the degree of correlation in the data.

To assess if the pattern observed in the MDA-MB-231 cell line holds true when we compare more distantly related samples of different tissue origins, we analyzed the Illumina Human Body Map 2.0 data on 16 human tissues (see ‘Materials and Methods’ section). We performed pairwise comparisons of alternative splicing profiles of all possible tissue pairs. Between any two tissues, we observed a high overall correlation in the estimated exon inclusion levels of alternatively spliced cassette exons, with the Pearson correlation coefficient ranging from 0.89 to 0.97 (Supplementary Figure S1). These data further justify the use of the multivariate uniform prior, even for the comparison between more divergent samples representing different tissue types.

Simulation study of MATS

We evaluated the performance of MATS with a simulation study. Specifically, we generated a simulated RNA-Seq data set with a mixture of data points representing non-differentially spliced exons and differentially spliced exons. The threshold of the exon inclusion level difference between two samples was set as 10% in our simulation study (i.e. $|\psi_1 - \psi_2| > 0.1$). We generated data for non-differentially spliced exons by sampling the exon inclusion level of an exon in sample 1 from a uniform (0, 1) distribution, and randomly added or subtracted a small variation drawn from a uniform (0, 0.1) distribution to obtain the exon inclusion level in sample 2. We generated data for differentially spliced exons by sampling the exon inclusion level of an exon in sample 1 from a uniform (0, 1) distribution, and randomly added or subtracted a large variation drawn from a uniform (0.1, 1) distribution to obtain the exon inclusion level in sample 2. For all simulated exons, if the variation added to or subtracted from the exon inclusion level in sample 1 caused the exon inclusion level in sample 2 to be above 1 or below 0, the sampling step for the variation was repeated until the exon inclusion level in sample 2 was within the [0,1] range. In the simulated data, we generated a total of 5000 data points in which 95% represented non-differentially spliced exons and 5% represented differentially spliced exons (Figure 4A). After the exon inclusion levels were simulated for 5000 exons, we generated 5 simulated data sets, by setting the total inclusion + skipping isoform junction counts per exon and sample as 100, 200, 500,

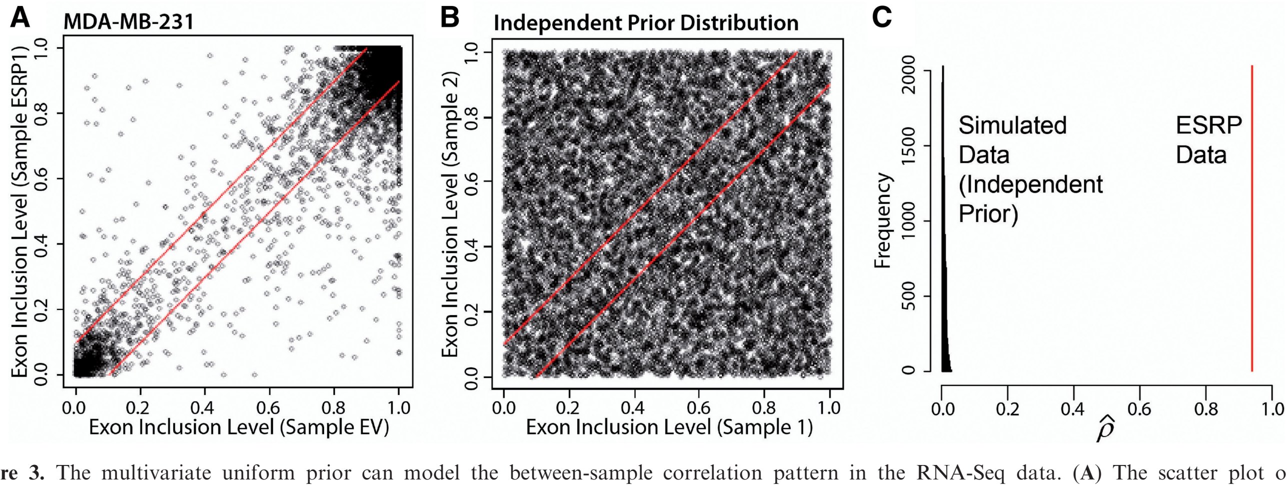

Figure 3. The multivariate uniform prior can model the between-sample correlation pattern in the RNA-Seq data. (**A**) The scatter plot of the estimated exon inclusion levels of 12 890 alternatively spliced cassette exons in the ESRP1 and EV samples. Only exons with at least 20 reads mapped to one of the three exon–exon junctions in both samples are included in the plot. (**B**) The scatter plot of the exon inclusion levels in two samples simulated from two independent uniform priors. In (A and B), the two red lines define the area where $|\psi_1 - \psi_2| \leq 0.1$. (**C**) The MCMC estimate of the correlation parameter ρ can capture the correlation pattern in the data. For the ESRP data, $\hat{\rho}$ is 0.93 (the red vertical line). For the 10 000 simulated data sets from independent uniform priors, $\hat{\rho}$ is distributed close to zero.

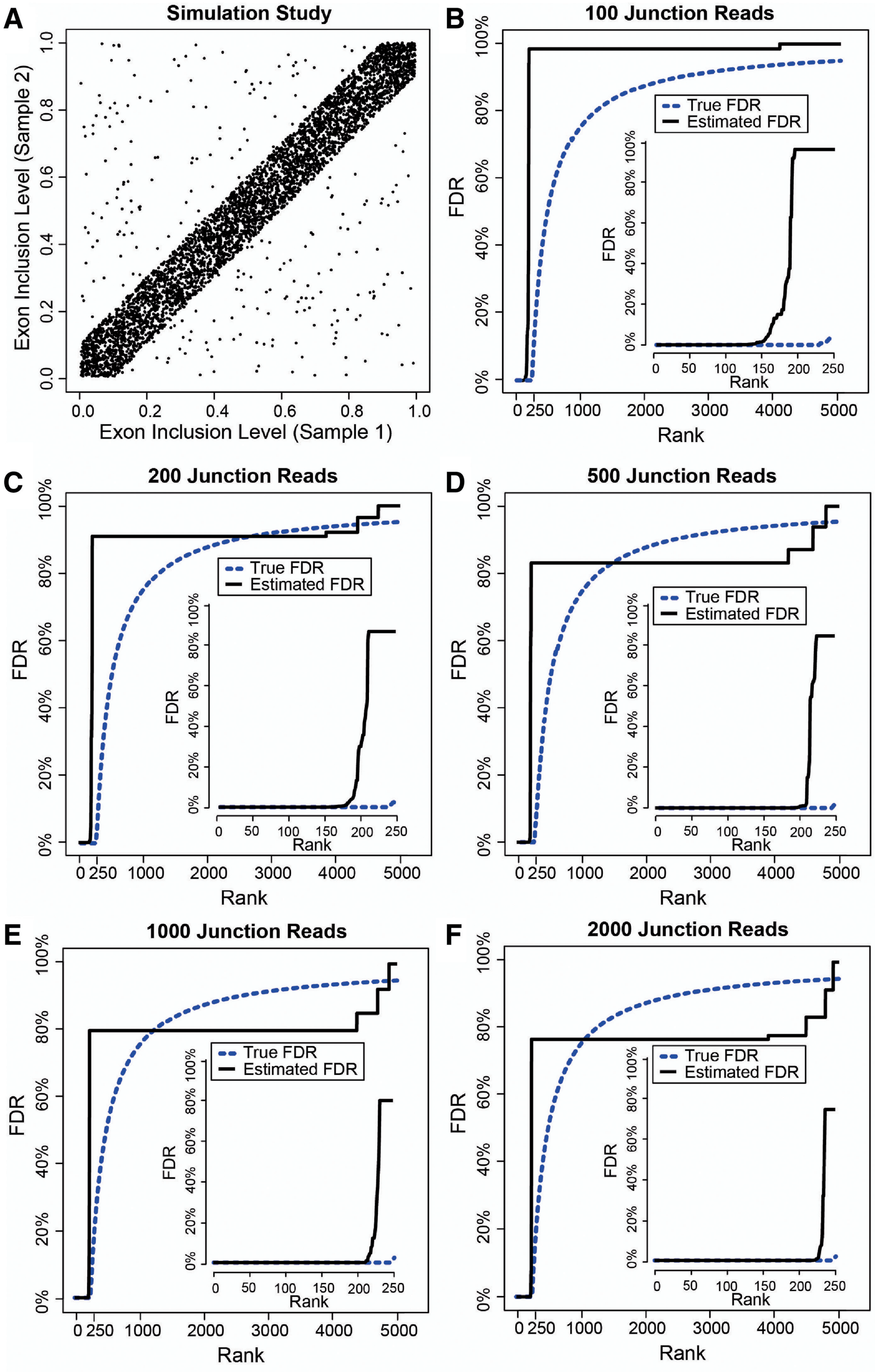

Figure 4. Simulation study of MATS. (**A**) Simulated exon inclusion levels of 5000 exons in two samples. A total of 95% of the data points are simulated from the null hypothesis ($|\psi_1 - \psi_2| \leq 0.1$) and 5% are simulated from the alternative hypothesis ($|\psi_1 - \psi_2| > 0.1$). (**B–F**) MATS FDR estimates on simulated data with the exon inclusion levels from (A) and the total junction count per exon and sample as 100 (B), 200 (C), 500 (D), 1000 (E) and 2000 (F). In each panel, exons are rank sorted by MATS FDR estimates in ascending order. The zoomed-in figure shows the FDR estimates of the top 250 exons by MATS.

1000 and 2000 respectively. The inclusion isoform count of an exon in a sample was then calculated as the product of its simulated exon inclusion level and the total junction count. After the simulation data sets were generated, we used MATS to calculate the *P*-value and FDR of each exon, and compared the estimated FDRs to the true FDRs. As shown in Figure 4B–F, although the estimated FDRs were generally more conservative (i.e. with higher values) than the true FDRs, the overall curve of the estimated FDR followed the trend of the true FDR curve, especially for exons ranked by MATS among the top 250 (i.e. the number of true positives in our simulated data set). We observed a sharp increase in the estimated FDR when the MATS rank of differential alternative splicing approached 250, consistent with the total number of true positives in the simulated data. Moreover, the number of exons with MATS FDR value of close to 0 increased progressively with increasing simulated junction counts (see the zoomed-in figures of Figure 4B–F), reflecting the influence of RNA-Seq depth on the sensitivity of detecting differential alternative splicing events. We note that non-differentially spliced exons constitute 95% of the simulated data points. For such exons, the correct FDR estimates should be high FDR values.

To further evaluate the performance of the MATS algorithm especially the benefit of the correlation parameter ρ, we conducted another simulation study to compare MATS with a simplified MATS Bayesian model in which ρ is fixed at 0 (i.e. independent prior), as well as the Fisher exact test. As in the previous simulation study, we generated 5000 data points in which 95% represented non-differentially spliced exons and 5% represented differentially spliced exons. To mimic the overall distribution of junction counts in real data sets, for each simulated exon its total junction counts (i.e. inclusion + skipping isoform junction counts) in sample 1 and 2 respectively were randomly sampled from the MDA-MB-231 ESRP1 data set by taking the counts of a randomly selected alternatively spliced cassette exon in the ESRP1 and EV samples. For each of the three methods tested, we calculated the true positive rate and false positive rate under sliding P-value cutoffs from 0 to 1. We then generated the receiver operating characteristic (ROC) curve for each method as the true positive rate versus false positive rate plot, and calculated the area under curve (AUC) for each method (Figure 5). MATS had the highest AUC of 0.96, significantly better than the simplified MATS Bayesian model with $\rho = 0$ (AUC = 0.88, DeLong's Test P = 6.8e-8). MATS also significantly outperformed the Fisher exact test (AUC = 0.90, DeLong's Test $P < 2.2\text{e-}16$). Additionally, we note that even the simplified MATS Bayesian model (with $\rho = 0$) outperformed the Fisher exact test in the most critical area of the ROC curve where the false positive rate was low (Figure 5). This indicates that by testing for *difference* (with a threshold) instead of *equality*, the test statistics is better at separating true positives from false positives.

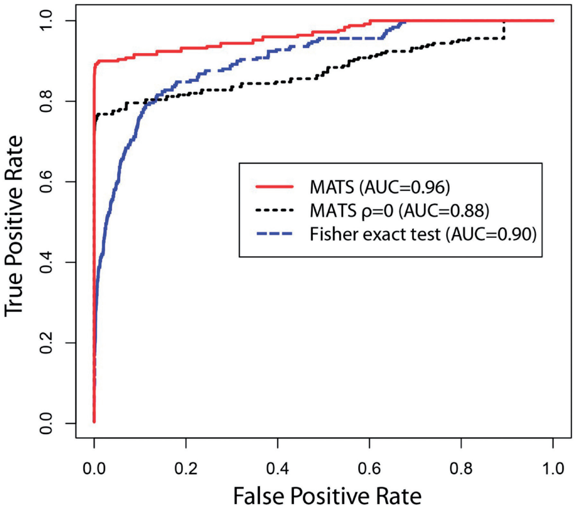

Figure 5. Simulation study to compare MATS, a simplified MATS Bayesian model in which ρ is fixed at 0 (i.e. independent prior), and the Fisher exact test. MATS significantly outperforms the other two methods based on the AUC of the ROC curve (i.e. the true positive rate versus false positive rate plot).

MATS analysis of ESRP1-regulated differential alternative splicing

To evaluate the performance of MATS on a real data set, we used MATS to detect ESRP1-regulated differential alternative splicing events using our RNA-Seq data on the MDA-MB-231 cell line with ectopic expression of ESRP1 or an empty vector (EV) control. For each of the 18 859 alternatively spliced cassette exons (defined as exons with at least one inclusion read and one skipping read in these two samples), we calculated the Bayesian posterior probability, P-value and FDR for $|\psi_1 - \psi_2| > 0.1$. Among 240 exons with MATS FDR of <10%, all (100%) had posterior probability of >0.85, including 239 (99.6%) and 234 (97.5%) with posterior probability of >0.9 and >0.95 respectively. Figure 6 illustrates a differentially spliced exon (exon 7 of *SPNS1*) identified by MATS. Based on the RNA-Seq read counts we estimated an exon inclusion level of 77% in the EV sample and 27% in the ESRP1 sample, with a FDR value of 4.6e-4 (Figure 6A). These predictions matched RT-PCR results of *SPNS1* exon 7 splicing in these two samples (Figure 6B).

To assess the overall accuracy of our FDR estimates, we selected 164 exons covering a broad range of MATS FDR values (Supplementary Table S1) and tested their splicing patterns by RT–PCR. Of all the exons tested by RT–PCR, 111 exons had at least 10% difference in the exon inclusion levels between the two samples with the direction of change matching the RNA-Seq predictions. This yielded an overall validation rate of 68%. To assess whether the validation rate correlated with MATS FDR estimates, we divided the full list of 164 exons into four cohorts according to the estimated FDR values, and calculated the RT–PCR validation rate for each cohort. We observed a progressive decrease in the RT–PCR validation rate for cohorts with increasing FDR values (Figure 7). The first cohort had 92 exons with FDR estimates between 0 to 10%. In this cohort, 79 exons were validated by RT–PCR as differentially spliced, yielding a high validation rate of 86%. The second, third and fourth cohorts corresponded to exons with FDR estimates between 10% and 30%, between 30% and 60%, and between 60% and 100%. These three cohorts had RT–PCR validation rates of 73%, 55% and 36%, respectively (Figure 7). These results indicate that MATS can generate experimentally meaningful FDR estimates to help biologists with the interpretation of RNA-Seq predictions and the design of follow-up experiments. There was a sharp increase in the estimated FDR value after the initial list of top 240–406 exons (Figure 7), with ~98% of the exons having a FDR

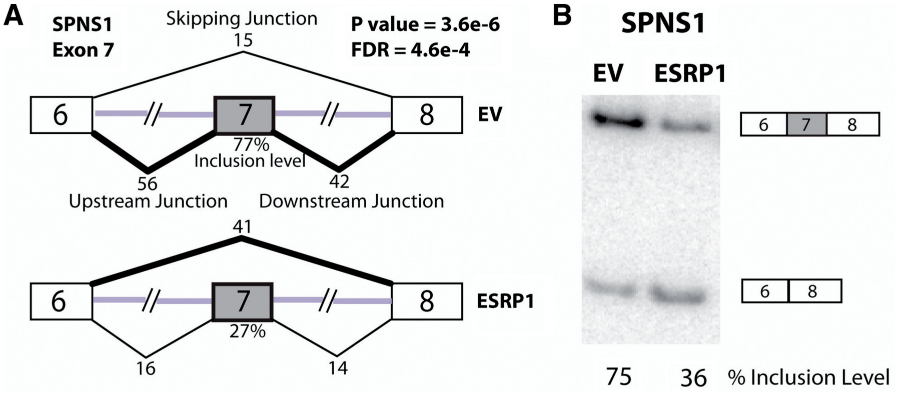

Figure 6. RNA-Seq and RT–PCR analysis of *SPNS1* exon 7 splicing. (**A**) RNA-Seq junction counts and MATS result of *SPNS1* exon 7 in the EV and ESRP1 samples. (**B**) RT–PCR result of *SPNS1* exon 7 in the EV and ESRP1 samples.

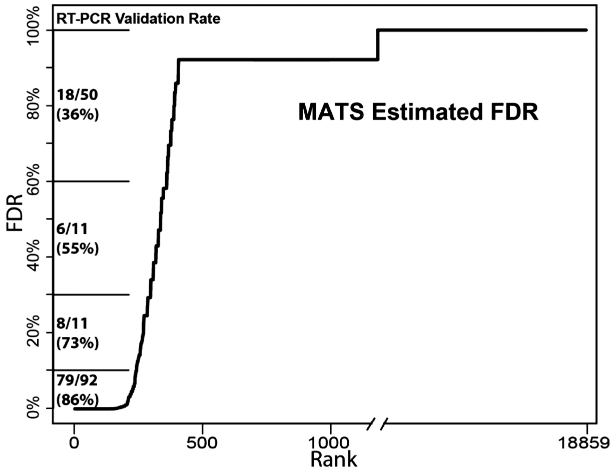

Figure 7. RT–PCR validation of 164 exons covering a broad range of MATS FDR values. All exons analyzed by MATS are rank sorted by FDR estimates (*y*-axis) in ascending order. The 164 exons tested by RT–PCR are divided into four non-overlapping cohorts according to the FDR estimates. The validation rate for each cohort is shown.

of ≥90%. This was similar to the shape of the FDR distribution in the simulation study (Figure 4), probably reflecting the number of ESRP1-regulated exons in the human genome as well as the percentage of which that can be detected at the current RNA-Seq depth. Of note, among the 164 exons tested by RT–PCR, 17 had a MATS FDR of 100%. Only 1 of the 17 exons had more than 10% change in the RT–PCR estimated exon inclusion levels with the direction of change matching the RNA-Seq prediction, yielding a low validation rate of only 6% that closely matched the estimated FDR. This indicates that MATS can provide reliable FDR estimates for the full range of possible FDR values.

Since our exon–exon junction database includes all known exon-exon junctions observed in Ensembl transcripts, as well as hypothetical exon–exon junctions obtained by all possible pairwise fusions of exons within genes, we can detect and analyze novel exon skipping events not supported by existing transcript annotations. This is important considering the prevalence of tissue- and condition-specific alternative splicing (32). Of all 18 859 alternatively spliced cassette exons, 5373 represent known events and 13 486 represent novel events. Of the 240 significant events with MATS FDR<10%, 140 represent known events while 100 represent novel events. The percentage of events called as differentially spliced is significantly higher for known events (2.6%; 140/5373) than for novel events (0.7%; 100/13486). Interestingly, of the 92 RT–PCR tested cassette exons with FDR<10%, the RT–PCR validation rate of differential alternative splicing is higher for novel events (98%; 43/44) than for known events (75%; 36/48).

The MATS algorithm can be naturally extended to other types of alternative splicing events such as alternative 5′ or 3′ splice sites and mutually exclusive exon usage. In the analysis of alternative 5′ or 3′ splice sites, the counts of junction reads that uniquely support the two competing splice sites can serve as the input for MATS. These counts can be used to estimate the 'inclusion level' of any specific splice site. All subsequent analysis steps are identical to the analysis of differential exon skipping. To illustrate this, we applied MATS to 1571 alternative 5′ splice site events and 2383 alternative 3′ splice site events in the ESRP1 data set. With a FDR cutoff of 5%, 13 events (9 alternative 5′ splice sites and 4 alternative 3′ splice sites) were identified to undergo significant differential splicing (>10% change) between the ESRP1 and EV samples. The small number of detected differential alternative 5′ or 3′ splice sites is consistent with an early study using the Affymetrix exon 1.0 array (33). One possible explanation is that ESRP1 regulates a small number of such events in the transcriptome. Nonetheless, we tested five events by RT–PCR, of which three were validated to have at least 10% change in splice site inclusion level (Supplementary Table S2). Supplementary Figure S2 shows the example of a validated differential alternative 5' splice site event in exon 4 of *HNRNPH3*.

MATS analysis of switch-like alternative splicing between brain and 15 other tissues

MATS has the flexibility to detect exons with the extreme 'switch-like' differential alternative splicing pattern (see 'Materials and Methods' section). To illustrate this function, we used MATS to detect switch-like differential alternative splicing events between the brain and each of the 15 other tissues in the Illumina Human Body Map 2.0 data set. For each pairwise comparison, we tested if the exon inclusion level of an exon switches from less than 1/3 in one tissue to more than 2/3 in the other tissue (i.e. ($\psi_1 < 1/3$ and $\psi_2 > 2/3$) or ($\psi_1 > 2/3$ and $\psi_2 < 1/3$). With a FDR cutoff of <50%, a total of 229 exons were identified to have the switch-like differential alternative splicing pattern between the brain and at least one other tissue. Prior studies have revealed sequence features of such 'tissue-switched' cassette exons characteristic of functional alternative splicing events (3,22). A unique feature of tissue-switched exons is that they are much more likely to be exact multiples of 3 nt in length, thus alternative splicing adds or removes a modular peptide segment of the final protein product while preserving the downstream open reading frame (i.e. 'frame-preserving', as opposed to 'frame-switching' for exons not exact multiples of 3 nt in length). Consistent with these findings, of the 229 switch-like exons detected by MATS between brain and 15 other tissues, 70% are frame-preserving, compared to 42% for other alternatively spliced cassette exons without switch-like differential alternative splicing ($P = 0$; see Supplementary Figure S3).

DISCUSSION

We present MATS, a new method to detect differential alternative splicing events from RNA-Seq data. A major advantage of MATS over existing methods is that it allows flexible hypothesis testing of differential alternative splicing patterns. Most of the published work attempted to identify differential alternative splicing events by testing the equality of the exon inclusion levels between samples (11,15,16,18–20). Some also applied a secondary filter (without statistical testing) on the change in the estimated exon inclusion levels (20). MATS provides a Bayesian statistical framework to directly test the hypothesis and evaluate the statistical significance that the absolute difference in exon inclusion levels between two samples exceeds any user-defined threshold. This allows researchers to select the magnitude of splicing changes suitable for specific research goals in a rigorous statistical setting. We assessed the performance of MATS using simulated data and real RNA-Seq data sets. In the RNA-Seq analysis of ESRP1 regulated alternative splicing, we obtained a high RT–PCR validation rate of 86% for candidate exons with MATS FDR <10%. Additionally, over the full list of RT–PCR tested exons, the MATS FDR estimates matched well with the experimental validation rate (Figure 7). The MATS framework is also applicable to other null hypotheses of interest. For example, MATS can be used to test the hypothesis that an exon exhibits the extreme 'switch-like' differential alternative splicing pattern.

A novel feature of MATS is the multivariate uniform prior that models the between-sample correlation in exon splicing patterns. In both the ESRP data and the Human Body Map 2.0 data, the degree of correlation between samples is high, resulting in a high estimated value of the correlation parameter ρ for the multivariate prior model. The high between-sample correlation observed in real RNA-Seq data is consistent with the mechanism of splicing regulation in eukaryotic cells. Splicing is a complex process mediated by extensive interactions among *cis* regulatory elements and *trans* acting regulators (34). Most splicing regulators may control the splicing of up to several hundred exons via sequence-specific protein–RNA interactions (35). Perturbing a specific component of the splicing regulatory pathway usually changes the splicing activity of a small subset of alternatively spliced exons, while the majority of alternatively spliced exons remain unaffected. We also note that when no correlation exists in the data (see Figure 3 for the simulated data), the estimate of ρ by the MCMC procedure is close to zero. Taken together, our analysis indicates that the multivariate uniform prior model is flexible enough to accommodate different degrees of between-sample correlation in the RNA-Seq data. Moreover, our simulation study (Figure 5) indicates that incorporating the correlation parameter ρ in the MATS model improves the performance of the algorithm.

The MATS software and documentation as well as the raw ESRP1 RNA-Seq data are freely available at http://intron.healthcare.uiowa.edu/MATS/. The scripts provided online include the core MATS program to calculate the posterior probability, *P*-value and FDR of differential alternative splicing from the input junction counts, as well as accessory scripts to generate junction counts and detect alternative splicing events from raw RNA-Seq data. To facilitate data analysis, we also provide databases of pre-compiled exon-exon junctions based on the Ensembl (28) or UCSC Known Genes (36) annotations. We suggest that MATS can be used either as a stand-alone software for differential alternative splicing analysis of RNA-Seq data, or as part of an existing RNA-Seq analysis pipeline to calculate the statistical significance of a user-defined differential splicing pattern using junction counts generated by other mapping procedures or transcript annotation databases. We note that a number of recent studies have reported biases in RNA-Seq data such as the non-uniform distribution of RNA-Seq reads along mRNA transcripts, and have proposed methods to adjust raw RNA-Seq read counts by correcting for such biases (18,37–41). Additionally, it has been demonstrated that in paired-end RNA-Seq, the distribution of insert size between the two ends can be utilized to improve the assignment of reads to specific transcript isoforms (16). Thus, it is possible to use an appropriate method to adjust raw RNA-Seq read counts and refine the counts of isoform-specific junctions, prior to the hypothesis testing of differential alternative splicing by MATS. It should also be noted that although the analysis in this manuscript is mostly focused on exon skipping events (i.e. differential inclusion/skipping of an entire exon), exon skipping is only one subtype of alternative splicing events. The MATS algorithm can be readily applied to junction counts generated for other types of alternative splicing events, as illustrated for alternative 5′ or 3′ splice sites on the ESRP1 data set (Supplementary Figure S2).

MATS currently performs two-group comparison of two samples, with one sample per group without within-group replicates. This is the typical experimental setup in most published RNA-Seq studies of alternative splicing including studies of splicing regulators (16,20,42), largely due to the high cost of RNA-Seq to achieve sufficient depth for splicing analysis. However, as the cost of high-throughput sequencing continues to decline, we anticipate that it will soon become feasible and common for researchers to incorporate biological replicates in RNA-Seq studies of alternative splicing (43). Medical researchers may soon be able to generate RNA-Seq data across a large number of healthy and diseased specimens, with the depth sufficient for quantifying splicing in each individual sample. Thus, an important future direction is to extend the statistical framework to incorporate the use of RNA-Seq replicates in detecting differential alternative splicing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures S1–S3 and Supplementary Tables S1–S2.

ACKNOWLEDGEMENTS

We wish to thank Keyan Zhao for comments, and Collin Tokheim for technical assistance. We thank Gary Schroth for early access to the Illumina Human Body Map 2.0 data. We thank David Eichmann, Ben Rogers and the University of Iowa Institute for Clinical and Translational Science (NIH grant UL1RR024979) for computer support.

FUNDING

National Institutes of Health (R01GM088342 and P30DK054759 to Y.X., R01GM088809 to R.P.C., R01CA120988 to J.H.); National Science Foundation (DMS-0805491, Career award DMS-1055286 to Q.Z.); a junior faculty grant from the Edward Mallinckrodt Jr Foundation (to Y.X.); National Institutes of Health T32 postdoctoral fellow training grant (T32HL007638 to J.W.P.). Funding for open access charge: NIH R01GM088342.

Conflict of interest statement. None declared.

REFERENCES

1. Keren,H., Lev-Maor,G. and Ast,G. (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.*, **11**, 345–355.
2. Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
3. 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.
4. 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.
5. Cooper,T.A., Wan,L. and Dreyfuss,G. (2009) RNA and disease. *Cell*, **136**, 777–793.
6. Heyd,F. and Lynch,K.W. (2011) DEGRADE, MOVE, REGROUP: signaling control of splicing proteins. *Trends Biochem. Sci.*, **36**, 397–404.
7. Buchner,D.A., Trudeau,M. and Meisler,M.H. (2003) SCNM1, a putative RNA splicing factor that modifies disease severity in mice. *Science*, **301**, 967–969.
8. Ingram,E.M. and Spillantini,M.G. (2002) Tau gene mutations: dissecting the pathogenesis of FTDP-17. *Trends Mol. Med.*, **8**, 555–562.
9. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.
10. 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.
11. Xu,Q. and Lee,C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res.*, **31**, 5635–5643.
12. Wang,Z., Lo,H.S., Yang,H., Gere,S., Hu,Y., Buetow,K.H. and Lee,M.P. (2003) Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res.*, **63**, 655–657.
13. Gupta,S., Zink,D., Korn,B., Vingron,M. and Haas,S.A. (2004) Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC Genomics*, **5**, 72.
14. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
15. Griffith,M., Griffith,O.L., Mwenifumbo,J., Goya,R., Morrissy,A.S., Morin,R.D., Corbett,R., Tang,M.J., Hou,Y.C., Pugh,T.J. *et al.* (2010) Alternative expression analysis by RNA sequencing. *Nat. Methods*, **7**, 843–847.
16. Katz,Y., Wang,E.T., Airoldi,E.M. and Burge,C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.
17. Shen,S., Lin,L., Cai,J.J., Jiang,P., Kenkel,E.J., Stroik,M.R., Sato,S., Davidson,B.L. and Xing,Y. (2011) Widespread establishment and regulatory impact of Alu exons in human genes. *Proc. Natl Acad. Sci. USA*, **108**, 2837–2842.
18. Srivastava,S. and Chen,L. (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.*, **38**, e170.
19. Lalonde,E., Ha,K.C., Wang,Z., Bemmo,A., Kleinman,C.L., Kwan,T., Pastinen,T. and Majewski,J. (2011) RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res.*, **21**, 545–554.
20. Brooks,A.N., Yang,L., Duff,M.O., Hansen,K.D., Park,J.W., Dudoit,S., Brenner,S.E. and Graveley,B.R. (2011) Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res.*, **21**, 193–202.
21. 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.
22. Xing,Y. and Lee,C.J. (2005) Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet.*, **1**, e34.
23. Warzecha,C.C., Jiang,P., Amirikian,K., Dittmar,K.A., Lu,H., Shen,S., Guo,W., Xing,Y. and Carstens,R.P. (2010) An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. *EMBO J.*, **29**, 3286–3300.
24. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.*, **57**, 289–300.
25. Modrek,B. and Lee,C. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased rate of exon creation/loss. *Nat. Genet.*, **34**, 177–180.
26. Zhu,C.Y., Byrd,R.H., Lu,P.H. and Nocedal,J. (1997) Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale

bound-constrained optimization. *ACM Trans. Math. Software*, **23**, 550–560.

27. Byrd,R.H., Lu,P.H., Nocedal,J. and Zhu,C.Y. (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, **16**, 1190–1208.
28. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
29. Warzecha,C.C., Sato,T.K., Nabet,B., Hogenesch,J.B. and Carstens,R.P. (2009) ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. *Mol. Cell*, **33**, 591–601.
30. 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.
31. Lu,Z.X., Jiang,P., Cai,J.J. and Xing,Y. (2011) Context-dependent robustness to 5′ splice site polymorphisms in human populations. *Hum. Mol. Genet.*, **20**, 1084–1096.
32. Kalsotra,A. and Cooper,T.A. (2011) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.*, **12**, 715–729.
33. Warzecha,C.C., Shen,S., Xing,Y. and Carstens,R.P. (2009) The epithelial splicing factors ESRP1 and ESRP2 positively and negatively regulate diverse types of alternative splicing events. *RNA Biol.*, **6**, 546–562.
34. Wang,Z. and Burge,C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, **14**, 802–813.
35. Chen,M. and Manley,J.L. (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.*, **10**, 741–754.
36. Hsu,F., Kent,W.J., Clawson,H., Kuhn,R.M., Diekhans,M. and Haussler,D. (2006) The UCSC known genes. *Bioinformatics*, **22**, 1036–1046.
37. 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.
38. Schwartz,S., Oren,R. and Ast,G. (2011) Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS ONE*, **6**, e16685.
39. Li,J., Jiang,H. and Wong,W.H. (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.*, **11**, R50.
40. Roberts,A., Trapnell,C., Donaghey,J., Rinn,J.L. and Pachter,L. (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.*, **12**, R22.
41. Bohnert,R. and Ratsch,G. (2010) rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.*, **38**, W348–W351.
42. Luco,R.F., Pan,Q., Tominaga,K., Blencowe,B.J., Pereira-Smith,O.M. and Misteli,T. (2010) Regulation of alternative splicing by histone modifications. *Science*, **327**, 996–1000.
43. Hansen,K.D., Wu,Z., Irizarry,R.A. and Leek,J.T. (2011) Sequencing technology does not eliminate biological variability. *Nat. Biotechnol.*, **29**, 572–573.