Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Array Platform Modeling and Analysis (A)

Mapping Splicing Quantitative Trait Loci in RNA-Seq

Cheng Jia, Yu Hu, Yichuan Liu and Mingyao Li

Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

ABSTRACT

BACKGROUND: One of the major mechanisms of generating mRNA diversity is alternative splicing, a regulated process that allows for the flexibility of producing functionally different proteins from the same genomic sequences. This process is often altered in cancer cells to produce aberrant proteins that drive the progression of cancer. A better understanding of the misregulation of alternative splicing will shed light on the development of novel targets for pharmacological interventions of cancer.

METHODS: In this study, we evaluated three statistical methods, random effects meta-regression, beta regression, and generalized linear mixed effects model, for the analysis of splicing quantitative trait loci (sQTL) using RNA-Seq data. All the three methods use exon-inclusion levels estimated by the PennSeq algorithm, a statistical method that utilizes paired-end reads and accounts for non-uniform sequencing coverage.

RESULTS: Using both simulated and real RNA-Seq datasets, we compared these three methods with GLiMMPS, a recently developed method for sQTL analysis. Our results indicate that the most reliable and powerful method was the random effects meta-regression approach, which identified sQTLs at low false discovery rates but higher power when compared to GLiMMPS.

CONCLUSIONS: We have evaluated three statistical methods for the analysis of sQTLs in RNA-Seq. Results from our study will be instructive for researchers in selecting the appropriate statistical methods for sQTL analysis.

KEYWORDS: alternative splicing, quantitative trait loci, RNA-Seq

SUPPLEMENT: Array Platform Modeling and Analysis (A)

CITATION: Jia et al. Mapping Splicing Quantitative Trait Loci in RNA-Seq. *Cancer Informatics* 2014:13(S4) 35–43 doi: 10.4137/CIN.S13971.

RECEIVED: April 18, 2014. **RESUBMITTED:** July 23, 2014. **ACCEPTED FOR PUBLICATION:** July 25, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Original Research

FUNDING: This work was supported by the National Institutes of Health (R01HG004517, R01HG005854, R01GM097505, R01HG006465, and R01GM1008600 to ML). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: mingyao@mail.med.upenn.edu

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Background

Alternative splicing, a post-transcriptional process that allows multiple messenger RNA (mRNA) isoforms to be produced by a single gene, is a regulated process, and a major mechanism for generating protein diversity. During this process, specific exons of a gene can be either included or excluded from the mature mRNAs, leading to structurally and functionally distinct proteins. In multicellular organisms, alternative splicing is a prevalent phenomenon, which has been estimated to occur in over 90% of the human genes.[1]

Alternative splicing is often altered in cancer cells to produce aberrant proteins that drive the progression of cancer.[2–5] Genome-wide studies have identified more than 15,000 splicing variants associated with a wide range of cancers.[6–8] During oncogenesis, alternative splicing can affect genes involved in promoting cell migration, activating cell growth, maintaining hormone responsiveness, curbing apoptosis, and evading chemotherapy.[9,10] A number of factors can contribute to the misregulation of alternative splicing, including the disruption of either *cis*-acting elements within the affected gene or *trans*-acting factors that are required for normal splicing. A better understanding of this misregulation can help identify novel targets for pharmacological intervention of cancer.

Several studies have demonstrated the regulatory role of single nucleotide polymorphisms (SNPs) on the splicing patterns of mRNA precursors. These SNPs have been termed splicing quantitative trait loci (sQTL). One of the critical steps in studying the mechanisms and regulation of alternative splicing is the identification of these loci, which previously has been achieved via high-throughput technologies such as microarrays. Using samples from lymphoblastoid B cell lines, brain, or peripheral blood mononuclear cell,[11–14] several studies have demonstrated the functional importance of alternative splicing in a variety of normal and diseased tissues in human. Interestingly, this body of work has also collectively suggested that *cis*-acting sQTLs are prevalent while *trans*-acting sQTLs are less common. Recently, RNA sequencing (RNA-Seq), a high-throughput sequencing-based approach, has also been employed to study sQTLs.[15–17] Because of the improved accuracy of RNA-Seq in gene expression quantification over microarrays and its ability to measure isoform-specific gene expression, it has become an increasingly popular approach for studying alternative splicing.

Analysis of sQTLs using RNA-Seq is challenging because the characterization of alternative splicing relies on isoform-specific gene expression, which has to be estimated statistically. To date, only a few methods have been developed to identify sQTLs from RNA-Seq data. One simple approach is to perform linear regression in which the percentage of exon read counts over total gene read counts is treated as the quantitative trait, and the SNP genotype is treated as the independent variable.[15] However, this model fails to account for the variability in RNA-Seq read counts and hence can lead to false positive results. Recently, Zhao et al developed GLiMMPS, a generalized linear mixed effects model approach that takes into account the variation of exon-specific read coverage as well as the overdispersion of read counts.[18] Although GLiMMPS has shown significant improvement over simple linear regression, this method has several shortcomings. First, by considering only splice junction reads, GLiMMPS ignores reads that align to the body of alternative exon and those that align to flanking constitutive exons. Previous studies have shown that both types of reads are informative for alternative splicing inference.[19] Second, GLiMMPS cannot be extended to integrate the paired-end nature of RNA-Seq data. Third, GLiMPPS treats the estimated exon-inclusion level as a random effect, while a more desirable and mathematically accurate approach is to explicitly model the variance associated with the exon-inclusion level estimation. Fourth, GLiMPPS relies on the assumption that RNA-Seq reads are uniformly distributed along transcripts. Various studies have shown that RNA-Seq reads are rarely uniformly distributed, and the ignorance of this phenomenon can lead to biased estimates of isoform-specific gene expression.[20]

An ideal method for sQTL analysis should be able to directly model the variation of exon-inclusion level estimates between samples, utilize extra information embedded in paired-end RNA-Seq data, and adjust for non-uniform read distribution. In this study, we evaluated three statistical methods, including random effects meta-regression, beta regression, and generalized linear mixed effects model, for the analysis of sQTLs. All the three methods used exon-inclusion level estimated by the PennSeq algorithm as input.[20] PennSeq is a recently developed statistical method that utilizes paired-end reads and allows for non-uniform read distribution. Using both simulated and real RNA-Seq datasets, we demonstrated that the best performing method is the random effects meta-regression approach, which shows low false discovery rates (FDRs) but higher power when compared to GLiMMPS.

Materials and Methods

Estimation of exon-inclusion level. One vital step in sQTL analysis is to estimate exon-inclusion level, which is defined as the proportion of mRNAs that originates from the exon-inclusion isoform, ie, the longer isoform including the exon that is otherwise skipped, among all transcripts from the same gene. The estimated exon-inclusion level is often treated as a quantitative trait and used in a regression framework to identify sQTLs. GLiMMPS only uses junction reads mapped to splice junctions and estimates the exon-inclusion level as the fraction of the number of inclusion splice junction reads to the total number of junction reads. However, as shown in Figure 1A, reads aligning to the alternative exon body and those that align to the flanking constitutive exons are also informative for exon-inclusion level estimation, because higher expression of the exon-inclusion isoform will increase the density of reads in the alternative exon relative to the flanking constitutive exons.[19] To utilize all available information, we chose to use PennSeq,[20] a recently developed non-parametric-based approach, to estimate exon-inclusion level. The PennSeq algorithm considers all mapped reads in a given exon-trio, which is composed of the alternative exon and the flanking constitutive exons. Additionally, PennSeq takes into account the paired-end nature of the data and allows the exon-inclusion isoform and the exon-exclusion isoform to have their own non-uniform read distributions. The exon-inclusion level is then estimated as the relative expression of the exon-inclusion isoform over the total expression of the two isoforms.

Random effects meta-regression for analysis of sQTLs. The goal of random effects meta-regression is to synthesize results from multiple studies, accounting for variability in the effect estimates across studies by explicitly allowing for different sources of variability: within- and between-study variation.[21] This parallels perfectly with sQTL analysis in that within-study variation represents the variance introduced in exon-inclusion level estimation and between-study variation represents the variation in exon-inclusion levels across samples. This analogy motivated us to explore random effects meta-regression as a means to identify sQTLs.

Let Y_i denote the estimated exon-inclusion level of an exon-trio for subject i ($i = 1,\ldots,n$), and σ_{0i} denotes the standard error of the estimated exon-inclusion level. Both Y_i and σ_{0i} can be obtained from programs that estimate isoform-specific gene expression (eg, PennSeq[20] or Cufflinks[22]). The SNP

genotype is denoted by G_i, which takes values of 0, 1, and 2, counting the number of minor alleles. To fit the data using random effects meta-regression, the exon-inclusion levels are transformed using the logit function so that their distribution is approximately normal. We approximate the standard error of logit(Y_i) using the delta method:

$$Var\left[\log it(Y_t)\right] = \sigma_{1i}^2 \approx \frac{\sigma_{0i}^2}{\left[E(Y_i)(1-E(Y_i))\right]^2}.$$

With the above notation, the random effects meta-regression model (Fig. 1B) can be expressed as:

$$\text{logit}(Y_i) = \beta_0 + \beta_1 G_i + u_i + e_i,$$

where u_i represents the estimation uncertainty of logit(Y_i), and e_i is the random error due to the remaining differences between exon-inclusion levels across samples. For this random effects model, we assume: (1) $u_i \sim N(0, \sigma_{1i}^2)$, (2) $e_i \sim N(0, \sigma_e^2)$, (3) u_i and e_i are uncorrelated, ie, Cov(u_i, e_i) = 0, and (4) the n observations are independent. If the variance of u_i is set to zero, ie, there is no estimation uncertainty of exon-inclusion level, then this random effects model reduces to the standard linear regression model. To test the null hypothesis of no association between the SNP genotype and the exon-inclusion level, ie, H_0:β_1 0, a Wald test is performed. We can carry out statistical inference using standard statistical software, such as R (http://www.r-project.org) or Stata (Stata Corp, College Station, TX). In our analysis, we used the *metafor* package in R.[21]

Beta regression for analysis of sQTLs. Since exon-inclusion level takes values between 0 and 1, it is natural to assume that it follows a beta distribution. With this assumption, we can identify sQTLs using beta regression.[23] In contrast to the random effects meta-regression, which requires logit transformation on the exon-inclusion estimates, beta regression can model the exon-inclusion level directly, producing results that are readily interpretable. The beta regression model is based on an alternative parameterization of the beta distribution in terms of a mean parameter μ and a precision parameter φ. We assume Y_i, the exon-inclusion level for subject i, follows a beta distribution, B(m_i, φ), where $E(Y_i) = m_i$ and $Var(Y_i) = m_i\,(1 - m_i)/(1 + \varphi)$. The beta regression model (Fig. 1C) can be expressed as:

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 G_i.$$

We can perform beta regression using the *betareg* package in R, and test H_0:$\beta_1 = 0$ using a Wald test. Since the variance of Y_i is a function of m_i, the beta regression model can naturally account for heteroscedasticity of exon-inclusion levels.

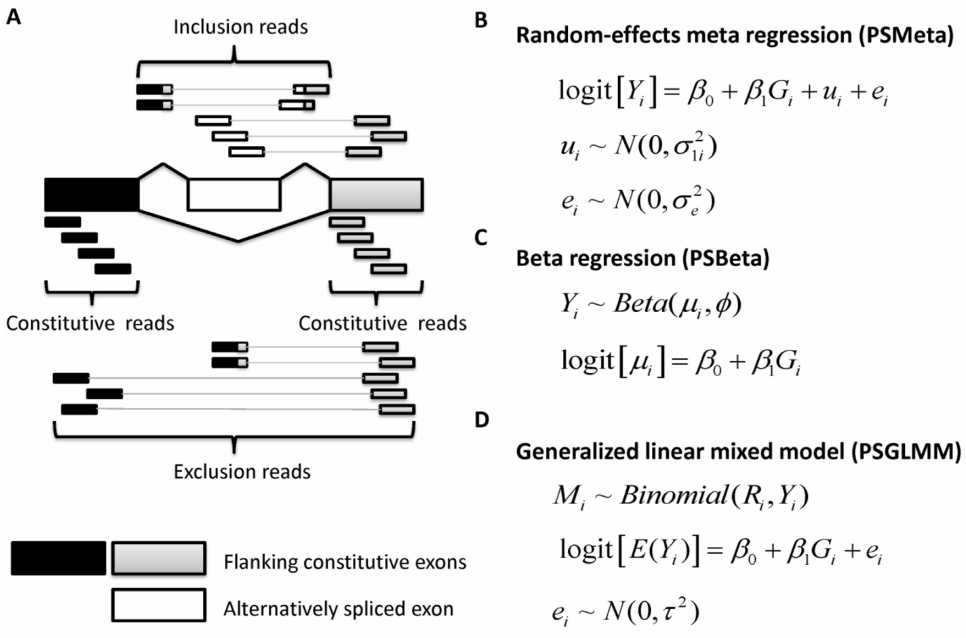

Figure 1. (**A**) An exon-trio is composed of two flanking constituitive exons and an alternatively spliced exon. RNA-Seq reads aligning to the body of the alternatively spliced exon or to splice junctions involving this exon support the inclusive isoform, whereas reads joining the two constitutive exons support the exclusive isoform.[19] GLiMMPS uses splice-junction reads only, but does not use reads mapping to the alternative exon body and those mapping to the flanking constitutive exon body. In contrast, PennSeq uses all available reads in the exon-trio. (**B**) PSMeta: random effects meta-regression. Y_i is the estimated exon-inclusion level from PennSeq for subject i. G_i is the SNP genotype for subject i. u_i represents the estimation uncertainty of logit(Y_i), and e_i is the random error due to the remaining differences between exon-inclusion levels across samples. (**C**) PSBeta: beta regression, where μ_i is the mean parameter for subject i, and φ is the dispersion parameter. (**D**) PSGLMM: generalized linear mixed effects model. R_i is the total number of reads mapped to the exon-trio in subject i. M_i is the number of reads originating from the exon-inclusion isoform for subject i.

Generalized linear mixed effects model for analysis of sQTLs. GLiMMPS is based on generalized linear mixed effects model in which the dependent variable is the exon-inclusion level, estimated exclusively from reads spanning splice junctions, which only represent part of the information on alternative splicing embedded in RNA-Seq data. Therefore, it is expected that using estimates from PennSeq, which incorporates other alternative splicing informative reads omitted by GLiMMPS, would yield higher power. To fit the generalized linear mixed effects model, we first calculate the total number of reads mapped to a given exon-trio, denoted by R_i, and then estimate the number of reads originating from the exon-inclusion isoform by $M_i = R_i \times Y_i$, where Y_i is the exon-inclusion level obtained from PennSeq. With M_i, R_i, and G_i, we can directly fit the data using the generalized linear mixed effects model (Fig. 1D), which can be written as:

$$\text{logit}\left[E(Y_i)\right] = \beta_0 + \beta_1 G_i + e_i ,$$

where $e_i \sim N(0,\tau^2)$. Same as GLiMMPS, we carry out the analysis using the *lme4* package in R.

RNA-Seq data simulation. To evaluate the performance of the aforementioned methods in sQTL identification, we conducted simulation studies and compared their empirical power to that of GLiMMPS. Flux Simulator was used to simulate a series of paired-end RNA-Seq experiments *in silico*.[24] The human genome sequence (hg19, NCBI build 37) was downloaded from UCSC Genome Browser together with the coordinates of all isoforms in the RefGene table. We selected genes with at least three exons and two isoforms, and further required that the selected genes do not overlap with each other. For each selected gene, we kept the longest isoform and generated a shorter isoform by randomly removing an interior exon from the longest isoform, resulting in 4,710 exon-trios in the final list. We simulated SNP genotype data following Hardy–Weinberg equilibrium and assumed a SNP minor allele frequency (MAF) of 0.4. The exon-inclusion level was determined by the formula, $Y_i = \text{logit}^{-1}(-0.35 + \beta_1 G_i + e_i)$ where $e_i \sim N(0, 0.05^2)$. For each exon-trio, Y_i was used to calculate the number of molecules for the exon-inclusion isoform and the exon-exclusion isoform. We then simulated data with 50% of the exon-trios having sQTLs in which β_1 was set to log(1.2), and the remaining 50% having no sQTLs in which β_1 was set to zero. Based on the total number of RNA molecules, the Flux Simulator assigns an abundance value for each isoform following a mixed power/exponential law. Additionally, the Flux Simulator simulates common sources of systematic bias in the abundance and distribution of reads by *in silico* library preparation and sequencing. We simulated 120 individuals with 10 million 76 bp paired-end reads per individual. For each simulated dataset, the RNA-Seq reads were mapped to the human reference genome using Tophat,[25] and exon-inclusion levels were estimated using PennSeq.

RNA-Seq datasets and genotype data. We downloaded the RNA-Seq data produced by Lappalainen et al.[17] This dataset includes 91 lymphoblastoid B cell lines from the CEPH (CEU) population in the HapMap project. Each sample has approximately 10 million 75 bp paired-end reads, which were already mapped to the reference human genome (hg19, NCBI build 37) using the JIP pipeline. We downloaded the Phase 1 genotype data for 79 CEU samples generated by the 1000 Genomes Project.[26] The number of subjects who had both RNA-Seq and DNA genotype data is 78. To search for sQTLs, we identified all exon-trios in autosomal chromosomes and restricted analysis to *cis*-sQTLs, because various studies have shown that *trans*-sQTLs are less common.[15,17] Specifically, for each exon-trio, we restricted our analysis to SNPs within 200 kb on each side of the trio. For quality control purpose, we removed SNPs with Hardy–Weinberg P value < 0.0001 and genotype missingness $>5\%$. Because of the small sample size of the available data, we also removed SNPs with MAF <0.2. Multiple testing adjustment was performed with the Benjamini–Hochberg algorithm and an SNP was declared to be an sQTL if the FDR-adjusted P value was less than 0.05.

Results

Comparison of exon-inclusion level estimation. First, we compared the exon-inclusion levels estimated by GLiMMPS and PennSeq based on simulated data. Because of the narrow range of the exon-inclusion levels under the null model, we focused on those exon-trios from the alternative model in which the exon-inclusion level was influenced by an sQTL. For each of the 120 simulated individuals, we calculated the Pearson correlation coefficient between the estimated and the true values of the exon-inclusion levels. As expected, PennSeq yielded more accurate estimate than GLiMMPS. Among the 120 individuals, 102 (85%) had higher correlation coefficients in PennSeq than in GLiMMPS. The improvement in accuracy was also reflected in the root mean squared error, calculated as $\sqrt{\sum (\hat{Y} - Y)^2 / m}$, where m is the total number of exon-trios and the summation was taken over all exon-trios. The mean for root mean squared error of GLiMMPS was 0.16, whereas the mean for PennSeq was 0.13, which is significantly smaller than GLiMMPS (two-sample t-test P value $< 2.2 \times 10^{-16}$).

Comparison of FDR and power. Next, we compared the FDR of random effects meta-regression (denoted by PSMeta), beta regression (denoted by PSBeta), generalized linear mixed effects model with PennSeq estimates (denoted by PSGLMM), and GLiMPPS. We analyzed all 120 simulated individuals for sQTLs. To evaluate the impact of sample size, we generated samples of reduced sample size by randomly picking 60 and 90 individuals out of the 120. All evaluated methods had FDRs well below the 5% nominal level for all sample sizes we considered (Fig. 2A). However, GLiMMPS appeared to be overly conservative compared to the other

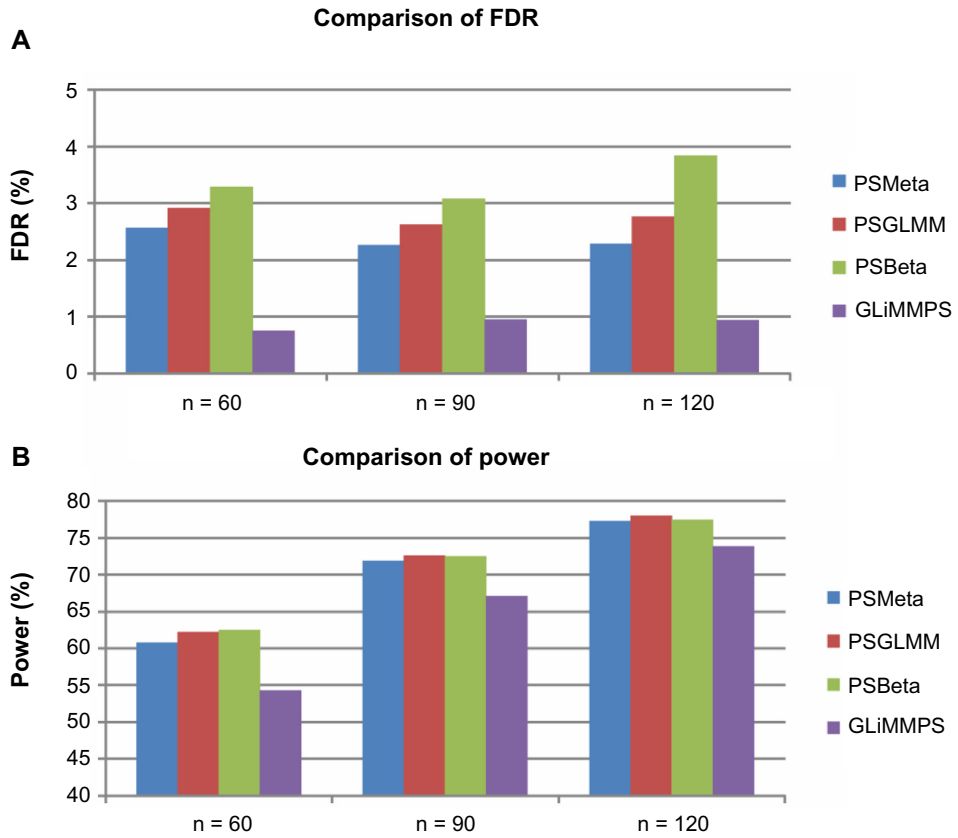

Figure 2. Comparison of false discovery rate and power using all simulated exon-trios. (**A**) The FDR was calculated as the fraction of false discoveries (ie, an SNP with FDR-adjusted *P* value < 0.05 but was not designated as an sQTL SNP in the simulation) among all discoveries (ie, an SNP with FDR-adjusted *P* value < 0.05). (**B**) Power was calculated as the fraction of SNPs with FDR-adjusted *P* values < 0.05 among all designated sQTL SNPs.

methods. Its estimated FDR was generally below 1%, which was several times smaller than the expected 5%.

Our main interest was to compare the power of various types of regressions with GLiMMPS. Unsurprisingly, due to its conservativeness, GLiMMPS had the lowest power among all models we evaluated (Fig. 2B). For example, when the sample size was 60, the power of GLiMMPS was 54%, whereas the powers of PSMeta, PSBeta, and PSGLM were all above 60%, which represent 12–15% improvement over GLiMMPS. The loss of power for GLiMMPS became less pronounced when sample size gets larger. When the sample size was 90, the power improvement of the other methods over GLiMMPS was around 7–8%, and when the sample size was 120, the power improvement was approximately 5–6%.

Impact of non-uniform read distribution. The assumption of uniform read distribution is one of the significant limitations of GLiMMPS. In order to appraise the real-life applicability of a method, understanding the impact of sequencing coverage non-uniformity on its performance is a critical step. In the GLiMMPS publication, Zhao et al informally evaluated the impact by introducing a random scaling factor in junction read counts.[18] However, read count is a simple summary of the original RNA-Seq data and is unlikely to capture all of the variations in non-uniformity present in raw reads. Using raw RNA-Seq data generated by the Flux Simulator, we can directly evaluate the impact of non-uniformity on the power of the various approaches. To quantify the degree of non-uniformity, we extracted the fraction of coverage, defined as the fraction of the transcript that is covered by reads, from the output of the Flux Simulator.

Based on this measure, we calculated the mean fraction of coverage across all samples and focused on those transcripts with mean coverage less than 1/3. A large fraction of the transcript body of these genes was not covered, which would lead to severe non-uniform read distribution. Figure 3A shows that the FDRs of PSMeta and GLiMMPS were under control, but PSBeta and PSGLMM had slightly inflated FDRs when sample size was 90 or 120. The power of all methods dropped substantially (Fig. 3B), especially for GLiMMPS. Compared to the power obtained from all simulated exon-trios, the loss of power for PSMeta, PSBeta, and PSGLMM was between 34 and 38%, whereas the power loss was 44% for GLiMMPS. Moreover, the power improvement of the other three methods over GLiMMPS was also more pronounced, especially when sample size was small. With 60 subjects, the power improvement of PSMeta over GLiMMPS was 23%, which is twice of the power improvement when all exon-trios were considered. These results suggest that when non-uniformity is a concern, using exon-inclusion levels obtained from junction reads alone can lead to substantial loss of power.

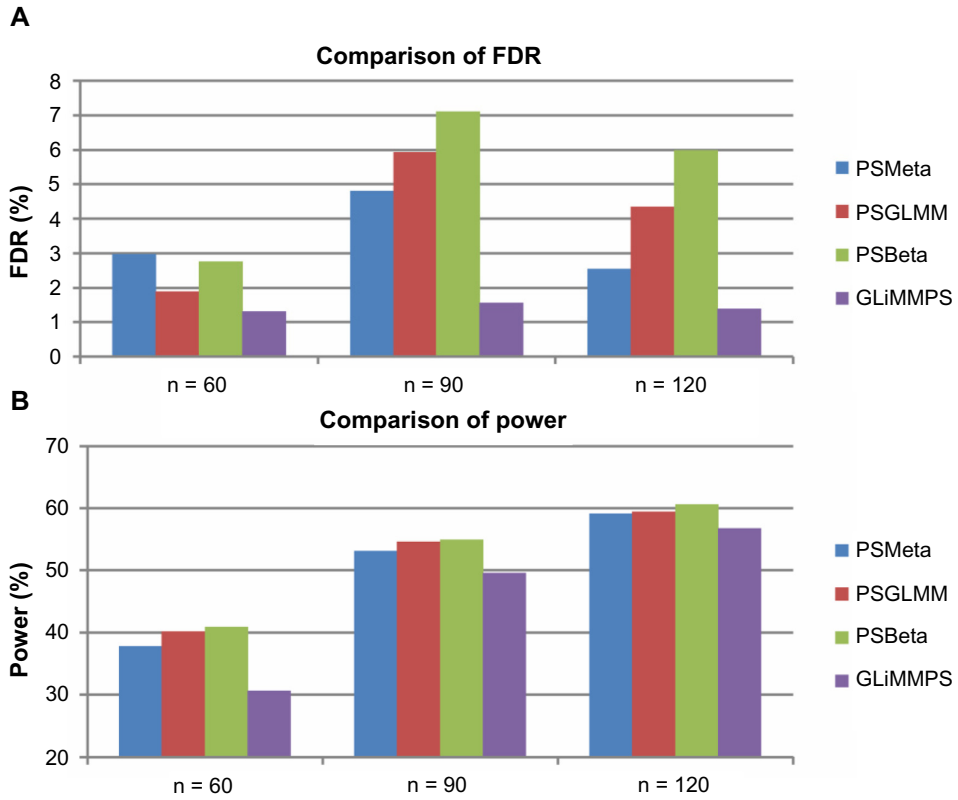

Figure 3. Impact of non-uniform read distribution. Only exon-trios with average percent transcript coverage less than 1/3 were included in the analysis, where the average percent transcript coverage was calculated across all 120 simulated subjects. (**A**) The FDR was calculated as the fraction of false discoveries (ie, an SNP with FDR-adjusted *P* value < 0.05 but was not designated as an sQTL SNP in the simulation) among all discoveries (ie, an SNP with FDR-adjusted *P* value < 0.05). (**B**) Power was calculated as the fraction of SNPs with FDR-adjusted *P* values < 0.05 among all designated sQTL SNPs.

Application to the real RNA-Seq dataset on CEU samples. To further assess the performance of different methods, we analyzed a real human RNA-Seq dataset on CEU lymphoblastoid cell line samples. We carried out sQTL analysis for common SNPs (MAF >0.2) within 200 kb from alternatively spliced exons because previous studies have suggested that sQTLs are often near the target exons.[12,15] To get reliable results, we only focused on those exon-trios in which the number of mapped reads is greater than 10 in all 78 subjects. Eventually, we were left with 3,694 exon-trios (in 2,070 genes), 640,105 SNPs, and 1,590,722 exon-trio-SNP pairs in the final comparison.

We found that the two generalized linear mixed effects model based approaches, PSGLMM and GLiMMPS, failed to converge for a large number of exon-trio-SNP pairs. Among the 1,590,722 exon-trio-SNP pairs we considered, GLiMMPS failed to converge for 35.75% of the pairs and PSGLMM failed for 50.07%. In contrast, PSBeta failed to converge for 0.92% of the pairs, and PSMeta failed to converge for only 1.75%. For the 549,095 pairs that failed in GLiMMPS, we observed an excess of *P* values less than 0.05 for PSMeta (binomial test *P* value $< 2.2 \times 10^{-16}$) and PSBeta (binomial test *P* value $< 2.2 \times 10^{-16}$).

To characterize and compare the performance of these methods in greater detail, we generated quantile–quantile plots with the *P* values generated by each model (Fig. 4). We also examined the numbers of sQTL SNPs identified by each method. We did not consider PSBeta and PSGLM due to their tendency of generating false-positive results. By performing FDR adjustment of the nominal *P* values on exon-trio-SNP pairs that converged for both PSMeta and GLiMMPS at the 5% level, PSMeta identified 7,845 significant pairs, involving 6,513 unique sQTL SNPs for 361 exon-trios located in 286 genes, whereas GLiMMPS identified 7,392 significant pairs, involving 6,233 unique sQTL SNPs for 271 exon-trios in 215 genes (Table 1). The number of sQTL SNPs that overlapped between PSMeta and GLiMMPS was 4,244. The number of sQTL SNPs that were identified by PSMeta but were missed by GLiMMPS was 2,269, whereas the number of sQTL SNPs that were identified by GLiMMPS but were missed by PSMeta was only 1,989. Figure 5 shows four randomly selected pairs that were identified by PSMeta but were missed by GLiMMPS. Consistent with our observations in simulated data, the estimated exon-inclusion levels from GLiMMPS showed less variation than estimates obtained from PennSeq, which could contribute the loss of power for GLiMMPS.

Next, we compared the performance of PSMeta and GLiMMPS independently by conducting FDR adjustment

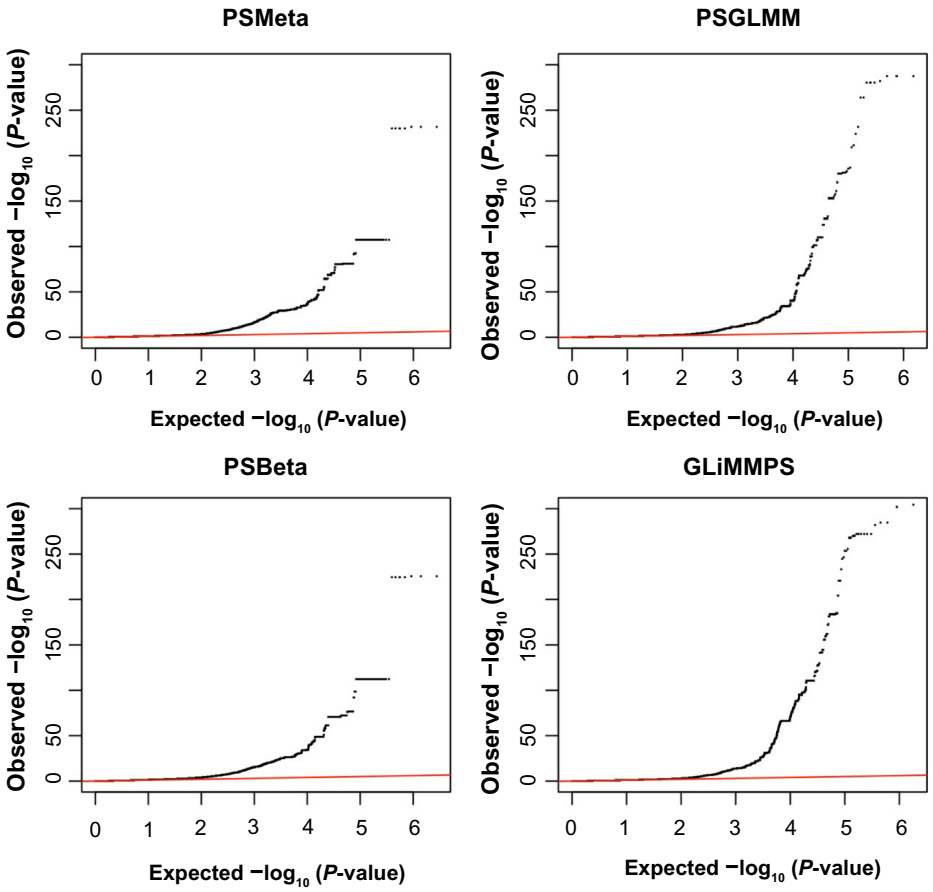

Figure 4. Comparison of quantile–quantile plots based on the CEU RNA-Seq data. Displayed are valid results from each method, ie, those exon-trio-SNP pairs that failed to converge were eliminated.

on exon-trio-SNP pairs that converged for each individual method. Since PSMeta converged for a significantly larger proportion of the pairs, we detected notably more sQTL SNPs in PSMeta as compared to GLiMMPS: 22,150 exon-trio-SNP pairs, containing 16,757 sQTL SNPs for 624 exon-trios in 447 genes for PSMeta, but only 7,409 exon-trio-SNP pairs, involving 6,251 sQTL SNPs for 272 exon-trios in 216 genes for GLiMMPS (Table 1). This result suggests that due to its intrinsic computational advantage, PSMeta greatly outperformed GLiMMPS in genome-wide search for sQTL SNPs.

Discussion

The advent of RNA-Seq has equipped us with a powerful tool to systematically search for sQTLs that regulate the pattern of alternative splicing. In sQTL analysis using RNA-Seq data, it is important to account for exon-inclusion level estimation uncertainty, directly model variation in the precision of exon-inclusion level estimates between samples, and allow for non-uniform read distribution along transcripts. In this study, we evaluated three statistical methods, including random effects meta-regression, beta regression, and generalized linear mixed effects model, for the analysis of sQTLs. In contrast to GLiMMPS, which uses junction reads only to quantify exon-inclusion levels, we used PennSeq,[20] a statistical method that utilizes all available reads and allows for non-uniform read distribution. Using both simulated and real RNA-Seq datasets, we demonstrated that PSMeta is the best performing method, and identified sQTLs at low FDRs but higher power when compared to GLiMMPS.

Table 1. Comparison of the numbers of sQTL SNPs identified by PSMeta and GLiMMPS based on the CEU RNA-Seq dataset. Displayed are the numbers of exon-trio-SNP pairs, exon-trios, and genes that contain a significant sQTL SNP.

COMPARISON	METHOD	EXON-TRIO-SNP PAIRS	UNIQUE sQTL SNPs	EXON-TRIOS	GENES
Converged for both methods	PSMeta	7,845	6,513	361	286
	GLiMMPS	7,392	6,233	271	215
Converged for each individual method	PSMeta	22,150	16,757	624	447
	GLiMMPS	7,409	6,251	272	216

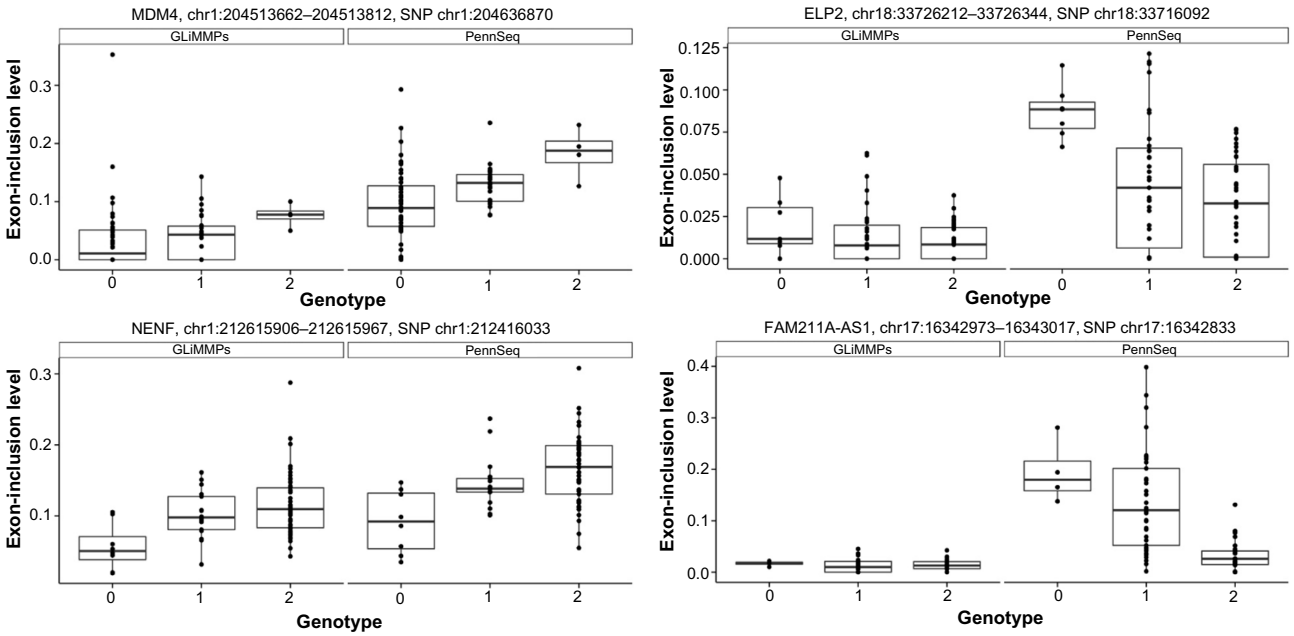

Figure 5. Illustrative examples of sQTL SNPs identified by PSMeta but were missed by GLiMMPS. The horizontal axis shows the genotype (represented by 0, 1, 2, counting the number of minor alleles) at each sQTL SNP, and the vertical axis shows the estimated exon-inclusion levels from PennSeq and GLiMMPS.

The main reason for power improvement of PSMeta over GLiMMPS is due to the efficient use of additional RNA-Seq read information in exon-inclusion level estimation. Closer examination of the simulated data showed that the exon-inclusion levels using junction reads only were less well estimated as compared to PennSeq, which uses all available reads including those from flanking constitutive exons. Another reason is that GLiMMPS cannot explicitly model paired-end data structure, but PennSeq can effectively utilize paired-end read information in its modeling. In paired-end RNA-Seq data with tight distribution of insert size, reads mapped to flanking constitutive exons can provide useful information about the exon-inclusion level. By using the generalized linear mixed effects model with estimates obtained from PennSeq, we confirmed that the power loss of GLiMMPS was due to the use of less-accurate estimate of exon-inclusion levels.

We also examined the impact of non-uniformity on the performance of different methods. The power of all methods decreased for exon-trios that demonstrate severe non-uniformity. Among the four methods we evaluated, PSGLMM and PSBeta had slightly inflated FDRs. The FDRs of both PSMeta and GLiMMPS were under control, but PSMeta had greater power. Overall, PSMeta appeared to be the most reliable yet powerful method for sQTL analysis.

In this study, we only focused on exon-skipping events, but the framework we presented can be easily generalized to examine other types of alternative splicing, including intron retention, alternative 5′ splice site, alternative 3′ splice site, and mutually exclusive exons. For example, for alternative splicing events that involve alternative 5′ splice site, we can treat the relative abundance of the isoform with alternative 5′ splice site as the quantitative trait in random effects meta-regression. Analysis for events that involve intron retention, alternative 3′ splice site, and mutually exclusive exons can be performed in a similar fashion.

In our analysis, we estimated the exon-inclusion levels first and then identified sQTLs using regression-based methods. This two-stage approach might be less powerful than identifying sQTLs using a one-stage approach, which avoids estimating exon-inclusion levels directly. We are currently pursuing extensions in this direction. Another possible direction of future research is to consider the overall splicing pattern of a gene by simultaneously considering the relative abundances of all isoforms of the gene. This analysis will give a single test statistic for each gene instead of one statistic for each exon-trio. We expect this approach to have increased statistical power due to its reduced burden of multiple testing.

Conclusions

In summary, we have evaluated three statistical methods for the analysis of sQTLs in RNA-Seq. As shown by both simulations and the analysis of real data, the most robust method is PSMeta, a random effects meta-regression-based approach. An appealing feature of PSMeta is that it can be easily implemented using existing software packages. Results from our study will be instructive for researchers in selecting the appropriate statistical methods for sQTL analysis.

List of Abbreviations

RNA-Seq, RNA sequencing; sQTL, splicing quantitative trait loci; MAF, minor allele frequency; FDR, false discovery rate.

Author Contributions

Designed the study: ML. Conducted the analysis: CJ, YH, and YL. Wrote the manuscript: CJ, ML. Made critical revisions: CJ, ML. All authors reviewed and approved the final manuscript.

REFERENCES

1. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470–6.
2. Menon R, Zhang Q, Zhang Y, et al. Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer. *Cancer Res*. 2009;69(1):300–9.
3. Menon R, Omenn GS. Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. *Cancer Res*. 2010;70(9):3440–9.
4. Menon R, Roy A, Mukherjee S, Belkin S, Zhang Y, Omenn GS. Functional implications of structural predictions for alternative splice proteins expressed in Her2/neu-induced breast cancers. *J Proteome Res*. 2011;10(12):5503–11.
5. Menon R, Im H, Zhang EY, et al. Distinct splice variants and pathway enrichment in the cell-line models of aggressive human breast cancer subtypes. *J Proteome Res*. 2014;13(1):212–7.
6. He C, Zhou F, Zuo Z, Cheng H, Zhou R. A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis. *PLoS One*. 2009;4(3):e4732.
7. Venables JP, Klinck R, Bramard A, et al. Identification of alternative splicing markers for breast cancer. *Cancer Res*. 2008;68(22):9525–31.
8. Shapiro IM, Cheng AW, Flytzanis NC, et al. An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genet*. 2011;7(8):e1002218.
9. Skotheim RI, Nees M. Alternative splicing in cancer: noise, functional, or systematic? *Int J Biochem Cell Biol*. 2007;39(7–8):1432–49.
10. Venables JP. Unbalanced alternative splicing and its significance in cancer. *Bioessays*. 2006;28(4):378–86.
11. Kwan T, Benovoy D, Dias C, et al. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet*. 2008;40(2):225–31.
12. Coulombe-Huntington J, Lam KC, Dias C, Majewski J. Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet*. 2009;5(12):e1000766.
13. Heinzen EL, Ge D, Cronin KD, et al. Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol*. 2008;6(12):e1.
14. Fraser HB, Xie X. Common polymorphic transcript variation in human disease. *Genome Res*. 2009;19(4):567–75.
15. Pickrell JK, Marioni JC, Pai AA, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. 2010;464(7289):768–72.
16. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*. 2010;464(7289):773–7.
17. Lappalainen T, Sammeth M, Friedländer MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501(7468):506–11.
18. Zhao K, Lu ZX, Park JW, Zhou Q, Xing Y. GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biol*. 2013;14(7):R74.
19. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*. 2010;7(12):1009–15.
20. Hu Y, Liu Y, Mao X, et al. PennSeq: accurate isoform-specific gene expression quantification in RNA-Seq by modeling non-uniform read distribution. *Nucleic Acids Res*. 2014;42(3):e20.
21. Huizenga HM, Visser I, Dolan CV. Testing overall and moderator effects in random effects meta-regression. *Br J Math Stat Psychol*. 2011;64(pt 1):1–19.
22. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511–5.
23. Simas AB, Barreto-Souza W, Rocha AV. Improved estimators for a general class of beta regression models. *Comput Stat Data Anal*. 2010;54(2):348–66.
24. Griebel T, Zacher B, Ribeca P, et al. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res*. 2012;40(20):10073–83.
25. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105–11.
26. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56–65.