Quantitative Biology 2014, 2(2): 71–79
DOI 10.1007/s40484-014-0031-9

Research article

Exon expression QTL (eeQTL) analysis highlights distant genomic variations associated with splicing regulation

Leying Guan[1,4], Qian Yang[1], Mengting Gu[2], Liang Chen[3] and Xuegong Zhang[1,2,*]

[1] MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST, Department of Automation, Tsinghua University, Beijing 100084, China
[2] School of Life Sciences, Tsinghua University, Beijing 100084, China
[3] Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA
[4] Department of Physics, Tsinghua University, Beijing 100084, China
* Correspondence: zhangxg@tsinghua.edu.cn

Received July 22, 2014; Revised September 10, 2014; Accepted September 12, 2014

Alternative splicing is a ubiquitous mechanism of post-transcriptional regulation of gene expression and produces multiple isoforms from the same genes. Expression quantitative trait loci (eQTL) has been a major method for finding associations between gene expression and genomic variations. Differences in alternative splicing isoforms are resulted from differences in the expression of exons. We propose to use exon expression QTL (eeQTL) to study the genomic variations that are associated with splicing regulation. A stringent criterion was adopted to study gene-level eQTLs and exon-level eeQTLs for both *cis*- and *trans*- factors. From experiments on an RNA-sequencing (RNA-Seq) data set of HapMap samples, we observed that compared with eQTLs, more eeQTL *trans*-factors can be found than *cis*-factors, and many of the eeQTLs cannot be found at the gene level. This work highlights that the regulation of exons adds another layer of regulation on gene expression, and that eeQTL analysis is a new approach for investigating genome-wide genomic variations that are involved in the regulation of alternative splicing.

Keywords: eeQTL; eQTL; alternative splicing; *trans*-factor; association; regulation

INTRODUCTION

The expression of a gene as measured by the abundance of its mRNA transcripts has been taken as a basic unit for the quantitative study of the gene's function and regulation. Gene transcription is regulated by transcription factors interacting with genomic elements in the promoter and enhancer regions. Genomic variations that affect those elements or transcription factors are an important cause of variations in gene expression. Expression quantitative trait locus (eQTL) analysis studies the association of genomic variations with gene expression variations, and is a key to decode the molecular network that regulates the expression of genes [1–6]. Genome-scale eQTL analyses have discovered many genomic variations that can affect gene expression, including *cis*-eQTLs located inside or close to the gene which highlight genomic regions with regulatory elements, and *trans*-eQTLs that are far away from the gene or on different chromosomes which highlights regions associated with transcription factors that regulate the gene or possible long-range DNA interactions, e.g., [3,7–15].

Most human genes are composed of multiple exons and the procedure of transcription is accompanied by splicing. Alternative splicing is the mechanism that allows a gene to produce multiple isoforms of transcripts by using different combinations of exons in the splicing. This is an important mechanism of co- or post-transcriptional regulation in humans and other high organisms. Different isoforms and differences in their quantitative proportions can have important biologic consequences [16–19]. It has been widely observed that many alternative splicing

events are associated with complex diseases such as cancers [20–22]. For example, the spleen tyrosine kinase gene (*SYK*) has an effect on breast cancer when it undergoes aberrant alternative splicing [11]. Later its alternative splicing is found as a regulator of mitosis and cell survival, and has effects on multiple types of tumors [23–27]. Alternative splicing was found as a rare event when it was discovered in 1970s. For a long period, it had been understood that alternative spliced genes may compose about 40-60% of human genes [16]. Recent advancement in next-generation sequence (NGS) for RNAs (RNA-Seq) experiments has found that actually most human genes can have alternative splicing [10,28,29]. However, the study of alternative splicing regulation in eQTL studies is still at the starting phase.

Alternative splicing adds another layer of regulation to the system of gene expression. With the ubiquitous existence of alternative isoforms, the concept of gene expression becomes less well defined. As different isoforms of the same gene can have differences in their functions, it is natural to use the expression of each isoform to replace the "gene expression" in earlier studies. Computational biologists have developed bioinformatics tools that can map RNA-Seq reads to isoforms instead of genes [30–35], infer new isoforms that have not been annotated in databases [36–38], and estimate the expression levels of isoforms [9,39–42]. The tasks are challenging and cannot be exempt from errors as the current NGS technology only provides short sequencing reads, and cannot obtain reads of the full-length transcript directly. Sequenced short reads can be the mixture of multiple isoforms, besides being affected by noises and biases. Unique solutions are not mathematically guaranteed or accomplishable in many situations, and the inferences or estimations usually depend on assumptions that might not fit the biologic truth. Therefore, exon-centric methods have also been proposed to detect genes that show differential splicing between compared samples [20,43]. Such methods do not depend on annotations or assumptions about splicing isoforms, and leave the task of inferring and estimating isoform expression to the downstream analysis of only the fewer detected genes. Similar to this idea, junction reads have also been used to connect two spliced exons to define concepts such as exon-inclusion levels or splicing levels to quantify the relative expression of isoforms [28,44].

Microarrays had been the major technique to measure gene expression as well as genomic variations for the last decade, and were also the major technique for eQTL studies [45–51]. RNA-Seq revolutionized the technology and provides higher resolution and accuracy in both measuring genomic variations and RNA expressions. Several recent studies have used RNA-Seq data and sequencing-based single nucleotide polymorphisms (SNPs) data to perform eQTL studies [3,4,52–54]. They investigated associations of SNPs with gene expressions at the gene level, the splicing level or the exon level with RNA-Seq data, aiming to reveal more complex regulatory relations of genetic variations with gene expression. Hull et al. still using microarray data, reported that splicing patterns of exons depend on the single nucleotide polymorphisms distributed in flanking introns or exons [45]. Pickrell et al. used RNA-Seq data to discover a large number of eQTLs and splicing QTLs, and concluded that an exon's inclusion is affected by the variation within and near the consensus splice sites [3]. Montgomery et al. conducted an eQTL study at gene, transcript and exon levels with RNA-Seq data of 60 European individuals and identified more eQTLs than with microarrays [9], which showed the potential of using exons as the functional unit in eQTL study. Heinzen et al. carried out eQTL mapping and exon eQTL mapping in human primary cells with exon-level microarrays and suggested that "splicing effects may be of more phenotypic significance than overall gene expression changes" [47]. Lalonde et al. identified many isoform eQTLs which are located near splice site and influence the splicing of cassette exons [52]. Recently, Lappalainen et al. systematically mapped *cis*-eQTLs for exon quantifications that can capture both gene expression and splicing variation, and reported a large number of *cis*-regulatory eQTLs of various types, including gene eQTLs, exon eQTLs, transcript ratio QTLs, miRNA eQTLs and transcribed repeat eQTLs. They showed that the genetic loci affecting transcript structures are largely independent of gene eQTLs and they are both common in humans [54]. These studies bring new insights on the complexity of transcription and splicing regulation, and also highlight the importance to study gene expression at the level of exons.

We had proposed to use exon expression as a basic unit in studying transcriptomes with RNA-Seq data, especially for studying differential splicing patterns [55]. Although isoforms are the basic unit of function in the current understanding, all isoforms are composed of expressed exons, and all analyses on gene expression and isoform expression are based on short reads mapped to exons and their junctions. Due to the limited read length, noises and biases in sequencing data and the intrinsic complexity of possible isoform compositions, inferring isoforms and estimating their expressions can introduce extra inaccuracy or uncertainty. On the other hand, splicing can be viewed as the regulation of exon expression on top of the transcriptional regulation of the whole gene. This understanding has been confirmed from the observations in several recent studies, e.g., [54]. In this work, we study the association of genomic variations with exon expression variations using RNA-Seq data and call it exon-expression QTL mapping or eeQTL mapping. The

expression of an exon is regulated by both transcription and splicing. We conducted both gene-level eQTL and exon-level eeQTL mapping for both *cis*- and *trans*- loci, based on an RNA-Seq data set of the samples from the HapMap Project [56]. Stringent computational protocols were adopted to ensure a low false positive rate in the discovery. We found that there are a noticeable amount of exons that have significant eeQTLs that are not eQTLs of their host gene, especially *trans*-eeQTLs that are distant from the target gene or are on a different chromosome. Further study on such eeQTLs will bring new understanding to the regulation of exon expression that is independent with the transcriptional regulation.

DATA AND METHODS

The RNA-Seq data we used in this study were from the study by Pickrell et al. [3]. They sequenced RNA from 69 lymphoblastoid cell lines derived from unrelated Nigerian individuals. The individuals are genotyped by the International HapMap Project [56]. We used the release 27 of the HapMap genotypes obtained from http://www.hapmap.org. We left out individuals with $> 50\%$ missing genotype values and also SNPs with more than 3 missing values (5%) among the individuals. After the filtering, the data used in this study include 54 individuals and about 14 million SNPs. There were still missing values of around 1% in the remaining data after this filtering. We used the software package BIMBAM [57] to do missing data imputation in the data. It is based on the fastPHASE model [58], and we set the parameters as to run the EM algorithm 5 times with 20 steps per run. After the filtering and imputation, we got the genotype value of 0, 1 or 2 at each studied SNP site.

The RNA-Seq reads were mapped to the genome using gene models of the Ensembl database in the original work [3]. Following the procedures used by Pickrell et al. [3], we applied a series of pre-processing steps on RNA-Seq read counts at exons and genes to get the normalized gene and exon expression data. The steps include GC-content correction, correction for possible batch effects, principal component analysis to remove confounders, and two rounds of normalization. Sequence coverage can be influenced by the GC-content of the region which can cause biases in estimating gene and exon expression. We binned exons according to their GC contents and applied smoothing on them in each sequencing lane [3,59,60]. As the data were sequenced at two centers, and some were at different concentrations, we also adopted the correction step as described in [3] to compensate for differences between the two centers as well as different concentrations. Quantile normalization was applied on the expression data to make them follow a Gaussian distribution, which is a prerequisite of the ANOVA method used for detecting associations. Then principal component analysis was applied on the data to remove unmeasured confounders in the correlation between different individuals. We used the same setting of removing 16 principal components which has been reported to give the largest number of eQTLs in downstream analysis in [3]. Another round of quantile normalization was applied on the residuals.

We are equally interested in eQTLs and eeQTLs in local regions of the gene and in the whole genome. Therefore, unlike the strategy used in [3] and most other eQTL literature that detected local associations and genome-wide associations separately with different models and parameters, we used ANOVA to test for the associations of gene and exon expressions with genotypes of all the studied genome-wide SNPs. This will give associations with SNPs of different distances equal opportunity to be detected. We controlled the false discovery rate (FDR) to be less than 25%. The FDR was estimated via permutation and was controlled for possible associations on the whole genome. The permutation was done by randomly shuffling the expression of genes and exons, and we detected associations of the shuffled gene and exon expression with genotypes using the same method. Note that the whole-genome genes or exons for each individual were handled together during the shuffling. Therefore, the dependence among genes and the dependence among markers were kept during the permutation. Any detected association on the permuted data will be false discovery. As the expression of genes and exons has been normalized, we can treat all genes and exons using the same threshold of the p-value. At a given p-value level, we calculate the ratio of the number of discoveries obtained on the permuted data to the number of discoveries on the real data, and use the p-value at which this ratio becomes 25% as the threshold for controlling FDR $\leqslant 25\%$ on the real data.

In most existing work on eQTL, candidate SNPs were restricted to a nearby region of the gene for *cis*-eQTL study as the discovery power will be too low if all SNPs on the genome is considered. This made the standards for calling local associations and genome-wide associations very different, and is a major reason why many *cis*-eQTLs could be identified but *trans*-eQTLs were few. In our study, we took all SNPs on the genome equally for eQTL and eeQTL study with the same model and same threshold. This gives equal opportunity for discovering *cis*- and *trans*-loci. But since the number of candidate genome-wide SNPs are several magnitudes larger than those in the *cis*-region, it can be expected that the discovery power for *cis*-loci becomes lower than existing reports.

RESULTS

The key question we investigated was the possible differences between exon-expression QTLs (eeQTLs) and gene expression QTLs (eQTLs). Therefore, we focused on only the 929 genes that have been reported to have eQTLs on this data set in the study of Pickrell et al. [3] to limit the computational workload of this study. In the original work, the authors focused on SNPs located within 200 kb of target genes and identified 929 genes that have local eQTLs at the FDR level of 10%. These genes contain a total of 9,552 exons. We searched both local and distant eQTLs and eeQTLs within these genes by a more stringent criterion, with SNPs on the whole genome as candidates. Expanding to whole-genome SNPs makes the power of detecting eQTLs and eeQTLs much lower, but it can give equal opportunities for eQTLs and eeQTLs at different distances from the target genes. For gene expression, we reported only 411 significant eQTL associations (FDR $\leqslant$ 25%) that involve 77 genes and 411 SNPs. For exon expression, we found a total of 1,302 significant eeQTL associations, which belong to 138 genes and 408 SNPs.

Since we only called eQTLs or eeQTLs in a small fraction of the 929 genes that have been reported to have eQTLs, we double-checked the p-value of the gene in our study. We ranked the genes according to the minimal p-value of each gene with all SNPs. The largest minimal p-value among those genes, obtained on the gene ENSG00000209849, is 1.16e-05, which is at the same level with the p-value 1.96e-05 reported on this gene by Pickrell et al. [3] at http://eqtl.uchicago.edu. But since we used genome-wide SNPs in controlling for the multiple testing, the threshold for eQTL p-value is 1.93e-07 and for eeQTL p-value is 5.2e-08 for FDR $\leqslant$ 25%. This confirmed that the lower calling rate in our work is due to the more stringent criterion we used for calling significant associations.

Associations of mapping fall into different categories as *cis*- and *trans*- effects. In eQTL mapping literature, it is a common practice to define *cis*- and *trans*-eQTLs based on the genomic distance between target genes and SNPs, but there has been no precise definition of the threshold to discriminate distances of *cis*- and *trans*- loci. Some authors including Pickrell et al. used 200 kb as the boundary between *cis*- and *trans*-eQTL [3], and some others used 100 kb [1]. For eeQTLs, since they may be associated with splicing regulation, it is expected that *cis*-elements for splicing be closer to the exon and within the gene region. But a solid definition is still missing. In our study, we categorized three types of eQTLs and eeQTLs by considering the distance between gene/exon and its associated SNPs: *local* eQTLs and eeQTLs that are located close to the target gene on the same chromosome (distance $\leqslant$ 100 kb), *distant* eQTLs and eeQTLs that are located far away from the target gene (distance $>$ 100 kb) but are on the same chromosome, and *external* eQTLs and eeQTLs that are located on different chromosomes from the target gene. The external loci and distant loci most likely correspond to SNPs associated with *trans*-factors that regulate the expression of genes or exons by transcriptional or splicing regulation. The local loci are more likely to be associated with *cis*-elements of genes or exons that receive the regulation signals. Different thresholds for the distinction between local and distance loci might lead to different observations on the relative number of local or distant associations. We had experimented with different thresholds from 10 kb to 200 kb, and the general observations are consistent although the specific numbers will change.

Table 1 summarizes the numbers of significant associations we identified and the numbers of genes and SNPs involved with the associations of the three categories. We can observe that there are more local eQTLs than distant and external eQTLs. This is expected as all the discoveries were done within genes that have been reported to have local eQTLs in the neighborhood of the genes in the previous study [3]. We reported fewer eQTL/eeQTL genes because we adopted more stringent criteria in this work. However, when looking at the exon level, we observed that there are more genes detected with distant and external eeQTLs than genes with local eeQTLs (110 vs. 36), although the number of local associations is larger (245 vs. 1057). For gene level eQTL, we didn't see such a trend, and there are more genes and associations found at local loci than distant and external loci. This is more obvious if we calculate the ratio of the number of eeQTLs to the number of eQTLs

Table 1. Summary on numbers of significant eQTL and eeQTL associations, genes and SNPs

		All	Local loci	Distant loci	External loci
Number of associations	Gene level (eQTLs)	411	315	29	67
	Exon level (eeQTLs)	1302	1057	103	142
Number of genes	Gene level (eQTLs)	77	50	10	31
	Exon level (eeQTLs)	138	36	14	96
Number of SNPs	Gene level (eQTLs)	411	315	29	67
	Exon level (eeQTLs)	408	244	33	141

identified in the same category, as shown in Table 2. This is a strong indication that more trans-regulatory elements can be identified when we look at the expression at the exon level instead of the gene level.

From Tables 1 and 2, we can see that in general, more associations can be found at exon level than at gene level, and so is the number of genes involved in the associations. However, for local loci which are very possibly *cis*-elements of regulation, although 3-fold more associations can be found at the exon level, the number of SNPs involved and the number of genes identified at exon level is less than the numbers identified at the gene level. This observation brought some surprise, since if there is no alternative splicing events and if the read distribution on the exons of the same gene is uniform, genes that show significance eQTLs should also be detected at most of their exons. Even if there is alternative splicing event, the associations detected at gene level should still be reflected at some of the constitutive exons of the genes. The possible reason is that since most genes are composed of multiple exons, the signal of gene expression, which sums up the signals of all its exons, is stronger and more robust to noises. Once the signal of gene expression is spread among its multiple exons, especially for those genes with lower expression and therefore less read coverage at each exon, the signal becomes weaker and less powerful to detect significant associations. We have observed that associations with genes of lower expression levels are less likely to be detected at their exons. Besides, the number of candidate exons is about 10 times larger than the number of candidate genes. This makes the p-value threshold for calling an eeQTL more stringent than that for calling an eQTL at the same FDR level.

Table 3 gives another view of the overlap of genes discovered at the gene level and at the exon level. If a gene that has an eQTL is also detected to have eeQTL for at least one exon of that gene, we call it as a shared gene. If a gene with an eQTL is not detected to have any eeQTL for any of its exons, we call the association as a gene-only association. If a gene with no eQTL is detected to have an eeQTL for one of its exons, we call the association as an exon-only association. Table 3 summarized the number of genes of those situations in the local, distant and external categories. Similarly, the eQTLs and eeQTLs can also be categorized as shared, gene-only and exon-only loci. If an eQTL for a gene is detected as an eeQTL of at least one exon of the gene, or equally if an eeQTL of one exon is also detected as an eQTL for the gene hosting the exon, we call it as a shared locus. If an eQTL of a gene is not detected as a significant eeQTL for any of its exons, we call the locus as gene-only locus. And if an eeQTL is detected for an exon but is not detected as an eQTL for the gene hosting the exon, we call it as an exon-only locus. Table 4 summarized that number of loci of those situations in the three categories. Note that some of the shared genes might have different loci for the eQTL and eeQTL, and some of the gene-only and exon-only loci may be of the same gene. It can happen that a gene is detected with an eQTL and one of its exons is also detected with an eeQTL (so the gene is a shared gene), but the eQTL and eeQTL are not on the same loci (so the loci are one gene-only locus and one exon-only locus).

Tables 3 and 4 further strengthened the observation from Tables 1 and 2. For the majority of genes that have been detected to be associated with some SNPs in this data set, the associations were only detected at the exon expression level (106 among $183 = 32 + 106 + 45$). Also many detected SNPs only have associations with expression of exons but not of the whole genes (247 among 658). Especially, while most published results on eQTL and/or splicing eQTL had put more attention on *cis*-loci and reported few *trans*-loci, we observed that when we study the expression at the exon level and put equal attention on all genome-wide SNPs, more *trans*-

Table 2. Ratios of numbers of eeQTLs over eQTLs in each category

	All	Local loci	Distant loci	External loci
Number of associations	3.17	3.36	3.55	2.12
Number of genes	1.79	0.72	1.40	3.10
Number of SNPs	0.99	0.77	1.14	2.10

Table 3. Numbers of genes as shared, gene-only and exon-only

Categories	Shared*	Exon-only	Gene-only
Local	22	14	28
Distant	4	10	6
External	5	91	26
All	32	106	45

* For the "local," "distant" and "external" categories, a gene is counted as "shared" only when the eQTL and eeQTL are in the same category. But in "all," a gene is counted as "shared" if only it has both an eQTL and an eeQTL, not necessarily of the same category. Similar strategy was used for the "exon-only" and "gene-only" classes, and also in Table 4.

Table 4. Numbers of loci detected as shared, gene-only and exon-only

Categories	Shared	Exon-only	Gene-only
Local	147	97	168
Distant	12	21	17
External	1	140	66
All	161	247	250

eeQTLs (external and distant eeQTLs) can be found than *cis*-eeQTLs (local eeQTLs) in numbers. This phenomenon has not been observed at the gene level eQTL, indicating that they are hard to be detected at gene expression level. The distant and external eeQTL SNPs are very likely linked with *trans*-factors that regulate gene splicing.

The definition of local vs. distant loci on the same chromosome is based on the distance between the SNP of the gene with which eQTL or eeQTL is detected. The threshold is set as 100 kb in the above experiments. We did a series of experiments with different settings of the threshold, and found that for the choices from 10 kb to 200 kb, the specific numbers corresponding to those in Tables 1-4 have some changes but the overall observation on the trends and on the comparison of eQTL vs. eeQTL results does not change.

In Figs. 1 and 2, we picked up a few examples that have exon-only eeQTLs detected to illustrate the situation that exon-level associations cannot be detected at the gene level. Figure 1 is the example of the gene ENSG00000077984 (gene *CST7*) on chr20 with the SNP rs1036333 on chr2. A significant eeQTL was found in the 2nd exon of the gene with this SNP with p-value of 2.20e-08 but significant eQTL was not found at the gene level. The SNP rs1036333 was in the intron region of the gene *PLCL1* on chr2. We can see significant differences in the expressions between the genotypes at the exon level but not the gene level. Figure 2 is the example of ENSG00000160392 (gene *C19* or *F47*) on chr19 with the SNP rs11882778, about 1.1 Mb away from the gene on the same chromosome. Most of the samples are of the CT or CC genotypes. The two groups have very similar distribution in the gene expression but a significant difference can be seen in the exon expression. The genes with detected eeQTLs and eQTLs are listed in the Supplementary File 1 (for eeQTLs) and Supplementary File 2 (for eQTLs).

DISCUSSION

The ubiquitous existence of alternative splicing events in human genes has made people to pay more attention to the expression and regulation of alternative isoforms. Isoforms are composed of combinations of exons and the quantitative regulation of isoforms must be implemented by the regulation of exons in the splicing procedures. On the other hand, current sequencing technology can only

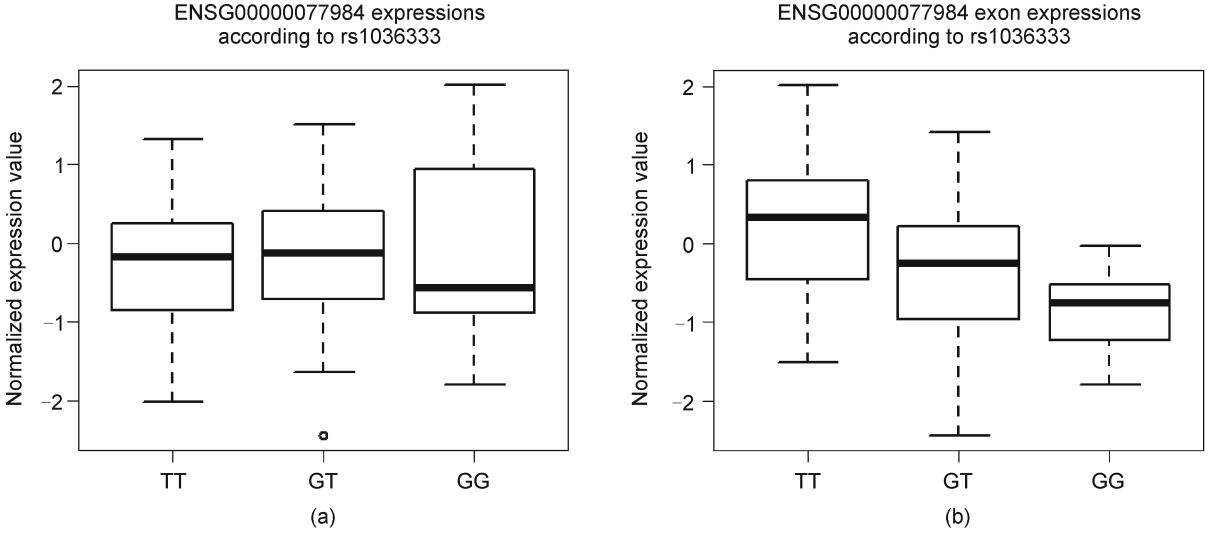

Figure 1. The association of gene ENSG00000077984 on chr20 with SNP rs1036333 on chr2. The horizontal axis is the genotype and the vertical axis is the normalized expression value. (a) Box-plots of gene expression of different genotypes. (b) Box-plots of the expression of the 2nd exon of different genotypes. The numbers of samples of each genotype are: TT: 21, GT: 26, GG: 7, NN: 0. Significant QTL was detected for the exon expression but not for the gene expression.

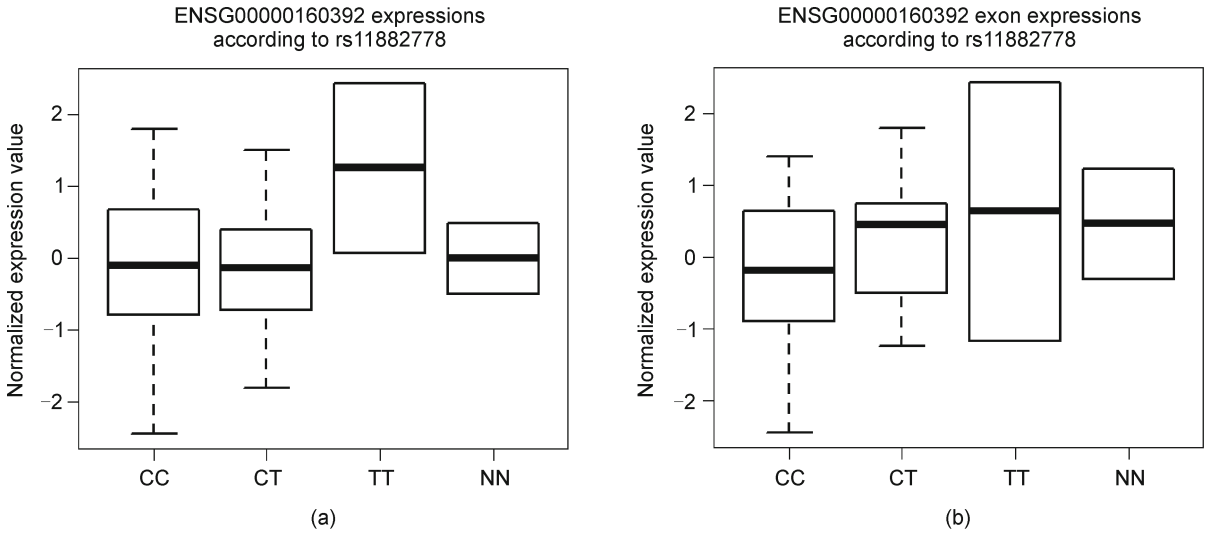

Figure 2. The association of gene ENSG00000160392 with SNP rs11882778, about 1,100 kb from each other on chr19. The horizontal axis is the genotype and the vertical axis is the normalized expression value. (a) Box-plots of gene expression of different genotypes. (b) Box-plots of expression of the 8th exon of different genotypes. The numbers of samples of each genotype are: CC: 32, CT: 18, TT: 2, NN: 2. Significant QTL was detected for the exon expression but not for the gene expression.

measure short reads and cannot cover whole mRNAs. The quantitative estimation of isoform expression is based on the read counts on the exons and their junctions. Therefore, it is more natural to study the expression and regulation of alternative splicing isoforms at the exon level instead of at the gene or isoform level. In this work, we proposed to use exon expression levels to replace gene expression levels in eQTL study and called this strategy as eeQTL or exon expression QTL. We followed the existing methods for data preprocessing and QTL mapping, but corrected for multiple testing by controlling the false discovery rate for all genome-wide SNPs instead of only the candidate SNPs in a selected region. This introduces no pre-assumption on the location of the regulatory loci and gives equal opportunity for discovering *cis*- and *trans*- factors. The computational experiments show that for gene-level eQTL study, more *cis*-loci are detected than *trans*-loci, but the proportion of identified *trans*-loci is larger than existing studied which reported mostly *cis*-loci. On the other hand, for exon-level eeQTL study, there are significantly more *trans*-loci being found than *cis*-loci. Many of the exon-level eeQTLs do not show any significant associations with SNPs at the gene level. These observations suggest that regulation of exons by *trans*-factors adds another layer of regulation after transcriptional regulation, and exon expression QTL study can be a powerful approach for detecting such regulatory factors.

The presented work is still quite preliminary since it only analyzed the genes that have been previously reported to have eQTL in this data set, and also the criterion adopted in this work is very conservative. Also the possible functional association of the discovered eeQTL SNPs with alternative splicing has not been further explored. But the observations can be a proof of concept for a more systematic survey for this direction. As more RNA-Seq data in multiple tissues accompanied with genotype data are becoming available, it's the time for a complete study of genetic variations that are associated with and possibly responsible for the regulation of alternative splicing in cooperation with transcriptional regulation by integrating eQTL studies at the gene level, the exon level and the splicing junction level.

From the methodological viewpoints, there are still much open questions for eeQTL studies. Variations in exon expression levels can be affected by changes in steady-state gene expression level, or changes in exon splicing activities, or both. As discussed above, the signals at exon levels are also weaker and more sensitive to sequencing noises and biases than at gene levels. Methods for deconvoluting signals of the transcription regulation and splicing regulations need to be studied. Using junction reads to calculate splicing QTLs or transcript QTLs is another strategy for mapping genomic variations associated with splicing regulation, e.g., [54,61,62]. For genes and isoforms with reasonable coverage on junction reads, such methods can identify splicing variations more efficiently. It can be expected that by integrating observations from eQTL, eeQTL and splicing QTL studies, we will be able to gain a more systematic understanding on the nature of genomic regulations on gene expression and alternative splicing.

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at DOI 10.1007/s40484-014-0031-9.

ACKNOWLEDGEMENTS

This work is partially supported by the National Basic Research Program of China (2012CB316504), the Hi-tech Research and Development Program of China (2012AA020401), NSFC Grant (91010016), and the National Institute of General Medical Sciences (R01GM097230).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Leying Guan, Qian Yang, Mengting Gu, Liang Chen and Xuegong Zhang declare that they have no conflict of interests.
This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Gilad,Y., Rifkin, S. A. and Pritchard, J. K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet., 24, 408–415
2. Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S. and Cheung, V. G. (2004) Genetic analysis of genome-wide variation in human gene expression. Nature, 430, 743–747
3. Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J. B., Stephens, M., Gilad, Y. and Pritchard, J. K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature, 464, 768–772
4. Majewski, J. and Pastinen, T. (2011) The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet., 27, 72–79
5. Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., et al. (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature, 422, 297–302
6. Rockman, M. V. and Kruglyak, L. (2006) Genetics of global gene expression. Nat. Rev. Genet., 7, 862–872
7. Xia, K., Shabalin, A. A., Huang, S., Madar, V., Zhou, Y. H., Wang, W., Zou, F., Sun, W., Sullivan, P. F. and Wright, F. A. (2012) seeQTL: a searchable database for human eQTLs. Bioinformatics, 28, 451–452
8. Yang, T. P., Beazley, C., Montgomery, S. B., Dimas, A. S., Gutierrez-Arcelus, M., Stranger, B. E., Deloukas, P. and Dermitzakis, E. T. (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. Bioinformatics, 26, 2474–2476
9. Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R. and Dermitzakis, E. T. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature, 464, 773–777
10. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009) Mapping complex disease traits with global gene expression. Nat. Rev. Genet., 10, 184–194
11. Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol., 6, e107
12. Myers, A. J., Gibbs, J. R., Webster, J. A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P., et al. (2007) A survey of genetic human cortical gene expression. Nat. Genet., 39, 1494–1499
13. Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., et al. (2007) Population genomics of human gene expression. Nat. Genet., 39, 1217–1224
14. Veyrieras, J. B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M. and Pritchard, J. K. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet., 4, e1000214
15. Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H., et al. (2010) Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. PLoS ONE, 5, e10693
16. Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A. and Soreq, H. (2005) Function of alternative splicing. Gene, 344, 1–20
17. Graveley, B. R. (2001) Alternative splicing: increasing diversity in the proteomic world. Trends Genet., 17, 100–107
18. Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. Nat. Genet., 30, 13–19
19. Brett, D., Pospisil, H., Valcárcel, J., Reich, J. and Bork, P. (2002) Alternative splicing and genome complexity. Nat. Genet., 30, 29–30
20. Gardina, P. J., Clark, T. A., Shimada, B., Staples, M. K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., et al. (2006) Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics, 7, 325
21. Venables, J. P. (2004) Aberrant and alternative splicing in cancer. Cancer Res., 64, 7647–7654
22. Garcia-Blanco, M. A., Baraniak, A. P. and Lasda, E. L. (2004) Alternative splicing in disease and therapy. Nat. Biotechnol., 22, 535–546
23. Wang, L., Duke, L., Zhang, P. S., Arlinghaus, R. B., Symmans, W. F., Sahin, A., Mendez, R. and Dai, J. L. (2003) Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer. Cancer Res., 63, 4724–4730
24. Goodman, P. A., Wood, C. M., Vassilev, A., Mao, C. and Uckun, F. M. (2001) Spleen tyrosine kinase (Syk) deficiency in childhood pro-B cell acute lymphoblastic leukemia. Oncogene, 20, 3969–3978
25. Nakashima, H., Natsugoe, S., Ishigami, S., Okumura, H., Matsumoto, M., Hokita, S. and Aikou, T. (2006) Clinical significance of nuclear expression of spleen tyrosine kinase (Syk) in gastric cancer. Cancer Lett., 236, 89–94
26. Prinos, P., Garneau, D., Lucier, J. F., Gendron, D., Couture, S., Boivin, M., Brosseau, J. P., Lapointe, E., Thibault, P., Durand, M., et al. (2011) Alternative splicing of SYK regulates mitosis and cell survival. Nat. Struct. Mol. Biol., 18, 673–679
27. Feng, H., Qin, Z. and Zhang, X. (2013) Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. Cancer Lett., 340, 179–191
28. 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
29. 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

30. Marco-Sola, S., Sammeth, M., Guigó, R. and Ribeca, P. (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. Nat. Methods, 9, 1185–1188
31. Chen, L. Y., Wei, K. C., Huang, A. C., Wang, K., Huang, C. Y., Yi, D., Tang, C. Y., Galas, D. J. and Hood, L. E. (2012) RNASEQR—a streamlined and accurate RNA-seq sequence analysis program. Nucleic Acids Res., 40, e42
32. Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105–1111
33. Wu, J., Anczuków, O., Krainer, A. R., Zhang, M. Q.and Zhang, C. (2013) OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucleic Acids Res., 41, 5149–5163
34. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15–21
35. Wang, L., Wang, X., Wang, X., Liang, Y. and Zhang, X. (2011) Observations on novel splice junctions from RNA sequencing data. Biochem. Biophys. Res. Commun., 409, 299–303
36. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol., 28, 511–515
37. Roberts, A., Pimentel, H., Trapnell, C. and Pachter, L. (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics, 27, 2325–2329
38. Li, W., Feng, J. and Jiang, T. (2011) IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. J. Comput. Biol., 18, 1693–1707
39. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol., 31, 46–53
40. Ma, X. and Zhang, X. (2013) NURD: an implementation of a new method to estimate isoform expression from non-uniform RNA-seq data. BMC Bioinformatics, 14, 220
41. Jiang, H. and Wong, W. H. (2009) Statistical inferences for isoform expression in RNA-Seq. Bioinformatics, 25, 1026–1032
42. Wu, Z., Wang, X., Zhang, X. (2011) Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. Bioinformatics, 27, 502–508
43. Richard, H., Schulz, M. H., Sultan, M., Nürnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., et al. (2010) Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. Nucleic Acids Res., 38, e112
44. Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R. and Shoemaker, D. D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science, 302, 2141–2144
45. Hull, J., Campino, S., Rowlands, K., Chan, M. S., Copley, R. R., Taylor, M. S., Rockett, K., Elvidge, G., Keating, B., Knight, J., et al. (2007) Identification of common genetic variation that modulates alternative splicing. PLoS Genet., 3, e99
46. Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., Hudson, T. J., Sladek, R. and Majewski, J. (2008) Genome-wide analysis of transcript isoform variation in humans. Nat. Genet., 40, 225–231
47. Heinzen, E. L., Ge, D., Cronin, K. D., Maia, J. M., Shianna, K. V., Gabriel, W. N., Welsh-Bohmer, K. A., Hulette, C. M., Denny, T. N. and Goldstein, D. B. (2008) Tissue-specific genetic control of splicing: implications for the study of complex traits. PLoS Biol., 6, e1
48. Coulombe-Huntington, J., Lam, K. C., Dias, C. and Majewski, J. (2009) Fine-scale variation and genetic determinants of alternative splicing across individuals. PLoS Genet., 5, e1000766
49. Lee, Y., Gamazon, E. R., Rebman, E., Lee, Y., Lee, S., Dolan, M. E., Cox, N. J. and Lussier, Y. A. (2012) Variants affecting exon skipping contribute to complex traits. PLoS Genet., 8, e1002998
50. Ramasamy, A., Trabzuni, D., Gibbs, J. R., Dillman, A., Hernandez, D. G., Arepalli, S., Walker, R., Smith, C., Ilori, G. P., Shabalin, A. A., et al., (2013) Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies. Nucleic Acids Res., 41, e88
51. Mozhui, K., Wang, X., Chen, J., Mulligan, M. K., Li, Z., Ingles, J., Chen, X., Lu, L. and Williams, R. W. (2011) Genetic regulation of Nrnx1 expression: an integrative cross-species analysis of schizophrenia candidate genes. Transl. Psychiatr., 1, e25
52. 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
53. Sun, W. and Hu, Y. (2013) eQTL mapping using RNA-seq data. Stat. Biosci., 5, 198–219
54. Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al., (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501, 506–511
55. Wang, W., Qin, Z., Feng, Z., Wang, X. and Zhang, X. (2013) Identifying differentially spliced genes from two groups of RNA-seq samples. Gene, 518, 164–170
56. The International HapMap Consortium. (2003) The international HapMap project. Nature, 426, 789–796
57. Guan, Y. and Stephens, M. (2008) Practical issues in imputation-based association mapping. PLoS Genet., 4, e1000279
58. Scheet, P. and Stephens, M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet., 78, 629–644
59. Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res., 19, 1586–1592
60. Boeva, V., Zinovyev, A., Bleakley, K., Vert, J.-P., Janoueix-Lerosey, I., Delattre, O. and Barillot, E. (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics, 27, 268–269
61. Zhao, K., Lu, Z. X., Park, J. W., Zhou, Q. and Xing, Y. (2013) GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. Genome Biol., 14, R74
62. Monlong, J., Calvo, M., Ferreira, P. G. and Guigó, R. (2014) Identification of genetic variants associated with alternative splicing using sQTLseekeR. Nat. Commun., 5, 4698

