The Author(s) *BMC Genomics* 2017, **18**(Suppl 1):963
DOI 10.1186/s12864-016-3258-1

BMC Genomics

RESEARCH Open Access

ulfasQTL: an ultra-fast method of composite splicing QTL analysis

Qian Yang[1†], Yue Hu[1†], Jun Li[2] and Xuegong Zhang[1,3,4*]

From The 27th International Conference on Genome Informatics
Shanghai, China. 3-5 October 2016

Abstract

Background: Alternative splicing plays important roles in many regulatory processes and diseases in human. Many genetic variants contribute to phenotypic differences in gene expression and splicing that determine variations in human traits. Detecting genetic variants that affect splicing phenotypes is essential for understanding the functional impact of genetic variations on alternative splicing. For many situations, the key phenotype is the relative splicing ratios of alternative isoforms rather than the expression values of individual isoforms. Splicing quantitative trait loci (sQTL) analysis methods have been proposed for detecting associations of genetic variants with the vectors of isoform splicing ratios of genes. We call this task as composite sQTL analysis. Existing methods are computationally intensive and cannot scale up for whole genome analysis.

Results: We developed an ultra-fast method named ulfasQTL for this task based on a previous method sQTLseekeR. It transforms tests of splicing ratios of multiple genes to a matrix form for efficient computation, and therefore can be applied for sQTL analysis at whole-genome scales at the speed thousands times faster than the existing method. We tested ulfasQTL on the data from the GEUVADIS project and compared it with an existing method.

Conclusions: ulfasQTL is a very efficient tool for composite splicing QTL analysis and can be applied on whole-genome analysis with acceptable time.

Keywords: Alternative splicing, Genetic variants, sQTL, Genome-wide, Ultra-fast method

Background

The human genome contains about 3 billion base pairs, and there are only about 0.1% differences between two individuals' genome [1]. These genetic variants largely contribute to human multiple phenotypes [2]. Genome-wide association studies (GWAS) have identified many genetic loci that are associated with diseases. Understanding how these variants exert their effects still remains to be a big challenge [3]. It has been observed that many of the effects are through variations in the expression of genes and pathways, especially RNA splicing [4].

Alternative splicing is an important mechanism in the regulation of gene expression. High-throughput RNA-sequencing (RNA-seq) data have shown that most human genes undergo alternative splicing [5, 6], and it has been reported that many alternative splicing events are associated with many complex diseases [7–10]. Expression quantitative trait loci (eQTL) analysis is an effective approach for studying the association between genetic variants and gene expression [11–16]. This strategy has been extended to the analysis of association of alternative splicing genes with genetic variants [15, 17–29]. This is called splicing quantitative trait loci (sQTL) analysis, including exon-level sQTL and isoform-level sQTL. For exon-level sQTL study, researchers take exon expression, exon inclusion level or junction expression as the quantitative phenotype to perform sQTL analysis against genetic variants [15, 17, 20, 23–26, 28, 29]. Exons in one gene are not independent and they compose multiple isoforms through

* Correspondence: zhangxg@tsinghua.edu.cn
†Equal contributors
[1]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST, Department of Automation, Tsinghua University, Beijing 100084, China
[3]School of Life Sciences, Tsinghua University, Beijing 100084, China
Full list of author information is available at the end of the article

alternative splicing. In some cases, changes in the splicing pattern of a gene cannot be observed by changes in inclusion levels of individual exons [27]. The expression of each individual isoform can also be used as the quantitative phenotype for sQTL study [18, 20, 21]. Coulombe-Huntington et al. [19], Lappalainen et al. [23] and Battle et al. [22] used the isoform ratio as the quantitative trait for sQTL analysis, which controls the effects of overall gene expression and tests the relative abundances of isoforms. But they took each isoform ratio as a phenotype and did not consider the correlations between isoforms of the same gene. In many situations, besides the expression of each isoform, compositions and relative proportions of alternative isoforms of the same gene play important roles. Monlong et al. [27] proposed to use the splicing ratios of all isoforms of the same gene as a composite phenotype to take into consideration such correlations. In this way, the studied phenotype is not only the relative abundance of each isoform but also the correlated structure of the alternative splicing gene. We call this as composite splicing QTL analysis. They developed an R package sQTLseekeR to implement this strategy, which describes alternative splicing events by a vector of splicing ratios [27]. They compared their method with other univariate sQTL methods that are based on exons or isoforms, and showed that sQTLseekeR is more capable of detecting associations that cannot be found by univariate exon-based method [27].

sQTLseekeR is based on tests on every gene-variant pair. Considering the tens of thousands genes and millions of genetic variants on the whole genome, the computational speed of sQTLseekeR prohibits it to be applied for analyzing all the genes and variants at the whole-genome scale. In their original work, they only applied it for analyzing variants located within 5 kb of each gene [27]. This largely limits the scope of questions that can be addressed with the method. Alternative splicing is regulated by both cis-elements and trans-factors [30]. More computationally efficient methods are in critical need for building the full picture of both cis- and trans-regulations of alternative splicing.

In this paper, we developed a method named ulfasQTL for ultra-fast composite sQTL analysis. It transforms vectors of splicing ratios to a spherical coordinate system and uses a matrix-based computation to test multiple genes and variants at the same time. This can dramatically boost the computational speed. We applied the proposed method and compared it with sQTLseekeR on data from the GEUVADIS project [23] to evaluate its performance and test its feasibility for genome-scale computation. Results show that ulfasQTL is several orders faster and can be readily used for genome-wide studies for associations between the alternative splicing structures of all genes and all variants in the genome.

Methods

Definition of splicing-QTL

Suppose a gene has n isoforms, and their expressions are $x_1, x_2, \ldots, x_n$. The splicing ratios of isoforms are their proportions in the total expression of the gene:

$$p_i = x_i / \sum_1^n x_i, i = 1, 2, \ldots, n.$$

Let a variant's genotype be g, and $g = 0$, 1 or 2. Our goal is to detect associations between the genotype of a variant (the value of g) and the splicing pattern of a gene. For splicing pattern of a gene, we focus on the splicing ratios p_i, $i = 1, 2, \ldots, n$ of the isoforms but not the gene's total expression $\sum_1^n x_i$.

The splicing pattern of a gene is described by the vector $(p_1, p_2, \ldots, p_n)$. Figure 1 shows a simple example of a gene with 3 isoforms. The patterns of the splicing ratios are very different among samples of different genotypes of the variant, which indicates that the variant is a splicing-QTL or sQTL of the gene.

The sQTLseekeR method

The sQTLseekeR method by Monlong et al. [27] uses a distance-based approach to detect composite sQTLs for each gene-variant pair. For one gene, each sample's phenotype is the vector of the splicing ratios of all its isoforms. So each sample can be treated as a point in this vector space. All samples of a dataset are divided into the three or two groups according to their genotypes at a variant locus. sQTLseekeR calculates the variability of splicing ratios of a gene between and within

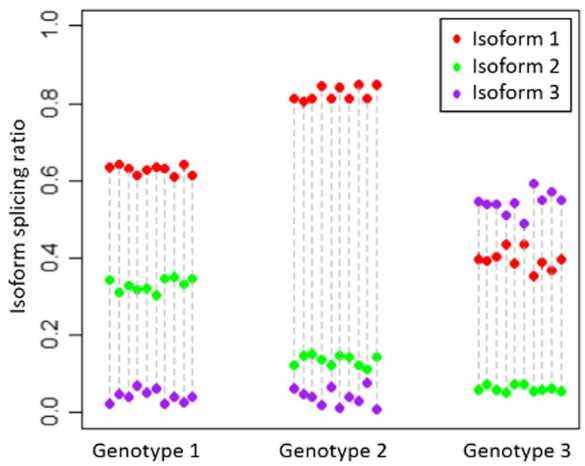

Fig. 1 An example case of splicing-QTL. The gene has 3 isoforms. The splicing ratios of the three isoforms of the same sample are shown as points of different colors linked by a *dashed line*. Samples with the same genotype are shown together. We can see that the distribution patterns of splicing ratios are different between different genotypes, which indicates that this variant is associated with the alternative splicing pattern of this gene, and therefore it is a sQTL of the gene

groups using the Hellinger distance. For a gene containing n isoforms, the Hellinger distance between sample a and b is defined as

$$d_H(a,b) = \sqrt{\sum_{i=1}^{n}\left(\sqrt{p_{ia}}-\sqrt{p_{ib}}\right)^2},$$

where p_{ia} is the splicing ratio of isoform i in sample a, and p_{ib} is the splicing ratio of isoform i in sample b. The variability is defined as the sum of squared distances (SS) between the samples and their centroid,

$$SS = \sum_{j=1}^{N} d_H^2(j,\boldsymbol{c}),$$

where $\boldsymbol{c}$ is the centroid, and N is the number of samples in this group. The within-group variability SS_w is defined as

$$SS_W = SS = \sum_{j=1}^{N} d_H^2(j,\boldsymbol{c}).$$

The between-group variability SS_B can be obtained by

$$SS_B = SS_T - SS_W,$$

where SS_T is the total variability $SS_T = \sum_{i=1}^{L} d_H^2(c_i,\boldsymbol{c})$, L is the number of variant's genotype groups, c_i is the centroid of each genotype group, and $\boldsymbol{c}$ is the overall centroid of all samples.

The Anderson test [31] is used to compute a pseudo F-ratio score to measure the relative differences between within-group and between-group distances,

$$F = [SS_B/(L-1)]/\left[SS_W/\left(\sum_{i=1}^{L} N_i - L\right)\right],$$

where L is the number of groups and N_i is the number of samples in group i. They used a direct method to calculate the pseudo F-ratio score by considering matrix of distances between every pair of samples instead of using centroids in the definition [31]. The null distribution of the F-score is approximated via simulation to get the FDR (false discovery rate) of the tests.

Different genes contain different numbers of isoforms so their splicing ratio vectors are of different dimensions. Also different genetic variants divide samples with different grouping. Therefore, sQTLseekerR needs to test each gene against each variant individually. It is very time-consuming and infeasible for analyses at whole-genome scales.

The ulfasQTL method

The goal of our ulfasQTL method is to detect composite sQTLs for all gene-variant pairs on the whole genome efficiently. The core strategy is to compute the statistics of associations for a large number of gene-variant pairs concurrently within a single run of the test. A matrix-based test for multiple independent phenotype-variant pairs is adopted to achieve the high computational efficiency, and we introduced a coordination transform on the splicing ratio vector to make the tests in the matrix independent. The test results on the ratios belonging to the same gene are then combined to produce the final statistics on the gene. We describe the details below.

Suppose there are n isoforms in a gene, and their splicing ratios are $p_1, p_2, \ldots p_n$, respectively. There is the constraint that $\sum_{i=1}^{n} p_i = 1$, and so the degrees of freedom of the vector $(p_1, p_2, \ldots p_n)$ is $n - 1$. Thus, we cannot directly perform association analysis for all isoforms in a gene by adding the statistics up as the test for the gene because of their dependence. We need to transform the n splicing ratios to a set of n-1 independent variables. We propose to do this transformation using the idea of "n-sphere". Firstly, let

$$q_i = \sqrt{p_i},\ \ i = 1, 2, \ldots, n,$$

then one sample can be represented by the vector $(q_1, q_2 \ldots, q_n) = \left(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_n}\right)$ in a n-dimensional Cartesian coordinate system. We convert it to coordinates in a spherical coordinate system $(\rho, \phi_1, \phi_2, \ldots, \phi_{n-1})$, where ρ is the length of the vector, defined as $\rho = \sqrt{q_1^2 + q_2^2 + \ldots + q_n^2}$, and $\phi_1, \phi_2, \ldots, \phi_{n-1}$ are the angles between the vector and n-1 of the Cartesian axes, defined as

$$\begin{aligned}
\phi_1 &= arccos\frac{q_1}{\sqrt{q_1^2 + q_2^2 + \ldots + q_n^2}},\\
\phi_2 &= arccos\frac{q_2}{\sqrt{q_2^2 + q_3^2 + \ldots + q_n^2}},\\
&\ldots,\\
\phi_{n-1} &= \frac{arccos q_{n-1}}{\sqrt{q_{n-1}^2 + q_n^2}}.
\end{aligned}$$

In this way, the original n splicing ratios are converted to n-1 independent variables $\phi_1, \phi_2, \ldots, \phi_{n-1}$. We call them *converted splicing components* for convenience. The order of the $q_1, q_2, \ldots, q_n$ in the above transformation can be arbitrary. We order them from the largest to the smallest mean values across the samples in our implementation.

Figure 2 illustrates how the spherical coordinate system works with an example gene. In the example, the gene contains three isoforms. The original constraint $\sum_{i=1}^{3} p_i = 1$ on the splicing ratios becomes

$$q_1^2 + q_2^2 + q_3^2 = 1$$

on the q_i's. In the spherical coordinate system, we always have $\rho = 1$ regardless of the values of p_1, p_2, and p_3. The two angles in the spherical coordinate system, ϕ_1 and ϕ_2, on the other hand, are independent from each other.

In [32], Shabalin proposed a matrix-based method Matrix eQTL for fast eQTL computation. It can test all gene-variant pairs together by choosing appropriate test

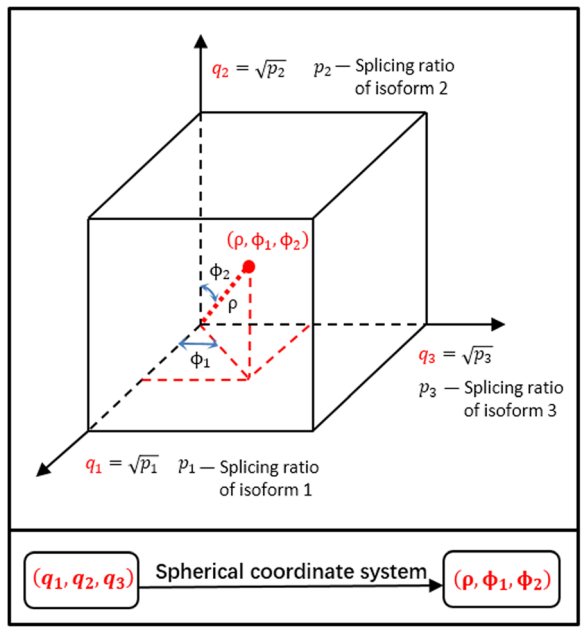

Fig. 2 An illustrative example of spherical coordinate system application. The example gene contains 3 isoforms, and their splicing ratios in a sample are p_1, p_2, p_3 with the constraint $\sum_{i=1}^{3} p_i = 1$. We let $q_i = \sqrt{p_i}$, $i = 1, 2, 3$ and present this sample as a point (the *red point* in the plot) in the 3-dimensional rectangular coordinate system. We then apply spherical coordinate system in which this sample can also be regarded as a point (ρ, ϕ_1, ϕ_2) in a spherical coordinate system. Because of the constraint that $\sum_{i=1}^{3} p_i = 1$, the spherical coordinates of this sample is $(1, \phi_1, \phi_2)$, in which ϕ_1 and ϕ_2 are independent from each other

statistics and applying matrix operations to calculate their test statistics values in parallel. It implements both linear regression model and ANOVA model for eQTL analysis. Matrix eQTL can detect associations between two variables, but our goal is to detect associations between vectors and variables. After spherical coordinate transformation, we converted a vector into a set of mutually independent variables, and then adopted this matrix-based strategy in ulfasQTL to implement massive tests on the converted splicing components ϕ_i's in a matrix. Suppose we want to do tests on m genes and k variants of l samples in a single run, the expression of these genes can be represented by a matrix G_{m*l} and the genotypes of the variants can be represented by a matrix V_{k*l}. Now we do the tests on the converted splicing components instead of the expression values. So we build the matrix Φ of all converted splicing components of the m genes. The dimension of this matrix is $t*l$, where t is the total number of independent splicing components of the m genes, which equals to the total number of isoforms of these genes minus the number of genes. The columns (samples) of the matrix Φ_{t*l} and matrix V_{k*l} are matched with each other.

Here is the detailed method of Matrix eQTL for linear regression model, and the method of Matrix eQTL for ANOVA model is similar to linear regression model [32]. We assumed that the association between splicing component ϕ and variant v is linear.

$$\phi = \alpha + \beta v + \epsilon, \quad \text{where } \epsilon \sim \text{i.i.d. } N(0, \sigma^2)$$

For linear regression model, Matrix eQTL chose the absolute value of sample correlation $|r| = |\text{cor}(\phi, v)|$ as the test statistic which can has equal power and can be computed faster than other test statistics. Then Matrix eQTL performed standardization preprocessing procedures which do not change the correlation.

$$\sum \phi_i = 0, \sum \phi_i^2 = 1, \sum v_i = 0, \sum v_i^2 = 1$$

So Matrix eQTL computed the test statistics by the inner product $\langle g, v \rangle$ between vectors ϕ and v as follows.

$$r_{gv} = \text{cor}(g, v) = \frac{\sum (g_i - \bar{g})(v_i - \bar{v})}{\sqrt{\sum (g_i - \bar{g})^2 (v_i - \bar{v})^2}} = \sum g_i v_i = \langle g, v \rangle$$

Matrix eQTL can greatly simplify the computation of test statistics by the multiplication of the two preprocessed matrices $\Phi_{t*l} \quad V_{k*l}^T$ [32]. In this way, the computational load can be reduced dramatically. The key assumption for this fast computation is the rows (components) in the matrix are independent with each other, which is guaranteed by the spherical coordinate transformation.

Matrix eQTL can conduct either linear regression or ANOVA based on the obtained correlations, and report the t-test statistics or F-test statistics of all associations between each converted splicing component and each variant. After getting all test statistics for each component-variant pair, we combine results from all components of the same gene-variant pair to get the test statistic for the gene-variant pair. We convert t-test statistics or F-test statistics of the component-variant pairs to z values that follow the standard normal distribution. Finally, we get the test statistic s for each gene-variant pair by

$$s = \sum_{i=1}^{n-1} z_i^2.$$

It follows a chi-square distribution with degrees of freedom n-1, i.e., $s \sim \chi_{n-1}^2$, and the p-value for each gene-variant pair can be obtained accordingly. We can then convert the p-values to false discovery rates (FDRs) using the q-value method [33]. We developed a software package ulfasQTL to implement the above method, which calls for the MatrixEQTL package [32] in the matrix calculation. Figure 3 shows the basic flowchart of the whole method (the left panel) and a detailed illustrative example (the right panel).

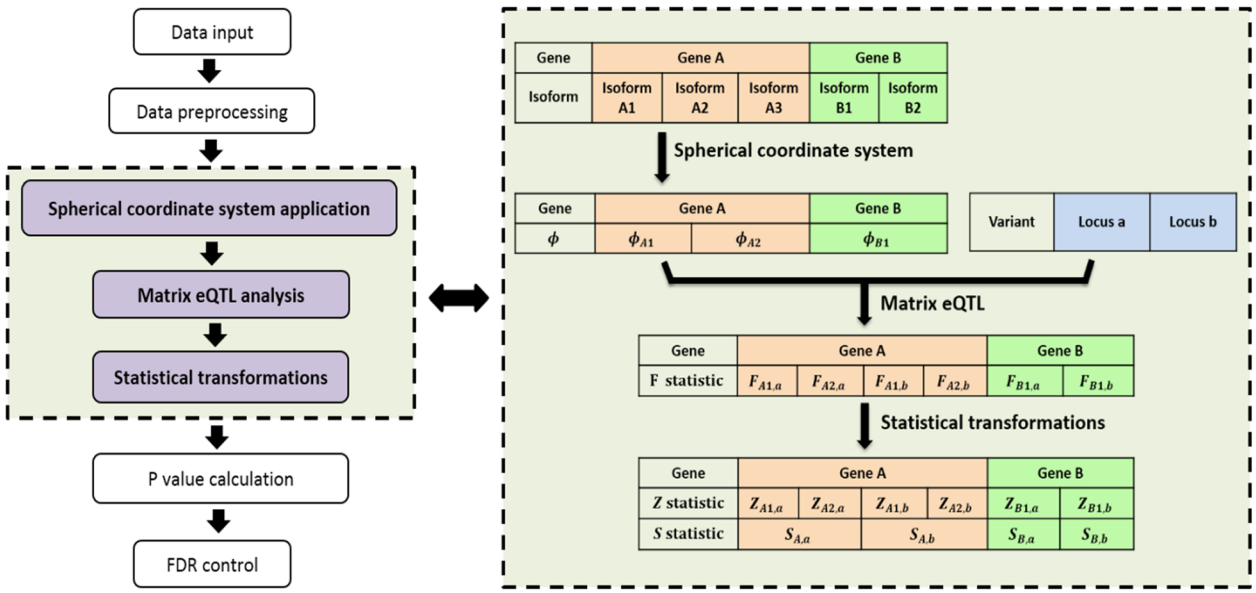

Fig. 3 The diagram of ulfasQTL. The left panel is basic flowchart and the right panel shows an illustrative example of core steps. There are in seven steps in the flowchart and the 3 steps shown in purple are core steps of ulfasQTL. We have two matrices after data preprocessing. One for the genotype data (genotypes of each sample at loci a and b) and one for the isoform expression data (isoforms A1, A2, A3 of gene A and isoforms B1, B2 of gene B of each sample). Then isoform expression matrix is converted to splicing components matrix (φ_{A1}, φ_{A2}, φ_{B1} for each sample) by the spherical coordinate system. We then send genotype data and splicing components matrix into MatrixeQTL (ANOVA model) to get the *F* statistics ($F_{A1,a}$, $F_{A2,a}$, $F_{A1,b}$, $F_{A2,b}$, $F_{B1,a}$, $F_{B2,b}$) for each component-variant pair. Finally, we apply statistical transformations and get corresponding *Z* statistics according to *F* statistics, and merge *Z* statistics into *S* statistics ($S_{A,a}$, $S_{A,b}$, $S_{B,a}$, $S_{B,b}$) for each gene-variant pair

When applying the method on very large datasets like genome-wide analyses, the dataset can be too large to be fit into computer memory. In such cases, we split dataset into smaller subsets and calculate them in multiple runs. For example, in the experiments reported below, we take all genes together but split the variants into smaller files, each containing 1000 variants.

The ulfasQTL package was developed using R and C++. It includes C++ codes for data preprocessing and the spherical coordinate transformation, and R codes for Matrix eQTL analysis and the calculation of *p*-values and FDRs. The package can be downloaded at http://bioinfo.au.tsinghua.edu.cn/software/ulfasQTL/.

The computational complexity

The computing time of ulfasQTL is consumed mostly by two major steps. One is the computation with MatrixEQTL for calculating the correlation matrix of components and variants. The dimension of component matrix is $t*l$, the dimension of variant matrix is $k*l$. So the time complexity of this step is $O(k*t*l)$, where k is the total number of variants, t is the total number of converted splicing components and l is the total number of samples. The second major step is that after getting the MatrixEQTL output, we need to sort the component-variant pairs by both splicing components and variants to make sure pairs from the same gene and the same variant stay together. We need to sort the pairs twice for that purpose. We use the mergesort method as it is one of the fastest stable sorting method. The time complexity of mergesort is $O(t*k*\log(t*k))$. Therefore, the total time complexity of ulfasQTL is $O((l + \log(t*k))*t*k)$.

For large datasets that need to be split into multiple smaller datasets, the computation on the multiple datasets can be assigned to multiple kernels or computers, which provides an easy and efficient way of doing large-scale sQTL analysis in parallel.

For each gene-variant pair, sQTLseekeR calculates the within-group variability and the between-group variability to get the Anderson test statistic for the pair. The complexity for this step is $O(l^2)$. The test method used by sQTLseekeR is sensitive to the homogeneity of the variabilities or dispersions of the compared groups. The test power may decrease when dispersions of the groups are very different. So sQTLseekeR needs an extra step to filter such variants to avoid potential false sQTLs. The method for this filtering is similar to ANOVA, but the distance measurement is different from Euclidan distance. They applied principle component analysis (PCA) to the data and calculated the Euclidan distances between group members and the group centroid on the principal components. The time complexity of computing the eigenvalue in PCA of a $l*l$ dimensional matrix is $O(l^3)$, and computing the within-group variability is

$O(l^2)$. For all phenotype-variant pairs of m genes and k variants, the total time complexity of the above steps is $O(l^3*m*k)$. After getting the F score of a candidate pair and this filtering step, sQTLseekeR performs an approximation of permutations for each gene to calculate the significance of the F score. The computational complexity of this step is $O(l^3*\mathrm{m})$. So the overall complexity of sQTLseekeR is at the level of $O(l^3*m*k)$.

Results

Data

We applied ulfasQTL on the data of lymphoblastoid cell lines of 462 individuals published in [23] to study its performance. The transcripts expression data are from the GEUVADIS project [23] and the genotype data are from 1000 Genomes Project Phase I dataset 1 [1]. The dataset includes individuals from European population (CEU, FIN, GBR, TSI) or African population (YRI). For isoform expression data, at first we added a small number to the expression data to avoid the occurrence of 0's in the denominator. Next we computed the splicing ratios of each isoform of all genes, and only considered active isoforms with splicing ratios larger than a given threshold. Genes with less than two active isoforms after this step were filtered out. Then we calculated the splicing variability for each gene and removed genes whose splicing variability are less than 0.01. For each gene, we used samples whose gene expression is over 0.01 RPKM. For genotype data, we kept variants that have at least 2 genotype groups in the samples and each group has at least 5 samples. Groups with less than 5 samples are set to *NA* to make sure that they are not taken into consideration in the test. We picked up samples which have both expression data and genotype data, and made the samples' order identical in two data files.

We conducted 3 experiments, Experiments I, II and III. Experiments I and II were on small-scale datasets to study the performance of ulfasQTL and to compare it with sQTLseekeR. Experiments III was on a genome-scale dataset to test the feasibility of ulfasQTL on big data. The experiments were done on a desktop computer with CPU of Intel Core i7-4790 k(4GHz) and 16GB DDR3 RAM, running 64 bit Ubuntu and 64 bit R 3.2.3.

The computational efficiency

Experiments I and II were on a small dataset on which both ulfasQTL and sQTLseekeR can work. In Experiment I, we randomly picked 1000 variants and 407 genes containing a total of 1000 isoforms in Chr.1. We performed sQTL analysis using both methods to compare the computational efficiency and results of the two methods. It took 13,680 s (3.8 h) for sQTLseekeR to complete the computation, while the ulfasQTL only used 2.3 s to complete the computation. ulfasQTL works about 6000 times faster than sQTLseekeR.

In Experiment II, we choose 400 genes and 180,446 variants which are all located at 1–13,000,000 in Chr.1. After preprocessing on the expression data and variant data, 160 genes and 76,779 variants were kept, which gave 12,284,640 candidate gene-variant pairs. ulfasQTL accomplished all the computation in 901 s (15.0 min). We applied sQTLseekeR on these 160 genes with only the variants that are located within 5 kb of each gene as in the original work. This gave a total of 8560 candidate gene-variant pairs. sQTLseekR used 4492 s (~1.25 h) to complete these computations.

Experiment III was on all genes and genetic variants on Chr.1 to test the feasibility of ulfasQTL for genome-scale analyses. There are in total 5172 genes and 1,900,188 variants after screening. The total number of gene-variant pairs which need to be tested are $9.8\mathrm{x}10^9$. On the same desktop computer as in the first experiment, ulfasQTL can give the result of a split subset of 5172 genes and 1000 variants in about 45 s. The analysis on the whole task took 87,112 s (~24.20 h).

Applying sQTLseekeR on the data of Experiment III is impractical due to the heavy computing cost. In the original sQTLseekeR paper [27], the authors reported that they ran sQTLseekeR separately in each sub-population on this dataset, and each sub-population contains about 10,012 genes and 140 variants per gene on average. The analysis of ~1,400,000 gene-variant pairs took about 4 h using 16 cores (2Gb 2.70GHz nodes). Based on these reports, we can estimate that it would take about 1169 days or 3.2 years on a similar cluster if sQTLseekeR were to be used to analyze the data in Experiment III.

Comparison of *p*-values

We compared the results of ulfasQTL and sQTLseekeR in Experiments I and II to obtain better understanding on the similarities and differences between the tests used by the two methods. In Experiment I, after data preprocessing there were 359 candidate variants and 140 candidate genes that were analyzed by both ulfasQTL and sQTlseekeR. They composed 50,260 candidate associations to be tested by ulfasQTL and sQTLseekeR. sQTLseekeR adopted some further filtering on the genes and only tested 47,069 of the candidate associations. We used these 47,069 candidate associations to study the relationship of p-values reported by the two methods. Figure 4 shows the scatter plots of the p-values of the two methods on the same points (candidate associations) in the range of $p < 0.1$, $p < 0.05$ and $p < 0.01$. The Spearman correlations of the two p-values are 0.58, 0.56 and 0.73, respectively, for candidate associations with p-values less than 0.1, 0.05 and 0.01. We can observe that ulfasQTL tends to be more conservative

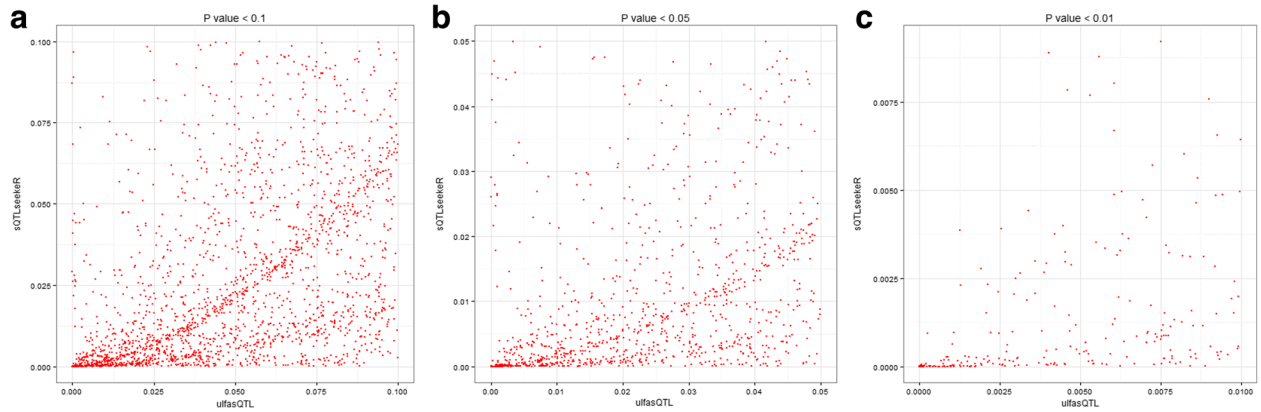

Fig. 4 *P*-values of the two methods on the same points (candidate associations) in the range of $p < 0.1$, $p < 0.05$ and $p < 0.01$ in Experiment I. The *horizontal axes* are *p*-values of associations detected by ulfasQTL and the *vertical axes* are *p*-values of associations detected by sQTLseekeR

and tends to produce slightly larger *p*-values for most of the data. Note that the data in this experiment were a randomly selected subset of genes and SNPs on Chr.1. We can expect that most of the candidates would not be significant. The more conservative *p*-value obtained with ulfasQTL presents an advantage over existing method not only on the higher computational efficiency, but also on the possible lower false discoveries.

In Experiment II, 160 candidate genes and 76,779 candidate variants were analyzed by ulfasQTL, and 160 candidate genes and variants located within 5 kb from them were analyzed by sQTLseekeR. After preprocessing, we got a total of 8560 candidate associations that have *p*-values reported by both methods. Figure 5 shows the scatter plots of the *p*-values of the two methods on the same points in the range of $p < 0.1$, $p < 0.05$ and $p < 0.01$. The Spearman correlations of the two *p*-values are 0.60, 0.69 and 0.67, respectively, for candidate associations with *p*-values less than 0.1, 0.05 and 0.01. We can see that the general trends of relations of the *p*-values are the same in Experiments I and II, while the correlation between the results of the two methods is higher in Experiment II. Experiment I was on randomly selected genes and variants so it can be expected that most of the gene-variants pairs are not significantly associated. On the other hand, candidate variants compared in Fig. 5 in Experiment II were all within 5 kb of the candidate genes, which are more likely to have significant sQTLs. The higher correlation between *p*-values of the two methods implies that the two methods agrees better with each other on true association signals.

Discussion

There are several directions that need further investigation. We used ANOVA to test the hypothesis in the method based on two underlying assumptions. The first one is the distribution of data should be normal distribution or close to normal distribution. We can see that the distribution of converted splicing components may not always meet the assumption. The other one is ANOVA assumes homogeneity

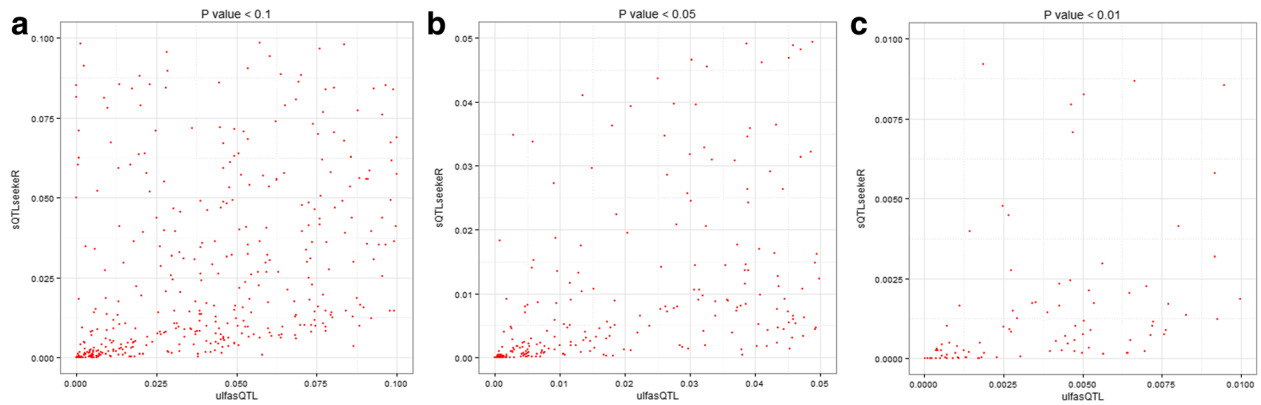

Fig. 5 *P*-values of the two methods on the same points (candidate associations) in the range of $p < 0.1$, $p < 0.05$ and $p < 0.01$ in Experiment II. The *horizontal axes* are *p*-values of associations detected by ulfasQTL and the *vertical axes* are *p*-values of associations detected by sQTLseekeR

among groups, which may be violated when the sample size of one group is small. Such situations can cause false positives. The preprocessing to add a small value to the denominator also may cause false results for some special cases when all isoforms are not expressed in some samples. Therefore, after applying ulfasQTL on genome-wide candidates, users may use slower single-gene based methods only on the reported results to further validate the significance if necessary, or to check homogeneity (such as using Bartlett's test) of different genotype groups.

Composite splicing QTL involves the collaborative regulation of multiple isoforms. Comparing to the traditional univariate isoform- or exon-based splicing QTL analysis, golden-standard validation data is less available. Monlong et al. [27] illustrated a few examples of composite splicing QTLs, but due to the small scale of their work, the examples cannot be taken as standard. Actually, when applied on a larger range of candidate variations with sQTLseekeR on fewer genes, we observed that some examples became no longer significant after multiple test correction. This may be due to the nature that splicing composite variation is associated by the multiple genetic factors. The ability to conduct genome-wide study of composite sQTL by ulfasQTL can help to better investigate both cis- and trans- factors that can be associated with splicing composite variation, and it will be of great interest if methods can be developed for finding associations of composite splicing phenotypes with multiple genomic variation loci.

Conclusions

We developed a new method ulfasQTL for ultra-fast splicing QTLs analysis of splicing patterns that are associated with genetic variants. This is the first time that coordination conversion is used for decomposing composite splicing pattern to a set of independent components. This conversion allows for the simultaneous computation on many genes in a matrix form. Experiments on small- and large-scale data show that it is several thousand times faster than the existing method for splicing QTL, and is efficient for splicing QTL analysis at the whole-genome scale.

Declaration

This article has been published as part of BMC Genomics Volume 18 Supplement 1, 2016: Proceedings of the 27th International Conference on Genome Informatics: genomics. The full contents of the supplement are available online at http://bmcgenomics.biomedcentral.com/articles/supplements/volume-18-supplement-1.

Funding

This work is partially supported by the National Basic Research Program of China (2012CB316504) and NSFC Grant 91010016 to XZ, and the National Institute of General Medical Sciences (R01GM097230) to JL. Publication was funded by the National Basic Research Program of China (2012CB316504).

Availability of data and materials

Testing data and R implementations are available for free at http://bioinfo.au.tsinghua.edu.cn/software/ulfasQTL/.

Authors' contributions

QY and XZ initiated the project. QY and JL developed the method. YH wrote the codes and analyzed the algorithmatic complexity. QY and YH implemented the experiments. QY and XZ wrote the manuscript with inputs from all authors. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST, Department of Automation, Tsinghua University, Beijing 100084, China. [2]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA. [3]School of Life Sciences, Tsinghua University, Beijing 100084, China. [4]Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China.

Published: 25 January 2017

References

1. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.
2. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16:197–212.
3. Manolio TA. Genomewide Association Studies and Assessment of the Risk of Disease. N Engl J Med. 2010;363:166–76.
4. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y, Pritchard JK. RNA splicing is a primary link between genetic variation and disease. Science. 2016;352:600–4.
5. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40:1413–5.
6. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. Deciphering the splicing code. Nature. 2010;465:53–9.
7. Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy. Nat Biotech. 2004;22:535–46.
8. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. Function of alternative splicing. Gene. 2005;344:1–20.
9. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S. Function of alternative splicing. Gene. 2013;514:1–30.
10. Kornblihtt AR, Schor IE, Allo M, Dujardin G, Petrillo E, Munoz MJ. Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat Rev Mol Cell Biol. 2013;14:153–65.
11. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003;422:297–302.
12. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004;430:743–7.
13. Rockman MV, Kruglyak L. Genetics of global gene expression. Nat Rev Genet. 2006;7:862–72.
14. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 2008;24:408–15.
15. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464:768–72.
16. Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet. 2011;27:72–9.
17. Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna KV, Gabriel WN, Welsh-Bohmer KA, Hulette CM, Denny TN, Goldstein DB. Tissue-specific genetic control of splicing: implications for the study of complex traits. PLoS Biol. 2008;6:e1000001.

18. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. Genome-wide analysis of transcript isoform variation in humans. Nat Genet. 2008;40:225–31.
19. Coulombe-Huntington J, Lam KCL, Dias C, Majewski J. Fine-scale variation and genetic determinants of alternative splicing across individuals. PLoS Genet. 2009;5:e1000766.
20. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010;464:773–7.
21. Lalonde E, Ha KCH, Wang Z, Bemmo A, Kleinman CL, Kwan T, Pastinen T, Majewski J. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. Genome Res. 2011;21:545–54.
22. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2013;24:14–24.
23. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PAC, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501:506–11.
24. Zhao K, Lu Z-x, Park JW, Zhou Q, Xing Y. GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. Genome Biol. 2013;14:R74-R74.
25. Guan L, Yang Q, Gu M, Chen L, Zhang X. Exon expression QTL (eeQTL) analysis highlights distant genomic variations associated with splicing regulation. Quantitative Biology. 2014;2:71–9.
26. Hassan MA, Butty V, Jensen KDC, Saeij JPJ. The genetic basis for individual differences in mRNA splicing and APOBEC1 editing activity in murine macrophages. Genome Res. 2014;24:377–89.
27. Monlong J, Calvo M, Ferreira PG, Guigó R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. Nat Commun. 2014;5:4698.
28. Ongen H, Dermitzakis ET. Alternative Splicing QTLs in European and African Populations. Am J Hum Genet. 2015;97:567–75.
29. Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, Johnson AD, Levy D, O'Donnell CJ. Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat Genet. 2015;47:345–52.
30. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA. 2008;14:802–13.
31. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001;26:32–46.
32. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28:1353–8.
33. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat. 2003;31:2013–35.