G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2016;14(4):211-215
https://doi.org/10.5808/GI.2016.14.4.211

ORIGINAL ARTICLE

Investigation of Splicing Quantitative Trait Loci in *Arabidopsis thaliana*

Wonseok Yoo, Sungkyu Kyung, Seonggyun Han, Sangsoo Kim*

Department of Bioinformatics and Life Science, Soongsil University, Seoul 06978, Korea

The alteration of alternative splicing patterns has an effect on the quantification of functional proteins, leading to phenotype variation. The splicing quantitative trait locus (sQTL) is one of the main genetic elements affecting splicing patterns. Here, we report the results of genome-wide sQTLs across 141 strains of *Arabidopsis thaliana* with publicly available next generation sequencing datasets. As a result, we found 1,694 candidate sQTLs in *Arabidopsis thaliana* at a false discovery rate of 0.01. Furthermore, among the candidate sQTLs, we found 25 sQTLs that overlapped with the list of previously examined trait-associated single nucleotide polymorphisms (SNPs). In summary, this sQTL analysis provides new insight into genetic elements affecting alternative splicing patterns in *Arabidopsis thaliana* and the mechanism of previously reported trait-associated SNPs.

Keywords: QTL mapping, RNA-seq, sQTLs

Introduction

Arabidopsis thaliana is a metaphyte and angiosperm growing in Europe, Asia, and Africa. This is a small plant that is one of the widely popular model organisms, because this plant has a small genome and can be cultivated easily. Thus, much genetic research has been carried out with *Arabidopsis thaliana* [1, 2]. Recently, in *Arabidopsis thaliana*, genetic components associated with traits or diseases have been reported through genome-wide association studies (GWAS) studies [3]. The GWAS is a powerful approach to discover genetic elements associated with various traits. The GWAS was based on 107 phenotypes. This can be grouped into 4 categories: 23 flowering, 23 defense, 18 ionomics, and 43 development traits. Although GWASs have been successful, the molecular mechanisms of genetic elements discovered through the GWAS have been elusive [4]. The molecular mechanisms of previously reported trait-associated single nucleotide polymorphisms (SNPs) in *Arabidopsis thaliana* have been also elusive. Expression quantitative trait loci (eQTLs) influencing gene expression levels and splicing quantitative trait loci (sQTLs) changing alternative splicing patterns are useful genetic elements to explain the trait-associated SNPs. In *Arabidopsis thaliana*, there have been many studies to identify eQTLs, and these accumulated eQTLs have been greatly helpful for explaining trait-associated loci [5-7]. However, there have been few studies of the genome-wide investigation of sQTLs in *Arabidopsis thaliana*. Since sQTLs can change alternative splicing patterns and consequently affect various traits, those sQTLs can constitute crucial elements for the interpretation of trait-associated SNPs [8]. Here, we discover candidate genome-wide sQTLs in *Arabidopsis thaliana* using the IVAS package, which is a user-friendly R/Bioconductor package, to discover sQTLs with genotypes and fragments per kilobase million (FPKM) of a transcript dataset [9]. We analyzed 141 previously published matched genotype data and an RNA-seq dataset downloaded in the 1001 Genomes Data Center [10] and the Gene Expression Omnibus (GEO) database [11], respectively. To obtain the FPKM of transcripts, we processed and aligned raw RNA-seq data. As a result of the discovery using IVAS, we identified 1,694 SNPs. Furthermore, among those SNPs, we found 96 candidate sQTLs that overlapped with the list of previously published trait-associated loci [3], and of 96 sQTLs, 25 sQTLs that led to a large difference in expression ratio of alternatively targeted exons across genotypes of each sQTL were finally selected. In

Received August 1, 2016; Revised September 5, 2016; Accepted October 16, 2016

***Corresponding author:** Tel: +82-2-820-0457, Fax: +82-2-824-4383, E-mail: sskimb@ssu.ac.kr

summary, we reported 1,694 genome-wide sQTLs and 25 candidate trait-associated sQTLs in *Arabidopsis thaliana*. This list of those sQTLs may be a useful dataset for sQTL utilities in *Arabidopsis thaliana*, as well as its neighboring species.

Methods

The overall pipeline

An overview of our research is given in Fig. 1. We downloaded genotype and RNA-seq datasets measured in *Arabidopsis thaliana* from the 1001 Genomes Data Center [10] and the GEO database (GSE43858) [11]. The genotype and RNA-seq datasets include 171 and 160 strains, respectively. Among them, we selected a matched dataset of 141 strains. The raw RNA-seq data were filtered by quality control and mapped to a reference sequence using SHRiMP2 software, which is an alignment tool for high accuracy and sensitivity at a very reasonable speed [12]. An average mapping rate was about 80%. After mapping to the reference sequence, we filtered out mapped reads with low mapping quality using Samtools [13]. In order to obtain FPKM values, we estimated quantifications from mapped reads using Cufflinks [14] with the reference GTF file downloaded from the Ensembl Genomes website. With the preprocessed genotype and RNA-seq datasets, we carried out IVAS for the discovery of genome-wide sQTLs. Afterwards, we found candidate sQTLs that overlapped with trait-associated loci in the list of the 107 phenotypes database [3].

Association test for identification of genome-wide sQTLs in *Arabidopsis thaliana*

In order to discover genome-wide sQTLs in *Arabidopsis thaliana*, we carried out an association test using the IVAS package [9]. For each exon with at least two isoforms, IVAS tested associations between the alternatively targeted exon and the genotype of SNPs located in the exon and flanking introns. We filtered the result based on a cutoff at false discovery rate (FDR) = 0.01 and median values of exon ratio across each genotype > 0.1.

Interpretation of trait-associated SNPs with candidate sQTLs

We downloaded the list of loci associated with the 107 phenotypes of 4 categories [3] ($p < 0.05$). We identified and tabulated the trait-associated loci overlapping with the candidate sQTLs obtained by IVAS. In order to yield a more biologically meaningful result, we kept those trait-associated sQTLs achieving a high differential expression ratio of an alternatively targeted exon across genotypes of each sQTL (the third quantile of exon ratios of one genotype $>$ the first quantile of exon ratios of the other genotype).

Results

Processing for the genotype and RNA-seq datasets

We downloaded a genotype dataset in 1001 Genomes Data Center [10] and an RNA-seq dataset in the GEO database (GSE43858) [11] in order to identify genome-wide sQTLs in *Arabidopsis thaliana*. In order to increase the accuracy of the mapping of RNA-seq reads, we manually filtered out low-quality reads and aligned them to the *Arabidopsis thaliana* reference genome. The aligned reads were estimated by Cufflinks [14]. As a result, we obtained and analyzed the quantifications of 41,621 isoforms and 5,833,535 SNP markers of the 141 matched strains.

Discovery of sQTLs in *Arabidopsis thaliana*

Using the IVAS package [9], we analyzed the preprocessed dataset. For each alternatively spliced exon, we carried out an association test between a given exon expression and genotype of an SNP located in the given exon and flanking introns. As a result, we found 1,694 candidate sQTLs by a cutoff at an FDR = 0.01. Fig. 2A displays an exemplary boxplot of an sQTL, located in chr3:20,536,243, which

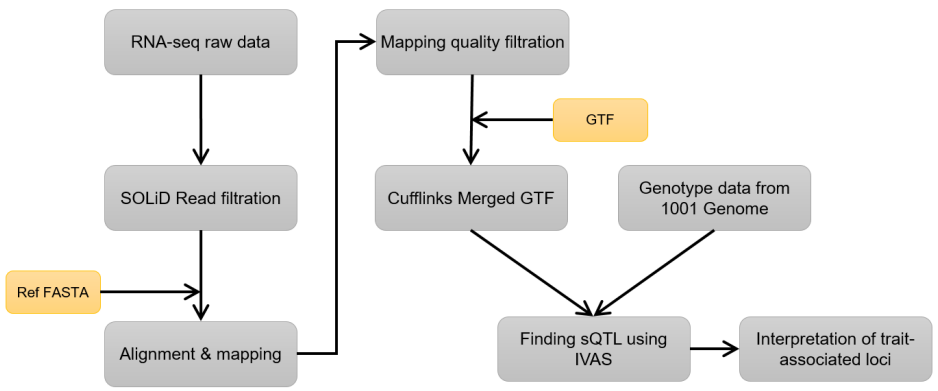

Fig. 1. The overview of our analysis. The RNA-seq raw reads were filtered by a quality score, and then, the filtered reads were mapped to a reference genome using SHRiMP2. After mapping, filtration with mapping quality was carried out using Samtools. The filtered mapped reads and genotype data were analyzed by IVAS in order to obtain the candidate splicing quantitative trait loci.

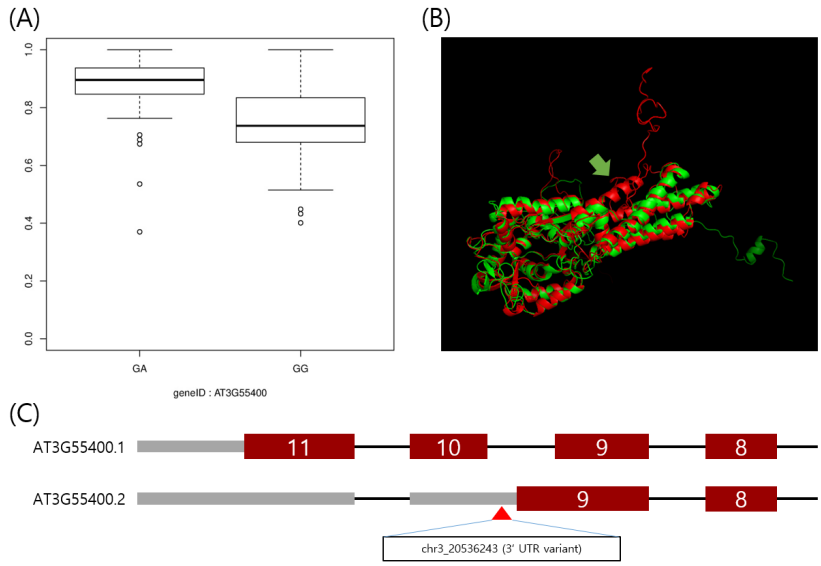

Fig. 2. Example of the utility of splicing quantitative trait loci for interpretation of trait-associated loci. (A) Boxplot of expression ratio of the exon with alternative 3-prime splicing site between phenotypes of the polymorphism, located in chr3:20536243. The expression ratio is higher in the heterozygote. (B) This is the predicted protein structure of the AT3G55400 gene using the RaptorX web-based tool [16]. The alternative 3-prime splicing site leads to the loss of alpha helices (yellow arrow). (C) The single nucleotide polymorphism, a 3′-untranslated region variant, likely leads to an alternative 3′ splicing site.

shows the association between its genotype and an expression ratio of the AT3G55400 ninth exon.

The sQTLs and trait-associated SNPs

Here, we interpreted the molecular mechanism of previously published trait-associated SNPs into sQTLs. The published trait-associated SNPs were downloaded from the 107 phenotypes database [3]. This database includes 107 phenotypes of 4 categories. We looked for the sQTLs that overlapped with SNPs in the database resource and selected SNPs having a significant association with phenotypes. As a result, we found 96 candidate sQTLs that were able to explain the mechanism of trait-associated SNPs. Among the 96 candidate sQTLs, we finally selected 25 sQTLs affecting high differential expression of an alternatively spliced target exon across genotypes. We show an example of a trait-associated sQTL in Fig. 2. The SNP, which resides in chr3:20,536,243, affects the ninth exon of AT3G55400, *OVA1*. That is, the expression ratio of the exon with the alternative 3′ splice site is higher in the heterozygote (AG) than in the homozygote (GG). *OVA1* encodes a methionine-tRNA ligase and has an association with ovule development, and its polymorphism was reported to be associated with the pedicel [3]. We report the full list of trait-associated sQTLs in Table 1.

Discussion

Here, we report 1,694 genome-wide sQTLs in *Arabidopsis thaliana* with a high-throughput RNA-seq dataset. RNA-seq using next generation sequencing provides an accurate method for annotation and quantification of exons than the exon array method, and this is greatly helpful for the investigation of sQTLs [15]. *Arabidopsis thaliana* is a model organism to study the genetic elements of plants [1]. Thus, an sQTL survey in *Arabidopsis thaliana* as a model organism has importance for understanding the regulation of alternative splicing of plants. However, there are only a few studies on sQTLs in plant. Thus, our result may be a great resource to explain the genetic regulation of alternative splicing patterns. Furthermore, the sQTLs can be useful for explaining known trait-associated SNPs. A recent study showed that genetic markers with various traits have been discovered in *Arabidopsis* [3]. However, their molecular mechanisms have remained elusive. sQTLs are able to alter alternative splicing patterns, and consequently, they can lead to phenotypic variations [5-8]. Thus, our result can be useful in interpreting trait-associated SNPs as sQTLs. For example, the *OVA1* gene encodes a functional protein for methionine-tRNA ligase, having several alpha-helices. The polymorphism, in chr3:20,536,243, was reported to have an association with the flower pedicel [3]. However, this molecular mechanism has been not elucidated. Through our result, the SNP was shown to have an association with an alternative 3′ splicing site of the ninth exon of the *OVA1* gene. We predicted the three-dimensional structure of proteins translated by transcription, including normal exons and the alternative 3′ splicing site of the ninth exon, respectively, using the RaptorX Structure Prediction tool, a web-based protein structure prediction tool [16]. As a result, translation of the transcript with the alternative 3′ splicing site showed a loss of 2 alpha helices (Fig. 2B). It is tempting to speculate that the SNP, in chr3:20,536,243, which is known as a trait-associated locus, is able to lead to translation of abnormal

Table 1. The candidate trait-associated sQTLs

SNP	Chromosome	TargetExon[a]	Type[b]	p-value	Gene	Method	Representative phenotype category[c]
chr1_2437747	1	2437678-2437833	SE	2.24E-11	AT1G07890	lm	Development
chr1_2437747	1	2437681-2437833	SE	2.44E-06	AT1G07890	lm	Development
chr1_4985421	1	4985428-4985656	A3SS	1.68E-06	AT1G14570	lm	Flowering
chr1_4985421	1	4985491-4985656	A3SS	1.68E-06	AT1G14570	lm	Flowering
chr1_9663440	1	9663959-9664053	A3SS	0.000399265	AT1G27752	lm	Flowering
chr1_9663440	1	9663970-9664053	A3SS	0.000399265	AT1G27752	lm	Flowering
chr1_20706598	1	20706649-20706893	A5SS	6.89E-10	AT1G55450	lm	Development
chr1_20706598	1	20706676-20706890	A5SS	6.89E-10	AT1G55450	lm	Development
chr1_25423641	1	25423001-25423287	A3SS	4.31E-11	AT1G67800	lm	Ionomics
chr2_18206454	2	18206433-18206480	SE	0.000247694	AT2G43970	lm	Ionomics
chr2_18226575	2	18226917-18227893	A3SS	3.12E-05	AT2G44060	lm	Development
chr2_18226575	2	18226920-18227893	A3SS	3.12E-05	AT2G44060	lm	Development
chr3_1740690	3	1740700-1740906	A3SS	6.08E-08	AT3G05840	lm	Development
chr3_1740690	3	1740729-1740906	A3SS	6.08E-08	AT3G05840	lm	Development
chr3_6903994	3	6903833-6903989	A5SS	9.33E-46	AT3G19860	lm	Ionomics
chr3_11979328	3	11978927-11979037	SE	3.18E-20	AT3G30390	lm	Development
chr3_16477273	3	16477020-16477128	A5SS	5.47E-07	AT3G45050	lm	Defense
chr3_16477273	3	16477020-16477165	A5SS	2.27E-13	AT3G45050	lm	Defense
chr3_20536243	3	20536084-20536481	IR	7.65E-11	AT3G55400	lm	Flowering
chr4_146351	4	146791-146928	A3SS	1.02E-12	AT4G00335	lm	Development
chr4_146351	4	146797-146928	A3SS	1.02E-12	AT4G00335	lm	Development
chr4_12007691	4	12007782-12007894	A3SS	7.69E-09	AT4G22890	lm	Development
chr4_12007691	4	12007788-12007894	A3SS	1.14E-13	AT4G22890	lm	Development
chr4_14266090	4	14266023-14266105	SE	6.64E-07	AT4G28910	lm	Development
chr4_14266090	4	14266023-14266144	SE	6.64E-07	AT4G28910	lm	Development
chr4_17645238	4	17644985-17645449	IR	0.003370022	AT4G37550	lm	Flowering
chr4_18014940	4	18014903-18014975	SE	2.80E-32	AT4G38510	lm	Flowering
chr5_106647	5	105976-107405	IR	9.61E-11	AT5G01260	lm	Development
chr5_12077987	5	12078034-12078207	A3SS	7.08E-10	AT5G32440	lm	Flowering
chr5_12077987	5	12078062-12078207	A3SS	7.08E-10	AT5G32440	lm	Flowering
chr5_16472181	5	16472007-16472117	A5SS	3.06E-27	AT5G41140	lm	Development
chr5_16472181	5	16472007-16472138	A5SS	3.06E-27	AT5G41140	lm	Development
chr5_18699432	5	18699385-18699429	A5SS	2.89E-07	AT5G46110	lm	Development
chr5_18757787	5	18757716-18757785	A5SS	1.83E-30	AT5G46250	lm	Defense
chr5_20970838	5	20970623-20971712	IR	1.03E-06	AT5G51630	lm	Ionomics
chr5_22244548	5	22243864-22243986	A5SS	3.07E-15	AT5G54760	lm	Development
chr5_26000452	5	26000811-26000927	A3SS	2.10E-06	AT5G65080	lm	Flowering

sQTL, splicing quantitative trait locus; SNP, single nucleotide polymorphism.
[a]The alternatively target exons.
[b]SE, skipped exon; A3SS, alternative 3′ splice site; A5SS, alternative 5′ splice site; IR, intron retention.
[c]Phenotypes associated with the polymorphisms (the first column).

protein and consequently affect the flower pedicel. Further biochemical studies of this alternative spliced transcription may be able to validate our prediction. Our result can explain only a small portion of trait-associated SNPs. However, these sQTLs can be a powerful resource for the interpretation of more trait-associated loci that will be found in the coming years. Furthermore, since there have been few investigations of genome-wide sQTLs in *Arabidopsis thaliana*, our result can be a reference resource of sQTLs for its neighboring species.

Acknowledgments

The financial support of this work was made available by the National Research Foundation of Korea (NRF-2012M-3A9D1054705), funded by the Ministry of Education, Science, and Technology, and by the Rural Development Administration of Korea (PJ01167402).

References

1. Meinke DW, Cherry JM, Dean C, Rounsley SD, Koornneef M. *Arabidopsis thaliana*: a model plant for genome analysis. *Science* 1998;282:662, 679-682.
2. Van Norman JM, Benfey PN. *Arabidopsis thaliana* as a model organism in systems biology. *Wiley Interdiscip Rev Syst Biol Med* 2009;1:372-379.
3. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, *et al*. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 2010;465:627-631.
4. Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet* 2009;10:318-329.
5. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 2010;6: e1000888.
6. West MA, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, *et al*. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* 2007;175:1441-1450.
7. Zhang X, Cal AJ, Borevitz JO. Genetic architecture of regulatory variation in *Arabidopsis thaliana*. *Genome Res* 2011;21: 725-733.
8. Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* 2007;8:749-761.
9. Han S, Jung H, Lee K, Kim H, Kim S. Genome wide discovery of genetic variants affecting alternative splicing patterns in human using bioinformatics method. *Genes Genomics (in press)*.
10. Weigel D, Mott R. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol* 2009;10:107.
11. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, *et al*. Patterns of population epigenomic diversity. *Nature* 2013;495:193-198.
12. David M, Dzamba M, Lister D, Ilie L, Brudno M. SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics* 2011;27:1011-1012.
13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078-2079.
14. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010; 28:511-515.
15. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* 2010;7:1009-1015.
16. Kallberg M, Margaryan G, Wang S, Ma J, Xu J. RaptorX server: a resource for template-based protein structure modeling. *Methods Mol Biol* 2014;1137:17-27.