Advances in
Biotechnology & Microbiology
ISSN 2474-7637

Juniper Publishers
key to the Researchers

Editorial
Volume 12 Issue 2 - January 2019
DOI: 10.19080/AIBM.2019.12.555833

Adv Biotech & Micro

Allele-specific genomic variations and transcriptomic research on quantitative phenotyping in plants

Kan Liu*

Department of Computer Science and Engineering, University of Nebraska-Lincoln, USA

Submission: December 01, 2018; **Published:** JAnuary 11, 2019

***Corresponding author:** Kan Liu, Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, USA

Keywords: promoter region, splicing region, coding regions, Major Allele Frequency, homologous regions, Variant Call Format, mechanisms, random sequencing, Haplotype information, accuracy, correlation structure, global pattern, significance

Abbreviations: QEL: Quantitative Expression Loci; ASE: Allele-Specific Expression; MAF: Major Allele Frequency; VCF: Variant Call Format

Editorial

Genomes and interactions among gene products with other molecules are the physical fundamentals of biological systems. This is especially true for research in plants, which usually have complicated genomes and many important traits like yield, plant height, and stress tolerance are quantitative. Many quantitative traits are usually controlled by more than one Quantitative expression Loci (QTL) through the regulation of gene expression. The expression of a gene could be associated with a genetic variant far away from it, which is called trans-eQTL (expression QTL) or be affected by a local variant, which is named cis-eQTL. With a widespread existence throughout the plant genome, cis-acting genetic variants have been proven to account for a larger proportion of variation in gene expression. However, it is challenging for identifying cis-eQTL in the population with sequencing data because the power of eQTL mapping is either constrained by sample size or reduced by confounding factors. For organisms with a diploid genome, the information of Allele-Specific Expression (ASE) which could provide more direct evidence of cis-eQTL is often ignored or discarded due to the unavailability of haplotype information and mapping bias.

The next-generation sequencing provided a huge potential to study genome structures and gene regulation/interactions, and it helps to link complex traits and underlying biological mechanisms from the perspective of genetic variants and gene expression regulation. Recent years, the lower cost of next-generation sequencing enables a large amount of genotype information to be produced and various strategies have been proposed to control the mapping bias. ASE analysis can provide more direct evidence of the existence of cis-eQTL compared with traditional eQTL mapping. The power of ASE analysis depends on its internally controlled system, in which transcript abundance from different alleles is compared within individuals. In such a system, noises from the environment, batch effect or trans-acting variants would not affect the transcript abundance comparison since they exert the same effect on the expression of different alleles. RNA-seq can not only measure the abundance of transcripts but also provide information on genetic variants required to differentiate between paternal transcripts and maternal transcripts.

Gene expression could be affected by genetic variants and is a major way whereby genetic variants exert their influence over traits. Depending on the physical distance from the regulated genes, genetic variants could be put into two categories, cis-genetic variation, and trans-genetic variation. Typically, cis-genetic variation is located within 1MB each side of the transcribed region while tran-genetic variation is located much farther away (more than 5MB upstream or downstream of transcribed region) or even at different chromosome [1]. cis-genetic variation could affect the transcription initiation, rate, and transcript stability by altering promoter region, splicing region, or coding regions. Many publications have demonstrated a larger effect size of cis-genetic variation on gene expression [2] Contributing to the advancement of micro-array and next-generation sequencing, the variation of total read counts in the population could be used to map eQTL with eGWAS [3].

However, the power of eGWAS relies on the sample size and most times could be biased by many factors like environmental interference, batch effect, and population structure. Although those interference factors could be efficiently eliminated or well explained by specified model terms using sophisticated experimental designs and various statistical procedures were developed, some intrinsic characterizations of eQTL analysis still prevent them from being a powerful *cis*-eQTL detector [4-7]. For instance, the discrimination between *cis*-eQTL and trans-eQTL depends on the physical distance, which is difficult for people to find a cutoff to make an unambiguous separation. In addition, the effect of an allele on transcript abundance tends to be masked

by other molecular mechanisms like negative-feedback control. Rare variations which play an important role in gene expression usually fail to pass the criterion of Major Allele Frequency (MAF) and thus are excluded from the eGWAS study [8].

The deficiency of traditional eQTL mapping in analyzing *cis*-acting genetic variants could be made up by Allele-Specific Expression (ASE) analysis. In diploid organisms, ASE refers to the differential expression levels between paternal and maternal copies of the same transcript, distinguished by heterozygous sites within the transcript. The most attractive feature of ASE analysis is that the expression levels of two alleles are compared within the same sample, excluding extra trans-acting genetic and environmental noises that would rather increase variations among individuals in eQTL analysis [9,10]. Since the variation identified in ASE analysis only affects the transcription process of the local allele, we are endowed with more confidence in determining *cis*-eQTLs from trans-eQTLs. Research in cancer area also suggests that, besides common *cis*-eQTL, the ASE analysis has the ability to detect rare regulatory variation [11,12]. What is more, the technology has progressed from single gene qRT-PCR to the next-generation sequencing, which has scaled up the ASE analysis to the whole-genome level. One thing to be noted in ASE analysis is that the imbalanced expression between haplotypes could be due to epigenetic mechanisms like imprinting. Typically, extra experiments like family trio study would be performed to confirm the situation of epigenetic effect.

Both the characterization of genetic variations and the calculation of allelic read counts could be biased by mapping reads to the haploid reference genome. As we know, the most fundamental step in next-generation sequencing analysis is mapping short reads onto the reference genome. During the mapping process with the haploid reference genome, reads overlapping indels positions tend to suffer from severe mapping bias. To be specific, the mapping process tends to keep reads with reference allele and discard reads with alternative allele, even with many aligner tools supporting gapped alignment. The same issue also exists for reads with SNP features. Combined with the random sequencing error nearby, the reads from variant alleles are prone to be discarded or aligned to a similar incorrect genome region [13,14]. The direct consequence of this intrinsic mapping bias would be underestimated reads with alternative alleles and this underestimation could, in turn, affect the downstream analysis like new variant discovery, genotype calling, and association research.

In order to reduce the mapping bias caused by the universal reference genome, researchers came up with different strategies from the aspect of refining the mapping reference. The most straightforward way is to mask all known SNP positions with the ambiguity nucleobase 'N', eliminating the intrinsic difference between the reference genome and the alternative genome. Since both references reads and allele reads can map equally to the reference genome, 'N-masking' method can significantly reduce mapping bias [15]. However, it was reported that 'N-masking' suffers from low overall mapping success rate when there exists moderate mapping error. Moreover, as the number of masked sites for one read increases, the sensitivity of mapping could be severely affected especially for homologous regions, and the correct discrimination from homologous regions requires information from masked regions.

Dewey et al proposed to do the read mapping against ethnically concordant major allele reference genome. Although a significant improvement could be seen with Dewey's method, the reference used was not able to represent all the variants found in the population. Another practical idea is to build personalized reference genome which was mainly applied to reduce the bias appeared in allele-specific mapping with RNA sequencing data and Chip-Seq data and has been further developed to improve the accuracy of genotype calling with WGS or WES data [14]. The construction of the personalized diploid reference genome was realized by modifying the custom haploid reference genome with known individual genetic variants (SNPs, Indels, and SVs). There are some well-known tools for this purpose. Vcf2diploid, as the first part of AlleleSeq pipeline, takes Variant Call Format (VCF) files with information of genetic variants as input and output two complete haplotype genomes one for paternal haplotype and the other for maternal haplotype. In the meantime, annotation files, splice-junction library and map files which record relative positions between paternal, maternal and reference haplotype are produced [16].

The shortcoming of vcf2diploid is the large size of the reference genome produced, which might not be so efficient during mapping and thus incompatible with some mapping aligner. Ref Editor is another tool designed to improve mapping accuracy through constructing personalized reference genomes. Compared with AlleleSeq which builds the whole allele chromosome, Ref Editor adopts a more efficient strategy which creates "mini chromosome". For the homozygous wild-type allele, nothing was changed; for the homozygous mutant type allele, the nucleotides in the reference genome are edited; for heterozygous genotype, a short sequence overlapping the SNP or Indel position was created and named as "mini chromosome". The original chromosome, along with the "mini chromosome", is used as the personalized reference genome [14].

Despite the great advancement in both technologies and methods, little change has been made in ASE research when it comes to plant area; as ever, sequencing reads are simply mapped to the haploid reference and a naïve binomial test is used to prove the significance. All the above issues, combined with the complexity of plant genome, would introduce non-ignorable bias to the available discovery system, reducing the reliability of the identified genetic variants. Moreover, most research was limited to the general description of the global pattern of cis-acting genetic variants, lacking association with phenotype and deep exploration of the underlying molecular mechanisms. This is mainly due to the complex genetic correlation between genetic variants and multiple causal variant candidates and partly due to

How to cite this article: Kan Liu. Allele-specific genomic variations and transcriptomic research on quantitative phenotyping in plants. Adv Biotech & Micro. 2019; 12(2): 555833. DOI: 10.19080/AIBM.2019.12.555833.

the high false positive rate in a single test. Therefore, developing new computational pipeline that are specifically for plants which would increase the accuracy in isolating functional cis-acting variants with ASE information and will integrate information from variation annotation, correlation structure, and phenotype to further characterize the biological function of identified cis-eQTLs.

References

1. Nica AC, Dermitzakis ET (2013) Expression quantitative trait loci: present and future. Philos Trans R Soc Lond B Biol Sci 368(1620): 20120362.
2. Kabakchiev B, Silverberg MS (2013) Expression Quantitative Trait Loci Analysis Identifies Associations Between Genotype and Gene Expression in Human Intestine. Gastroenterology 144(7): 1488-1496. e3.
3. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science, 296 (5568): 752-755.
4. Listgarten J, Kadie C, Schadt EE, Heckerman D (2010) Correction for hidden confounders in the genetic analysis of gene expression 107 (38): 16465-16470.
5. Zhou X, Stephens M (2012) Genome-wide Efficient Mixed Model Analysis for Association Studies. Nat Genet 44(7): 821-824.
6. Shabalin AA (2012) Matrix eQTL: Ultra-fast eQTL analysis via large matrix operations. Bioinformatics 1353-1358.
7. Gong J, Mei S, Liu C, Xiang Y, Ye Y (2018) Pancan QTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. Nucleic Acids Res D971–D976.
8. Pastinen T (2010) Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev Genet 11(18): 533-538.
9. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10(3): 184-194.
10. Pastinen T, Hudson TJ (2004) Cis-acting regulatory variation in the human genome. Science 306(5696): 647-650.
11. Montgomery SB, Lappalainen T, Gutierrez-Arcelus M, Dermitzakis ET (2011) Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet 7(7): e1002144.
12. Pastinen T (2010) Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev Genet 11(8): 533-538.
13. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, et al. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25(24): 3207-3212.
14. Yuan S, Johnston HR, Zhang G, Li Y, Hu YJ, et al. (2015) One Size Doesn't Fit All-RefEditor: Building Personalized Diploid Reference Genome to Improve Read Mapping and Genotype Calling in Next Generation Sequencing Studies. PLoS comput biol 11(8): e1004448.
15. Krueger F, Andrews SR (2016) SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. F1000Res 5: 1479.
16. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, et al. (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol 7: 522.

DOI: 10.19080/AIBM.2019.12.555833

How to cite this article: Kan Liu. Allele-specific genomic variations and transcriptomic research on quantitative phenotyping in plants. Adv Biotech & Micro. 2019; 12(2): 555833. DOI: 10.19080/AIBM.2019.12.555833.