Gene 533 (2014) 366–373

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Methods paper

Genome-wide identification of allele-specific effects on gene expression for single and multiple individuals

Shaojun Zhang [a,1], Fang Wang [b,1], Hongzhi Wang [a,1], Fan Zhang [b], Bin Xu [b], Xia Li [b,*], Yadong Wang [a,**]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[b] Department Of Bioinformatics, Harbin Medical University, Harbin 150086, China

ARTICLE INFO

Article history:
Accepted 8 September 2013
Available online 11 October 2013

Keywords:
Allele-specific expression
RNA-seq
Maximum likelihood model
Populations

ABSTRACT

The analysis of allele-specific gene expression (ASE) is essential for the mapping of genetic variants that affect gene regulation, and for the identification of alleles that modify disease risk. Although RNA sequencing offers the opportunity to measure expression at allele levels, the availability of powerful statistical methods for mapping ASE in single or multiple individuals is limited. We developed a maximum likelihood model to characterize ASE in the human genome. Approximately 17% of genes displayed an allele-specific effect on gene expression in a single individual. Simulations using our model gave a better performance and improved robustness when compared with the binomial test, with different coverage levels, allelic expression fractions and random noise. In addition, our method can identify ASE in multiple individuals, with enhanced performance. This is helpful in understanding the mechanism of genetic regulation leading to expression changes, alternative splicing variants and even disease susceptibility.

1. Introduction

Allele-specific gene expression (ASE) is the representation of the two alleles of a given gene in the corresponding mRNA. Normal development and cellular processes require the ratio of expression of the two alleles to be different from the allelic representation in genomic DNA (50:50). However, the precise mechanisms by which allele-specific gene expression occurs are not yet understood and there may be multiple mechanisms. Studies of expression quantitative trait loci (eQTLs) have shown that ASE usually reflects cis-acting genetic polymorphisms (Stranger et al., 2007), whereas trans-genetic regulatory or epigenetic mechanisms are relatively rare (Stranger et al., 2005; Zeller et al., 2010). It is generally believed that cis-regulatory polymorphism is the primary source of phenotypic difference and is associated with many diseases. The functional cis-regulatory variation can be mapped by measurement of ASE, using statistical or experimental approaches (Campino et al., 2008; Pastinen et al., 2005; Serre et al., 2008; Verlaan et al., 2009). In addition, although monoallelic expression is relatively rare, epigenetic mechanisms of allelic expression, such as imprinted genes, can also be detected by measuring ASE (Babak et al., 2008).

The precise identification of ASE genes has been the focus of much attention. Studies using the Illumina Allele-Specific Expression BeadArray platform and quantitative sequencing of real-time polymerase chain reaction (RT-PCR) products showed that differential allelic expression is a widespread phenomenon, which affects the expression of 20% of human genes in individuals of European descent (Serre et al., 2008). In addition, quantitative measurements of allelic expression in different HapMap populations (60 Caucasians of Northern and Western European origin (CEU), 45 unrelated Chinese individuals from Beijing University (CHB), 45 unrelated Japanese individuals from Tokyo (JPT), and 60 Yoruba from Ibadan, Nigeria (YRI)), using the Illumina BeadChips, found that approximately 18% of human genes showed differential allelic expression (Dimas et al., 2008). Statistical analyses of the Illumina BeadChip data have been used to identify genome regions that exhibit ASE. These analyses included the integration of z-score computations and a machine learning approach, based on hidden Markov models (Wagner et al., 2010). Recently, high-throughput RNA sequencing (RNA-seq) has provided a platform-independent method, similar to the microarray approach, which has allowed identification of the genetic regulatory variants at the transcript, isoform and allele levels. Statistical approaches have been proposed to characterize ASE on the basis of RNA-seq data. The binomial exact test has been applied to single nucleotide polymorphism (SNP) to test whether the expression of a reference allele was

Abbreviations: ASE, allele-specific gene expression; eQTLs, expression quantitative traits loci; RT-PCR, real-time polymerase chain reaction; CEU, Caucasians of Northern and Western European origin; CHB, Chinese individuals from Beijing University; JPT, Japanese individuals from Tokyo; YRI, Yoruba from Ibadan, Nigeria; RNA-seq, high-throughput RNA sequencing; SNP, single nucleotide polymorphisms; FDR, false discovery rate; AUC, the area under the curve; HLA-DPA1, major histocompatibility complex, class II, DP alpha 1; MHC, major histocompatibility complex; HBV, hepatitis B virus; UTR, untranslated region; ROC, receiver operating characteristic.

* Correspondence to: X. Li, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, China. Tel./fax: +86 451 86615922.
** Correspondence to: Y. Wang, Center for Biomedical Informatics, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, 150001, China. Tel.: +86 138 0450 5289; fax: +1 317 278 9217.
E-mail addresses: lixia6@yahoo.com (X. Li), ydwang@hit.edu.cn (Y. Wang).
[1] These authors contributed equally to this work.

0378-1119/$ – see front matter.
http://dx.doi.org/10.1016/j.gene.2013.09.029

greater than or less than 0.5 (Degner et al., 2009). In addition, Nothnagel et al. (2011) developed a statistical framework, based on the likelihood ratio test, to examine allele imbalance of single SNPs in RNA-seq data, which allows for allele miscalls (Nothnagel et al., 2011). A Bayesian hierarchical model has been developed by Skelly et al. (2011), using RNS-seq data from a diploid hybrid of two diverse *Saccharomyces cerevisiae* strains, which can test for ASE in both a SNP and a gene (Skelly et al., 2011).

Although some statistical approaches have been developed to test for ASE, using RNA-seq data, they mainly focus on a single SNP or a single individual. To address the lack of statistical methods for detecting ASE from high-throughput RNA-seq data, we developed a maximum likelihood model to characterize ASE from individuals and populations. In a single individual approximately 17% of genes showed ASE or variable ASE, with a false discovery rate (FDR) of 7.50%. Together with simulation experiments, our method is accurate and robust for the detection of different allelic fractions, and reads coverage levels and random noise. Furthermore, we identified more ASE genes in populations. These data provide insights into the genetic mechanism of cis-acting regulatory variants and the inconsistent effects of regulatory variants observed in different individuals.

2. Materials and methods

2.1. Human reference genome construction of SNP data

Phased variant sets were obtained from 1000 genome projects (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/integrated_call_sets), which included phased genotypes from NA12891, NA12892 and CEU individuals (lymphoblastoid samples from HapMap individuals from the CEPH—Centre d'Etude du Polymorphism Human). All heterozygote SNP genome locations were mapped and phase information was converted to the Browser Extensible Data (BED) format. The mitochondrial chromosome, Y chromosome and random genome supercontigs were excluded from the following analysis. Raw DNA sequencing data were mapped to the hg19 human reference genome sequence (GRCh37) and the BAM (Binary Alignment/Map) data were downloaded (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/data/). The data were then transformed from BAMs to Sequence Alignments/Maps (SAMs), using SAM tools. Each of the alleles was mapped to SAM alignments and allele read counts were made according to genome location of SNPs and phase information.

2.2. Allele-specific expression SNP processing

We obtained approximately 10.1 Gb of sequence for NA12891 and NA12892 by RNA-seq data, produced from high throughput sequencing (Lalonde et al., 2011). The CEU RNA-seq data sets were obtained from Montgomery et al. (2010). All raw reads were mapped to the hg19 genome sequence using Bowtie2 software (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml), with local parameters (—D 15-R 2-N 0-L 20-i S,1,0.75) that allowed a maximum of two mismatches in a seed alignment (Bowtie2 web manual) (Langmead and Salzberg, 2012). The Bowtie2 software was used to search for multiple alignments, report the best and print alignments in SAM format. In order to obtain allele-specific read counts of SNPs in each gene, SNPs were grouped according to gene annotations given by Ensembl. We examined any genic SNPs overlapping a mapped read. Reads were assigned, using Perl software, according to SNP phase information from the 1000 genome projects. This resulted in allele-specific read counts for SNPs in each gene. Allelic read counts were also obtained for heterozygous SNPs from 45 CEU individuals to estimate ASE for the population. In addition, spliced reads were processed by a spliced read mapper for RNA-seq (TopHat) to obtain allelic read counts as described above.

2.3. Maximum likelihood statistical models

We denoted y_{ij} as the allelic read counts of SNP j in gene i and N_{ij} as the read counts of SNP j in gene i. Under the null hypothesis of balanced allelic expression, y_{ij} should follow the binomial distribution, B (N_{ij}, 0.5). It is expected that the distribution of allelic fractions in genomic DNA will approximate be a binomial distribution, with a probability of 0.5. There was some divergence in the distribution of allelic fractions in genomic DNA from the binomial distribution (Fig. 1). It is suggested that a minority of SNPs are biased in genomic DNA. In previous studies, a small proportion of SNPs is biased toward one of the two alleles, and they identified the presence of flanking sequences sharing identity with another region of genome as one factor contributing to this read-mapping bias at some SNPs in humans (Degner et al., 2009; Skelly et al., 2011). The read-mapping bias at some SNPs may be the same for both genomic DNA and RNA-seq data due to the same cell and mapping method, so we tolerated the biased SNPs. When the probability of y_{ij} was not fixed, the binomial distribution was not applicable for allelic

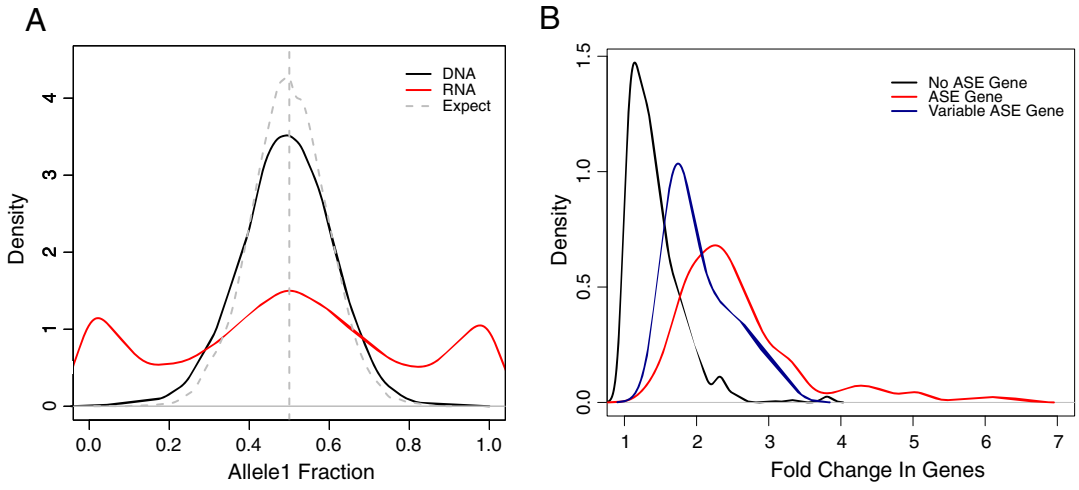

Fig. 1. The bias of allelic fraction. (A) Distribution of allelic fraction in genomic DNA and RNA data. Black solid line: the probability density curve of the allelic fraction in genomic DNA data. Red solid line: the probability density curve of the allelic fraction in the RNA data. Gray dashed curve: the probability density curve of the expected allelic fraction for allele balance, B (N, 0.5). Gray dashed vertical line: the expected fraction of allele (p = 0.5). (B) Magnitude of ASE for genes in NA12891. Red: fold-changes in ASE genes; Blue: fold-changes in genes with variable ASE; Black: fold-changes in genes with no ASE. The fold-changes were computed as the logarithm of the mean of the allelic fraction, at lower levels for all SNPs in each gene.