Bioinformatics, 2016, 1–7
doi: 10.1093/bioinformatics/btw416
Advance Access Publication Date: 13 July 2016
Original Paper

OXFORD

Gene expression

cisASE: a likelihood-based method for detecting putative *cis*-regulated allele-specific expression in RNA sequencing data

Zhi Liu[1], Tuantuan Gui[1], Zhen Wang[1], Hong Li[1], Yunhe Fu[2], Xiao Dong[3,*] and Yixue Li[1,4,5,6,*]

[1]Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China, [2]Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China, [3]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA,[4]School of Life Science and Technology, Shanghai Jiaotong University, Shanghai 200240, China, [5]Shanghai Center for Bioinformation Technology, Shanghai Industrial Technology Institute, Shanghai 201203, China and [6]Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai 200438, China

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on October 27, 2015; revised on May 16, 2016; accepted on June 24, 2016

Abstract

Motivation: Allele-specific expression (ASE) is a useful way to identify *cis*-acting regulatory variation, which provides opportunities to develop new therapeutic strategies that activate beneficial alleles or silence mutated alleles at specific loci. However, multiple problems hinder the identification of ASE in next-generation sequencing (NGS) data.

Results: We developed cisASE, a likelihood-based method for detecting ASE on single nucleotide variant (SNV), exon and gene levels from sequencing data without requiring phasing or parental information. cisASE uses matched DNA-seq data to control technical bias and copy number variation (CNV) in putative *cis*-regulated ASE identification. Compared with state-of-the-art methods, cisASE exhibits significantly increased accuracy and speed. cisASE works moderately well for datasets without DNA-seq and thus is widely applicable. By applying cisASE to real datasets, we identified specific ASE characteristics in normal and cancer tissues, thus indicating that cisASE has potential for wide applications in cancer genomics.

Availability and Implementation: cisASE is freely available at http://lifecenter.sgst.cn/cisASE.

Contact: biosinodx@gmail.com or yxli@sibs.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Allele-specific gene expression (ASE) refers to the differential expression of two alleles in a diploid genome. ASE is primarily a result of the associated cis-element variation and allele-specific epigenetic modifications (Pastinen, 2010). Germline ASEs are usually associated with the existence of regulatory elements, which exert a pronounced effect on human phenotypic polymorphism and common diseases. Somatic ASEs (ASEs of somatic mutations) in tumors are often genetically linked to tumor drivers. Thus, ASE studies provide an opportunity for the development of new therapeutic strategies that activate beneficial alleles or silence mutated alleles at specific loci (Huang *et al.*, 2012).

Next-generation sequencing (NGS) allows for the genome-wide identification of ASE (Heap *et al.*, 2010; Lee *et al.*, 2013; Smith, 2013); however, several problems exist. The first problem is technical and intrinsic allele bias, which we and others have previously demonstrated (Degner *et al.*, 2009; Liu *et al.*, 2014). For example, reads carrying single nucleotide variants (SNVs) are less likely to be correctly mapped than those carrying reference genotype (Degner *et al.*, 2009). Another example is that, a large proportion of previously identified ASEs, especially in tumor samples, have been found a result of copy number variations (CNVs) rather than cis-regulation (Tuch *et al.*, 2010). The second problem is that gene-level ASE detection usually requires phased-SNVs or parental genomes (Montgomery *et al.*, 2010; Rozowsky *et al.*, 2011; Skelly *et al.*, 2011; Zhang *et al.*, 2009), which are usually not available. Although methods to phase single-nucleotide polymorphisms (SNPs) into individual haplotypes exist (Browning and Browning, 2007; Howie *et al.*, 2009), these methods cannot be applied to somatic mutations in tumors. The third problem relates to testing ASE with statistical confidence. Two simple statistical methods, i.e. binomial and chi-square tests, have been frequently used in SNV-level ASE detection (Ge *et al.*, 2009; Heap *et al.*, 2010; Zhang *et al.*, 2009); however, these methods do not make full use of the information of these complex sequencing datasets. Although several tools have been developed for ASE detection, none has addressed all of these concerns (Mayba *et al.*, 2014; Pandey *et al.*, 2013; Rozowsky *et al.*, 2011; Skelly *et al.*, 2011).

To overcome the problems outlined above, we proposed a new computational method and developed a software tool, cisASE, based on a likelihood ratio test. cisASE uses DNA-seq data to make site-by-site adjustments for RNA allele imbalance assessment to reduce the effects of technical bias and CNV, and reports ASE putatively caused by *cis*-regulation. cisASE also considers the base quality of each base to reduce the influence of sequencing error. We tested cisASE on both simulated and real datasets, and compared it with other methods (Mayba *et al.*, 2014; Skelly *et al.*, 2011). cisASE exhibits significantly improved accuracy and computational speed compared with existing methods.

We applied cisASE to public colon tumor datasets, and observed several important features, i.e. germline ASE hotspots of human leukocyte antigen (HLA) loci, cancer somatic ASE genes in focal adhesion and extracellular matrix (ECM)–receptor interaction pathways. These findings revealed a landscape of germline and somatic ASE in colon tumors and highlighted broad applications of our cisASE method in ASE detection.

2 Materials and methods

2.1 Simulated data

We selected normal tissue data ($n = 46$) of a human colon dataset ($n = 92$) (Seshagiri *et al.*, 2012) as the basis for generating a simulated dataset (Supplementary Material) because this dataset contains high-quality matched DNA-seq and RNA-seq data with high sequencing depth. The dataset was also used to test our method (described in the following sections).

First, fold changes of adjusted RNA allele counts of the real data were calculated using the following equation,

$$fc = \frac{\sum_{i=1}^{n} RNAref / \sum_{i=1}^{n} RNAalt_i}{\sum_{i=1}^{n} DNAref / \sum_{i=1}^{n} DNAalt_i} \quad (1)$$

where n is the number of SNVs for a specific feature (SNV, exon or gene). The distributions of log2 fc and log2 DNA depth were then modeled using *Laplace* (Supplementary Fig. S1) and normal distributions, respectively. For each simulated SNV, its DNA sequencing depth was drawn from the normal distribution, and its RNA sequencing depth was sampled from the real data. Both DNA and RNA reference allele counts were assumed to exhibit a binomial distribution$r \sim B(n, p)$, where n is the sequencing depth of DNA-seq or RNA-seq, and p is the frequency of reference allele.

To simulate data without CNV (non-CNV data), DNA reference allele counts were subjected to a binomial distribution with parameter p estimated from real DNA-seq data. To mimic the proportion of ASEs in real data, we sampled fc for each SNV from the above *Laplace* distribution. When the sampled fc exceeded a certain cutoff (1.5, 2, 3, 4 or 5), we generated a true positive ASE SNV, whose RNA reference allele count was subject to a binomial distribution with parameter p determined as the fraction of DNA reference allele after adjustment by the sampled fc (illustrated in Supplementary Fig. S2). Otherwise, we generated a true negative site with the parameter p of the binomial distribution setting as its DNA reference allele fraction. We simulated 1000 genes for each fc cutoff, and the number of SNVs in each gene was also sampled from the real data.

To simulate data carrying CNV (CNV data), DNA reference allele count of each SNV was assumed to have a binomial distribution, with a DNA ratio of reference (R) and alternative (A) alleles n(R):n(A)=2, 3 or 4, for DNA copy numbers of 2, 3 and 4, respectively.

Next, to assess the performance of cisASE at different sequencing depths, we simulated another non-CNV dataset with fixed depth (10–20, 20–30, 30–40 and 40–50) and fixed number of SNVs per gene (1, 2, 3, 4 or 5), resulting in datasets with 125 strata (Supplementary Material).

2.2 Real sequencing datasets

We collected three real sequencing datasets to test cisASE. First, we obtained RNA and Exon sequencing data of 46 pairs of matched human colon tumor-normal samples from the European Genome-Phenome Archive (EGA) under accession number EGAS00001000288 (Seshagiri *et al.*, 2012). Second, RNA-seq and DNA-seq data for a well-studied individual NA12878 were obtained from the Gene Expression Omnibus (GEO) under accession number GSE30401 (Rozowsky *et al.*, 2011) and the ENA with accession number ERR194147, respectively. The phased genomic variants of individual NA12878 in hg19 coordinates were obtained from Rozowsky *et al.* (2011). Additionally, a mouse dataset including exon and RNA sequencing data was downloaded from the European Nucleotide Archive (ENA) under accession numbers PRJEB5321 and PRJEB5320, respectively (Castle *et al.*, 2014). DNA-seq and RNA-seq reads were mapped to the reference genome using bowtie2, and all duplicated reads were removed to avoid the effects of polymerase chain reaction (PCR) duplication.

2.3 Methods

The allele ratios of reference and alternative alleles in normal samples deviate from expected value (1:1) (Supplementary Fig. S3) because of technical bias in strand capture, sequencing process and mapping. Additionally, in cancer samples, because of somatic structure variations, e.g. CNV, the bias is more severe (Supplementary Fig. S3). To detect putative cis-regulated ASE regardless of bias and CNV, we use DNA-seq data for site-by-site adjustment based on the assumption of Zhang *et al.* (2009) that technical bias in the

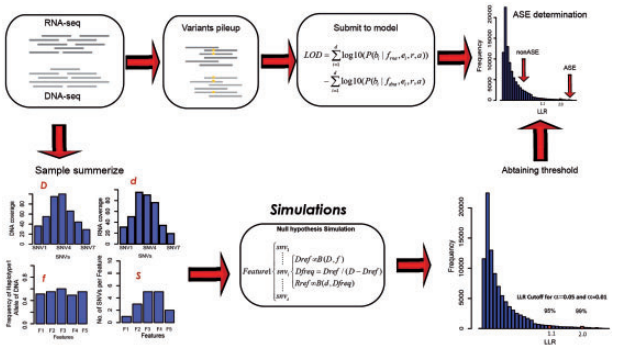

Fig. 1. Overview of cisASE algorithm. Mapped RNA-seq and DNA-seq data are piled up using samtools. LLR of the null and alternative models is calculated for each feature, i.e. SNV, exon or gene. Simulated DNA and RNA allele counts are generated from the data to produce a null distribution of the LLR, which is used to define an LLR cutoff with respect to a particular significance level. The LLR of each measured feature is compared to the LLR cutoff to determine whether it is an ASE

measurement of the allele ratio for DNA is the same for RNA. cisASE pipeline is illustrated in Figure 1, including following steps: (1) pileup files are generated from DNA-seq and RNA-seq data (mapped bam files after duplication removal) using samtools (Supplementary Material) (Li *et al.*, 2009). (2) The pileup files are submitted to cisASE, and log-likelihood ratio score (LLR) for each feature (SNV, exon or gene) is calculated. (3) For each input sample, the null distribution of LLRs is simulated from the input itself, and the LLR cutoffs corresponding to alpha levels of 0.05 and 0.01 are determined from the null distribution. (4) The LLRs are compared with a cutoff at a specific alpha level to determine whether the features are ASEs.

2.3.1 Model for SNV-level LLR calculation

We take standard pileup files as input and calculate the probability of observing base b_i on read i ($i = 1 \ldots d$, where d is the coverage of base b_i) at each SNV following Cibulskis *et al.* (2013),

$$P(b_i|f,e_i,r,a) = \begin{cases} f.\dfrac{e_i}{3} + (1-f)(1-e_i) & \text{if } b_i = r \\ f(1-e_i) + (1-f).\dfrac{e_i}{3} & \text{if } b_i = a \\ \dfrac{e_i}{3} & \text{otherwise} \end{cases} \quad (2)$$

where r and a refer to the reference and alternative allele respectively. f refers to the fraction of the alternative allele. e_i is the probability of errors in that base call (each base has an associated Phred-like quality score, q_i, where $e_i = 10^{-q_i/10}$). We assume that the sequencing errors are independent for each read and that substitution errors are uniformly distributed and occur with probability $e_i/3$.

To assess allelic imbalance, we explain the observations with the following two hypotheses,

$$\begin{aligned} H_0 &: f = DNAalt/(DNAref + DNAalt) \\ H_1 &: f = RNAalt/(RNAref + RNAalt) \end{aligned} \quad (3)$$

where *DNAref/alt* and *RNAref/alt* represent the allele read counts of the reference and alternative alleles in the DNA and RNA sequencing data, respectively. ASE detection is performed by comparing the likelihoods of two hypotheses. When the LLR, as defined below, exceeds a decision threshold, we identify the SNV as an ASE.

$$\begin{aligned} LLR &= \log10\left(\frac{L(H_1|X)}{L(H_0|X)}\right) \\ &= \sum_{i=1}^{d}\log10(P(b_i|f_{rna},e_i,r,a)) - \sum_{i=1}^{d}\log10(P(b_i|f_{dna},e_i,r,a) \end{aligned} \quad (4)$$

2.3.2 Model for gene/exon-level LLR calculation

For genes harboring n informative heterozygous SNVs, the hypotheses are modified as follows,

$$\begin{aligned} H_0 &: f = \frac{\sum_{i=1}^{n} DNAhap2}{\sum_{i=1}^{n}(DNAhap1 + DNAhap2)}; \\ H_1 &: f = \frac{\sum_{i=1}^{n} RNAhap2}{\sum_{i=1}^{n}(RNAhap1 + RNAhap2)} \end{aligned} \quad (5)$$

where *DNAhap1/2* and *RNAhap1/2* represent the allele counts of the haplotype 1/2 allele. We implemented the pseudo-phasing process applied by Mayba *et al.* (2014) in cisASE. This process estimates the major and minor haplotypes by assigning SNV-level alleles with higher RNA read counts to the 'major' haplotype (hap1) and those with lower RNA read counts to the 'minor' haplotype (hap2). This process is applied to both gene-level and exon-level ASE detection when no phasing information is provided. Alternatively, users can provide their own phasing information to cisASE.

One problem in gene-level ASE detection is that alternative splicing can affect the ASE identification when skip efficiency differs between two haplotypes. cisASE can accept annotation files containing labels for constitutive exons (parameter – C1) and considers only SNVs within constitutive exons when measuring gene-level ASE. A chi-square test is applied to check the consistency of the allele ratios for all SNVs within a gene. For multi-SNV genes, a small P-value from the chi-square test indicates that individual SNVs differ significantly in their allelic ratios. This inconsistency can arise because of differences in ASE between various transcript isoforms of a gene or counting errors for SNVs prone to technical bias. Candidate ASE genes with this inconsistency (e.g. P-value of the chi-square test <0.05) should be filtered out by users.

2.3.3 Generating a null distribution

To determine a decision threshold to reject the null hypothesis, we perform the following simulation to calculate the threshold for a specific significance level a.

For each input dataset, we simulate paired DNA and RNA data from the dataset itself. The simulated DNA and RNA sequencing depths are randomly sampled from the input. Read counts of both the simulated DNA and RNA reference alleles are assumed to follow a binomial distribution$r \sim B(n,p)$, where p is the average DNA or RNA reference allele frequency of the input. Base qualities and the number of SNVs in each feature (i.e. SNV, exon or gene) are also sampled from the input, and sequencing errors (0.001 errors per base) are added to the simulated reads. The LLR of each simulated feature is calculated according to Equation (4).

By repeating the simulation S times, we estimate a null distribution of the LLR and determine a cutoff corresponding to significance level *a* according to Equation (6),

$$P(\Lambda(X) \leq \eta | H_0) = \alpha \quad (6)$$

The default simulation repeat time S is set to 2000, which is reasonable for obtaining a robust cutoff (Supplementary Fig. S4; discussed in the Supplementary Material), and the LLR cutoffs at alpha levels of 0.05 and 0.01 are provided.cisASE can also be applied to datasets without parallel DNA-seq data. In this situation, the pre-existing DNA bias can be set by users; otherwise, the pre-existing DNA bias is assumed to be the same as the mean of the RNA bias.

In addition, when simulation is disabled by users, we recommend SNV-level LLR cutoffs of 0.85 and 1.47 and gene-level LLR cutoffs of 0.82 and 1.50 for alpha levels of 0.05 and 0.01, respectively, based on a simulation of 92 sets of high-quality sequencing data (Seshagiri *et al.*, 2012). To test the robustness of the recommended LLR cutoff, we applied it to Seshagiri's data (Seshagiri *et al.*, 2012) and NA12878, and used the recommended LLR and the LLR calculated from the dataset itself as the cutoff to define ASE genes. This process revealed 4.2 and 2.3% differences in the Seshagiri's data and NA12878, respectively (see the Supplementary Material for details), indicating that, with the recommended LLR, cisASE has acceptable performance.

2.3.4 Implementation of the chi-square test

We also implemented chi-square test, a widely used test that uses count information of both DNA and RNA for ASE identification (Heap *et al.*, 2010; Zhang *et al.*, 2009), in cisASE. We adjust the DNA and RNA counts for the chi-square test using base qualities. The adjusted counts for each base is defined as $1 - e_i$, where e_i is the error rate of base *i*. In the absence of DNA input, we simulate the DNA reference and alternative allele counts for each SNV from the RNA sequencing depth and assume a DNA bias as mentioned in the Section 2.3.3. The allelic counts are summed for each haplotype for multi-SNV genes and exons for testing.

3 Method evaluation

3.1 Evaluation in simulated data

To evaluate the effectiveness of taking DNA-seq data as a reference for reducing bias, we ran cisASE with the following four settings, using simulated non-CNV data and compared the performance (Fig. 2 and Supplementary Table S1):

A. cisASE default (adopt site-by-site bias adjustment with DNA-seq data);

B. set no DNA bias;

C. set DNA bias as the mean reference allele fraction of all SNVs in the DNA-seq data; and

D. set DNA bias as the mean reference allele fraction of all SNVs in the RNA-seq data.

The highest performance was achieved when using setting A. The performances of cisASE with settings B and C were similar (overlapping in Fig. 2) and inferior to that obtained with setting A. These results suggest that bias is a local event and global adjustment does not perform better than no adjustment. The performance was worst when we estimated DNA bias from RNA-seq data (setting D). Because of overdispersion (Skelly *et al.*, 2011) and uneven sequencing depth of RNA-seq data, allele bias in a small percentage of highly expressed transcripts may shift the mean bias away from 0.5, resulting in higher false positive and false negative rates. Therefore, we recommend running cisASE with setting B, in which the DNA reference allele fraction is set to the expected value (0.5), for experiments without matched DNA-seq data. Of note, in this case, users cannot distinguish between cis-regulated ASE and CNV-resulted ASE.

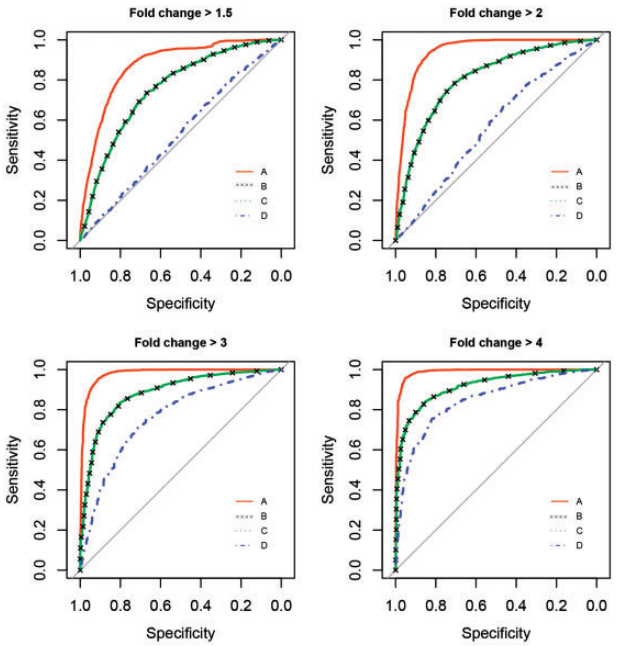

Fig. 2. The performance of cisASE in simulated data with four different settings. We applied cisASE to simulated data with four different settings for pre-existing DNA bias. (A) site-by-site bias calculated from DNA-seq; (B) no bias; (C) mean bias of the DNA-seq data; and (D) mean bias of the RNA-seq data. B and C are overlapped

We compared cisASE with MBASED (Mayba *et al.*, 2014), which, to the best of our knowledge, is the only currently published ASE-detection method that allows for gene-level ASE detection without phasing information (Supplementary Table S2). We applied MBASED to our simulated data with its default settings. For simulated CNV data, cisASE substantially outperformed MBASED (Fig. 3a). As illustrated in the table in Figure 3a, when a SNV is located in a CNV region with a DNA ratio of 3:1 and its RNA allele ratio is 1:1, MBASED produced a false negative result, whereas for an RNA allele ratio of 3:1, MBASED produced a false positive result. cisASE also performed better with simulated non-CNV data (Fig. 3b). Additionally, the computational speed of cisASE was faster than that of MBASED. On average, for a sample with 3500 SNVs, the running time for MBASED was 460 min, and cisASE was seven times faster, i.e. 68 min.

Next, we compared our method with the chi-square test. The results showed that, although the two methods exhibited comparable specificity, the sensitivity of cisASE was much higher than that of the chi-square test (Supplementary Fig. S5). We also measured the accuracy of the chi-square test relative to that of cisASE. cisASE outperformed the chi-square test in each stratum of the simulated data, especially when the fold change threshold of the simulated data was low (fc = 1.5), indicating its superior performance in detecting mildly biased ASE and ASE at low sequencing depth (Fig. 4).

3.2 Evaluation in real data

We assessed the performance of cisASE in paired human colon tumor-normal samples (Seshagiri *et al.*, 2012). We randomly selected 16 of the 46 pairs of paired data to compare the methods

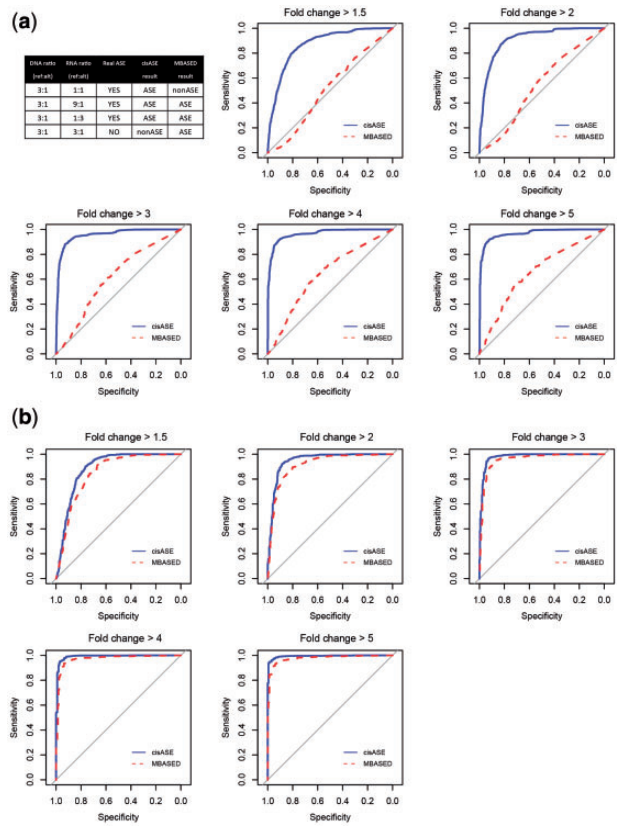

Fig. 3. Comparison of cisASE with MBASED in simulated data. cisASE performs better than MBASED in each stratum of the simulated data with (a) and without (b) CNV

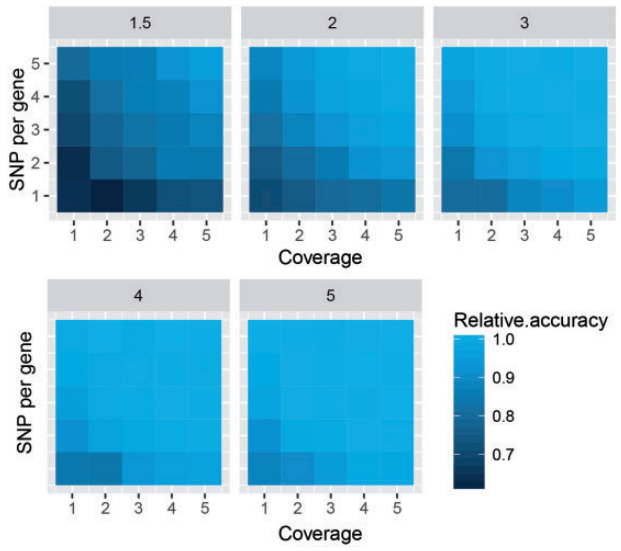

Fig. 4. Comparison of cisASE with the chi-square test in simulated data. The colors represent accuracy of the chi-square test relative to that of cisASE

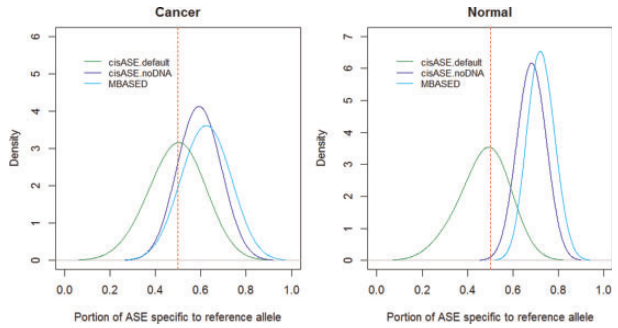

Fig. 5. Distribution of the percentage of ASE overexpressing reference alleles in a real dataset. For each cancer ($N = 16$) and normal sample ($N = 16$), we calculated the percentage of ASE specific to the reference allele (ASE with overexpressed reference allele)

because some of the methods have speed limitations. Ideally, without technical and assessment bias, we would expect a balanced proportion of ASE SNVs over-expressing reference allele and alternative allele (0.5 versus 0.5). For ASE SNVs identified in the data, the proportion of ASEs over-expressing reference alleles was 0.50 when using cisASE and 0.67 when using MBASED (Mayba *et al.*, 2014) (Fig. 5). We also compared the performance of cisASE with and without DNA-seq data as an input for bias adjustment (Fig. 5). These results indicate that cisASE efficiently decreased the false positive results caused by pre-existing bias, especially bias toward the reference allele.

Next, we applied cisASE to a mouse colon carcinoma cell line (Castle *et al.*, 2014). We identified 43 ASE SNVs with cisASE, covering all 31 SNVs identified by chi-square test at the same significance level (P-value = 0.01). As demonstrated in Supplementary Figure S5, for the SNV-level ASE detection, even for the most biased SNVs (fc = 5), sensitivity of the chi-square test was not as high as that of cisASE until the sequencing depth increased to 40. Of the 12 ASE SNVs only detected by cisASE, 10 had sequencing depths less than 40 for DNA or RNA, and one had a sequencing depth less than 60. The remaining one had a chi-square P-value that fell slightly short of the chosen cutoff of 0.01 (Supplementary Table S3). These results were consistent with our conclusions based on the simulated data, suggesting that cisASE has higher sensitivity, especially in cases of low sequencing coverage.

We further applied cisASE to a high-quality phased individual, NA12878, from the 1000 Genomes Project. First, we assessed the performance of cisASE with and without known phasing information. RNA and DNA sequencing data were pre-processed analogously to our previous dataset (Supplementary Material). Overall 2295 genes were tested, including 963 (42%) with >1 heterozygous locus. In total, 274 and 291 genes (Supplementary Table S4) were classified as ASE genes at a significance level of 0.05 with and without the known haplotype as input, respectively. Among these genes, 273 were shared. Similar to previous reports (Mayba *et al.*, 2014), pseudo-phasing had a high recovery rate of the true haplotype, and 98.0% (950/963) of the multi-SNV genes were correctly classified by cisASE when no phasing information was supplied. These results indicated that cisASE is an accurate method, even without phasing information.

We then compared cisASE with MBASED(Mayba *et al.*, 2014) and Skelly's method (Skelly *et al.*, 2011) (see the Supplementary Material for details). Skelly's method requires known haplotype data as input and incorporates DNA-seq information as the overall adjustment for bias. In this comparison, haplotypes were supplied to all three methods. A total of 274 (significance level = 0.05, adjusted MAF $\geq$ 0.7), 114 ($P \leq 0.05$, MAF $\geq$ 0.7) and 137 [posterior P(ASE) >0.95, posterior median MAF $\geq$ 0.7] ASE genes were detected by cisASE, MBASED and Skelly's method, respectively (Supplementary Table S5). cisASE identified 85% ASE genes reported

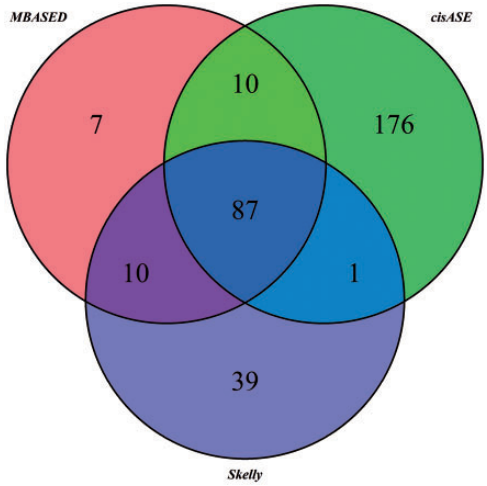

Fig. 6. Comparison of cisASE with other methods in individual NA12878

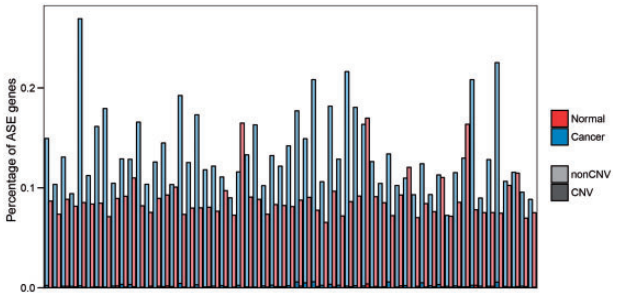

Fig. 7. Comparison of ASE in paired tumor and normal samples

by MBASED and 64% of those identified by Skelly's method but was more sensitive (Fig. 6 and Supplementary Fig. S6). Additionally, cisASE required much less computation time, i.e. 90 min compared to 360 min for Skelly's method and 660 min for MBASED.

4 Applications

To demonstrate the application of cisASE, we applied it to a panel of 46 paired tumor-normal samples (Seshagiri *et al.*, 2012).

4. 1 Higher cis-regulated ASE level in tumor samples

For methods that do not use DNA-seq as a reference for bias adjustment, a high rate of CNV in tumor tissues (Shlien and Malkin, 2009) distorts the ASE detection results. For example, according to Mayba *et al.* (2014), the ASE genes identified by MBASED in tumor samples are mainly driven by large-scale genomic alterations, especially CNVs. However, the ASEs reported by cisASE should be unaffected by CNV. Applying cisASE to colon samples revealed that only 1% of the ASE genes in tumor samples are located within CNV regions (Fig. 7) and that the CNV genes are not significantly enriched in ASE genes (Fisher test P-value = 0.273). This result validates the effectiveness of cisASE in identifying *cis*-regulated ASEs.

By comparing tumor and normal samples, we found that tumor samples exhibited 1.8-fold-higher ASE rates than normal samples ($P = 9.961\text{e-}06$). This result indicates that the *cis*-regulation of gene expression is more extensive in tumors.

4.2 Recurrent germline ASE and ASE hotspots in normal and tumor samples

In the human colon dataset, we found that most ASE SNVs are sample-specific (Supplementary Fig. S7). Low-frequency ASEs in the population randomly distribute across chromosomes. However, some hotspots exist (Supplementary Fig. S8 and Table S6), e.g. HLA regions (Supplementary Table S7). MHC loci are some of the most genetically variable coding loci in mammals, including humans (Parham and Ohta, 1996). In normal tissues, polymorphisms in the HLA loci have been found to be associated with a large number of human phenotypic traits and common diseases (Shiina *et al.*, 2009). In tumors, HLA is frequently altered in comparison to the tissue from which it originates (Yip *et al.*, 2000). Our identification of ASE hotspots in HLA loci indicated that the HLA loci are enriched with regulatory polymorphisms, which contribute to phenotypic polymorphism and disease in humans.

4.3 Highly overexpressed somatic mutant alleles in tumor samples

Somatic ASEs, which are caused by linked somatic events in the regulatory regions of genes or the somatic mutation itself, may contribute to tumor progression. Because of the high genetic heterogeneity of tumors, matched DNA-seq data are required to identify somatic ASEs with high accuracy.

We observed that 75% of the somatic ASEs overexpressed their mutant alleles. The significantly up-regulated expression level of the mutant alleles (Fisher exact test P-value < 2.2e-16, odds ratio = 2.99) suggests the existence of widespread selective pressure to produce more mutant proteins in tumors. Of the 25% somatic ASEs, which underexpressed their mutant alleles, 73% lost the mutant allele expression completely. This is possibly because potential tumor suppressors are involved in suppressing the expression of these somatic mutations.

By mapping somatic mutations to genes, we found that genes with underexpressed mutant alleles are significantly enriched in focal adhesion and ECM-receptor interaction pathways (Supplementary Table S8). Abnormal behaviors of these two pathways have been reported to play important roles in cancer progression in many cancer types (Lu *et al.*, 2012; Nagano *et al.*, 2012; Stewart *et al.*, 2004), including colon cancer (Lascorz *et al.*, 2011; Peng *et al.*, 2015). Our ASE study reveals detailed information about how somatic mutations in these genes affect gene expression and identifies possible therapeutic targets.

5 Discussion

In this work, we present a novel flexible computational method, cisASE, to detect ASE at the SNV, exon and gene levels. cisASE makes full use of information from DNA-seq and RNA-seq data to achieve an unbiased estimation of ASE. By default, cisASE adopts a site-by-site adjustment for ASE identification by using the DNA allele ratio. This procedure reduces the artifacts from technical and mapping bias (Degner *et al.*, 2009) and CNVs (Mayba *et al.*, 2014), making cisASE efficient for detecting putative cis-regulated ASEs. In the absence of matched DNA-seq data, cisASE performed moderately well. Additionally, the implementation of a pseudo-phasing function in cisASE allows the incorporation of unphased SNV data into an individual exon, gene or isoform.

To evaluate the performance of cisASE for ASE detection, we performed extensive analyses based on both simulated and real data from multiple resources, i.e. human and mouse colon tumor cell

lines and a phased 1000 Genomes Project individual. cisASE outperformed the chi-square test in terms of accuracy, especially for genes with low sequencing depth. Additionally, cisASE exhibits greatly increased accuracy and computation speed compared with other previous methods based on complex models.

By applying cisASE to a panel of paired tumor-normal human colon samples, we identified HLA loci as the most significant 'hotspot' of germline ASEs in both tumor and normal samples. In tumor samples, somatic ASE showed widespread overexpression of mutant alleles, indicating oncogenic roles of the mutant genes. These results confirmed the ability of cisASE to detect tumor-associated ASEs.

Acknowledgements

The authors thank Mr. Jie Bi of Shanghai Institutes for Biological Sciences for useful advice on web development and Dr. Guoqing Zhang of Shanghai Center for Bioinformation Technology for managing the websites.

Funding

This work was supported by the National Basic Research Program of China (2011CB910204, 2011CB510102 and 2010CB529200), the National Key Scientific Instrument and Equipment Development Project of China (2012YQ03026108), the National Key Technology Support Program of China (2013BAI101B09) and the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA12000000).

Conflict of Interest: none declared.

References

Browning,S.R. and Browning,B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.

Castle,J.C. *et al.* (2014) Immunomic, genomic and transcriptomic characterization of CT26 colorectal carcinoma. *BMC Genomics*, **15**, 190.

Cibulskis,K. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.

Degner,J.F. *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.

Ge,B. *et al.* (2009) Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.*, **41**, 1216–1222.

Heap,G.A. *et al.* (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.*, **19**, 122–134.

Howie,B.N. *et al.* (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.

Huang,H.S. *et al.* (2012) Topoisomerase inhibitors unsilence the dormant allele of Ube3a in neurons. *Nature*, **481**, 185–189.

Lascorz,J. *et al.* (2011) Consensus pathways implicated in prognosis of colorectal cancer identified through systematic enrichment analysis of gene expression profiling studies. *PLoS One*, **6**, e18867.

Lee,R.D. *et al.* (2013) Large-scale profiling and identification of potential regulatory mechanisms for allelic gene expression in colorectal cancer cells. *Gene*, **512**, 16–22.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu,Z. *et al.* (2014) Comparing computational methods for identification of allele-specific expression based on next generation sequencing data. *Genet. Epidemiol.*, **38**, 591–598.

Lu,P. *et al.* (2012) The extracellular matrix: a dynamic niche in cancer progression. *J. Cell Biol.*, **196**, 395–406.

Mayba,O. *et al.* (2014) MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol.*, **15**, 405.

Montgomery,S.B. *et al.* (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773–777.

Nagano,M. *et al.* (2012) Turnover of focal adhesions and cancer cell migration. *Int. J. Cell Biol.*, **2012**, 310616.

Pandey,R.V. *et al.* (2013) Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Mol. Ecol. Resour.*, **13**, 740–745.

Parham,P. and Ohta,T. (1996) Population biology of antigen presentation by MHC class I molecules. *Science*, **272**, 67–74.

Pastinen,T. (2010) Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.*, **11**, 533–538.

Peng,G. *et al.* (2015) Transcriptome profiling of the cancer and adjacent non-tumor tissues from cervical squamous cell carcinoma patients by RNA sequencing. *Tumour Biol.*, **36**, 3309–3317.

Rozowsky,J. *et al.* (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.*, **7**, 522.

Seshagiri,S. *et al.* (2012) Recurrent R-spondin fusions in colon cancer. *Nature*, **488**, 660–664.

Shiina,T. *et al.* (2009) The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.*, **54**, 15–39.

Shlien,A. and Malkin,D. (2009) Copy number variations and cancer. *Genome Med.*, **1**, 62.

Skelly,D.A. *et al.* (2011) A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.*, **21**, 1728–1737.

Smith,R.M. (2013) Whole transcriptome RNA-Seq allelic expression in human brain. *BMC Genomics*, **14**, 571.

Stewart,D.A. *et al.* (2004) Changes in extracellular matrix (ECM) and ECM-associated proteins in the metastatic progression of prostate cancer. *Reprod. Biol. Endocrinol.*, **2**, 2.

Tuch,B.B. *et al.* (2010) Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS One*, **5**, e9317.

Yip,D. *et al.* (2000) Immunomodulation therapy in colorectal carcinoma. *Cancer Treat. Rev.*, **26**, 169–190.

Zhang,K. *et al.* (2009) Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods*, **6**, 613–618.