ORIGINAL RESEARCH
published: 27 November 2018
doi: 10.3389/fgene.2018.00570

A Genome-Wide Study of Allele-Specific Expression in Colorectal Cancer

Zhi Liu[1]*, Xiao Dong*[2]** and Yixue Li*[3,4,5]***

[1] Department of Epidemiology and Biostatistics, Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, School of Public Health, Nanjing Medical University, Nanjing, China, [2] Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, United States, [3] Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [4] Shanghai Center for Bioinformation Technology, Shanghai Industrial Technology Institute, Shanghai, China, [5] Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai, China

OPEN ACCESS

Edited by:
Chuan Lu,
Aberystwyth University,
United Kingdom

Reviewed by:
Shihao Shen,
University of California, Los Angeles,
United States
Younghee Lee,
University of Utah, United States
Yanfeng Zhang,
HudsonAlpha Institute
for Biotechnology, United States

***Correspondence:**
Xiao Dong
biosinodx@gmail.com
Yixue Li
yxli@sibs.ac.cn

Specialty section:
This article was submitted to
Bioinformatics and Computational
Biology,
a section of the journal
Frontiers in Genetics

Received: *19 July 2018*
Accepted: *06 November 2018*
Published: *27 November 2018*

Citation:
Liu Z, Dong X and Li Y (2018) A Genome-Wide Study of Allele-Specific Expression in Colorectal Cancer.
Front. Genet. 9:570.
doi: 10.3389/fgene.2018.00570

Accumulating evidence from small-scale studies has suggested that allele-specific expression (ASE) plays an important role in tumor initiation and progression. However, little is known about genome-wide ASE in tumors. In this study, we conducted a comprehensive analysis of ASE in individuals with colorectal cancer (CRC) on a genome-wide scale. We identified 5.4 thousand genome-wide ASEs of single nucleotide variations (SNVs) from tumor and normal tissues of 59 individuals with CRC. We observed an increased ASE level in tumor samples and the ASEs enriched as hotspots on the genome. Around 63% of the genes located there were previously reported to contain complex regulatory elements, e.g., human leukocyte antigen (HLA), or were implicated in tumor progression. Focussing on the allelic expression of somatic mutations, we found that 37.5% of them exhibited ASE, and genes harboring such somatic mutations, were enriched in important pathways implicated in cancers. In addition, by comparing the expected and observed ASE events in tumor samples, we identified 50 tumor specific ASEs which possibly contributed to the somatic events in the regulatory regions of the genes and significantly enriched known cancer driver genes. By analyzing CRC ASEs from several perspectives, we provided a systematic understanding of how ASE is implicated in both tumor and normal tissues and will be of critical value in guiding ASE studies in cancer.

Keywords: allele-specific expression, colorectal cancer, *cis*-regulatory variation, somatic mutation, tumor

BACKGROUND

Allele-specific expression (ASE) refers to the phenomenon that occurs in diploid or polypoid genomes, where two or more alleles of a gene has an imbalanced expression (Kwaepila et al., 2006; Ge et al., 2009; Heap et al., 2010; Tung et al., 2011). It is common in both humans (Lo et al., 2003) and other organisms (Tung et al., 2011; Graze et al., 2012; Hasin-Brumshtein et al., 2014), and potentially contributes to multiple phenotypes and complex traits (Frazer et al., 2009). Because of the intrinsic power of using two alleles of a gene in the same individual, as controls to reduce the background genetic and environmental effects, ASE is also an accurate and sensitive marker for *cis*-regulatory variation (Pastinen, 2010). For example, an ASE can indicate a heterozygous variant within the translated region, resulting in a modified or truncated protein (Kukurba et al., 2014); at a regulatory site, it can cause differential binding of transcription factors or epigenetic modifiers (Prendergast et al., 2012; Reddy et al., 2012); or at a splice site or UTR, it can affect transcript processing (Li et al., 2012).

Allele-specific expressions are also frequently observed in tumors (Valle et al., 2008; Curia et al., 2012; Walker et al., 2012; Wei et al., 2013). ASE was first proposed as a direct approach for connecting a genotype to disease susceptibility in 2002 (Yan et al., 2002). In 2013 it was discovered that ASE, at the death-associated protein kinase 1 (*DAPK1*) gene locus, was potentially predisposed to chronic lymphocytic leukemia (CLL) using a single-nucleotide primer extension (SNuPE) and MALDI-TOF mass spectrometry (Wei et al., 2013). In colorectal cancer (CRC), a decrease in expression of one adenomatous polyposis coli tumor suppressor (*APC*) gene allele, leads to the development of familial adenomatous polyposis (Curia et al., 2012). In addition to *APC*, ASE of transforming growth factor beta receptor 1 (*TGFBR1*), which leads to reduced expression of the gene, can also cause an increased risk of CRC (Valle et al., 2008). In addition to the association with cancer risk, ASE also affects the prognosis and outcome of cancer patients. For example, the monoallelic expression of TP53 and IDH1 was found to determine the oncogenic progression and survival in brain tumors (Walker et al., 2012).

With the development of large-scale transcriptome sequencing, the systematic analysis of the ASE in the transcriptome was achieved at the single nucleotide resolution (Tuch et al., 2010; Smith et al., 2013). To date, several studies have reported genome-wide ASE, in human, mice and cell lines, and identified hundreds of genes exhibiting ASE (Heap et al., 2010; Smith et al., 2013; Hasin-Brumshtein et al., 2014). However, little is known about genome-wide patterns of ASE in tumor tissues. In this study, we carried out an ASE study in a cohort of 59 patients with CRC (Seshagiri et al., 2012) and revealed the comprehensive landscape of ASE in CRC patients.

MATERIALS AND METHODS

Data Preprocessing

RNA and Exon sequencing data of matched human colorectal tumor-normal samples was downloaded from the European Genome-Phenome Archive (EGA) under accession number EGAS00001000288 (Seshagiri et al., 2012). Fifty-nine pairs of samples correctly processed were retained for ASE analysis.

Quality controlled DNA and RNA sequencing data was mapped with bowtie2 (Langmead and Salzberg, 2012) with default parameters to report the best alignment. The base qualities were then recalibrated using the procedure recommended by GATK (DePristo et al., 2011).

Somatic mutations were called with both Mutect (Cibulskis et al., 2013) and Varscan (Koboldt et al., 2009), and the intersection was considered a reliable result and used for the following analysis. Germline SNVs were called using the GATK best practices from DNA sequencing data, and filtered using the flowing four criteria to obtain a final SNV list ready for ASE analysis.

(1) SNVs cluster together;
(2) SNVs covered by less than 20 reads;
(3) SNVs located within repeated regions;
(4) SNVs located within non-coding regions;
(5) SNVs were identified as a somatic mutation in exon sequencing data.

Allele counts for each germline SNV and the somatic mutation in DNA and RNA sequencing data, were generated with SAMtools (Li et al., 2009) in a pileup format.

The list of germline SNV and somatic mutation, as well as the corresponding pileup files were subjected to cisASE for ASE identification, respectively.

ASE Identification

Allele-specific expression SNVs and genes were identified by the cisASE pipeline (Liu et al., 2016). SNVs with a coverage of less than 10 in RNA or DNA sequencing data were filtered. SNVs or genes with a log likelihood ratio (LLR) value more than the LLR cutoff, at a significance level of 0.01, were defined as ASE. In addition, genes with a heterogeneity *p*-value less than 0.05, which indicates inconsistent ASE status of SNVs within the gene, were excluded from further analysis.

Identifying ASE Hotspots

Allele-specific expression counts and frequency was calculated in consecutive sliding windows with fixed sizes along the genome. We randomly assigned ASE labels to the SNVs across chromosomes, according to the total ASE frequency. By repeating this process 1000 times, we obtained a null distribution of ASE density in each of the sliding windows. A *p*-value was derived by counting the number of times the number of ASEs in the window after perturbation, exceeded the observed ASE, and adjusted it with an add-one smoothing. These *p*-values were then corrected for multiple tests using the Benjamini-Hochberg method.

Group of ASE Somatic Mutation

We mapped the ASE somatic mutations to genes and then classified the genes into two categories, i.e., genes with over-expressed mutant alleles and genes with under-expressed mutant alleles. Genes harboring multiple somatic mutations with conflicting mutant allele expression, were excluded from the following analysis. Gene expression profiles were generated with tophat2 (Kim D. et al., 2013) and cufflinks (Trapnell et al., 2012) software. For genes in each group, we compared the FPKM value of both tumor and normal tissues of patients with the somatic mutation, and defined the FPKM fold change of 2 and 1/2 as the threshold of up-regulated and down-regulated expression in tumor samples. This resulted in three groups for each category according to the gene expression.

Identifying Somatic ASE Genes

We counted the number of ASE somatic events (s_i) and the number of total tested pairs (t_i) in a population of 118 individuals, for each gene. We refer to the ASE somatic event (s_i) as the gene showing ASE in a tumor sample but not in the matched normal sample. In addition, we refer to the tested pairs (t_i) as the sample pairs, and the gene is tested in both matched tumor-normal samples. The expected ASE somatic event rate was calculated by the following equation,

$$f = \sum_{i=1}^{n} S_{\mathrm{i}} \Big/ \sum_{i=1}^{n} t_{\mathrm{i}},$$

where n is the number of genes.

The expected number of the ASE somatic event for each gene, was calculated as the product of the total tested pairs and the expected ASE somatic event rate (f^*t_i). A p-value was obtained for each gene using the Poisson distribution and the observed and expected number of ASE somatic events ($P[X \geq x]$). These p-values were corrected for multiple testing using the Benjamini-Hochberg method and genes that had a corrected p-value <0.05 were called a somatic ASE gene.

RESULTS

Increased ASE Level in Tumor Samples

We identified SNV and gene level ASEs in both normal and tumor tissues of 59 CRC patients with our previously developed pipeline for ASE identification (Kukurba et al., 2014; **Figure 1** and **Supplementary Table S1**). The major steps included sequence alignment, variant calling, ASE detection using cisASE (Kukurba et al., 2014), and further filtration (see section "Materials and Methods" for details). We detected 431 (*SD* = 133.3) SNV-level ASEs per normal tissue and 477 (*SD* = 181.6) per tumor tissue, and 137 (*SD* = 39.3) and 216 (*SD* = 108.5) gene-level ASEs per normal and tumor tissue, respectively (**Supplementary Table S1**). The frequency of ASE SNVs (a ratio of number of ASE SNVs to number of non-ASE SNVs) in normal tissue is in agreement with previous studies (Zhang et al., 2009; Skelly et al., 2011).

We compared the portion of sites exhibiting an ASE in tumors with its matched normal tissues. On average, 20.0% of the SNVs in tumor samples and 16.8% in normal samples exhibited an ASE (two-tailed paired t-test, p-value = 7.1e−09), indicating a significantly higher ASE level in tumor samples than in normal samples. When only testing the SNVs identified in tumor and normal tissues, the results were the same, i.e., a significantly higher portion of the ASE in tumor samples (21.6%) than in the normal samples (18.1%; two-tailed paired t-test, p-value = 1.2e−04; **Figure 2A**).

For each tumor-normal pair, we found that 68% of the ASEs are either normal (29%) or tumor (39%) specific (**Figure 2B**). And the remaining 32% of the ASEs are shared by both the tumor and normal samples (**Figure 2B**), most of which had the same ASE direction. This indicates that the majority of ASEs (about 2/3) are dynamic in tumorigenesis while the other 1/3 ASEs are consistent.

Next, we identified genes with recurrent ASE events in tumor and normal samples. There were 94 and 95 genes with ASE events in at least 20% of the tumor and normal samples, respectively, of which 63 genes were shared by both tumor and normal samples (common ASE genes) (**Supplementary Table S2** and **Figure 3**). The allele ratio of recurrent of ASE genes was significantly segregated from the background and the total pool of ASE genes (**Supplementary Figure S1**) and the average major haplotype allele ratio of common ASE genes was 0.92.

The ASE genes that was mostly recurrently observed in both tumor and normal samples had a high allele imbalance, such as AP3P1, BCLAF1, STED8, PRIM2, IL32, SEC22, and MAP2K3 (**Figure 3A**). The recurrent ASE genes in tumor samples include Chromosome-Associated Kinesin KIF4B, spindle and kinetochore associated complex subunit 3 (SKA3) and so on. We also found that the ASE of TP53 was specifically and recurrently observed in tumor samples (observed in 12 tumor samples and 1 normal sample). There were 32 recurrent ASE genes observed in normal samples. For example, PYY, CD177, PEG3, and FAM83D, were observed in more than 11 normal samples, while less than 3 were observed in tumor samples.

The ASE Hotspots in the Normal and Tumor Genome

Variants on the *cis*-regulatory element on the genome, tend to affect the expression of one or more genes nearby (Pastinen, 2010), and if the variation is heterozygous, the genes regulated by it will exhibit an ASE, therefore we prioritized the existence of such variations by identifying clusters of the ASEs on genomic regions. We calculated the ASE density and frequency in the tumor and normal samples, by using a sliding window approach with a window size of 100k base pair (bp) and a step size of 10 kbp. Windows with an adjusted p-value <0.05 were kept, and overlapping windows were manually checked and merged, to get focal hotspot regions.

We identified 32 ASE hotspots in normal samples (**Supplementary Table S3**) and 27 in tumor samples (**Supplementary Table S4**), affecting a total of 57 genes (**Supplementary Figure S2**), which resulted in 4.0% (723 out of 17866) and 4.4% (748 out of 17866) of the ASE SNVs identified in normal and tumor samples, respectively. There were 21 genes located within hotspots identified in both normal and tumor samples, as well as 22 and 14 genes located within the hotspots specific to tumor and normal samples, since the tumor or normal differential expression might result in a different power of ASE detection. We checked the expression of all these genes in tumor and normal samples (**Supplementary Table S5**), and found no difference of the tumor and normal FPKM ratio among the three groups of genes (Kruskal–Wallis test p-value = 0.07), indicating the difference of the ASE hotspots in tumor and normal samples did not result from the different detection powers, due to the expression difference. In addition, one gene with relatively low expression (PRSS1, FPKM < 0.1) was excluded (**Figure 4**).

To investigate the biological process affected by the ASE, we conducted the GO and KEGG enrichment analyses for the ASE genes. The ASE genes shared by tumor and normal tissues were

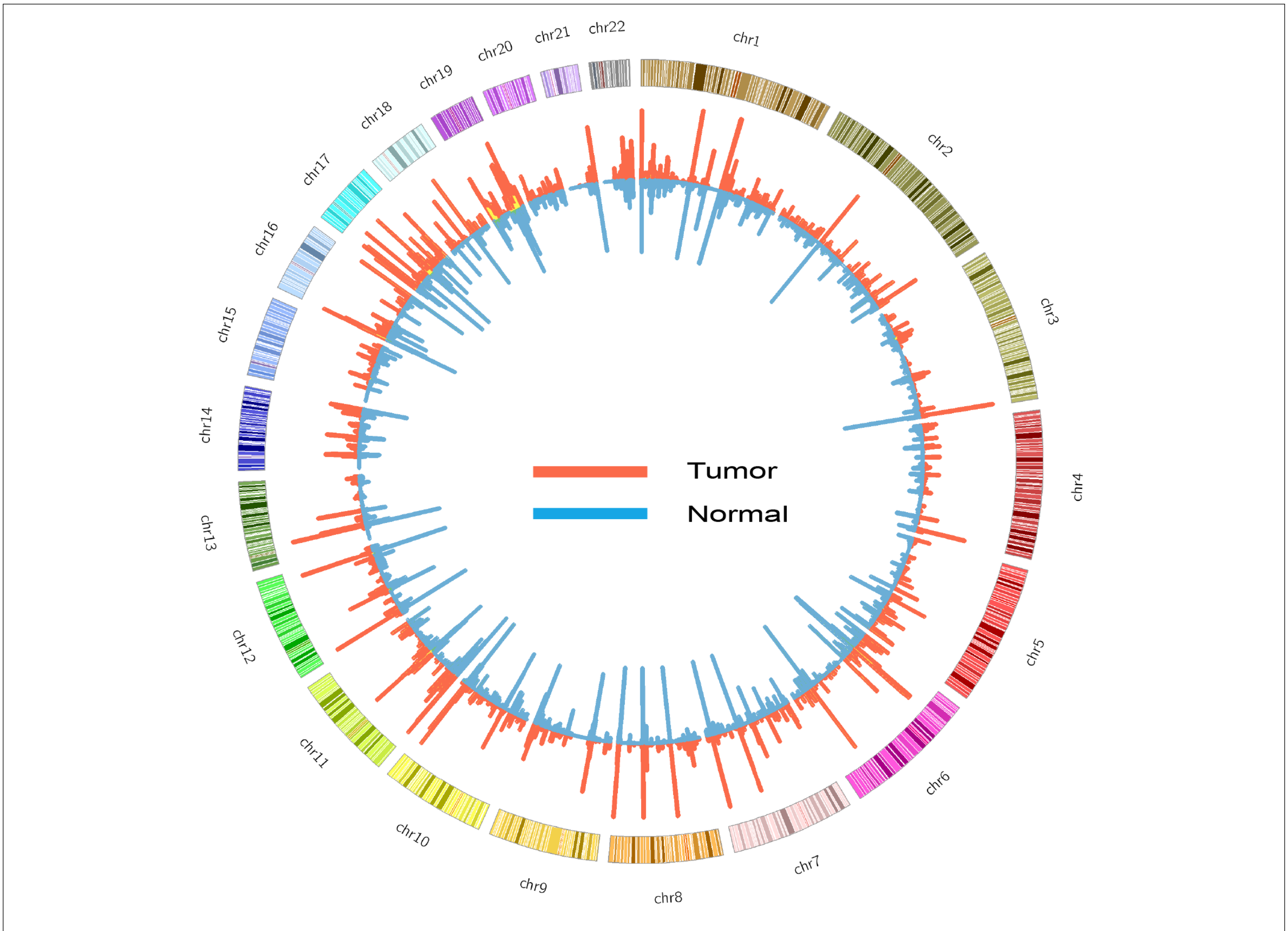

FIGURE 1 | Circos diagram of ASE detected in tumor (red) and normal (blue) samples. The height of histogram indicates the counts of ASE in a window of 1 million base pair.

significantly enriched in antigen processing. The significantly enriched GO and KEGG terms of tumor specific ASE genes were closely associated with immune activity. However, the normal ASE genes were not enriched in specific functions. The results impose the possibility that the ASE plays a role in maintaining regular immune activities, and an excessive ASE event was activated in tumor tissues.

Among the genes shared in normal and tumor tissues, several genes involved in polymorphism, or which were reported to be related to cancer predisposition or progression were included, such as, the human leukocyte antigen (HLA) on chromosome 6, members of the mucin gene family (MUC) on chromosome 3, and the MAP2K3 and CDC27 locus, which is involved in the cell proliferation and cell cycle on chromosome 7. Three members of the carcinoembryonic antigen (CEA) family (CEACAM5, CEACAM6, and CEACAM7) were also included.

Though 35.6% (21 out of 59) of the ASE genes were common in both normal and cancer tissues, it was reported that the change in the allele ratio of the ASE can also lead to phenotypic diversity, for example, a study reported that the proportion of the JAK2 V617F mutant allele in RNA levels is significantly associated with distinct subtypes of BCR/ABL-negative myeloproliferative neoplasms (MPNs) (Kim H.R. et al., 2013). Therefore, we tested whether there are similar cases in ASE genes between the tumor and normal tissues. We found that four out of the 21 shared ASE genes (HLA-A, HLA-B, HLA-C, and CEACAM7) and showed significant differences in the allele ratios between tumor and normal tissues (paired *t*-test; **Supplementary Table S6**). Three of the four genes belong to the HLA family, i.e., HLA-A and HLA-B, and HLA-C, and all revealed a lower allelic heterozygosity in tumor tissues (**Figure 5**). Loss of heterozygosity (LOH) of the HLA loci was reported in many cancers (Maleno et al., 2002; Wang et al., 2006; Zollikofer et al., 2014). In our case, these loci are heterozygous at the DNA level, while at the mRNA level, one of the copies showed a significantly reduced expression compared to the other one. The results suggest the possibility that in tumor tissues, the allele-specific regulation on the transcriptional level may lead to a similar consequence as the LOH.

The other 18 shared ASE genes, showed no difference in the allele ratio, between normal and tumor tissues (**Supplementary Figure S3**), indicating that most of the shared ASEs are conserved during tumorigenesis. However, because the normal tissues we

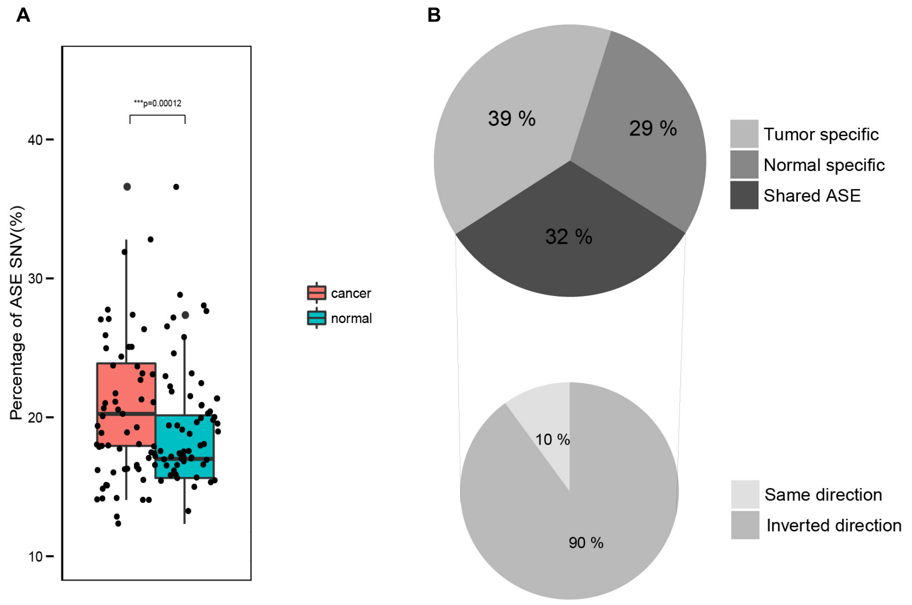

FIGURE 2 | Comparison of ASE in normal and tumor tissues. **(A)** The fraction of ASE in normal and tumor samples. **(B)** Intersection of ASE identified in paired normal and tumor samples.

studied were obtained from CRC patients, ASE genes shared by tumor and normal tissues can be involved either in normal physiological functions or associated with tumor predisposition. Since it is hard to obtain gut tissue samples from healthy people, we cannot distinguish these two possibilities.

Of the twenty-two genes located in the tumor-specific hotspots (**Supplementary Table S4**), several were reported to play an important role in tumor progression. For example, over-expression of the FAT1 was observed in different tumors including in DCIS breast cancer (Kwaepila et al., 2006), melanoma (Sadeqzadeh et al., 2011) and leukemia (de Bock et al., 2012). MKI67 is a protein that is widely used as a marker for cell proliferation, and its increased expression in human cancer specimens generally denotes an aggressive phenotype (Cidado et al., 2016). The observed allele specific expression of these phenotypes may help to prioritize the underlying mechanisms which contribute to the abnormal expression in tumors. Furthermore, 14 genes (ACSF3, AHNAK, APOBR, CCBL2, CLN3, EPPK1, FAM104B, FUT2, HLA-DRA, HLA-G, MUC12, NBPF1, RASIP1, RBMXL1, and SLC25A5) were located in the normal-specific hotspots (**Supplementary Table S3**), which suggests that precise control of the ASE may be important for maintaining the normal function of cells. These results might provide opportunities for mining new therapy targets.

Overexpressed Allele With Somatic Mutations in Tumors

Somatic mutations (missense mutations and non-sense mutations) within the coding region may lead to abnormal protein products. However, the impact of a heterozygous coding SNV depends on whether the SNV-containing allele is transcribed to the RNA. In addition, clinical therapy-selection for targeted drugs, often assay mutations using DNA as an analyte, such as KRAS assays designed to identify responders to anti-EGFR therapy (Allegra et al., 2009). However, if the mutant allele is selectively lost in the transcript, the mutation may not have a therapeutic impact and the merit of using a DNA-based assay for clinical decision-making may be problematic. The above are the major reasons for us to further analyze the allelic expression of somatic mutations in tumors. A genome-wide study in mouse tumor cell lines reported that mutations are transcribed in proportion to their DNA allele frequency (Castle et al., 2014). However, to our knowledge, a genome-wide study of the relationship between DNA and RNA mutation allele frequency in tumor tissues, has not been done.

We found that 37.5% of the 2,754 somatic mutations exhibit an ASE in the colorectal tissues (**Figure 6A**), which is more than two times higher than that for germline polymorphisms (18%) (**Figure 6B**). This indicates a significant imbalance of the allelic expression for somatic mutations. Furthermore, 78% of the ASE somatic mutations over-expressed mutant alleles, comparing to a proportion of 52% for germline polymorphisms (chi-square test p-value $<2.2e{-}16$). The results reveal that gene copies with somatic mutations have prevailing expression superiority compared to the wild type ones in tumor tissues.

Next, we explored the functional significance of the ASE somatic mutations with a different mutant/wild-type allele expression pattern. We mapped the ASEs to genes and classified them into six groups according to the alteration of both mutant allele expression and total gene expression in tumor tissues (**Figure 7** and **Supplementary Table S7**; Materials and Methods).

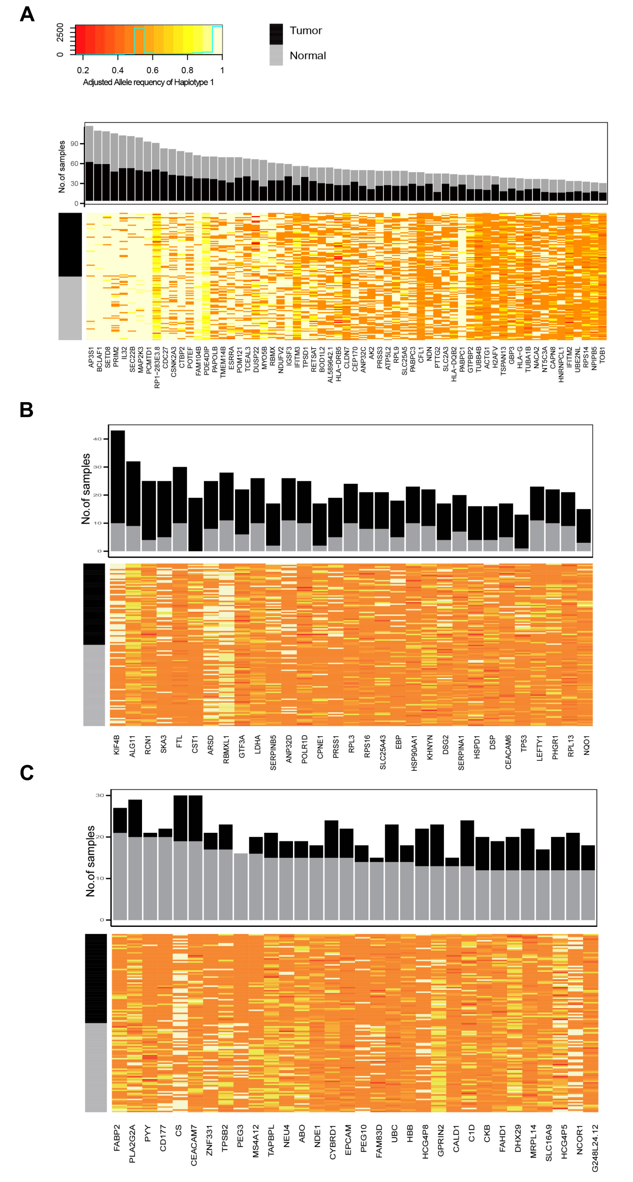

FIGURE 3 | The frequency of recurrently observed ASE genes and heatmap of their allele ratio among samples. **(A)** Genes shared by tumor and normal samples, **(B)** ASE genes recurrently observed in tumor samples; **(C)** ASE genes recurrently observed in normal samples.

Ideally, if an ASE somatic mutation is functional, the direction of the mutant allele expression change should be the same as the direction of the gene expression change in tumor cells, compared with normal cells (Group a and f in **Figure 7**). Genes which exhibited the ASE somatic mutation but an unchanged total gene expression (Group b and e in **Figure 7**) might be regulated by other trans-regulatory factors, and the effects of the ASE were buffered. Those conflicting with the somatic allele expression and tumor gene expression (Group c and d in **Figure 7**) were possible artificial results, or the ASE was a random event without functional significance.

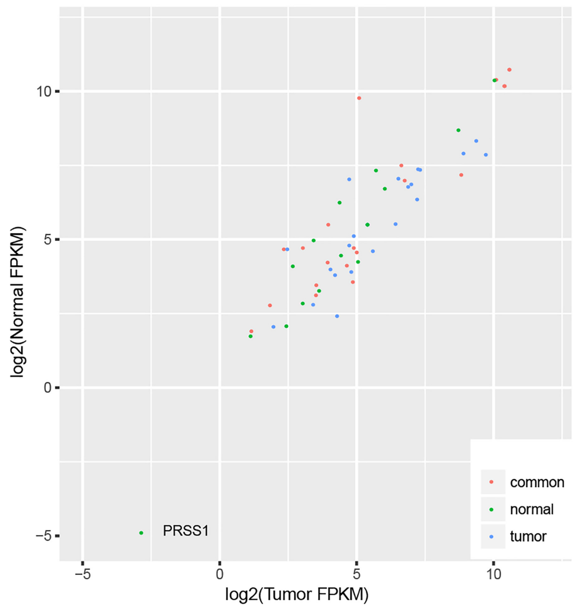

FIGURE 4 | Gene expression in tumor and normal samples for genes located in ASE hotspots.

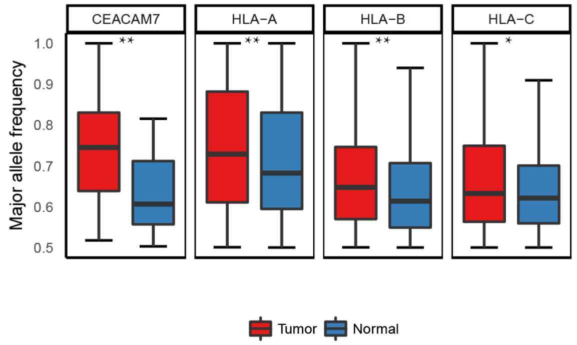

FIGURE 5 | Allele ratio of the four shared ASE genes which exhibited differential allele ratio in tumor and normal samples. *$p < 0.05$, **$p < 0.01$.

As expected, only the genes in Groups a were farely significantly enriched in KEGG pathways (Du et al., 2014; **Table 1**). Group a, which contain genes over-expressing mutant allele and showing an up-regulated gene expression level in tumor samples compared with the matched normal sample, is enriched in the DNA replication and mismatch repair pathways. Dysfunction of the DNA replication and DNA mismatch repair pathways are implicated in many cancer types (Boyer et al., 2016; Puigvert et al., 2016), which initiates cancer or promotes cancer cell proliferation (Padmanabhan et al., 2004; Dudderidge et al., 2007). The average mutant allele fraction for the genes enriched in these two pathways is 80%, indicating a widely over-expressed mutant allele. This suggests that in tumor tissues, genes involved in the DNA replication and DNA mismatch repair pathways, tend to selectively express mutant proteins with abnormal functions,

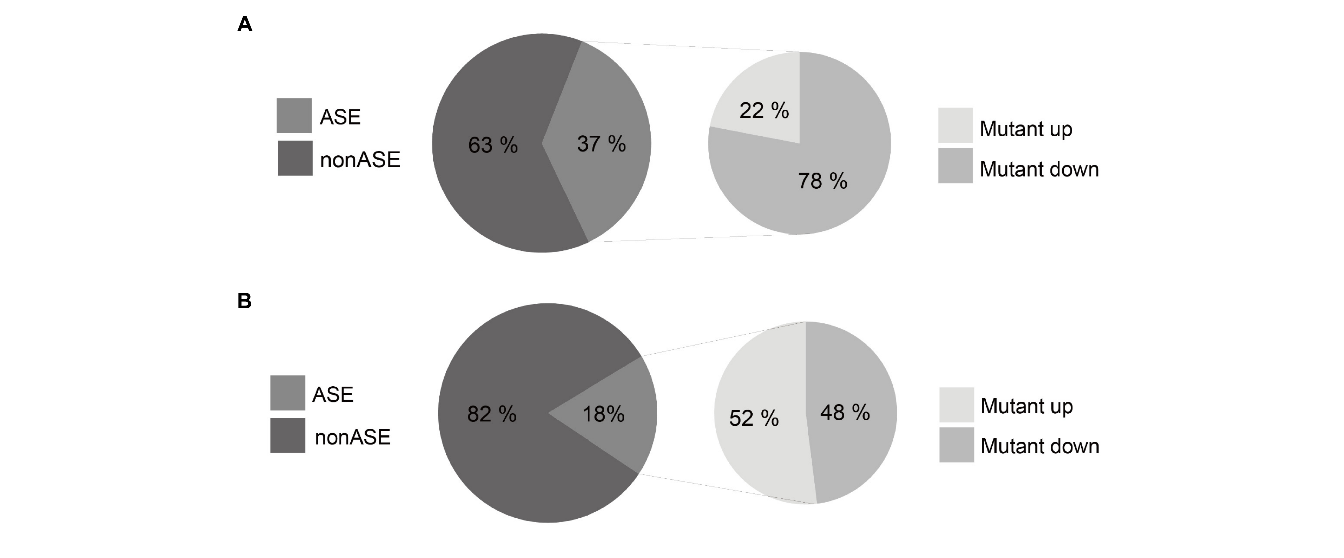

FIGURE 6 | Statistics of ASE for somatic mutations **(A)** and germline polymorphisms **(B)**.

which may compete with normal proteins to disrupt normal signal pathways, or decrease the dosage of normal proteins for normal functions.

Genes in Group f, which contain genes with under-expressed mutant alleles and down-regulated gene expression in tumor samples, compared with matched normal ones, are enriched in the focal adhesion signal pathway. The genes enriched in the focal adhesion pathway showed limited mutant allele fractions only 10% of the two alleles, suggesting that mutation-containing alleles are effectively silenced by epigenetic and chromatin modification mechanisms (Jaenisch and Bird, 2003) or mutation-containing transcripts are degraded by activating RNA surveillance mechanisms (Rehwinkel et al., 2006), resulting in an overall decrease of gene expression levels and thus an abnormal signal pathway.

Somatic ASE Genes Are Enriched in Known Cancer-Related Genes

Genes specifically exhibiting the ASE in cancer tissues are likely linked to somatic variations in regulatory regions. In order to detect genes with an excess of somatic *cis*-regulatory events, we used matched tumor and normal samples to identify genes specifically and significantly exhibiting ASE in tumor samples (which we defined as the "somatic ASE gene"). We found 50 somatic ASE genes (**Supplementary Table S8**), which significantly enriched TCGA pan-cancer drivers (Gonzalez-Perez et al., 2013) (five pan-cancer drivers p-value = 0.010) and CRC drivers (Gonzalez-Perez et al., 2013) (two CRC drivers p-value = 0.04), indicating that the tumor specific ASE genes analysis catches known cancer genes, and has the potential to be a complementary method for driver detection. Next we compared the somatic ASE gene with differential expressed genes (DEG) between tumor and normal samples (**Supplementary Table S9**), and found that they significantly enriched in DEGs (fisher exact test p-value = 5.0e−07, odds ratio = 3.22).

DISCUSSION

The ASE is a measure of the effect of genetic variants on gene expression, that does not require any assumption on the genetic structure of the population studied, and hence a direct measurement of how gene-expression changes at the individual level (Yan et al., 2002). The development of next generation sequencing technologies and our unbiased computation method cisASE (Kukurba et al., 2014) have enabled us to characterize this genome-wide landscape of the ASE in tumor and normal tissues of CRC patients from diverse perspectives.

The higher incidence of the ASE in tumor samples than that of normal samples is consistent with the fact that gene expression in tumor cells is under more complicated *cis*-regulation (Maurano et al., 2012). Furthermore, 29 and 39% of the ASE SNVs were specific to either normal or cancer samples, respectively, indicating both the gain and loss of *cis*-regulatory variation as possible contributors to tumor initiation or development. We also observed a high percentage (32%) of ASEs shared by normal and tumors tissues of patients, which might be a mixture of CRC preposition sites, as well as sites where ASE play a role in maintaining regular cellular function. Since it is difficult to obtain gut tissue samples from healthy people to distinguish these two categories of ASE, some researchers suggest using blood samples from normal healthy people (Valle et al., 2008). However, *cis*-regulatory variation is a tissue dependent feature, so is the ASE (GTEx Consortium, 2015), therefore, using a different tissue as control might result in high false discovery rates. Creative and accurate methods are needed to further explore cancer risk sites from regular sites.

By summarizing the ASE in a region-based fashion, we identified the ASE hotspots under true and recurrent *cis*-regulation in the studies samples. Although the majority of the ASE hotspots, including the HLA loci, were shared by both normal and tumor tissues, four of the HLA genes revealed a significant lower heterozygosity in the tumor tissues

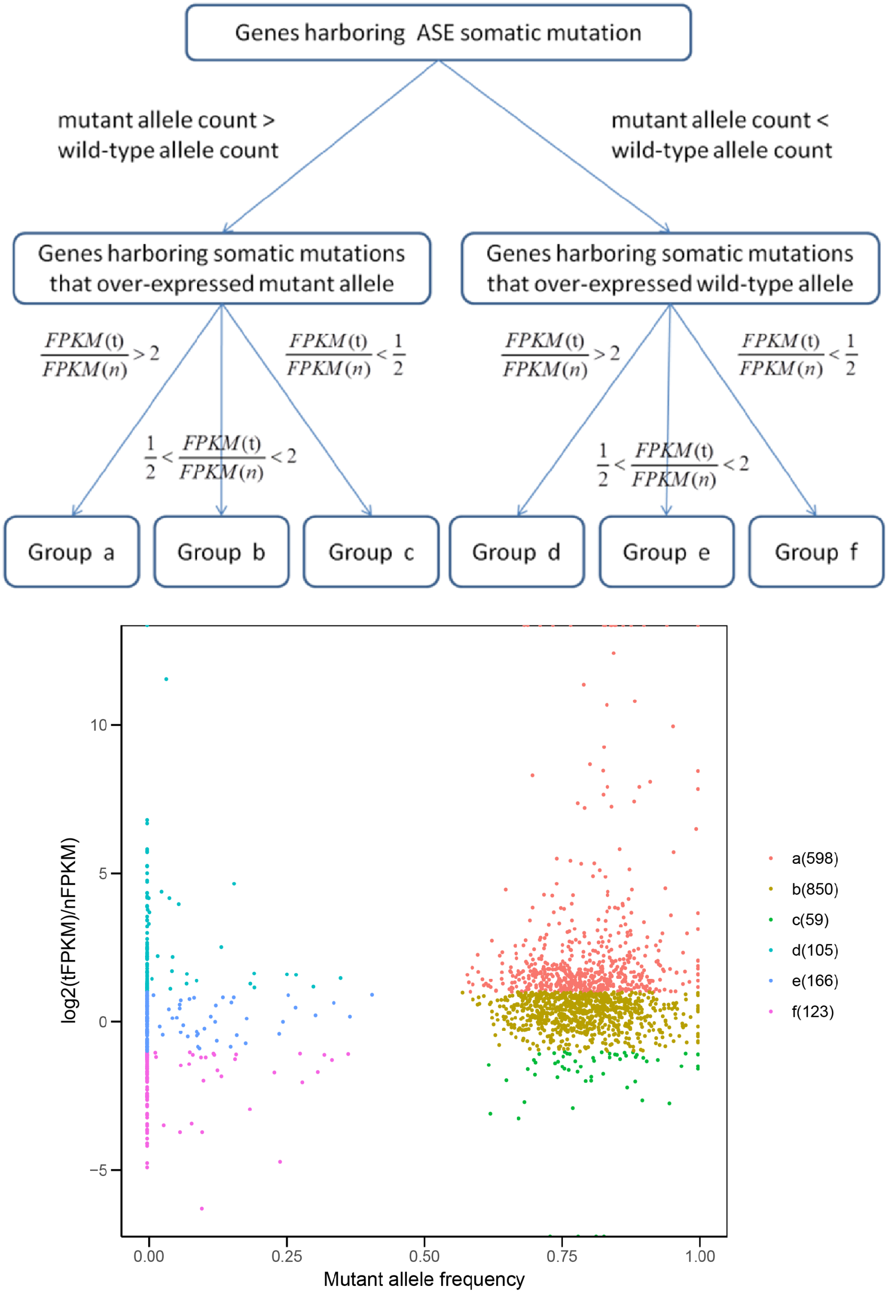

FIGURE 7 | Groups of genes harboring ASE somatic mutations. FPKM (t) and FPKM (n) represent the FPKM value of gene in tumor and its matched normal tissue, respectively.

TABLE 1 | Enriched KEGG pathways for genes in group a and f.

Type a			
Term	ID	Adj.pvalue	Genes
DNA replication	hsa03030	0.019631059	RFC3, MCM7, RFC1, POLD2, MCM3, RNASEH2A
Mismatch repair	hsa03430	0.026586385	EXO1, RFC3, RFC1, POLD2, MLH3
Type f			
Term	ID	Adj.pvalue	Genes
Focal adhesion	hsa04510	0.017378214	TLN1, TNC, COL6A3, ZYX, THBS1, FLNA, MYLK

compared with normal tissues. The LOH in the short arm of chromosome 6 is the most frequent mechanism contributing to the HLA haplotype loss in human cancer, which is a tumor escape mechanisms from the host's immune surveillance system (So et al., 2005). The selective expression of one allele of the HLA gene might be another mechanism that contributes to the HLA haplotype loss in cancer.

One category of targeted drugs, is the targeting of specific genes with or without certain somatic mutations, such as osimertinib targeting at EGFR (with EGFR T790M mutation) and afatinib targeting at EGFR(with EGFRL858R mutations) in non-small cell lung cancer, vemurafenib targeting at BRCA(with BRAF V600 mutation) in melanoma, and panitumumab targeting at EGFR (with KRAS will type) in CRC. A DNA assay is usually used to test whether a specific gene mutation codes the target. However, an RNA level expression is not necessary a faithful replication of the DNA. We found that 38% of the somatic ASE exhibited the ASE, indicating that the DNA-assay based therapy-selection might be problematic. Somatic mutations and mutant allele that followed the same direction as the total gene expression, i.e., Group a and f, were enriched in important signal pathways involved in tumor initiation and progression. However, mutations belonging to other groups may also have biological implications, are not significantly enriched in the KEGG pathways, since we cannot exclude the possibility that, in some cases, homeostatic or feedback mechanisms act to constrain the total expression so that an imbalance in allelic expression does not change the total output.

Somatic ASE genes were regulated by *cis*-regulatory elements with somatic variations, which may be the driver mutation implicated in cancers, the fact that the identified somatic ASE genes enriched pan-cancer and CRC driver genes, support this speculation.

In this study, we focused on the ASE of protein coding regions. However, in recent years, lncRNAs were reported to be involved in gene regulation and other cellular processes (Quinn and Chang, 2016). With an ASE analysis, Almlof et al. (2014) found that 22.9% (258 out of 1122) of intergenic lncRNAs were regulated by *cis*-rSNP in human primary monocytes, which is comparable to our analysis. Though the number of lncRNAs exceeded the protein coding genes, because of a much lower expression (Iyer et al., 2015), a higher sequencing depth and more sensitive detector is required to quantify ASE in lncRNAs more efficiently.

CONCLUSION

By applying the ASE studies in CRC patients, we found a higher incidence of the ASE in tumor tissues, which implicated more complicated *cis*-regulation in tumors. ASEs under recurrent *cis*-regulation were enriched as hotspots on the genome and the majority of the genes (∼63%) involved in the hotspots, were previously reported to have complex regulatory elements, or were implicated in tumor progression. In addition, the ASE analysis of somatic mutation revealed a significant increased ASE rate for somatic mutations, and genes harboring such somatic mutations were enriched in important pathways implicated in CRC (DNA replication and focal adhesion). Furthermore, the somatic ASE genes analysis catches known cancer genes.

In summary, this study provides a systematic understanding of how the ASE is implicated in tumors and a schema of the application of the ASE studies in patients with cancerous tumors.

DATA AVAILABILITY

The datasets supporting the conclusions of this article are included within the article and its additional files. Raw RNA and Exon sequencing data were downloaded from the European Genome-Phenome Archive (EGA) under accession number EGAS00001000288 by proper application.

AUTHOR CONTRIBUTIONS

ZL and XD conceived the study and wrote the manuscript. ZL carried out all the analysis in this study. YL supervised the study and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Key R&D Program of China (2016YFC0901704, 2017YFA0505500, 2017YFC0907505, and 2017YFC0908405).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00570/full#supplementary-material

FIGURE S1 | The allele ratio of recurrent of ASE genes was significantly segregated from the background and the total pool of ASE genes.

FIGURE S2 | ASE frequency across chromosomes.

FIGURE S3 | The comparison of allele ration of 18 shared ASE genes between normal and tumor tissues.

TABLE S1 | SNV and gene levels of ASE in the normal and tumor tissues.

TABLE S2 | Genes with recurrent ASE events in tumor and normal samples.

TABLE S3 | ASE Hotspots in normal.

TABLE S4 | ASE Hotspots in Cancer.

TABLE S5 | Expression for genes in hotspots region.

TABLE S6 | Differences in allele ratios between tumor and normal tissues.

TABLE S7 | Allele ratio and FPKM for somatic ASE and involved genes.

TABLE S8 | Somatic ASE genes.

TABLE S9 | Gene expression of somatic ASE genes in tumor and normal samples.

REFERENCES

Allegra, C. J., Jessup, J. M., Somerfield, M. R., Hamilton, S. R., Hammond, E. H., Hayes, D. F., et al. (2009). American society of clinical oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. *J. Clin. Oncol.* 27, 2091–2096. doi: 10.1200/Jco.2009.21.9170

Almlof, J. C., Lundmark, P., Lundmark, A., Ge, B., Pastinen, T., Cardiogenics Consortium, et al. (2014). Single nucleotide polymorphisms with cis-regulatory effects on long non-coding transcripts in human primary monocytes. *PLoS One* 9:e102612. doi: 10.1371/journal.pone.0102612

Boyer, A. S., Walter, D., and Sørensen, C. S. (2016). DNA replication and cancer: from dysfunctional replication origin activities to therapeutic opportunities. *Semin. Cancer Biol.* 3, 16–25. doi: 10.1016/j.semcancer.2016.01.001

Castle, J. C., Loewer, M., Boegel, S., Tadmor, A. D., Boisguerin, V., de Graaf, J., et al. (2014). Mutated tumor alleles are expressed according to their DNA frequency. *Sci. Rep.* 4:4743. doi: 10.1038/Srep04743

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31, 213–219. doi: 10.1038/nbt.2514

Cidado, J., Wong, H. Y., Rosen, D. M., Cimino-Mathews, A., Garay, J. P., Fessler, A. G., et al. (2016). Ki-67 is required for maintenance of cancer stem cells but not cell proliferation. *Oncotarget* 7:6281. doi: 10.18632/oncotarget.7057

Curia, M. C., De Iure, S., De Lellis, L., Veschi, S., Mammarella, S., White, M. J., et al. (2012). Increased variance in germline allele-specific expression of APC associates with colorectal cancer. *Gastroenterology* 142, 71.e1–77.e1. doi: 10.1053/j.gastro.2011.09.048

de Bock, C. E., Ardjmand, A., Molloy, T. J., Bone, S. M., Johnstone, D., Campbell, D. M., et al. (2012). The Fat1 cadherin is overexpressed and an independent prognostic factor for survival in paired diagnosis-relapse samples of precursor B-cell acute lymphoblastic leukemia. *Leukemia* 26, 918–926. doi: 10.1038/leu.2011.319

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Du, J. L., Yuan, Z. F., Ma, Z., Song, J., Xie, X., and Chen, Y. (2014). KEGG-PATH: kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Mol. Biosyst.* 10, 2441–2447. doi: 10.1039/c4mb00287c

Dudderidge, T. J., McCracken, S. R., Loddo, M., Fanshawe, T. R., Kelly, J. D., Neal, D. E., et al. (2007). Mitogenic growth signalling, DNA replication licensing, and survival are linked in prostate cancer. *Br. J. Cancer* 96, 1384–1393. doi: 10.1038/sj.bjc.6603718

Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* 10, 241–251. doi: 10.1038/nrg2554

Ge, B., Pokholok, D. K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D. J., et al. (2009). Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.* 41, 1216–1222. doi: 10.1038/ng.473

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., et al. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* 10, 1081–1082. doi: 10.1038/Nmeth.2642

Graze, R. M., Novelo, L. L., Amin, V., Fear, J. M., Casella, G., Nuzhdin, S. V., et al. (2012). Allelic imbalance in Drosophila hybrid heads: exons, isoforms, and evolution. *Mol. Biol. Evol.* 29, 1521–1532. doi: 10.1093/molbev/msr318

GTEx Consortium (2015). The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Hasin-Brumshtein, Y., Hormozdiari, F., Martin, L., van Nas, A., Eskin, E., Lusis, A. J., et al. (2014). Allele-specific expression and eQTL analysis in mouse adipose tissue. *BMC Genomics* 15:471. doi: 10.1186/1471-2164-15-471

Heap, G. A., Yang, J. H., Downes, K., Healy, B. C., Hunt, K. A., Bockett, N., 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. doi: 10.1093/hmg/ddp473

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* 47, 199–208. doi: 10.1038/ng.3192

Jaenisch, R., and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* 33(Suppl.), 245–254. doi: 10.1038/ng1089

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/Gb-2013-14-4-R36

Kim, H. R., Choi, H. J., Kim, Y. K., Kim, H. J., Shin, J. H., Suh, S. P., et al. (2013). Allelic expression imbalance of JAK2 V617F mutation in BCR-ABL negative myeloproliferative neoplasms. *PLoS One* 8:e52518. doi: 10.1371/journal.pone.0052518

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., et al. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25, 2283–2285. doi: 10.1093/bioinformatics/btp373

Kukurba, K. R., Zhang, R., Li, X., Smith, K. S., Knowles, D. A., Tan, M. H., et al. (2014). Allelic expression of deleterious protein-coding variants across human tissues. *PLoS Genet.* 10:e1004304. doi: 10.1371/journal.pgen.1004304

Kwaepila, N., Burns, G., and Leong, A. S. (2006). Immunohistological localisation of human FAT1 (hFAT) protein in 326 breast cancers. Does this adhesion molecule have a role in pathogenesis? *Pathology* 38, 125–131. doi: 10.1080/00313020600559975

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–U354. doi: 10.1038/Nmeth.1923

Li, G., Bahn, J. H., Lee, J. H., Peng, G., Chen, Z., Nelson, S. F., et al. (2012). Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res.* 40:e104. doi: 10.1093/nar/gks280

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liu, Z., Gui, T. T., Wang, Z., Li, H., Fu, Y., Dong, X., et al. (2016). cisASE: a likelihood-based method for detecting putative cis-regulated allele-specific expression in RNA sequencing data. *Bioinformatics* 32, 3291–3297. doi: 10.1093/bioinformatics/btw416

Lo, H. S., Wang, Z. N., Hu, Y., Yang, H. H., Gere, S., Buetow, K. H., et al. (2003). Allelic variation in gene expression is common in the human genome. *Genome Res.* 13, 1855–1862. doi: 10.1101/gr.1006603

Maleno, I., Lopez-Nevot, M. A., Cabrera, T., Salinero, J., and Garrido, F. (2002). Multiple mechanisms generate HLA class I altered phenotypes in laryngeal carcinomas: high frequency of HLA haplotype loss associated with loss of heterozygosity in chromosome region 6p21. *Cancer Immunol. Immunother.* 51, 389–396. doi: 10.1007/s00262-002-0296-0

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195. doi: 10.1126/science.1222794

Padmanabhan, V., Callas, P., Philips, G., Trainer, T. D., and Beatty, B. G. (2004). DNA replication regulation protein Mcm7 as a marker of proliferation in prostate cancer. *J. Clin. Pathol.* 57, 1057–1062. doi: 10.1136/jcp.2004.016436

Pastinen, T. (2010). Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.* 11, 533–538. doi: 10.1038/nrg2815

Prendergast, J. G., Tong, P., Hay, D. C., Farrington, S. M., and Semple, C. A. (2012). A genome-wide screen in human embryonic stem cells reveals novel sites of allele-specific histone modification associated with known disease loci. *EpigeneticsChromatin* 5:6. doi: 10.1186/1756-8935-5-6

Puigvert, J. C., Sanjiv, K., and Helleday, T. (2016). Targeting DNA repair, DNA metabolism and replication stress as anti-cancer strategies. *FEBS J.* 283, 232–245. doi: 10.1111/febs.13574

Quinn, J. J., and Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* 17, 47–62. doi: 10.1038/nrg.2015.10

Reddy, T. E., Gertz, J., Pauli, F., Kucera, K. S., Varley, K. E., and Newberry, K. M. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* 22, 860–869. doi: 10.1101/gr.131201.111

Rehwinkel, J., Raes, J., and Izaurralde, E. (2006). Nonsense-mediated mRNA decay: target genes and functional diversification of effectors. *Trends Biochem. Sci.* 31, 639–646. doi: 10.1016/j.tibs.2006.09.005

Sadeqzadeh, E., de Bock, C. E., Zhang, X. D., Shipman, K. L., Scott, N. M., Song, C., et al. (2011). Dual processing of FAT1 cadherin protein by human melanoma cells generates distinct protein products. *J. Biol. Chem.* 286, 28181–28191. doi: 10.1074/jbc.M111.234419

Seshagiri, S., Stawiski, E. W., Durinck, S., Modrusan, Z., Storm, E. E., Conboy, C. B., et al. (2012). Recurrent R-spondin fusions in colon cancer. *Nature* 488, 660–664. doi: 10.1038/nature11282

Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* 21, 1728–1737. doi: 10.1101/gr.119784.110

Smith, R. M., Webb, A., Papp, A. C., Newman, L. C., Handelman, S. K., and Suhy, A. (2013). Whole transcriptome RNA-Seq allelic expression in human brain. *BMC Genomics* 14:571. doi: 10.1186/1471-2164-14-571

So, T., Takenoyama, M., Mizukami, M., Ichiki, Y., Sugaya, M., Hanagiri, T., et al. (2005). Haplotype loss of HLA class I antigen as an escape mechanism from immune attack in lung cancer. *Cancer Res.* 65, 5945–5952. doi: 10.1158/0008-5472.Can-04-3787

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Tuch, B. B., Laborde, R. R., Xu, X., Gu, J., Chung, C. B., Monighetti, C. K., et al. (2010). Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS One* 5:e9317. doi: 10.1371/journal.pone.0009317

Tung, J., Akinyi, M. Y., Mutura, S., Altmann, J., Wray, G. A., and Alberts, S. C. (2011). Allele-specific gene expression in a wild nonhuman primate population. *Mol. Ecol.* 20, 725–739. doi: 10.1111/j.1365-294X.2010.04970.x

Valle, L., Serena-Acedo, T., Liyanarachchi, S., Hampel, H., Comeras, I., Li, Z., et al. (2008). Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. *Science* 321, 1361–1365. doi: 10.1126/science.1159397

Walker, E. J., Zhang, C., Castelo-Branco, P., Hawkins, C., Wilson, W., Zhukova, N., et al. (2012). Monoallelic expression determines oncogenic progression and outcome in benign and malignant brain tumors. *Cancer Res.* 72, 636–644. doi: 10.1158/0008-5472.CAN-11-2266

Wang, X. C., Zhang, J. Q., Shen, Y. Q., Miao, F. Q., and Xie, W. (2006). Loss of heterozygosity at 6p21.3 underlying HLA class I downregulation in gastric cancer. *J. Exp. Clin. Cancer Res.* 25, 115–119.

Wei, Q. X., Claus, R., Hielscher, T., Mertens, D., Raval, A., Oakes, C. C., et al. (2013). Germline allele-specific expression of DAPK1 in chronic lymphocytic leukemia. *PLoS One* 8:e55261. doi: 10.1371/journal.pone.0055261

Yan, H., Yuan, W. S., Velculescu, V. E., Vogelstein, B., and Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science* 297, 1143–1143. doi: 10.1126/science.1072545

Zhang, K., Li, J. B., Gao, Y., Egli, D., Xie, B., Deng, J., et al. (2009). Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods* 6, 613–618. doi: 10.1038/nmeth.1357

Zollikofer, C., Ringhoffer, M., Kündgen, L., Fürst, D., and Schrezenmeier, H. (2014). Complete loss of HLA class I heterozygosity in a patient with acute myeloid leikemia. *Oncol. Res. Treat.* 37, 135–135.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.