www.oncotarget.com Oncotarget, 2020, Vol. 11, (No. 4), pp: 462-479

Research Paper

Tumor suppressor genes and allele-specific expression: mechanisms and significance

Evan A. Clayton[1,2], Shareef Khalid[1,2], Dongjo Ban[1,2], Lu Wang[2,3], I. King Jordan[1,2,3,4] and John F. McDonald[1,2]

[1]Integrated Cancer Research Center, School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA

[2]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA

[3]PanAmerican Bioinformatics Institute, Cali, Colombia

[4]Applied Bioinformatics Laboratory, Atlanta, GA, USA

Correspondence to: *John F. McDonald,* ***email:*** *john.mcdonald@biology.gatech.edu*

Keywords: *allele-specific expression; alternative splicing; antisense RNA; cancer; tumor-suppressor genes*

Received: *November 21, 2019* ***Accepted:*** *January 13, 2020* ***Published:*** *January 28, 2020*

ABSTRACT

Recent findings indicate that allele-specific expression (ASE) at specific cancer driver gene loci may be of importance in onset/progression of the disease. Of particular interest are loss-of-function (LOF) of tumor suppressor gene (TSGs) alleles. While LOF tumor suppressor mutations are typically considered to be recessive, if these mutant alleles can be significantly differentially expressed relative to wild-type alleles in heterozygotes, the clinical consequences could be significant.

LOF TSG alleles are shown to be segregating at high frequencies in world-wide populations of normal/healthy individuals. Matched sets of normal and tumor tissues isolated from 233 cancer patients representing four diverse tumor types demonstrate functionally important changes in patterns of ASE in individuals heterozygous for LOF TSG alleles associated with cancer onset/progression. While a variety of molecular mechanisms were identified as potentially contributing to changes in ASE patterns in cancer, changes in DNA copy number and allele-specific alternative splicing possibly mediated by antisense RNA emerged as predominant factors.

In conclusion, LOF TSGs are segregating in human populations at significant frequencies indicating that many otherwise healthy individuals are at elevated risk of developing cancer. Changes in ASE between normal and cancer tissues indicates that LOF TSG alleles may contribute to cancer onset/progression even when heterozygous with wild-type functional alleles.

INTRODUCTION

The long-standing belief that cancer is a genetic disease driven by mutations in a select set of oncogenes and/or tumor suppressor genes (aka, "cancer driver" genes) [1–3], has been augmented in recent years to incorporate the auxiliary contribution of changes in a variety of regulatory controls [4–6]. Findings indicate that these additional regulatory controls may, in at least some instances, manifest as allele-specific expression (ASE) at specific cancer driver gene loci [7, 8]. ASE is the phenomenon whereby two or more gene alleles are differentially expressed with respect to one another [9]. The potential clinical consequences of ASE have been previously documented [10] including emerging evidence for the potential contribution of ASE to cancer [8, 11].

If cancer driver mutations can be transcriptionally repressed/de-repressed in an allele-specific manner, it follows that mutations in these genes may be necessary but not always sufficient for onset and progression of the disease. For example, cancer driver mutations may, to a greater or lesser extent, be repressible and thus segregating

at higher than expected frequencies in populations of normal healthy individuals. In addition, regulatory modulations in the ASE of cancer driver mutations may themselves, in at least some instances, be a significant contributor to cancer onset and progression. Of particular interest, in this regard, are those genes where loss-of-function (LOF) mutations have been shown to drive cancer onset/progression. This class of cancer driver genes is commonly known as tumor suppressor genes (TSGs) because a functional wild-type allele is considered sufficient to "suppress" the cancer driver effect of LOF alleles in heterozygotes. While LOF tumor suppressor mutations are typically considered to be recessive [12], if these mutant alleles can be significantly differentially expressed relative to wild-type alleles in heterozygotes, the clinical consequences could be significant.

In this study, we first demonstrate that LOF TSG alleles are segregating in world-wide populations of normal/healthy individuals at relatively high frequencies, thereby establishing the potential importance of these genes in pre-disposing otherwise healthy individuals to cancer. To directly evaluate the possible contribution of ASE of tumor suppressor LOF alleles in cancer onset/progression, we analyzed matched sets of normal and tumor tissues isolated from 233 cancer patients representing four diverse tumor types. The results indicate that there are functionally important changes in ASE in individuals heterozygous for LOF TSG alleles associated with cancer onset/progression. While a variety of molecular mechanisms were identified as potentially contributing to changes in ASE in cancer, changes in DNA copy number and allele-specific alternative splicing possibly mediated by antisense RNA emerged as predominant factors.

RESULTS

Tumor suppressor mutations are abundant in human populations

The Catalogue Of Somatic Mutations In Cancer (COSMIC) is the world's largest database of somatic mutations associated with cancer onset and progression [13]. To determine the extent to which cancer associated mutations are segregating in the general human population, the genomic locations of all coding mutations in COSMIC census genes were intersected with sequence variants identified in individuals comprising the Phase 3 release of the One Thousand Genomes Project (1KGP). The Phase 3 release catalogues all of the genetic variants present in 2504 putatively healthy individuals, representing a diversity of racial and ethnic groups randomly selected from 26 human populations around the world.

Remarkably, all individuals in the 1KGP were found to contain at least 31 homozygous and 68 heterozygous COSMIC census mutations (Supplementary Figure 1). In total, 2,296 and 3,123 COSMIC census mutations were found in oncogenes and tumor suppressor genes, respectively, in healthy individuals. However, since the functional significance of all COSMIC mutations is not yet known and the fact that gain-of-function (dominant) mutations are difficult to unambiguously identify [14], we focused our subsequent analyses on COSMIC mutations in TSGs that could be definitively classified as deleterious (*i. e.*, non-sense, frame-shift, deletion mutations), along with all missense mutations predicted to be damaging by both The Sorting Intolerant from Tolerant (SIFT) [15] and Polymorphism Phenotyping v2 (PolyPhen-2) [16] algorithms. Employing this more conservative metric, 448 LOF COSMIC census mutations (28 truncating, 420 missense predicted damaging) in TSGs were identified (Supplementary Dataset 1), of which ~93% of individuals carried at least one (Figure 1A). These 448 LOF mutations mapped to 137 different TSGs in at least one individual and four of these TSGs, Cbl Proto-Oncogene C (*CBLC*), Cadherin 11 (*CDH11*), Leucine Zipper Like Transcription Regulator 1 (*LZTR1*), and Tet Methylcytosine Dioxygenase 2 (*TET2*) had LOF mutations in >25% of the population (Figure 1B). Collectively, these findings indicate that genetic variants previously characterized as "cancer driver" mutations are segregating at relatively high frequencies in populations of individuals not afflicted with the disease.

A minority (<20%) of TSGs display genetic profiles in the cancer samples that are consistent with Knudson's two-hit hypothesis

Given the relative abundance of TSG LOF alleles in human populations, we utilized The Cancer Genome Atlas (TCGA) database [17] to explore the possible contribution of these genes to cancer onset and/or progression by examining matched sets of cancer and normal tissues collected from 233 cancer patients representative of four diverse cancer types (breast invasive carcinoma, head and neck squamous cell carcinoma, lung adenocarcinoma, and thyroid carcinoma). According to a model first proposed by Alfred Knudson in 1971 [18], newly arising LOF TSG mutant alleles, being recessive, can be carried by normal cells with little significant negative effect. According to this model, acquisition of a second LOF mutation in the alternate wild-type allele is pre-requisite for tumor onset.

To test this hypothesis in our dataset, we genotyped all samples and identified TSGs that were heterozygous for a LOF mutation in normal tissues but that have acquired a secondary LOF mutation in the wild-type allele in the tumor samples. In total we found that only 46 of the 233 cancer patients (19.7%) were associated with acquisition of homozygosity in cancer for LOF alleles at TSG loci consistent with Knudson's "two-hit" hypothesis. These results indicate that the vast majority of TSGs heterozygous for wild-type and LOF alleles in normal tissues remain heterozygous in tumor tissue. However, if

recessive LOF alleles can be significantly overexpressed relative to the wild-type alleles in an ASE fashion, LOF TSGs may be significant contributors to cancer onset/progression even in the heterozygous state.

The proportion of LOF mutations displaying ASE is significantly elevated in cancer tissue samples

To explore the possible contribution of ASE in matched sets of normal and cancer tissues, we employed DNA-seq data from the TCGA database to identify all heterozygous sites in the exome and subsequently leveraged complementary RNA-seq data to compare the expression of wild-type or "reference" (ref) vs LOF mutant or "alternative" (alt) alleles at those loci (Supplementary Figure 2).

The proportion of COSMIC census mutations in TSGs displaying ASE was found to be significantly higher in the cancer relative to normal tissues for breast invasive carcinoma, head and neck squamous cell carcinoma, and lung adenocarcinoma ($P < 3.11 \times 10^{-10}$) (Figure 2). Thyroid carcinoma was the only cancer type not displaying a significant difference, perhaps because these cancers are typically associated with a relatively low mutation rate [19] resulting in relatively fewer genetic alterations.

To determine if this regulatory change was limited to TSG loci, we computed ASE for all heterozygous single nucleotide polymorphisms (het-SNPs) exome-wide. We found that all genes, on average, contain a significantly higher proportion of het-SNPs displaying ASE in breast, lung, head and neck ($P < 3.46 \times 10^{-15}$) and thyroid ($P < 0.005$) tumors than normal samples (Figure 2). Thus, dysregulation in cancer, at least as manifest by ASE, is not limited to TSGs but extends to genes not previously identified as being implicated in tumorigenesis.

Differences in patterns of ASE between normal and tumor tissues includes but is not limited to TSGs

Changes in the relative expression of wild-type (ref) alleles vs. mutant (alt) alleles between normal and cancer tissues may manifest in one of six alternative ASE patterns (Figure 3A): Pattern 1: No significant difference in ASE (ref=alt) in normal tissues but significant ASE (ref<alt) in cancer tissues; Pattern 2: Significant ASE in normal tissues (ref>alt) but no significant ASE (ref=alt) in cancer tissues; Pattern 3: Significant ASE in normal sample (ref>alt) and significant ASE in tumor sample (ref<alt); Pattern 4: No significant ASE in normal tissues (ref=alt) but significant ASE in cancer tissues (ref>alt); Pattern 5: Significant ASE (ref<alt) in normal tissues but no significant ASE in cancer tissues (ref=alt); Pattern 6: Significant ASE in normal (ref<alt) and in cancer tissues (ref>alt). Patterns 1-3 are potentially of the most significance to cancer onset/progression because, in each case, the expression of the cancer driver LOF mutant (alt) allele is expressed at a higher level than the wild-type allele in cancer tissues.

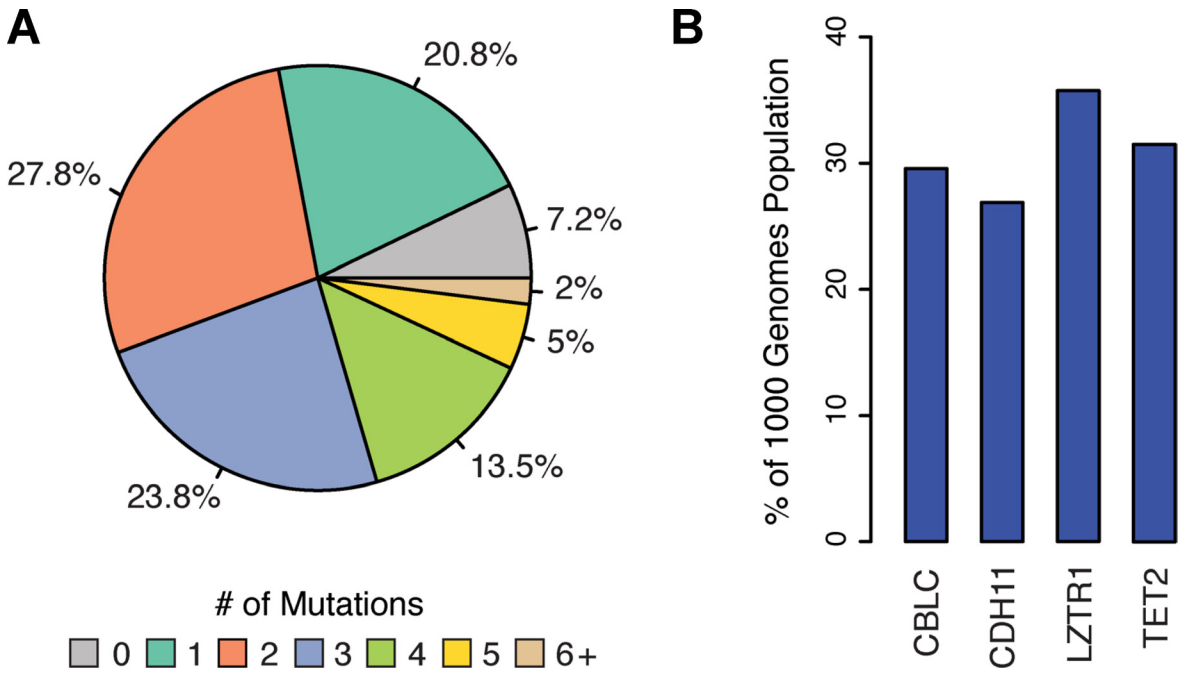

Figure 1: Distribution of LOF COSMIC census mutations in TSGs of the 1KGP. Cancer associated mutations were identified in the 1000 genomes population (1KGP) as detailed in the Materials and Methods. (**A**) Pie chart depicting the percent of the 1KGP containing deleterious cancer associated mutations in at least one TSG. (**B**) Four TSGs most frequently mutated (LOF) in 1KGP.

The observed changes in ASE between matched sets of normal and cancer tissues for each of the 233 patients, grouped into their respective Patterns, is presented in Table 1. A significant percentage of mutations in TSGs were found to display various patterns of ASE (FDR = 5%, $P < 0.005$; breast 14.9%, head and neck 16.0% and lung 19.6%) with Patterns 1 and 4 being the most predominant (see also Figure 3B). Thyroid cancer again stood out as an outlier where only 4.1% of mutations in TSGs were found to display ASE with Patterns 1 (2.3%) and 5 (1.3%) being nearly equally abundant.

When the analysis was extended to include all transcribed genes ("%Total SNPs" in Table 1), a similar trend was observed, where 14.8%, 16.6% and 18.3% of all het-SNPs were found to display ASE in breast, head and neck and lung cancers, respectively. Thyroid cancer was again an outlier with only 4.5% of all transcribed genes displaying ASE (Figure 3C). Collectively these results indicate that changes in ASE in cancer are widespread and not limited to TSGs.

To explore this apparent dysregulation of COSMIC census mutations in TSGs further, we aggregated our SNP ASE data to quantify ASE of the entire allele of a gene by employing the Meta-analysis Based Allele-Specific Expression Detection (MBASED) protocol [8]. We found 14.4%, 17.9%, 20.4% and 5.7% of all TSGs show ASE in breast, head and neck, lung and thyroid cancers, respectively (Supplementary Table 1). These results are consistent with the relative levels of ASE associated with individual SNPs in these cancers with Pattern 1 again emerging as a predominant pattern (9.1%, 11.1%, 13.2%, and 2.8%) (Supplementary Table 1).

One example of those TSGs displaying ASE in cancer is the Human Leukocyte Antigen A1 gene (*HLA-A*). *HLA-A* has been previously identified as a hotspot for ASE activity [20], likely due to the high genetic variability that is well-documented in the major histocompatibility complex [21]. We detected ASE in the *HLA-A* gene in 20% of patient samples including nucleotide positions not previously reported to display ASE [22].

Another example is Tumor Protein P53 (*TP53*) that displayed the most changes in ASE within the breast cancer patients (57.9% of all patients) displaying Pattern 4 63.6% of the time (Figure 4). Additionally, breast cancer implicated TSGs Breast cancer type 1 susceptibility protein (*BRCA1*) and Cadherin 1 (*CDH1*) were found to display changes in ASE in 15.4% and 32.4% of breast tumors, respectively, frequently displaying Pattern 1 (Figure 4). Interestingly, Zinc Finger Protein 331 (*ZNF331*) was the only TSG predominately displaying Pattern 2 (Figure 4). A previous study [23] has shown *ZNF331* to display large amounts of ASE in breast cancer, citing genomic imprinting as a possible explanation [24].

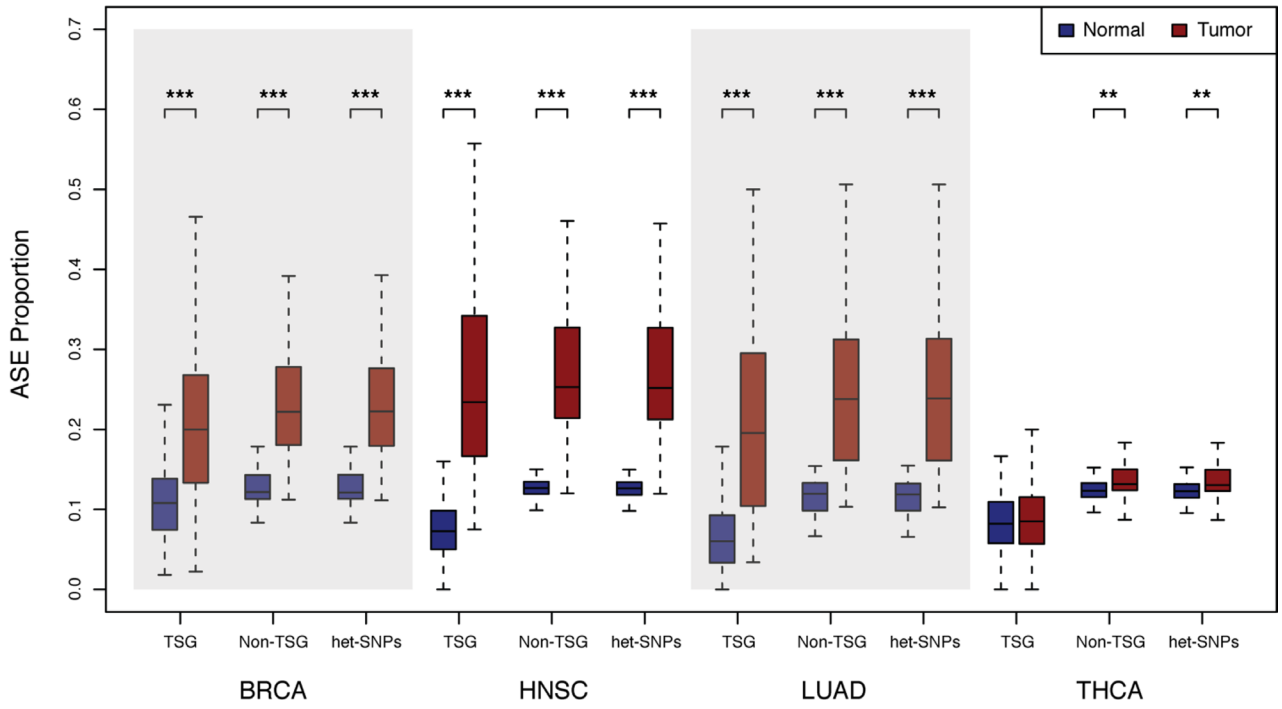

Figure 2: Distribution of the proportion of ASE loci. Allele counts were generated for normal and primary tumor tissue pairs for breast invasive carcinoma (BRCA), head and neck squamous cell carcinoma (HNSC), lung adenocarcinoma (LUAD), and thyroid carcinoma (THCA) via analysis of RNA-Seq as described in the Materials and Methods section. Boxplots show the distribution of the proportion of heterozygous COSMIC Census mutations in tumor suppressor genes (TSGs), all heterozygous SNPs in non-tumor suppressor genes (Non-TSG) and all heterozygous single nucleotide polymorphisms exome-wide (All het-SNPs) with significant ASE (FDR = 5%, $P < 0.005$) in normal (blue) and tumor (red) samples ($^{***}P < 3.46 \times 10{-}15$; $^{**}P < 0.005$).

The four TSGs: *CBLC*, *CDH11, LZTR1*, and *TET2* previously shown to be most frequently mutated in 1KGP (Figure 1B) were also observed to display changes in ASE in breast cancer (Figure 4). Similar trends in the frequency of ASE Patterns among TSGs were observed in head and neck, lung, and thyroid cancers, with thyroid again sporting the least amount of ASE (Supplementary Figure 3).

Changes in DNA allelic ratios may explain up to 35% of the observed changes in ASE between normal and cancer samples

Perhaps the most straight-forward explanation of the observed changes in ASE in cancer is that it reflects the underlying changes in allele counts on the DNA level. For example, it is known that the duplication or deletion of

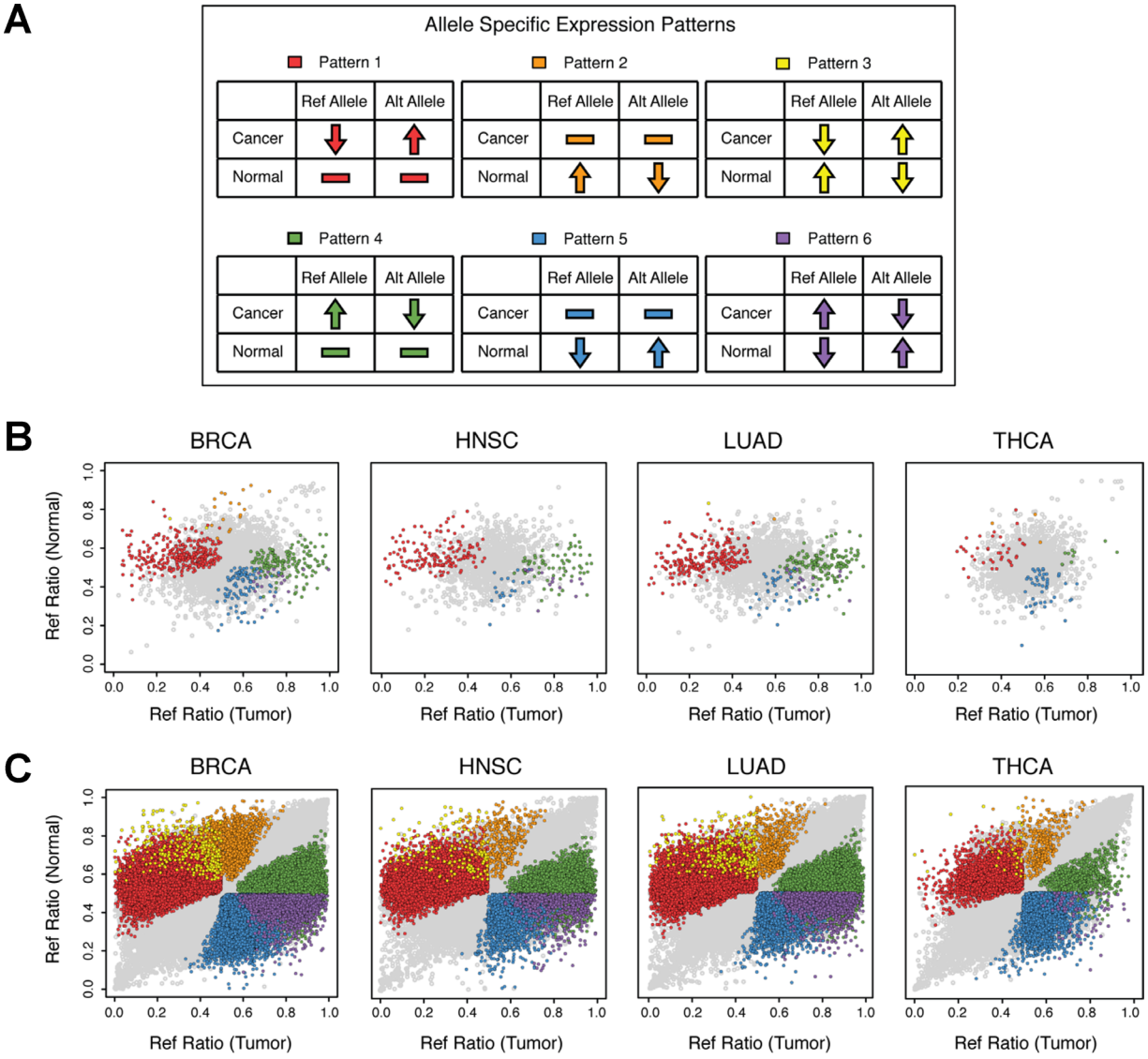

Figure 3: ASE SNP patterns. Allele counts were generated for normal and primary tumor tissue pairs for breast invasive carcinoma, head and neck squamous cell carcinoma, lung adenocarcinoma and thyroid carcinoma via analysis of RNA-Seq as described in the Materials and Methods section. Sites demonstrating significantly different ASE ratios ($P < 0.05$) between normal and tumor sample pairs are color coded by expression pattern as demonstrated in the top panel. (**A**) Six ASE patterns of interest were analyzed; Pattern 1: No significant difference in ASE (ref=alt) in normal tissues but significant ASE (ref<alt) in cancer tissues; Pattern 2: Significant ASE in normal tissues (ref>alt) but no significant ASE (ref=alt) in cancer tissues; Pattern 3: Significant ASE in normal sample (ref>alt) and significant ASE in tumor sample (ref<alt); Pattern 4: No significant ASE in normal tissues (ref=alt) but significant ASE in cancer tissues (ref>alt); Pattern 5: Significant ASE (ref<alt) in normal tissues but no significant ASE in cancer tissues (ref=alt); Pattern 6: Significant ASE in normal (ref<alt) and in cancer tissues (ref>alt). Significant ASE (FDR = 5%, $P < 0.005$) was determined using a binomial test within samples in order to group loci into patterns. (**B**) Reference allele ratios (ref/total) for all COSMIC Census loci in TSGs intersecting normal and tumor sample pairs, for 233 TCGA participants are shown here. (**C**) Reference allele ratios (ref/total) for all loci intersecting normal and tumor sample pairs, for 233 TCGA participants are shown here.

Table 1: Percent of SNPs displaying ASE in 233 TCGA patients

	Pattern	% Total SNPs	% All COSMIC	% TSG	% Oncogene	% Fusion
BRCA	1	7.3	7.6	7.3	7.0	8.3
	2	0.4	0.8	0.3	0.4	1.4
	3	0.1	0.2	0.0	0.0	0.3
	4	3.8	4.1	3.7	3.4	4.8
	5	2.6	2.5	2.8	2.2	2.7
	6	0.5	0.5	0.8	0.4	0.4
	No ASE	85.2	84.4	85.1	86.6	82.1
HNSC	1	8.7	9.7	10.5	10.1	8.3
	2	0.4	0.4	0.0	0.0	0.9
	3	0.2	0.3	0.0	0.0	0.6
	4	4.6	4.5	3.7	4.2	4.7
	5	2.1	1.7	1.3	1.2	2.0
	6	0.7	1.0	0.5	0.9	1.0
	No ASE	83.4	82.5	83.9	83.5	82.4
LUAD	1	9.7	10.7	9.2	9.8	12.0
	2	0.3	1.0	0.1	0.0	2.0
	3	0.2	1.0	0.0	0.0	1.9
	4	5.4	6.8	7.0	6.5	6.9
	5	1.9	2.3	2.2	1.6	2.7
	6	0.7	0.8	1.1	0.6	0.9
	No ASE	81.7	77.4	80.4	81.5	73.5
THCA	1	2.1	2.3	2.3	1.7	2.6
	2	0.2	0.5	0.0	0.1	1.0
	3	0.0	0.0	0.0	0.0	0.0
	4	0.5	0.8	0.5	0.4	1.1
	5	1.6	1.9	1.3	2.0	2.1
	6	0.1	0.0	0.0	0.0	0.0
	No ASE	95.5	94.5	95.9	95.8	93.2

alleles on the DNA level can contribute to ASE in cancer [7, 25]. In addition, the polyclonal heterogeneity of most tumors can manifest as an imbalance in DNA allele counts and associated ASE changes in analyses carried out on bulk tumor samples.

To explore the extent to which changes in DNA copy number may be contributing to the observed ASE in the samples, we downloaded whole-exome sequencing data (WXS) for nine randomly selected patients representing each of the three cancer types displaying the highest level of ASE (breast invasive carcinoma, lung adenocarcinoma, and head & neck squamous cell carcinoma) (Supplementary Table 2). We ensured these individuals displayed ASE in COSMIC genes (Supplementary Figure 4) and that their ASE was evenly spread throughout the genome (Supplementary Figure 5). We found that, on average, 35.2% (45.7% Breast, 25.8% Head and Neck, and 20.5% Lung) of ASE genes displayed DNA allele counts that correlated with RNA allele counts (Table 2). Further investigation of these samples, however, revealed that only 10% of these genes displayed copy number duplications potentially accounting for their ASE (Figure 5A). Collectively these findings indicate that while, on average, a large fraction of the observed changes in ASE may be accounted for by corresponding changes in DNA allele counts, many instances of ASE in the cancer samples are likely attributable to allele-specific changes in gene regulation.

Allele-specific cis-regulatory variation may account for a relatively small fraction of observed changes in ASE between normal and cancer samples

Allele-specific regulatory changes in gene expression could be explained by sequence variation mapping to *cis*-regulatory regions located up- or down-stream of affected genes [26–28]. To explore the extent to which allele-specific *cis*-regulatory variation may

account for ASE in cancer, we identified expression-quantitative trait loci (eQTLs) present in six of the nine patients' normal and tumor samples using the Genotype-Tissue Expression Project's (GTEx) single tissue *cis*-eQTL data available for breast and lung tissue [29]. eQTLs are regions of the genome containing DNA sequence variants previously established to regulate gene expression levels [30]. Genes previously established to be regulated by at least one eQTL are classified as eGenes [31].

Genes displaying ASE in our study were found to be significantly enriched for eGenes relative to genes not displaying ASE ($P = 0.018$) (Supplementary Figure 6). This finding was pronounced for breast ($P = 2.56 \times 10^{-6}$) and lung cancer ($P = 6.22 \times 10^{-4}$) patients (Supplementary Figure 6). However, collectively only 24% of genes displaying ASE in our dataset are eGenes and just 3% of ASE eQTLs are ASE-specific. Moreover, we found that the expression slope of an eQTL is not often correlated with the allelic expression of a gene (1.8%; Figure 5B; Supplementary Table 3). For example, consider the heterozygous eQTL variant (rs10654) mapping to the 3′ UTR of the *NUP54* gene in both normal and tumor samples of breast cancer patient 2 (TCGA-BH-A0BW). Despite being heterozygous, this eQTL is not associated with ASE in the normal tissue but is associated with ASE in cancer tissue where the alternative haplotype is overexpressed and in phase with the highly expressed alternative eQTL allele (Supplementary Figure 7A).

We also pursued the eQTLs differing in genotype between normal and tumor samples for specific evidence of *cis*-regulation. While infrequent, we did find several notable cases where eQTL genotypes correlated with ASE. Shown in Supplementary Figure 7B, is a model for how *cis*-eQTLs may be responsible for the intragenic ASE we observed. In this particular example, three separate eQTLs within a 50bp region (rs34176173, rs12085114, rs34016668) are found ~4.8k base pairs from the 3′ UTR of the gene *NME7* in breast invasive carcinoma patient 3 (TCGA-BH-A0DT). The eQTL is homozygous for the ref allele in the normal sample that does not show ASE, and heterozygous

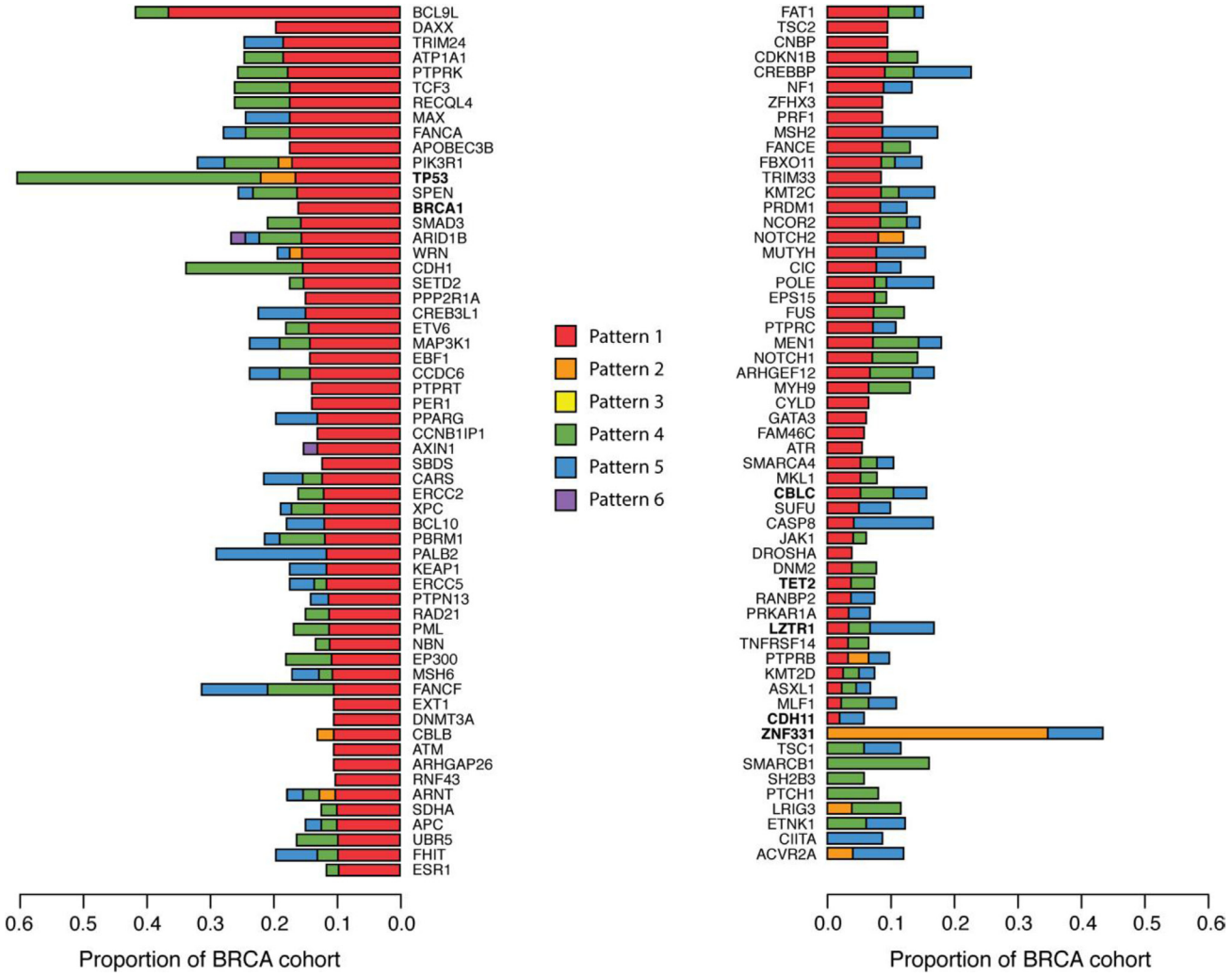

Figure 4: Tumor suppressor genes with ASE in breast cancer patients. Gene level ASE was computed as described in the Materials and Methods section. The proportion of breast cancer patients with ASE in 115 TSGs are shown here, colored by ASE Pattern.

Table 2: ASE patterns potentially explained by DNA counts

Patient	ASE SNPs explained by DNA counts	Total ASE SNPs	Percentage of ASE correlated
Breast 1	526	1336	39.4
Breast 2	719	1411	51.0
Breast 3	177	367	48.2
Head & Neck 1	275	930	29.6
Head & Neck 2	179	674	26.6
Head & Neck 3	217	993	21.9
Lung 1	56	233	24.0
Lung 2	47	155	30.3
Lung 3	11	167	6.6
Total	**2207**	**6266**	**35.2**

in the tumor sample. The eQTL alternative allele that is associated with high expression of NME7 is present on the alt haplotype being overexpressed. Further, all three eQTLs are in linkage disequilibrium with the ASE SNP ($r > 0.42$) suggesting they segregate together. We found four additional cases where *cis*-eQTLs could account for ASE but none of these were associated with COSMIC census genes.

Collectively, the above findings indicate that while allele-specific *cis*-regulatory variation may account for some instances of ASE, it alone does not explain the vast majority (>75%) of instances of ASE in our dataset.

Changes in methylation do not appear to be a major contributor to the observed changes in ASE between normal and cancer samples

Another possibility is that ASE is regulated epigenetically. For example, it has been previously suggested that epigenetic inactivation of one of the two alleles could result in ASE [32]. Epigenetic effects across chromosomes are often regionally associated with CpG repeats or "CpG islands" [33]. To determine if genes displaying ASE in our dataset display evidence of regional chromosomal clustering, we visualized the genomic locations of ASE for nine patients on a genome ideogram (Supplementary Figure 5). The results provide no evidence for regional chromosomal clustering indicative of regional epigenetic effects.

To further search for evidence of epigenetic involvement in ASE in our dataset, we analyzed global DNA methylation in normal and cancer tissues since this is a common mechanism by which gene transcription can be repressed epigenetically [34, 35]. Methylation data were downloaded from TCGA for seven of the nine patients described above and used to compare genes that had a significant change in methylation with genes showing a significant change in ASE. We found that only 10.2% of genes displaying ASE also displayed significant differences (>1.3-fold) in methylation between normal and tumor tissues (Figure 5C; Supplementary Table 3). Although these results indicate that changes in methylation are not likely to be playing a significant role in the ASE detected in patient samples, the analysis cannot be considered definitive because the methylation data provided by TCGA are not allele specific.

A significant fraction of changes in ASE between normal and cancer may be a reflection of underlying alternative-splicing events

A recent study has implicated allele-specific alternative splicing as a potentially significant factor in ASE [36]. For example, consider a scenario where an allele-specific exon-skipping event occurs more frequently in a cancer tissue than normal (Supplementary Figure 8). This would result in a negligible difference in the level of transcripts containing the wild-type (ref) and LOF mutant (alt) allele in normal but significantly fewer transcripts containing the wild-type allele ("T allele", in Supplementary Figure 8) in cancer.

To explore the possibility that allele-specific alternative splicing may be contributing to the observed ASE in patient samples, we leverage previously computed isoform counts for TCGA patient data [37]. Specifically, we sought to determine if there is a significant increase in exon skipping in genes displaying ASE. The results indicate that 46% of SNPs displaying changes in ASE between normal and cancer correlate with an increased frequency of exon-skipping events (i. e., ≥1.5-fold increase in expression of reads consistent with exon-skipping events) (Table 3; Figure 5D).

While these results suggest that allele-specific alternative splicing may be a significant contributor to ASE, it does not provide a mechanism as to how two variant alleles from the same gene may be alternatively spliced. One possibility is that the point mutations or

indels that distinguish mutant LOF (alt) alleles from wild-type (ref) alleles map to consensus splice sites or other *cis*-regulatory locations known to be involved in the splicing process [38]. However, of the 100,852 SNPs associated with changes in ASE between normal and tumor, only 1.4% (1,418/100,852) map to consensus splice sites (716 in acceptor G, 702 in donor AG) (Supplementary Dataset 2).

A second possible mechanism that may explain how two variant alleles from the same gene may be alternatively spliced emerges from previous studies showing that splicing events can be experimentally induced *in vivo* by exposing primary transcripts to even small fragments of antisense RNAs that pair with known splice sites in the primary transcript [39, 40]. We reasoned that if such allele-specific antisense RNAs are being differentially produced in normal and cancer tissues, it may explain observed differences in allele-specific splicing and consequent differences in ASE.

To test this hypothesis, we estimated the levels of antisense RNAs mapping to splice sites adjacent to allele-specific alternative-splice events. The results presented in Table 3 demonstrate a notable increase in levels of antisense RNA in genes displaying allele-specific alternative-splice events associated with ASE. For example, Figure 6A depicts a case where the *ADAM15* gene displays ASE in breast cancer patient 3 (TCGA-BH-A0DT). The ADAM15 protein is known to display tumor suppressive activities when it is released as an exosomal component [41], and abnormal expression and dysregulation of alternative splicing in *ADAM15* has been previously associated with breast cancer [42]. Previous studies have also shown that four ADAM15 isoforms varying by the sequence of the cytoplasmic domains, display variable effects *in vitro*. The shortest isoform, ADAM-15D, arises due to loss of exons 19 to 21 causing a reading frame shift in exons 22 and 23 when compared with the other three isoforms. The variant lacks proline-rich modules and has a distinct sequence of

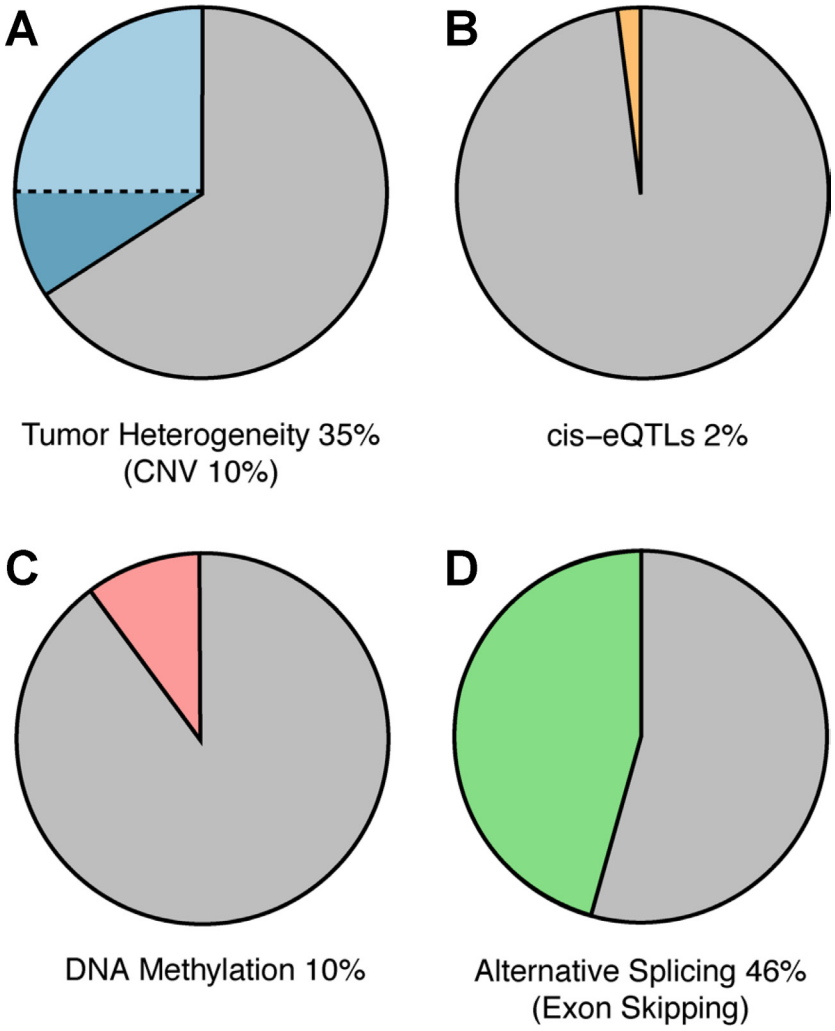

Figure 5: Mechanisms of ASE. Potential underlying mechanisms for ASE were explored as outlined in the Materials and Methods. Pie chart depicts amount of ASE that could be attributed to (**A**) tumor heterogeneity and copy number variation (CNV, darker blue) (**B**) *cis*-eQTLs, (**C**) DNA methylation and (**D**) exon-skipping via computational analysis.

Table 3: ASE SNPs with differentially expressed exon-skipping events and enrichment of antisense RNA

ASE Pattern	Total ASE SNPs	ASE SNPs w/ 1.5× ISO1[a]	Percentage of ASE correlated	ASE SNPs w/ 1.5× ISO1 and 1.5× AS[b]	Percentage of ISO1 correlated w/ antisense
1	591	278	47.0	191	32.3
2	13	5	38.5	1	7.7
3	4	4	100.0	4	100.0
4	500	217	43.4	155	31.0
5	59	18	30.5	7	11.9
6	71	51	71.8	51	71.8
Total	1238	573	**46.3**	409	**33.0**

[a]Reads supporting ISO1 (isoform 1 genelet) are split-reads spanning the two flanking exons adjacent to the skipped exon
[b]RNA reads with predicted antisense orientation.

37 amino acids. As shown in Figure 6B, we observe an increase in antisense RNA mapping to acceptor (1.7×) and donor (1.8×) sites in this patient's tumor. We have identified an exon-skipping event (exon 19), consistent with the ADAM-15D isoform. The increase in antisense RNA correlates with this isoform's expression, which is substantially higher (3.8×) in the patient's tumor sample when compared to normal and could explain ASE at this locus (Figure 6C).

Another example is illustrated in Supplementary Figure 9A, where the Lysyl Oxidase Like 2 (*LOXL2*) gene displays ASE at the *rs1051146* locus in breast cancer patient 1 (TCGA-BH-A0B3), which overlaps with an exon-skipping event. *LOXL2* has accumulated numerous reports that document its role in cancer formation and proliferation of breast cancer [43, 44]. Further, research has shown that a short isoform of LOXL2 missing exon 13 can regulate cancer cell migration and invasion through a dissimilar mechanism compared to its canonical form [45]. Here, we observe more antisense RNA mapping to acceptor (8.7×) and donor (9.2×) sites (Supplementary Figure 9B) and increased skipping of exon 6 (9.9×) (Supplementary Figure 9C) in breast cancer patient 1's tumor sample, both correlated with an increase in ASE.

Tenascin C (*TNC*) is a gene belonging to a family of extracellular matrix (ECM) glycoproteins that is known to be overexpressed in cancer cells. Studies have shown that remodeling of ECM in cancer can affect cellular interaction as ECM influences behavior of the cells [46, 47]. One specific study has shown that a TNC isoform containing exons 14 and 16 but not 15 is upregulated in breast cancer, which leads to increased cell invasion and proliferation [48]. In breast cancer patient 3, *TNC* displays changes in ASE at *rs17819466* inside exon 15 (Supplementary Figure 10A). Antisense RNA mapping to acceptor (2.1×) and donor (2.6×) sites are elevated in the tumor sample (Supplementary Figure 10B), as are split-reads spanning exons 14 and 16 (5.5×) (Supplementary Figure 10B).

Collectively, these results suggest that antisense RNA mediated alternative splicing may be a significant factor in accounting for our observed changes in ASE between normal and cancer samples.

DISCUSSION

Cancer is a complex disease not only from the perspective of the number and diversity of genes involved but also because of the existence of extensive regulatory variation controlling the expression of these genes. One manifestation of these regulatory controls is allele-specific expression (ASE) at specific cancer driver gene loci [7]. If cancer driver mutations can be transcriptionally repressed/de-repressed in an allele-specific manner, they may be segregating at higher than expected frequencies in populations of normal healthy individuals. In an initial effort to explore this possibility, we conducted a computational analysis of functionally significant cancer driver mutations in a sampling of normal healthy human populations across the world (2.5 thousand genomes comprising the 1000 Genome Project (1KGP) [49]). While relatively few confirmed dominant oncogene mutations were found to be segregating in these populations, 93% of healthy individuals sampled were found to carry functionally significant loss-of-function (LOF) cancer driver mutations at one or more tumor suppressor gene loci (21% of individuals carry 1 mutant allele; 28% carry 2, 24% carry 3, 13% carry 4, 5% carry 5, 2% carry >6). This encompassed 448 LOF mutations (averaging 3.2 LOF mutations per TSG), 420 of which are computationally predicted to be deleterious without experimental validation. In contrast, we found that the frequency of such LOF mutations is higher in random samplings of non-TSGs (Supplementary Table 4) as well as in TSGs from normal tissues in TCGA patients (Supplementary Table 5) consistent with the idea of negative selection against LOF somatic mutations that affect TSGs.

While the frequencies of LOF TSGs we detected are higher than what has been typically reported for specific TSGs [50, 51], they are not unprecedented. For example, among the most intensively studied TSGs is the *RB1* gene that is associated with inherited childhood retinoblastoma [18]. Although the frequency of individuals heterozygous for LOF RB1 alleles ("carriers") in human populations is generally reported to be ≤ 5% [51], considerable variability exists among ethnic groups/populations. For example, in a study of select Asian populations, the frequency of carriers of LOF RB1 alleles was reported to be as high as 34% in specific ethnic populations [52].

The "two-hit" hypothesis proposes that individuals heterozygous for a LOF tumor suppressor allele will not typically develop cancer unless an additional LOF mutation occurs in the gene's functional partner allele [18]. While the two-hit hypothesis has been successfully employed to account for many instances of inherited cancers associated with tumor suppressor genes [53, 54], a number of examples have been identified in recent years that are inconsistent with Knudson's "two-hit" hypothesis [55–57]. For example, it is now known that not all children afflicted with retinoblastoma are homozygous for the LOF RB1 allele [58, 59] and this condition has, in several cases, been

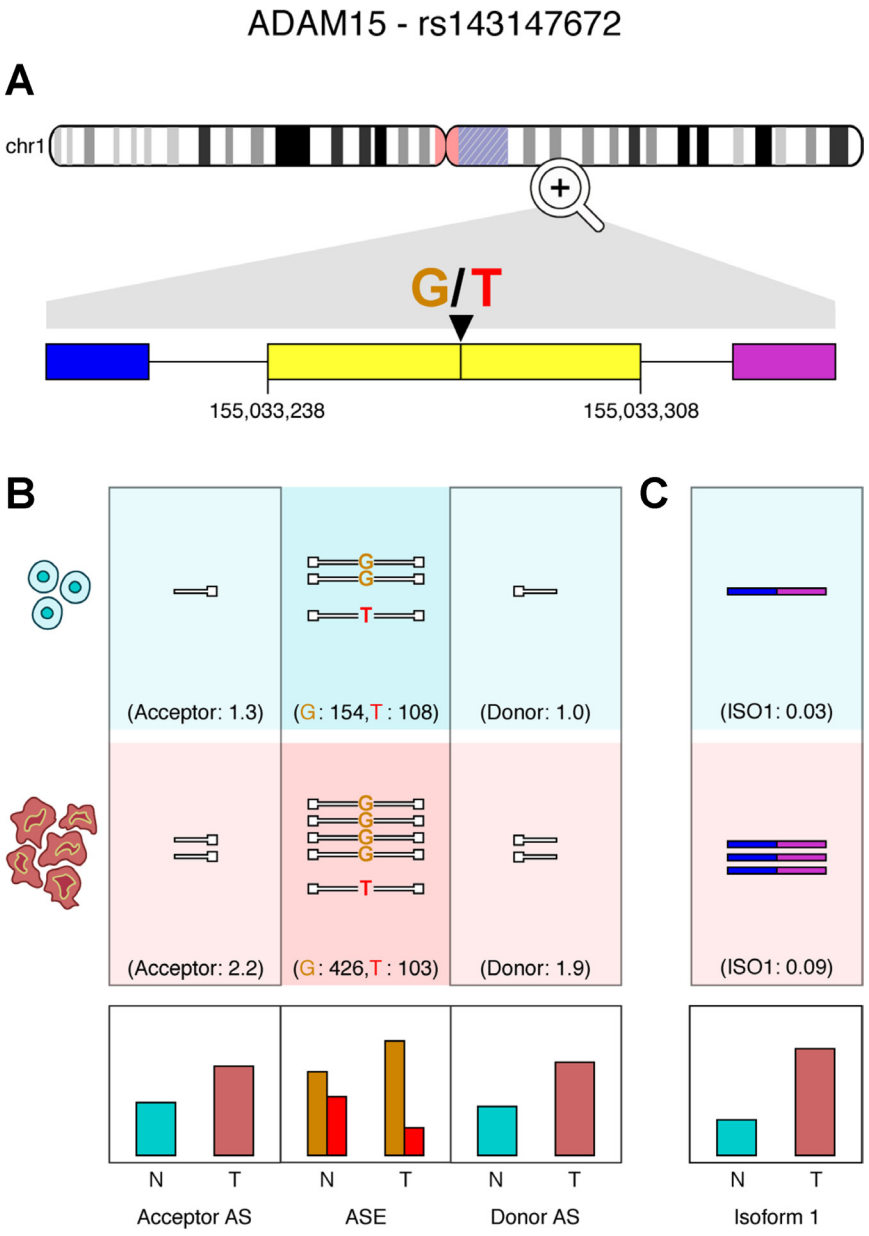

Figure 6: ADAM15 exon skipping correlates with ASE in a breast adenocarcinoma patient. (**A**) An exon-skipping event in exon 19 of ADAM15 in a breast cancer patient (TCGA-BH-A0DT). (**B**) Antisense reads ( ) mapping to donor and acceptor sites are quantified, alongside the ASE locus within the exon (N = normal; T = tumor; AS = antisense RNA). (**C**) Quantification of reads supporting the isoform missing exon 19. Relative expression plots are shown for antisense RNA, ASE and isoforms below.

associated with aberrant expression of unlinked regulatory genes [60]. While evaluating the "two-hit" model in our dataset, we found that < 20% of patients acquired a second LOF mutation in cancer tissues as predicted by the "two-hit" model. This finding is consistent with a growing body of evidence that the mechanisms underlying the contribution of TSGs to cancer onset and progression are often more complex than originally envisioned [61, 62].

A primary goal of our study was to evaluate the potential significance of changes in ASE of TSGs between normal and cancer, and to explore the molecular mechanisms that may underly this process. While searching the TCGA database for evidence of ASE we found that COSMIC census mutations in TSGs display significantly ($P < 3.11 \times 10^{-10}$) more ASE in tumors compared to matched normal tissues in breast, head and neck, and lung cancers. Our finding that this change is not limited to COSMIC genes but extends to genes not previously associated with cancer, implies a general loss of regulatory control in cancer. Evidence for such a global loss in regulatory control in cancer has been previously reported [63, 64]. This difference in ASE between tumors and matched normal tissues was substantially less and not significant within TSGs in thyroid carcinoma. Further research will be required to unequivocally determine the basis for this discrepancy. However, one possibility is that thyroid cancer's inherently low mutation rate [19] allows for its transcriptional regulation to remain more intact.

Of the six possible Patterns of change in ASE between normal and cancer tissues, we found that Pattern 1 (*i. e*., no ASE in normal tissue but expression of mutant allele (alt) > expression of wild-type (ref) allele in cancer tissue) was one of the most commonly observed Patterns across cancer types. This finding is consistent with the hypothesis that LOF TSG alleles may be contributing significantly to cancer onset/progression even in the heterozygous state.

One possible explanation of the observed changes in ASE patterns between normal and cancer tissue is that it is structural in nature, *i. e*., the consequence of differences in allele counts attributable to, for example, loss of heterozygosity (LOH) or the polyclonal heterogeneity characteristic of most tumors [25]. To test this possibility, we compared RNA allele counts with DNA allele counts in the same patient samples. We found that on average, 35% of genes displaying ASE had DNA allele counts that correlated with RNA allele counts. These results are consistent with prior findings indicating that a significant fraction of ASE can be accounted for by underlying differences in DNA allelic content [25]. It should be noted that structural changes introduced by somatic mutations that create premature termination codons and/or induce nonsense-mediated RNA decay could also be a contributing factor [65]. Nevertheless, collectively our results indicate that, at least with respect to our patient samples, differences in ASE between normal and cancer tissues is not merely structurally based but likely attributable to allele-specific differences in gene expression.

Another mechanism of regulatory change of emerging significance in cancer is epigenetics (83). In a preliminary effort to explore the possible contribution of epigenetics to global changes in patterns of ASE, we analyzed methylation data for patient samples from TCGA. We found that only 10% of genes displaying ASE also displayed significant differences (>1.3-fold) in methylation between the normal and tumor tissue samples. Because changes in methylation are generally considered to be a reliable indicator of epigenetic-associated changes in gene expression [66], our results suggest that changes in methylation may not be playing a predominant role in the regulation of ASE in our patient samples.

Allele-specific differences in gene expression may also be attributable to variant *cis*-regulatory sequences located up- or down-stream from the respective alleles' coding regions. Such *cis*-regulatory variation is commonplace and is often identified by utilizing QTL mapping methodologies [67]. We employed the Genotype-Tissue Expression Project's (GTEx) single tissue *cis*-eQTL database to explore the extent to which allele-specific *cis*-regulatory variation may account for ASE in patient samples. We found that only 24% of genes displaying ASE are eGenes, more of which explain ASE in normal (38%) than tumor (21.8%) samples. Moreover, only 1.8% of ASE haplotypes were found to be in phase with an eQTL indicating that *cis*-regulatory variation is not a likely explanation of the majority of instances of ASE in our dataset.

Having failed to identify a mechanism of transcriptional level regulation that could explain the majority of observed instances of ASE in our dataset, we turned our attention to the potential influence of post-transcriptional regulation on ASE. One post-transcriptional mechanism of growing prominence in cancer biology is alternative splicing [68]. The primary RNA products of genes are processed at the post-transcriptional level by alternative RNA splicing resulting in multiple RNA isoforms per gene. If alternate RNA isoforms are generated on an allele-specific basis (allele-specific alternative splicing), it could manifest itself as differences in ASE. To explore the possibility that allele-specific alternative splicing could be contributing to changing patterns of ASE in cancer, we examined isoform counts associated with our TCGA patient data [37]. We found that almost half (46%) of SNPs displaying ASE in patient samples were indeed associated with exon skipping. However, further studies that employ a more granular isoform-specific quantification around ASE loci will be needed to fully understand the workings of an allele-specific alternative splicing mechanism.

While the potential functional significance of alternative splicing in cancer has been long noted [69], the

mechanisms underlying the phenomenon remain poorly understood. Because the genes displaying changes in ASE are not associated with *cis*-regulatory mutations in splice acceptor/donor sites, we focused our attention on possible *trans*-regulatory mechanisms. One possibility is that one or more of the regulatory proteins or RNAs associated with the spliceosome could be mutated or otherwise dysregulated in cancer resulting in aberrant splicing patterns [70]. However, the fact that our observed allele-specific alternative splicing was limited to only a subset of genes suggested that the underlying mechanism was of a more targeted nature.

One possibility was suggested from previous studies showing that splicing events can be experimentally induced *in vivo* by exposing primary transcripts to antisense RNAs that pair with known splice sites in the primary transcript [71, 72]. Indeed, there is growing evidence that *de novo* expression of antisense RNAs may play a significant role in the induction of alternate-splice variants [73] and that this may be a significant factor in cancer onset/progression. Our results are generally consistent with this hypothesis and suggest that allele-specific alternative splicing, possibly mediated by changes in the expression of antisense RNAs, may play a significant role in the induction of changes in ASE patterns in cancer. Further studies inducing ASE *in vitro* via use of antisense oligonucleotides will be needed to validate this hypothesis.

MATERIALS AND METHODS

Cancer associated mutation identification in 1000 genomes population

Using the BEDTools program [74], the genomic locations of all coding mutations in COSMIC census genes (v82) [75] were intersected with a VCF file containing all sequence variants called from the 2,504 individuals of the Phase 3 release of the 1000 Genomes Project (1KGP) [49]. The distribution of these cancer associated mutations was determined for all intersecting mutations including the subset of deleterious mutations. Variant effects were annotated using Variant Effect Predictor (VEP) using the Ensembl 91 release [76]. Mutations were considered to be deleterious if they were non-sense, frameshift, splice acceptor/donor mutations, or whole gene deletion mutations. Missense mutations predicted deleterious by both SIFT [15] and Polyphen2 [16] were also scored as deleterious mutations. Moreover, we removed any mutation that had been labeled as benign or likely benign by ClinVar [77].

Genotyping and variant calling with WXS and variant annotation

Genotyping was implemented from WXS. SAMtools mpileup output was fed to VarScan's mpileup2snp function in order to call variants [78]. Only reads with mapping quality > 14 were counted. Further, to call a variant, a position must have met a minimum read depth of 8, minimum allelic depth of 2 and variant allele frequency threshold of 0.2. The default *p*-value of 0.01 was used for calling variants. Variants were annotated using VEP with the same criteria mentioned as above.

Allele specific expression (ASE) analysis

Counting allele-specific reads

Indexed RNA-Seq BAM files along with filtered heterozygous variants were passed to GATK's ASEReadCounter tool [79]. At this step, only reads with minimum mapping quality and base quality scores of 20 and 30, respectively, were counted. Also, minimum depths of 20 reads per site and four reads per allele were applied. With the aim of inferring biological significance, resulting allele counts were annotated with rsid using Kaviar. Subsequently, gene names associated with particular SNPs were fetched from dbSNP using EDirect [80]. The fraction of reads containing the reference allele over the total number of reads at a given position (Ref Ratio) was calculated for all heterozygous SNPs. Custom scripts were written to perform allele specific expression analysis.

Accounting for mapping bias

When mapping RNA-seq reads to the reference genome, reads overlapping a SNP that contain the alternative allele tend to map less frequently than those reads containing the reference allele. This allelic mapping bias has been well documented and presents challenges in ASE analysis [81]. Degner *et al*. demonstrated that the reliability in ASE estimation is greatly dependent on the capability to control for reference mapping bias [82]. To limit this bias, we first removed sites known to be susceptible to mapping bias. We did so by removing all sites with 50bp mapability < 1 based on the UCSC mapability track [83]. To correct for any residual bias, we calculated the genome-wide allelic ratios for all nucleotide pairs and used them in place of 0.5 as the expected allelic ratio in the binomial test (Supplementary Figure 11) as previously done by Lappalainen *et al* [84].

ASE-analysis

Using the allele counts for every heterozygous position that met our filtering requirements, we performed a binomial test to identify whether the ratio of reference and alternative read counts differed significantly from the corresponding expected proportion between those alleles. Expected ratios were inflated slightly from 0.5 based on the observed allele counts within our population as described in the previous section. We classified a site as an ASE SNP if its binomial *p*-value was less than 0.005

and corrected for a false discovery rate (FDR) of 5%. Gene level ASE was determined by aggregating ASE information from all heterozygous SNPs within a gene as outlined by the MBASED protocol [8]; ASE genes were classified with a major allele frequency (MAF) greater than 0.7 and *p*-value less than 0.05 (FDR 5%). To label significant ASE genes with Patterns we pseudo-phased them by creating a major haplotype consisting of the alleles with higher RNA read counts. If a haplotype contained more reference SNPs it was labelled as the reference haplotype and *vice versa* for the alternative. If the number of reference and alternative SNPs on each haplotype were the same, the haplotype was labelled as ambiguous.

Differences in ASE between normal and cancer tissue groups, were evaluated by comparing the distributions of the proportion of SNPs with ASE within each collection. The statistical significance levels of the observed difference in ASE between normal and tumor tissues for both COSMIC census mutations and all heterozygous SNPs were evaluated by comparing these distributions using the non-parametric Mann-Whitney *U* test.

When comparing SNPs intersecting paired normal and tumor samples, we applied a combined binomial-Fisher test to determine if ASE patterns were significant. Three ASE patterns of interest were analyzed; Pattern 1: No significant difference in ASE (ref=alt) in normal tissues but significant ASE (ref<alt) in cancer tissues; Pattern 2: Significant ASE in normal tissues (ref>alt) but no significant ASE (ref=alt) in cancer tissues; Pattern 3: Significant ASE in normal sample (ref>alt) and significant ASE in tumor sample (ref<alt); Pattern 4: No significant ASE in normal tissues (ref=alt) but significant ASE in cancer tissues (ref>alt); Pattern 5: Significant ASE (ref<alt) in normal tissues but no significant ASE in cancer tissues (ref=alt); Pattern 6: Significant ASE in normal (ref<alt) and in cancer tissues (ref>alt). All Patterns are visualized in Figure 3A. Significant ASE (FDR = 5%, $P < 0.005$) was determined using a binomial test within samples and subsequently a Fisher's exact test ($P < 0.05$) when comparing two samples. Both tests were applied to increase stringency and validity of results.

The analyses used to determine mechanisms of ASE are outlined in detail in the Supplementary File: Materials and Methods.

Second site loss-of-function mutations

Filtered heterozygous sites in tumor suppressor genes (TSGs) of all 233 patients in normal and tumor samples were phased using SHAPEIT [85]. Loss-of-function mutations were defined as stop gained, frameshift, splice acceptor/donor, start lost and stop lost mutations. Deleterious missense mutations predicted to be damaging/deleterious by SIFT [15] and Polyphen2 [16] were also considered loss of function in TSGs. Patients with a secondary site loss-of-function mutation were defined as having a heterozygous mutation in the normal sample and either: 1) the same mutation homozygous in the tumor sample, 2) a new loss of function mutation on the opposite allele in the tumor sample (*i. e.* compound heterozygote), or 3) a DNA segment with loss of allele at the locus in the tumor sample. Segments of DNA with loss of allele were identified using FACETS [86].

Detailed materials and methods are included in the Supplementary File.

CONCLUSIONS

We have shown that LOF TSGs are segregating in human populations at significant frequencies suggesting that many otherwise healthy individuals are at elevated risk of developing cancer. Changes in ASE between normal and cancer tissues indicates that LOF TSG alleles may contribute to cancer onset/progression even when heterozygous with wild-type functional alleles. While a variety of molecular mechanisms were identified as potentially contributing to changes in ASE between normal and cancer, differences in DNA counts and allele-specific alternative splicing emerged as predominant factors.

Abbreviations

TSGs: Tumor suppressor genes; 1KGP: 1000 Genomes Project; ASE: Allele-specific expression; LOF: Loss of function; COSMIC: Catalogue of somatic mutations in cancer; TCGA: The Cancer Genome Atlas.

Author contributions

EAC contributed to the design of the experiment, leadership of the data analysis and in the writing of the manuscript; SK contributed to the design of the experiment, data analysis and writing of Methods; DB contributed to the data analysis and writing of Methods; LW contributed to the data analysis; IKJ contributed to the design of the experiment; and, JFM contributed to the design of the experiment, data analysis and in the writing of the manuscript.

ACKNOWLEDGMENTS

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING

This research was supported by the Ovarian Cancer Institute (Atlanta), Northside Hospital (Atlanta), the Deborah Nash Endowment Fund, and National Institute of Health Bioinformatics Training Grant: CRP 10-2012-03.

REFERENCES

1. Wunderlich V. Early references to the mutational origin of cancer. Int J Epidemiol. 2007; 36:246–247. https://doi.org/10.1093/ije/dyl272. [PubMed]
2. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med. 2004; 10:789–799. https://doi.org/10.1038/nm1087. [PubMed]
3. Knudson AG. Two genetic hits (more or less) to cancer. Nat Rev Cancer. 2001; 1:157–162. https://doi.org/10.1038/35101031. [PubMed]
4. Zhang Y, Yang L, Kucherlapati M, Chen F, Hadjipanayis A, Pantazi A, Bristow CA, Lee EA, Mahadeshwar HS, Tang J, Zhang J, Seth S, Lee S, et al. A pan-cancer compendium of genes deregulated by somatic genomic rearrangement across more than 1,400 cases. Cell Reports. 2018; 24:515–527. https://doi.org/10.1016/j.celrep.2018.06.025. [PubMed]
5. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013; 152:1237–1251. https://doi.org/10.1016/j.cell.2013.02.014. [PubMed]
6. Michalak EM, Burr ML, Bannister AJ, Dawson MA. The roles of DNA, RNA and histone methylation in ageing and cancer. Nat Rev Mol Cell Biol. 2019; 20:573–589. https://doi.org/10.1038/s41580-019-0143-1. [PubMed]
7. Ongen H, Andersen CL, Bramsen JB, Oster B, Rasmussen MH, Ferreira PG, Sandoval J, Vidal E, Whiffin N, Planchon A, Padioleau I, Bielser D, Romano L, et al. Putative cis-regulatory drivers in colorectal cancer. Nature. 2014; 512:87–90. https://doi.org/10.1038/nature13602. [PubMed]
8. Mayba O, Gilbert HN, Liu J, Haverty PM, Jhunjhunwala S, Jiang Z, Watanabe C, Zhang Z. MBASED: allele-specific expression detection in cancer tissues and cell lines. Genome Biol. 2014; 15:405. https://doi.org/10.1186/s13059-014-0405-3. [PubMed]
9. Buckland PR. Allele-specific gene expression differences in humans. Hum Mol Genet. 2004; 13:R255–R260. https://doi.org/10.1093/hmg/ddh227. [PubMed]
10. Sigurdsson MI, Saddic L, Heydarpour M, Chang TW, Shekar P, Aranki S, Couper GS, Shernan SK, Seidman JG, Body SC, Muehlschlegel JD. Allele-specific expression in the human heart and its application to postoperative atrial fibrillation and myocardial ischemia. Genome Med. 2016; 8:127. https://doi.org/10.1186/s13073-016-0381-1. [PubMed]
11. Liu Z, Dong X, Li Y. A genome-wide study of allelespecific expression in colorectal cancer. Front Genet. 2018; 9:570. https://doi.org/10.3389/fgene.2018.00570. [PubMed]
12. Sherr CJ. Principles of tumor suppression. Cell. 2004; 116:235–246. https://doi.org/10.1016/S0092-8674(03)01075-4. [PubMed]
13. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015; 43:D805–D811. https://doi.org/10.1093/nar/gku1075. [PubMed]
14. Flanagan SE, Patch AM, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomarkers. 2010; 14:533–537. https://doi.org/10.1089/gtmb.2010.0036. [PubMed]
15. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003; 31:3812–3814. https://doi.org/10.1093/nar/gkg509. [PubMed]
16. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013; Chapter 7:Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76. [PubMed]
17. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, and Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45:1113–1120. https://doi.org/10.1038/ng.2764. [PubMed]
18. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A. 1971; 68:820–823. https://doi.org/10.1073/pnas.68.4.820. [PubMed]
19. Khatami F, Tavangar SM. A review of driver genetic alterations in thyroid cancers. Iran J Pathol. 2018; 13:125–135. https://doi.org/10.30699/ijp.13.2.125. [PubMed]
20. Liu Z, Gui T, Wang Z, Li H, Fu Y, Dong X, Li Y. cisASE: a likelihood-based method for detecting putative cis-regulated allele-specific expression in RNA sequencing data. Bioinformatics. 2016; 32:3291–3297. https://doi.org/10.1093/bioinformatics/btw416. [PubMed]
21. Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. PLoS One. 2011; 6: e14643. https://doi.org/10.1371/journal.pone.0014643. [PubMed]
22. Brandt DY, Aguiar VR, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. G3 (Bethesda). 2015; 5:931–941. https://doi.org/10.1534/g3.114.015784. [PubMed]
23. Gao C, Devarajan K, Zhou Y, Slater CM, Daly MB, Chen X. Identifying breast cancer risk loci by global differential allele-specific expression (DASE) analysis

in mammary epithelial transcriptome. BMC Genomics. 2012; 13:570. https://doi.org/10.1186/1471-2164-13-570. [PubMed]

24. Daelemans C, Ritchie ME, Smits G, Abu-Amero S, Sudbery IM, Forrest MS, Campino S, Clark TG, Stanier P, Kwiatkowski D. High-throughput analysis of candidate imprinted genes and allele-specific gene expression in the human term placenta. BMC Genet. 2010; 11:25. https://doi.org/10.1186/1471-2156-11-25. [PubMed]

25. Tuch BB, Laborde RR, Xu X, Gu J, Chung CB, Monighetti CK, Stanley SJ, Olsen KD, Kasperbauer JL, Moore EJ, Broomer AJ, Tan R, Brzoska PM, et al. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. PLoS One. 2010; 5:e9317. https://doi.org/10.1371/journal.pone.0009317. [PubMed]

26. Hasin-Brumshtein Y, Hormozdiari F, Martin L, van Nas A, Eskin E, Lusis AJ, Drake TA. Allele-specific expression and eQTL analysis in mouse adipose tissue. BMC Genomics. 2014; 15:471. https://doi.org/10.1186/1471-2164-15-471. [PubMed]

27. Pastinen T, Hudson TJ. Cis-acting regulatory variation in the human genome. Science. 2004; 306:647–650. https://doi.org/10.1126/science.1101659. [PubMed]

28. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev Genet. 2010; 11:533–538. https://doi.org/10.1038/nrg2815. [PubMed]

29. Aguet F, Brown AA, Castel S, Davis JR, Mohammadi P, Segre AV, Zappala Z, Abell NS, Fresard L, Gamazon ER. Local genetic effects on gene expression across 44 human tissues. bioRxiv. 2016. https://doi.org/10.1101/074450.

30. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015; 16:197–212. https://doi.org/10.1038/nrg3891. [PubMed]

31. GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet. 2013; 45:580–585. https://doi.org/10.1038/ng.2653. [PubMed]

32. Wagner JR, Ge B, Pokholok D, Gunderson KL, Pastinen T, Blanchette M. Computational analysis of whole-genome differential allelic expression data in human. PLoS Comput Biol. 2010; 6:e1000849. https://doi.org/10.1371/journal.pcbi.1000849. [PubMed]

33. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011; 25:1010–1022. https://doi.org/10.1101/gad.2037511. [PubMed]

34. Siegfried Z, Eden S, Mendelsohn M, Feng X, Tsuberi BZ, Cedar H. DNA methylation represses transcription *in vivo*. Nat Genet. 1999; 22:203–206. https://doi.org/10.1038/9727. [PubMed]

35. Bird AP, Wolffe AP. Methylation-induced repression—belts, braces, and chromatin. Cell. 1999; 99:451–454. https://doi.org/10.1016/S0092-8674(00)81532-9. [PubMed]

36. Romanel A. Allele-specific expression analysis in cancer using next-generation sequencing data. Methods Mol Biol. 2019; 1878:125–137. https://doi.org/10.1007/978-1-4939-8868-6_7. [PubMed]

37. Kahles A, Lehmann KV, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, Stegle O, Kohlbacher O, Sander C, Caesar-Johnson SJ. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. Cancer Cell. 2018; 34:21–24.e6. https://doi.org/10.1016/j.ccell.2018.07.001. [PubMed]

38. Jayasinghe RG, Cao S, Gao Q, Wendl MC, Vo NS, Reynolds SM, Zhao Y, Climente-González H, Chai S, Wang F, Varghese R, Huang M, Liang WW, et al. Systematic analysis of splice-site-creating mutations in cancer. Cell Rep. 2018; 23:270–81. e3. https://doi.org/10.1016/j.celrep.2018.03.052. [PubMed]

39. Morrissy AS, Griffith M, Marra MA. Extensive relationship between antisense transcription and alternative splicing in the human genome. Genome Res. 2011; 21:1203–1212. https://doi.org/10.1101/gr.113431.110. [PubMed]

40. McClorey G, Fall AM, Moulton HM, Iversen PL, Rasko JE, Ryan M, Fletcher S, Wilton SD. Induced dystrophin exon skipping in human muscle explants. Neuromuscul Disord. 2006; 16:583–590. https://doi.org/10.1016/j.nmd.2006.05.017. [PubMed]

41. Lee HD, Koo BH, Kim YH, Jeon OH, Kim DS. Exosome release of ADAM15 and the functional implications of human macrophage-derived ADAM15 exosomes. FASEB J. 2012; 26:3084–3095. https://doi.org/10.1096/fj.11-201681. [PubMed]

42. Ortiz RM, Karkkainen I, Huovila AP. Aberrant alternative exon use and increased copy number of human metalloprotease-disintegrin ADAM15 gene in breast cancer cells. Genes Chromosomes Cancer. 2004; 41:366–378. https://doi.org/10.1002/gcc.20102. [PubMed]

43. Salvador F, Martin A, López-Menéndez C, Moreno-Bueno G, Santos V, Vázquez-Naharro A, Santamaria PG, Morales S, Dubus PR, Muinelo-Romay L, López-López R, Tung JC, Weaver VM, et al. Lysyl oxidase–like protein LOXL2 promotes lung metastasis of breast cancer. Cancer Res. 2017; 77:5846–5859. https://doi.org/10.1158/0008-5472.CAN-16-3152. [PubMed]

44. Wu L, Zhu Y. The function and mechanisms of action of LOXL2 in cancer (Review). Int J Mol Med. 2015; 36:1200–1204. https://doi.org/10.3892/ijmm.2015.2337. [PubMed]

45. da Silva MR, Moreira GA, Gonçalves da Silva RA, de Almeida Alves Barbosa É, Pais Siqueira R, Teixera RR, Almeida MR, Silva Júnior A, Fietto JL, Bressan GC. Splicing regulators and their roles in cancer biology and therapy. Biomed Res Int. 2015; 2015:150514. https://doi.org/10.1155/2015/150514. [PubMed]

46. Tsunoda T, Inada H, Kalembeyi I, Imanaka-Yoshida K, Sakakibara M, Okada R, Katsuta K, Sakakura T, Majima Y, Yoshida T. Involvement of large Tenascin-C splice variants in breast cancer progression. Am J Pathol.

2003; 162:1857–1867. https://doi.org/10.1016/S0002-9440(10)64320-9. [PubMed]

47. Guttery DS, Shaw JA, Lloyd K, Pringle JH, Walker RA. Expression of tenascin-C and its isoforms in the breast. Cancer Metastasis Rev. 2010; 29:595–606. https://doi.org/10.1007/s10555-010-9249-9. [PubMed]

48. Hancox RA, Allen MD, Holliday DL, Edwards DR, Pennington CJ, Guttery DS, Shaw JA, Walker RA, Pringle JH, Jones JL. Tumour-associated tenascin-C isoforms promote breast cancer cell invasion and growth by matrix metalloproteinase-dependent and independent mechanisms. Breast Cancer Res. 2009; 11:R24. https://doi.org/10.1186/bcr2251. [PubMed]

49. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stutz AM, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015; 526:75–81. https://doi.org/10.1038/nature15394. [PubMed]

50. Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb Perspect Biol. 2010; 2: a001008. https://doi.org/10.1101/cshperspect.a001008. [PubMed]

51. Lesueur F, Song H, Ahmed S, Luccarini C, Jordan C, Luben R, Easton DF, Dunning AM, Pharoah PD, Ponder BA. Single-nucleotide polymorphisms in the RB1 gene and association with breast cancer in the British population. Br J Cancer. 2006; 94:1921–1926. https://doi.org/10.1038/sj.bjc.6603160. [PubMed]

52. Kadam-Pai P, Su XY, Miranda JJ, Soemantri A, Saha N, Heng CK, Lai PS. Ethnic variations of a retinoblastoma susceptibility gene (RB1) polymorphism in eight Asian populations. J Genet. 2003; 82:33–37. https://doi.org/10.1007/BF02715879. [PubMed]

53. Knudson AG Jr, Strons LC. Mutation and cancer: a model for Wilms' tumor of the kidney. J Natl Cancer Inst. 1972; 48:313–324. [PubMed]

54. Knudson AG Jr, Strong LC. Mutation and cancer: neuroblastoma and pheochromocytoma. Am J Hum Genet. 1972; 24:514. [PubMed]

55. Berger AH, Knudson AG, Pandolfi PP. A continuum model for tumour suppression. Nature. 2011; 476:163–169. https://doi.org/10.1038/nature10275. [PubMed]

56. Paige AJ. Redefining tumour suppressor genes: exceptions to the two-hit hypothesis. Cell Mol Life Sci. 2003; 60:2147–2163. https://doi.org/10.1007/s00018-003-3027-6. [PubMed]

57. Tucker T, Friedman JM. Pathogenesis of hereditary tumors: beyond the "two-hit" hypothesis. Clin Genet. 2002; 62:345–357. https://doi.org/10.1034/j.1399-0004.2002.620501.x. [PubMed]

58. Tomar S, Sethi R, Sundar G, Quah TC, Quah BL, Lai PS. Mutation spectrum of RB1 mutations in retinoblastoma cases from Singapore with implications for genetic management and counselling. PLoS One. 2017; 12:e0178776. https://doi.org/10.1371/journal.pone.0178776. [PubMed]

59. Dommering CJ, Mol BM, Moll AC, Burton M, Cloos J, Dorsman JC, Meijers-Heijboer H, van der Hout AH. RB1 mutation spectrum in a comprehensive nationwide cohort of retinoblastoma patients. J Med Genet. 2014; 51:366–374. https://doi.org/10.1136/jmedgenet-2014-102264. [PubMed]

60. Rushlow DE, Mol BM, Kennett JY, Yee S, Pajovic S, Thériault BL, Prigoda-Lee NL, Spencer C, Dimaras H, Corson TW. Characterisation of retinoblastomas without RB1 mutations: genomic, gene expression, and clinical studies. Lancet Oncol. 2013; 14:327–334. https://doi.org/10.1016/S1470-2045(13)70045-7. [PubMed]

61. Slattery ML, Herrick JS, Mullany LE, Samowitz WS, Sevens JR, Sakoda L, Wolff RK. The co-regulatory networks of tumor suppressor genes, oncogenes, and miRNAs in colorectal cancer. Genes Chromosomes Cancer. 2017; 56:769–787. https://doi.org/10.1002/gcc.22481. [PubMed]

62. Sung J, Turner J, McCarthy S, Enkemann S, Li CG, Yan P, Huang T, Yeatman TJ. Oncogene regulation of tumor suppressor genes in tumorigenesis. Carcinogenesis. 2005; 26:487–494. https://doi.org/10.1093/carcin/bgh318. [PubMed]

63. Goodarzi H, Elemento O, Tavazoie S. Revealing global regulatory perturbations across human cancers. Mol Cell. 2009; 36:900–911. https://doi.org/10.1016/j.molcel.2009.11.016. [PubMed]

64. Cordero D, Sole X, Crous-Bou M, Sanz-Pamplona R, Pare-Brunet L, Guino E, Olivares D, Berenguer A, Santos C, Salazar R, Biondo S, Moreno V. Large differences in global transcriptional regulatory programs of normal and tumor colon cells. BMC Cancer. 2014; 14:708. https://doi.org/10.1186/1471-2407-14-708. [PubMed]

65. Lindeboom RG, Supek F, Lehner B. The rules and impact of nonsense-mediated mRNA decay in human cancers. Nat Genet. 2016; 48:1112–1118. https://doi.org/10.1038/ng.3664. [PubMed]

66. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003; 33:245–254. https://doi.org/10.1038/ng1089. [PubMed]

67. Bickel RD, Kopp A, Nuzhdin SV. Composite effects of polymorphisms near multiple regulatory elements create a major-effect QTL. PLoS Genet. 2011; 7:e1001275. https://doi.org/10.1371/journal.pgen.1001275. [PubMed]

68. Escobar-Hoyos L, Knorr K, Abdel-Wahab O. Aberrant RNA Splicing in cancer. Annual Review of Cancer Biology. 2019; 3:167–185. https://doi.org/10.1146/annurev-cancerbio-030617-050407.

69. Oltean S, Bates DO. Hallmarks of alternative splicing in cancer. Oncogene. 2013; 33:5311–5318. https://doi.org/10.1038/onc.2013.533. [PubMed]

70. El Marabti E, Younis I. The Cancer Spliceome: reprograming of alternative splicing in cancer. Front

Mol Biosci. 2018; 5:80. https://doi.org/10.3389/fmolb.2018.00080. [PubMed]

71. Sazani P, Kole R. Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing. J Clin Invest. 2003; 112:481–486. https://doi.org/10.1172/JCI200319547. [PubMed]

72. Havens MA, Hastings ML. Splice-switching antisense oligonucleotides as therapeutic drugs. Nucleic Acids Res. 2016; 44:6549–6563. https://doi.org/10.1093/nar/gkw533. [PubMed]

73. Lieberger B, Pinon Hofbauer J, Wally V, Arzt C, Hainzl S, Kocher T, Murauer EM, Bauer JW, Reichelt J, Koller U. RNA trans-splicing modulation via antisense molecule interference. Int J Mol Sci. 2018; 19:762. https://doi.org/10.3390/ijms19030762. [PubMed]

74. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014; 47:11.12.1–34. https://doi.org/10.1002/0471250953.bi1112s47. [PubMed]

75. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018; 18:696–705. https://doi.org/10.1038/s41568-18-0060-1. [PubMed]

76. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17:122. https://doi.org/10.1186/s13059-016-0974-4. [PubMed]

77. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014; 42:D980–D985. https://doi.org/10.1093/nar/gkt1113. [PubMed]

78. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012; 22:568–576. https://doi.org/10.1101/gr.129684.111. [PubMed]

79. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; 43:10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43. [PubMed]

80. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2011; 39:D52–D57. https://doi.org/10.1093/nar/gkq1237. [PubMed]

81. Hodgkinson A, Grenier JC, Gbeha E, Awadalla P. A haplotype-based normalization technique for the analysis and detection of allele specific expression. BMC Bioinformatics. 2016; 17:364. https://doi.org/10.1186/s12859-016-1238-8. [PubMed]

82. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009; 25:3207–3212. https://doi.org/10.1093/bioinformatics/btp579. [PubMed]

83. Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, Ribeca P. Fast computation and applications of genome mappability. PLoS One. 2012; 7:e30377. https://doi.org/10.1371/journal.pone.0030377. [PubMed]

84. Lappalainen T, Sammeth M, Friedlander MR. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501:506–511. https://doi.org/10.1038/nature12531. [PubMed]

85. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011; 9:179–181. https://doi.org/10.1038/nmeth.1785. [PubMed]

86. Shen R, Seshan V. FACETS: Allele-specific copy number and clonal heterogeneity analysis tool estimates for highthroughput DNA sequencing. Nucleic Acids Res. 2016; 44:e131. https://doi.org/10.1093/nar/gkw520. [PubMed]