doi:10.1093/nar/gks280

Identification of allele-specific alternative mRNA processing via transcriptome sequencing

Gang Li[1], Jae Hoon Bahn[1], Jae-Hyung Lee[1], Guangdun Peng[1], Zugen Chen[2], Stanley F. Nelson[2,3,4] and Xinshu Xiao[1,4,*]

[1]Department of Integrative Biology and Physiology, [2]Department of Human Genetics, [3]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine and [4]Molecular Biology Institute, University of California Los Angeles, Los Angeles, CA 90095, USA

Received November 16, 2011; Revised March 2, 2012; Accepted March 14, 2012

ABSTRACT

Establishing the functional roles of genetic variants remains a significant challenge in the post-genomic era. Here, we present a method, allele-specific alternative mRNA processing (ASARP), to identify genetically influenced mRNA processing events using transcriptome sequencing (RNA-Seq) data. The method examines RNA-Seq data at both single-nucleotide and whole-gene/isoform levels to identify allele-specific expression (ASE) and existence of allele-specific regulation of mRNA processing. We applied the methods to data obtained from the human glioblastoma cell line U87MG and primary breast cancer tissues and found that 26–45% of all genes with sufficient read coverage demonstrated ASE, with significant overlap between the two cell types. Our methods predicted potential mechanisms underlying ASE due to regulations affecting either whole-gene-level expression or alternative mRNA processing, including alternative splicing, alternative polyadenylation and alternative transcriptional initiation. Allele-specific alternative splicing and alternative polyadenylation may explain ASE in hundreds of genes in each cell type. Reporter studies following these predictions identified the causal single nucleotide variants (SNVs) for several allele-specific alternative splicing events. Finally, many genes identified in our study were also reported as disease/phenotype-associated genes in genome-wide association studies. Future applications of our approach may provide ample insights for a better understanding of the genetic basis of gene regulation underlying phenotypic diversity and disease mechanisms.

INTRODUCTION

Recent advances in sequencing technologies have enabled an extraordinary expansion of the catalogs of genetic variants in disease genomes or across populations. However, significant challenges still exist in establishing the functional roles of such variants. To date, only a minority of genetic variants identified by genome-wide association studies (GWASs) elicits protein-coding changes. A large number of variants are expected to influence *cis*-regulation of gene expression (1). Thus far, the most common approach used to predict regulatory variants is the method of expression quantitative trait loci (eQTL) mapping (1). In this approach, massive-scale parallel expression assays are required to identify statistical associations between genotypes and gene expression in populations with a diverse genetic background (2,3). Such studies often focus on the association between genetic variants and whole-gene expression levels, without differentiating isoforms resulted from alternative mRNA processing.

Allele-specific expression (ASE) is an attractive alternative method to infer the existence of *cis*-acting regulatory variants (4). In an ASE study, the relative proportion of mRNA expression levels of two alleles of a heterozygous variant is measured in the same cellular environment within the same subject (4,5). Thus, a major advantage of the method is that the alternative alleles serve as within-sample controls of each other, eliminating environmental or *trans*-acting influences that alter gene expression and making it optimal for detecting *cis*-acting differences. If the regulatory variants are located in intronic or untranscribed regions, those in the mRNAs may serve as markers for the existence of causal variants. Identification of autosomal ASE might be the most direct method to identify functional *cis*-regulation, which can be followed-up by detailed experimental analyses.

[*]To whom correspondence should be addressed. Tel: +1 310 206 6522; Fax: +1 310 206 9184; Email: gxxiao@ucla.edu

However, observation of ASE in a gene does not normally suggest which type of *cis*-regulatory mechanism is responsible for ASE but rather that such mechanisms exist. *Cis*-acting regulation by genetic variants may affect different aspects of gene expression, e.g. transcription, alternative mRNA processing or mRNA stability. Genetic control of transcription often results in changes in whole-gene expression levels, which have been the focus of many eQTL studies. Other mechanisms, such as alternative mRNA processing, were much less often examined despite their known importance in example genes (6–8). Results from large-scale exon array studies showed that genetic influence on alternative mRNA processing could add remarkable complexity to molecular diversity (9). However, investigations of such relationships using microarrays normally require a large number of subjects and arrays to ensure statistical power.

Here, we present methods to analyze transcriptome sequencing (RNA-Seq) data and demonstrate that data of a single subject enabled identification of many genes and alternatively processed regions that are under genetic influence. RNA-Seq provides concurrent allelic and gene expression data. Thus, it allows expression analyses at different levels including single-nucleotide, alternatively processed mRNA isoforms and whole-gene levels. Integrative analysis of the information at multiple levels allows in-depth understanding of the transcriptome, an advantage of RNA-Seq rarely exploited in previous work. This advantage enabled us to develop pipelines to first identify ASE patterns followed by inference of their potential involvement in alternative mRNA processing including alternative splicing, alternative 3′ processing and alternative 5′ initiation. We applied this method to two human cancer data sets, our in-house RNA-Seq data obtained from the glioblastoma cell line U87MG and a public RNA-Seq data set from a breast cancer patient. Our study demonstrated that ASE analysis of individual samples via RNA-Seq can provide substantial insights about the genetic control of gene expression, a potentially much more cost-effective approach than existing methods relying on massive-scale parallel expression assays of a large number of samples.

MATERIALS AND METHODS

Cell culture, RNA purification and RNA-Seq data acquisition

U87MG cells were purchased from American Type Culture Collection (ATCC) and maintained in DMEM high glucose medium supplemented with pyruvate, L-glutamine and 10% fetal bovine serum (FBS) (Hyclone). Total RNA was isolated using the mirVana kit (Ambion), according to the manufacturer's instructions. We used the standard Illumina protocol to prepare libraries for RNA-Seq (http://www.illumina.com/support/documentation.ilmn). Briefly, 10 µg total RNA was first processed via poly-A selection and fragmentation. We generated first-strand cDNA using random hexamer-primed reverse transcription and subsequently used it to generate second-strand cDNA using RNase H and DNA polymerase. Sequencing adapters were ligated using the Illumina Paired-End sample prep kit. Fragments of ~200 bp were isolated by gel electrophoresis, amplified by 15 cycles of PCR and sequenced on the Illumina Genome Analyzer IIx (Cofactor Genomics) in the paired-end sequencing mode (2 × 60 nt reads).

RNA-Seq reads mapping

The same mapping methods as in our previous work (10) were used. Briefly, reads were mapped to the human genome and Ensembl-defined transcriptome using multiple tools including Bowtie (11), BLAT (12) and Tophat (13). Two reads in a pair were mapped separately. Alignments of a read with more than 12 mismatches were discarded. Read pairs were then examined for uniqueness and correct pairing. A uniquely mapped pair was required to have less than six mismatches on each read and not to map to anywhere else in the genome as a pair with less than or equal to 12 mismatches each. Since the genomic locations of heterozygous single nucleotide variants (SNVs) were provided by whole-genome sequencing of U87MG (14), we corrected the number of mismatches in reads harboring the non-reference allele of an SNV such that reads with SNVs were treated without a bias. Only uniquely paired reads were used for subsequent analyses. In addition, we removed all duplicate reads (those mapped to the same genomic locations as a pair) except the one with the best quality score in the mismatch positions (if any).

Identification of ASE of SNVs

For each heterozygous SNV, we first obtained the number of RNA-Seq reads mapped to its alleles. Since the first read position was observed to have relatively large sequencing errors in our data, we excluded reads whose SNVs were located at the first nucleotide. We then calculated the allelic ratio defined as the number of reads mapped to the reference allele divided by the total number of reads covering an SNV. To identify ASE patterns, we used the Chi-square Goodness-of-Fit test to determine if the allelic ratio deviates from the expected ratio 0.5 (i.e. when the two alleles are equally expressed). SNVs were excluded if they are potentially in regions with copy number variants determined by the read depth of the genome sequencing data (14,15). In this analysis, only SNVs with at least 20 RNA-Seq reads were included to reach adequate statistical power (see 'Results' section). Significant ASE patterns were determined using a false discovery rate (FDR) cutoff of 5% based on a modified Benjamini–Hochberg method (16,17) to account for possible correlations of ASE patterns in a gene. The FDR was also estimated using biological replicates (see 'Results' section) or an explicit simulation procedure. In this procedure, for each heterozygous SNV location, we randomly assigned each mapped read in the data set to either allele (with equal probability). Following this randomization, the ASE patterns were identified as described above and an FDR was calculated.

Identification of allele-specific alternative mRNA processing

To elucidate the type of *cis*-regulatory mechanisms potentially involved in generating the ASE patterns, we designed a pipeline, namely ASARP (allele-specific alternative mRNA processing), that synergistically analyzes expression profiles of genes, alternative exons, alternative UTR regions and SNVs obtained from the same RNA-Seq data set (Supplementary Figure S1 and Supplementary Methods). This method allowed us to determine if the *cis*-regulatory mechanisms influencing a gene belonged to one of the following categories: allele-specific whole-gene-level regulation, allele-specific alternative splicing (ASAS), allele-specific alternative polyadenylation (ASAP) and allele-specific transcriptional initiation (ASTI). Whole-gene-level regulation can be simply distinguished from the other categories because such regulation affects expression of all SNVs in a gene. In contrast, the latter three types of regulation act upon local regions within mRNAs and thus only affect SNVs in nearby regions. As a result, genes under whole-gene-level regulation were defined as those in which SNVs with enough power were all identified with ASE patterns. To define genes belonging to one of the categories related to ASARP, we required that (i) the SNV of interest (target SNV) is located in a region (exon or UTR) that is alternatively processed as evidenced in the RNA-Seq data, EST or mRNA data or public databases of gene definitions; (ii) there exists in the same gene non-ASE SNVs with adequate statistical power; (iii) the allelic ratio of the target SNV is significantly different from that of the non-ASE SNV. These criteria exclude the possibility that ASE was resulted from more global regulation other than the local mRNA processing events (alternative splicing or alternative UTR generation). Note that the above filters do not establish causality of the SNV in inducing the observed ASE pattern. Details of this method are described in Supplementary Methods.

To estimate the FDR among genes identified with ASAS, ASAP, ASTI or gene-level regulation, we generated randomized read counts for all SNVs (with N reads $\geq$ 1) in a gene, similarly as for the FDR estimation of ASE. This procedure controls for the read coverage of each gene and SNV and maintains the read distribution in alternative and constitutively processed regions. Using the randomized read counts, the same framework described above was applied and an FDR was estimated for each category of events.

Splicing reporter assays

Genomic regions surrounding candidate SNVs associated with ASAS were amplified by PCR using Taq 2× Master Mix (NEB) with 100 ng of U87MG genomic DNA according to the manufacture's instruction. Primers were designed such that the amplification product spans the associated exonic region and ~250 nt intronic regions on either side. The primers contained HindIII or SacII restriction site for subcloning purposes. Genotypes of SNVs were confirmed via Sanger sequencing of the PCR products (Genewiz) that were sub-cloned into the pZW1 splicing reporter containing cloning sites between two green fluorescent protein (GFP) exons (18). Final constructs were sequenced to ensure that a pair of plasmids containing the two alternative alleles of the SNV was obtained.

The splicing reporter constructs were transfected into 5×10^4 U87MG cells per well in 6-well plates using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instruction. The transfected cells were incubated for 48 h before total RNAs were isolated using RNeasy Plus mini kit (Qiagen). Reverse transcription was performed using 500 ng of total RNA and the SuperScript III first-strand synthesis kit (Invitrogen). PCR was then carried out using 1/20th of the cDNA product, primers targeting the two GFP exons in the splicing reporter, and Taq 2× Master Mix (NEB) with 250 pmol of Cy5-dCTP (GE Healthcare). Quarter amount of PCR products was separated on 6% native polyacrylamide gel. Fluorescence scanning was conducted using a Typhoon 9400 imager (GE Healthcare). Expression levels of splicing isoforms were estimated using the ImageQuant software (GE Healthcare). Inclusion level (% inclusion) of the studied exon was calculated as the intensity ratio of upper/(upper + lower) bands. To visualize the DNA size marker, the scanned gel was post-stained with 1× SYBR Safe DNA gel stain in 1× TBE buffer (Invitrogen).

RESULTS

Mapping of RNA-Seq reads

We obtained paired-end RNA-Seq data using the Illumina Genome Analyzer IIx platform and standard RNA-Seq protocols for four biological replicates of U87MG RNA. A total of ~108 million pairs of reads (2 × 60 nt in length) were acquired. We used a read-mapping strategy as developed in our previous work (10) to achieve unbiased mapping of the reads expressing variant bases relative to the reference genome. This method applied a 'double-filtering' scheme to examine mismatches in the reads relative to the references to remove mapping errors due to the existence of highly homologous regions in the mammalian genome. Such errors may create false positive predictions of allele-specifically expressed SNVs and/or biased estimates of the allelic ratios.

About 53 million (~49%) were mapped uniquely and their distributions in exons, introns and intergenic regions are shown in Table 1. In addition, a small fraction of reads (2%) were mapped to intronic regions where there is clear evidence of a novel exon (Supplementary Methods). Compared to traditional mapping where two mismatches are allowed on each read, approximately 16 million pairs of reads (15% of all) were discarded due to the stringent filters for mismatches and uniqueness in mapping.

In the U87MG genome, we extracted a total of 1 116 235 heterozygous SNVs based on high-throughput genome sequencing (14). Among these SNVs, 33 122 (3.0% of all) were located within Ensembl-annotated exons (either coding or non-coding), of which 17 205 (51.9%) were covered by at least one RNA-Seq read in our data. As expected, coverage of heterozygous SNVs

Table 1. RNA-Seq read-mapping results

Raw reads	Total, *N*	Unique pairs *n* (%)	Multiple pairs *n* (%)	Low-quality pairs *n* (%)	No pairs *n* (%)	Unmapped reads
	107 626 587	53 162 291 (49)	43 538 740 (40)	1 637 731 (2)	3 312 997 (3)	5 974 828 (6)
Unique pairs	**Total, *N***	**Exons *n* (%)**	**Exon–exon junctions *n* (%)**	**Introns *n* (%)**	**Intergenic regions *n* (%)**	**Novel exons *n* (%)**
	53 162 291	34 295 289 (65)	11 221 489 (21)	5 336 129 (10)	1 213 980 (2)	1 095 404 (2)

Number of pairs of reads is shown in each category. Unique pairs: reads mapped uniquely as a pair; multiple pairs: read pairs mapped to multiple genomic locations; low-quality pairs: best mapping results containing more than five mismatches on either read; no pairs: no valid pairing found for the read pair; unmapped: one or both reads were unmappable. Only reads in the 'Unique pairs' category were used for further analyses. Distribution of such reads in different genomic regions is shown. Known gene structures were defined by combining annotations in Ensembl, RefSeq, UCSC, Gencode and Vega genes. Novel exons were identified using our in-house algorithms (Supplementary Methods).

was dependent on the total number of mapped reads and the expression levels of corresponding genes (Supplementary Figure S2). For genes with a minimum expression level of 1 RPKM [reads per kilobase of exon model per million mapped reads (19)], 13 292 (84.7%) of all 15 686 heterozygous SNVs in these genes were covered by at least one RNA-Seq read.

Evaluation of the mapping results

Since the U87MG cells were derived from a female patient, we expected to observe patterns of X-inactivation (as the cells are monoclonal). We identified 22 heterozygous SNVs on the X chromosome that were associated with at least two reads. Twenty of them had monoallelic expression (i.e. all reads mapped exclusively to one of the two alleles). If we assume that SNVs with reads mapped to both alleles were associated with mapping errors, then 3 out of 2010 reads (0.15%) were mapped incorrectly.

In analyzing ASE of genetic variants in RNA-Seq reads, previous work observed that significant bias exists in the read-mapping results that favors reads harboring the reference allele of heterozygous SNVs (4,20–22). To evaluate whether such bias exists in our mapping, we examined the allelic ratios (defined as the number of reads with the reference allele divided by the total number of reads per SNV) of heterozygous SNVs (Figure 1A). In the absence of mapping bias, the average allelic ratio is expected to be 0.5 assuming ASE is only present in a small fraction of SNVs. As shown in Figure 1A, our results confirmed an average allelic ratio of 0.5, supporting the effectiveness of the mapping strategy. In contrast, if read mapping were carried out by allowing two mismatches on each read, as in traditional methods, a statistically significant bias toward the reference allele was detected (Supplementary Figure S3A). Note that the local peaks in Figure 1A at allelic ratios of about 0.33 and 0.66 were due to the prevalence of SNVs with low read coverage (specifically, with 1:2 or 2:1 read counts for the two alleles). The corresponding peaks were not observed if SNVs with three reads in total were excluded (Supplementary Figure S3B).

We next examined the allelic ratios of heterozygous SNVs within the same constitutively spliced exon (Figure 1B). This analysis only included SNVs with at least 20 reads to ensure adequate statistical power (see below). As expected, these values are highly correlated despite the possible compromise of mapping accuracy due to the closeness of multiple SNVs. This finding further attests to the validity of our mapping approach. In addition, the allelic ratios of all heterozygous SNVs (with at least 20 reads) are highly correlated across biological replicates (Figure 1C). This result supports the assumption that expression bias associated with alternative alleles of SNVs presented here represents a biological phenomenon that is not influenced much by variations in cell culture and other experimental techniques. Therefore, in the subsequent analyses, we combined data from all biological replicates to maximize the statistical power.

Identification of ASE in RNA-Seq data

To identify ASE events, we tested the null hypothesis of equal expression of the alternative alleles of a heterozygous SNV. SNVs were excluded if they were potentially in regions with copy number variants determined by the read depth of the genome sequencing data (14,15). The power to detect a significant ASE event is dependent on the number of reads associated with an SNV, as shown in Figure 2A. For example, if our goal is to identify an allelic ratio of 0.8:0.2 (either reference/variant or variant/reference allele, two-sided test) with ~75% power, then a minimum of 20 reads are needed for each SNV at an FDR of 5% (Figure 2A). Thus, a deeper RNA-Seq coverage can enable better power in detecting ASE patterns. To illustrate the dependence of this power requirement on the amount of available reads, we randomly sampled (with replacement) all the mapped reads and examined the read coverage of heterozygous SNVs (Figure 2B). This simulation offers a reasonable estimate of the requirement of sequencing depth since the available mapped reads in this study enabled coverage of most SNVs in expressed genes (≥1 RPKM) (Supplementary Figure S2B). As the number of reads increases, the number of SNVs that meet the power requirement approaches a plateau (at ~200 million mapped reads for $N \geq 20$) as a result of the limited number of expressed genes.

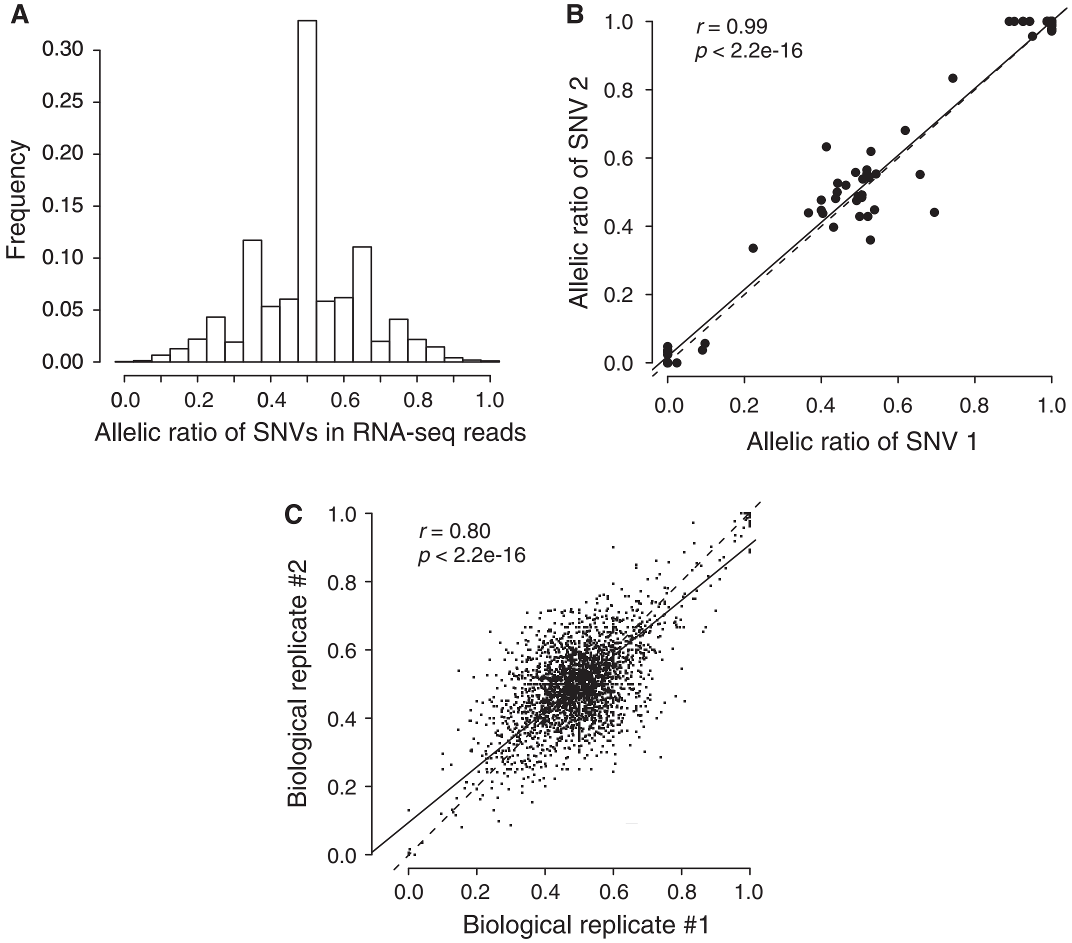

Figure 1. Evaluation of the allelic ratios calculated from RNA-Seq. (**A**) Distribution of allelic ratios (no. of reads containing the reference allele/total no. of reads) at heterozygous SNVs with reads from both alleles (mean: 0.500, median: 0.5, $P = 0.11$, binomial test). (**B**) Scatter plot of the allelic ratios of pairs of heterozygous SNVs (with ≥20 reads) located in the same constitutive exons (502 pairs with many overlaps in the plot). Only SNV pairs whose phase can be inferred from the RNA-Seq reads were included in this analysis. Pearson correlation coefficient and *P*-values are shown. The solid line shows the linear regression of the data points and the dashed line denotes the diagonal line. (**C**) Similar as (B), but for all heterozygous SNVs with ≥20 reads in both biological replicates.

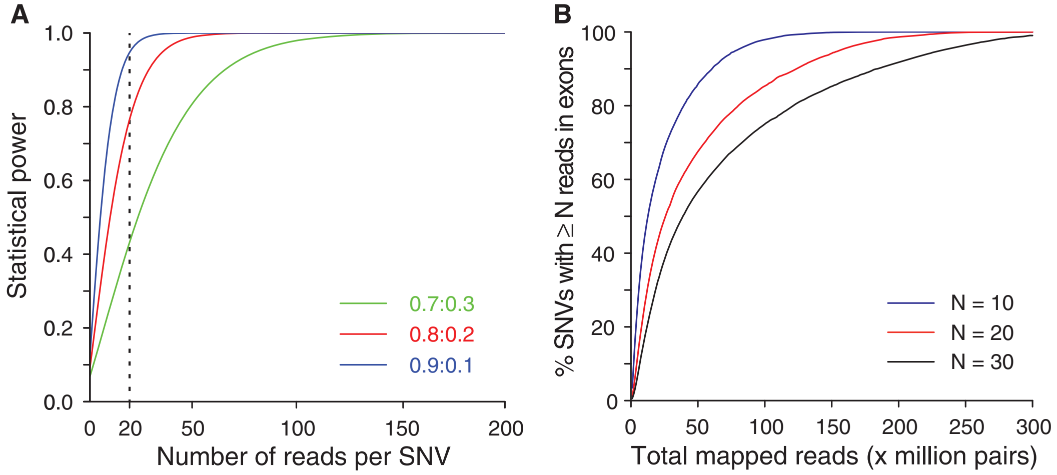

Figure 2. Statistical power and read coverage for ASE analysis. (**A**) Number of reads per SNV required to reach levels of statistical power (Chi-square Goodness-of-Fit test, $q \leq 0.05$) in the detection of allelic ratios of 0.7:0.3, 0.8:0.2 and 0.9:0.1 in the RNA-Seq reads. (**B**) Simulated results for percentage of SNVs with adequate power (*N* reads ≥ 10, 20 or 30, respectively) as a function of total mapped reads. The percentages were calculated against all exonic heterozygous SNVs with average coverage located in genes expressed at ≥ 1 RPKM in the original RNA-Seq data.

In our data, 7784 heterozygous SNVs (in 4553 genes) had a minimum of 20 reads. Among these SNVs, 1494 (19.2%, in 1172 genes) were identified to have allele-specific patterns using the Chi-square Goodness-of-Fit test at an FDR cutoff of 5% (see 'Materials and Methods' section, Supplementary Table S1). Similar results were obtained with the requirement of a minimum of 10 or 30 reads per SNV (Table 2). Therefore, our analyses showed that ~15–21% of SNVs or 20–26% of genes are associated with ASE patterns in U87MG cells, similar to the estimated percentage of *cis*-regulatory SNVs in previous studies (23–25). Table 2 reports the distribution of SNVs with ASE patterns in various types of genomic regions. The majority of these SNVs are located in coding exons or 3′-UTR regions and a small number in introns, novel exons or intergenic regions, consistent with the data acquisition protocol of RNA-Seq.

Quality of the ASE results

To evaluate the quality of the ASE events, we first examined the locations and quality scores of the corresponding SNVs in the reads. As shown in Supplementary Figure S4A, the SNVs with ASE patterns are distributed along the reads without significant positional bias. In addition, no significant difference was observed between the quality scores of the read bases corresponding to SNVs with and without ASE patterns (Supplementary Figure S4B). Thus, ASE identification was unlikely affected by sequencing errors which are relatively frequent near the ends of the reads and are often associated with low quality scores.

As another estimate of the FDR of the identified ASE events, we analyzed the difference in the allelic ratios of SNVs in biological replicates shown in Figure 1C. We found that 98 SNVs had significantly different allelic ratios in the two samples at a relatively large *P*-value cutoff ($P < 0.05$). Thus, the FDR among the identified ASE events is up to 6.6% (which could be an overestimate due to the relaxed *P*-value cutoff and the fact that biological replicates may possess natural variations). A read-randomization method was also applied to estimate the FDR independently (see 'Materials and Methods' section). This procedure resulted in 100 ASE events out of the 7784 SNVs with adequate power (N reads ≥ 20). Thus, the FDR of our results is ~6.7% based on this analysis.

Table 2. ASE of SNVs and the associated genes

Read coverage of SNVs	$N \geq 10$	$N \geq 20$	$N \geq 30$
SNVs, *n*	12 493	7784	6041
SNVs with ASE, *n* (%)	1831 (14.7)	1494 (19.2)	1268 (21.0)
Genes, *n*	5981	4553	3866
Genes with ASE, *n* (%)	1156 (19.3)	1172 (25.7)	961 (24.9)
Location of ASE SNVs, *n* (%)			
Coding exons	542 (29.6)	504 (33.7)	454 (35.8)
Non-coding exons	130 (7.1)	83 (5.6)	58 (4.6)
Introns	194 (10.6)	88 (5.9)	46 (3.6)
3′-UTR	663 (36.2)	627 (42.0)	571 (45.0)
5′-UTR	93 (5.1)	74 (5.0)	58 (4.6)
Novel exons	103 (5.6)	60 (4.0)	39 (3.1)
Intergenic regions	106 (5.8)	58 (3.9)	42 (3.3)

Numbers of SNVs and associated genes satisfying different power requirements and those identified with ASE patterns are shown. Power requirements were specified in terms of the number of reads covering an SNV ($N \geq 10$, $N \geq 20$ and $N \geq 30$). The distribution of SNVs with ASE patterns was determined using known gene structures defined by combining annotations in Ensembl, RefSeq, UCSC, Gencode and Vega genes. Novel exons were identified using our in-house algorithms (Supplementary Methods). Non-coding exons refer to exons in non-coding genes or non-coding transcripts of coding genes.

In our previous project (10), the same mapping approach was used to study RNA editing events revealed by RNA-Seq data. We showed that the allelic ratios of alternative bases of the RNA variants calculated based on RNA-Seq are highly concordant with those estimated by clonal sequencing. The correlation coefficient was 0.88 if at least 20 reads were required to cover the variant location. Without the knowledge of the genome sequences, expressed SNVs and RNA editing events are indistinguishable in the RNA-Seq reads. Thus, the result in our previous study also serves as a validation of the accuracy of the allelic ratios estimated in this study.

Our ASE results have significant overlaps with those reported in related studies. In a recent article, ASE patterns in a human lymphoblast cell line were analyzed using RNA-Seq (26). Among the SNVs reported with ASE patterns, 181 had adequate statistical power (N reads ≥ 20) and 53 demonstrated ASE in our study (overlap significance $P = 0.0006$, hypergeometric test). We also compared our results with those reported by *cis*-eQTL studies (repository at http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/). No significant overlap was observed between the exact SNVs associated with *cis*-eQTL and those with ASE. However, a relatively significant overlap was found between the genes hosting these two types of SNVs. Specifically, 973 genes with *cis*-eQTL contained SNVs with adequate power in our study, of which 271 genes harbored SNVs with ASE patterns ($P = 0.05$, hypergeometric test). These findings might be explained by the fact that both types of studies identify genes under *cis*-regulation, but not necessarily the exact SNV causing such regulatory mechanisms.

Analysis of association of ASE with alternative mRNA processing

ASE patterns identified above may be the results of *cis*-regulation affecting different aspects of gene expression, e.g. transcription, mRNA processing or stability. To elucidate the underlying regulatory mechanisms, we examined the gene and isoform expression patterns in the RNA-Seq data to relate them to the expression of SNVs. We found distinct ASE patterns within genes that may allow possible association with four types of *cis*-regulation (Figure 3). The first category consists of genes with global ASE demonstrated by all heterozygous SNVs. For example, the gene *L-RAP* has six SNVs that passed the power requirement (N reads ≥ 20) and all of them showed significant allelic expression bias (Figure 3A). In this case, it is very likely that a mechanism (such as allele-specific transcription factor binding) affecting whole-gene-level expression exists that resulted in allelic

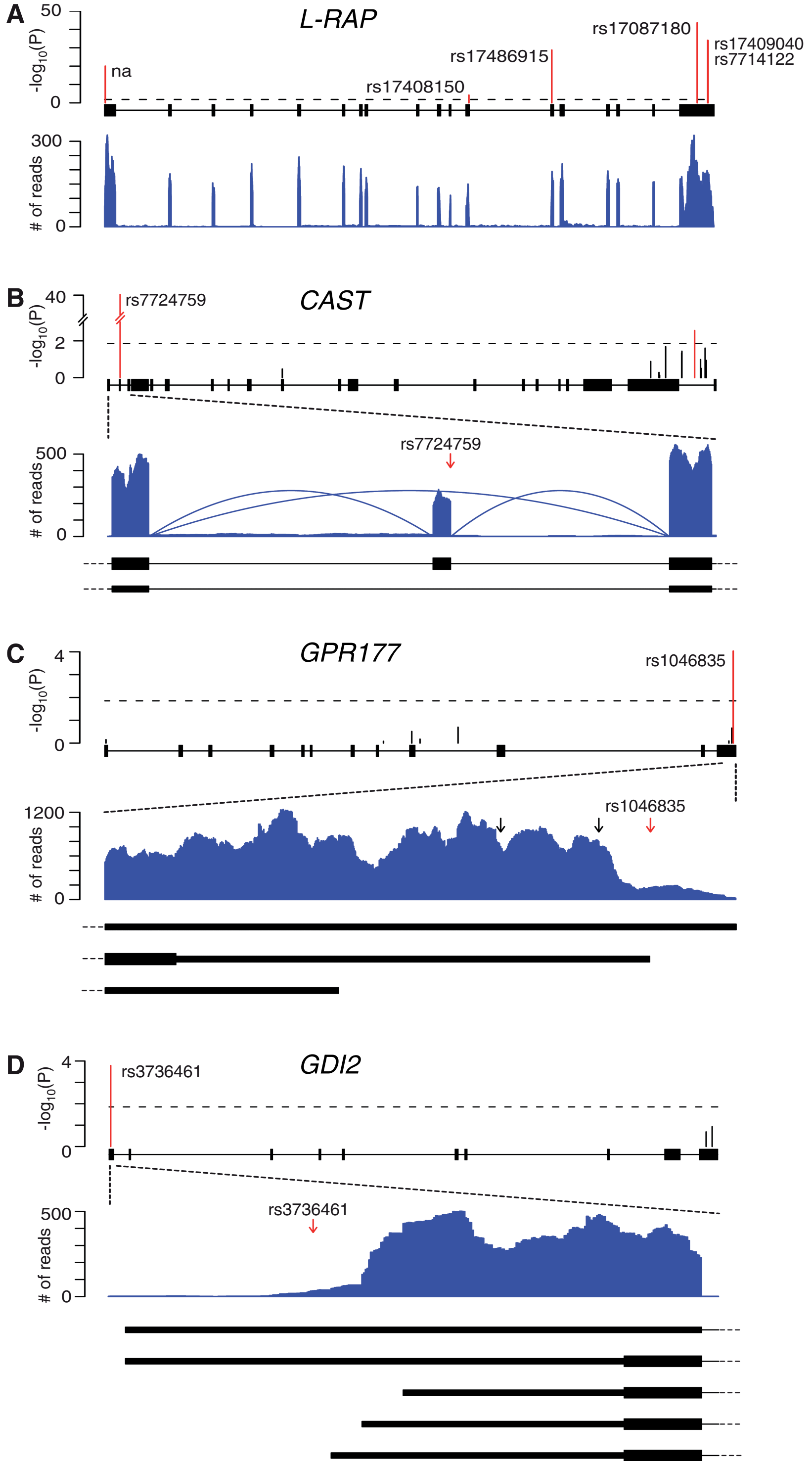

Figure 3. Example genes potentially associated with different types of *cis*-regulatory mechanisms. In the *P*-value panel, red and black vertical lines represent significant and insignificant *P*-values in the ASE analysis, respectively. Ensembl exons (black boxes) in any known isoforms are shown. For illustration purpose, the complete gene structure may not be shown. For all examples, transcript start sites are at the left of the exon–intron structure. Read distributions are shown only for the indicated regions except for (A). The arcs represent RNA-Seq reads mapped to spliced junctions. (**A**) *Cis*-regulation affecting overall gene expression level. (**B**) ASAS. (**C**) ASAP. (**D**) ASTI.

expression of the entire gene. The other three categories pertain to allele-specific mRNA processing, specifically, ASAS, ASAP and ASTI. In these cases, expression of only a fraction of SNVs in a gene is altered in the presence of allele-specific regulation that changes local transcript isoform structures.

An example of a potential ASAS event in the gene *CAST* is shown in Figure 3B. There are 12 SNVs in this gene that passed the power requirement for ASE analysis. Only two of them demonstrated significant allelic expression bias, thus excluding the possibility that a regulatory mechanism impinging on the whole-gene expression level explains the observed ASE patterns. The most significant ASE pattern occurs in the SNV rs7724759. We found that at least one other exonic SNV (non-ASE) are located in the same transcript isoforms as rs7724759 according to the RefSeq annotation. Hence, it is unlikely that allele-specific regulation of whole-isoform expression is the cause of its ASE. An examination of the RNA-Seq read distribution and annotated gene structure (Figure 3B) shows that this SNV resides in an alternative cassette exon. Since reads overlapping this SNV can only be obtained when the exon is included in the mRNA, the ASE pattern means that exon inclusion is significantly associated with only one allele (the reference allele in this example). Thus, this is a case of allele-specific exon inclusion most likely due to *cis*-regulation of alternative splicing (i.e. ASAS). Of course, it is not clear based on this data alone which SNV is the causal one underlying this observation (see below for related experimental studies).

Another category of possible allele-specific *cis*-regulatory mechanism is ASAP occurring in genes with alternative 3′-UTRs. We found examples (Figure 3C) where an SNV located in the extended UTR region has allelic bias in its expression, whereas other SNVs in the same gene/isoform do not have significant ASE. In this case, the most likely mechanism accounting for the observed ASE pattern is one that leads to ASAP. Similarly, in genes with alternative 5′-UTRs, we observed ASE patterns that are possibly due to the ASTI mechanism (Figure 3D).

Based on the above observations, it is possible to distinguish the potential regulatory mechanisms of ASE patterns using RNA-Seq data. We thus classified genes into the aforementioned categories via an automatic pipeline named ASARP (see 'Materials and Methods' section and Supplementary Figure S1). This analysis compares the expression patterns of all heterozygous SNVs of a gene and combines SNV expression with expression of alternatively processed mRNA regions as illustrated by the examples in Figure 3. Note that the specific SNVs associated with ASAS, ASAP or ASTI events were not required to have statistically significant ASE because, despite a large allelic bias, such SNVs often fail to pass the power requirement due to the fact that they reside in regions that by definition are sometimes absent in the mRNA. Altogether, 488 genes were classified into the aforementioned categories (Supplementary Table S2). Specifically, 197 genes showed ASE at the whole-gene-level and 291 genes demonstrated allele-specific mRNA processing patterns. As summarized in Figure 4, ASAS events constitute the largest category (66%) among all allele-specific mRNA processing events considered in this study, followed by the ASAP events. It should be noted that some genes identified with ASE difference at the whole-gene level might also show allele-specific mRNA processing patterns, which is not considered in the above results. We estimated the FDR of the above analyses for each type of event using read-randomization similarly as for the ASE events (see 'Materials and Methods' section). The numbers of genes identified with ASAS, ASAP and ASTI in the randomized data were 34, 16 and 1, yielding an FDR of 18, 12 and 1%, respectively.

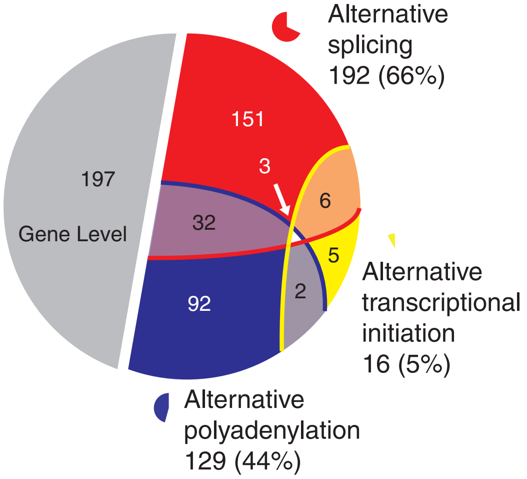

Figure 4. Number of genes with ASE patterns in the U87MG cells classified into different categories. A total of 488 genes were included that could be classified using our approach (see 'Materials and Methods' section). The numbers shown outside the pie chart represent the total number of genes in each category. The percentages for the alternative mRNA processing events were calculated relative to the union of all 291 genes with such events. 'Gene Level' events are not included in the percentage calculation because they are not comparable with the mRNA processing events (e.g. the latter only applies to genes with alternative mRNA processing). Since some genes may be classified into more than one category, the sum of the percentages of all types may be larger than 100%.

Molecular analysis of the causal SNVs for ASAS events

In the above section, we identified genes or alternatively processed regions that are possibly genetically regulated. The causal SNVs responsible for these observations can be obtained via detailed molecular analyses. Here, we use four ASAS events as examples given the prevalence of this type of event. For a putative ASAS event, we sub-cloned the SNV-harboring exon (separately for each allele) and the surrounding intronic regions from the U87MG DNA into a minigene expression vector (18). This minigene contains three exons with the middle exon being the SNV-harboring one (Figure 5). Since it is known that many splicing regulatory elements are located within or close to the exons (27,28), we only included the ∼250 bp intronic regions immediately flanking the exons. A pair of constructs was made for each SNV carrying the two alternative alleles. We transiently transfected the constructs into U87MG cells and examined the spliced isoforms using RT-PCR with fluorescence-labeling followed by gel electrophoresis (see 'Materials and Methods' section).

Figure 5 shows the differential splicing patterns associated with alternative alleles of the corresponding SNV in each of the four genes we tested. In the *CAST* gene, the SNV identified in our analysis (Figure 3B) is a known single nucleotide polymorphism (SNP) (id: rs7724759) located at the last base of the exon. The two alleles of the SNV lead to a significant difference in the

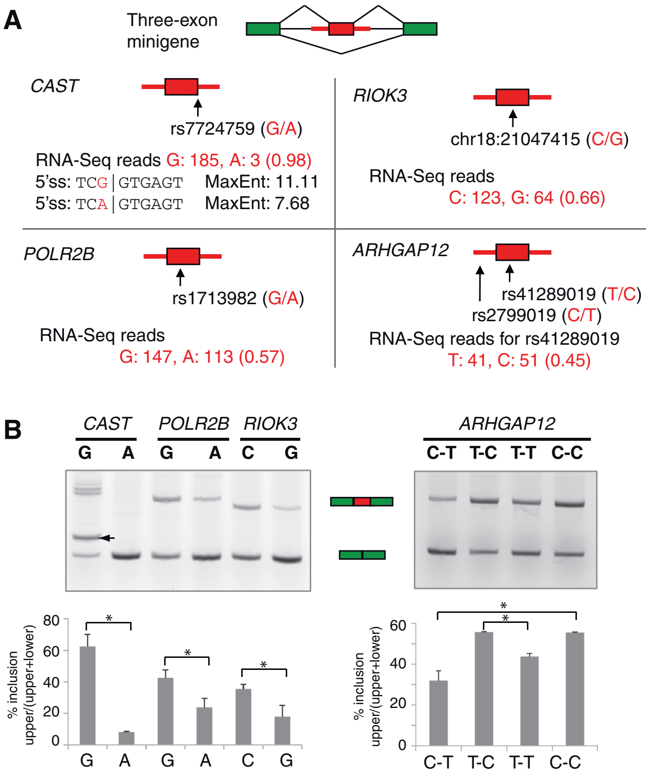

Figure 5. Molecular analysis of the causal SNVs in predicted ASAS events. (**A**) The three-exon minigene is shown that contains the tested exon of each gene and its flanking intronic regions (~250 nt). For each gene, the SNV associated with the ASAS pattern is shown, together with the number of RNA-Seq reads corresponding to each allele of the SNV (in red). The allelic ratios (as defined in Figure 1) are shown in parentheses. For the *CAST* gene, the SNV is located in a 5′ ss, the sequences and MaxEnt scores (31) of the alternative versions of the splice site are shown. (**B**) Semi-quantitative RT-PCR of total RNA from U87MG cells transfected with the minigenes with primers targeted to flanking exons [exons in green in (A)]. All transfections were repeated four times. Levels of exon inclusion are shown below (determined as the ratio of intensity of the upper band to the sum of upper and lower bands). For all genes, the upper band represents exon inclusion form and the lower band represents exon skipping form (for the *CAST* gene, arrow points to the exon inclusion form). For the *ARHGAP12* gene, the genotypes of the two SNVs are shown (in the order of intronic–exonic). Among the four pair-wise comparisons where the genotypes differ by one nucleotide (i.e. C–T versus C–C, C–T versus T–T, T–C versus T–T and T–C versus C–C), only two were statistically significant ($*P \leq 0.01$, Student's *t*-test).

strength of the 5′ ss (Figure 5A). Since this SNV is the only sequence variant between the two minigene constructs (confirmed by Sanger sequencing), we can conclude that it is the causal genetic variant that is responsible for the observed splicing difference (Figure 5B), most likely due to the alteration in 5′ ss strength. Our finding is consistent with that reported in (29).

The *POLR2B* gene, which encodes the second largest subunit of RNA polymerase II, has an exonic SNP (rs1713982) identified with ASAS pattern and it is the only sequence variant between the two constructs used in our experiments. As shown in Figure 5B, the G allele of this SNP is associated with an exon inclusion level higher than that of the A allele. Interestingly, binding sites (exonic splicing enhancers) of known splicing factors SRSF1, SRSF2 and SRSF5 were predicted to overlap the G allele, but not the A allele, according to ESEfinder (30), possibly explaining the cause of enhanced exon inclusion by the G allele. Similarly, the *RIOK3* gene has an SNV (genomic coordinate: chr18:21047415, not a known SNP) located in an alternative cassette exon, which is the only variant between the two constructs. Its G allele caused a higher level of exon skipping than the C allele (Figure 5B). However, no known exonic regulatory enhancers or silencers can explain this splicing difference, which may indicate that our current knowledge of splicing motifs is not yet complete.

Different from the above examples, the constructs for the *ARHGAP12* gene contained two sequence variants (rs41289019 in the exon and rs2799019 in the intron, Figure 5A). Although the exonic SNP is the one associated with an ASAS pattern, it is not clear which SNP(s) causes the expression variation. We thus made four constructs to encompass all allelic combinations of the two SNPs and compared the splicing patterns associated with different allelic combinations. As shown in Figure 5B, the exonic SNP has a predominant effect in inducing the observed splicing changes, thus most likely being the cause underlying the ASE pattern. However, there is no known splicing regulatory motif that can explain the allele-specific behavior of splicing. Importantly, this observation reflects the strength of our method, that is, it can detect existence of a splicing-altering SNP which could not have been predicted by examination of known splicing signals.

Application of the methods to breast cancer RNA-Seq

ASE and *cis*-regulation of mRNA processing can be highly tissue-, cell type- or disease-specific (32,33). To further demonstrate the usage of our methods, we analyzed another set of RNA-Seq data obtained from primary breast cancer tissues published in a previous study (34). This study conducted high-coverage sequencing of both the genome and the transcriptome of a metastatic lobular breast cancer specimen. We analyzed the genome sequencing data using the Short Oligonucleotide Analysis Package (SOAP) (35) and identified a total of 1 760 963 high-confidence heterozygous SNVs. For the RNA-Seq data, the same read mapping and analysis methods were used as presented above. The mapping results are shown in Supplementary Table S3 and the overall distribution of allelic ratios of reads mapped to heterozygous SNVs revealed no significant mapping bias (Supplementary Figure S5).

Among the 14 570 heterozygous SNVs (in 4433 genes) that satisfy the power requirement (N reads ≥ 20), 4052 (in 2001 genes) were identified with ASE patterns. Thus, the proportions of SNVs and genes demonstrating ASE are 27.8 and 45.1%, respectively. We next classified the genes according to the predicted *cis*-regulatory mechanisms, using the same approach as for U87MG cells. As shown in Figure 6, 225 genes demonstrated ASE due to whole-gene-level regulation and 605 genes showed ASARP. ASAS events again constitute the largest category (80%) among all three types of alternative mRNA processing events, followed by the ASAP events. This distribution

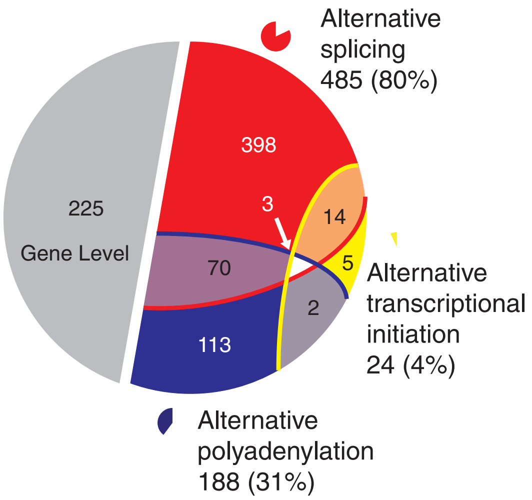

Figure 6. Number of genes with ASE patterns in the breast cancer data classified into different categories. Similar as Figure 4; a total of 830 genes were included that could be classified using our approach (see 'Materials and Methods' section).

of ASE patterns in different mechanistic categories is largely similar as that in the U87MG cells.

Comparison of allele-specific gene and alternative RNA expression in U87MG and breast cancer

The prevalence of ASE in breast cancer cells (27.8% SNVs and 45.1% genes) is higher than that in the U87MG cell line (19.2 and 25.7% for SNVs and genes, respectively). To determine if these differences were due to the difference in sequencing depth in the two data sets, we randomly sampled 4000 SNVs with adequate power in each data set. The sampling was controlled such that the SNVs had one-to-one match between the two data sets in terms of read coverage and hence statistical power. A total of 744 (17%) SNVs in the U87MG data were identified with ASE patterns, whereas a much higher number (1329, 32%) of ASE SNVs were resulted from the breast cancer data. Thus, the more frequent occurrence of ASE in breast cancer was unlikely an artifact related to the level of read coverage.

Genetic backgrounds and expression profiles of the U87MG and breast cancer samples are vastly different. Among all heterozygous SNVs, only 18 613 were common to both samples, of which 753 (4%) satisfied the power requirement (N reads ≥ 20) in both. Among the 753 SNVs, 144 and 176 were identified with ASE patterns, respectively, in the U87MG and breast cancer data. Thus, the proportions of ASE in the two data sets are similar among the shared SNVs. More importantly, 43 SNVs had allelic expression bias common to both samples, demonstrating a significant overlap ($P = 0.017$, hypergeometric test). This result suggests that there exists cell type-independent ASE given common genetic backgrounds (i.e. SNVs), which also serves as a support for the validity and reproducibility of our methods in identifying ASE. Among the common SNVs in U87MG and breast cancer data, only a small fraction could be categorized into different types of events in both samples (Supplementary Table S4) as a result of the stringent criteria in our categorization scheme (Supplementary Methods). Nevertheless, the categories that the SNVs belong to are concordant between the two samples ($P = 0.005$ against the null hypothesis that the categories of the shared SNVs are independent in the two samples, Supplementary Table S4).

Allele-specific gene and alternative RNA expression and GWASs

Recent GWASs reported a large number of SNPs associated with phenotypes of various diseases. However, the mechanisms underlying most SNP-disease associations remain unknown. Here, we demonstrate that our methods of RNA-Seq analysis may shed light on the functional impacts of GWAS SNPs. We first examined whether any GWAS SNPs (reported at http://www.genome.gov/gwastudies/) were identified with ASE patterns in our study. A total of 10 disease/phenotype-associated SNPs were found in the U87MG and breast cancer ASE results (Supplementary Table S5). For example, a SNP rs4770433 in the gene *SACS* was reported to be associated with the protein level of *IL12* (36), a gene involved in the immune response and tumor growth (37). Mutations in *SACS* are known to cause a neurodegenerative disorder, autosomal recessive spastic ataxis of Charlevoix–Saguenay (38). In the U87MG data, this gene was identified to be under allele-specific *cis*-regulation at the whole-gene level. Another example is the SNP rs6749447 in the gene *STK39*, which was one of the first SNPs identified in GWASs to be associated with hypertension (39). *STK39* encodes a serine/threonine kinase that possibly functions in the cellular stress response pathway. In our study, this gene was identified to be under ASAS regulation in the breast cancer data.

Instead of being the causal genetic basis, GWAS SNPs may only serve as markers of disease-genotype associations. We next investigated whether the genes identified in our study were reported in GWASs, regardless of the specific SNPs involved. A total of 64 and 129 genes in the U87MG and breast cancer data, respectively, overlap those resulted from GWASs (Supplementary Table S6). Among these genes, 52, 107, 44 and 5 demonstrated allele-specific gene-level expression, ASAS, ASAP and ASTI, respectively, in at least one data set. As an example, the gene *LITAF* was identified with ASAS patterns in both samples. It contains a locus associated with the QT interval duration of the heart (40) and was reported to exert inhibitory effects on tumor growth (41). The above results suggest that our methods may potentially enable a better understanding of various disease mechanisms.

DISCUSSION

Recent genotyping, exome and genome sequencing projects generated an extraordinary list of genetic variants across human populations and diseases. Most of these variants may not have a significant function.

Thus far, identification of the functional variants remains a significant challenge. Ultimately, the biological impact and functionality of these variants need to be examined and validated experimentally. Nevertheless, bioinformatic predictions that can narrow down the search for causal/functional variants are in high demand to guide experimental studies effectively. Here, we presented methods to analyze RNA-Seq data that enabled identification of genes and alternatively processed regions whose expression is under the regulation of genetic variants. Compared to previous methods (e.g. eQTL analysis), our approach utilizes RNA-Seq data of a single subject to provide insights that were only possible using massive-scale parallel expression assays of a large number of subjects. In addition, different from eQTL and other recent ASE studies, our approach not only identifies ASE patterns, but also predicts the functional mechanisms of genetic variants in specific categories of *cis*-regulation of gene expression, which provides essential information to facilitate discovery of causal variants as shown by our experimental studies.

The strength of our methods rooted from the unique advantages of RNA-Seq. First, RNA-Seq provides mRNA sequence information at single-nucleotide resolution. With enough read coverage, RNA-Seq can potentially interrogate all expressed SNVs of a gene, thereby providing a powerful tool for ASE studies. The simultaneous quantification of single-nucleotide expression and exon/gene expression is another advantage of RNA-Seq because a single data set can provide not only allelic expression of SNVs, but also whole-gene and alternative isoform expression. In this sense, RNA-Seq is a cost-effective approach for studies of genetic controls of gene expression.

To demonstrate the utilities of the methods, we analyzed RNA-Seq data of two different types of cancer samples. Read-mapping bias of the alternative alleles of SNVs (20–22) was removed using our previously developed mapping strategy (10). Our results suggest that 26–45% of genes demonstrated ASE patterns in the studied cancer cells at an FDR of ~5%. We also demonstrated that the *cis*-regulatory mechanisms underlying ASE may be inferred from RNA-Seq data for hundreds of genes in each sample at FDRs <20%. We chose parameters in this analysis such that a relatively relaxed FDR was reached since these predictions provide candidate events that can be further examined for disease relevance (such as in GWAS results) and molecular validations. Thus, a relatively large repository of candidates to start with may be beneficial. Nevertheless, the parameters can be adjusted (e.g. a smaller *P*-value cutoff of the Fisher's exact test, Supplementary Methods) if a lower FDR is desired. In addition, approaches other than FDR analysis (e.g. Bonferroni correction) may be used to account for multiple hypothesis testing, which may change the number and statistical stringency of the final results.

The ASE profiles of shared SNVs of the two data sets overlap significantly, despite their substantial difference in the types of cells and diseases involved. This observation confirms that genetic factors play an important role in gene regulation. On the other hand, the prevalence of ASE differs between the two samples, with the breast cancer data showing a much higher percentage of genes with ASE. This difference may be explained by the considerable difference in their genetic backgrounds or, possibly, expression or functional difference of *trans*-acting factors regulating the ASE patterns. For the small number of SNVs predicted to be under allele-specific regulation at the levels of whole gene or mRNA processing in both samples, their associated categories of potential *cis*-regulatory mechanisms are highly concordant in the two samples (Supplementary Table S4). This finding indicates that studies of ASE in a specific cell type may be extrapolated to infer functional roles of disease-associated SNPs or mutations even if the cell type is not directly related to the disease. Indeed, we found many common genes between those with allele-specific regulation and those reported in recent GWASs. Combining the two samples, 182 such genes were involved in GWAS disease association, confirming that *cis*-regulation is an essential aspect in the function of genetic variants.

Although RNA-Seq allows *de novo* identification of biallelically expressed SNVs, we utilized known SNVs in the respective samples to avoid the complication by RNA editing events (10). Although whole-genome sequencing data are not yet available for many samples, knowledge of heterozygous SNVs can be readily obtained from exome sequencing or microarray analysis. With the extraordinary improvement in high-throughput technologies in recent years, an unprecedented amount of transcriptome sequencing data is becoming available. Bioinformatic analyses that can examine and integrate such data sets to address a wide variety of biological questions are highly desirable. In this study, we demonstrated that analyses of RNA-Seq data revealed a large number of allele-specific events potentially associated with different types of *cis*-regulatory mechanisms of gene expression. Such studies may provide a solid foundation to facilitate further investigations of the genetic basis of human diseases.

ACCESSION NUMBERS

RNA-Seq data were submitted to the Gene Expression Omnibus with ID GSE29738.

Software is available for download at: http://www.ibp.ucla.edu/research/xiao/Software.html

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables S1–S6, Supplementary Figures S1–S5, Supplementary Methods and Supplementary References [34,35,42,43].

ACKNOWLEDGEMENTS

We thank members of the Xiao laboratory for helpful comments on this article and B. O'Connor for helping with the genome sequence data of U87MG. We are grateful to the Bowie and Clarke labs for generously allowing us to use their equipment and the Eisenberg lab (especially D. Cascio and L. Goldschmidt) for providing help on computational clusters. This article was prepared using a limited access data set (of the human primary breast cancer samples) obtained from BCCA and does not necessarily reflect the opinions or views of BCCA.

FUNDING

The National Institutes of Health (R21DA027039); Alfred P. Sloan Foundation Research Fellowship; March of Dimes Foundation (5-FY10-486 to X.X.); and an American Heart Association Postdoctoral Fellowship to J.H.L. Funding for open access charge: Sloan Foundation Research Fellowship.

Conflict of interest statement. None declared.

REFERENCES

1. Cookson,W., Liang,L., Abecasis,G., Moffatt,M. and Lathrop,M. (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184–194.
2. Pastinen,T., Ge,B. and Hudson,T.J. (2006) Influence of human genome polymorphism on gene expression. *Hum. Mol. Genet.*, **15(Spec. No. 1)**, R9–R16.
3. Rockman,M.V. and Kruglyak,L. (2006) Genetics of global gene expression. *Nat. Rev. Genet.*, **7**, 862–872.
4. Pastinen,T. (2010) Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.*, **11**, 533–538.
5. Yan,H., Yuan,W., Velculescu,V.E., Vogelstein,B. and Kinzler,K.W. (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.
6. Pagani,F., Raponi,M. and Baralle,F.E. (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl Acad. Sci. USA*, **102**, 6368–6372.
7. Cunninghame Graham,D.S., Manku,H., Wagner,S., Reid,J., Timms,K., Gutin,A., Lanchbury,J.S. and Vyse,T.J. (2007) Association of IRF5 in UK SLE families identifies a variant involved in polyadenylation. *Hum. Mol. Genet.*, **16**, 579–591.
8. Graham,R.R., Kyogoku,C., Sigurdsson,S., Vlasova,I.A., Davies,L.R., Baechler,E.C., Plenge,R.M., Koeuth,T., Ortmann,W.A., Hom,G. *et al.* (2007) Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl Acad. Sci. USA*, **104**, 6758–6763.
9. Kwan,T., Benovoy,D., Dias,C., Gurd,S., Provencher,C., Beaulieu,P., Hudson,T.J., Sladek,R. and Majewski,J. (2008) Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.*, **40**, 225–231.
10. Bahn,J.H., Lee,J.H., Li,G., Greer,C., Peng,G. and Xiao,X. (2012) Accurate Identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.*, **22**, 142–150.
11. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
12. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
13. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
14. Clark,M.J., Homer,N., O'Connor,B.D., Chen,Z., Eskin,A., Lee,H., Merriman,B. and Nelson,S.F. (2010) U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet.*, **6**, e1000832.
15. Abyzov,A., Urban,A.E., Snyder,M. and Gerstein,M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.
16. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Series B*, **57**, 289–300.
17. Ventura,V., Paciorek,C.J. and Risbey,J.S. (2004) Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *J. Climate*, **17**, 4343–4356.
18. Xiao,X., Wang,Z., Jang,M., Nutiu,R., Wang,E.T. and Burge,C.B. (2009) Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat. Struct. Mol. Biol.*, **16**, 1094–1100; PMC2766517.
19. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
20. Degner,J.F., Marioni,J.C., Pai,A.A., Pickrell,J.K., Nkadori,E., Gilad,Y. and Pritchard,J.K. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.
21. Heap,G.A., Yang,J.H., Downes,K., Healy,B.C., Hunt,K.A., Bockett,N., Franke,L., Dubois,P.C., Mein,C.A., Dobson,R.J. *et al.* (2009) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.*, **19**, 122–134.
22. McDaniell,R., Lee,B.K., Song,L., Liu,Z., Boyle,A.P., Erdos,M.R., Scott,L.J., Morken,M.A., Kucera,K.S., Battenhouse,A. *et al.* (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, **328**, 235–239.
23. Lo,H.S., Wang,Z., Hu,Y., Yang,H.H., Gere,S., Buetow,K.H. and Lee,M.P. (2003) Allelic variation in gene expression is common in the human genome. *Genome Res.*, **13**, 1855–1862.
24. Hoogendoorn,B., Coleman,S.L., Guy,C.A., Smith,K., Bowen,T., Buckland,P.R. and O'Donovan,M.C. (2003) Functional analysis of human promoter polymorphisms. *Hum. Mol. Genet.*, **12**, 2249–2254.
25. Serre,D., Gurd,S., Ge,B., Sladek,R., Sinnett,D., Harmsen,E., Bibikova,M., Chudin,E., Barker,D.L., Dickinson,T. *et al.* (2008) Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet.*, **4**, e1000006.
26. Rozowsky,J., Abyzov,A., Wang,J., Alves,P., Raha,D., Harmanci,A., Leng,J., Bjornson,R., Kong,Y., Kitabayashi,N. *et al.* (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.*, **7**, 522.
27. Wang,Z. and Burge,C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, **14**, 802–813.
28. Xiao,X. and Lee,J.H. (2010) Systems analysis of alternative splicing and its regulation. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **2**, 550–565.
29. Coulombe-Huntington,J., Lam,K.C., Dias,C. and Majewski,J. (2009) Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet.*, **5**, e1000766.
30. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
31. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
32. Heinzen,E.L., Ge,D., Cronin,K.D., Maia,J.M., Shianna,K.V., Gabriel,W.N., Welsh-Bohmer,K.A., Hulette,C.M., Denny,T.N. and Goldstein,D.B. (2008) Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol.*, **6**, e1.
33. Kwan,T., Grundberg,E., Koka,V., Ge,B., Lam,K.C., Dias,C., Kindmark,A., Mallmin,H., Ljunggren,O., Rivadeneira,F. *et al.* (2009) Tissue effect on genetic control of transcript isoform variation. *PLoS Genet.*, **5**, e1000608.
34. Shah,S.P., Morin,R.D., Khattra,J., Prentice,L., Pugh,T., Burleigh,A., Delaney,A., Gelmon,K., Guliany,R., Senz,J. *et al.* (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**, 809–813.
35. Li,R., Li,Y., Fang,X., Yang,H., Wang,J. and Kristiansen,K. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, **19**, 1124–1132.

36. Melzer,D., Perry,J.R., Hernandez,D., Corsi,A.M., Stevens,K., Rafferty,I., Lauretani,F., Murray,A., Gibbs,J.R., Paolisso,G. *et al.* (2008) A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.*, **4**, e1000072.
37. Robertson,M.J. and Ritz,J. (1996) Interleukin 12: basic biology and potential applications in cancer treatment. *Oncologist*, **1**, 88–97.
38. Engert,J.C., Dore,C., Mercier,J., Ge,B., Betard,C., Rioux,J.D., Owen,C., Berube,P., Devon,K., Birren,B. *et al.* (1999) Autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS): high-resolution physical and transcript map of the candidate region in chromosome region 13q11. *Genomics*, **62**, 156–164.
39. Wang,Y., O'Connell,J.R., McArdle,P.F., Wade,J.B., Dorff,S.E., Shah,S.J., Shi,X., Pan,L., Rampersaud,E., Shen,H. *et al.* (2009) From the cover: whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proc. Natl Acad. Sci. USA*, **106**, 226–231.
40. Pfeufer,A., Sanna,S., Arking,D.E., Muller,M., Gateva,V., Fuchsberger,C., Ehret,G.B., Orru,M., Pattaro,C., Kottgen,A. *et al.* (2009) Common variants at ten loci modulate the QT interval duration in the QTSCD Study. *Nat. Genet.*, **41**, 407–414.
41. Zhou,J., Yang,Z., Tsuji,T., Gong,J., Xie,J., Chen,C., Li,W., Amar,S. and Luo,Z. (2011) LITAF and TNFSF15, two downstream targets of AMPK, exert inhibitory effects on tumor growth. *Oncogene*, **30**, 1892–1900.
42. Lee,J.H., Gao,C., Peng,G., Greer,C., Ren,S., Wang,Y. and Xiao,X. (2011) Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. *Circ. Res.*, **109**, 1332–1341.
43. Li,R., Yu,C., Li,Y., Lam,T.W., Yiu,S.M., Kristiansen,K. and Wang,J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.