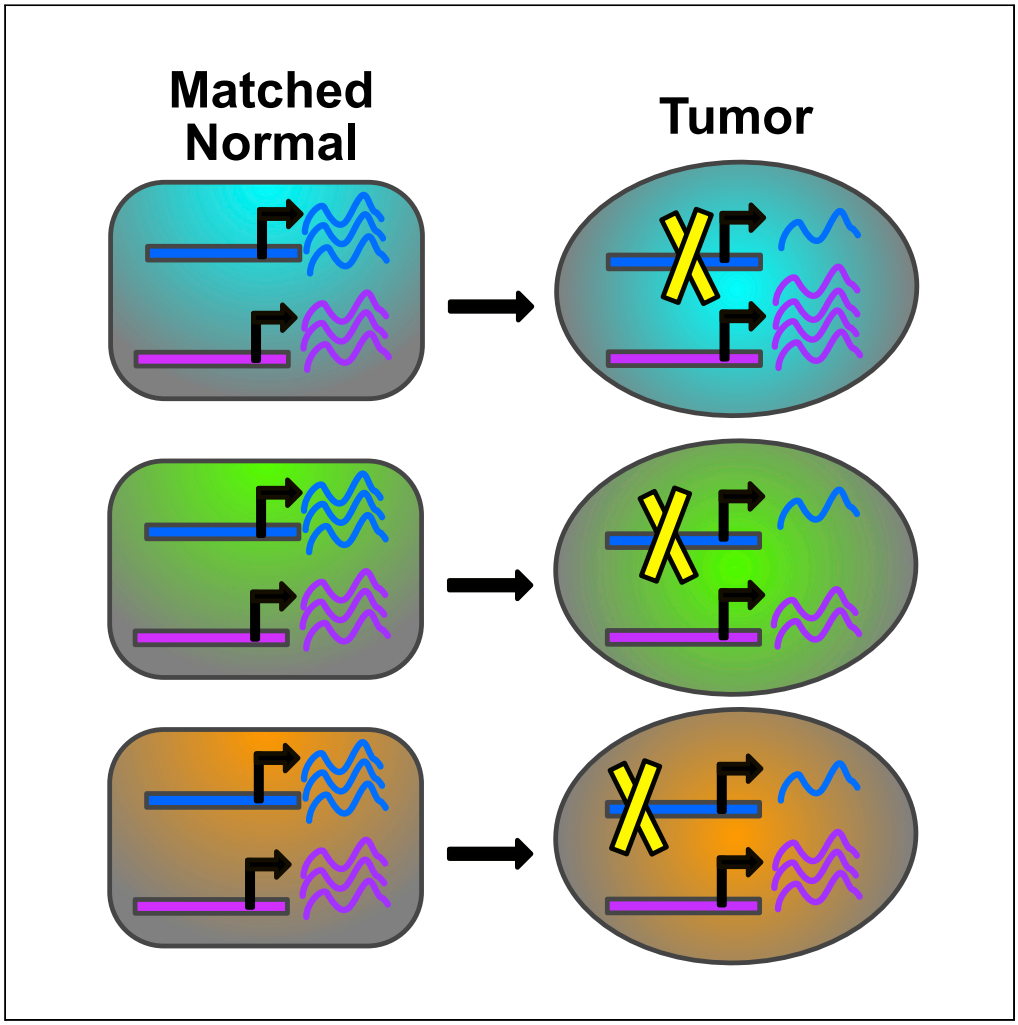
iScience

CellPress
OPEN ACCESS

Article

Cis-regulatory mutations with driver hallmarks in major cancers



Zhongshan Cheng, Michael Vermeulen, Micheal Rollins-Green, Brian DeVeale, Tomas Babak

brian.deveale@ucsf.edu (B.D.)
tomas.babak@queensu.ca (T.B.)

HIGHLIGHTS

Enrichment of functional non-coding somatic mutations predicts drivers

Elevated variant allele frequencies are consistent with roles in tumorigenesis

Putative non-coding drivers disrupt transcription factor binding motifs

Predicted drivers associate with increased oncogene and decreased TSG expression

Cheng et al., iScience *24*, 102144
March 19, 2021
https://doi.org/10.1016/j.isci.2021.102144

iScience

CellPress
OPEN ACCESS

Article

Cis-regulatory mutations with driver hallmarks in major cancers

Zhongshan Cheng,[1,3] Michael Vermeulen,[1,4] Micheal Rollins-Green,[1] Brian DeVeale,[2,5,6,*] and Tomas Babak[1,5,*]

SUMMARY

Despite the recent availability of complete genome sequences of tumors from thousands of patients, isolating disease-causing (driver) non-coding mutations from the plethora of somatic variants remains challenging, and only a handful of validated examples exist. By integrating whole-genome sequencing, genetic data, and allele-specific gene expression from TCGA, we identified 320 somatic non-coding mutations that affect gene expression in *cis* (FDR<0.25). These mutations cluster into 47 cis-regulatory elements that modulate expression of their subject genes through diverse molecular mechanisms. We further show that these mutations have hallmark features of non-coding drivers; namely, that they preferentially disrupt transcription factor binding motifs, are associated with a selective advantage, increased oncogene expression and decreased tumor suppressor expression.

INTRODUCTION

Identification of somatic mutations that contribute to tumorigenesis is an essential step to understanding disease prognosis and developing therapies (Gerstung et al., 2015; Kulik et al., 1989; Verhaak et al., 2010). Despite extensive exome and genome sequencing efforts, a substantial proportion of causal or driver mutations (called *drivers* from here on) are thought to be unknown (Kandoth et al., 2013; Cancer Genome Atlas Research Network, 2014, 2015; Nik-Zainal et al., 2016). On average, 22.2% of tumor samples within each cancer type do not harbor coding mutations in any of 144 common driver genes (Schroeder et al., 2014). Moreover, since multiple drivers are typically involved (Vogelstein et al., 2013), even tumors with well-characterized mutations likely harbor additional causal alterations (Beerenwinkel et al., 2007; Merid et al., 2014; Sjoblom et al., 2006; Vogelstein et al., 2013). Mutations in *cis*-regulatory elements (CREs) are postulated to comprise a large fraction of the undiscovered drivers (Sjoblom et al., 2006). However, despite the availability of hundreds of complete tumor genomes, only a few non-coding drivers have been experimentally validated (Table S1).

Distinguishing drivers from passengers outside coding regions requires overcoming several known challenges: the search space is orders of magnitude larger, functional impact cannot be predicted from amino acid changes (especially gain-of-function alterations), mutation rates are higher (Poulos et al., 2015), and positive selection pressure on relative growth is relaxed. These challenges have been partially overcome by associating mutations with disruption or acquisition of transcription factor binding sites (Kalender Atak et al., 2017; Mathelier et al., 2015; Melton et al., 2015; Svetlichnyy et al., 2015; Weinhold et al., 2014), altered mRNA abundance (Fredriksson et al., 2014), clinical data (Smith et al., 2015; Weinhold et al., 2014), and evolutionary conservation (Carter et al., 2009; Foo et al., 2015; Fu et al., 2014; Piraino and Furney, 2017). Combinations of these features have also been weighed to prioritize putative drivers and determine significant mutational hotspots (Fu et al., 2014; Kalender Atak et al., 2017; Piraino and Furney, 2017; Puente et al., 2015; Weinhold et al., 2014).

Since the tumorigenic role of a non-coding driver is likely exerted through a *cis*-change in gene expression (Khurana et al., 2016), mapping genes whose expression is impacted by *cis*-acting regulatory effects has significant promise. Allele-specific expression (ASE), where one allele of a gene is more highly expressed than the other, is a powerful approach for detecting *cis*-regulatory effects, since *trans*-regulatory effects impact both alleles equally (Fraser, 2011). By comparing ASE in tumors to matched normal ASE ("diffASE"), it is further possible to distinguish somatic from germ-line effects. Ongen et al. applied this approach to identify 71 putative driver genes in colorectal cancer (Ongen et al., 2014). Furthermore, after predicting

[1]Department of Biology, Queen's University, Kingston, ON K7L 3N6, Canada

[2]The Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, Center for Reproductive Sciences, University of California, San Francisco, San Francisco, CA 94143, USA

[3]Present address: Applied Bioinformatics Center, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

[4]Present address: Department of Medical Genetics, University of British Colombia, Vancouver, BC V6T 1Z4, Canada

[5]These authors contributed equally

[6]Lead contact

*Correspondence: brian.deveale@ucsf.edu (B.D.), tomas.babak@queensu.ca (T.B.)

https://doi.org/10.1016/j.isci.2021.102144

functional non-coding variants by prioritizing those generating *de novo* transcription factor binding motifs, Atak et al. showed that many of these somatic mutations were associated with ASE (Kalender Atak et al., 2017). In practice, however, the sparse availability of matched tumor and normal gene expression and genetic data poses a significant limitation. Just 7.7% of The Cancer Genome Atlas (TCGA) tumor samples have matched normal RNA-Seq data (Figure S1A).

Here we show that the vast majority of differential ASE is acquired in tumors, enabling us to dispense with the matched normal requirement and expand our survey 13-fold. We interrogated all whole-genome sequenced non-coding somatic mutations across 1,165 TCGA patients and identified 47 putative drivers as mutated CREs on the basis of robust association with ASE in tumors. The driver role of these mutations is further supported by functional disruption of transcription factor binding sites and elevated variant allele frequencies. This functional catalog of non-coding features significantly expands our knowledge of non-coding tumor driver biology.

RESULTS

Survey of breast invasive carcinoma reveals that differential ASE is due to ASE in tumors

We initially focused on breast invasive carcinoma (BRCA) since it is the cancer type with the largest set of matched tumor and normal RNA-seq data accompanied by whole-genome sequence (WGS) in TCGA (Figure S1A). Measuring ASE relies on counting RNA-Seq reads that map over heterozygous single-nucleotide polymorphisms (SNPs) (Figure 1A) detected by genotyping arrays. To maximize our sensitivity, we first imputed and phased SNPs using the 1000 Genome haplotypes (Howie et al., 2009) (Figure 1A), which on average increases the number of informative SNPs by 20%. We have previously shown that this is more accurate than relying on WGS, particularly where coverage is low (Figure S1B), and reduces false-positive SNPs that have a disproportionately high impact on estimates of ASE since all reads are assigned to one haplotype (Babak et al., 2015). Phasing also allowed us to combine allelic counts across SNPs within the same gene, which contributes to the improved accuracy (Babak et al., 2015) (and see Transparent methods and Figure S2 for details). We observed extensive diffASE in BRCA (Figure 1B).

Nearly all of the diffASE can be attributed to an increase of ASE in tumors relative to matched controls (Figure 1B). We reasoned that this trend may be due to higher clonality of tumors relative to matched normal tissue which would be expected to be more complex. We first considered whether loss of heterozygosity (LOH) may be a confounding factor. Since all BRCA tumors are female, a comparison of allelic expression between autosomes to the X chromosome could illuminate the contribution of clonality. X chromosomes are randomly inactivated across cells comprising normal tissue. Comparison with a clone derived from this tissue (where all cells retain monoallelic expression from the same allele) would yield strong diffASE for any expressed gene on chromosome X. If clonality was the dominant source of greater ASE in tumors, we would expect enrichment of highly ranked X-linked genes when evaluated for diffASE. This enrichment would not be expected if LOH was the dominant source. We indeed observed a high enrichment of X-linked genes (66/100) among the top diffASE genes, suggesting that these tumors are highly clonal (Figure S2D). When we performed the ASE analysis using only tumor expression (tumorASE), >98% of the diffASE events were recapitulated in tumorASE and >90% of diffASE events attributable to increased allelic bias in tumors (Figure 1C). Finally, neither CNVs nor methylation explained the majority of ASE events originating in tumors (Figures 1D, 1E, and S3). CNVs showed the stronger correlation but only account for about 11% of the ASE in tumors. These findings suggested that altered *cis*-regulatory mechanisms of gene expression might explain the observed ASE in tumors, and that this signal is a valuable starting point for identifying non-coding drivers.

Identification of mutations that explain ASE in tumors

The availability of WGS data for 113 BRCA RNA-seq tumor samples (Figure S1A, Table S2) allowed us to find specific mutations that are associated, and which may explain, the observed ASE in tumors. We evaluated common mutation callers and implemented a robust filtering scheme to yield high confidence somatic variants (Figures S4A–S4D; see Transparent methods for details). We then asked whether the presence or absence of these variants near a gene is associated with ASE of that gene across BRCA tumor samples. Unfortunately, using mutations 10 kb upstream of each transcription start site (TSS) as well as within each gene body did not yield associations that survived multiple test correction, even in this heavily surveyed cancer type. We chose the window because cis-regulatory variants are heavily enriched in the 10 kb window upstream of the TSS (Group et al., 2020). The high proportion of neutral mutations relative to genuine non-

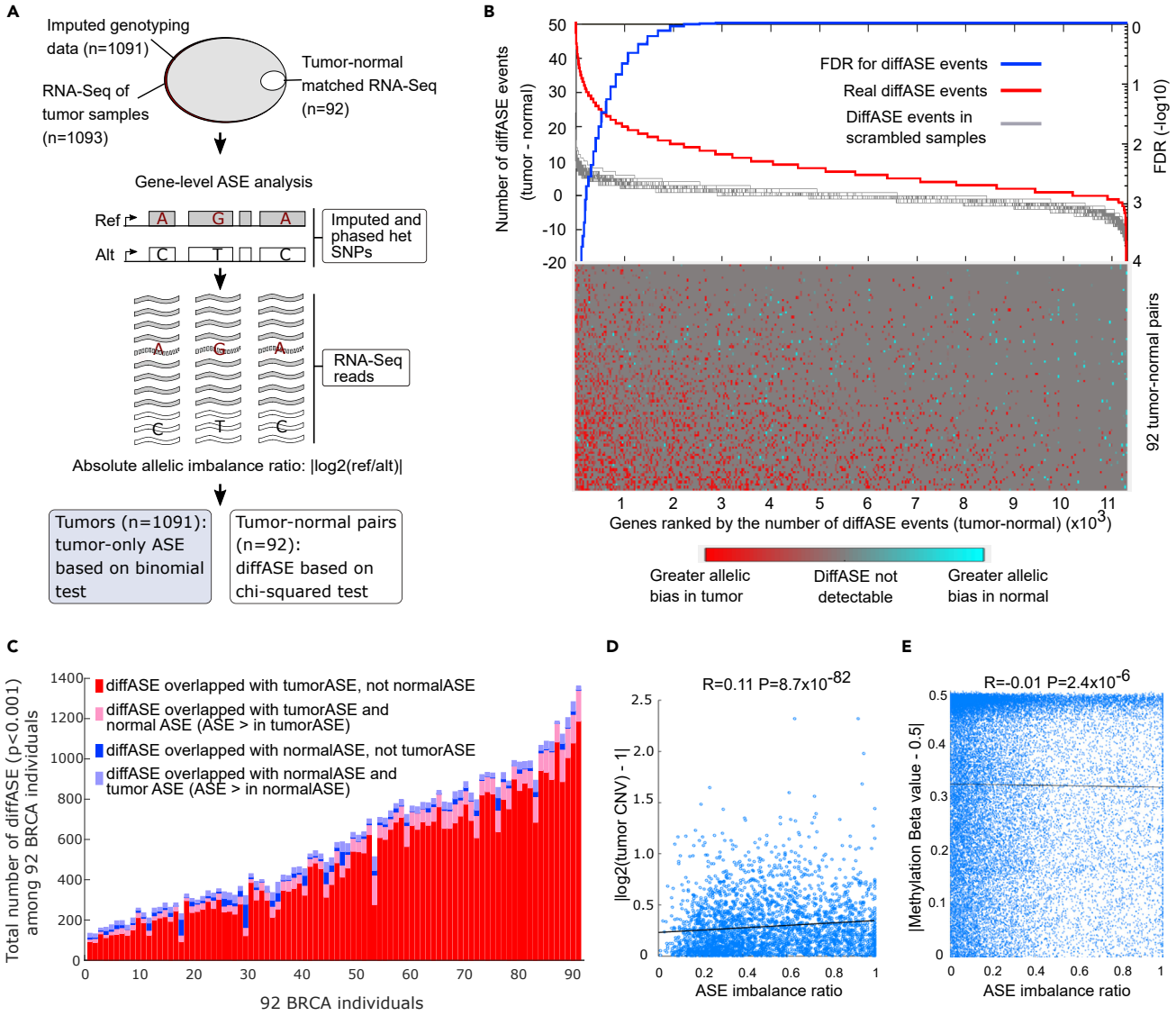


Figure 1. Allelic bias commonly arises in tumors independent of copy number variations (CNVs) and promoter methylation

(A) A schematic of the allele-specific expression (ASE) analysis strategy implemented on TCGA breast cancer samples. In brief, imputed genotyping data, tumor RNA-seq and cases where tumor-normal matched RNA-seq are assessed for gene-level ASE by calculating the allelic imbalance ratios for imputed and phased heterozygous single-nucleotide polymorphisms (SNPs). We report differential ASE (diffASE) between tumors and matched normals and ASE in tumors (tumorASE) in cases when matched normals are unavailable.

(B) diffASE events in breast cancer tumors exceed the background. A diffASE event is called between a tumor and its matched normal when the allelic ratio between them is $p < 0.001$ using a chi-squared test and the skew is greater in the tumor. Six hundred thirty two diffASE events were obtained when the diffASE events calculated with the actual sample identities were compared to the background obtained with 10,000 permutations of randomized normal/tumor identities (FDR<0.05 and greater ASE in the tumor than the matched normal, n = 92; for clarity, only 100 permutated data are displayed in the figure). The FDR reflects the proportion of permutations where the most significant diffASE event was obtained with the actual tumor/normal data.

(C) > 90% of diffASE events originate in breast cancer tumors. The overlap of diffASE events with ASE events in tumors and matched normal by individual ($p < 0.001$, binomial distribution, n = 92).

(D) In BRCA tumors, most ASE is not correlated with CNV. The Pearson correlation between linear regression of gene-level tumorASE and the absolute tumor CNV signal is significant ($R = 0.11$, $p = 1 \times 10^{-82}$, n = 92) but does not explain the majority of ASE. This analysis includes every gene exhibiting ASE (binomial test, $p < 0.001$) in an individual tumor and excludes all others. The CNV signal intensity is obtained from CNV microarrays. Only 10% of the genes are depicted for clarity.

(E) Gene-level diffASE is weakly correlated with the promoter (±2 kb from TSS) methylation (Pearson's linear correlation $R = -0.01$, $p = 2.4 \times 10^{-6}$, n = 92). As in Figure 1D, all genes exhibiting ASE (binomial test, $p < 0.001$) in an individual tumor were included. The methylation beta value is the ratio of methylated to total probe intensity. Only 10% of the genes are depicted for clarity.

See also Figures S1–S3.

coding drivers likely explains this result and necessitates an enrichment strategy for variants that are likely to have a functional impact.

The vast majority of previously validated non-coding driver mutations occur in 3′ UTRs, promoters, enhancers, and CTCF binding sites (Table S1). As these collectively encompass major sites of transcriptional regulation, we focused on somatic variants within these features and refined them using several publicly available annotation resources (see Transparent methods). To comprehensively map genomic regions where transcription is regulated, we also included an aggregate map of TF binding sites ('TF binding') and accessible chromatin (see Transparent methods, Table S3). For the enrichment analysis, we grouped the somatic mutations in these CREs by regulatory feature and asked if they were 10 kb upstream of a TSS or gene body of a gene exhibiting ASE. Since the active enhancers effecting target gene transcription vary by cell type and frequently regulate non-adjacent genes, we used the regulatory relationships previously defined for distinct cancer types by the association of chromatin accessibility and gene expression (cancer-specific) (Corces et al., 2018). Using active elements defined in cells matching the query tumor further reduces the search space and improves sensitivity by focusing on the active subset of enhancers in the query tumor (Perera et al., 2014). Putative drivers were identified by positive correlations between gene-level ASE and somatic regulatory mutations within a cis-regulatory feature (see Transparent methods and Figures 2A, S4A, and S5). This approach revealed candidate non-coding driver mutations regulating genes including some that had been previously implicated in breast cancer by coding variants.

Using these features, we found ten genes that are enriched for somatic mutations in regulatory elements that coincide with altered *cis*-regulation in breast cancer (FDR<0.25, n = 113). These include mutations in the CTCF binding sites of *DAAM1*, variants in the promoters of *UNC5B*, and in TF binding sites near *EHMT1, FHIT, GSN, ITPR3, NCOA3, TIMP3, VPS13B,* and *ZDHHC14* (Figure 2B). The most significant association was between variants in TF binding sites adjacent to *TIMP3* and its altered *cis*-regulation (Figure 2C). TIMP3 is an inhibitor of matrix metalloproteinases whose upregulation suppresses tumor growth (Anand-Apte et al., 1996). In the 3 tumors harboring enhancer mutations, dysregulation is evident from the ASE in mutated tumors compared to the remainder (Figure 2C, FDR = 0.11, $p = 1.4x10^{-8}$, Wilcoxon rank-sum test). Somatic non-coding mutations associated with dysregulation of *ITPR3* were the next most confident finding; their distribution in relation to *ITPR3* is shown in Figure 2D (FDR = 0.15, $p = 1.4 \times 10^{-6}$). ITPR3 mediates the release of intracellular calcium in response to IP3 (Yamamoto-Hino et al., 1994). It was recently implicated as the target of the tumor suppressor BAP1 that triggers apoptosis following exposure to genotoxic stress (Bononi et al., 2017). Gene set enrichment analysis revealed enriched interactions between the established breast cancer pathway and the genes dysregulated by these putative non-coding driver mutations (Figure S6, p = 0.044, KEGG 'Breast Cancer').

Somatic mutations in regulatory features are enriched for gene-level ASE in diverse tumors

To identify relevant non-coding somatic mutations in other cancer types, we applied our pipeline to 11 other cancer types that had a sufficient number of matched WGS, RNA-Seq, genotyping, and cancer-specific chromatin accessibility data (derived by ATAC-seq (Corces et al., 2018)) (Figure S1A). Overall we identified 320 mutations in 47 CREs associated with ASE of a nearby gene (Figures 2B and 3, Tables 1 and S4, FDR<0.25). We will collectively refer to these 47 mutated CREs as the "putative drivers". The top ranked putative driver by FDR was *SEMA6D* in stomach adenocarcinoma (STAD) (FDR = 0.01). *SEMA6D* promotes survival and anchorage independent growth of malignant pleural mesothelioma (Catalano et al., 2009). The second ranked putative driver by FDR was *CBLB* in acute myeloid leukemia (LAML) (FDR = 0.04). *CBLB* is an E3 ubiquitin ligase previously implicated in myeloid malignancies that helps to attenuate proliferative signals transduced by activated receptor tyrosine kinases (Makishima et al., 2009). Variants in the CTCF bound regions of *CBLB* were prevalent, occurring in 12.2% of tumors (n = 5/41). Other notable examples of putative drivers based on prevalence include the CTCF bound regions of *FHIT* in BRCA (11.5%; n = 13/113) and the CTCF bound region of *SEMA4D* in lung squamous cell carcinoma (LUAD) (11.5%; n = 26/226). The putative drivers were generally associated with consistently skewed ASE across tumors (Figure S8). The transcript abundance of most genes exhibiting cis-dysregulation in association with somatic variants was unaffected (Figure 3F). Moreover, the coding regions of genes exhibiting altered cis-regulation are free of nonsense mutations, consistent with the ASE dysregulation occurring at the level of transcription rather than being a secondary consequence of nonsense-mediated decay. The majority of genes impacted by our putative drivers have been implicated previously in cancer and compelling cases for their driver mechanistic roles are explored further below (see Discussion).

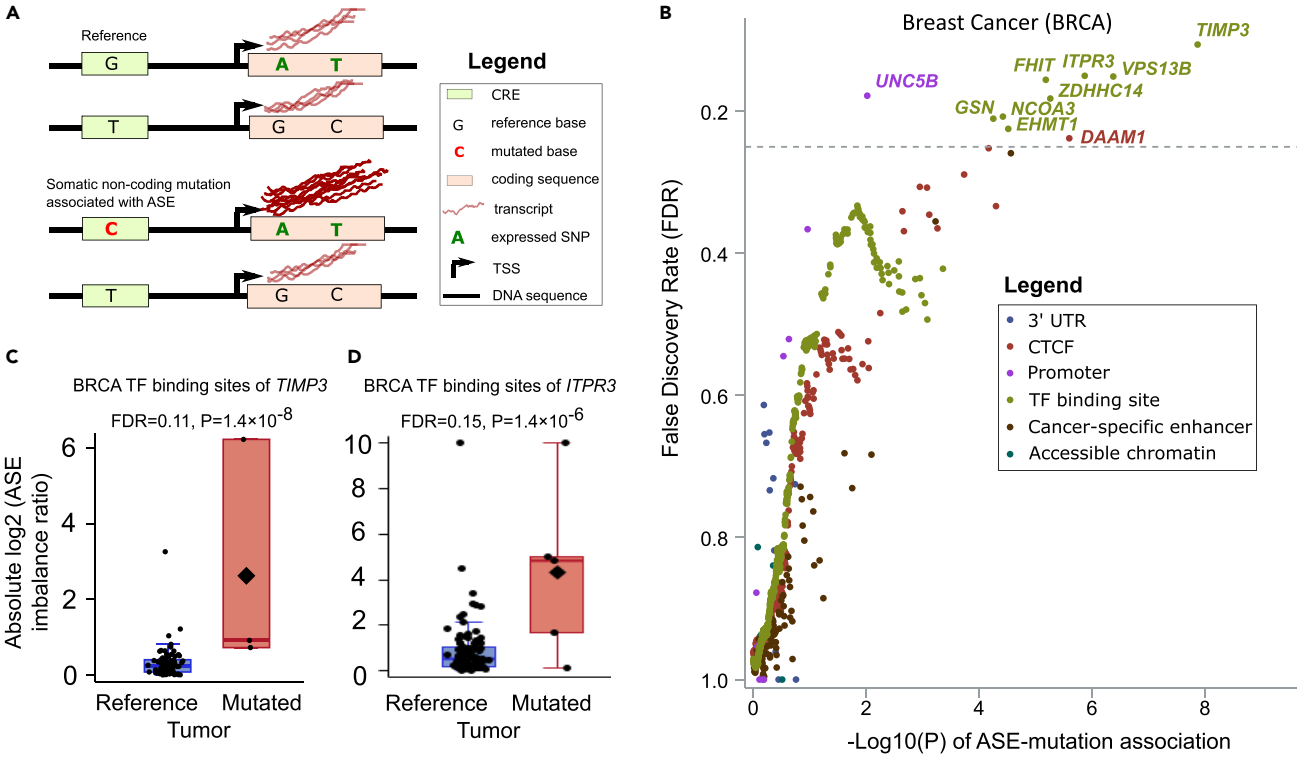


Figure 2. Seven regulatory features harboring somatic mutations are enriched for ASE in breast cancer

(A) A schematic of somatic mutations in cis-regulatory elements causing allele-specific expression.

(B) The significance of gene-level associations between mutated regulatory features and ASE in relation to FDR in breast cancer. The association of gene-level ASE was evaluated with a Wilcoxon rank-sum test (n = 113). The FDR is calculated as depicted in Figure S5 and detailed in the Transparent methods. The association was performed genome-wide.

(C) The ASE ratio of putative BRCA driver *TIMP3* ($p = 1.4 \times 10^{-8}$, FDR = 0.11, Wilcoxon rank-sum test, n = 113). The boxplot is delimited by the first and the third quartile, the whiskers encompass minimum and maximum data, while the diamonds and dots represent the medians and raw absolute gene-level ASE, respectively.

(D) The ASE ratio of putative BRCA driver *ITPR3* ($p = 1.4 \times 10^{-6}$, FDR = 0.15, Wilcoxon rank-sum test, n = 113). Boxplot features as in 'C.'

See also Figures S4–S6 as well as Table S4.

Candidate drivers have elevated variant allele frequencies

By definition, driver mutations confer a selective advantage to the cells in which they occur. Variant allele frequency (VAF) measures the fraction of alleles in a sample in which the variant is present. Hence, if a mutation confers a selective advantage to the cell in which it occurs, its VAF would be higher, on average, than passenger mutations that arose coincidentally. A corollary being that mutations with increased VAF occurred early enough during tumor evolution for this selective advantage to manifest as increased VAF. To ask whether the putative driver mutations conferred a selective advantage, we compared the normalized VAF of all putative drivers to all non-coding mutations that were not enriched for ASE ($p > 0.5$). As a positive control, we used known coding driver mutations (Schroeder et al., 2014). As expected, we found that the VAF of known coding drivers (n = 116) was, on average, higher than background mutations in coding regions (Figure 4A, p value = 2.9×10^{-6}, n = 2,971). Importantly, we found that the VAF of our candidate drivers was also higher (Figures 4A and 4B), an effect that is independent of CNV based on the stable ratio of adjacent heterozygous SNPs (Tables 1 and S4).

Candidate drivers disrupt transcription factor binding motifs

To further explore functional evidence supporting our non-coding driver mutations (Table 1), we asked whether they may be impacting DNA binding of transcription factors. Two features, specifically, might be expected to reflect this type of mechanism; TF binding sites and cancer-specific enhancers. We therefore limited our analysis to these features. Transcription factor binding affinities are typically represented by a generalized position-weight matrix (PWM) that represents a motif and a probability of observing any of

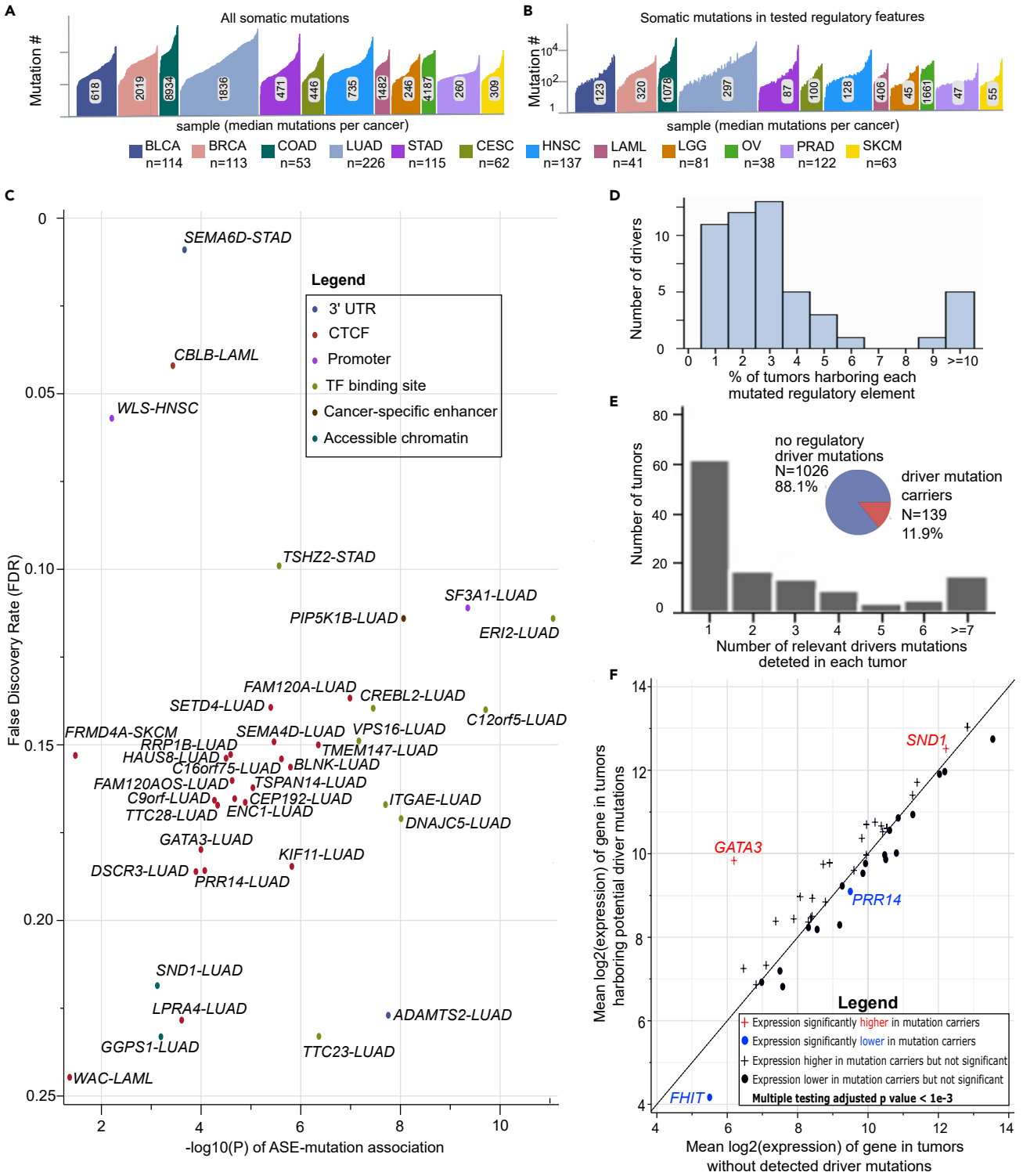


Figure 3. Thirty-nine regulatory features harboring somatic mutations are enriched for ASE across 11 additional cancer types

(A) The number of somatic mutations in each tumor as well as the median number for each type of cancer.

(B) The number of somatic mutations in regulatory features that were tested in the ASE-Mutation association in each tumor as well as the median number for each type of cancer.

Figure 3. *Continued*

(C) The significance of gene-level associations between mutated regulatory features and ASE in relation to FDR in 11 additional cancer types. The association of gene-level ASE was evaluated with a Wilcoxon rank-sum test (see Figure S5 and Transparent methods for detail and Figure S7D for 'n' evaluated in each cancer). The associations were performed genome-wide and independently on each cancer type.

(D) The percentage of tumors where each mutated regulatory element is found. The percentage of mutated regulatory elements is presented for the cancer type where mutated regulatory elements were associated with ASE.

(E) The distribution of mutated regulatory features across tumors. The inset illustrates that the majority of samples do not harbor driver mutations.

(F) The abundance of most putative driver genes does not change.

See also Figures S7 and S8 as well as Table S4.

the four bases at each position in that motif. These probabilities are typically constructed from observed frequencies of genuine binding events and can be represented by bit-scores. A bit-score of 2 implies that a particular base is always found at that position. We first scanned the genome for 392 transcription factor motifs and overlapped them with mutations (not filtered on ASE). Each motif had between 2 and 2,000 mutations (Figure 5A) and the number of mutations directly correlated with the number of motifs found in the genome (Figure 5B inset). As a preliminary analysis, we asked whether driver mutations (Table 1) are enriched in any putative transcription factor motifs. We observed a compelling enrichment of mutations among nucleotides important for transcription factor binding (Figure 5B). The challenge with relying on PWMs exclusively to identify transcription factor binding is that there is typically insufficient information to distinguish genuine binding sites from the many possible motif sequence matches in the genome. Here, we relied on the built-in component of our analysis to only consider mutations in functionally annotated regions. To further enrich for genuine transcription factor binding sites, we took advantage of ATAC-Seq data (Corces et al., 2018). Specifically, we considered motifs within open chromatin (i.e. in ATAC-Seq peaks) that were within 5 kb of a TSS. As expected the difference between reference and mutated bases tended to disrupt overall binding affinity (i.e. shift in all LUAD-ATAC-TSS-filtered mutations is overall negative; Figure 5C). Interestingly, driver mutations have an even stronger shift (Figure 5D). We did not see a significant difference in any other features, as expected, and also note that the numbers of driver mutations severely restricted our statistical power when exploring subsets of driver mutations outside of LUAD. It is also possible that improving binding (i.e. a positive delta-bit) could cause ASE, but we are not powered to explore this possibility.

The expression changes of putative driver genes are consistent with roles in tumorigenesis

Next, we considered how cis-regulatory drivers might contribute to tumorigenesis. The mechanism of non-coding drivers might be ectopic target expression or altered target expression kinetics; however, since altered target abundance is a common driver mechanism (Bailey et al., 2018), we leveraged DepMap to ask whether the upregulated target genes associated with predicted drivers were oncogenic and vice-versa (Figure 6A) (Meyers et al., 2017).

In an effort to reveal the contribution of genes targeted by predicted drivers to tumorigenesis, we asked how their deletion affects cellular expansion (fitness). As expected, we found that deletion of coding oncogenes decreased fitness and deletion of TSGs increased fitness (Figure 6B, 2-sided unpaired t test for unequal variance, $p < 1 \times 10^{-200}$ (oncogenes), $p = 9.2 \times 10^{-89}$ (TSGs)). Deletion of target genes associated with predicted drivers mirrored the effects of known coding drivers: deletion of upregulated genes led to decreased fitness, and fitness increased following deletion of genes downregulated in association with predicted drivers (Figures 6B–6D, 2-sided unpaired t test for unequal variance, $p = 4.9 \times 10^{-185}$ (oncogenes), $p = 1.7 \times 10^{-80}$ (TSGs)).

The deletion of target genes in cell lines harboring predicted driver mutations is consistent with tumorigenic roles. Evaluating the oncogenic and suppressive potential of predicted driver genes across all CCLE lines provided a robust measurement of each gene's fitness contribution; however it does not capture the role of target genes in lines where the target gene might be a driver. To identify CCLE lines potentially driven by the predicted non-coding drivers, we screened CCLE lines for identical mutations. In 3 of 4 cell lines harboring variants identical to predicted driver mutations, deletion of target genes had the expected impact on fitness (e.g. Figures 6C and 6D). While most drivers are not associated with a clinical outcome (Smith and Sheltzer, 2018), we also found predicted drivers associated with expression changes relevant to patient survival. For example, predicted driver variants decreased *SEMA4D* expression, which is marginally associated with worse overall survival (Figure 6E, $p = 0.06$). Collectively, these analyses are consistent with many predicted drivers promoting tumorigenesis by altering target abundance.

Table 1. Annotated catalog of the 47 putative driver hotspots

Cancer	Regulatory feature	Gene symbol	P	FDR	Carriers	ASE-CNV assoc. P
BRCA	CTCF	*DAAM1*	2.5×10^{-6}	0.24	5	8.3×10^{-1}
BRCA	Promoter	*UNC5B*	9.6×10^{-3}	0.18	4	1.0#
BRCA	TF binding site	*EHMT1*	3.0×10^{-5}	0.22	3	$*4.6 \times 10^{-2}$
BRCA	TF binding site	*FHIT*	6.6×10^{-6}	0.16	13	5.0×10^{-1}
BRCA	TF binding site	*GSN*	5.6×10^{-5}	0.21	3	3.6×10^{-1}
BRCA	TF binding site	*ITPR3*	1.3×10^{-6}	0.15	5	7.2×10^{-1}
BRCA	TF binding site	*NCOA3*	3.8×10^{-5}	0.21	3	1.0
BRCA	TF binding site	*TIMP3*	1.3×10^{-8}	0.11	3	4.7×10^{-1}
BRCA	TF binding site	*VPS13B*	4.2×10^{-7}	0.15	3	1.5×10^{-1}
BRCA	TF binding site	*ZDHHC14*	5.5×10^{-6}	0.18	6	4.8×10^{-1}
HNSC	Promoter	*WLS*	6.1×10^{-3}	0.06	3	$*6.7 \times 10^{-3}$
LAML	CTCF	*CBLB*	3.6 × 10-4	0.04	4	1.0#
LAML	CTCF	*WAC*	4.3×10^{-2}	0.24	3	1.0#
LUAD	3′ UTR	*ADAMTS2*	1.7×10^{-8}	0.23	4	3.6×10^{-1}
LUAD	Cancer-specific enhancer	*PIP5K1B*	8.6×10^{-9}	0.11	5	2.1×10^{-1}
LUAD	CTCF	*BLNK*	1.6×10^{-6}	0.16	6	4.7×10^{-1}
LUAD	CTCF	*C16orf75*	2.4×10^{-6}	0.15	5	1.0#
LUAD	CTCF	*C9orf95*	5.4×10^{-5}	0.17	8	1.0#
LUAD	CTCF	*CEP192*	9.1×10^{-6}	0.16	3	9.0×10^{-1}
LUAD	CTCF	*DSCR3*	1.2×10^{-4}	0.19	6	3.8×10^{-1}
LUAD	CTCF	*ENC1*	1.3×10^{-5}	0.17	6	3.2×10^{-1}
LUAD	CTCF	*ERI2*	6.3×10^{-8}	0.18	4	2.9×10^{-1}
LUAD	CTCF	*FAM120A*	1.0×10^{-7}	0.14	7	7.5×10^{-1}
LUAD	CTCF	*FAM120AOS*	2.3×10^{-5}	0.16	5	9.7×10^{-1}
LUAD	CTCF	*GATA3*	1.0×10^{-4}	0.18	5	9.5×10^{-1}
LUAD	CTCF	*HAUS8*	3.1×10^{-5}	0.15	2	2.3×10^{-1}
LUAD	CTCF	*KIF11*	1.5×10^{-6}	0.18	12	1.4×10^{-3}
LUAD	CTCF	*LPAR1*	2.4×10^{-4}	0.23	24	6.0×10^{-1}
LUAD	CTCF	*TMEM147*	4.4×10^{-7}	0.15	3	1.0#
LUAD	CTCF	*PRR14*	8.4×10^{-5}	0.19	5	1.0#
LUAD	CTCF	*RRP1B*	2.5×10^{-5}	0.15	3	1.0#
LUAD	CTCF	*SEMA4D*	3.4×10^{-6}	0.15	26	1.4×10^{-2}
LUAD	CTCF	*SETD4*	4.0×10^{-6}	0.14	5	1.3×10^{-2}
LUAD	CTCF	*TSPAN14*	2.1×10^{-5}	0.17	4	7.1×10^{-1}
LUAD	CTCF	*TTC28*	4.6×10^{-5}	0.17	8	7.0×10^{-2}
LUAD	Accessible chromatin	*GGPS1*	6.4×10^{-4}	0.23	3	4.7×10^{-1}
LUAD	Accessible chromatin	*SND1*	7.5×10^{-4}	0.22	3	8.8×10^{-1}
LUAD	Promoter	*SF3A1*	4.4×10^{-10}	0.11	3	5.8×10^{-1}
LUAD	TF binding site	*C12orf5*	1.9×10^{-10}	0.14	3	1.0#
LUAD	TF binding site	*CREBL2*	3.5×10^{-8}	0.14	4	7.7×10^{-1}
LUAD	TF binding site	*DNAJC5*	9.7×10^{-9}	0.17	3	8.9×10^{-3}
LUAD	TF binding site	*ITGAE*	1.9×10^{-8}	0.17	11	1.9×10^{-1}
LUAD	TF binding site	*TTC23*	4.3×10^{-7}	0.23	6	2.2×10^{-2}
LUAD	TF binding site	*VPS16*	6.8×10^{-8}	0.15	7	4.3×10^{-1}
SKCM	CTCF	*FRMD4A*	3.3×10^{-2}	0.15	3	$*3.7 \times 10^{-2}$

(*Continued on next page*)

Table 1. *Continued*

Cancer	Regulatory feature	Gene symbol	P	FDR	Carriers	ASE-CNV assoc. P
STAD	3′ UTR	*SEMA6D*	2.1×10^{-4}	0.01	3	3.9×10^{-1}
STAD	TF binding site	*TSHZ2*	2.7×10^{-6}	0.10	4	$*1.5 \times 10^{-2}$

Note: *Among the 7 genes where ASE is associated with CNV, tumors coincidently harboring driver mutations and CNV occurred in *SEMA4D*, *SETD4*, *DNAJC5*, and *TTC23* of LUAD and *TSHZ2* of STAD. *SEMA4D* was only nominally significant (p = 0.03) after exclusion of CNV carriers. *TSHZ2* had one tumor carrying both driver mutation and CNV; however, the association between mutations and ASE remained significant after excluding the tumor for *TSHZ2* ($p = 2.2 \times 10^{-5}$). *SETD4*, *DNAJC5*, and *TTC23* were not significant after excluding CNV carriers in the ASE-Mut association, with p values were 0.20, 0.48, and 0.76, respectively. #When a driver gene does not have multiple samples harboring CNV for the association test between ASE and CNV, the association is assigned as p = 1.

CNV, copy number variation; P, ASE-mutation association p value; ASE-CNV Assoc P, ASE-CNV association p value; FDR, false discover rate; Carriers, driver mutations carriers.

DISCUSSION

The association between mutated CREs and gene expression altered in cis revealed 47 clusters of non-coding mutations across 12 cancer types that exhibit the hallmarks of driver mutations. These 47 mutation hotspots significantly expand the landscape of putative non-coding cancer drivers. Prior approaches did not reveal the majority of our findings, although there is partial overlap with previous non-coding driver discoveries. For example, we found enriched *cis*-regulatory mutations in the CTCF binding sites of *DAAM1* in BRCA. DAAM1 is a member of the formin protein family activated by Disheveled binding (Liu et al., 2008). It regulates cytoskeletal dynamics through its control of linear actin assembly (Li et al., 2011). Regulatory mutations in *DAAM1* were recently implicated in invasiveness of melanoma (Zhang et al., 2018). Our findings also overlap previous reports in that somatic mutations in other regulatory regions of the same genes in the same type of cancer have been implicated as drivers. For example, mutations in the splice-acceptor site of *GATA3* were previously implicated in LUAD (Hornshoj et al., 2018). Here we implicated promoter mutations in *GATA3* in LUAD. This overlap suggests that the consequences of mutated regulatory features may overlap in these cases, and that combining the association of distinct features that regulate the same gene may increase sensitivity.

Many of the genes impacted by the putative non-coding drivers discovered here (Tables 1 and S4) have been previously implicated in cancer biology. The predicted non-coding drivers disproportionately impact established coding drivers (hypergeometric, p = 0.006). The target genes impacted by non-coding drivers include the COSMIC genes *CBLB*, *FHIT*, *GATA3*, and *SND1*. The genes associated with mutated CREs in BRCA illustrate how driver roles clearly tie into the established functions of the dysregulated genes. *NCOA3* is a transcriptional co-activator that is alternatively known as Amplified in Breast 1 (*AIB-1*) after its amplification and increased abundance was discovered in breast cancer (Anzick et al., 1997). *NCOA3* enhances estrogen-dependent transcription (Anzick et al., 1997). In this analysis, *NCOA3* was neither amplified nor elevated in abundance in the tumors with putative driver mutations, but our approach still implicated it in BRCA based on the association of somatic mutations with its dysregulation (Table 1). *EHMT1* provides insight into the observation that the total abundance of most genes dysregulated in association with non-coding mutations is unchanged. *EHMT1* represses transcription by methylating H3K9 residues in conjunction with *EHMT2* (*G9a* in mice) (Tachibana et al., 2005). *EHMT1/2* complexed with *E2F6*, and polycomb proteins preferentially occupy promoters in G_0 phase of the cell cycle and is associated with cellular quiescence (Ogawa et al., 2002; Tachibana et al., 2005), suggesting that non-coding mutations associated with *EHMT1* might disrupt its coordination with the cell cycle as opposed to its abundance. Conversely, *FHIT* is one of the few putative drivers that is dysregulated and differentially expressed. Consistent with the observed decrease in expression, *FHIT* is an established tumor suppressor gene (Waters et al., 2014).

Extensive cis-regulatory changes occur during tumorigenesis that are unrelated to copy number variation. In contrast with previous reports in different tumor types, CNVs were not responsible for the majority of altered *cis*-regulation in BRCA tumors (Mayba et al., 2014). Somatic regulatory variants are a major source of altered cis-regulation in tumors. Somatic variants in non-coding regions that are enriched for altered *cis*-regulation were found in 11.9% of the tumors analyzed. This high-prevalence is predicted by multi-hit models as well as divergent phenotypes between tumors with common known drivers. While many of the associations involve genes thought to be involved in tumorigenesis, the implication of specific mutations and regulatory features is a mechanistic advance. Indeed, we are not aware of any of the specific mutated regulatory features reported here previously being implicated as drivers of tumorigenesis.

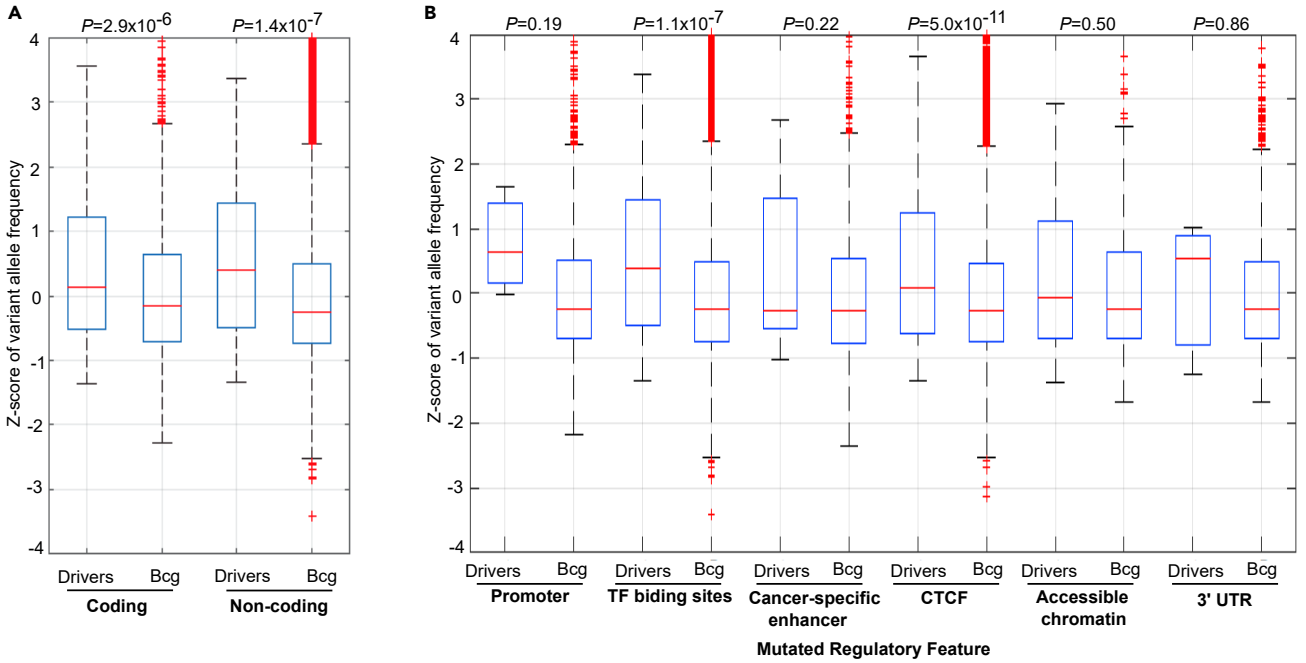


Figure 4. Variant allele frequency (VAF) of putative non-coding drivers suggests positive selection

VAF was calculated as the fraction of all sequencing reads covering variant with mutation and was normalized against all mutations within each patient to account for differences in tumor heterogeneity.

(A) Coding Drivers (n = 116) represent all mutations within known driver genes that yield a functional amino acid change and Coding Bcg (n = 2,971) represents identically selected mutations in all other coding genes (Schroeder et al., 2014). Putative non-coding drivers represent all mutations from Table S4 (n = 320) and Non-coding Bcg represents all non-coding mutations not enriched for ASE (n = 122,603). Both Coding and Non-coding VAFs are positively shifted relative to background ($p = 2.9 \times 10^{-6}$, $p = 1.4 \times 10^{-7}$, 2-tailed Student's t-test, equal variance). The boxplots are delimited by the first and the third quartile, red lines indicate medians, whiskers encompass minimum and maximum data and red points indicate outliers. Please see Transparent methods for more details.

(B) Same as (A) with putative non-coding drivers divided by feature. The number of putative non-coding driver and background mutations, as well as the p value (2-tailed Student's t-test, equal variance) comparing the VAF between putative driver and background mutations for each feature are: promoters (n = 10, n = 1,747, p = 0.19), cancer-specific enhancers (n = 5, n = 2,329, p = 0.22), CTCF binding sites (n = 208, n = 26,128, $p = 5.03 \times 10^{-11}$), TF binding sites (n = 88, n = 50,202, $p = 1.12 \times 10^{-7}$), accessible chromatin (n = 6, n = 583, p = 0.50), 3′ UTRs (n = 8, n = 1,661, p = 0.86).

Although TCGA and other emerging cancer data now include >1000 available genomes, illuminating the complete set of non-coding drivers will require a substantially broader collection. Even with the approach employed here of focusing on functional somatic variants with underlying evidence of gene expression regulation, we found ourselves limited by statistical power, especially in cancer types with fewer than 100 genomes. Deeper genome sequencing with longer reads will also improve driver detection sensitivity by enabling phasing of mutations with the direction of ASE. This would allow more evidence to be used to prioritize genuine drivers (e.g. disruption of an activating transcription factor binding site should reduce expression of that allele). This was generally not possible with the current available data since accurate phasing of somatic variants more than a few hundred base pairs away from the gene would require long-read technology or much deeper coverage. Improved matching of the regulatory features to each cell type will also improve sensitivity. When possible, cellular context was prioritized throughout these analyses to account for context-specific aspects of gene-regulation. For example, enhancers were matched to the cancer type being analyzed (Corces et al., 2018), and each cancer was separately analyzed in parallel, however, enhancer to gene maps are still incomplete and will no doubt improve with more chromatin accessibility readouts expand. In any case, we believe our approach here, made freely available as a dockerized pipeline (see Transparent methods) will be a powerful tool for taking advantage of these emerging resources and building on our discoveries.

Limitations of the study

The greatest limitation currently hindering our ability to apply our approach to broadly map cis-acting regulatory mutations across cancer is the limited availability of matched WGS and RNA-Seq data. Most RNA-

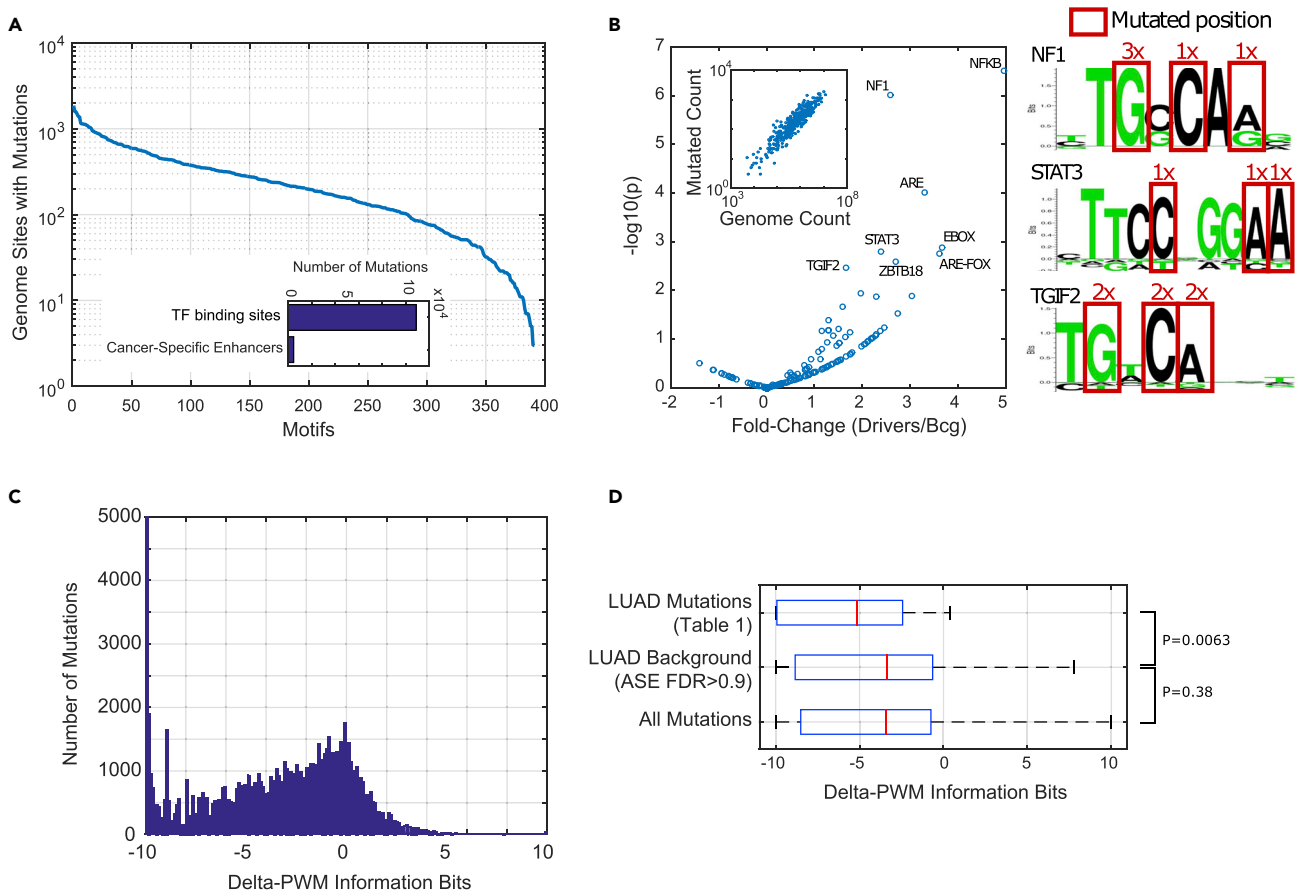


Figure 5. Driver mutations disrupt transcription factor binding motifs

(A) Number of overlapping mutation/motifs across 392 genome-mapped position-weight matrices (PWMs). TF binding site mutations outnumber cancer-specific enhancer driver mutations (inset).

(B) Putative driver mutations are overrepresented in several genome-mapped motif sets and disrupt important binding residues within the motifs of several transcription factors. The enrichment of mutations disrupting transcription factor binding motifs relative to background mutations was evaluated using a chi-squared test.

(C) The frequency at which mutations enhance or disrupt transcription factor binding motifs, evaluated as changes in PWM bits. Most mutations lead to a lower-affinity PWM.

(D) LUAD driver mutations (TF binding site and cancer-specific enhancers merged) within high-likelihood functionally bound motifs (ATAC-Seq support, within 5 kb of TSS; n = 18) result in a stronger shift than matched background (equivalent selection criteria except no association with ASE; n = 4,179). p values are two-tailed and computed using a Wilcoxon rank-sum test. The boxplots are delimited by the first and the third quartile, red lines indicate medians, and whiskers encompass minimum and maximum data.

Seq samples in TCGA do not have WGS data at all, and many WGS have relatively low-coverage data that makes identifying somatic mutations in regulatory regions difficult. Statistical power is simply not there to detect meaningful associations with just dozens (sometimes hundreds) of samples having WGS within each cancer type. A second, related limitation, is the current difficulty in phasing somatic variants captured only by WGS, into haplotypes against which ASE is ascertained. This adds complexity to characterizing the putative causal mechanism of ASE associated with a specific mutation. Long reads are a potential solution, but even increasing the coverage of paired-end short reads via WGS could dramatically improve phasing from overlapping reads.

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Brian DeVeale (brian.deveale@ucsf.edu).

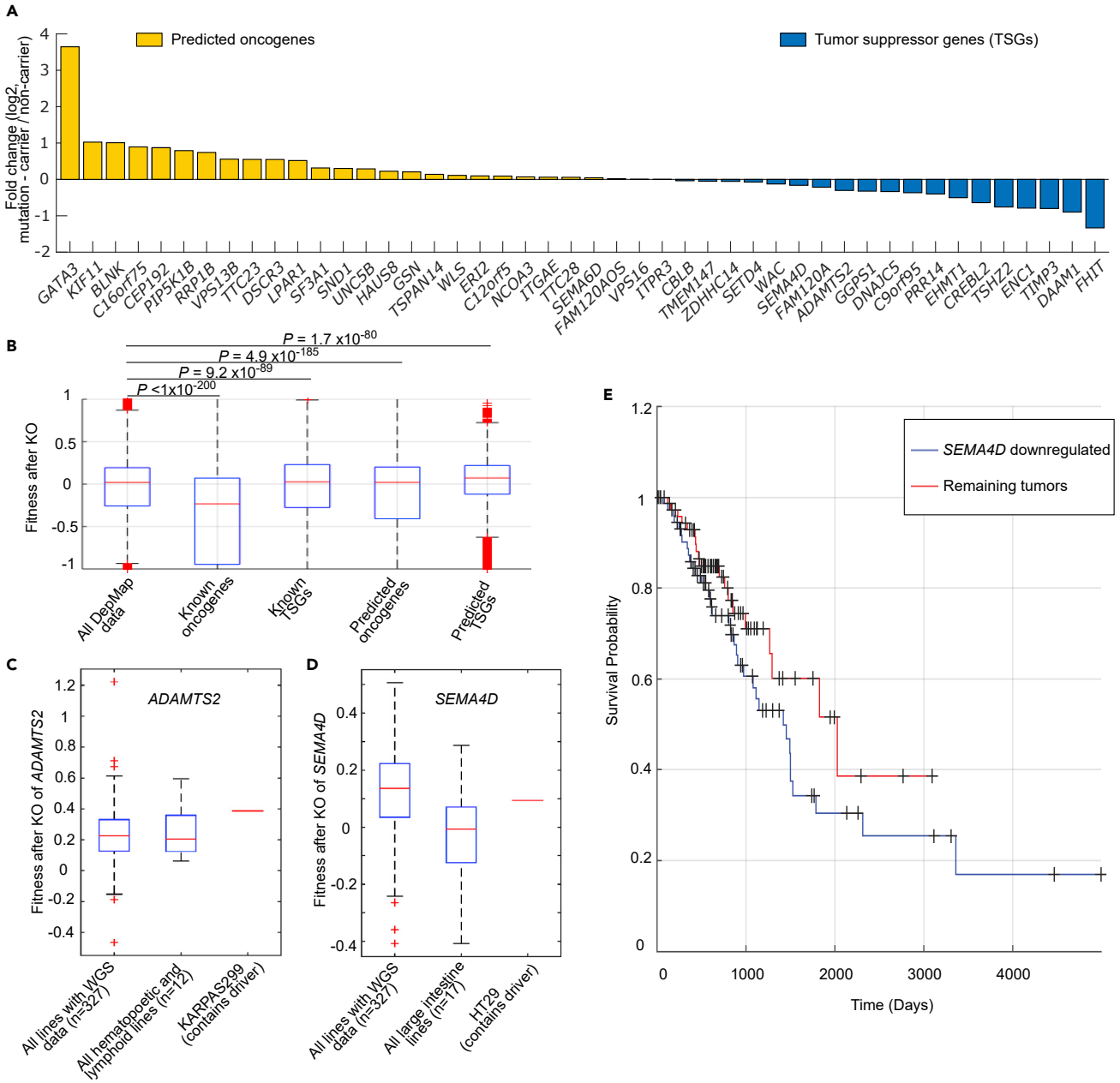


Figure 6. Transcript abundance of regulatory variant targets align with roles in tumorigenesis

(A) The expression change associated with putative driver mutations. The fold-change of target genes compares expression in tumors harboring putative drivers relative to the remaining tumors of the same type. Upregulated and downregulated target genes were inferred to be oncogenic and tumor suppressors respectively.

(B) The fitness impact of deleting predicted driver genes. Deletion of 40 coding oncogenes decreases fitness, while deletion of 91 coding. tumor suppressor genes (TSGs) increases fitness across 808 cell lines. Deletion of target genes associated with predicted non-coding drivers had equivalent fitness effects as coding drivers: deletion of putative oncogenes reduced fitness (n = 26) and increased fitness of putative TSGs (n = 20, 2-sided unpaired t test for unequal variance, p values are as shown). The boxplots are delimited by the first and the third quartile, red lines indicate medians, whiskers encompass minimum and maximum data and red points indicate outliers.

(C and D) The fitness impact of deleting putative TSGs *ADAMTS2* and *SEMA4D*. The boxplot features are the same as panel 'B'.

(E) Kaplan-Meier plot of LUAD patients (n = 515) with low expression of *SEMA4D* (n = 75) compared to the remaining patients. Low expression of *SEMA4D* is marginally associated with worse overall survival (p = 0.06; log rank test; '+' indicate censored data).

Materials availability

All of the data analyzed in this study was generated and made accessible by TCGA.

Data and code availability

We have made all of the code scripted and used in this analysis freely publicly available. Details are described in the Transparent methods section. Our Driver-ASE package is available via GitHub (https://github.com/MichealRollins-Green/Driver-ASE) and as a Docker image (https://hub.docker.com/r/mikegreen24/driver-ase). All raw gene-level ASE and somatic mutations called in this analysis can be accessed via Mendeley Data: https://data.mendeley.com/datasets/4kx5sfx9vz/2.

Driver-ASE uses data or software provided by the following websites:

UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTables), Genomic Data Commons (https://gdc.cancer.gov/), Genomic Data Commons (https://portal.gdc.cancer.gov/legacy-archive/search/f), The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov), PLINK (www.cog-genomics.org/plink), NIH Roadmap Epigenomics Mapping Consortium (www.roadmapepigenomics.org), SAMtools (www.htslib.org), overlapSelect (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/overlapSelect), Varscan2 (http://massgenomics.org/varscan), impute2 (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html), and shapeit (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html).

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102144.

ACKNOWLEDGMENTS

We would like to thank the funding of the Canadian Cancer Society, Canada, which enabled this research.

This study was supported by Canadian Cancer Society (CBCF grant BC-RG-15-2, PI: T.B.). No funding sources were involved in study design, data collection and interpretation, or the decision to submit the work for publication.

AUTHOR CONTRIBUTIONS

T.B. and B.D. designed the study. T.B, B.D., Z.C., M.V., and M.R. analyzed the data and wrote the manuscript.

Ethics approval and consent to participate.

We agreed to and followed the TCGA data use agreement.

Consent for publication TCGA obtained informed consent from participants for all of the donated specimens. TCGA uses the informed consent guidelines developed by NCI and NHGRI as detailed on their website.

DECLARATION OF INTERESTS

The authors have no conflicts of interest to declare.

Received: July 11, 2020
Revised: September 2, 2020
Accepted: January 25, 2021
Published: March 19, 2021

REFERENCES

Anand-Apte, B., Bao, L., Smith, R., Iwata, K., Olsen, B.R., Zetter, B., and Apte, S.S. (1996). A review of tissue inhibitor of metalloproteinases-3 (TIMP-3) and experimental analysis of its effect on primary tumor growth. Biochem. Cell Biol. *74*, 853–862.

Anzick, S.L., Kononen, J., Walker, R.L., Azorsa, D.O., Tanner, M.M., Guan, X.Y., Sauter, G., Kallioniemi, O.P., Trent, J.M., and Meltzer, P.S. (1997). AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. Science *277*, 965–968.

Babak, T., DeVeale, B., Tsang, E.K., Zhou, Y., Li, X., Smith, K.S., Kukurba, K.R., Zhang, R., Li, J.B., van der Kooy, D., et al. (2015). Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. Nat. Genet. *47*, 544–549.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. Cell *173*, 371–385.e18.

Beerenwinkel, N., Antal, T., Dingli, D., Traulsen, A., Kinzler, K.W., Velculescu, V.E., Vogelstein, B., and Nowak, M.A. (2007). Genetic progression and the waiting time to cancer. PLoS Comput. Biol. *3*, e225.

Bononi, A., Giorgi, C., Patergnani, S., Larson, D., Verbruggen, K., Tanji, M., Pellegrini, L., Signorato, V., Olivetto, F., Pastorino, S., et al. (2017). BAP1 regulates IP3R3-mediated Ca(2+) flux to mitochondria suppressing cell transformation. Nature *546*, 549–553.

Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550.

Cancer Genome Atlas Research Network. (2015). The molecular taxonomy of primary prostate cancer. Cell *163*, 1011–1025.

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. *69*, 6660–6667.

Catalano, A., Lazzarini, R., Di Nuzzo, S., Orciari, S., and Procopio, A. (2009). The plexin-A1 receptor activates vascular endothelial growth factor-receptor 2 and nuclear factor-kappaB to mediate survival and anchorage-independent growth of malignant mesothelioma cells. Cancer Res. *69*, 1485–1493.

Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898.

Foo, J., Liu, L.L., Leder, K., Riester, M., Iwasa, Y., Lengauer, C., and Michor, F. (2015). An evolutionary approach for identifying driver mutations in colorectal cancer. PLoS Comput. Biol. *11*, e1004350.

Fraser, H.B. (2011). Genome-wide approaches to the study of adaptive gene expression evolution: systematic studies of evolutionary adaptations involving gene expression will allow many fundamental questions in evolutionary biology to be addressed. Bioessays *33*, 469–477.

Fredriksson, N.J., Ny, L., Nilsson, J.A., and Larsson, E. (2014). Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat. Genet. *46*, 1258–1263.

Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X.J., Yip, K.Y., Khurana, E., and Gerstein, M. (2014). FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. *15*, 480.

Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Porta, M.G., Jadersten, M., Dolatshad, H., Verma, A., Cross, N.C., Vyas, P., et al. (2015). Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. Nat. Commun. *6*, 5901.

Group, P.T.C., Calabrese, C., Davidson, N.R., Demircioglu, D., Fonseca, N.A., He, Y., Kahles, A., Lehmann, K.V., Liu, F., Shiraishi, Y., et al. (2020). Genomic basis for RNA alterations in cancer. Nature *578*, 129–136.

Hornshoj, H., Nielsen, M.M., Sinnott-Armstrong, N.A., Switnicki, M.P., t Juul, M., Madsen, T., Sallari, R., Kellis, M., Orntoft, T., Hobolth, A., et al. (2018). Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. NPJ Genom Med. *3*, 1.

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

Kalender Atak, Z., Imrichova, H., Svetlichnyy, D., Hulselmans, G., Christiaens, V., Reumers, J., Ceulemans, H., and Aerts, S. (2017). Identification of cis-regulatory mutations generating de novo edges in personalized cancer gene regulatory networks. Genome Med. *9*, 80.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M.A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. Nat. Rev. Genet. *17*, 93–108.

Kulik, G.I., Pel'kis, F.P., and Korol, V.I. (1989). Adaptation of the body to alkylating anti-tumor substances. Eksp. Onkol. *11*, 34–38.

Li, D., Hallett, M.A., Zhu, W., Rubart, M., Liu, Y., Yang, Z., Chen, H., Haneline, L.S., Chan, R.J., Schwartz, R.J., et al. (2011). Dishevelled-associated activator of morphogenesis 1 (Daam1) is required for heart morphogenesis. Development *138*, 303–315.

Liu, W., Sato, A., Khadka, D., Bharti, R., Diaz, H., Runnels, L.W., and Habas, R. (2008). Mechanism of activation of the formin protein Daam1. Proc. Natl. Acad. Sci. U S A *105*, 210–215.

Makishima, H., Cazzolli, H., Szpurka, H., Dunbar, A., Tiu, R., Huh, J., Muramatsu, H., O'Keefe, C., Hsi, E., Paquette, R.L., et al. (2009). Mutations of e3 ubiquitin ligase cbl family members constitute a novel common pathogenic lesion in myeloid malignancies. J. Clin. Oncol. *27*, 6109–6116.

Mathelier, A., Lefebvre, C., Zhang, A.W., Arenillas, D.J., Ding, J., Wasserman, W.W., and Shah, S.P. (2015). Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. Genome Biol. *16*, 84.

Mayba, O., Gilbert, H.N., Liu, J., Haverty, P.M., Jhunjhunwala, S., Jiang, Z., Watanabe, C., and Zhang, Z. (2014). MBASED: allele-specific expression detection in cancer tissues and cell lines. Genome Biol. *15*, 405.

Melton, C., Reuter, J.A., Spacek, D.V., and Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat. Genet. *47*, 710–716.

Merid, S.K., Goranskaya, D., and Alexeyenko, A. (2014). Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. BMC Bioinformatics *15*, 308.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat. Genet. *49*, 1779–1784.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54.

Ogawa, H., Ishiguro, K., Gaubatz, S., Livingston, D.M., and Nakatani, Y. (2002). A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells. Science *296*, 1132–1136.

Ongen, H., Andersen, C.L., Bramsen, J.B., Oster, B., Rasmussen, M.H., Ferreira, P.G., Sandoval, J., Vidal, E., Whiffin, N., Planchon, A., et al. (2014). Putative cis-regulatory drivers in colorectal cancer. Nature *512*, 87–90.

Perera, D., Chacon, D., Thoms, J.A., Poulos, R.C., Shlien, A., Beck, D., Campbell, P.J., Pimanda, J.E., and Wong, J.W. (2014). OncoCis: annotation of cis-regulatory mutations in cancer. Genome Biol. *15*, 485.

Piraino, S.W., and Furney, S.J. (2017). Identification of coding and non-coding mutational hotspots in cancer genomes. BMC Genomics *18*, 17.

Poulos, R.C., Sloane, M.A., Hesson, L.B., and Wong, J.W. (2015). The search for cis-regulatory driver mutations in cancer genomes. Oncotarget *6*, 32509–32525.

Puente, X.S., Bea, S., Valdes-Mas, R., Villamor, N., Gutierrez-Abril, J., Martin-Subero, J.I., Munar, M., Rubio-Perez, C., Jares, P., Aymerich, M., et al. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature *526*, 519–524.

Schroeder, M.P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2014). OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. Bioinformatics *30*, i549–555.

Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., et al. (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268–274.

Smith, J.C., and Sheltzer, J.M. (2018). Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. Elife *7*, e39217.

Smith, K.S., Yadav, V.K., Pedersen, B.S., Shaknovich, R., Geraci, M.W., Pollard, K.S., and De, S. (2015). Signatures of accelerated somatic evolution in gene promoters in multiple cancer types. Nucleic Acids Res. *43*, 5307–5317.

Svetlichnyy, D., Imrichova, H., Fiers, M., Kalender Atak, Z., and Aerts, S. (2015). Identification of high-impact cis-regulatory mutations using transcription factor specific random forest models. PLoS Comput. Biol. *11*, e1004590.

Tachibana, M., Ueda, J., Fukuda, M., Takeda, N., Ohta, T., Iwanari, H., Sakihama, T., Kodama, T., Hamakubo, T., and Shinkai, Y. (2005). Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3-K9. Genes Dev. *19*, 815–826.

Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98–110.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Waters, C.E., Saldivar, J.C., Hosseini, S.A., and Huebner, K. (2014). The FHIT gene product: tumor suppressor and genome "caretaker". Cell. Mol. Life Sci. *71*, 4577–4587.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat. Genet. *46*, 1160–1165.

Yamamoto-Hino, M., Sugiyama, T., Hikichi, K., Mattei, M.G., Hasegawa, K., Sekine, S., Sakurada, K., Miyawaki, A., Furuichi, T., Hasegawa, M., et al. (1994). Cloning and characterization of human type 2 and type 3 inositol 1,4,5-trisphosphate receptors. Recept. Channels *2*, 9–22.

Zhang, W., Bojorquez-Gomez, A., Velez, D.O., Xu, G., Sanchez, K.S., Shen, J.P., Chen, K., Licon, K., Melton, C., Olson, K.M., et al. (2018). A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. Nat. Genet. *50*, 613–620.

iScience, Volume *24*

Supplemental information

Cis-regulatory mutations with driver hallmarks in major cancers

Zhongshan Cheng, Michael Vermeulen, Micheal Rollins-Green, Brian DeVeale, and Tomas Babak

Supplementary figures and legends

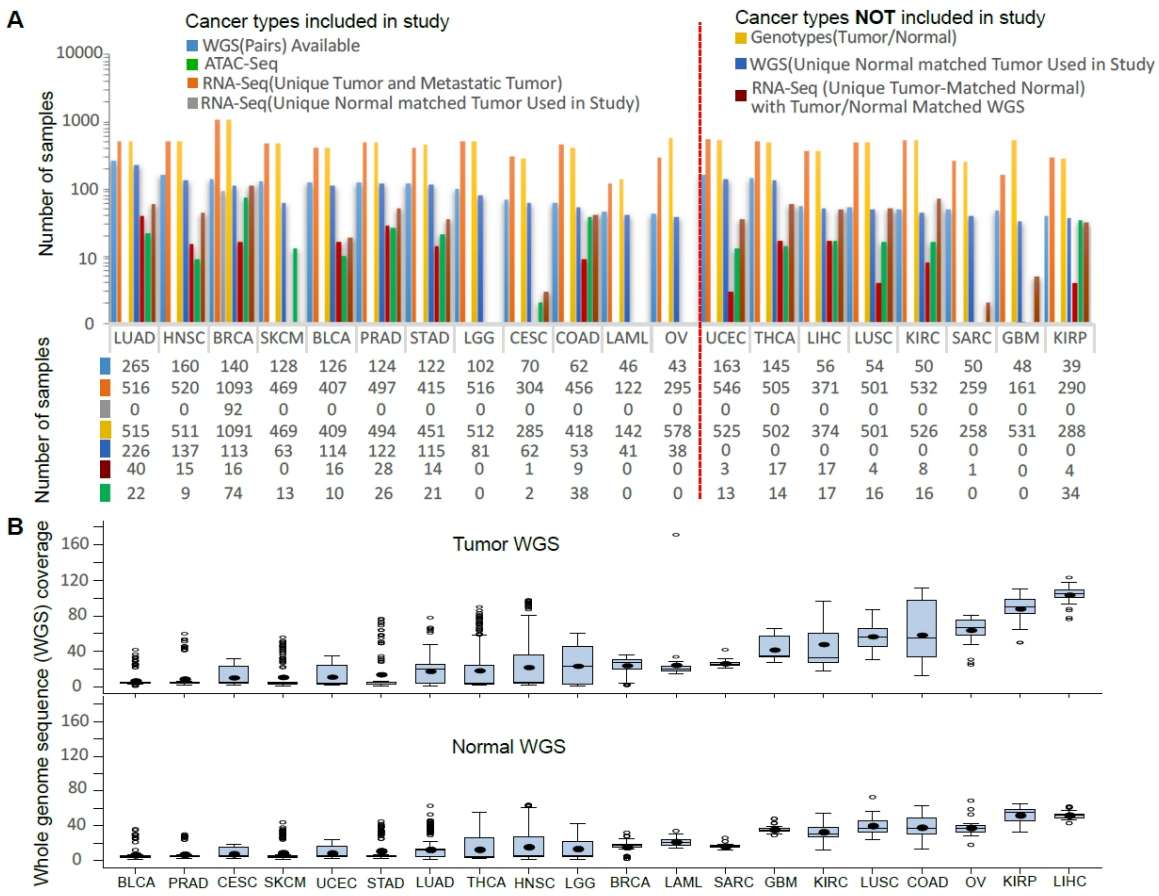


Figure S1. TCGA sample summary for ASE analysis. Related to Figure 1. (A) A summary of relevant TCGA samples available for analysis. (B) Whole genome sequence (WGS) coverage for tumor and matched normal across 20 cancer types (estimated based on bam size of WGS).

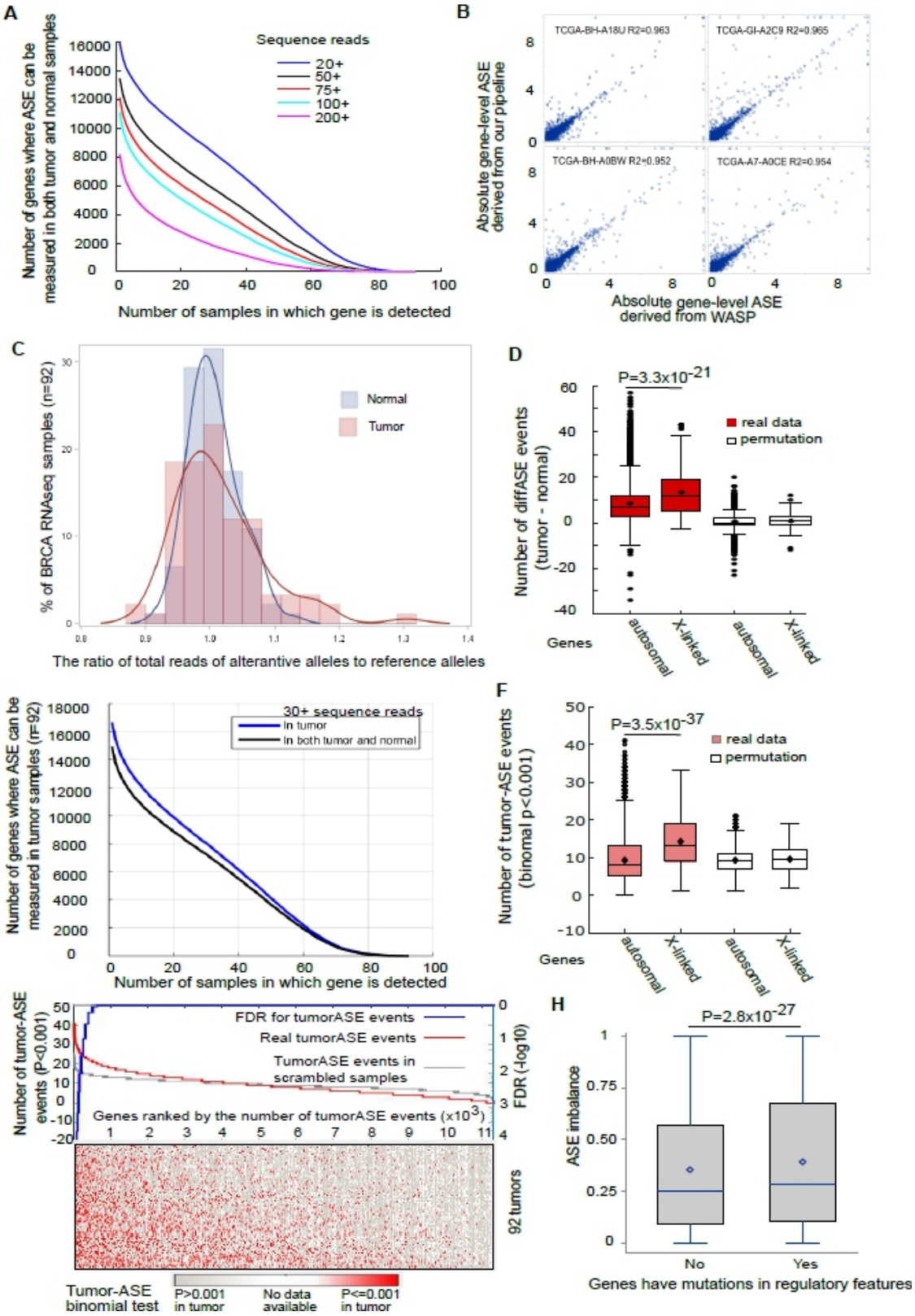

A
Number of genes where ASE can be measured in both tumor and normal samples
Sequence reads
20+
50+
75+
100+
200+
Number of samples in which gene is detected
B
TCGA-BH-A18U R2=0.963
TCGA-GI-A2C9 R2=0.965
TCGA-BH-A0BW R2=0.952
TCGA-A7-A0CE R2=0.954
Absolute gene-level ASE derived from our pipeline
Absolute gene-level ASE derived from WASP
C
% of BRCA RNAseq samples (n=92)
Normal
Tumor
The ratio of total reads of alterantive alleles to reference alleles
D
P=3.3x10-21
real data
permutation
Number of diffASE events (tumor - normal)
Genes
autosomal
X-linked
autosomal
X-linked
E
Number of genes where ASE can be measured in tumor samples (n=92)
30+ sequence reads
In tumor
In both tumor and normal
Number of samples in which gene is detected
F
P=3.5x10-37
real data
permutation
Number of tumor-ASE events (binomal p<0.001)
Genes
autosomal
X-linked
autosomal
X-linked
G
Number of tumor-ASE events (P<0.001)
FDR for tumorASE events
Real tumorASE events
TumorASE events in scrambled samples
Genes ranked by the number of tumorASE events (x10^3)
FDR (-log10)
92 tumors
Tumor-ASE binomial test
P>0.001 in tumor
No data available
P<=0.001 in tumor
H
P=2.8x10-27
ASE imbalance
No
Yes
Genes have mutations in regulatory features

Figure S2. diffASE and tumorASE events exceed background in BRCA tumors with matched RNA-seq samples (n=92). Related to Figure 1. (A) The number of genes where diffASE can be calculated per sample at various read thresholds. (B) The concordance in gene-level ASE when TCGA aligned and WASP-filtered reads are input to our pipeline (van de Geijn et al., 2015). The Pearson correlation is shown for four representative BRCA RNA-Seq samples and evaluated on genes with ≥30 reads. (C) The distribution of the ratio of reference:alternative aligned reads per sample. (D) X-linked diffASE events are enriched relative to autosomal events. The number of X-linked diffASE events exceeds the number of autosomal events in real, but not permuted data (P=3.3x10^{-21}, ANOVA). (E) A comparison of the number of genes where diffASE and tumorASE (n=92 tumor/normal RNA-seq) can be calculated per BRCA sample at a threshold of 30 reads. (F) X-linked tumorASE events for the 92 tumor RNA-seq samples are enriched relative to autosomal events. The number of X-linked and autosomal tumorASE events (P=3.5x10^{-37}, ANOVA). (G) tumorASE events in BRCA (n=92 RNA-seq) exceed the number expected by chance due to the distribution of the data. tumorASE events are those where the allelic ratio is P<1x10^{-3} using a binomial test. 241 tumorASE events with FDR<0.05 were obtained when the tumorASE events calculated with the actual sample identities were compared to the background obtained with 10,000 permutations of randomized tumor identities. The FDR reflects the proportion of permutations where the most significant tumorASE event was obtained with the actual tumor data. (H) Gene-level ASE is imbalanced among those harboring mutations in CTCF or TF binding sites, promoters, 3' UTR, accessible chromatin, and cancer-specific enhancers (2.8x10^{-27}, ANOVA, n=113 samples where tumor RNA-seq and matched tumor/normal WGS data were available). Only 16 of these samples overlap with those used to compute diffASE in Fig. 1.

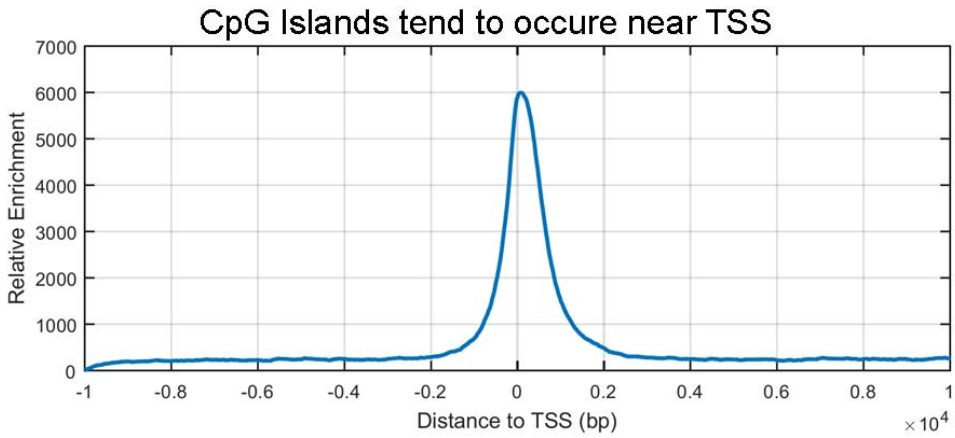


Figure S3. Most CpG islands are located within 2 kb of TSS. Related to Figure 1. The distribution of 52,383 CpG islands from the UCSC browser (hg19) plotted relative to TSS.

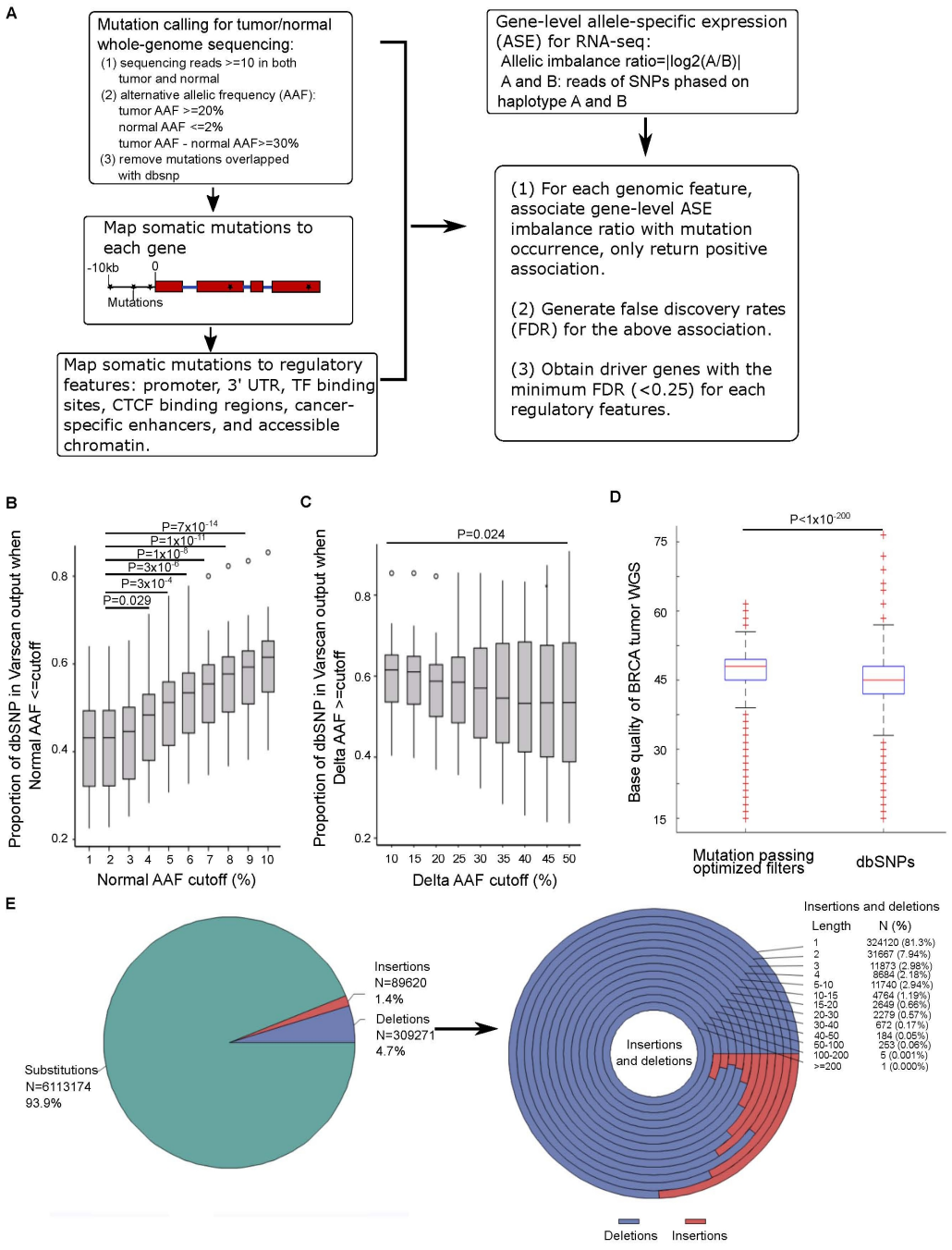


Figure S4. Optimized parameters deplete somatic mutation calls of rare germline SNPs. Related to Figure 2. (A) A schematic of the analysis workflow used to associate allelic imbalance with mutations in regulatory regions. Our analyses identified 47 drivers at a false discovery rate (FDR)<0.25 (see Table S4 for detailed output). 12 tumors were analyzed: breast cancer (BRCA), Bladder Urothelial *Carcinoma* (BLCA), Cervical Squamous Cell *Carcinoma* (CESC), Colon Adenocarcinoma (COAD), Head and Neck Squamous

Cell *Carcinoma* (HNSC), Acute Myeloid Leukemia (LAML), Low Grade Glioma (LGG), Lung Adenocarcinoma (LUAD), Ovarian cancer (OV), Prostate Adenocarcinoma (PRAD), Skin Cutaneous Melanoma (SKCM), Stomach Adenocarcinoma (STAD). (B, C) The effect of filtering parameters on the fraction of false-positive mutations (SNPs) called by Varscan2 was determined using BRCA WGS samples (n=48 matched tumor/normal BRCA WGS samples). (B) The proportion of false positive mutations called by Varscan2 (i.e. those that are actually SNPs) increases when normal AAF increases (normal AAF at a threshold of 2% was significantly different when compared to normal AAF thresholds ranging from 4% to 10%, n=48 BRCA WGS samples, Duncan's new multiple range test). (C) The proportion of SNPs among mutations called by Varscan2 decreases when the delta alternative allele frequency (delta AAF) between tumor and matched normal WGS increases (≥10 versus ≥50, P=0.024, n=48 BRCA samples, Duncan's new multiple range test). (D) The base quality score of mutations and dbSNPs called with the optimized filters in a representative BRCA samples (TCGA-CI-A2C9). (E) The total number of somatic mutations, including substitutions, insertions, and deletions for all 12 cancer types (left panel). The insertions and deletions grouped by the length of insertion and deletion (right panel). All of these somatic mutations were obtained by applying the optimized WGS filters listed in (A).

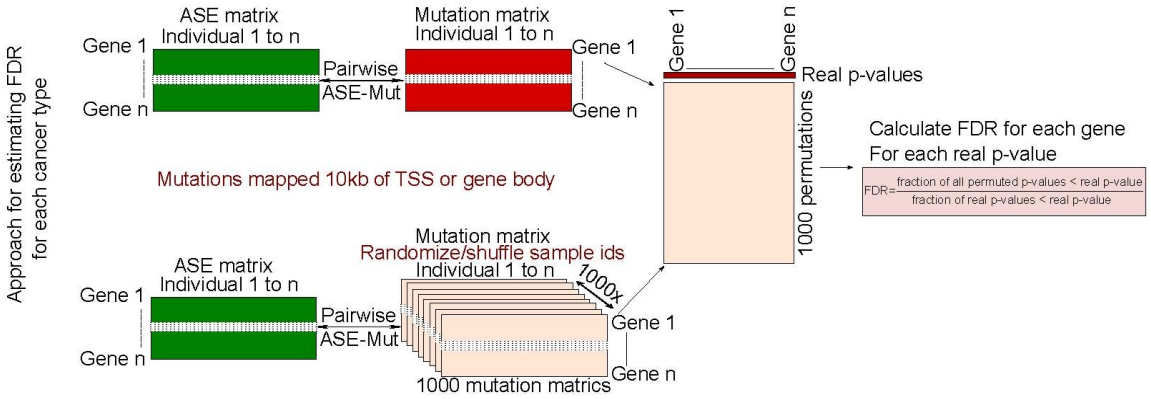


Figure S5. Schematic of the FDR calculation used to evaluate ASE-Mutation associations. Related to Figure 2.

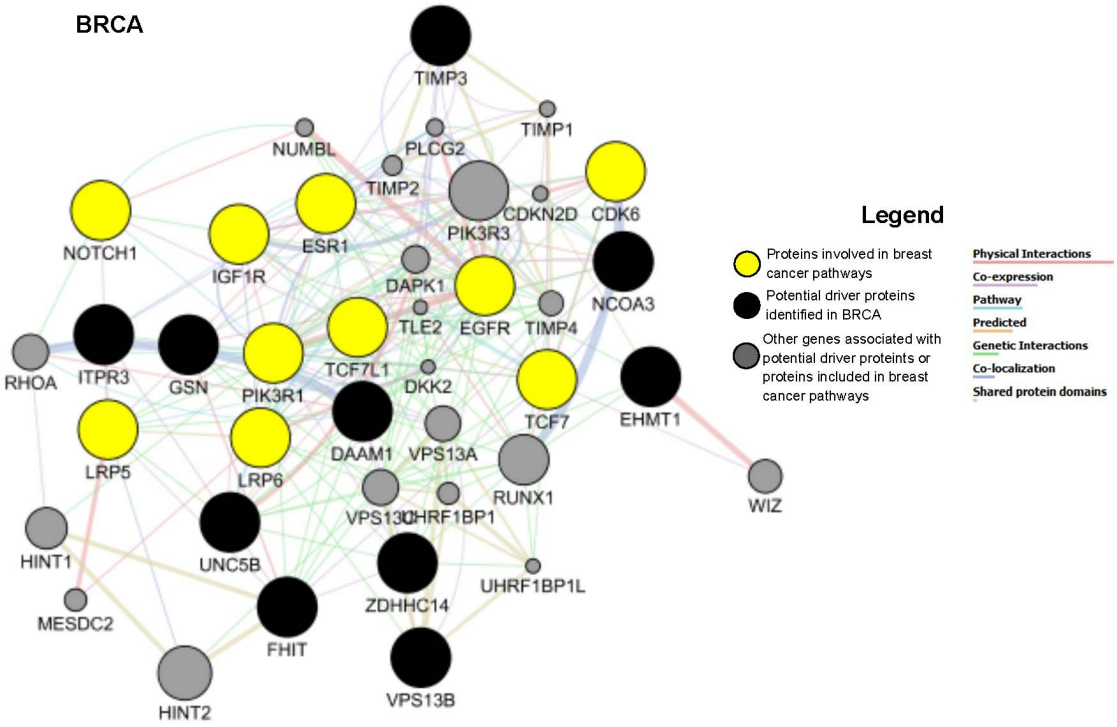


Figure S6. Network of putative and known breast cancer drivers. Related to Figure 2. The KEGG breast cancer pathway (hsa05224) was enriched in GSEA analysis of our predicted breast cancer drivers (*P*=0.044). To visualize the result, we integrated it with the putative non-coding drivers revealed by association of somatic mutations with ASE in this study. The networks were integrated using GeneMANIA (Warde-Farley et al., 2010) and visualized using Cytoscape 3.7 (Shannon et al., 2003).

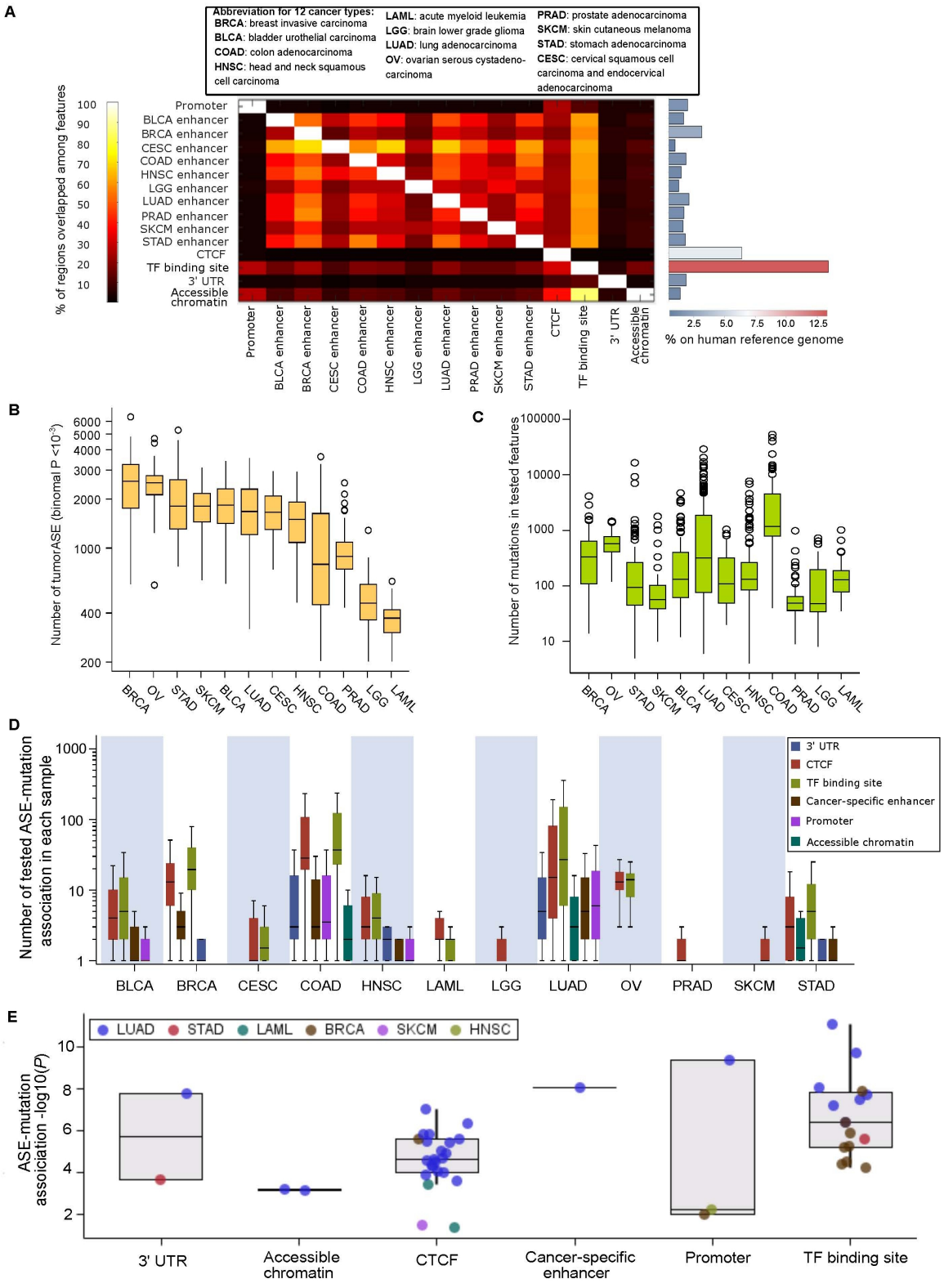


Figure S7. Characterizing six regulatory features and tumor ASE events, as well as mutations residing in these regulatory features, for 12 cancer types. Related to Figure 3. (A)

The heatmap displays the pairwise overlap among the six regulatory features analyzed: CTCF, TF binding sites (ChIP-seq), promoters, 3' UTR, accessible chromatin, and cancer-specific enhancers. Reading to the right of the diagonal indicates how much of the feature listed in each row is overlapped by the feature listed in each column. Conversely, to the left of the diagonal indicates how much of the feature listed in each column is overlapped by the feature listed in each row. For example, 20-70% of each enhancer track is overlapped by the CTCF and TF binding sites, but these enhancers occupy <10% of the CTCF and TF binding sites. The bar plots for each feature on the right side of the heatmap demonstrates the % of the human reference genome that the feature occupies. (B) The average number of tumorASE (binomial $P<10^{-3}$) events per tumor among 1,165 tumor RNA-seq samples plotted by the 12 types of cancers. (C) The average number of somatic mutations residing in the four regulatory features among 1,165 tumor/normal WGS divided into the 12 cancer types. (D) The number of somatic mutations in each regulatory feature distinguished by different colors for each of the 12 cancer types. In each boxplot, the horizontal line represents the median. The boxes are delimited by the first and the third quartile. (E) The distribution of putative driver significance plotted by the type of regulatory feature that is mutated.

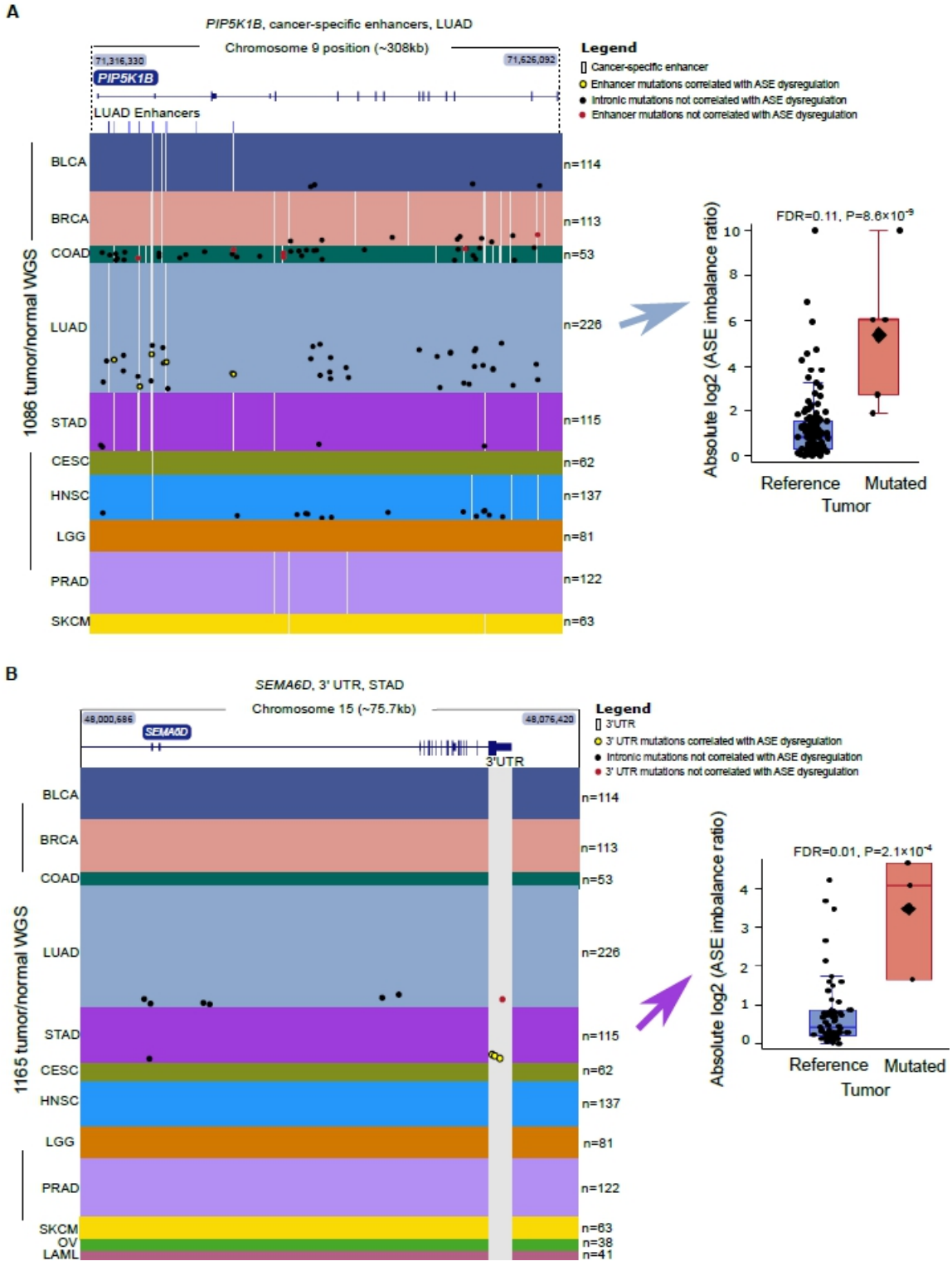


Figure S8. Non-coding variants have consistent impact on cis-regulation of *PIP5K1B* and *SEMA6D*. Related to Figure 3. (A) Left panel: The distribution of putative LUAD driver mutations in cancer-specific enhancers of *PIP5K1B* across 10 cancer types available of cancer-

specific enhancer features (n=1,086 tumors). Cancer-specific enhancers were not available for OV and LAML. Each row represents a sample and each column represents 1 bp. Right panel: The *PIP5K1B* ASE ratio in LUAD (P=8.6×10^{-9}, FDR=0.11, Wilcoxon rank sum test, n=225). (B) As in 'A' but depicting the distribution of putative STAD driver mutations in the 3'UTR of *SEMA6D* across 12 cancer types (n=1,165 tumors). Right panel: The *SEMA6D* ASE ratio (P=2.1×10^{-4}, FDR=0.01, Wilcoxon rank sum test, n=112). The boxplot is delimited by the first and the third quartile, while the diamonds and dots represent the medians and raw absolute gene-level ASE, respectively.

Transparent Methods

Genotyping and imputation Genome-wide Affymetrix 6.0 genotype array datasets from normal blood samples were downloaded as Birdseed files from GDC Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive/search/f) for all 5,875 patients from 12 cancer types. Among these patients, only the 1,165 where tumor RNA-seq and matched tumor/normal WGS data were available were included in the downstream association between gene-level ASE and mutation occurrence (see Fig. S1A). These datasets were annotated with Affymetrix annotation files and converted into base-level genotypes. To minimize allelic mapping bias we excluded SNPs with more than 2 polymorphisms or those where 2 SNPs conflicted at the same site on the same strand in phased 1000 Genomes Project Phase1 v3 data. Affymetrix 6.0 arrays genotype nearly 1 million SNPs. Typically ~25% of these sites are heterozygous and only a small fraction falls within expressed regions (mean=12,468). To increase the number of SNPs available to resolve ASE, we imputed and phased genotypes as previously described (Kulik et al., 1989). In brief, genotyping data were transformed into PLINK binary format and subjected to pre-phasing with Shape-IT software (v2.r790) (Gerstung et al., 2015) using the 1000 Genomes Project Phase1 v3 data as the reference, then imputed and phased using Impute2 software (v2.3.2) (Verhaak et al., 2010). We imputed with default parameters and used phased 1000 Genomes Project Phase1 v3 data as the reference panel. For each individual, heterozygous SNPs with genotype probability ≥ 0.95 were retained, as well as the allelic status within phased haplotypes. This provided an average of 25% more heterozygous SNPs per individual.

RNA-seq Matched tumor/normal RNA-Seq BAM files were downloaded from GDC Legacy Archives for all 1,165 patients across these 12 cancer types. SAMtools (Network, 2015) mpileup was used to calculate the reads at each heterozygous SNP site in all the RNA-seq BAM files. All these RNA-seq BAMs were aligned by TCGA (Nik-Zainal et al., 2016) using Mapsplice (Network, 2014) against the GRCh37-lite human reference genome. Default SAMtools mpileup settings were used to count reads at heterozygous SNPs except for read depths exceeding 8,000, in order to reduce the bias of bases showing excessive depth and to conserve computational resources.

Gene-level ASE To generate gene-level ASE ratios, heterozygous SNPs in each individual were mapped to a custom human transcript track generated by aggregating Ensembl (v80), UCSC and NCBI transcripts. Gene-level ASE was calculated by summing allelic counts from all heterozygous SNPs within the same haplotype by gene (Kandoth et al., 2013). Notably, the increased number of expressed heterozygous SNPs provided by imputation increased the

proportion of genes assayable (≥30 reads) for gene-level ASE from 50% to ~90%. To determine the impact that aligning heterozygous alleles to the reference genome had on evaluation of ASE, we assessed how alignment filtered by WASP influenced gene-level ASE (Schroeder et al., 2014). RNA-seq BAM files were aligned to hg19 using Mapsplicer and then reads at heterozygous exonic SNPs were counted using Samtools mpileup. Then we filtered reads with alignment bias using WASP and compared gene-level ASE between pre- and post-filter inputs among genes with ≥ 30 reads (Fig S2B).

Differential ASE To distinguish somatic from physiological ASE (random monoallelic silencing, imprinting etc.) we performed differential (diff)ASE analysis. diffASE is the difference in gene-level allelic bias between the normal and the tumor expression profiles (e.g. 50:50 versus 60:40, Chi-squared P=0.1). An event was either tumor- or normal-specific, depending on which sample deviated further from 50:50 allelic expression. diffASE events where the ratio between two alleles is more skewed in the tumor than in the normal are of primary interest. The FDR of diffASE events (tumor - normal) was calculated as the fraction of samples where diffASE events were not skewed in matched normal samples at a threshold P≤0.001. The FDR for the number of diffASE events arising in tumors was generated with 10,000 permutations of sample labels of tumor and normal samples, where each gene is evaluated relative to the permuted background of the same gene (sum(real diffASE events > permuted diffASE events)/10,000). For tumor samples where matched normal RNA-seq was not available, 'tumorASE' was assessed relative to the binomial distribution (P≤0.001). The FDR of tumorASE events was assessed with 10,000 permutations of the 92 samples (sum(tumorASE events, P≤0.001 > permuted ASE events, P≤0.001)/10,000. To compare the outcome of diffASE and tumorASE, we filtered for genes ≥30 reads in at least half of both the tumor and normal samples. The overlap between diffASE (Chi-squared P<0.001) and tumorASE (binomial distribution, P<0.001), is far greater than with normalASE (binomial distribution, P<0.001), indicating that >90% of diffASE originates in tumors (Fig. 1C). Thus we included gene-level ASE from tumor RNA-seq without a matched normal sample to dramatically increase the sample size (Fig. S1A).

Mutation calling To identify somatic mutations, we used Varscan2 (Vogelstein et al., 2013) in conjunction with custom filters. WGS data of 12 cancer types were downloaded from GDC Legacy Archive as 1,165 matched tumor/normal BAM files that were pre-aligned to the GRCh37 reference build (hg19) by TCGA using BWA (extensive sample information is available in Table S3 and Fig. S1A,B). We required that WGS samples had matched tumor/normal files, as well as corresponding genotype array, copy-number, and tumor RNA-seq data for inclusion. Additionally,

only WGS tumor/normal pairs aligned to the GRCh37 reference build (hg19) were included in our analysis. For SKCM, we used metastatic tumor samples (sample type 6 in the TCGA database), and primary tumor samples for the remaining cancer types (sample type 1). The sequence read counts at each site were obtained from WGS BAM files aligned to the GRCh37-lite human reference genome with the SAMtools (Vogelstein et al., 2013) mpileup. The base quality alignment (BAQ) computation of SAMtools was turned off with the parameter '-B' as it is too stringent for variant calling, and read map quality > 0 was set with '-q 1'. Single nucleotide substitutions, insertions and deletions were simultaneously called using Varscan2 somatic caller with the default base quality setting of >15 (Merid et al., 2014). Data were processed using a 1,052-core Linux cluster at the High-Performance Computing Virtual Laboratory (HPCVL) (Kingston, Ontario).

The parameters for identifying somatic mutations were focused on single-nucleotide substitutions, as well as small insertions and deletions as opposed to structural variations. Somatic mutations were generated by Varscan2 with default settings recommended by (Beerenwinkel et al., 2007), including minimum mutation frequency ≥0.2 and somatic p-value ≤0.05. Somatic mutations generated by Varscan2 were initially filtered with two criteria: (1) a minimum read depth of 10 for both the tumor and matched normal, and (2) alleles with a mutation frequency exceeding 0.2 in tumor but less than 0.1 in matched normal (Fig. S4A). However, since the mutations called with this approach included rare germline SNPs, we implemented custom filters to deplete them.

The custom filters were chosen based on optimization on 48 randomly selected BRCA tumor/normal WGS. Pearson correlation analysis was performed among 48 BRCA WGS samples to determine which Varscan2 parameters contributed to dbSNP150s being called mutations. These included the number of normal reads, tumor reads, alternative allele frequency (AAF) in tumors (Tumor AAF), AAF in normal (Normal AAF), and Delta AAF (i.e., Tumor AAF – Normal AAF). The extent of germline SNPs contaminating Varscan-called somatic mutations was assessed as a proportion and Pearson correlation R^2 relative to dbSNP150s from UCSC and mapped by genomic coordinates to both dbSNP alleles. This assessment was run among 48 randomly selected BRCA tumor/normal WGS. Using these metrics of contamination, each parameter (Tumor AAF, Normal AAF and Delta AAF was assessed through a range of values. This analysis revealed that two parameters, Normal AAF (R^2=0.37, $P<1\times10^{-4}$) and Delta AAF (R^2=-0.13, $P<1\times10^{-4}$), were primarily responsible for the high proportion of dbSNP150s in Varscan2.

This optimization supported the use of an AAF ≤ 2% in matched normal samples since the fraction of dbSNP150s increased when the AAF was higher (Fig. S4B). It also supported a

requirement that the AAF exceed 20% in tumors. Finally, we required that the difference between the AAF in tumors and matched normal samples exceed 30% based on the optimization (Fig. S4C). Notably, the base quality of mutations and dbSNPs called with these filters were equivalent, supporting the veracity of the mutation calls (Fig. S4D). These filters yielded single-nucleotide substitutions, as well as small insertions and deletions (Fig. S4E).

To identify variants in the CCLE lines identical to those we predicted as non-coding drivers, we aligned reads using the default settings of bowtie2 and then called mutations using the default settings of Strelka2 (Kim et al., 2018).

Effect of copy number variations on gene-level ASE To evaluate the effect of CNV on gene-level ASE, raw Affymetrix CNV data were downloaded from GDC Legacy Archive for 1,091 BRCA tumors. CNV data were then annotated with 'GenomeWideSNP_6.cn.na35.annot.csv.zip' downloaded from Affymetrix home page and mapped to Ensembl genes. Gene-level CNV signals were calculated by averaging the signals of all CNVs mapped to the gene. Finally, the absolute CNV signal, |log2(CNV signal)-1|, for each gene was correlated with its corresponding value of gene-level allelic imbalance (reads ≥30) to determine the influence of CNV on ASE in tumors. We applied this process to 92 tumors analyzed in the diffASE analysis, calling significant associations between gene-level ASE and the absolute CNV signal when $P<0.05$.

To remove the confounding effect of CNVs among associations between gene-level ASE and regulatory mutations, we assessed the correlation of each gene associated in our analysis with CNV signal. We also determined the association between gene-level ASE and CNV exclusively among putative driver mutation carriers to differentiate the effect of these mutations from that of CNV on gene-level ASE. We applied this filter for BRCA and other 11 cancer types, and found that none of candidate driver genes displayed significant association ($P<0.05$) with CNV when only the samples containing candidate driver mutations were considered.

To ask whether the CNV contributed to the association between mutated regulatory features and ASE of individual genes, we asked if CNV and ASE were correlated for each putative driver (Table 1 and Table S2). ASE and CNV did not correlate across tumors for the majority of putative drivers (40 of 47). For the 7 genes where there was an association, we asked if it was dependent on CNV. The CNVs impacting *FRMD4A*, *EHMT1* and *WLS* did not occur in the same tumors as the somatic variants. Some CNV coincided with somatic variants in the tumors where association of *SETD4*, *DNAJC5*, *TTC23*, *SEMA4D* and *TSHZ2* (STAD) were found. Hence the association between mutations and ASE is independent of CNV in the majority (40/47) putative drivers.

Effect of methylation on gene-level ASE To determine the effect of methylation on gene-level ASE, we downloaded methylation beta values for 1,091 BRCA tumors from GDC Legacy Archive. These methylation data were converted into bed format and mapped to Ensembl genes. The average methylation beta value was determined for each gene including a 2 kb region upstream and downstream of each gene to encompass the promoter. ASE imbalance values were then correlated with the absolute methylation signal, |average methylation beta value - 0.5|, on a gene-by-gene basis to determine the influence of methylation on gene-level ASE. This analysis was applied to all genes for 92 tumor RNA-seq samples involved in the diffASE analysis. Significant correlations between gene-level ASE and the absolute methylation beta value were called at $P<0.05$.

Selection of *cis*-regulatory features We surveyed major *cis*-regulatory features for *cis*-regulatory variants. These included TF binding sites (Encode ChIP-seq peaks clustered V3, 2013), CTCF binding sites and DNase hypersensitive regions (both from GM12878 cells), and 3' UTRs that were all obtained from the UCSC database. The track of TF binding sites was downloaded from UCSC Genome Browser and derived by collapsing multiple ChIP-seq maps of TF binding (Table S4). We also interrogated promoters (Roadmap Project) and cancer-specific enhancers (Sjoblom et al., 2006). These were defined based on the presence of peaks: promoters were defined as H3K4me3+ regions (signal in ≥10/127 tissues/cell types from the NIH Roadmap Epigenomics Mapping Consortium), while cancer-specific enhancers were defined by association between accessible chromatin and gene-expression changes in specific cancers (Sjoblom et al., 2006).

Association of mutations and gene-level ASE Only samples with corresponding WGS, genotyping array and tumor RNA-seq data were included in the ASE-mutation association analysis. To test association between mutations and gene-level ASE across various regulatory features, these data were imported separately for each cancer type into MATLAB 2014a (The MathWorks Inc., Natick, MA, 2014) and analyzed as schematized in Figure S4A and S5.

First, the somatic mutations were mapped based on proximity to promoters and enhancers as well as other features, including TF and CTCF binding sites. Using these annotations, promoters comprise 1.6% of the genome, enhancers 1.4% (ranging from 0.8% in LGG to 2.7% in BRCA), TF binding sites 13.2% and CTCF binding 6.0% of the genome. Somatic mutations were binned as present (=1) or absent (=0) among the regulatory features. The overlap among these regulatory features ranges from 0.04% to 85% (Fig. S7A). For example, CTCF and TF binding sites occupy 20-70% of the enhancer feature, while the enhancers occupy <10% of the CTCF binding sites.

Second, somatic mutations, including single-nucleotide alterations, insertions and deletions were mapped to nearby genes. The genomic coordinates of each gene were defined as beginning 10 kb upstream of the TSS and gene body of each gene. These settings were applied to test mutation association within each regulatory region (see Fig. S7B-D for a summary of the number of mutations mapped to the different regulatory features). Third, each gene (*i*) containing somatic mutations and also with summed heterozygous SNP allelic counts ≥30 reads was analyzed for gene-level allelic imbalance (*P*) by using the read counts of the two haplotypes (H_a, H_b) of each gene (*i*), in each sample (*n*), with the following formula, $P_i=|\log2(H_a/H_b)|$, with $P_i>10$ assigned as 10. The gene-level ASE varies considerably between different cancer types (Fig. S7B, binomial distribution, $P<0.001$). To increase sample size in the association test, we included all gene-level ASE, regardless of their binomial p-values.

Finally, the significance of association between a gene's allelic imbalance and mutations in each annotated region was determined in MATLAB using a Wilcoxon rank sum test. To obtain robust results we only ran the association when both mutation carriers (n≥3) and non-carriers (n≥3) had gene-level imbalance values derived from summing ≥30 reads from all heterozygous sites. ASE events positively or negatively correlated with mutations were retained in the analysis to focus on allelic imbalance resulting from dysregulation.

We permutated the samples to determine the false discovery rate (FDR) for the association between gene-level ASE and the occurrence of somatic mutations in each genomic feature (Poulos et al., 2015; Weinhold et al., 2014). For mutations residing in specific genomic regions, all pairs of gene-level ASE and mutations were randomized 1,000 times to generate association p-values that reflect the distribution of mutations in each CRE with ASE of each gene in each cancer. The FDR for each gene was then calculated as depicted in Fig. S5. Regulatory features that could be associated with multiple genes were included in all possible associations. When independent mutations were found within the same feature of the same sample, they were collapsed to a single mutation for the association. Finally, if multiple regulatory features were enriched for ASE of the same gene, only the most significant association with the smallest FDR was retained. Hence the reported associations (FDR<0.25) are cases where the mutations in a particular CRE coincide with ASE of a nearby gene more frequently than expected by the distribution of mutations and ASE for a CRE/gene pair in a particular cancer.

VAF VAF was calculated independently for each mutation that we identified in WGS data as the fraction of all sequencing reads covering the variant that were mutated. VAF was z-scaled within each patient using all mutations detected in that patient in order to normalize out the effect of tumor heterogeneity. There were 105,826 mutations that met filters and 833 individuals. Putative

non-coding driver (FDR<0.25) and background sets (FDR>0.5) were selected on basis of association with ASE. Significance was assessed using a Student's t-test (2-tailed, equal variance).

Effect on transcription factor binding sites The human genome was scanned for transcription factor binding sites using HOMER (scanMotifGenomeWide.pl using default settings for 392 motifs in the HOMER package) (Melton et al., 2015). Motif-enrichment (Fig. 5B) was assessed using a chi-square test, comparing mutated/non-mutated motif counts between drivers (gene ASE FDR<0.25) and background (gene ASE FDR>0.9). All LUAD somatic mutations (Table 1) were overlapped with each motif and the difference in bit-scores (i.e. "delta-bit" using the PWM; maximum is 2) between the reference and mutated bases was calculated. Delta-bit scores associated with genes under ASE vs. no-ASE were compared. Only single nucleotide substitutions within 5 kb of a TSS and within open chromatin detected in at least one LUAD ATAC-Seq sample (using peak calls within (Mathelier et al., 2015)). A mutation could be considered more than once if two or more transcription factor binding motifs were present.

Driver impact on fitness. CRISPR pooled screening data and Cancer Cell Line Encyclopedia (CCLE) gene expression data was downloaded from DepMap portal. Expression data and pooled CRISPR fitness data was available for 16,863 genes across 808 cell lines. Fitness effects were only considered for expressed genes (log2(FPKM)>2; 43% of fitness data corresponds to genes that are not expressed and was not considered here). No other filtering was applied. Fitness effects for 91 known TSGs and 40 known oncogenes were compared with our predicted drivers (Schroeder et al., 2014). We used screen the WGS of CCLE cell lines for mutations identical to our predicted drivers (Cancer Cell Line Encyclopedia and Genomics of Drug Sensitivity in Cancer, 2015).

Driver-ASE We have implemented our analysis methods for gene-level ASE and somatic mutation calling into a Perl package named Driver-ASE, which is available at GitHub (https://github.com/MichealRollins-Green/Driver-ASE). All MATLAB scripts to test association between mutations and gene-level ASE are also included in Driver-ASE. All of the dependencies required to run Driver-ASE are contained in a Docker (http://www.docker.com) image found here: https://hub.docker.com/r/mikegreen24/driver-ase. Docker is required to run Driver-ASE and the instructions Docker installation can be found here: https://docs.docker.com/engine/installation. Instructions to set up a Docker image are in the description section of the Docker page.

Data availability Driver-ASE uses data or software provided by the following websites: UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTables), Cancer Genomics Hub

(https://cghub.ucsc.edu), Genomic Data Commons (https://gdc.cancer.gov), The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov), PLINK (www.cog-genomics.org/plink), NIH Roadmap Epigenomics Mapping Consortium (www.roadmapepigenomics.org), SAMtools (www.htslib.org), overlapSelect (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64.v287), Varscan2 (http://massgenomics.org/varscan), impute2 (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html) and shapeit (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html).

All raw gene-level ASE and somatic mutations called by Varscan2 can be freely accessed via Mendeley Data (https://data.mendeley.com/datasets/4kx5sfx9vz/1).

References

Beerenwinkel, N., Antal, T., Dingli, D., Traulsen, A., Kinzler, K.W., Velculescu, V.E., Vogelstein, B., and Nowak, M.A. (2007). Genetic progression and the waiting time to cancer. PLoS Comput Biol *3*, e225.

Cancer Cell Line Encyclopedia, C., and Genomics of Drug Sensitivity in Cancer, C. (2015). Pharmacogenomic agreement between two cancer cell line data sets. Nature *528*, 84-87.

Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Porta, M.G., Jadersten, M., Dolatshad, H., Verma, A., Cross, N.C., Vyas, P.*, et al.* (2015). Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. Nature communications *6*, 5901.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A.*, et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.

Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Kallberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P.*, et al.* (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods *15*, 591-594.

Kulik, G.I., Pel'kis, F.P., and Korol, V.I. (1989). Adaptation of the body to alkylating anti-tumor substances. Eksp Onkol *11*, 34-38.

Mathelier, A., Lefebvre, C., Zhang, A.W., Arenillas, D.J., Ding, J., Wasserman, W.W., and Shah, S.P. (2015). Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. Genome Biol *16*, 84.

Melton, C., Reuter, J.A., Spacek, D.V., and Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat Genet *47*, 710-716.

Merid, S.K., Goranskaya, D., and Alexeyenko, A. (2014). Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. BMC bioinformatics *15*, 308.

Network, C.G.A.R. (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543-550.

Network, C.G.A.R. (2015). The Molecular Taxonomy of Primary Prostate Cancer. Cell *163*, 1011-1025.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C.*, et al.* (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47-54.

Poulos, R.C., Sloane, M.A., Hesson, L.B., and Wong, J.W. (2015). The search for cis-regulatory driver mutations in cancer genomes. Oncotarget *6*, 32509-32525.
Schroeder, M.P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2014). OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. Bioinformatics *30*, i549-555.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.
Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N.*, et al.* (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268-274.
van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods *12*, 1061-1063.
Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P.*, et al.* (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98-110.
Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546-1558.
Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T.*, et al.* (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res *38*, W214-220.
Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet *46*, 1160-1165.