Allele-specific expression is selected in tumorigenesis, results from epigenetic changes and has prognostic relevance

Guillermo Palou-Márquez[1], Pere Pericot-Masdevall*[1], and Fran Supek*[1,2,3]

1 Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute for Science and Technology (BIST), 08028 Barcelona, Spain

2 Biotech Research and Innovation Centre, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

3 Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

* Authors contributed equally

Correspondence to: fran.supek@bric.ku.dk

Abstract

Allele-specific expression (ASE) is the differential abundance in levels of mRNAs that originated from the paternal and maternal copies of a gene. Such allelic imbalances can contribute to phenotypic variation and influence disease traits, including cancer. There is common ASE in tumors that results from somatic copy-number alterations (CNAs) at the DNA level, but there also exist other causes of ASE: *cis*-acting genetic or epigenetic variation that can lead to differential expression between the two alleles. However, the latter, non-CNA mechanisms of ASE remain understudied in cancer, as well as their role in tumor evolution and impact on clinical outcomes. By integrating a wide variety of genomic and transcriptomic pan-cancer data from the TCGA project, we show that ASE favoring the preferential expression of the mutant allele in some driver genes is subject to positive selection, and that these events are associated with worse overall survival across all cancer types. We found that the impact of ASE triggered by non-CNA causes is substantial, and we propose that some instances of *cis*-ASE are explained by the epigenetic changes affecting alleles differently. Furthermore, as a second mechanism, we find that splicing-altering mutations are selected in various cancer genes and result in ASE. We anticipate that the study and understanding of the role of mutant allele imbalances at the mRNA level can help understand epigenetic changes during cancer evolution, as well as identify new prognostic markers and therapeutic approaches that target altered allelic expression in tumors.

Introduction

In diploid genomes such as the human, each gene bears two copies referred to as alleles, each of which is inherited from one of the two parents. Allele-specific expression (ASE) is the differential abundance of mRNA transcribed from the two allelic copies of a gene[1], which are in general expected to be balanced. However, allelic imbalance of gene expression is a known occurrence in the human genome[2–5]. A well-known mechanism generating ASE in a defined set of genes is genomic imprinting, in which only a specific parental copy of affected genes is expressed and the other is silenced through epigenetic mechanisms[6]. In more general terms, allelic imbalances at the mRNA level are known to commonly arise from dosage differences at the DNA level resulting from copy number alterations (CNAs), which are fairly rare in the germline however occur commonly in somatic evolution in the context of cancer[7–10]. Other known causes of ASE include *cis*-acting genetic variation - occurrence of single nucleotide variants (SNVs) that alter gene regulatory elements leading to ASE, and can also alter splicing and/or trigger nonsense-mediated mRNA decay (NMD)[11] of the variant-bearing allele[4,5]. The comparison of allelic frequencies between the genome (DNA-level, representing CNA effects) and the transcriptome (RNA-level) can quantify the effect of *cis*-acting, non-CNA mechanisms causing ASE by altering transcription activity or transcript processing[12].

Allelic imbalances of gene expression can contribute to phenotypic variation, and are known to have an impact on disease traits[13,14], including cancer[8,9,15,16], where they are a common event[16]. ASE is anticipated to have a causal effect on tumorigenesis, as pervasive ASE has been observed in individual examples of cancer driver genes[17,18], and more systematically[19], which are commonly subject to CNA events. One example of how ASE could drive cancer would be in tumor-suppressor genes (TSGs), in which the recessive loss-of-function (LoF) mutations[20] may require the other allele to be silenced in some way. This is typically understood to happen via a loss of heterozygosity (LOH) event affecting gene dosage at the DNA level, leading to a significant functional impact[21], however it may in principle similarly occur via epigenetic silencing causing the second-hit affecting the gene.

As in TSGs, also in some oncogenes (OGs) the CNAs causing dosage increases at the DNA-level of oncogenic gain-of-function (GoF) mutant alleles confer a selective advantage to the tumor[10]. However, in contrast to ASE resulting from the well-studied CNA and LOH events in cancer genes, the ASE resulting from *cis*-acting effects of genetic or epigenetic somatic variation on tumor evolution remain understudied[17,22,23]. A particular hypothesis of interest is whether the non-CNA mechanisms generating mRNA-level allelic imbalances (i.e. ASE) can modulate selective pressures acting on driver genes, steering cancer evolution in an analogous way to CNA.

Here, by integrating whole-exome sequencing (WES) and RNA-seq data from tumors in 8,809 TCGA patients[24], spanning 31 different cancer types, we aim to gain insights into the selective pressures that act on the ASE in driver gene mutations. We find that non-CNA driven ASE is common in cancer genes and that it is under selection. We further investigate possible mechanistic causes, showing that allele-specific epigenetic effects often underlie the ASE, as well as prognostic utility where ASE on driver mutations is a predictor of aggressive tumors. Another cause of ASE are splicing-altering mutations in gene coding regions.

Results

Quantifying the effect of ASE in cancer

Throughout this study, we conceptualize that mechanisms generating ASE occur in two stages: i) CNA events that alter dosage at the DNA level (henceforth, CNA-ASE), and ii) other mechanisms altering allelic expression at the RNA level independently of DNA dosage, which may be caused by *cis*-acting genetic or epigenetic variation (henceforth, *cis*-ASE). The actual, observed differential abundance between allelic transcripts at the RNA level (henceforth, net-ASE), is therefore the composite of both CNA-ASE and *cis*-ASE. Thus, there are three different paths by which allelic imbalance could occur, with respect to a somatic mutation (schematic in Fig. 1A), involving imbalance of the mutant allele only at DNA level, only at mRNA level, or at both; the latter two cases constitute ASE (we note that the former, DNA level-only case, is here also referred to as "CNA-ASE", for consistency of terminology, although it does not involve expression as such). In these various mechanisms both the DNA and the mRNA levels of the allele imbalancing process could act independently of each other, or together, generating ASE of driver mutations.

Our approach consisted in separating the known effect of CNAs from the effects of *cis*-acting variation on overall ASE of each somatic SNV mutation in cancer exomes. We aim to quantify the relative contributions of the well-understood CNA-ASE and the less explored *cis*-ASE, and the implications of *cis*-ASE on cancer evolution via positive selection affecting driver genes. To this end, we first quantified the effect of CNA-ASE by calculating the log2 odds ratio (OR) determined between the expected allele counts given heterozygosity (50% corrected by tumor sample purity), and the DNA allele counts obtained from WES (see Methods; schematic in Fig. 1A). In other words, this measurement labeled CNA-ASE represents the imbalance at the DNA level only, and does not consider RNA levels. Next, we quantified the effect of *cis*-ASE by the log2(OR) between the DNA variant allele counts in WES data and the RNA-level allele counts obtained from RNA-seq data, in both cases normalizing the mutant allele by the wild-type allele. The summation of the two effects (the DNA-level – CNA-ASE – and RNA-level corrected by DNA-level – *cis*-ASE) estimates the net-ASE, or the overall imbalance observed in mRNA reads, normalized to the expectation given a heterozygous somatic SNV. Using matched exome and transcriptome data from 8,809 tumors spanning 31 cancer types from TCGA we obtained 854,785 total somatic mutations suitable for analysis (see Methods).

Non-CNA ASE contributes substantially to allelic imbalance in cancer transcriptomes

We hypothesized that mechanisms other than CNA may have a substantial impact on overall ASE on somatic variant-bearing alleles in the context of cancer. We quantified the overall contribution of this *cis*-ASE as in a recent study[18] (see Methods).

In a pan-cancer analysis, we estimated that *cis*-ASE (i.e. effects other than CNA) explained 55.6% of the total net-ASE variability (Fig. 1B). We observed a considerable amount of differences between cancer types regarding the contribution of *cis*-ASE on overall ASE variability, ranging from 17.3% to 70.2% (Fig. 1B). Among the cancer types where *cis*-ASE effect was lower and the CNA-ASE was the predominant mechanism underlying net-ASE was ovarian serous cystadenocarcinoma (OV), adrenocortical carcinoma (ACC), as well as thyroid carcinoma (THCA). On the other extreme, in cancers such as diffuse large B-cell lymphoma (DLBC) or

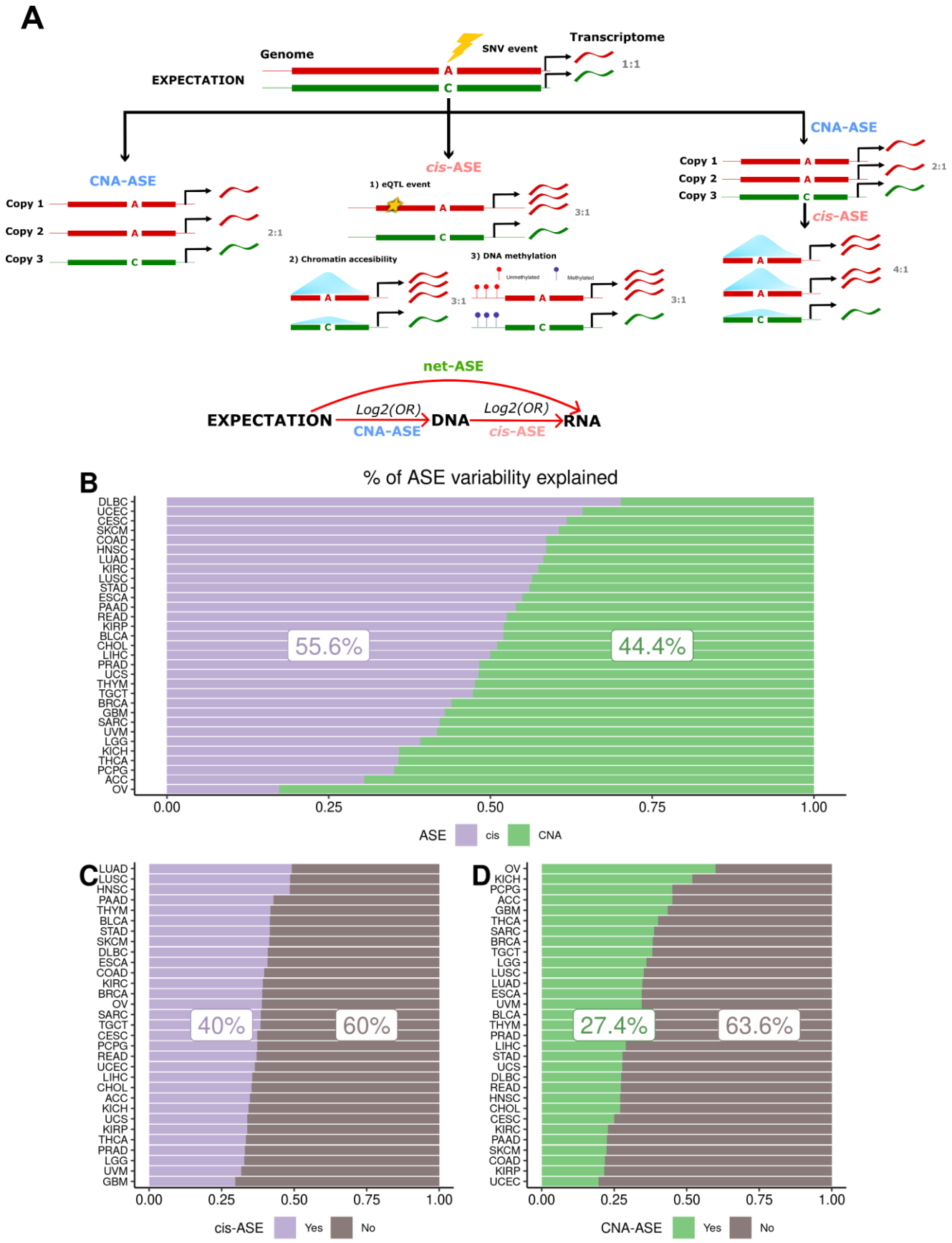
A
Genome
SNV event
Transcriptome
EXPECTATION
1:1
CNA-ASE
Copy 1
Copy 2
Copy 3
2:1
cis-ASE
1) eQTL event
3:1
2) Chromatin accesibility
3) DNA methylation
Unmethylated
Methylated
4:1
net-ASE
EXPECTATION
Log2(OR)
CNA-ASE
DNA
Log2(OR)
cis-ASE
RNA
B
% of ASE variability explained
55.6%
44.4%
ASE
cis
CNA
C
40%
60%
cis-ASE
Yes
No
D
27.4%
63.6%
CNA-ASE
Yes
No

Figure 1. Allele-specific expression (ASE) interpreted as a two-stage process
A, Illustration of three different paths by which ASE could favor the expression of the mutant allele after an SNV event. CNAs and *cis*-acting variation could either act additively (right panel), or independently of each other (left and middle). It includes a graphical summary (bottom) of our approach to quantify the respective effects of CNA-ASE and *cis*-ASE. The sum of these effects results in the total net-ASE observed at the RNA level, adjusted for sample purity. **B**, Relative contributions of *cis*-ASE and CNA-ASE to the overall net-ASE variation of all exome-wide somatic SNVs in TCGA, stratified by cancer type. It illustrates that ASE caused by non-CNA mechanisms has a major impact on overall ASE and is pervasive in cancer genomes (55.6% of variability). **C-D**, Percentage of exome-wide SNVs exhibiting a significant amount of *cis*-ASE (**C**) and CNA-ASE (**D**), determined as *abs(log2(OR)) > 1*, stratified by cancer type.

uterine corpus endometrial carcinoma (UCEC), *cis*-ASE had the most influence on mRNA imbalance in expression of somatically mutated alleles.

To further understand these differences, we assessed what proportion of exome-wide somatic SNVs showed an appreciable amount of *cis*-ASE, defined as *abs(cis-ASE) > 1* (i.e. two-fold difference between alleles), as well as CNA-ASE, defined as abs*(CNA-ASE) > 1.* We found that *cis*-ASE had an effect on 40.0% of total mutations, on average, across all TCGA cancer types (Fig. 1C), while CNA-ASE had an effect on 27.4% of mutations (Fig. 1D). Similarly as above, the proportions varied notably depending on cancer type: the proportion of somatic variants that were affected by CNA-ASE went as high as 60.0% in the case of OV, while this same number went as low as 19.7% in the case of UCEC (Fig. 1D). In comparison to CNA-ASE, the percentage of variants with *cis*-ASE was more consistent across cancer types, from 29.7% to 49.2%.

Our comprehensive analysis reports higher estimates than previous smaller scale studies. For instance, Correia et al.[18] found that *cis* effects comprised only 16% of the ASE in breast cancer, compared to 84% attributed to CNA. To validate these findings, we applied similar filters and data stratification, specifically analyzing only BRCA and *PIK3CA* somatic variants (see Methods). This approach yielded a qualitative agreement, with estimated 23% contribution from *cis*-ASE (Supp. Fig. S1A). Another study, the PCAWG transcriptomic analysis of 1,188 RNA-seq samples[8], calculated the lower-bound estimate of *cis*-ASE contribution at 15.7%. Their approach relies on a logistic regression model with explicitly provided *cis*-ASE putative causal variables, thus not capturing unknown effects on *cis*-ASE. In agreement, when we employed to our dataset a similar approach, meaning a linear regression model incorporating both CNA and multiple putatively causal non-CNA variables to predict our net-ASE (see Methods), we arrived at a comparable of 17.3% relative contribution for *cis*-ASE variables (Supp. Fig. S1B). These findings highlight the significant role of mechanisms other than CNA in generating ASE, and indicate that such *cis*-ASE is more prevalent in cancer genomes than anticipated[8,18].

To support the variability in CNA-ASE between cancer types (Fig. 1D), we correlated their ranks by their CNA-ASE variability with their total CNA burden (see Methods). This yielded a Spearman rank correlation coefficient of 0.29 (Supp. Fig. S2).

Selection on ASE as a driver of tumorigenesis

We hypothesized that ASE driven by somatic mutations provides a selective advantage during cancer evolution. For downstream analyses, we classified each somatic SNV in gene coding regions according to their effect, namely nonsense variants, missense variants, and effectively synonymous (ES) variants. The latter uses the pathogenicity predictor VARITY_ER[25,26] to

reclassify low-impact (non-pathogenic) missense variants (with score < 0.5) and merge them together with the synonymous variants into the ES group. This aims to increase the signal-to-noise in missense variants, while increasing sample size for the synonymous variants baseline group. Individual genes were also grouped as OGs (n = 233), TSGs (n = 246), according to Cancer Gene Census (CGC)[27]; we further had a group of passenger genes that were not in CGC to use as a negative control (n = 15,572). Essential genes were taken from the CEG2 set determined from CRISPR screens on cell lines[28] (n = 626). The *TP53* gene was treated independently, due to the large number of variants therein, which provided statistical power for individual analyses, as well as due to its particular relevance in the context of our study. We refer to "cognate genes" as those preferentially mutated in a specific cancer type, according to the MutPanning catalog of driver genes[29]. Therefore, driver genes are classified as cognate for some cancer types, while being noncognate for others.

Driver TSGs have an enrichment of missense variants with positive ASE

To evaluate whether ASE showed signs of selection acting upon its presence in cancer genes, we performed a conditional dN/dS test based on ASE (ASE-dNdES, see methods). In this test, we assessed the proportion of variants with significant ASE (either positive, >1, or negative, < -1, in separate tests) in missense (N) against effectively synonymous variants (ES), for each gene category (see Methods). For nonsense variants, we analyzed its proportions in comparison to the other two SNV types (Fig. 2A). As expected, variants showing negative ASE were enriched for nonsense mutations, relative to those showing no ASE or positive ASE (net-ASE: OR = 3.33; *cis*-ASE: OR = 5.49, both p < 2e-16, Fisher test). This is explained by the activity of NMD, a surveillance pathway that degrades mRNAs containing premature termination codons (PTCs), a known cause of ASE[30,31]. Further as expected, we do not find a nonsense enrichment when checking CNA-ASE, as its negative ASE is not due to NMD (NMD acts on the mRNA, not on DNA levels) but rather due to a copy number deletion of the mutated allele (OR = 0.87, p = 3.7e-24). Henceforth, we thus excluded nonsense mutations to avoid overt effects of NMD on ASE, and in subsequent comparisons we contrast the impactful missense mutations versus the ES mutations. Next, we considered passenger genes as a control, which showed no observable differences in enrichment between positive net-ASE and no net-ASE or negative net-ASE, for the missense mutations (p = 0.26, Fisher test). Also, as expected and regardless of ASE, cognate OGs and TSGs showed an overall increase in the proportion of missense variants over ES with respect to their noncognate pair (OR = 7.85 for OGs and 2.63 for TSGs, both p < 2e-16, Fisher test), which suggests selection on these missense variants.

More interestingly for our hypothesis, in the case of cognate TSGs, including *TP53*, missense mutations were considerably more prevalent within positive ASE than in the no ASE and negative ASE groups (net-ASE: OR = 4.65, p = 7.1e-151; *cis*-ASE: OR = 2.64, p = 7.8e-63; CNA-ASE: OR = 5.25, p = 5.9e-134). This same idea was further confirmed when quantifying the effect of ASE values stratified by SNV functional category: we observed a notable imbalance towards the mutated allele on missense variants compared to ES variants in the case of cognate TSGs and in the case of *TP53*, and similarly but more subtly so in the case of cognate OGs (Supp. Fig. S3A-C for net-ASE, *cis*-ASE and CNA-ASE separately). Additionally, we observed a modest correlation between ASE and the VARITY functional impact scores for missense mutations in cognate TSGs (net-ASE: r = 0.40, p = 8.2e-185; *cis*-ASE: r = 0.25, p = 3.0e-70; CNA-ASE: r = 0.36, p = 1.9e-143; *Pearson* correlation; Supp. Fig. S3D). The above suggests that missense variants in driver genes showing preferential expression of the mutant allele are subject to stronger positive selection than driver gene variants not showing ASE of the mutant allele,

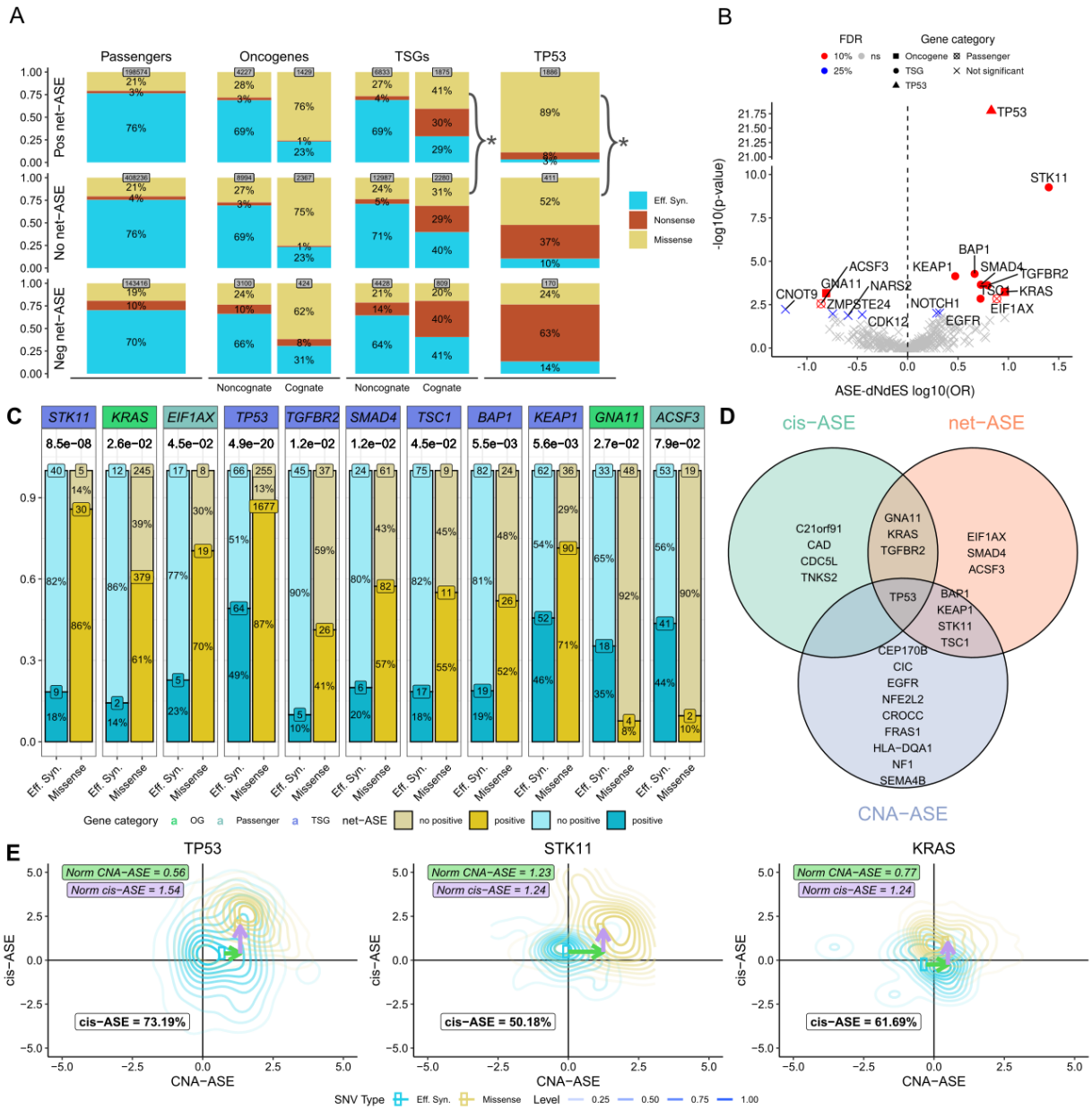

Figure 2. ASE is subject to selection in cancer genomes, potentially contributing to tumorigenesis

A, Relative frequencies of effectively synonymous (Eff. Syn or ES), missense and nonsense somatic SNVs, stratified by net-ASE status and gene categories: passengers, oncogenes (OGs), tumor suppressor genes (TSGs) and *TP53*. OGs and TSGs were further divided into cognate and noncognate. For clarity, "positive net-ASE" refers to mutations where ASE is greater than 1, "negative net-ASE" denotes ASE less than -1, and "no net-ASE" represents ASE values between -1 and 1. **B**, Volcano plot for the conditional gene-level ASE-dNdES test, displaying the odds ratios (ORs) on X-axis and raw *p*-values on Y-axis, both in log10 scale, highlighting significant genes under FDR < 10% or potential candidates at FDR < 25%. **C**, Proportion of SNVs with positive net-ASE in ES compared to missense variants for 11 significant genes from (B) that exhibit selection on the effect of net-ASE. The number of variants and the FDR is also shown. **D**, Overlap of significant hits from the ASE-dNdES test in which selection was inferred for either net-ASE, *cis*-ASE and CNA-ASE. **E**, Two-dimensional distribution densities of CNA-ASE and *cis*-ASE in ES variants and missense variants, for *TP53*, *STK11* and *KRAS*. The axial displacement of their centroids on one axis relative to the other represents the amount of selective contribution that each of the mechanisms (either CNA or *cis*) had over the final net-ASE.

highlighting the key role of ASE in tumor evolution. This test of selection, when applied to our CNA-ASE estimates, recapitulates the known effects whereby CNA affect selection on driver genes[10,32], and when applied to *cis*-ASE further suggests that the non-CNA causes of ASE can have a similar impact on selection on driver mutations as the CNA have.

ASE in known driver genes is subject to selection

To further identify individual driver genes subject to potential selection on positive ASE we again performed the ASE-dNdES but at individual gene level; here we considered jointly the no net-ASE and negative ASE.

We focused on known cancer genes from the CGC (excluding translocation-only genes) and additional genes identified by MutPanning[29] that are significantly mutated in two or more cancer types, leading to 330 genes (Subset 1). Furthermore, in addition to these known driver genes, we assessed all genes genome-wide for mean ASE differences (Subset 2), that is, evaluating whether the ASE on missense variants exceeded that on ES variants (across 2,394 eligible genes after filtering, see Methods). This pre-selection, performed at a permissive FDR threshold of 50%, identified 43, 20, and 57 genes with putative signatures of selection on net-ASE, *cis*-ASE, and CNA-ASE, respectively.

After merging the sets of genes obtained according to criteria above and after further filtering (see Methods), we performed separate ASE-dNdES tests for net-ASE (267 known drivers, Subset 1, and 43 pre-selected, Subset 2), CNA-ASE (196 known drivers and 57 pre-selected) and *cis*-ASE (271 known drivers and 20 pre-selected). At a FDR < 10%, we identified significant selection on ASE favoring the mutant allele in 11 net-ASE (Fig. 2B-C), 8 *cis*-ASE (Supp. Fig. S4 A-B) and 14 CNA-ASE (Supp. Fig. S4 C-D) genes. No inflation was observed in Q-Q plots, indicating the *p*-values are well calibrated (Supp. Fig. S5). The significant genes include well known cancer driver genes such as *TP53*, *KRAS* and *KEAP1*, suggesting that overall ASE on these genes is under selection. Among the net-ASE genes, *TP53* exhibited the strongest signal: 87% of missense variants showed positive net-ASE compared to 49% of synonymous variants (OR = 6.77, FDR = 6e-21, Fisher test, Fig. 2C); we note that *TP53* has common synonymous variants that disrupt splicing and so would result in ASE[33]. Along with *TP53*, the genes *STK11*, *BAP1*, *KEAP1*, *TGFBR2*, *SMAD4*, *KRAS*, *EIF1AX,* and *TSC1* also revealed positive selection, with 41-86% of missense variants exhibiting positive net-ASE, in contrast to only 10%-46% ES variants (range across the 9 genes, Fig. 2C). Conversely, *ACSF3 and GNA11*, the latter a recognized oncogene, displayed a signature of negative selection on the effect of positive net-ASE, where the fraction of missense mutations (8-10%) was decreased when compared to ES (35-44%).

Interestingly, tumor-suppressor genes were among the most common cancer gene categories with respect to oncogenes (Supp. Fig. S4E): net-ASE (63% TSGs vs 18% OGs), CNA-ASE (43% vs 7%) and *cis*-ASE (both at 25%). The other genes we tested, which are not known driver genes identified through our mean ASE differences pre-selection test (Subset 2), were present in net-ASE (9.1%, 1 gene: *ACSF3*), and were enriched in *cis*-ASE (50%, 4 genes: *CAD*, *C21orf91*, *CDC5L* and *TNKS2*) and CNA-ASE (29%, 4 genes: *CEP170B*, *CROCC*, *HLA-DQA1*, *SEMA4B*).

Upon examining the overlap among genes across the different ASE selected gene sets, we observed that *TP53* was the only gene to be significant in the CNA-ASE and the *cis*-ASE test, suggesting a convergence of both CNA and non-CNA mechanisms of ASE in *TP53* (Fig. 2D). However, it was more common to find genes exclusively in a single ASE category, indicating that

selective pressures act on ASE predominantly either by CNA or by non-CNA mechanism. Specifically, there were 4 unique genes in *cis*-ASE (one positively selected: *CDC5L*; 3 negatively selected: *TNKS2*, *CAD*, *C21orf91*) and 10 in CNA-ASE (4 positively selected: *EGFR*, *CIC*, *NFE2L2*, *NF1*, and 5 negatively selected: *CEP170B*, *CROCC*, *HLA-DQA1*, *FRAS1*, *SEMA4B*) (Fig. 2D). When relaxing the FDR threshold of our test from 10% to 50% the overlap of genes between CNA-ASE and *cis*-ASE increased from 4.7% (1 gene: *TP53*) to 9.4% (5 genes: *TP53*, *BAP1*, *BCOR*, *KEAP1*, *STK11*), suggesting the possibility of additional genes where both ASE mechanisms may co-occur and possibly synergize.

Within the 11 significant net-ASE selected genes, we aimed to dissect how CNA-ASE and *cis*-ASE contribute individually to the net imbalance favoring the mutant allele in the case of pathogenic variants, normalizing to a baseline of non-pathogenic (ES) variants. We used the mean ASE difference between missense and ES for CNA-ASE and *cis*-ASE separately, and compared them to the mean ASE difference in net-ASE, as a measure of the respective relative contributions of each ASE mechanism. In the case of *TP53*, *STK11* and *KRAS* (Fig. 2E), as well as all other selected genes (Supp. Fig. S6), the mutant allele is selectively favored by both CNA-ASE and *cis*-ASE. Remarkably, the majority of the signal came not from the expected CNA-based mechanisms, but instead from the non-CNA mechanisms that we refer to as *cis*-ASE. We found that *cis*-ASE contributed to 73.2% of the difference in net-ASE mean between missense and ES variants in the case of *TP53*. However, the proportion varied somewhat depending on the specific gene, with 50.2%, 61.7% and 71.87% of the difference attributed to *cis*-ASE in the case of *STK11, KRAS* and *TGFBR2* respectively. For the remaining 7 genes the range was in between 31.2% and 85.3%, with a mean of 53.31% *cis*-ASE contribution (Supp. Fig. S6).

Overall, allelic imbalances in gene expression favoring the mutant allele confer a selective advantage in various cancer driver genes. Importantly, we suggest that the role of *cis*-ASE, rather than CNA, in some genes is the dominant mechanism in generating the mRNA level allelic imbalances that are under positive selection.

Positive selection on ASE adversely impacts cancer prognosis

Given the signatures of positive selection associated with ASE, we hypothesized that ASE in driver genes could have an impact on cancer aggressiveness. To test this, we performed overall survival analyses in which we compared patients with tumors exhibiting selected, driver mutation ASE against patients with tumors with no detectable ASE. For this, we selected only samples bearing at least one missense mutation in any of our sets, encompassing 9 significantly positively selected net-ASE genes (we excluded *GNA11* and *ACSF3* that appeared to have negatively selected net-ASE). Tumor samples in which the average net-ASE of these missense mutations in known ASE driver genes was higher than 1 (indicating a 2-fold difference, expressed on a log2 scale) were classified as “affected” by ASE in driver genes, while the rest were labeled as “not affected”. In the Kaplan-Meier (KM) curves (Fig. 3A), the ASE affected group had a significantly poorer prognosis (p = 1.6e-6, log-rank test), where 36.0% of the affected patients have died during the follow-up period, while in comparison, only 25.6% of the non-affected patients did. Similar results were observed for *cis*-ASE (Supp. Fig. S7A, p = 1.8e-7) but not for CNA-ASE (Supp. Fig. S7B, p = 0.22), using their corresponding sets of positively selected ASE genes, 4 and 9 respectively. Considering the three ASE analyses, the median overall survival for the ASE affected group was 4.24 - 4.62 years, compared to 5.55 - 17.6 years for the ASE non-affected group; this analysis does not consider cancer type or other covariates.

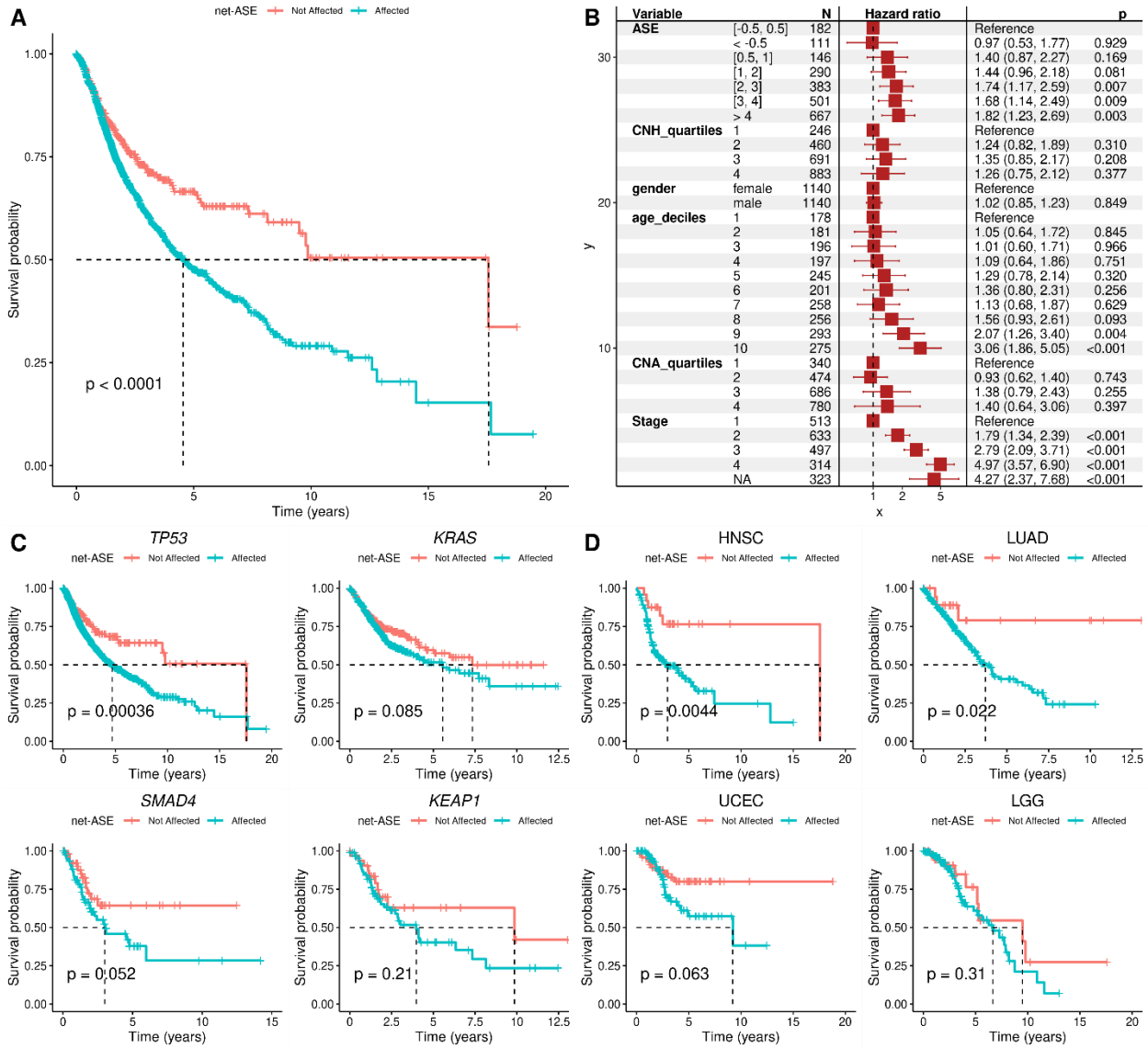

Figure 3. Positive selection on ASE has a detrimental effect on overall cancer prognosis

A, Kaplan-Meier (KM) curve of pan-cancer survival in patients exhibiting positive net-ASE in missense variants affecting positively selected net-ASE genes ("Affected"; blue), against those still bearing mutations in those genes, but not showing positive net-ASE ("Not affected"; red). **B,** Results from the Cox proportional hazards survival analysis, showing the estimated hazard ratios (HR) for each variable, along with the 95% confidence intervals (CI). For this, ASE variable was discretized into seven groups for more detailed analysis: negative ASE (< -0.5); neutral/reference ([-0.5, 0.5]); and positive ASE across five ranges ([0.5, 1], [1, 2], [2, 3], [3, 4], and >4) to evaluate the impact of varying ASE intensity on survival. Despite not being shown, other covariates were considered in the model, including cancer subtype and sample purity. **C**, KM curves of pan-cancer survival in patients exhibiting positive net-ASE ("Affected"; blue) in missense variants affecting *TP53* (top left), *KRAS* (top right), *SMAD4* (bottom left), and *KEAP1* (bottom right), compared to those with mutations in these genes but without positive net-ASE ("Not affected"; red). From the set of 11 positively selected genes, we show only the top 4 with highest HRs. **D**, KM curves as described in (A) for specific cancer types, considering HNSC (top left), LUAD (top right), UCEC (bottom left), and COAD patients (bottom right).

To address this, we conducted a Cox proportional hazards survival analysis (see Methods), incorporating cancer type and additional covariates associated with cancer survival (Fig. 3B). We categorized our ASE variable into seven groups for more detailed analysis: negative ASE (< -0.5); neutral/reference ([-0.5, 0.5]); and positive ASE across five ranges ([0.5, 1], [1, 2], [2, 3], [3, 4], and >4). Covariates used in the model included the patient age (discretized into deciles), sex, and tumor stage. Additionally, the model also accounted for other relevant variables predicting survival, in particular the cancer type, intratumor heterogeneity (estimated as copy-number heterogeneity, CNH)[34], tumor sample purity, total CNA burden, and 86 examples of recurrent large-scale CNA events as estimated in a previous study[35]. After controlling for these covariates, the analysis revealed that patients categorized in the positive net-ASE ranges faced a progressively higher risk of death, compared to net-ASE non-affected patients, with hazard ratios (HR) ranging from 1.40 to 1.82 (95% CI 0.87-2.27, 1.23-2.69; p = 0.17 - 3e-3; Fig. 3B). Notably, lower ASE categories (ASE < 2) were associated with lower HRs (1.40-1.44), while higher categories (ASE > 2) corresponded to higher HRs (1.74-1.82). Conversely, the negative net-ASE category showed no association (HR = 0.97). This effect of positive net-ASE is substantial in comparison to known survival covariates: the HR of CNH was of 1.26 for the highest quartile compared to the lowest (p = 0.38), and of 3.06 for the highest decile of age when compared to the lowest (p = 1.1e-5). Patients within the 8th and 9th age deciles, ranging from 68 to 77 years old, exhibit hazard ratios comparable to those seen with our net-ASE variable, ranging from 1.56 to 2.07 (p = 9.3e-2 to 4e-3). This implies that the prognostic impact of net-ASE in tumor transcriptomes on cancer patient survival is similar to the effect of aging 29-38 years (compared to our reference age group, which was 39 years old). Similar results were observed specifically for *cis*-ASE across the five positive ASE categories (HR = 1.36-1.70; p = 0.11 - 1e-3, Supp. Fig. S7C). Conversely, the analysis for CNA-ASE showed less significant results, with HRs ranging from 1.04 to 1.44 (p = 0.81 - 2.5e-2, Supp. Fig. S7D), suggesting that the non-CNA mechanisms driving ASE are the ones primarily responsible for cancer aggressiveness.

ASE in individual genes and cancer types linked with cancer aggressiveness

Additionally, we set out to assess if the association of ASE with survival held for individual driver genes (from our positively selected gene sets) and cancer types (when a high number of mutations/samples was available); to simplify the model, we excluded the 86 variables representing arm-level CNA changes. For example, considering only patients harboring at least one *TP53* missense variant, those affected by ASE in *TP53* had a poorer prognosis than those who were not affected by ASE: the Cox model for net-ASE reports HR = 1.52 with p = 2.9e-2 (KM curve in Fig. 3C). Similar trends were observed for *TP53 cis*-ASE (HR = 1.4 with p = 8.4e-3, Supp. Fig. S8A) and less significant impact for *TP53* CNA-ASE (HR = 1.28 with p = 1.8e-2; Supp. Fig. S8B). A consistent signal was seen with *KRAS*, where KM curves indicated a nonsignificant trend toward poorer survival in net-ASE, (HR = 1.22 with p = 0.33 for Cox model; 8.5e-2 for KM curve, Fig. 3C) or significant in the *cis*-ASE (p = 3.5e-3 for KM curve, Supp. Fig. S8A). We saw large effect size trends in the case of *SMAD4* (p = 5.2e-2 for KM curve; HR = 7.75, p = 0.13) and *KEAP1* (p = 0.21 for KM curve; HR = 3.91, p = 0.16) for net-ASE (Fig. 3C). Of note, the significance of *TP53* in our analysis is enhanced by the availability of a larger number of variants, whereas the rest of the genes with fewer variants have less statistical power.

Additionally, within the CNA-ASE set, some positively selected genes demonstrated varied significance levels (Supp. Fig. S8B). *EGFR*, for example, followed similar trends as observed above but did not reach significant levels in the Cox models (p = 5.5e-2 for KM curve; HR = 1.28). In contrast, the lung cancer genes *KEAP1* and *NFE2L2* displayed inverse associations (*KEAP1*:

HR = 0.12 with p = 6.3e-3; *NFE2L2*: HR = 0.4 with p = 4.7e-2), suggesting that, contrary to the overall trend, in some individual genes there might be a protective effect of CNA-ASE on survival.

Next, we asked in which cancer types the ASE-survival associations were more prominent. In a cancer type-specific analysis, we observed prominent results for similar cancer types in net-ASE (KMs in Fig. 3D and Cox in Supp. Fig. 9A) and *cis*-ASE (Supp. Fig. S8C and S9B): head and neck squamous cell carcinoma (HNSC; KM p = 4.4e-3 and 8.6e-05, for net and *cis* ASE respectively), and lung adenocarcinoma patients (LUAD; KM p = 0.02 and 8.5e-3). Other major TCGA cancer types such as UCEC and low grade glioma (LGG) showed a similar trend in the KM curves, with the former being marginally significant (UCEC: p = 6e-2 and 5.7e-2; LGG: p = 0.31 and 0.52). For CNA-ASE, although the p-values were not significant, similar cancer types showed trends associating with survival (Supp. Fig. S8D and S9C).

In addition to our core set of positively selected net-ASE genes, we hypothesized that ASE in additional driver genes might also be prognostically relevant. To test this, we generated two additional sets of genes where net-ASE is putatively positively selected, identified by the ASE-dNdES test described above, at thresholds (unadjusted) $p < 0.05$ and $p < 0.1$, resulting in 24 and 40 genes, respectively. As expected from our pre-selection process, of these, 23/24 (96%) and 36/40 (90%) are known cancer genes found in either CGC or Mutpanning. Here, we compared the survival between those tumor samples showing positive net-ASE on average against the rest of tumors (requiring they have at least one missense variant in any cognate cancer driver gene). As expected, the KM curve showed significant decreased survival within the affected group when all cognate cancer genes were considered (KM p = 5.56e-19; Supp. Fig. S10A). However, as we progressively excluded our sets of selected net-ASE genes (p = 5.2e-4; Supp. Fig. S10B) and putatively selected net-ASE genes (p = 6.6e-3; Supp. Fig. S10C), the signal progressively diminished, until the survival difference between ASE and no-ASE groups was no longer significant when excluding our permissive set of 40 putatively ASE-positively selected genes (p = 0.21, Supp. Fig. S10D). Thus the set of survival-relevant net-ASE genes is larger than the 9 individually significant genes we prioritized above, and might include up to 40 genes; the individual genes might be reliably identified, and their effect sizes measured, in larger transcriptomic datasets going forward.

Allelic differences in chromatin accessibility can explain cis-ASE

The mechanisms that give rise to the selected *cis*-ASE in cancer genes are of interest. We hypothesized that differences in the activity of DNA regulatory elements between alleles (which can be measured as local chromatin accessibility via e.g. ATAC-seq) could generate such imbalances in gene expression. To test this, we used matched WES and ATAC-seq data from 381 patients from the TCGA consortium[24], and performed variant calling on the ATAC-seq as was done with RNA-seq reads (see Methods), in order to estimate allele-specific chromatin accessibility (ASCA) for the mutant alleles. We asked whether allelic imbalances in the chromatin accessibility would be able to explain allelic imbalances observed at the mRNA level (schematic in Fig. 4A).

We selected matched genetic variants between our ASE and ASCA datasets, and studied the association between *cis*-ASE and ASCA values. At a stringent coverage threshold (sum of number of WT + MUT reads > 20) for both RNA-seq and ASCA, we were able to analyze 782 somatic variants that overlapped with ATAC-seq peaks of sufficient intensity, meaning they were located in or near to active regulatory elements. We found that ASCA significantly predicted *cis*-

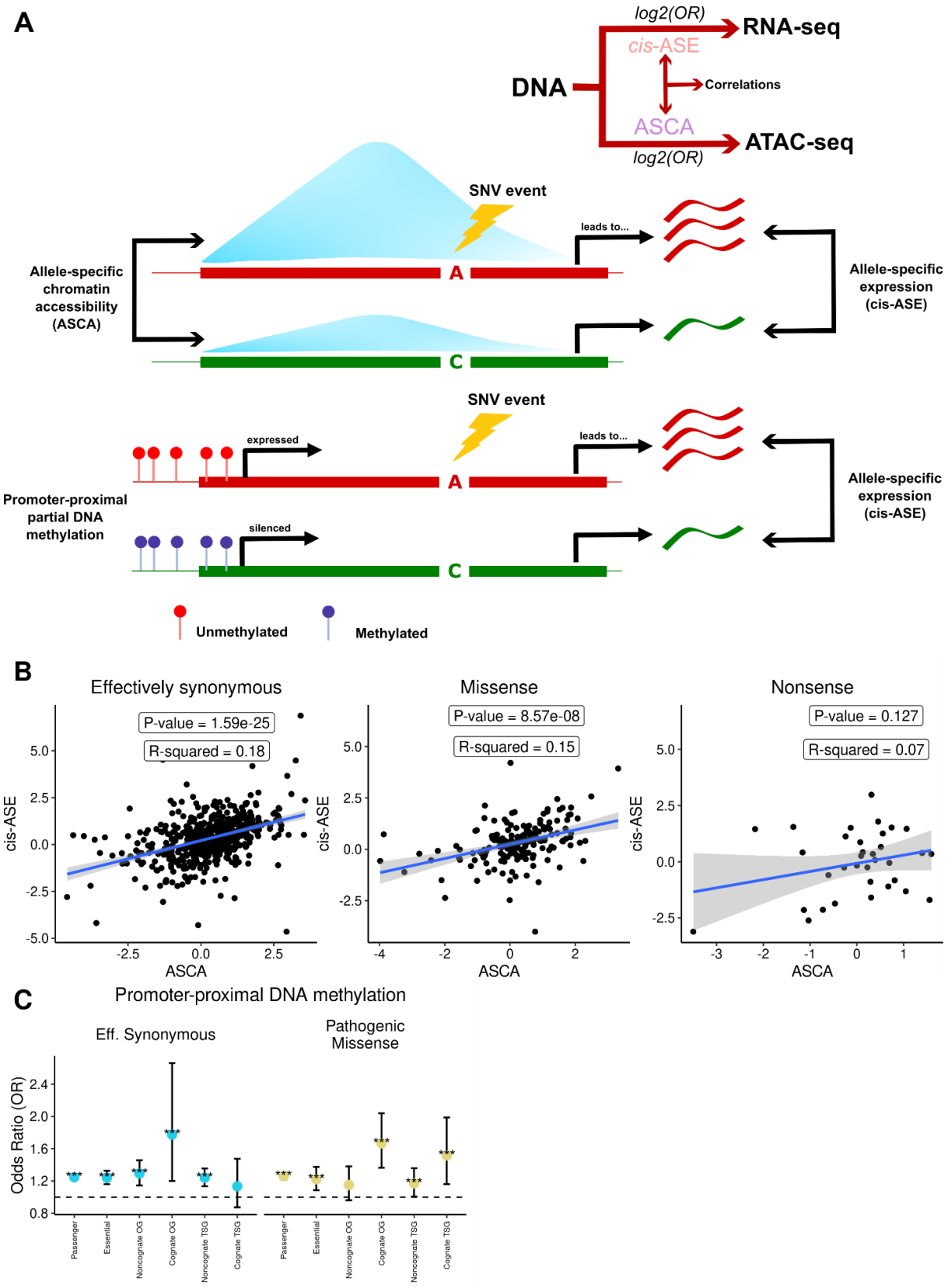
A
log2(OR)
RNA-seq
cis-ASE
DNA
Correlations
ASCA
ATAC-seq
log2(OR)
SNV event
leads to...
Allele-specific chromatin accessibility (ASCA)
A
C
Allele-specific expression (cis-ASE)
SNV event
expressed
leads to...
Promoter-proximal partial DNA methylation
A
silenced
C
Allele-specific expression (cis-ASE)
Unmethylated
Methylated
B
Effectively synonymous
P-value = 1.59e-25
R-squared = 0.18
Missense
P-value = 8.57e-08
R-squared = 0.15
Nonsense
P-value = 0.127
R-squared = 0.07
cis-ASE
ASCA
C
Promoter-proximal DNA methylation
Eff. Synonymous
Pathogenic Missense
Odds Ratio (OR)
Passenger
Essential
Noncognate OG
Cognate OG
Noncognate TSG
Cognate TSG

Figure 4. Allele-specific chromatin accessibility (ASCA) and promoter-proximal DNA methylation partially explains ASE caused by non-CNA mechanisms.
A, Schematic representation of our analysis (top), detailing how ASCA values are obtained as done with *cis*-ASE, and the subsequent correlation of these values with matched somatic variants. The diagram includes a graphical illustration showing how ASCA (middle) or DNA methylation (bottom) might influence the generation of ASE. **B**, Linear regressions of *cis*-ASE as a function of ASCA for matched ES (left), missense (middle) and nonsense mutations (right), with the 95% CI around the regression lines. **C**, Enrichment analysis comparing the occurrence of positive *cis*-ASE with partial methylation versus unmethylated status within promoter or promoter-proximal regions (specifically TSS200-1500 and the first 2kb of gene bodies) of genes. Analyses are stratified by gene categories including cognate and noncognate TSGs, OGs, passenger genes, and essential genes, showing ORs and their 95% CIs estimated by the Fisher tests.

ASE for missense variants (Pearson's $r = 0.38$; $p = 8.6e-8$; Fig. 4B), and similarly so for effectively synonymous variants ($r = 0.42$; $p = 1.6e-25$), suggesting that ASCA could account for approximately 15-18% of the *cis*-ASE variability.

Expectedly, this correlation did not extend to nonsense variants ($r = 0.13$; $p = 0.27$): due to the degradation of transcripts containing PTCs through NMD, transcript abundance is uncoupled from the effects of chromatin accessibility on transcription. We verified this pattern across different values of the technical parameter for coverage threshold (ranging from 5 to 25, Supp. Fig. S11). Excluding nonsense variants, we observed that 46.8% of variants exhibiting positive ASCA also showed positive *cis*-ASE, while only 3.7% showed negative *cis*-ASE (Supp. Fig. S12). Conversely, only 6.4% of variants with negative ASCA presented positive *cis*-ASE, while 27.7% exhibited negative *cis*-ASE.

We visualize three illustrative examples, showing, RNA-seq and ATAC-seq read alignments at a variant locus in a cancer gene, *NCOR1*, and loci in two passenger genes, *MID1IP1* and *RABIF* (Supp. Fig. S13). This shows an enrichment in the number of reads mapped to the mutated allele for both RNA-seq and ATAC-seq, in cases where both *cis*-ASE and ASCA exhibited log2(OR) values greater than 1, while CNA-ASE values were less than 0.5.

These findings highlight the association of allele-specific activity of regulatory elements with *cis*-ASE of somatic variants in cancer. One plausible cause of this is that the somatic mutation itself, or another genetic variant nearby, alters the activity of the regulatory element.

Altered methylation is associated with positive cis-ASE in oncogenes

To further evaluate the hypothesis that epigenetic influences during cancer evolution determine ASE, as suggested above, we analyzed DNA methylation data. In particular, we considered CpG sites near the TSS (see Methods for definition) jointly as a measure of gene activity levels, for each individual.

Given that positive *cis*-ASE reflects a preferential expression of the mutated allele, we hypothesized that the promoter regions of these genes might show partial methylation, associated with differential expression between alleles. To test this, we compared variants with positive *cis*-ASE to their methylation status, distinguishing between partially methylated (M-value > 30% and < 70% of gene's percentiles), consistent with one allele being less expressed, and unmethylated (M-value < 30% of gene's percentiles) promoter, consistent with both alleles being expressed.

We observed a stronger enrichment in cancer driver genes in cognate cancer types, meaning that there was a positive association between *cis*-ASE occuring in a tumor, and the partial DNA methylation of the promoter and promoter-proximal region. This was seen for pathogenic missense mutations in cognate OGs (OR = 1.67, CI = 1.36 - 2.04, p = 2.5e-7), and cognate TSGs (OR = 1.52, CI = 1.16 - 1.99, p = 1.8e-3), compared to the rest of gene categories (Fig. 4C): passengers, essentials and non-cognate TSGs and OGs (OR = 1.17 - 1.26, p = 1.3e-1 - 2.4e-40). Via a linear regression model, we estimated that for cognate OGs, each unit increase in DNA methylation was associated with an additional increase in *cis*-ASE of approximately 0.12 (in log2 scale) compared to passenger genes (p = 8.8e-11), suggesting an oncogene-specific effect. Notably, ES variants in cognate OGs also demonstrated a moderate association with partial methylation (OR = 1.77, CI = 1.20 - 2.66, p = 2.9e-3), possibly reflecting that some of the selected synonymous mutations in OGs[33] may act via altering allele transcriptional activity as reflected in promoter DNA methylation.

Given the observed enrichment in oncogenes and the associated partial methylation status, it's plausible that the WT allele is being selectively methylated (as a limitation of this test, DNA methylation is not phased with the mutation on which *cis*-ASE is measured). This epigenetic modification could suppress the expression of the WT allele, thereby boosting the impact of the mutated allele, conferring a GoF effect in these oncogenes.

Somatic mutations activate transcriptional chromatin states in cognate oncogenes

To further support the above-mentioned observations that epigenetic changes correlate with *cis*-ASE, we tested if the presence of a variant itself associates with the local epigenetic environment towards a more active state within the gene body. This change would be expected to occur in parallel with occurrence of open chromatin conditions in regulatory DNA, as measured by our ASCA test, and the partial methylation in or adjacent to the promoter region, as analyzed above. To test this hypothesis, we utilized the *Sei* tool[36], a neural network that predicts the chromatin states from local DNA sequence, to assess the impact of somatic SNVs in coding regions on chromatin states. Specifically, we focused on variants across four chromatin "sequence classes" associated with active transcription (TN1-TN4) in gene bodies.

We classified somatic mutations based on whether they activated ("TN activating") or deactivated ("TN inactivating" variant type) the *Sei* score for the transcriptional states TN1-TN4, and whether this change of transcriptional states associated with ASE. The distribution of *cis*-ASE values among these TN groups was broken down by SNV functional impact (ES and missense) and gene classification (including cognate and noncognate OGs and TSGs, and as controls passenger genes).

We observed a significant increase in positive *cis*-ASE values within the TN activating category among cognate OGs mutations (Fig. 5A), both for missense (mean *cis*-ASE TN activating = 0.61 vs mean *cis*-ASE TN inactivating = 0.48, p = 2.1e-13, t-test) and ES mutations (0.61 vs 0.43, p = 1.2e-6), suggesting a direct link between mutations that promote active transcription-like chromatin, and enhanced allele expression in that gene. This effect was not evident for missense mutations in noncognate OGs (0.5 vs 0.48, p = 0.22), suggesting the ASE-inducing epigenetic effect is selected in OGs. This may provide an explanation for a part of the selection on synonymous mutations seen in OGs[33] and the epigenetic effect may extend to some missense mutations as well. The association of TN activating mutations with ASE was modestly present in

cognate TSGs (0.88 vs 0.77, p = 5.7e-3) and also noncognate TSGs (0.55 vs 0.59, p = 1.5e-2), as well as to a lesser extent in passenger and essential genes (missense mutations in passenger genes: 0.46 vs 0.44, p = 3.7e-7; essential genes: 0.52 vs 0.55, p = 8.1e-5). This pattern supports the expectation that mutations which activate transcription are likely to correlate with positive *cis*-ASE, but it may be selected in OGs as the enhanced expression is more pronounced than other categories of genes.

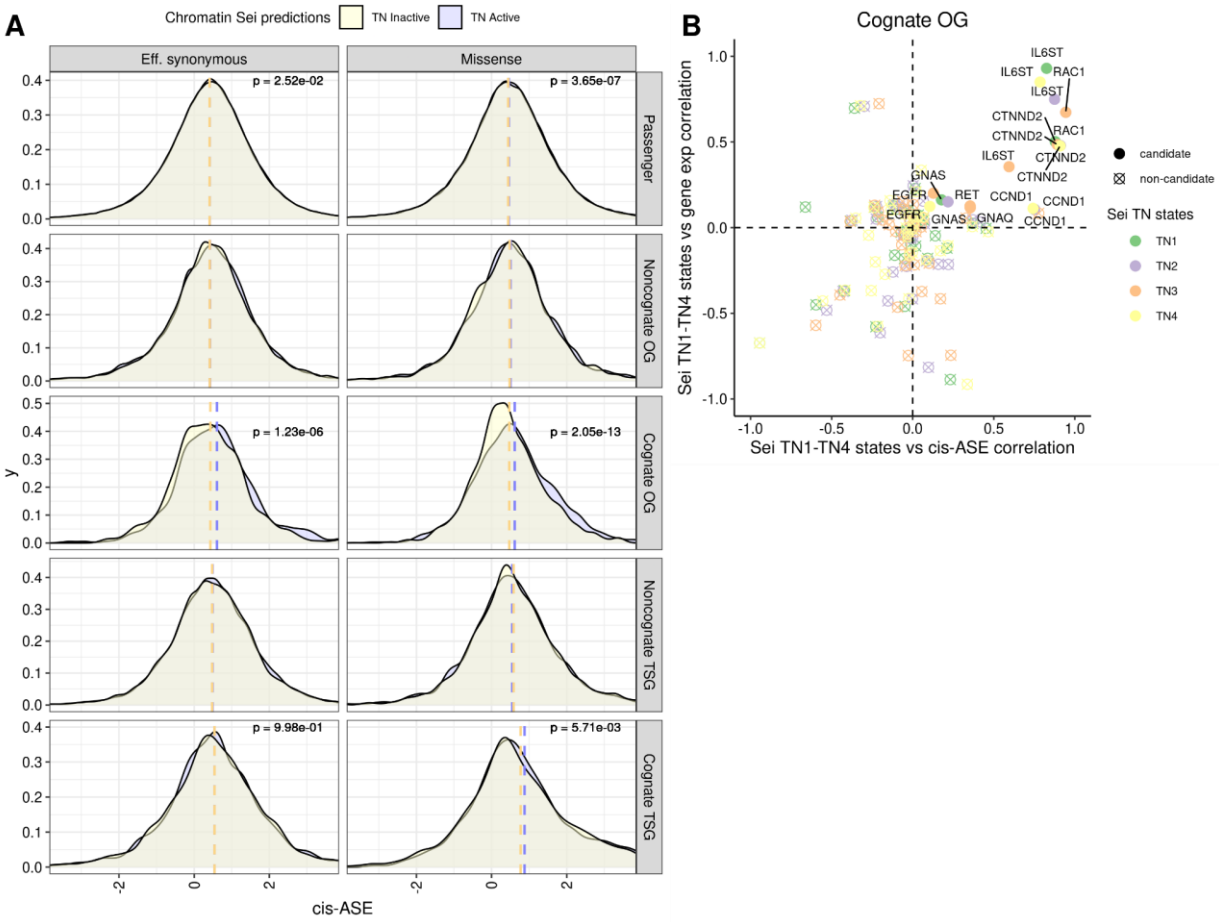

Figure 5. Activation of transcriptional chromatin states in cognate oncogenes.
A, Distribution of *cis*-ASE for somatic mutations (ES in left panel, missense in right panel) in different gene categories, including passenger genes, noncognate and cognate OGs and TSGs. Mutations are classified into TN Active (based on *Sei* scores > 0.0001) indicating active transcription chromatin states, and TN Inactive (*Sei* scores < -0.0001). Vertical dashed lines represent mean *cis*-ASE values for each *Sei* group. Differences in means between groups were tested using t-tests, with corresponding p-values displayed. **B,** Analysis of the correlation between *cis*-ASE and *Sei* TN1-4 states (X-axis), alongside the correlation between gene expression and Sei TN1-4 states (Y-axis), for each cognate oncogene and *Sei* TN state separately. This panel identifies correlations where both *cis*-ASE and gene expression are positively associated with *Sei* scores, suggesting transcriptional activation. OGs in the upper-right quadrant are highlighted (if both correlations > 0.1) as potential candidates experiencing selective pressure through gain-of-function effects.

Further, we investigated whether these chromatin state changes, evident in positive *cis*-ASE, also increased overall gene expression. In cognate OGs, we correlated *Sei* TN1-4 states scores with overall gene expression and with *cis*-ASE, where we expect bona fide activating mutations to show up as hits in both tests. From 39 cognate OGs, 8 genes showed R > 0.1 between Sei TN1-4 scores and both gene expression and *cis*-ASE (Fig. 5B).

Overall, this analysis supports that the changes in epigenetic states associated with ASE can derive from somatic variants that are upstream of the epigenetic change. Interestingly, some of the variants in the gene coding region may act primarily through the epigenetic change, modulating gene regulation rather than or in addition to changing protein structure.

Splicing-altering variants associate with a decreased expression of the mutated alleles

Somatic mutations altering splicing can have profound impacts on gene expression, for instance by resulting in frameshifts that trigger NMD[37,38]. We hypothesized that splicing changes might be an important mechanism underlying some of the variability observed in *cis*-ASE across different selected mutations. To test this, we used the *SpliceAI* tool[39] to estimate the probability of somatic variants affecting splicing at their location or nearby. *SpliceAI* provides delta score predictions (i.e. change in splice site activity) across four categories: splicing Donor Loss (DL), Donor Gain (DG), Acceptor Loss (AL), and Acceptor Gain (AG), measuring the differential impact on splicing between mutated and wild-type alleles. We also considered the maximum value of the four delta score *SpliceAI* types, termed "MaxDelta".

We categorized somatic variants with *SpliceAI* delta scores of ≥ 0.15 as having a high-impact on splicing while the rest were considered as neutral (5.3% of total variants; of note, here we consider the MaxDelta score). Analyzing the proportion across different categories of *cis*-ASE, we found that 4.4% of variants associated with positive *cis*-ASE had a high impact on splicing, similar to 5.0% for variants showing no *cis*-ASE. However 9.1% of variants exhibiting negative *cis*-ASE were predicted to be splicing altering.

In a pan-cancer analysis (Fig. 6A), a significant *cis*-ASE difference between variants with high versus low *SpliceAI* scores was found ($p < 1e-16$, Mann-Whitney *U* test) in each of the five splicing categories (AG, AL, DG, DL, MaxDelta), where higher scores were associated with negative *cis*-ASE values. This indicates that mutations that alter splicing, considered as a set, tend to reduce the expression of the mutated allele, however this may differ for particular genes or variants.

As a prominent example, we examined three known splice-disrupting exonic driver variants in the *TP53* gene[33,40]. Notably, the 3′ terminal nucleotide of *TP53* exon 4 (codon 125), a recurrent target for synonymous mutations, had a mean DG delta score of 0.15 and a DL of 0.7 across the 3 possible somatic synonymous variants at this codon. These occur in 32 total tumor samples in our dataset, from which 10 have a negative *cis*-ASE (< -1), compared to only 1 with negative CNA-ASE. Similarly, codon 224 of *TP53* exon 6, showed a DG of 0.46 and DL of 0.9 in one synonymous variant (n = 3, from which 2 have negative *cis*-ASE). Lastly, the codon 331 of *TP53* exon 9, exhibited a mean DL of 0.62 among two non-pathogenic (at the amino acid level) missense variants (n = 5, from which 2 have negative *cis*-ASE), providing an example where a nominally missense variant likely exerts its effect through splicing rather than through amino acid substitution.

Further analysis involved stratifying variants by gene categories and functional impact (missense or effectively synonymous). We tested for association between *cis*-ASE positive mutations and variants predicted to significantly impact splicing versus those with a neutral impact (Fig. 6B, see Methods). This analysis was repeated also on negative *cis*-ASE, where we expect more numerous associations based on the above, and additionally we also evaluated the differences in *cis*-ASE values between high-impact and neutral variants for the five *SpliceAI* delta score categories, as in the pan-cancer analysis (Fig. 6C for a subset and Supp. Fig. 14B for the complete analysis).

Tumor suppressor genes are often silenced by splice-altering missense and nonsense variants

A notable global enrichment in negative *cis*-ASE was observed across all five splicing delta scores (Fig. 6B). That the ASE direction would be negative is anticipated because many splicing motif disturbances result in out-of-frame splicing products, which contain premature stops and trigger NMD[37].

The most pronounced effects were seen for cognate TSGs, particularly in the ES variants for the *SpliceAI* MaxDelta (OR = 3.17, p = 9.4e-5) and across the other four splicing scores (mean OR = 4.23, range p = 5.6e-05 to 0.11). The differential *cis*-ASE values between the high-impact and neutral splicing impact variants showed similar trends for DG (median *cis*-ASE values of 0.06 vs 0.53, respectively, p = 6.45e-3), DL (-0.8 vs 0.53, p = 1.8e-5) (Fig. 6C). However, the two acceptor site delta scores, AL (0.17 vs 0.52, p = 0.16) and AG (0.72 vs 0.52, p = 0.62), have shown non-significant or contrary trends, suggesting that the effect of splicing-altering exonic variants on ASE occurs mainly through alteration of donor sites.

That synonymous variants could disrupt splicing patterns in cancer genes was reported previously[33,37,41]. Here, we observed an effect in the ES variants, and we next asked if this effect was attributable to synonymous variants only, or also to the missense mutations with low-impact at the protein level (non-pathogenic). We found a similar magnitude splicing effect on ASE resulting from the non-pathogenic missense mutations in the MaxDelta score (OR = 3.44, p = 4.8e-4; Supp. Fig. S14A) as for the synonymous group for MaxDelta (OR = 2.83, p = 3.9e-2). In addition, within cognate TSGs, 12.9% of non-pathogenic missense mutations with negative *cis*-ASE exhibited high MaxDelta scores (>0.15), in contrast to just 6.9% observed in passenger genes or 5.4% in noncognate TSGs. This 6.0–7.5% difference underscores a selective pressure favoring splice-altering missense mutations in TSGs. This indicates that these non-pathogenic missense mutations in TSGs are selected for their impact on splicing, rather than alterations to the amino acid sequence. For synonymous variants, the observed difference was 2.9-4%. Therefore, many missense mutations and also synonymous mutations can drive cancer through splicing effects.

In addition to the ES variants group, the pathogenic missense variants also demonstrated significant albeit lower enrichment in negative *cis*-ASE (MaxDelta score OR = 2.72; p = 2.1e-2), and the difference in *cis*-ASE between high-impact and neutral variants was not statistically significant (DG: median *cis*-ASE 0.62 vs 0.62; DL: 0.54 vs 0.62). This might be due to the predominance of positive *cis*-ASE values in pathogenic missense mutations within TSGs, which may mask the differences. By excluding mutations with positive *cis*-ASE (> 1), a more distinct trend was observed, showing more pronounced negative *cis*-ASE values for missense mutations with higher *SpliceAI* delta scores (MaxDelta: -0.24 vs 0.13, p = 2e-3). Therefore, even some

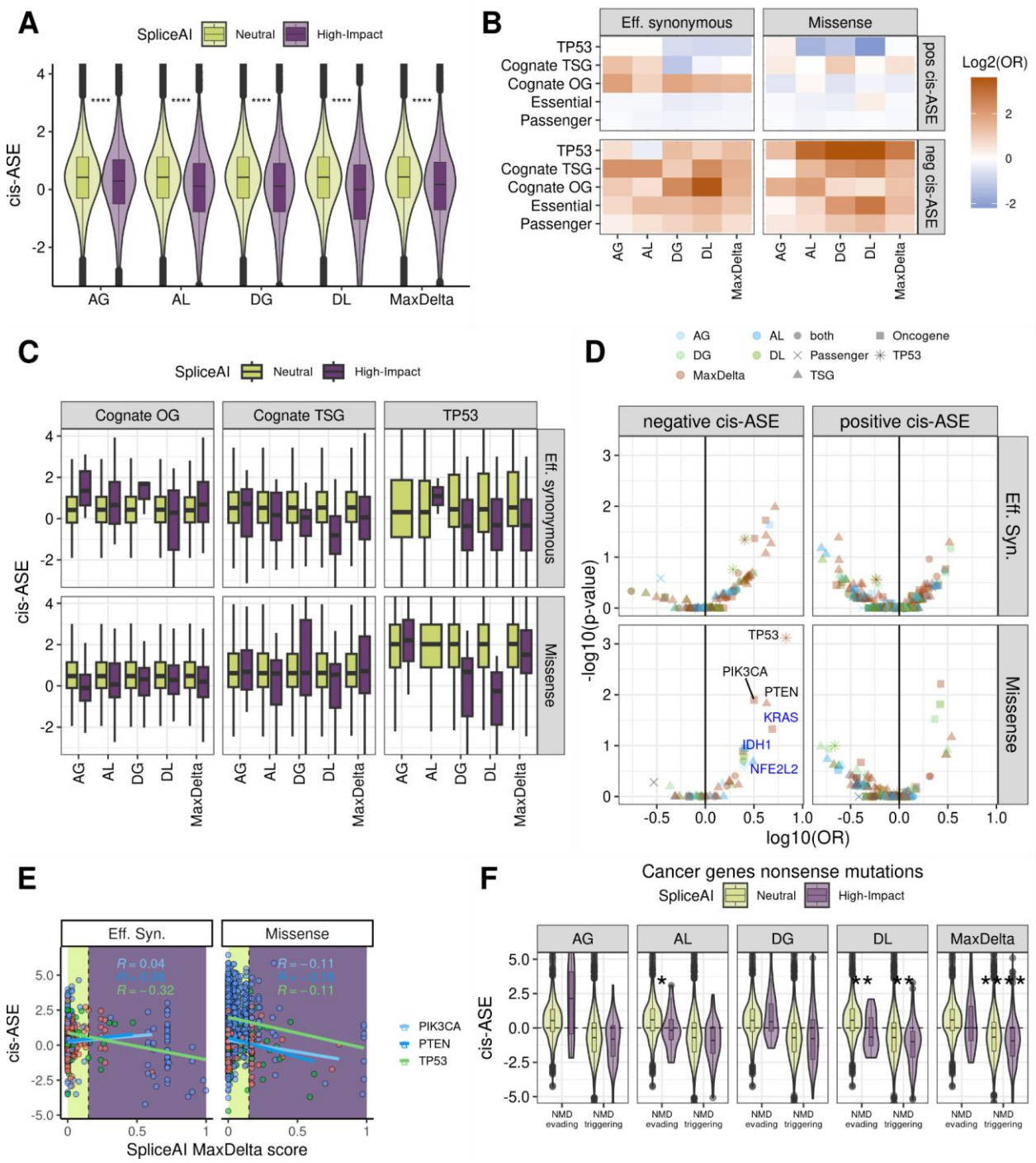

Figure 6. Association of splicing-altering variants with decreased expression of mutated alleles
A, Pan-cancer *cis*-ASE values categorized by four *SpliceAI* delta scores: Acceptor Loss (AL), Acceptor Gain (AG), Donor Loss (DL), and Donor Gain (DG). The MaxDelta category represents the maximum value among the four categories for each mutation. Mutations are further stratified into two groups based on their potential impact on splicing: "High-Impact" splicing for mutations with *SpliceAI* delta scores ≥ 0.15, and "Neutral" for scores < 0.15. **B**, Enrichment analysis displaying significant associations between the presence of somatic variants exhibiting either positive or negative *cis*-ASE and their predicted impact on splicing ("High-Impact" versus "Neutral"). The analysis is stratified by variant functional impact (ES and

missense, top panel) and gene classification (including cognate OGs and TSGs, passengers, essentials, and *TP53*). For clarity, only a subset of the gene categories is shown. This extends across all five *SpliceAI* delta scores. ORs are calculated using the Fisher Exact test and presented in a log2 scale, color-coded from low (blue) to high (red) values. Positive enrichment indicates a higher proportion of positive or negative *cis*-ASE mutations that are predicted to have a significant impact on splicing. **C**, Shows the distribution of *cis*-ASE values, stratified by SNV type and gene classification for TSGs, OGs, and *TP53*, across the five *SpliceAI* delta scores. It differentiates the "High-Impact" and "Neutral" mutation groups. **D**, Similar to B, but focuses on individual cancer genes, plotting raw *p*-values on the Y-axis against OR on the X-axis, both in log10 scale, for genes significant (FDR < 25%, black) or candidates (FDR < 50%, blue) in specific *SpliceAI* delta scores, if there were sufficient number of mutations. Essential genes are excluded from this analysis. **E**, Correlations between *SpliceAI* MaxDelta scores and *cis*-ASE specifically for *TP53, PTEN* and *PIK3CA* mutations, stratified by ES (left) and missense (right). Correlations are displayed as Pearson's *r.* "High-Impact" mutations are background-highlighted in magenta and "Neutral" in yellow. **F**, Similar as B, displays *cis*-ASE values for nonsense mutations, further split between NMD-triggering and NMD-evading, divided by *SpliceAI* delta scores and splicing impact group.

missense variants with (predicted) high pathogenicity on the amino acid sequence do alter splicing of TSGs in cancer.

Considering specifically the case of *TP53*, the enrichment of negative *cis*-ASE was seen in pathogenic missense variants (MaxDelta OR = 6.75, *p* = 7.6e-4; Fig 7B) and in ES (MaxDelta OR = 2.53, *p* = 4.5e-2) (Fig. 6C). This is consistent with prior work on exonic synonymous variants that disrupt splicing of *TP53* and more generally cause intron retention or exon skipping in TSG[33,37,41].

We further asked which other individual TSGs, in addition to *TP53*, exhibit similar links between splicing-altering variants and ASE, by performing a gene-level analysis (Fig. 6D). We examined the enrichment of negative *cis*-ASE with high-impact splicing scores across the five splicing delta score categories, including the global MaxDelta score. For ES variants, no individual gene reached statistical significance after FDR adjustment (< 25%), however five TSGs (*ATM*, *NCOR1*, *TP53*, *SPEN*, and *POLG*) had nominal *p* < 0.1 with ORs ranging from 2.53 to 5.21, and 95% CIs from 0.86 to 21.08 (Fig. 6D). For the pathogenic missense mutations, two TSGs were significant within the MaxDelta splicing category association with negative *cis*-ASE at FDR < 15%: *TP53* (OR = 6.75, CI = 2.08 - 20.71, FDR = 1.98%), and the *PTEN* (OR = 4.27, CI = 1.15 - 14.72, FDR = 19%). The negative correlation between *cis*-ASE values and *SpliceAI* MaxDelta scores for both genes (Fig. 6E), more clearly observed in missense (r = -0.18 to -0.11, p = 1.48e-6 - 3.7e-3) than ES (*r* = -0.32 to 0.08, *p* = 1.7e-4 - 0.56), further supported splicing changes as a cause of ASE. The negative *cis*-ASE would likely result from degradation of mis-spliced transcripts by NMD or related processes, supporting that coding variants in some TSGs can silence the genes at mRNA level.

Next, we tested the converse case of positive *cis*-ASE association and three TSGs emerged as suggestive results at nominal *p* < 0.1: *BUB1B* (MaxDelta: OR = 3.29, CI = 0.94 - 13.31, p = 5.18e-2) and *KMT2D* (DG: OR = 3.40, CI = 0.72 - 17.47, *p* = 6.8e-2) with ES mutations, and *BCOR* (MaxDelta OR = 3.45, CI = 0.82 - 17.7, *p* = 6.8e-3) with missense mutations. The positive *cis*-ASE on these TSG might act via bolstering the expression of dominant-negative acting variants or isoforms. Overall, in addition to *TP53*, other relevant tumor suppressor genes exhibit ASE due to splicing-disrupting mutations in their coding regions.

We further hypothesized the reduction in expression of mutated alleles is also observed in some nonsense variants due to splicing effects. This would be in addition to degradation by the NMD pathway due to the nonsense variant; therefore we tested splicing changes elicited specifically by the subset of nonsense mutations that evade NMD (categorized based on established genomic rules[31], see Methods). As expected, NMD-evading variants did not exhibit the strong negative *cis*-ASE that is typical of NMD-triggering nonsense variants (median *cis*-ASE of 0.55 vs -0.71, respectively; Fig. 6F). When further dividing these NMD-evading variants into high-impact and neutral *SpliceAI* delta score groups, the former exhibited a more pronounced negative *cis*-ASE, compared to the latter, in AL (-0.17 vs 0.59, p = 1.2-e2, Mann-Whitney U test) and DL delta scores (-0.67 vs 0.56, p = 6.7e-3). Interestingly, this effect was observed, albeit weaker, also in predicted NMD-triggering nonsense variants (MaxDelta: -0.93 vs -0.67, p = 6.3e-5; DL: -0.67 vs -1, p = 7.8e-2) suggesting some of these variants do partially escape NMD although not recognized by existing NMD rules. Thus, even if nonsense variants do escape NMD, their involvement in significant splicing alterations can still lead to decreased allele expression in TSG.

Splicing effects of exonic coding mutations are common in oncogenes and cause ASE

Next, we asked if the association to *SpliceAI* scores was coupled to ASE also in OGs. Interestingly, the effect on cognate OGs for ES variants was contrary to the trend in TSG: we identified an enrichment in positive *cis*-ASE across all four splicing alteration types (mean OR = 3.09, p = 2.6e-2 to 0.16; Fig. 6B) and consistently with the overall MaxDelta score (OR = 2.76, p = 9.2e-4). Specifically, ES variants with higher *SpliceAI* delta scores led to an increase in positive *cis*-ASE (Fig. 6C), particularly notable for AG (median of *cis*-ASE values 1.35 vs 0.42, p = 4.2e-3) and MaxDelta (0.68 vs 0.41, p = 2.3e-2), with consistent albeit non-significant trends in AL and DG categories (0.66 vs 0.43 p = 0.3; and 1.69 vs 0.43 p = 0.16, respectively). This pattern indicates that splicing changes may enhance the expression of the mutated allele in cognate OGs, in agreement with positive selection on synonymous mutations in OGs, proposed to act via splicing[33]. In contrast to the effect of ES mutations, no enrichment for pathogenic missense mutations in ASE was found (MaxDelta OR = 0.77, p = 0.22).

As we did for the TSGs above, we split the ES group into non-pathogenic missense and synonymous-only mutations. This revealed that the impact on positive *cis*-ASE was more pronounced for non-pathogenic missense mutations rather than for synonymous-only, showing a notable OR of 3.06 (p = 6.2e-4) in the MaxDelta splice type, compared to the similar trend was observed in the synonymous-only group (MaxDelta: OR = 1.70, p = 0.48; the limited number of variants prevented robust estimation of ORs). The non-pathogenic missense mutations had ASE OR of 5.03 (p = 2e-02) in both the AG and DL splicing types (Supp. Fig. S14A).

Next, in a gene-level analysis, we did not find oncogenes with a significant enrichment (at FDR < 15%) of positive *cis*-ASE for ES variants with high splicing scores (Fig. 6D). However, of the 24 total ES mutations with positive *cis*-ASE and high MaxDelta *SpliceAI* scores, 21 were non-pathogenic missense mutations from prominent oncogenes, including *FGFR2*, *ERBB3*, *ERBB2*, *PIK3CA*, *CTNNB1*, *FGFR3*, *BRAF*, and *EGFR*. Considered jointly, within cognate OGs, 9.8% of non-pathogenic missense mutations with positive *cis*-ASE exhibited high-impact splicing MaxDelta scores, in contrast to 3.9% observed in passenger genes or noncognate OGs. This 5.9% excess underscores a selective pressure in cognate OGs favoring splice-altering missense mutations, indicating that some missense mutations are selected for their impact on splicing rather than the alterations to the amino acid sequence. For synonymous variants, the observed

difference was 2.5%, supporting that they too are under selection for splicing effects that manifest as ASE.

In particular, among pathogenic missense variants, *IDH1* emerged as a candidate with a promising albeit non-significant association between splicing and positive *cis*-ASE (MaxDelta: OR = 2.67, CI = 1.27 - 5.42, raw p = 6.1e-3). Lastly, considering the ES variants the closest two oncogenes showing enrichment of negative *cis*-ASE, with nominal $p < 0.1$, were *PLCG1* and *TRRAP*, exhibiting ORs from 3.2 to 4.56 and CIs from 0.85 to 18.63. For the pathogenic missense mutations, the oncogene *PIK3CA* was significant (Fig. 6D) within the MaxDelta splicing category at FDR < 15% (OR = 3.17, CI = 1.15 - 7.92, FDR = 14%), and *KRAS* was on the borderline of significance (OR = 4.88, CI = 0.77 - 23.07, FDR = 33%).
Our observation of positive *cis*-ASE linked with ES mutations in OGs solidifies the notion that some changes without strong effects on protein sequence (synonymous and non-pathogenic missense) may be causal to carcinogenesis and that in OGs they have GoF effects. The ASE analysis also rules out confounding by trans-effects on splicing, suggesting that the effectively synonymous mutation itself is causal.

In summary, coding regions variants with high splicing impact scores can lead to an allele-specific reduction or increase in mRNA levels, by disrupting the splicing mechanism. This differential impact of splicing alterations can be exploited by cancer genes: in TSGs we observed a decreased expression of the mutated allele via *cis*-ASE, resulting in a loss-of-function. Conversely, OGs would leverage these splicing alterations to exert a gain-of-function effect, characterized by an increase in the expression of mutated allele.

Discussion

In this study, we performed an in-depth analysis of the genomic signatures, prevalence, mechanisms, and prognostic consequences of allele-specific expression on somatic mutations in cancer.

We estimated that the non-CNA causes of ASE, herein called *cis*-ASE, explained a very notable portion (55.6%) of the total pan-cancer ASE variability, which is several-fold higher than previous estimates[8,18]. In our case, we stringently accounted for imbalances at the DNA level and sample purity, and the analysis comprehensively included exome-wide somatic SNVs across >8000 tumors, while previous analyses focused on smaller number of genomes (~1000) from few cancer types[8], or on a single cancer type[9,15–17,42], or on individual genes[18]. This suggests that *cis*-acting genetic variation plays a critical role in the generation of allele expression imbalances in cancer genomes. Additionally, we observed the highest prevalence of *cis*-ASE-affected variants in lung tumors (both LUAD and LUSC) and also the related head-and-neck (HNSC), while ovarian cancer was the most affected by imbalances generated through CNA, in accord with a very high prevalence of somatic CNA events in ovarian carcinomas[43,44].

We demonstrated that this ASE is subject to selection in cancer genomes, with mRNA allele imbalances in favor of the mutant allele in some driver genes confer a selective advantage for cancer cells. This was observed mostly in the case of some TSGs, and most strikingly in *TP53*. We suggest that *cis*-ASE plays a key role in mimicking LOH events, therefore acting as a second-hit that increases the fitness effect of driver mutations. We identified specific genes displaying significant differences in 11 net-ASE, 8 *cis*-ASE, and 14 CNA-ASE genes, between the pathogenic missense (according to the VARITY score) and the effectively synonymous variants. Therefore, this conditional ASE-dNdES test that considers both DNA and RNA-level genetic variation can identify cancer genes; it further implicated additional genes such as *ACSF3*, *CAD*, *C21orf91*, *CDC5L* and *TNKS2* in cancer evolution via the significant allelic imbalance of the pathogenic mutations in mRNA. Within our set of positively selected genes, these differences in overall net-ASE resulted from an additive effect between CNA-ASE and *cis*-ASE, where in 8 out of 11 genes, *cis*-ASE contributed to more than 50% of the difference, reinforcing the non-CNA sources of ASE as pivotal in generating allelic imbalances that are selected for in cancer.

The imbalances favoring the mutant allele in missense somatic driver gene mutations predicts overall survival of cancer patients. This pertains exclusively to those driver genes where pathogenic missense variants are transcriptionally favored over the wild-type allele, compared to the non-pathogenic missense and synonymous. We anticipate that there exist more driver genes where ASE is relevant to cancer aggressiveness, with a saturation analysis of effects on survival suggesting about 30 candidate genes. Which individual genes and cancer types are more affected will be clarified with additional statistical power from larger sample sizes. These are warranted, since, the overall pan-cancer effect on survival was substantial for net-ASE and *cis*-ASE (but not for CNA-ASE), compared to other variables considered such as intratumor heterogeneity. This highlights the potential clinical implications of ASE as a prognosis biomarker of aggressive disease.

Next, we addressed the questions of the mechanisms underlying the observed *cis*-ASE, in the absence of (or, in addition to) CNA effects. While our data does not rule out the previously-reported effects of imprinting and of germline eQTL variants associating with ASE, these cases

are likely not widespread across many genes/individuals[8,14]. Instead, we suggest that epigenetic changes in the tumor generate many *cis*-ASE effects including ones we identify to be under selection because they associate with dN/dES ratios.

These epigenetic changes are observable in allelic differences in chromatin accessibility (as ATAC-seq), and in the DNA methylation near the promoter that associated with *cis*-ASE; both readouts reflect the altered activity of enhancers or promoters affecting transcription output. We estimate that ASCA explains around 15-20% of the *cis*-ASE variability (here, estimating only via ATAC-seq, for which allele-specific data was available). Our model suggests that differences in accessibility lead to the generation of allelic imbalances in expression, providing a basis for selective pressures to then act on somatic mutations present in one and not the other allele. Partial DNA methylation, associated with *cis*-ASE mostly in cognate oncogenes, suggests that the wild-type allele may be selectively methylated (as a limitation of this test, DNA methylation is not phased with the mutation on which *cis*-ASE is measured, thus this is an indirect observation). This could confer a GoF effect in OGs, analogous to the genetic alterations in OGs, in particular CNA deletions that preferentially remove the wild-type allele[29,45]. Similarly, the epigenetic effect on TSG ASE that we observed might bear an analogy to the putatively selected CNA gains in TSG that increase the dosage of the mutant allele[45].

Of note, our models do not establish the temporal order between the mutation and the epigenetic change, nor, importantly, do they require that the two are directly causally related. Instead, the mutation and epimutation could in principle occur independently, and their co-occurence in a cancer driver gene would be under selection. Epigenetic changes may occur due to trans-acting factors; indeed global epigenome changes are not uncommon in tumors[46–48] and could provide the second-hit event for various somatic mutations, generating drivers via mRNA expression imbalance. However, it is also well possible that the somatic mutation causes the epigenetic change. Indeed, by employing a machine learning method we predicted that the somatic mutation itself can shift the local epigenetic environment towards a more active transcribed chromatin state within the gene. This implies an increased expression of the mutated allele conferring a GoF effect, and consistently was stronger for cognate OGs. Thus, mutations with ASE-inducing epigenetic effect can be selected in OGs, possibly providing an explanation for a part of the selection on synonymous mutations seen in OGs[33] and the epigenetic effect may extend to some missense mutations as well.

In addition to epigenetic alterations, we identified another mechanism associated with somatic mutations that cause driver *cis*-ASE: splicing alterations. In cognate tumor suppressor genes particularly *TP53* and *PTEN*, high-impact splicing mutations generally reduce the expression of the mutated allele, noticeable both in the effectively synonymous and pathogenic missense mutations. More pronounced effects were observed in variants affecting donor sites, compared to acceptor sites; this might relate to donor sites being less conserved generally, as observed in that they are more frequently mutated than acceptor sites, in a 1.5:1 ratio[49]. Curiously, nonsense variants, even if they escape NMD, can cause splicing alterations thus diminishing allele expression.

Conversely to TSG, a different pattern emerged in OGs: effectively synonymous (including non-pathogenic missense and synonymous) mutations that affect splicing notably enhanced the expression of the mutated allele. This supports the hypothesis that positive selection on synonymous mutations in OGs operates through splicing mechanisms[33]. The association of positive *cis*-ASE with ES mutations in oncogenes reinforces the idea that mutations that do not

overtly affect protein sequences, can still play a causal role in cancer development, often manifesting as gain-of-function effects in OGs. The ASE analysis also rules out confounding by trans-effects on splicing, suggesting that the effectively synonymous mutation itself is causal to splicing changes.

Therefore, somatic mutations in coding regions can drive cancer via significant impacts on splicing, registering as allele-specific changes in mRNA levels with opposite directions for TSGs and OGs.

Overall, we find that *cis*-acting genetic variation in driver genes often associates with allele-specific expression. Positive selection acts on the presence of allelic imbalances that favor the abundance of pathogenic variants in mRNAs contributing to tumorigenesis and predicting detrimental prognostic outcomes in affected patients. We encourage for ASE to be regarded not only as a prognostic marker, but also a prospective therapeutic target in the context of personalized medicine, prioritizing cancer genes for pharmacological intervention. Overall, our study suggests that allele imbalances of mutations at the mRNA level should be explicitly considered in modeling tumor evolution, as well as highlighting the importance of interactions between the genome and epigenome somatic evolution.

Acknowledgements

We are grateful to Marcell Veiner for running the neural networks *Sei* and *spliceAI* on our somatic ASE data across the TCGA cohort. Further, we thank all members of the Genome Data Science lab for discussions. G.P.-M. was funded by an AGAUR FI fellowship, and F.S. was funded by the ICREA Research Professor program. Work in the lab of F.S. was supported by an ERC StG "HYPER-INSIGHT" (757700), Horizon2020 project "DECIDER" (965193), Spanish government project "REPAIRSCAPE", CaixaResearch project "POTENT-IMMUNO" (HR22-00402), the SGR funding of the Catalan government, and the Severo Ochoa Centers of Excellence award of the Spanish government to the hosting institution. The results published here are in whole or part based on data generated by the TCGA Research Network (https://www.cancer.gov/tcga).

Methods

TCGA datasets

We downloaded matched whole-exome sequencing (WES) aligned tumor and normal *bam* files and RNA-seq tumor raw *fastq* data from the The Cancer Genome Atlas (TCGA consortium)[24], using the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/), from which 8,809 samples spanning 31 different cancer types were retained for ASE analyses. Tumor purity estimates were also downloaded accordingly with TCGA data[24]. Copy number alteration (CNA) gene-level data, for arm-level, was obtained from the Broad Institute's GDAC Firehose portal (https://gdac.broadinstitute.org/). ATAC-seq *fastq* data for the analysis of allele-specific chromatin accessibility (ASCA) was also downloaded from TCGA through the GDC Data Portal, from which 381 samples spanning 23 different cancer types were retained. Additionally, DNA methylation data, represented as beta values for 486,427 CpGs across 8,689 samples and 32 cancer types, was retrieved using the *TCGAbiolinks* R package.

Sequencing data alignment

RNA-seq and ATAC-seq sequencing reads were aligned to the human genome *GRCh38.d1.vd1* using *STAR* v2.5.3a[50]. Reads with a minimum mapping quality of 255 were retained and allele-mapping bias was corrected using *WASP*[51]. An unbiased removal of duplicate reads was performed using *WASP* in-house script[52], which discards one of the duplicate reads at random.

Sequencing data variant calling and annotations

Variant calling was performed using *Strelka2* v2.9.10[53]. For WES, we compared normal and tumor *bam* files using *configureStrelkaSomaticWorkflow.py* with flags --exome --tumorBam and --normalBam. For RNA-seq and ATAC-seq, we used only tumor bam files, using *configureStrelkaGermlineWorkflow.py with flags --exome --rna --forcedGT*. This last argument forces the calling in the previously called somatic genotypes from WES. SNVs and INDELs were called separately in all three cases and merged afterwards. *VCF* files were annotated using ANNOVAR[54] (2020-06-07 version) with GENCODE v26 (ENSEMBL v88) and adding *gnomad211_exome* database to obtain the population minor allele frequencies (MAF) from gnomAD[55].

Data quality control processing

We selected heterozygous synonymous and missense SNVs, as well as nonsense variants, that had a total number of at least 5 RNA-seq (or ATAC-seq, depending on the analysis) read counts covering the locus (both mutated and wild-type), and a “PASS” at the FILTER column obtained from *Strelka2* WES variant calling. Additionally, missense mutations located within genes also bearing stopgain, frameshift indels, or mutations altering splicing in the same tumor sample were discarded.

Variant pathogenicity predictions

Variant pathogenicity predictions were obtained using VARITY[25], and downloaded from http://varity.varianteffect.org/. Missense variants were assigned an estimated pathogenicity score according to the “VARITY_ER” model. Variants mapping to multiple scores were assigned the

average of their predicted scores. Missense variants with a predicted VARITY_ER < 0.5, were relabeled, along with all synonymous SNVs, as effectively synonymous (ES) variants, while >= 0.5 VARITY_ER implies a pathogenic (impactful) missense variant.

Functional classification of genes

Data for the classification of individual genes according to their role in cancer was obtained from the Cancer Gene Census (CGC)[27]. CGC genes were classified according to their annotated role in cancer as TSGs, OGs, or both. Additionally, the *TP53* gene was treated as a separate group, as there were a high number of variants from that gene. Lists of essential genes were obtained from the CEG2 set determined from CRISPR screens on cell lines[28]. All remaining genes were classified as passenger genes.

Data on the association of genes with specific cancer types was obtained from the MutPanning catalog of significantly mutated driver genes[29]. Variants from cancer genes were classified as "cognate" on those cancer types with an existing association in MutPanning at FDR <= 25%, and "noncognate" otherwise.

ASE quantification from read counts

Allelic imbalances were calculated as depicted in Fig. 1A. More specifically, we calculated the differences in ratios of the mutated (MUT) to wild-type (WT) allele counts between WES, RNA-seq, as well as the expected ratios, in the form of log2 odds ratios (OR). This was in order to quantify the effect of ASE, in which the expected variant allele frequency (E) was estimated as:

$$\mathbf{E} = 0.5 * Tumor\ purity. \qquad \text{(Eq. 1)}$$

The effect sizes of *cis*-ASE, CNA-ASE and net-ASE were then quantified as follows:

$$\textbf{\textit{cis}-ASE} = log2\left(\frac{\boldsymbol{MUT_{RNA}\ /\ WT_{RNA}}}{\boldsymbol{MUT_{DNA}\ /\ WT_{DNA}}}\right) \qquad \text{(Eq. 2)}$$

$$\textbf{CNA-ASE} = log2\left(\frac{\boldsymbol{MUT_{DNA}\ /\ WT_{DNA}}}{\boldsymbol{E\ /\ (1-E)}}\right) \qquad \text{(Eq. 3)}$$

$$\textbf{net-ASE} = log2\left(\frac{\boldsymbol{MUT_{RNA}\ /\ WT_{RNA}}}{\boldsymbol{E\ /\ (1-E)}}\right) = cis\text{-ASE} + \text{CNA-ASE} \qquad \text{(Eq. 4)}$$

We also applied continuity correction by adding a pseudocount of 0.5 to all allele counts. Positive OR values indicate a preference for the expression of the mutated allele over the wild-type allele, while negative OR values suggest a greater relative expression of the wild-type allele compared to the mutated allele. We considered there was an appreciable amount of ASE whenever we observed more than a two-fold difference between allele counts (log2(OR) > 1). Based on these criteria, somatic variants were categorized into three groups for the scope of our analysis:

- Positive ASE: log2(OR) > 1
- Negative ASE: log2(OR) < -1
- No ASE: -1 < log2(OR) < 1

ASE variability estimates and comparison with prior studies

To quantify the overall contribution of *cis*-ASE and CNA-ASE to the overall mRNA imbalance (net-ASE), we used the formula from a recent study[18]:

$$\textit{cis}\text{-ASE explained variance} = \frac{\boldsymbol{var(cis_ASE) + cov(cis_ASE + CNA_ASE)}}{\boldsymbol{var(net_ASE)}} \quad \text{(Eq. 5)}$$

$$\text{CNA-ASE explained variance} = \frac{\boldsymbol{var(CNA_ASE) + cov(cis_ASE + CNA_ASE)}}{\boldsymbol{var(net_ASE)}} \quad \text{(Eq. 6)}$$

To replicate the findings from the Correia *et. al.* study[18] on *cis*-ASE contribution in BRCA patients and compare our findings to theirs, we adjusted our dataset using a similar filtering approach. We progressively applied more stringent filterings and then estimated our *cis*-ASE and CNA-ASE contributions using the above formulas. This included: (1) omitted adjustment by sample tumor purity; (2) additionally restricted the analysis to missense mutations; (3) additionally focusing only in BRCA patients; (4) additionally applying an RNA-seq coverage threshold of at least 30 counts (sum of WT and MUT alleles); (5) additionally focusing solely in the *PIK3CA* gene, all as per Correia *et al.*. We found that our *cis*-ASE contribution decreased from our original 54.6% to 23% (Supp. Fig. S1A) after the last filtering (5), which broadly aligns with the 16% reported by Correia *et. al.*[18] Therefore we infer the difference stems from a more comprehensive analysis in our case, including various cancer types and genes and variants, and purity adjustment.

To compare our results with the second study from PCAWG, a transcriptomic analysis of 1,188 RNA-seq samples from several cancer types[8], we emulated their model using a linear regression. It included variables akin to those they used but adapted for our dataset's focus on WES data rather than WGS, which limits the scope of detectable non-coding variants (especially germline eQTLs), specifically: (i) CNA-ASE, (ii) presence of rare germline variants (MAF < 1%, encompassing both exonic and intronic variants) as a binary indicator for any gene-sample pair, (iii) an imprinting binary variable based on known imprinted genes[56], (iv) the sum of somatic coding SNVs per gene-sample, and (v) the sum of somatic non-coding SNVs (excluding intronic, keeping UTR and splicing variants) per gene-sample pair. We then calculated the relative contribution of each variable by their effect size proportionally across the total effect sizes in the model. The results showed that CNA-ASE accounted for 82.7% of the total effect size explained by those variables that were tested (Supp. Fig. S1B). There were additional contributions from SNV-coding (6.37%), imprinting (5.36%), rare germline variants (5.14%), and non-coding SNVs (0.45%). This broadly aligns with the PCAWG findings where CNAs accounted for 86.1% of the explained effect size of the variables that were tested. Therefore, both approaches, which rely on explicitly defined putatively causal variables in a model, estimated the contributions of the remaining tested variables (apart from CNA) on ASE to be between 15.7% (PCAWG) and 17.3% (us). Crucially, this type of modeling can estimate the relative contribution of non-CNA-ASE (*cis*-ASE) only due to the known variables that were quantified and included in the model, but ignores the non-CNA-ASE (*cis*-ASE) stemming from unknown sources. This underestimates the relative contribution of the non-CNA-ASE, because it is likely an aggregate of multiple mechanisms, which may or may not have been measured and included in the model, for instance the cancer epigenetic evolution. In contrast to this analysis, our main analyses consider all systematic ASE signals that are not due to CNA, and find its contribution to be substantial and comparable to the CNA.

Additionally, for the non-CNA mechanisms, their quantification is likely to be more noisy (e.g. estimating the effects of non-coding variants is notoriously difficult), while large CNA can be estimated fairly accurately, further biasing the apparent effect in favor of CNA-ASE in this type of analysis that quantifies explainable variance. In contrast, in our analyses we quantify all systematic ASE, regardless of the source, and by a subtraction of the CNA-attributable ASE arrive at the 54.6% estimate of ASE from non-CNA sources. We note, nevertheless, that these estimates could be affected by various sources of noise.

Correlating CNA burden with CNA-ASE

CNA burden was calculated using gene-level and arm-level data, obtained from the Broad Institute's GDAC Firehose portal (https://gdac.broadinstitute.org/), with values ranging from -2 (deletion) to +2 (gain). We computed the total sum of CNAs per sample as an estimate of the overall CNA burden. The distributions of these burdens across samples for each cancer type are illustrated in Supp. Fig. S2, where cancer types are ranked by their increasing percentage contribution of CNA-ASE to net-ASE. Additionally, a Spearman correlation between ranks of cancer types by CNA-ASE contribution against their ranks based on the average CNA burdens revealed a trend where cancer types with greater CNA burdens exhibit higher contributions of CNA-ASE.

ASE-dNdES test for selection

To assess selection acting on cancer genes upon ASE, we implemented a conditional dN/dS-like test, termed ASE-dNdES. This analysis was conducted to determine whether pathogenic missense (N) variants exhibited higher enrichment levels of positive ASE (log2(OR) >1, compared to negative ASE and no ASE) compared to effectively synonymous (ES) mutations, indicative of selective pressures in cancer. We computed odds ratios (ORs) and *p*-values using Fisher's Exact test and conducted separate analyses for each type of ASE—net, *cis*, and CNA. The ASE-dNdES was applied at two levels: across gene categories and at the individual gene-level.

Gene category-level enrichment: The ASE-dNdES was applied across OGs, TSGs (including *TP53*), and passenger groups of genes, excluding nonsense mutations due to potential confounding effects from NMD. We analyzed nonsense mutations separately changing the direction of the ASE-dNdES test, that is, assessing the enrichment of negative ASE (log2(OR) < -1) compared to positive and no ASE categories, across all merged gene categories including essential genes.

Individual gene-level enrichment: We performed ASE-dNdES tests on a subset of cancer genes based on two filterings:

Subset 1) Initially we selected 550 known cancer genes from the CGC and then incorporated an additional 243 genes identified by MutPanning. From this pool, 463 genes were excluded because they were either associated with translocations (as per CGC) or appeared only in one cancer type in MutPanning leaving 330 mutational cancer driver genes for further analysis.

Subset 2) In addition to these 330 known driver genes, we assessed all genes exome-wide for mean ASE differences, that is, evaluating whether the log2(OR) ASE on missense variants exceeded that on ES variants, excluding nonsense variants. Genes were first filtered to ensure an enough count of ES (>10) and missense (>20) variants, leading to 2,394 eligible genes for this

pre-selection test. Using *t-tests,* we derived *p*-values and applied a permissive FDR threshold of 50% to identify 43, 20, and 57 genes with putative signatures of selection on net-ASE, *cis*-ASE, and CNA-ASE, respectively.

Further filtering involved removing genes lacking sufficient variants in either missense or ES categories, or positive ASE or negative ASE or no ASE categories (these filterings were done separately for net-ASE, CNA-ASE and *cis*-ASE categories). After merging the sets of genes obtained according to both criteria above, we performed separate ASE-dNdES tests for:

- **net-ASE:** Included 310 genes, combining 267 known driver genes (Subset 1) with 43 pre-selected ones (Subset 2).
- **CNA-ASE:** Included 253 genes, integrating 196 known driver genes with 57 pre-selected ones.
- ***cis*-ASE:** Including 291 genes, 271 known driver genes and 20 from the pre-selection test.

Following these ASE-dNdES tests, genes with an FDR < 10% were considered significantly selected, with a positive selection indicated by an OR of > 1 and negative selection by an OR of < 1.

Survival analyses

To perform survival analyses, we downloaded clinical data for TCGA from GDC (https://portal.gdc.cancer.gov/). The analyses were conducted using the *survival* package for R[57,58] and the *survminer* package[59] for Kaplan-Meier (KM) curves with *p*-values obtained through log-rank tests. Alive patients were censored at the end of their follow-up. Survival outcomes were evaluated for patients with somatic missense mutations in our significantly positively selected genes within three types of ASE separately: net-ASE (9 genes), *cis*-ASE (4 genes), and CNA-ASE (9 genes). The patients whose tumors exhibited ASE in these genes were compared against those without significant ASE in these genes. The ASE variable was stratified into seven bins for more detailed analysis: negative ASE (< -0.5); neutral/reference ([-0.5, 0.5]); and positive ASE across five ranges ([0.5, 1], [1, 2], [2, 3], [3, 4], and >4).

To account for multiple covariates, Cox proportional hazards models were employed at various levels—pan-cancer, individual cancer types, and individual genes—using the *survival* package in R. The models were applied across 31 distinct cancer types and also focused separately on each positively selected gene to assess the individual impacts. CNA data was obtained for all genes and cancer types separately, downloaded from http://firebrowse.org/. CNA burden was measured as the total number of CNA events in genes per sample (discretized to quartiles). Genetic intratumor heterogeneity data (estimated as copy-number heterogeneity, CNH) was obtained from a previous study[34], and was discretized to quartiles. Sex of the patient, tumor sample purity, tumor stage, and age of the patient (discretized as deciles) were also included in the model as covariates. A total of 86 recurrent large-scale or arm-level CNA events from a previous study[35] were also used as covariates in the pan-cancer analysis only. Results were plotted using the *forestmodel* package for R[57].

In circumstances where specific cancer types (LAML, THYM, and DLBC) lacked tumor purity estimates or when the Cox regression model failed to converge, those analyses were excluded. Additionally, models were discarded when hazard ratios (HRs) had values above 20 or confidence intervals (CIs) had “Infinite” or values above 15, signaling the regression analysis failed. When analysis was stratified by cancer type, we made appropriate adjustments, such as removing the

“sex” variable for sex-specific cancers or removing the “Stage” variable when it was missing in over 90% of samples in that cancer type.

ASCA analysis

Similarly as done with ASE, after performing variant calling on the ATAC-seq data, we selected variants with a total number of at least 5 ATAC-seq counts, and a “PASS” at the FILTER column obtained from WES variant calling. We quantified the imbalances at the chromatin accessibility level corrected by CNA events as described in section “ASE quantifications”, here using ATAC-seq counts instead of RNA-seq counts as in the ASE analysis. Hence, ASCA can be considered as analogous to *cis*-ASE signal, but for ATAC-seq data. We selected matched variants from WES, RNA-seq and ATAC-seq for our analysis, and kept those having more than 20 total (i.e. sum of WT and MUT alleles) RNA-seq and ATAC-seq reads. In order to test for the association between ASCA and *cis*-ASE we computed *Pearson* product-moment correlation coefficients.

Gene-level DNA methylation status

We sourced DNA methylation data from the TCGA, represented as a matrix of beta-values for ~450k CpG sites across all samples and cancers. Initial processing involved several filtering steps: (1) Removal of cross-reactive CpGs[60,61]; (2) Exclusion of probes with missing values (NAs) in more than 20% of samples, applied separately for each cancer type; (3) When multiple samples existed for an individual, average methylation values were computed.

Before analysis, beta-values were converted to M-values using the *beta2m* function from the *lumi* R package. We focused on CpGs near the transcription start site (TSS) to examine regions likely influencing transcription initiation. Specifically, we averaged M-values across CpGs located within "TSS1500", "TSS200", "1stExon", and "5’ UTR" categories per the Illumina 450k Manifest (https://support.illumina.com/downloads/infinium_humanmethylation450_product_files.html). We included additional probes within the first 2kb of the gene body as well, determined by distances from the TSS to CpGs using gene annotations from GENCODE v22 (https://zwdzwd.github.io/InfiniumAnnotation). Only CpGs within a 2kb range from the main isoform's TSS (using MANE as a reference isoform[62]), were included.

This approach yielded a single M-value per gene per individual, reflecting gene-specific DNA methylation levels near the promoter region. We then categorized this gene promoter-proximal DNA methylation status across the entire cohort based on percentile ranks separately per gene: samples with M-values in the lowest 30% for a given gene were labeled as unmethylated, those above the 70% percentile as methylated, and those in between as partially methylated (as would result from one allele being methylated and the other unmethylated). These methylation statuses were subsequently correlated with our *cis*-ASE values to explore potential relationships between methylation patterns and allele-specific expression.

Variant effect predictions on chromatin profiles by Sei

To predict the impact of mutations on chromatin state, we utilized the *Sei* tool[36], which uses deep neural networks to assess the chromatin-altering effects of genetic variants, within our ASE dataset (792,303 unique somatic mutations), with default parameters. *Sei* analyzes an extensive range of publicly available chromatin profiles—totaling 21,907—covering a wide spectrum of *cis*-regulatory elements, including transcription factor binding profiles, histone modification marks,

and chromatin accessibility profiles, providing insights with single-nucleotide resolution. For simplicity, we employed the simplified categorical classification by *Sei*, focussing on four distinct "sequence classes", associated with active transcription (TN1-TN4). We classified our somatic mutations based on whether they activated ("TN activating", *Sei* scores > 0.0001) or deactivated ("TN inactivating", *Sei* scores < -0.0001) separately for the transcriptional states TN1-TN4.

Variant effect predictions on splicing by SpliceAI and enrichment analysis

We evaluated the effect of splicing alterations on ASE in our dataset (792,303 unique somatic mutations), utilizing *SpliceAI*[39]. This tool employs a deep residual neural network to predict splicing changes at each position of a pre-mRNA transcript based on its genomic sequence. *SpliceAI* delta scores quantify changes in splice site activity due to mutations, across four categories: Donor Loss (DL), Donor Gain (DG), Acceptor Loss (AL), and Acceptor Gain (AG). These delta scores reflect the increase or decrease in splice site usage between the mutated and reference alleles. Additionally, we introduced a "MaxDelta" score representing the maximum delta score across the four types for each variant, to capture the most impactful potential splicing change. Mutations with *SpliceAI* delta scores ≥ 0.15 were classified as high splicing impact, while others were considered splicing-neutral.

Similarly as in the ASE-dNdES test, to explore the enrichment of positive and negative *cis*-ASE (log2(OR) >1 or < -1, respectively, both compared to no *cis*-ASE) among high-impact splicing versus neutral mutations, we conducted an enrichment analysis separately for each SNV type: effectively synonymous (ES) and pathogenic missense mutations (N). The ES group was further divided into non-pathogenic missense and synonymous variants and tested separately for the splicing test. Enrichment analyses were performed for each of the five *SpliceAI* delta score categories, calculating ORs and *p*-values using Fisher's Exact test.

This analysis was conducted at two levels:

Gene category-level enrichment: Applied across gene categories including cognate and noncognate OGs, TSGs, *TP53* as a separate category, essential genes, and passenger genes.

Individual gene-level enrichment: Focused on the same subset of known cancer driver genes used in our ASE-dNdES tests (here excluding genes from the pre-selected tests i.e. Subset 2). Following the enrichment analyses, FDR corrections were applied separately for each stratification—by SNV type, *SpliceAI* delta score category, and *cis*-ASE direction (positive or negative). Genes were considered significant where FDR < 15%. In instances where no significant genes were identified in a specific analysis, we used a nominal $p < 0.1$ to create a ranking list of putative candidate genes.

Classification of nonsense mutations into NMD-triggering and NMD-evading

Nonsense mutations often lead to nonsense-mediated mRNA decay (NMD) by creating premature termination codons (PTCs) on the coding sequence. However, some PTCs can evade this pathway, allowing mRNA to escape degradation. We classified nonsense mutations as NMD-triggering or NMD-evading based on established genomic rules[31]: (1) *55nt-rule:* nonsense mutations located ≤ 55 nt downstream from the last base of the penultimate exon are categorized as NMD-evading; (2) *Last exon rule:* Nonsense mutations within the last exon are considered NMD-evading; (3) *Start-proximal rule:* Nonsense mutations within the first 200 nt of the transcript are deemed NMD-evading.

References

1. Raghupathy, N. *et al.* Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics* **34**, 2177–2184 (2018).
2. Castel, S. E. *et al.* A vast resource of allelic expression data spanning human tissues. *Genome Biol.* **21**, 234 (2020).
3. Lo, H. S. *et al.* Allelic Variation in Gene Expression Is Common in the Human Genome. *Genome Res.* **13**, 1855–1862 (2003).
4. Pastinen, T. & Hudson, T. J. Cis-Acting Regulatory Variation in the Human Genome. *Science* **306**, 647–650 (2004).
5. Robles-Espinoza, C. D., Mohammadi, P., Bonilla, X. & Gutierrez-Arcelus, M. Allele-specific expression: applications in cancer and technical considerations. *Curr. Opin. Genet. Dev.* **66**, 10–19 (2021).
6. Ferguson-Smith, A. C. Genomic imprinting: the emergence of an epigenetic paradigm. *Nat. Rev. Genet.* **12**, 565–575 (2011).
7. Mayba, O. *et al.* MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol.* **15**, 405 (2014).
8. Calabrese, C. *et al.* Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).
9. Tuch, B. B. *et al.* Tumor Transcriptome Sequencing Reveals Allelic Expression Imbalances Associated with Copy Number Alterations. *PLOS ONE* **5**, e9317 (2010).
10. Bielski, C. M. *et al.* Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer. *Cancer Cell* **34**, 852-862.e4 (2018).
11. Lykke-Andersen, S. & Jensen, T. H. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.* **16**, 665–677 (2015).

12. Rhee, J.-K., Lee, S., Park, W.-Y., Kim, Y.-H. & Kim, T.-M. Allelic imbalance of somatic mutations in cancer genomes and transcriptomes. *Sci. Rep.* **7**, 1653 (2017).
13. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
14. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
15. Liu, Z., Dong, X. & Li, Y. A Genome-Wide Study of Allele-Specific Expression in Colorectal Cancer. *Front. Genet.* **9**, 570 (2018).
16. Batcha, A. M. N. *et al.* Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia. *Sci. Rep.* **9**, 11796 (2019).
17. Przytycki, P. F. & Singh, M. Differential Allele-Specific Expression Uncovers Breast Cancer Genes Dysregulated by Cis Noncoding Mutations. *Cell Syst.* **10**, 193-203.e4 (2020).
18. Correia, L. *et al.* Allelic expression imbalance of PIK3CA mutations is frequent in breast cancer and prognostically significant. *Npj Breast Cancer* **8**, 71 (2022).
19. Black, J. R. M. *et al.* RNA allelic frequencies of somatic mutations encode substantial functional information in cancers. 2023.03.09.531725 Preprint at https://doi.org/10.1101/2023.03.09.531725 (2023).
20. Sherr, C. J. Principles of Tumor Suppression. *Cell* **116**, 235–246 (2004).
21. Clayton, E. A. *et al.* Tumor suppressor genes and allele-specific expression: mechanisms and significance. *Oncotarget* **11**, 462–479 (2020).
22. Atak, Z. K. *et al.* Interpretation of allele-specific chromatin accessibility using cell state–aware deep learning. *Genome Res.* **31**, 1082–1096 (2021).
23. O'Neill, K. *et al.* Long-Read Sequencing of an Advanced Cancer Cohort Resolves Rearrangements, Unravels Haplotypes, and Reveals Methylation Landscapes. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.4743887 (2024).

24. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

25. Wu, Y. *et al.* Improved pathogenicity prediction for rare human missense variants. *Am. J. Hum. Genet.* **108**, 1891–1906 (2021).

26. Livesey, B. J. & Marsh, J. A. Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol. Syst. Biol.* e11474 (2023) doi:10.15252/msb.202211474.

27. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

28. Hart, T. *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 GenesGenomesGenetics* **7**, 2719–2727 (2017).

29. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).

30. MacArthur, D. G. *et al.* A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **335**, 823–828 (2012).

31. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).

32. Besedina, E. & Supek, F. Copy number losses of oncogenes and gains of tumor suppressor genes generate common driver events of human cancer. 2023.08.05.552104 Preprint at https://doi.org/10.1101/2023.08.05.552104 (2023).

33. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324–1335 (2014).

34. van Dijk, E. *et al.* Chromosomal copy number heterogeneity predicts survival rates across cancers. *Nat. Commun.* **12**, 3188 (2021).

35. Palou-Márquez, G. & Supek, F. Variable efficiency of nonsense-mediated mRNA decay across human tissues, tumors and individuals. 2024.02.29.582778 Preprint at

https://doi.org/10.1101/2024.02.29.582778 (2024).

36. Chen, K. M., Wong, A. K., Troyanskaya, O. G. & Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.* **54**, 940–949 (2022).
37. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* **47**, 1242–1248 (2015).
38. Jung, H., Lee, K. S. & Choi, J. K. Comprehensive characterisation of intronic mis-splicing mutations in human cancers. *Oncogene* **40**, 1347–1361 (2021).
39. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).
40. Varley, J. M. *et al.* Genetic and functional studies of a germline TP53 splicing mutation in a Li-Fraumeni-like family. *Oncogene* **16**, 3291–3298 (1998).
41. Shiraishi, Y. *et al.* Systematic identification of intron retention associated variants from massive publicly available transcriptome sequencing data. *Nat. Commun.* **13**, 5357 (2022).
42. Martínez-Ruiz, C. *et al.* Genomic–transcriptomic evolution in lung cancer and metastasis. *Nature* **616**, 543–552 (2023).
43. Tao, Z. *et al.* The repertoire of copy number alteration signatures in human cancer. *Brief. Bioinform.* **24**, bbad053 (2023).
44. Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. *Nature* **606**, 984–991 (2022).
45. Besedina, E. & Supek, F. Copy number losses of oncogenes and gains of tumor suppressor genes generate common driver mutations. *Nat. Commun.* **15**, 6139 (2024).
46. Du, Q. *et al.* Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat. Commun.* **10**, 416 (2019).
47. Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).

48. Salvadores, M. & Supek, F. Cell cycle gene alterations associate with a redistribution of mutation risk across chromosomal domains in human cancers. *Nat. Cancer* **5**, 330–346 (2024).

49. Krawczak, M. *et al.* Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum. Mutat.* **28**, 150–158 (2007).

50. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

51. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).

52. WASP/mapping at master · bmvdgeijn/WASP. *GitHub* https://github.com/bmvdgeijn/WASP.

53. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).

54. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

55. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

56. Morison, I. M., Ramsay, J. P. & Spencer, H. G. A census of mammalian imprinting. *Trends Genet.* **21**, 457–465 (2005).

57. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. (Springer, New York, 2000).

58. Therneau, T. M., until 2009), T. L. (original S.->R port and R. maintainer, Elizabeth, A. & Cynthia, C. survival: Survival Analysis. (2023).

59. Kassambara, A., Kosinski, M., Biecek, P. & Fabian, S. survminer: Drawing Survival Curves using ‘ggplot2’. (2021).

60. Chen, Y. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).

61. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).

62. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).

63. Richer, S. *et al.* Widespread allele-specific topological domains in the human genome are not confined to imprinted gene clusters. *Genome Biol.* **24**, 40 (2023).