Allele imbalance in mRNA is selected in tumorigenesis, results from epigenetic alterations 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

Allelic imbalance (AI) in levels of mRNAs that originated from the paternal and maternal copies of a gene can contribute to phenotypic variation and influence disease traits, including cancer. It is widely appreciated that AI at the DNA level, resulting from somatic copy number alterations (CNA) in tumors, generates second-hit events by which tumor suppressor genes are inactivated and oncogenes are activated; the somatic CNAs are also a known cause of AI at the mRNA level. Apart from CNA, other mechanisms could lead to AI in mRNA expression and also drive cancer evolution. By integrating genomic and transcriptomic pan-cancer data, we show that mRNA AI favoring the mutant allele in driver genes is subject to positive selection, generating second-hit events often independently of somatic CNA. The mRNA AI events associate with worse overall survival across all cancer types, outperforming other predictive markers. Various instances of mRNA AI are explained by the epigenetic changes in DNA accessibility and DNA methylation, affecting alleles differently. While we do not rule out that the driver somatic mutations themselves could induce allele-specific expression (we show an example of splicing-altering mutations, selected in various cancer genes), our study suggests that epigenetic alterations commonly act as second-hit events, generating mRNA allelic imbalances that drive cancer and identify new prognostic markers.

Introduction

Cancer genomes evolve by accumulation of somatic point mutations, structural variants (which often induce copy number alterations (CNA)), and epigenetic changes. During somatic evolution, healthy cells get transformed because cancer genes, including oncogenes (OGs) and tumor suppressor genes (TSGs), acquire driver point mutations, which largely act by changing the amino acid sequences of the encoded proteins; some structural variants also generate driver changes in a similar manner, by fusing or truncating proteins.

Similarly, cancer genes commonly acquire somatic CNA gains and losses, and epigenetic changes, of which promoter DNA hypermethylation is best studied. These types of alterations can act as tumor drivers by altering gene expression levels – the former, by changing DNA dosage, and the latter, by changing transcriptional output. An important research subject in tumor evolutionary studies is how these various types of driver changes interact with each other. This is epitomized by an observation, well-known from the pre-genomic era, that TSGs bearing damaging germline mutation often require a “second-hit” alteration such as to cause cancer, be it a somatic CNA deletion[1,2] or an epigenetic change[3]. Similarly, TSGs and also OGs bearing damaging somatic mutations are often additionally affected by CNA, acting as second-hit driver events[4,5].

In diploid genomes such as the human, each gene bears two copies referred to as alleles; somatic driver point mutations typically affect only one allele, with only rare observations of double somatic mutations in the same driver gene[6]. The somatic alterations that act via changing gene expression levels – CNA and epigenetic changes – can affect the mutant and the *wild-type* OG or TSG alleles differently, constituting an allelic imbalance (AI). The somatic CNA generate an AI at the DNA level, and favor increasing the proportion of the mutant allele of the driver gene by either CNA gains of the mutant or by CNA losses of the *wild-type* allele i.e. a loss-of-heterozygosity[5,7]. Effects of these DNA-level AI resulting from CNA propagate to the gene expression, resulting in mRNA-level AI. This was reported for impactful somatic mutations in individual examples of cancer genes (e.g. *PIK3CA*[8]) or cancer types (e.g. neuroblastoma[9]), and in a broader study of 1,188 RNA-seq cancer samples[10]. The mRNA-level AI can also be referred to as allele-specific expression (ASE) of the mutant transcripts; importantly here we do not imply that the studied mutations are in themselves necessarily causal to the imbalanced mRNA expression[11].

Apart from the cancer context and the somatic CNAs[4,10,12,13], allelic imbalance of gene expression is a known occurrence in the human genomic studies that consider germline variants[14–17]. While the mRNA AI can be often caused by *cis*-acting genetic variation – via occurrence of germline single nucleotide variants (SNVs) that alter gene *cis*-regulatory elements (such as enhancers and promoters)[16,17] or that alter coding regions to introduce stop codons and trigger nonsense-mediated mRNA decay (NMD)[18,19] – it can also arise by epigenetics. A well-known mechanism generating allele-specific expression in a defined set of human genes is genomic imprinting[20], in which only a specific parental copy of affected genes is expressed and the other is silenced through epigenetic mechanisms[21]. Beyond imprinting and chromosome X inactivation in females, there are reports of mitotically-stable, random monoallelic expression events affecting many

genes (reviewed in ref. [22]), associated with a histone mark signature[23], or close to topologically associated domains proximal to telomeres[24]. This suggests epigenetic effects may commonly generate mRNA-level AI in human cells, however the prevalence of these epigenetic AIs in cancer, their interactions with driver mutations and resulting impact on tumor evolution were less studied[25–27].

In this analysis, we compare allelic frequencies of driver somatic mutations between the tumor genome sequencing data (DNA-level, representing somatic CNA effects) and the tumor transcriptome (RNA-level), to dissect the effects of CNA versus other, non-CNA mechanisms causing mRNA AI by altering transcription activity or transcript processing[8,19].A particular hypothesis of interest is whether the non-CNA mechanisms generating mRNA-level AI can modulate selective pressures acting on driver genes, steering cancer evolution, in an analogous way to well-established mechanisms of CNA.

By integrating whole-exome sequencing (WES) and RNA-seq from 8,809 tumors[28], spanning 31 different cancer types, we find that non-CNA driven AI of somatic mutations is common in cancer genes, that it is under positive selection, and that it has strong prognostic accuracy of aggressive tumors. We further suggest that allele-specific epigenetic effects often underlie the non-CNA AI, and that these epigenetic changes largely do not result from the somatic mutation itself, but rather from a separate epimutation, which acts as a second-hit driver event on the cancer gene. Additionally, we find another cause of mRNA AI is not epigenomic, but rather post-transcriptional, and results from splicing-altering effects of driver missense and synonymous mutations in gene coding regions.

Results

Quantifying the DNA-level and mRNA-level AI on somatic driver mutations in cancer

Throughout this study, we conceptualize that mechanisms generating AI on cancer driver mutations can occur at two stages: i) somatic CNA events that alter dosage at the DNA level (henceforth, CNA-AI), and/or ii) other mechanisms altering allelic expression at the mRNA level independently of DNA dosage; We attribute this latter, non-CNA AI to epigenetic variation, or to genetic variation impacting regulatory regions, thus influencing transcriptional regulation. Hence, we refer to it as reg-AI (regulatory AI).

The observed differential abundance between allelic transcripts at the mRNA level (henceforth, mRNA-AI), is therefore the net composite of both CNA-AI and reg-AI. Thus, there are 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 dosage level via CNA-AI, only at mRNA dosage level via reg-AI, or a mixture of both (of note, we eschew the term “allele-specific expression”, because of an interpretation[11] that the imbalanced genetic variants were directly causal to the mRNA imbalance; we do not imply this in our study of somatic driver mutations).

Our approach consisted in separating the known effect of CNA-AI from the other effects on mRNA allelic expression imbalance, considering each somatic SNV mutation in cancer exomes. Prior studies of allelic imbalances in cancer compared transcriptomes of healthy versus cancerous tissues, quantifying AI of presumably neutral germline variants used as markers of allele-specific expression[4,10,12,13,29]. In contrast, our work considers somatic mutations, comparing DNA and mRNA AI of driver mutations against a baseline of DNA and mRNA AI of somatic passenger mutations in the same genes, comparing across a large cohort of matched tumor genomes and transcriptomes. This approach does not require mRNA data from tumor-adjacent healthy tissues, and so enables a systematic study across all driver genes in ~9000 tumors of ~30 cancer types. Recent studies of AI in somatic mutations focused on a single gene[8] or a single cancer type[8,9,27,29–31]; we provide a comprehensive comparison of AI of driver mutations across human cancers.

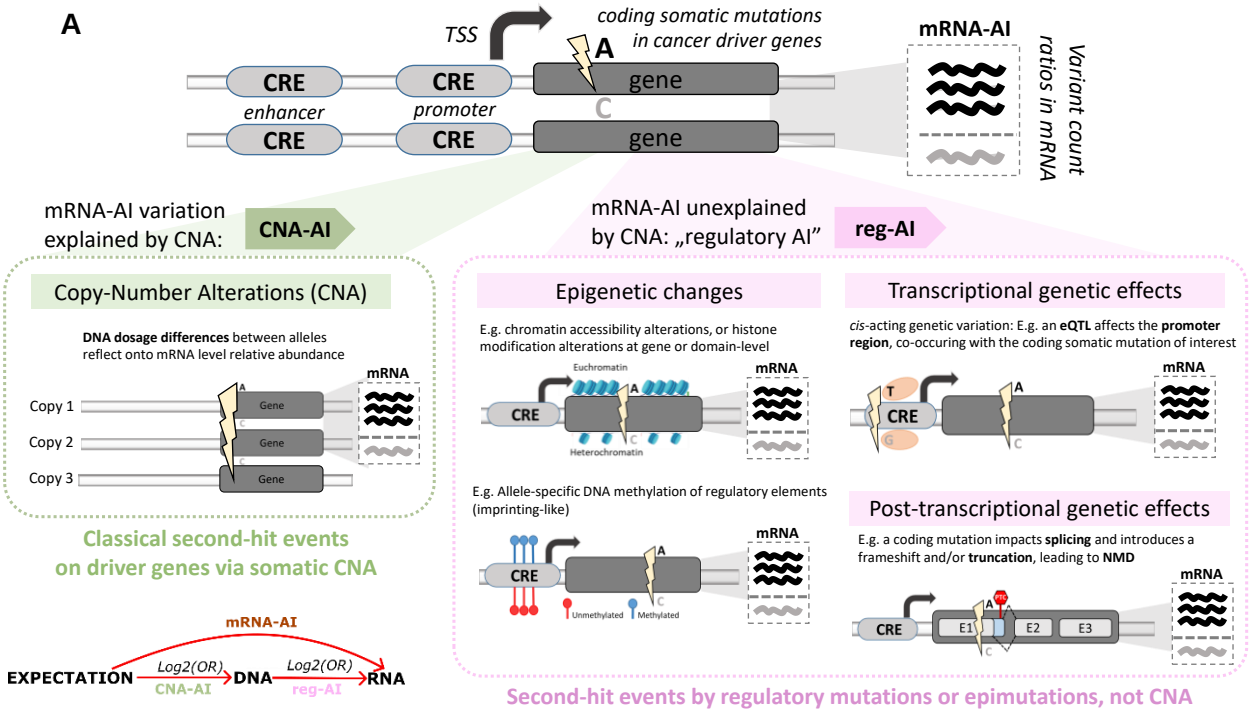
We aim to quantify the relative contributions of the well-understood CNA-AI and the less explored reg-AI towards the net mRNA-AI, and the implications of reg-AI on cancer evolution via positive selection affecting driver genes. To this end, we first quantified the DNA dosage effect of CNA-AI by comparing the expected allele fraction of the somatic mutation given heterozygosity (corrected by tumor sample purity), and the DNA allele fraction obtained from WES read counts (see Methods; schematic in Fig. 1A). Next, we quantified the effect of reg-AI by comparing the DNA variant allele counts in WES data and the RNA-level read counts obtained from RNA-seq data (adjusted for mapping effects of variants, see Methods), in both cases normalizing the somatic mutant allele by the *wild-type* allele. The summation of the two effects (measured as log base 2 odds ratios, $\log_2$(OR); see Methods) at the DNA-level – CNA-AI – and at the RNA-level corrected by DNA-level – reg-AI – estimates the net effect of mRNA-AI, or the overall imbalance observed in mRNA reads, relative to the expectation given a heterozygous somatic SNV with no CNA. Using matched exome and transcriptome data from 8,809 tumors spanning 31 cancer types from the TCGA project we obtained 854,785 total somatic mutations suitable for analysis (see Methods). In driver genes, the AI in these mutations will be contrasted between impactful mutations (nonsense, missense mutations with high VARITY score[32,33]) and less impactful mutations (missense with low VARITY score, and synonymous mutations) to estimate selection.

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

We hypothesized that mechanisms other than CNA may have a substantial impact on allelic mRNA expression imbalance of somatic variant alleles in cancer. We quantified the overall contribution of this reg-AI (see Methods).

In a pan-cancer analysis, we estimated that the non-CNA, “reg-AI” effects explained 55.6% of the total mRNA-AI variability (Fig. 1B). We observed differences between cancer types regarding the contribution of the non-CNA mechanisms on the overall mRNA-AI variability, ranging from 17.3% to 70.2% (Fig. 1B), however in all cancer types the reg-AI effects on somatic mutations were considerable. Among the major cancer types where the CNA-AI was the predominant mechanism underlying the mRNA-AI were ovarian serous cystadenocarcinoma (OV), thyroid carcinoma

(THCA), brain lower grade glioma and glioblastoma brain cancers (LGG and GBM), various sarcomas (SARC) and breast invasive carcinoma (BRCA). On the other extreme, cancer types where the non-CNA mechanisms (reg-AI) were dominant contributors to mRNA-AI were the diffuse large B-cell lymphoma (DLBC), uterine corpus endometrial carcinoma (UCEC), cervical squamous cell carcinoma (CESC), and skin melanoma (SKCM), reg-AI had the most influence on mRNA imbalance in expression of somatically mutated alleles. We note other major cancer types including colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KIRC) and lung cancers (adenocarcinoma: LUAD; squamous cell carcinoma: LUSC) also had a predominance of non-CNA, possibly epigenetic mechanisms generating mRNA imbalances on somatic mutations in exons.



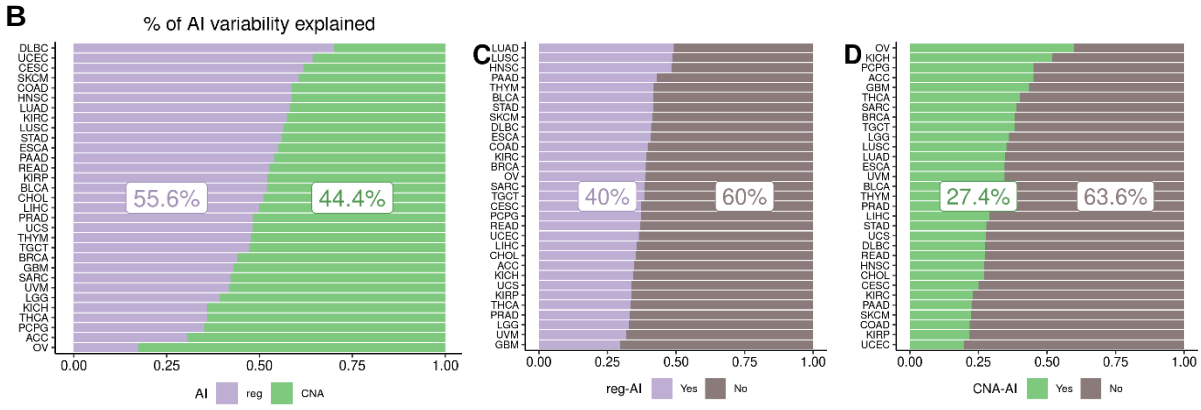


Figure 1. Determinants of allelic imbalance (AI) in cancer genomes

A, Illustration of two distinct paths through which AI may favor the expression of the mutant allele following an SNV event: (1) by CNAs alone (left) or (2) by genetic or epigenetic variations (right). We attribute this latter, non-CNA AI, to epigenetic changes and/or to genetic variations thant influence *cis*-regulatory elements (CREs), thereby affecting transcriptional regulation. We collectively refer to these non-CNA mechanisms as regulatory AI (reg-AI). Examples of epigenetic variation include histone modifications or chromatin accessibility or remodeling changes that occur without an underlying genetic variant (epimutation), as well as allele-specific DNA methylation, analogous to imprinting events. In some cases, a coding mutation can overlap with an intragenic CRE (e.g., an enhancer), disrupting a transcription factor (TF) binding site and producing downstream epigenetic effects. Alternatively, coding mutations may interfere with splicing or result in a nonsense mutation, both of which can introduce a premature termination codon (PTC), triggering nonsense-mediated decay (NMD). It includes a graphical summary (bottom left) of our approach to quantify the effects of CNA-AI and reg-AI. The sum of these two effects results in the total AI observed at the mRNA level (mRNA-AI), adjusted for sample purity. **B**, Relative contributions of reg-AI and CNA-AI to the overall mRNA-AI variation of all exome-wide somatic SNVs in TCGA, stratified by cancer type. It illustrates that AI caused by non-CNA mechanisms has a major impact on overall AI and is pervasive in cancer genomes (55.6% of variability). **C-D**, Percentage of exome-wide SNVs exhibiting a significant amount of reg-AI (**C**) and CNA-AI (**D**), determined as abs($\log_2$(OR)) > 1, stratified by cancer type.

To further quantify these differences, we assessed what proportion of exome-wide somatic SNVs showed an appreciable amount of reg-AI (defined as abs($\log_2$(OR)) > 1 i.e. two-fold difference between alleles), as well as CNA-AI abs($\log_2$(OR)) > 1. Remarkably, we found that the non-CNA, reg-AI mechanisms affected 40.0% of all somatic mutations, on average, across TCGA cancer types (Fig. 1C), while CNA-AI affected 27.4% of somatic mutations exome-wide (Fig. 1D). The fraction of somatic mutations affected by CNA-AI varied strongly across cancer types, being as high as 60.0% in OV (a very high CNA burden cancer type), and as low as 19.7% in the lower-CNA burden UCEC (Fig. 1D; Spearman correlation (R) of CNA-AI with CNA burden of cancer types is 0.29; Supp. Fig. S1). Unlike the CNA-AI, the percentage of variants affected by reg-AI was more consistent across cancer types, ranging from 29.7% to 49.2%, suggesting that the non-CNA mechanisms generating AI of somatic mutations in mRNA are widespread in human tumor tissues.

Our comprehensive analysis of exome-wide somatic mutations in 31 cancer types reports higher estimates of non-CNA causes of mRNA AI than previous, focused studies. For instance, Correia *et al.*[8] attributed 84% of allele-specific expression in the *PIK3CA* gene in breast cancer to CNA effects. To validate these findings[8], we applied similar filters and data stratification (see Methods). This approach, focusing on BRCA and *PIK3CA* somatic variants was broadly concurrent with previous analysis, where we estimated 23% contribution from reg-AI and 77% from CNA-AI (Supp. Fig. S2A). Another study of 1,188 RNA-seq samples in 27 cancer types[10], calculated the lower-bound estimate of reg-AI contribution at 15.7%. Their approach relies on a logistic regression model with explicitly-provided putatively causal variables to reg-AI, thus not quantifying unknown effects on reg-AI, e.g. resulting from epigenome remodeling in cancer[34–36]. In agreement, when we employed to our dataset a similar approach, meaning a linear regression model incorporating both CNA and various putatively causal non-CNA variables to predict our net

mRNA-AI (see Methods), we arrived at a comparable of 17.3% relative contribution for the known non-CNA variables (Supp. Fig. S2B). This indicates a comparatively larger role of the less-understood variables including global epigenome changes in tumors in generating the mRNA AI. These findings highlight the significant role of mechanisms other than CNA in generating mRNA AI on somatic mutations, which is more prevalent in cancer genomes than anticipated[8,10].

Selection on mRNA AI independent of CNA as a driver of tumorigenesis

We further hypothesized that this CNA-independent, mRNA-level AI superimposed onto somatic driver mutations can provide a selective advantage during cancer evolution, effectively generating second-hit events similarly as CNAs commonly do[4,5]. Indeed, epigenomic inactivation is a known second-hit event for inherited cancer risk variants in colon and breast cancers[3], therefore in principle somatic mutations might be affected by similar epigenetic two-hit mechanisms, which we comprehensively quantify here.

For our analyses quantifying genetic interactions of somatic mutations with AI in gene expression at the same locus, we classified each somatic SNV in gene coding regions according to their effect. We used the pathogenicity predictor VARITY_ER[32,33] to reclassify low-impact (potentially non-pathogenic) missense variants (with score < 0.5) and merge them together with the synonymous variants into an “effectively synonymous” (ES) group, increasing sample size for the synonymous baseline. Individual genes were also grouped as OGs (n = 233), TSGs (n = 246) and passenger genes not in Cancer Gene Census (CGC)[37] as a negative control (n = 15,572). Universally essential genes were taken from the CEG2 set determined from CRISPR screens on cell lines (n = 626)[38]. The *TP53* gene was treated independently, due to the large number of variants therein providing statistical power for individual analyses. We refer to “cognate genes” as those preferentially mutated in a specific cancer type, according to the MutPanning catalog[39]. 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 AI

To evaluate whether AI upon mutations is selected in cancer genes, we performed a conditional dN/dS-type test based on AI (AI-dNdES, see Methods). In this test, we assessed the proportion of variants with significant AI (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 AI were enriched for nonsense mutations, relative to those showing no AI or positive AI (mRNA-AI: OR = 3.33; non-CNA “reg-AI”: OR = 5.49, both p < 2e-16, Fisher test). This is explained by NMD, a surveillance pathway that degrades mRNAs containing premature termination codons, a well-known cause of allele-specific expression[19,40–42]. Henceforth, we excluded nonsense mutations to avoid confounding by NMD on AI, and in subsequent comparisons we contrast the impactful missense versus ES mutations.

Testing our hypothesis, we considered cognate TSGs (including *TP53*), in which missense mutations were considerably more prevalent within positive-AI group of tumors, than in the no-AI and negative-AI groups (mRNA-AI: OR = 4.65; reg-AI: OR = 2.64; CNA-AI: OR = 5.25; all $p <=$ 7.8e-63). As a control, passenger genes showed no significant differences in enrichment between positive mRNA-AI and no mRNA-AI or negative mRNA-AI for the missense mutations ($p = 0.26$, Fisher test).

To formally test this effect in driver genes resulted from selection, we quantified AI stratified by SNV functional impact category. We observed a notable imbalance towards the mutated allele on impactful missense variants, normalized to ES variants, in the case of cognate TSGs and in the case of *TP53* considered individually, and similarly but more subtly so in the case of cognate OGs (shown in Supp. Fig. S3A-C for overall mRNA-AI, and separately for reg-AI and CNA-AI). Consistently, AI correlated with the VARITY functional impact scores for missense mutations in cognate TSGs (mRNA-AI: Pearson R = 0.40; reg-AI: R = 0.25; CNA-AI: R = 0.36; all $p <=$ 3e-70; Supp. Fig. S3D). Thus, missense variants in driver genes showing preferential expression of the mutant allele are subject to stronger positive selection than driver gene variants not showing AI of the mutant allele, highlighting the key role of mRNA-level AI as second-hit events in tumor evolution. Our test of selection, when applied to our CNA-AI estimates, recapitulates the known interaction between CNA and mutations in driver genes[4,5], and when applied to reg-AI further suggests that the non-CNA causes of mRNA AI can have a comparably strong effect on selection on driver mutations.

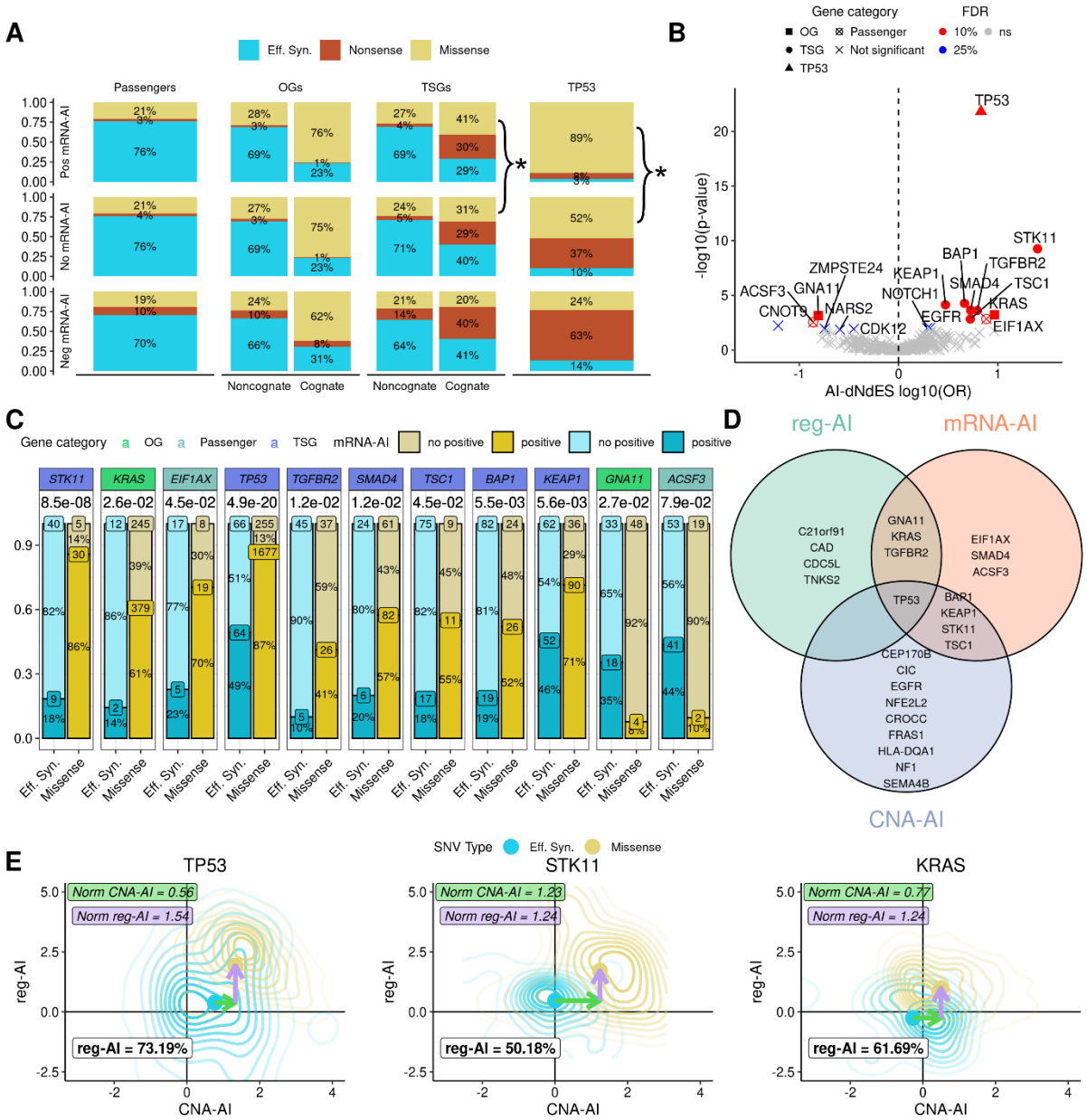


Figure 2. AI is subject to selection in cancer genomes, potentially contributing to tumorigenesis
A, Relative frequencies of effectively synonymous (Eff. Syn or ES), missense and nonsense somatic SNVs, stratified by mRNA-AI status and gene categories: passengers, oncogenes (OGs), tumor suppressor genes (TSGs) and *TP53*. OGs and TSGs were further divided into cognate and noncognate. For clarity, "positive mRNA-AI" refers to mutations where AI is > 1, "negative mRNA-AI" denotes AI < -1, and "no mRNA-AI" represents AI values between -1 and 1. **B**, Volcano plot for the conditional gene-level AI-dNdES test, displaying the odds ratios (ORs) on X-axis and raw *p*-values on Y-axis, both in $\log_{10}$ scale, highlighting significant genes under FDR < 10% or potential candidates at FDR < 25%. **C**, Proportion of SNVs with positive mRNA-AI in ES compared to missense variants for 11 significant genes from (B) that exhibit selection on the effect of mRNA-AI. The number of variants and the FDR is also shown. **D**, Overlap of significant hits from the AI-dNdES test in which selection was inferred for either mRNA-AI, reg-AI and CNA-AI. **E**, Two-dimensional distribution densities of CNA-AI and reg-AI 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 reg) had over the final mRNA-AI.

AI in known driver genes is subject to selection

To further identify individual driver genes subject to selection on positive AI (enriching the mutant allele in mRNA) we again performed our AI-dNdES test on somatic mutations, but at individual gene level; here, to increase power we considered the no net-AI and negative-AI groups jointly.

We pre-selected genes of interest for this test to reduce multiple-testing burdens. Firstly, we focused on known cancer genes from the CGC and additional genes identified by MutPanning[39] that are significantly mutated in two or more cancer types, leading to 330 genes (Subset 1). Secondly, in addition we tested all genes genome-wide for AI differences between missense and ES variants (Subset 2; testing across 2,394 eligible genes after filtering by number of variants, see Methods); a permissive threshold prioritized 43, 20, and 57 genes with tentative signatures of selection on mRNA-AI, reg-AI, and CNA-AI levels, respectively, and the candidate sets were further merged (both Subsets 1 and 2) and after further filtering (see Methods) subjected to significance testing using the dN/dES test on AI.

We performed separate AI-dNdES tests for the overall mRNA-AI (267 known drivers, Subset 1, and 43 candidates, Subset 2), and separately for the CNA-AI (196 drivers and 57 candidates) and the reg-AI (271 drivers and 20 candidates). At FDR < 10%, we identified significant positive selection on AI favoring the mutant allele in 11 mRNA-AI genes (Fig. 2B-C), and for 8 genes in reg-AI (Supp. Fig. S4 A-B) and 14 genes in CNA-AI (Supp. Fig. S4 C-D) genes. No inflation was observed in Q-Q plots, indicating the *p*-values are well calibrated (Supp. Fig. S5).

Among the genes where AI is a significant driver, *TP53* exhibited the strongest signal: 87% of missense variants showed positive mRNA-AI compared to 49% of neutral, ES variants (OR = 6.77, FDR = 6e-21, Fisher test, Fig. 2C); we note this estimate is conservative, because *TP53* has common synonymous variants that are under positive selection as they disrupt splicing[43], and so would result in AI, even though we classed all synonymous mutations as neutral. Alongside *TP53*, the genes *STK11*, *BAP1*, *KEAP1*, *TGFBR2*, *SMAD4*, *KRAS*, *EIF1AX,* and *TSC1* also revealed positive selection on AI, with 41-86% of missense variants exhibiting positive mRNA-AI, in contrast to only 10%-46% ES variants (range across the 8 genes, Fig. 2C). Conversely, *ACSF3 and GNA11*, the latter a recognized oncogene, displayed a signature of negative selection on the effect of positive (increased) mRNA-AI, where the fraction of impactful missense mutations (8-10%) appeared lower than ES (35-44%).

Tumor-suppressor genes were more common than oncogenes (Supp. Fig. S4E) in effects of CNA on AI: mRNA-AI (63% TSGs vs 18% OGs) and CNA-AI (43% vs 7%), however TSG and OG appeared similar in the non-CNA causes of AI (both at 25% for reg-AI). The other candidate genes

we prioritized through our mean AI differences pre-selection test (see above, Subset 2), were observed in mRNA-AI (*ACSF3*), reg-AI (*CAD*, *C21orf91*, *CDC5L* and *TNKS2*) and CNA-AI (*CEP170B*, *CROCC*, *HLA-DQA1*, *SEMA4B*) tests for selection at FDR < 10%, suggesting these genes have role in driving cancer as indicated by significant AIs in mRNA.

TP53 was the sole significant gene in both the CNA-AI and the reg-AI test, suggesting a convergence of both CNA and non-CNA mechanisms of AI in *TP53* (Fig. 2D). Other significant genes were significant by a single AI category test, indicating that selective pressures act on AI predominantly either via CNA or via non-CNA mechanisms on each gene (Fig. 2D). However, this subdivision of genes becomes less distinct at a more permissive FDR threshold, where we observe the overlap between the CNA-AI genes and reg-AI genes increases to 5 genes (*TP53*, *BAP1*, *BCOR*, *KEAP1*, and *STK11* at 50% FDR). This suggests that both AI mechanisms, CNA-dependent and CNA-independent may co-occur in these additional genes.

Within the 11 significant mRNA-AI selected genes, we aimed to dissect the relative contributions of CNA-AI versus the non-CNA, reg-AI, to the net mRNA imbalance favoring the mutant allele. We considered the AI difference between pathogenic (missense) and neutral (ES), for CNA-AI and reg-AI separately, and compared them to the AI difference in mRNA-AI. 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-AI and reg-AI. Remarkably, the majority of the signal came from the non-CNA mechanisms that we refer to as reg-AI; they contributed to 73.2% of the difference in mRNA-AI between missense and ES variants in *TP53*. This proportion varied somewhat depending on the specific gene, ranging between 31.2% and 85.3% reg-AI contribution to net mRNA-AI (Supp. Fig. S6). Overall, AI in gene expression favoring the mutant allele confer a selective advantage in various cancer driver genes. Non-CNA mechanisms are an important, and in some genes the dominant mechanism in generating the mRNA level AIs that are under positive selection in somatic evolution.

Positive selection on AI adversely impacts cancer prognosis

Given the signatures of positive selection on AI of driver somatic mutations, we hypothesized that this would be reflected in cancer aggressiveness. To test this, we performed overall survival analyses, in which we compared patients with tumors exhibiting AI on driver mutations, against patients with tumors with no detectable AI in drivers. For this, we selected only samples bearing at least one missense mutation in any of our gene sets, encompassing 9 significantly positively selected mRNA-AI genes (we excluded *GNA11* and *ACSF3* that appeared to have negatively selected mRNA-AI). Tumor samples in which the average mRNA-AI of missense mutations in these genes was > 1 (indicating a 2-fold difference, expressed on a $\log_2$ scale) were classified as “affected” by driver AI, while the rest were labeled as “not affected”. In the Kaplan-Meier (KM) curves (Fig. 3A), the AI 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 only 25.6% of the non-affected patients did. Similar results were observed for reg-AI (Supp. Fig. S7A, p = 1.8e-7) but, interestingly, not for CNA-AI (Supp. Fig. S7B, p = 0.22), using their corresponding

sets of positively selected AI genes (n = 4 and 9 genes, respectively). Considering the three AI analyses, the median overall survival for the AI affected group was 4.24 - 4.62 years, compared to 5.55 - 17.6 years for the AI non-affected group. Because this analysis does not consider cancer type or other covariates, we next conducted a Cox proportional hazards survival analysis (see Methods), incorporating cancer type and additional covariates known to be associated with cancer survival (Fig. 3B). We categorized our AI variable into negative AI (< -0.5 $\log_2$ OR units); neutral/reference ([-0.5, 0.5]); and positive AI 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 genetic heterogeneity (estimated as copy-number heterogeneity, CNH)[44], tumor sample purity, total CNA burden, and 86 individual recurrent large-scale CNA events as estimated previously[45]. After controlling for these covariates, the analysis revealed that patients categorized in the positive mRNA-AI ranges faced a progressively higher risk of death, compared to mRNA-AI 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), where higher HRs were positively associated with higher degree of AI, and negative mRNA-AI category showed no association (HR = 0.97).

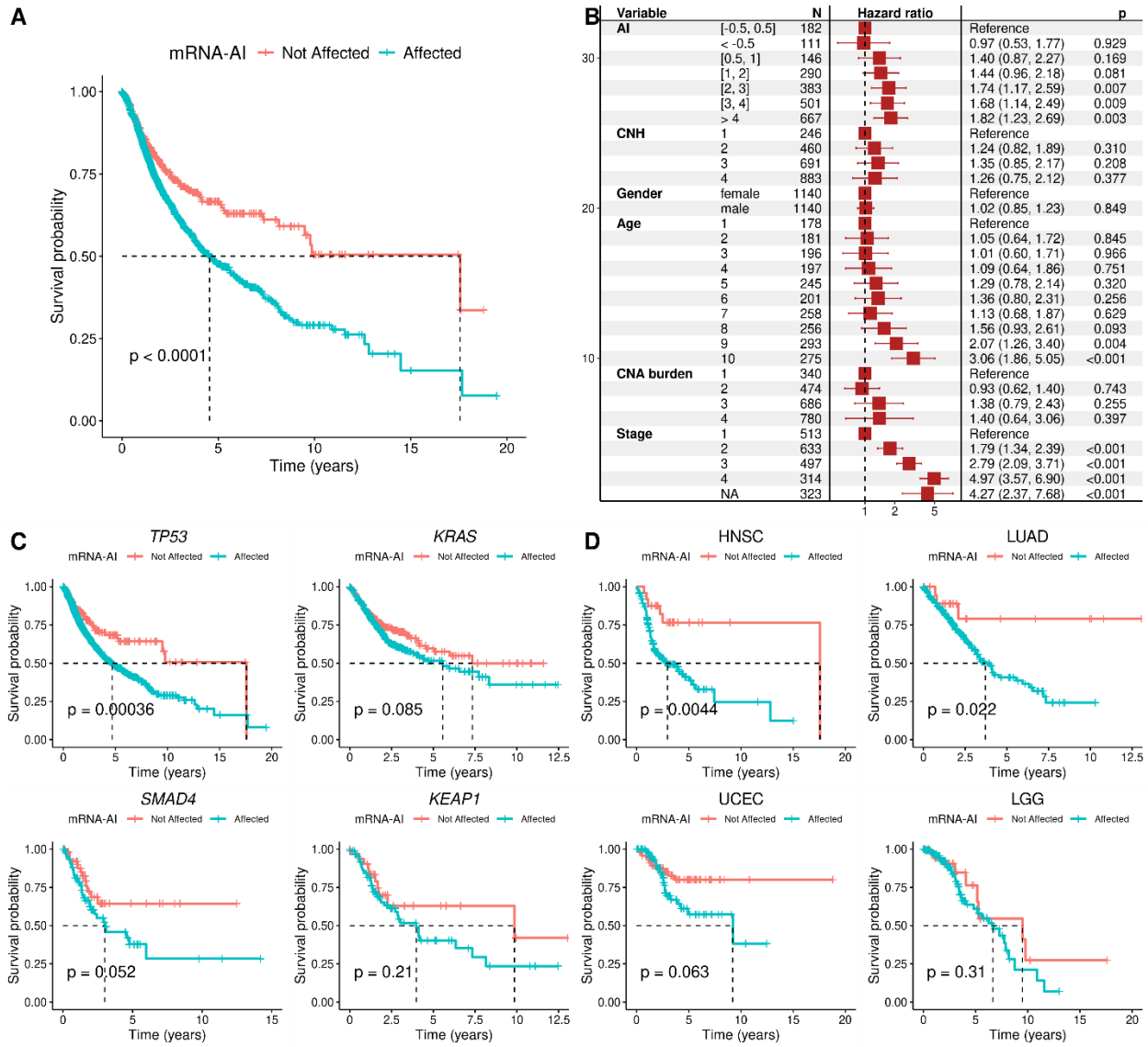


Figure 3. Positive selection on AI has a detrimental effect on overall cancer prognosis

A, Kaplan-Meier (KM) curve of pan-cancer survival in patients exhibiting positive mRNA-AI in missense variants affecting positively selected mRNA-AI genes ("Affected"; blue), against those still bearing mutations in those genes, but not showing positive mRNA-AI ("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, AI variable was discretized into seven groups for more detailed analysis: negative AI (< -0.5); neutral/reference ([-0.5, 0.5]); and positive AI across five ranges ([0.5, 1], [1, 2], [2, 3], [3, 4], and >4) to evaluate the impact of varying AI 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 mRNA-AI ("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 mRNA-AI ("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).

The effect size of positive mRNA-AI is substantial in comparison to known survival covariates: the HR of copy-number heterogeneity was of 1.26 for the highest quartile compared to the lowest (p = 0.38). Patients within the 8th and 9th age deciles, ranging from 68 to 77 years old, exhibit hazard ratios comparable to those seen with our mRNA-AI variable, ranging from 1.56 to 2.07 (p = 9.3e-2 to 4e-3). This implies that the prognostic impact of somatic mutation AI in tumor transcriptomes on cancer patient survival is comparable to the effect of being 29-38 years older (compared to our reference age group, which was 39 years old). Dissecting the mRNA-AI, we observed similar results specifically for reg-AI across the five positive AI categories (HR = 1.36-1.70; p = 0.11 - 1e-3, Supp. Fig. S7C). Conversely, the analysis for CNA-AI showed more modest 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 AI, rather than the CNA, are primarily responsible for cancer aggressiveness.

AI in individual genes and cancer types linked with cancer aggressiveness

Additionally, we asked if the association of AI 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). For example, considering only patients harboring at least one *TP53* missense variant, those affected by AI in *TP53* had a poorer prognosis than those who were not affected by AI: (Cox HR = 1.52 with p = 2.9e-2; KM curve in Fig. 3C). Similar trends were observed for *TP53* reg-AI (HR = 1.4 with p = 8.4e-3, Supp. Fig. S8A) and less significant impact for *TP53* CNA-AI (HR = 1.28 with p = 1.8e-2; Supp. Fig. S8B). A consistent signal was seen with *KRAS*, and large effect sizes with borderline significance were seen in *SMAD4* and *KEAP1* mRNA-AI associations with survival (Fig. 3C; separate results by CNA-AI versus reg-AI mechanism in Supp. Fig. S8).

Next, we asked in which cancer types the AI-survival associations were more prominent. A cancer type-specific analysis identified similar cancer types in the mRNA-AI (KMs in Fig. 3D and Cox in Supp. Fig. 9A) and reg-AI (Supp. Fig. S8C and S9B): head and neck squamous cell carcinoma (HNSC; KM p = 4.4e-3 and 8.6e-05, for mRNA-AI and reg-AI respectively), and LUAD (KM p = 0.02 and 8.5e-3). Other major TCGA cancer types such as UCEC and LGG showed a similar trend in the survival curves, with the former being marginally significant and latter non-significant (UCEC: p = 6e-2 and 5.7e-2; LGG: p = 0.31 and 0.52). Results from CNA-AI can be found in Supp. Fig. S8D and S9C.

In addition to our core set of driver genes where mRNA-AI on somatic mutations is positively selected, we hypothesized that AI in additional driver genes might also be prognostically relevant despite the genes not reaching significance in the initial test on AI, potentially due to statistical power issues. Indeed, when we repeated the mRNA-AI *versus* survival association test, we considered 40 genes significant only at a nominal p<0.10 in the AI test, finding that mRNA-AI in these genes is overall predictive (Supp. Fig. S10A-D). Thus, the set of survival-relevant mRNA-AI genes is likely larger than the 9 individually significant genes we prioritized above, and the individual genes might be reliably identified in larger transcriptomic datasets going forward.

Allelic differences in chromatin accessibility support epigenetic cause of reg-AI

The reg-AI signal in our data encompasses all types AI variation at the mRNA level (mRNA-AI) that cannot be explained by DNA-level AI originating from CNA (CNA-AI). Mechanisms that give rise to the selected reg-AI in cancer genes are of interest. Firstly, we considered the possibility that the reg-AI results from a somatic mutation (apart from CNA) affecting a regulatory region; this is analogous to germline eQTL variants often causing allele-specific expression by e.g. affecting transcription factor binding sites. A priori this explanation does not seem likely for our case of reg-AI selected cancer driver mutations, because driver somatic mutations affecting non-coding regulatory DNA are quite rare[10]. Known individual examples include *TERT* gene promoter driver mutations, and structural variants that translocate regulatory elements via promoter or enhancer hijacking[46,47]. We therefore considered the TCGA tumors from our original analyses that also had WGS data available (n = 777 tumors, from 22 cancer types; 44,373 WES variants available from those samples). We asked for how many examples where we observed AI in a driver gene we additionally observed a non-coding mutation near a gene promoter (conservatively, we include all variants ±1.5kb from the TSS, see Methods).

Among the 1,018 driver gene-sample pairs with exonic mutations exhibiting notable AI (defined as abs(reg-AI) > 1), 43.4% had a mutation (either SNV or indel) near the promoter . Comparing against a baseline of exonic mutations with no AI (abs(reg-AI) < 1), a similar proportion of cases (37.3%) had mutations near the promoter. This comparable percentage of co-occurrence suggests that TSS-proximal somatic variants are likely not the common drivers of mRNA-AI in cancer. Based on the excess over the baseline, we estimate promoter-affecting somatic variants might explain ~7% of cases of non-CNA, mRNA-level AI in driver genes.

Exploring further, we used a very stringent threshold for AI (i.e., examining reg-AI from < -4 to > 4), observing that higher intensity of positive reg-AI corresponded to increased co-occurrence of non-coding mutations near the TSS in cancer driver genes (from 45% at reg-AI > 1 to 63% at reg-AI > 4). In contrast, this co-occurrence percentage did not increase in passenger genes (from 37% at > 1 to 22% at > 4) (Supp. Fig. 11), implying selection on some cases of promoter-proximal mutations. This increased notably with more stringent thresholds for AI, from 8.3% at reg-AI > 1 to 41.3% at reg-AI > 4. In contrast, negative reg-AI values did not show significant differences between cancer and passenger genes, thus the selected effects of promoter-proximal mutations are only for overexpressing, but not for silencing an allele of a cancer driver gene. However, it's important to note that these cases remain relatively rare; in our dataset, the total number of cancer gene-tumor sample pairs with no AI is 1,518, 116 with reg-AI > 3, and 30 with reg-AI > 4. moreover, only a subset of these cases (reflected in the delta difference, see arrows in Supp. Fig. S11) exhibits an association between exonic AI and non-coding mutations near the TSS.

Therefore, the majority of gene-sample cases of reg-AI remain unexplained in this test; we propose they are more likely explained by epigenetic changes that occur irrespective of mutations. We acknowledge that it is possible that somatic mutations in distal enhancers affect gene

expression, however these are likely to have lower effects than the promoter-proximal mutations, which we tested above.

To further support that epigenetic causes underlies this CNA-independent expression AI phenomenon, here termed reg-AI, we sought signatures of allele-specific chromatin configuration. Specifically, we studied differential activity of DNA regulatory elements between alleles, which can be measured as local chromatin accessibility via e.g. ATAC-seq. To examine them, we used matched coding WES and ATAC-seq data from 381 TCGA patients[28], and we performed variant calling on the ATAC-seq as we have 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 AI and ASCA datasets, and studied the association between our reg-AI estimates 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 within the gene. We found that ASCA indeed significantly predicted reg-AI for impactful 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). Thus, the ASCA signal that we were able to measure in this analysis could predict reg-AI. 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. S12). Excluding nonsense variants, we observed that 46.8% of variants exhibiting positive ASCA also showed positive reg-AI, while only 3.7% showed negative reg-AI (Supp. Fig. S13). Conversely, only 6.4% of variants with negative ASCA presented positive reg-AI, while 27.7% exhibited negative reg-AI.

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. S14). 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 reg-AI and ASCA exhibited $\log_2$(OR) > 1, while CNA-AI < 0.5.

Overall, the link of allele-specific epigenomic state of regulatory elements, rather than mutations in promoter or proximal elements, more commonly explain the reg-AI we observe in somatic variants in cancer. Therefore, this somatic reg-AI would be analogous to imprinting, rather than to eQTLs, in that there need not be a genetic variant causing the allele-specific change in activity; instead, it would be an epimutation. Importantly, we do not rule out that the coding somatic variants that we observed AI on might in some cases alter the activity of a regulatory element overlapping the gene coding region, thereby causing imbalanced expression of itself.

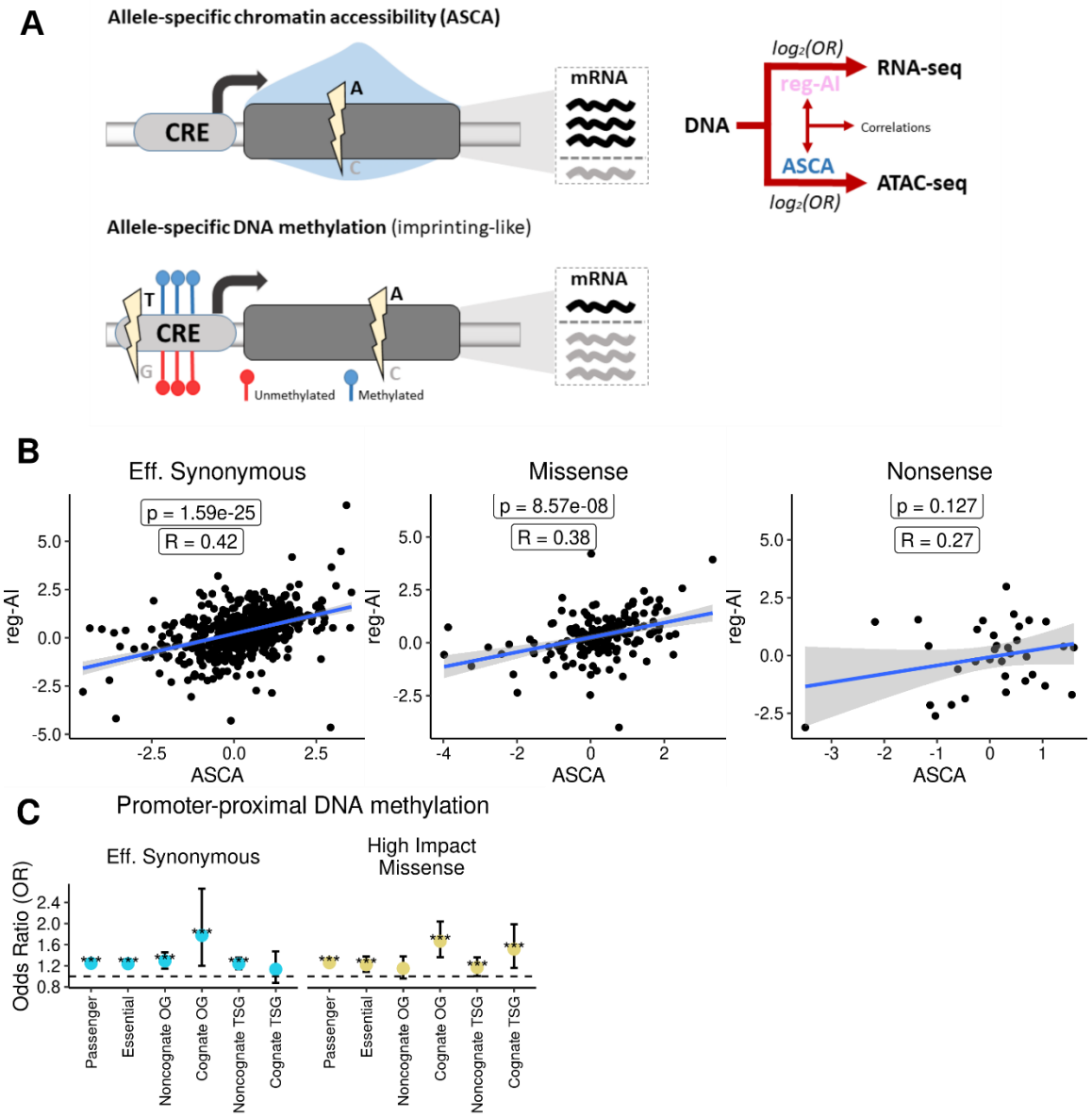


Figure 4. Allele-specific chromatin accessibility (ASCA) and promoter-proximal DNA methylation partially explains AI caused by non-CNA mechanisms.
A, Schematic representation of our analysis (top right), detailing how ASCA values are obtained as done with reg-AI, and the subsequent correlation of these values with matched somatic variants. The diagram includes a graphical illustration showing how ASCA (middle) or DNA methylation (bottom) might influence the generation of AI. **B**, Linear regressions of reg-AI 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 reg-AI 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.

Altered methylation is associated with positive reg-AI in oncogenes

To further evaluate the hypothesis that epigenetic influences during cancer evolution determine mRNA AI, 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 reg-AI 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 reg-AI 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 reg-AI occurring in a tumor, and the partial DNA methylation of the promoter and promoter-proximal region. This was seen for high impact 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 noncognate 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 reg-AI of approximately 0.12 (in $\log_2$ scale) compared to passenger genes (p = 8.8e-11), suggesting an oncogene-associated effect with potential to generate cancer driver AI. Notably, the ES variants (having lower impact on amino acid sequence) 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[43] 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 is likely that the WT allele is being selectively methylated; as a limitation of this test, DNA methylation is not phased with the mutation on which reg-AI is measured. This epigenetic modification could suppress the expression of the *wild-type* allele, thereby boosting the impact of the mutated allele, conferring a GoF effect in these oncogenes, analogous to CNA deletions which remove the *wild-type* allele in oncogenes[5].

Somatic mutations rarely directly activate chromatin states in driver genes

To further corroborate the above-mentioned observations that epigenetic changes underlie reg-AI, we tested for evidence that the occurrence of a coding variant itself could modify the local epigenetic environment towards a more active state, resulting in positive reg-AI. 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 in parallel with the partial methylation in or adjacent to the promoter region, as analyzed above. To test this hypothesis, we utilized the *Sei*[48] neural

network that predicts the chromatin states from local DNA sequence, to assess the impact of somatic SNVs in coding regions on *Sei* active transcription chromatin states (TN1-TN4, see Methods).

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

We observed a significant increase in positive reg-AI values within the TN activating category among cognate OGs mutations (Fig. 5A), both for missense (mean reg-AI TN activating = 0.61 vs mean reg-AI 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 chromatin states, and enhanced allele mRNA 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 AI-inducing epigenetic effect is selected in OGs. This may provide an explanation for a part of the selection on synonymous mutations seen in OGs[43] and the epigenetic effect may extend to some missense mutations as well. The association of TN activating mutations with AI 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 reg-AI, but it may be selected in OGs as the enhanced expression is more pronounced than other categories of genes.

Further, we investigated whether these chromatin state changes, evident in positive reg-AI, also increased overall gene expression. In cognate OGs, we correlated *Sei* TN1-4 states scores with overall gene expression and with reg-AI, 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 the gene expression and the reg-AI (Fig. 5B), meaning that in these genes (*EGFR*, *GNAS*, *RET*, *GNAQ*, *CCND1*, *CTNND2*, *IL6ST*, *RAC1*) there is a possibility that exonic mutations therein might alter gene expression.

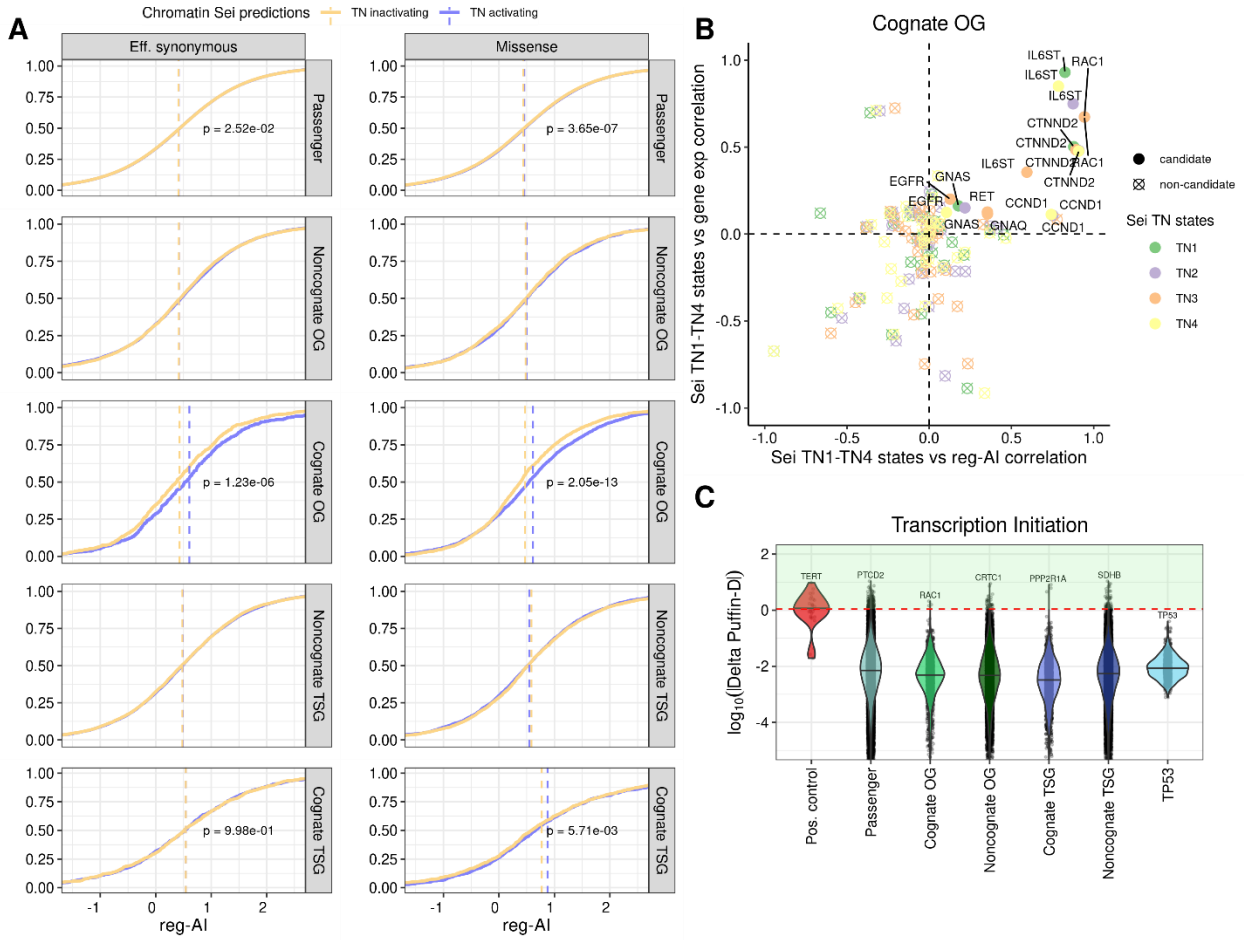


Figure 5. Activation of transcriptional chromatin states in cognate oncogenes.
A, Distribution of reg-AI for somatic mutations (ES in left panel, missense in right panel) in different gene categories, including passenger genes, noncognate and cognate OGs and TSGs. Mutations are classified into TN activating (based on *Sei* scores > 1e-04) indicating active transcription chromatin states, and TN inactivating (*Sei* scores < -1e-04). Vertical dashed lines represent mean reg-AI 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 reg-AI and *Sei* TN1-4 states (X-axis), alongside the correlation between gene expression and *Sei* TN1-4 states (Y-axis), for each cognate OG and *Sei* TN state separately (based on Pearson correlations). This panel identifies correlations where both reg-AI and gene expression are positively associated with *Sei* scores, suggesting transcriptional activation. OGs in the upper-right quadrant are highlighted (if both R > 0.1) as potential candidates experiencing selective pressure through gain-of-function effects.

As an additional test for direct effect of driver gene coding variants on transcriptional activity, we considered the Puffin-D neural net[49], which predicts activity of gene promoters, as GRO-Seq, PRO-Seq and CAGE scores, from DNA sequence in a 100kb-wide window. We considered the absolute change in Puffin-D scores (delta_Puffin-D) upon introducing the mutation in this DNA

sequence window (compared to the reference DNA sequence), integrating over the 100nt-long segment centered on the gene TSS (transcript model was via MANE Select, see Methods). As a positive control to ascertain relevant effect sizes in this test, we queried Puffin about the effects of the known non-coding) promoter somatic mutations in *TERT*[46,47], *CDC20*[50] and *FOXA1* genes[51]. We found that the delta_Puffin-D score of only 0.19-0.31% (for cancer genes) and 0.40-0.53% (for passenger genes) of the exonic variants tested was above median of the positive control variants (range across 3 scores from Puffin-D), and only 0.14-0.66% and 0.44-1.21% was above the 1st (lower) quartile of positive controls (Fig. 6C). Thus, perhaps expectedly, exonic mutations are predicted to very rarely have large effects on transcriptional output. An interesting exception are exonic mutations at the very 5' ends of genes (first 200 bp downstream of TSS), which can have large predicted effects (Supp. Fig. S15). Example exceptions are individual mutations in *ABI1*, *CD74*, *CNBP*, *CRTC1*, *ERCC4*, *ETV6*, *PPP2R1A*, *PPP6C*, *SDHB* and *SF3B1* driver genes with high predicted delta_Puffin-D scores (above Q4 of the positive control mutations). The median distance from these exonic, expression-altering mutations to the TSS was 60 bp. We highlight *PPP2R1A*, the only tumor suppressor gene with three distinct mutations in this test, located 60–74 bp downstream of the TSS (Supp. Fig. S16). These mutations have reg-AI values ranging from 0.82 to 1.69 (with the highest value in the cognate cancer type for this gene) indicating imbalanced expression, and delta_Puffin-D scores between 2.85 and 8.17 (Supp. Fig. S16).

Overall, the changes in epigenetic states associated with mRNA-level AI can, albeit rarely, derive directly from somatic variants driving the epigenetic change. Some somatic mutations in driver gene exonic regions may act additionally or primarily through the epigenetic change, modulating gene regulation. However, these rare examples do not explain the majority of mRNA AI of somatic mutations in driver genes that we observed.

Post-transcriptional cause of mRNA AI via splicing-altering mutations in cancer driver genes

Above, we suggested that the non-CNA sources of mRNA allelic imbalances observed in somatic mutations can result from epigenetic changes apparently unrelated with genetic mutations, or in rare cases from epigenetic changes that result from somatic mutations. Both of these types of events would regulate mRNA transcription, resulting in reg-AI. We additionally consider the possibility that this non-CNA, reg-AI might result from post-transcriptional regulation i.e. unrelated to epigenomics. A prime example are somatic mutations altering splicing, which can have profound impacts on gene expression in cancer, for instance by resulting in frameshifts that trigger NMD[19,42].

We hypothesized that splicing changes might be a common post-transcriptional mechanism underlying reg-AI. To test this, we used the *SpliceAI* neural network[52] 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, here termed "MaxDelta".

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

In a pan-cancer analysis (Fig. 6A), in each of the five *SpliceAI* splicing alteration categories (AG, AL, DG, DL, MaxDelta) a significant reg-AI difference between variants with high versus low *SpliceAI* scores was found (p < 1e-16, Mann-Whitney U test), where higher scores were associated with negative reg-AI values. This indicates that exonic mutations that alter splicing, considered as a set, tend to reduce the expression of the mutated allele; individual cases in the opposite direction are nonetheless possible.

As a prominent example, we examined three known splice-disrupting exonic driver variants in the *TP53* gene[43,53]. Notably, the 3′ terminal nucleotide of *TP53* exon 4 (codon 125), harboring recurrent synonymous mutations, had a mean DL delta score of 0.7, across the 3 possible somatic synonymous variants at this codon. These *TP53* mutations occur in a total of 32 tumor samples in our dataset, of which 10 exhibit a strong negative reg-AI (< -1), compared to only 1 with negative CNA-AI. Similarly, codon 224 of *TP53* exon 6, showed a DL of 0.9 in one synonymous variant (n = 3, of which 2 have negative reg-AI). Lastly, the codon 331 of *TP53* exon 9, exhibited a mean DL of 0.62 among two low-impact (at the amino acid level) missense variants (n = 5 occurrences, of which 2 have negative reg-AI), providing an example where a driver missense variant in *TP53* likely exerts its effect through splicing alteration, rather than through amino acid substitution.

Further analysis involved stratifying variants by gene categories (OGs, TSGs, essential, or passenger genes) and functional impact (missense or effectively synonymous). We tested for association between positive reg-AI mutations and variants predicted to impact splicing versus neutral variants (Fig. 6B and Supp. Fig. 17A, see Methods), and similarly so for negative reg-AI. Analyzing the same data differently, we also evaluated the differences in reg-AI values between high-impact and neutral variants by *SpliceAI* (Supp. Fig. 17B).

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

The notable global enrichment in negative reg-AI across splicing scores (Fig. 6B) was anticipated because many splicing motif disturbances result in out-of-frame splicing products, which usually premature stops and trigger NMD[42,54].

The most pronounced effects were seen for cognate TSGs, particularly in the low amino-acid impact ES variants for the *SpliceAI* MaxDelta (OR = 3.17, p = 9.4e-5, Fig. 6B-C). The differential

reg-AI values between the high-impact and neutral splicing variants showed higher effects for DG and, DL donor site alterations, while the two acceptor site delta scores, AL and AG have shown non-significant or contrary trends.

Synonymous variants can modulate splicing patterns in cancer genes[42,43,54]; here, we additionally asked if also the missense mutations can act similarly. We considered missense variants with low-impact at the protein level (in our ES variant set). Indeed, there was a splicing effect on AI resulting from these missense mutations (OR = 3.44, p = 4.8e-4 in the MaxDelta score), matching that of the synonymous mutations (OR = 2.83, p = 3.9e-2 for splicing MaxDelta) (Fig. 6B-C and Supp. Fig 17A).

We next asked if these splicing effects are selected during tumor evolution. Within cognate TSGs, 12.9% of low-protein impact missense mutations with negative reg-AI exhibited high *SpliceAI* scores (MaxDelta >0.15), in contrast to 6.9% in passenger genes, or 5.4% in TSGs in noncognate cancer types. This 6.0–7.5% difference quantifies the missense variants in TSGs that are under a selective pressure for splicing alterations, rather than or in addition to alterations to the amino acid sequence. For synonymous variants, the observed difference was 2.9-4.0%. Therefore, many missense somatic mutations drive cancer through splicing effects, as was proposed for synonymous coding mutations[43].

In addition to the ES variants group, the high protein impact missense variants also demonstrated some enrichment in negative reg-AI (MaxDelta score OR = 2.72; p = 2.2e-3, Fig 6B-C). For instance, in the *TP53* gene, the enrichment of negative reg-AI was seen in impactful missense variants (MaxDelta OR = 6.75, p = 7.6e-4) even more so than in ES variants (MaxDelta OR = 2.53, p = 4.5e-2) (Fig 7B-C and supp. Fig S17A-B). This is consistent with prior work on exonic variants that disrupt splicing of *TP53*; previously, synonymous variants were reported[42,43,54], and here we suggest missense variants can have substantial splicing effects.

We further asked which other TSGs exhibit links between splicing-altering variants and reg-AI at mRNA level (Fig. 6D). For ES variants associated with negative reg-AI, no individual gene reached statistical significance after FDR adjustment, however five TSGs (*ATM*, *NCOR1*, *TP53*, *SPEN*, and *POLG*) had suggestive signal (nominal $p < 0.1$) with ORs from 2.53 to 5.21 and for the impactful missense mutations, two TSGs were significant *TP53* (OR = 6.75, FDR = 2%), and the *PTEN* (OR = 4.27, FDR = 19%) (Fig. 6D). Consistently, there was a negative correlation between reg-AI values and *SpliceAI* MaxDelta scores for both genes (Supp. Fig S17D), observed in both for impactful missense (R = -0.18 to -0.11, p = 1.48e-6 - 3.7e-3) and ES variants (R = -0.32 to 0.08, p = 1.7e-4 - 0.56).

Curiously, some TSGs trended towards associating splicing variants with positive reg-AI: *BUB1B* (MaxDelta: OR = 3.29, p = 5.2e-2) and *KMT2D* (DG: OR = 3.40, p = 6.8e-2) with ES mutations, and *BCOR* (MaxDelta OR = 3.45, p = 6.8e-3) with high-protein impact missense mutations (Fig. 6D); these positive AI suggest dominant-negative acting TSG variants or isoforms. We also note that also the nonsense mutations in cancer driver genes can be splice-altering (Supp. Text S1

and Supp. Fig. S18). Overall, in addition to *TP53*, various other tumor suppressor genes exhibit selected splicing-disrupting mutations in their coding regions, which can manifest as mRNA-level AI.

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

Next, we asked if the association to *SpliceAI* scores was coupled to mRNA AI 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 reg-AI 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 reg-AI (Fig. 6B-C and Supp. Fig. S17C), particularly notable for AG and MaxDelta (0.68 vs 0.41, p = 2.3e-2, Supp. Fig. S17B), with consistent albeit non-significant trends in AL and DG categories. We compared the MaxDelta OR (2.76) to the baseline passenger OR (0.88), finding a marginally significant difference (p = 9.8e-2, Z-score, Fig. 6C).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[43]. In contrast to the effect of ES mutations, no enrichment for high impact missense mutations in AI was found (MaxDelta OR = 0.77, p = 0.22).

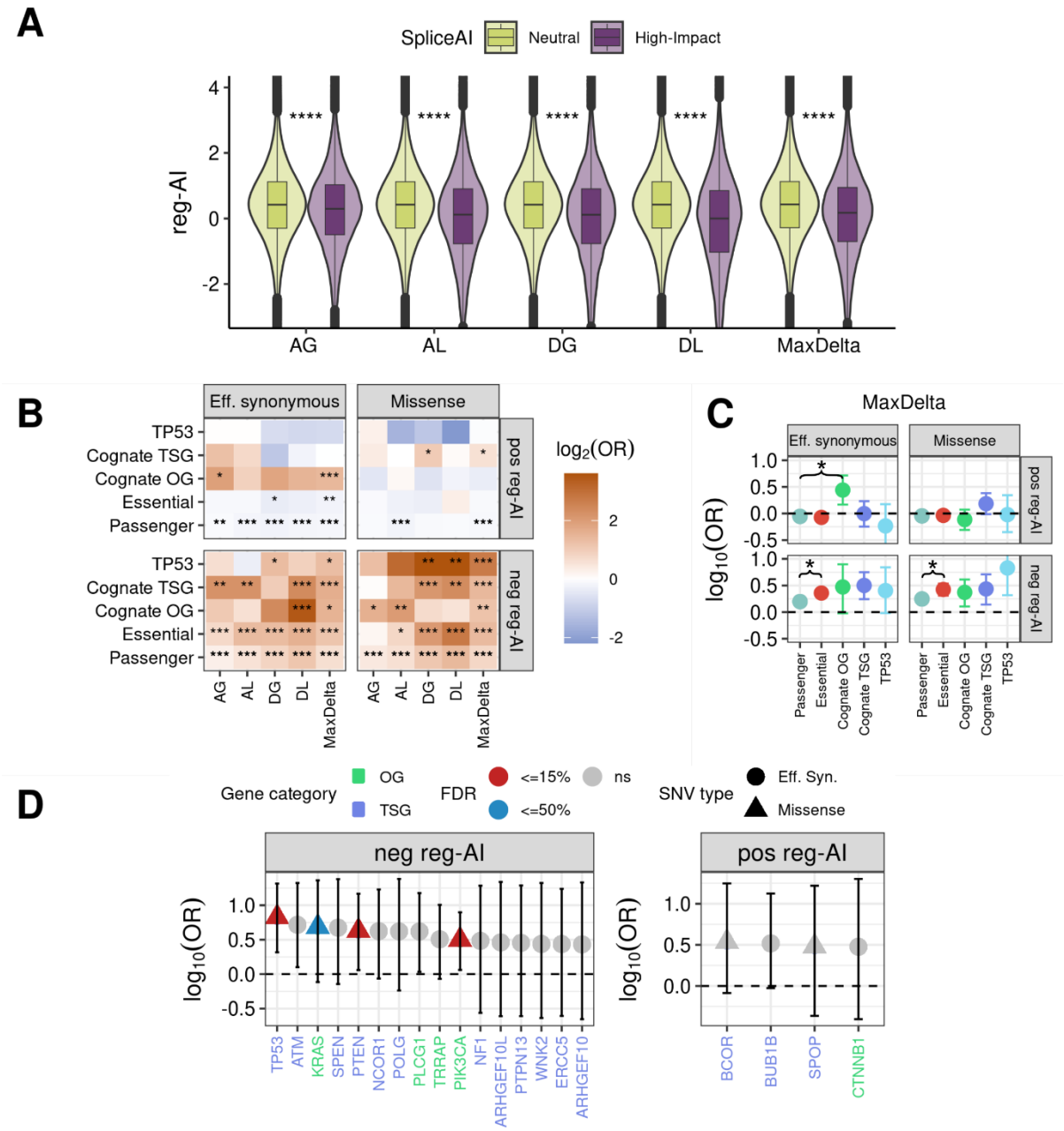


Figure 6. Association of splicing-altering variants with decreased expression of mutated alleles. **A,** Pan-cancer reg-AI values categorized by four *SpliceAI* delta scores: Acceptor Loss (AL), Acceptor Gain (AG), Donor Loss (DL), and Donor Gain (DG). The MaxDelta category represents the maximum value 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. *p*-values < 1e-16 from Mann-Whitney U test are denoted as ****. **B**, Enrichment analysis displaying significant associations between the presence of somatic variants exhibiting either positive or negative reg-AI 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 $\log_2$ scale, color-coded from low (blue) to high (red) values. FDRs are shown with *** indicating $p < 0.001$, ** $p < 0.05$ and * for $p < 0.15$. Positive enrichment indicates a higher proportion of positive or negative reg-AI mutations that are predicted to have a significant impact on splicing. **C**, ORs and CI at 95% from panel B, specifically for MaxDelta *SpliceAI* scores. Here, we additionally test for significant differences between each of the ORs compared to the "Passenger" gene ORs as a baseline. Nominal *p*-values < 0.1 from the Z-score are noted are *. **D**, Similar to panel A, an enrichment analysis was performed on individual cancer genes, separately for positive and negative reg-AI. The results are displayed as $\log_{10}$ OR (Y-axis) with its 95% CI, highlighting genes that were significant (FDR < 15%) or potential candidates (FDR < 50%) based on specific *SpliceAI* MaxDelta scores, provided there was a sufficient number of mutations. Enrichments are also split by ES and high-protein impact missense variants. Essential genes are excluded from this analysis. Figure shows the top 20 genes with highest OR.

Within ES mutations, the impact on positive reg-AI was more pronounced for the (low-protein impact) missense mutations (OR of 3.06, p = 6.2e-4 for MaxDelta), compared to the synonymous mutations (OR = 1.70, p = 0.48 for MaxDelta; the limited number of variants prevented robust estimation of ORs) (Supp. Fig 17A and C); links of missense mutations AI and splicing effects were strongly seen in both the AG and DL splicing alteration types (Supp. Fig S17A).

Next, in a gene-level analysis, we did not find oncogenes with a notably significant enrichment of positive reg-AI for ES or missense variants with high splicing scores (Fig. 6D); some individual examples of major oncogenes with tentative significance of splicing effects on AI are discussed in Supplementary Text S2. However, considered jointly, within cognate OGs, 9.8% of low-impact missense mutations with positive reg-AI exhibited high-impact splicing MaxDelta scores, in contrast to 3.9% observed in passenger genes or noncognate OGs. This 5.9% excess quantifies underscores a selective pressure in cognate OGs favoring splice-altering mutations, indicating that some missense mutations are selected for their impact on splicing. For synonymous variants in OGs, the observed difference was 2.5%, supporting that they too are under selection for splicing effects[43], that we found manifests as mRNA-level AI.

Our observation of positive reg-AI linked with ES mutations in OGs solidifies the notion that some changes without strong effects on protein may be causal to carcinogenesis, and that in OGs they have GoF effects[43]. The AI analysis also rules out confounding by trans-effects on splicing, suggesting that the effectively synonymous mutation itself is causal.

In summary, coding regions variants can lead to an allele-specific reduction or increase in mRNA levels, by disrupting splicing. This can be exploited by cancer evolution: in TSGs we generally observed a decreased expression of the mutated allele via reg-AI, resulting in a loss-of-function. Conversely, our data suggests OGs can leverage these splicing alterations to exert a gain-of-function effect, characterized by an increased expression of the mutated allele.

Discussion

In this study, we performed an in-depth analysis of the prevalence, mechanisms, and prognostic consequences of mRNA-level allelic imbalances of driver somatic mutations in cancer.

We estimated that the non-CNA causes of mRNA AI, here called reg-AI based on the suggested roles of epigenetic and genetic effects on transcription regulation, explained a very notable portion (55.6%) of the total pan-cancer variability in AI of somatic mutations, considerably higher than previous estimates[8,10]. We stringently accounted for imbalances at the DNA level and sample purity, and the analysis comprehensively included exome-wide somatic SNVs across all driver genes and across >8000 tumors, improving over previous smaller-scale studies[8,10,13,27,29–31]. We observed the highest prevalence of reg-AI-affected variants in lung tumors (both LUAD and LUSC) and also the related head-and-neck (HNSC), suggesting that these respiratory system-related tumors commonly harbor epigenetic changes that drive cancer by interacting with driver mutations. Somewhat expectedly, ovarian cancer was the most affected by AI generated through CNA, in accord with a very high prevalence of somatic CNA events in high-grade serous ovarian carcinomas[55,56].

Similar to the CNA interactions with driver changes[2,4,5,7], we demonstrated that the epigenetic AI is subject to somatic selection, with mRNA AI in favor of the mutant allele in some driver genes conferring a selective advantage by mimicking LOH events due to CNA. We identified specific genes displaying significant differences in 11 mRNA-AI and in a separate test 8 reg-AI genes, normalizing the impactful missense somatic mutations to a baseline of “effectively synonymous” (ES) somatic mutations. Methodologically, our AI-dNdES test considers both DNA and mRNA-level genetic variation but unlike various prior tests for allele-specific expression, does not necessitate RNA-Seq from tumor-matched healthy tissues from the same patient, and because of this it can be applied to considerably larger datasets.

Our test can identify cancer genes from AI of high impact somatic mutations in mRNA data; in addition to finding significant various TSG and prominently *TP53*, and individual examples of OGs, the test further implicated additional genes such as *ACSF3*, *CAD*, *C21orf91*, *CDC5L* and *TNKS2*. These differences in overall mRNA AI in selected genes resulted from an additive effect between CNA-AI and reg-AI, where in 8 out of 11 genes, reg-AI contributed to more than 50% of the difference, reinforcing the idea that non-CNA sources are pivotal in generating two-hit alterations that are selected for in cancer.

The mRNA AI favoring the mutant allele in impactful missense somatic driver gene mutations strongly predicts overall survival of cancer patients. We anticipate that there exist tens of additional driver genes where mRNA AI is relevant to cancer aggressiveness, which will be clarified with additional statistical power from additional cohorts with larger sample sizes. The overall pan-cancer effect on survival was substantial for mRNA-AI and for reg-AI (but interestingly not for CNA-AI), compared to other known correlates of patient survival such as intratumor heterogeneity.

Next, we addressed the questions of the mechanisms underlying the observed non-CNA AI of driver mutations at mRNA level. While our analysis of AI association with somatic variants does not explicitly address heritable effects, such as that of imprinting, and of germline eQTL variants generating mRNA AI, we infer these are likely not broadly relevant here, as imprinting does not affect many genes, and strong eQTLs do not affect many individuals[10,57]. Instead, we suggest that the epigenetic changes in the tumor that generate many reg-AI effects (including ones we identify to be under selection) are more alike to the "random monoallelic expression" effects[22–24]: somatic epimutations which are stable across cell divisions. These would generate a source of epigenetic heterogeneity in tumors that we show can form epistatic interactions with driver mutations, by analogy to CNA, increasing tumor fitness.

As a caveat of our analysis, we cannot fully exclude the effects of distal somatic mutations e.g. in enhancers located in intergenic regions or in introns, in generating the non-CNA mRNA-AI ("reg-AI") with driver potential. However, the somatic mutations in or near promoters are presumably more likely to be impactful than in distal elements, yet we found that even promoter-proximal mutations can explain at best only a small fraction of the non-CNA mRNA AI.

Similarly, it is possible albeit rare that the driver somatic mutation itself acts as a QTL and affects transcription of the target gene, e.g. by changing the local epigenomic environment and/or by promoting or blocking binding of a transcription factor. We used two modern machine learning predictors to suggest this is a rare occurrence: *Sei*[48] and Puffin-D tool[49], predicts there is at best only a modest association between a variant exhibiting AI, and variant effects transcription elongation histone marks (Sei), or transcription initiation signal at the promoter (Puffin-D). There may be individual examples where exonic mutations are predicted to have strong effects, particularly with 5'-most exonic mutations where Puffin predicts strong effects in some mutations; these exceptional cases might be followed up in future work.

Overall, we infer these mRNA AI changes are likely not downstream of somatic mutations, but rather results from independently occurring epigenetic changes. We observed examples thereof in allelic differences in chromatin accessibility (as ATAC-seq), and in the DNA methylation near the promoter that associated with reg-AI; both readouts reflect the altered activity of enhancers or promoters. Future studies using long-read sequencing of tumors will enable the somatic mutation to be phased with the epigenetic change in question[58], confirming whether the somatic driver mutation is indeed in *trans* with the inactivating promoter DNA methylation, and/or in *cis* with the activating, open chromatin state[59].

We would note that our models do not establish the temporal order between the mutation and the epigenetic change; regardless of their ordering, their co-occurrence in a cancer driver gene could be under selection. The epigenetic changes may be local epigenome alteration but may occur due to trans-acting factors; indeed global epigenome changes are not uncommon in tumors[34–36] and could provide the second-hit event for several somatic driver mutations at once, generating

drivers via mRNA expression imbalance by analogy to how arm-level/chromosome-level CNA are able to interact with various point mutations.

In addition to epigenetic alterations, as an additional result we identify that non-CNA AI of somatic mutations results from splicing-altering coding region mutations with driver potential. Altered splicing registers as differential mRNA abundance downstream because it changes coverage with RNA-Seq reads and/or triggers NMD in cases of frameshifts or inclusion of poison cassette exons.

In cognate tumor suppressor genes particularly *TP53* and *PTEN*, high-impact splicing mutations generally reduce the expression of the mutated allele, which was noticeable both in the ES mutations (synonymous and low-impact missense) but interestingly also in the high-impact missense mutations. 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: ES 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[43]. The association of positive mRNA-AI 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.

Overall, we find that impactful somatic mutations in driver genes often associates with allelic imbalances at the mRNA level. Positive selection favors the abundance of high impact mutations in mRNAs of at least 11 driver genes. These and dozens of additional genes may predict detrimental prognostic outcomes in affected patients, presumably by reflecting epigenome instability that contributes to intratumoral heterogeneity and generates two-hit driver events via epimutations. We encourage for mRNA AI of cancerous mutations to be regarded not only as a prognostic marker, but also a prospective therapeutic target in the context of personalized medicine, prioritizing patients for pharmacological intervention, similarly as was proposed for DNA-level AI resulting from somatic copy-number changes[4].

Acknowledgements

We are grateful to Marcell Veiner for running the neural networks *Sei* and *spliceAI* on our somatic AI 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)[28], 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 AI analyses. Tumor purity estimates were also downloaded accordingly with TCGA data[28]. 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[60].

For the non-coding mutations assessment, we downloaded whole-genome sequencing (WGS) files containing somatic variant calls for 777 tumor samples across 22 different cancer types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset (available at https://dcc.icgc.org/releases/PCAWG/Hartwig). These variant calls were processed using the Hartwig analytical pipeline v5 (https://github.com/hartwigmedical/pipeline5)[61].

Sequencing data alignment

RNA-seq and ATAC-seq sequencing reads were aligned to the human genome *GRCh38.d1.vd1* using *STAR* v2.5.3a[62]. Reads with a minimum mapping quality of 255 were retained and allele-mapping bias was corrected using *WASP*[63]. An unbiased removal of duplicate reads was performed using *WASP* in-house script[64], which discards one of the duplicate reads at random.

Sequencing data variant calling and annotations

Variant calling was performed using *Strelka2* v2.9.10[65]. 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[66] (2020-06-07 version) with GENCODE v26 (ENSEMBL v88) and adding *gnomad211_exome* database to obtain the population minor allele frequencies (MAF) from gnomAD[67].

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[32,33], 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)[37]. 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[38]. 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[39]. Variants from cancer genes were classified as "cognate" on those cancer types with an existing association in MutPanning at FDR <= 25%, and "noncognate" otherwise.

AI quantification from read counts

Allelic imbalances (AI) 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 $\log_2$ odds ratios (OR). This was in order to quantify the effect of AI, 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 reg-AI, CNA-AI and mRNA-AI were then quantified as follows:

$$\textbf{reg-AI} = log_2\left(\frac{\boldsymbol{MUT_{RNA}\ /\ WT_{RNA}}}{\boldsymbol{MUT_{DNA}\ /\ WT_{DNA}}}\right) \quad \text{(Eq. 2)}$$

$$\textbf{CNA-AI} = log_2\left(\frac{\boldsymbol{MUT_{DNA}\ /\ WT_{DNA}}}{\boldsymbol{E\ /\ (1-E)}}\right) \quad \text{(Eq. 3)}$$

$$\textbf{mRNA-AI} = log_2\left(\frac{\boldsymbol{MUT_{RNA}\ /\ WT_{RNA}}}{\boldsymbol{E\ /\ (1-E)}}\right) = \text{reg-AI} + \text{CNA-AI} \quad \text{(Eq. 4)}$$

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 AI whenever we observed more than a two-fold difference between allele counts (log_2(OR) > 1). Based on these criteria, somatic variants were categorized into three groups for the scope of our analysis:

- Positive AI: log_2(OR) > 1
- Negative AI: log_2(OR) < -1
- No AI: -1 < log_2(OR) < 1

AI variability estimates and comparison with prior studies

To quantify the overall contribution of reg-AI and CNA-AI to the overall mRNA imbalance (mRNA-AI), we used the formula from a recent study[8]:

$$\text{reg-AI explained variance} = \frac{\boldsymbol{var(reg_AI) + cov(reg_AI + CNA_AI)}}{\boldsymbol{var(mRNA_AI)}} \quad \text{(Eq. 5)}$$

$$\text{CNA-AI explained variance} = \frac{\boldsymbol{var(CNA_AI) + cov(reg_AI + CNA_AI)}}{\boldsymbol{var(mRNA_AI)}} \quad \text{(Eq. 6)}$$

To replicate the findings from the Correia *et. al.* study[8] on reg-AI contribution in BRCA patients and compare our findings to theirs, we adjusted our dataset using a similar filtering approach. We progressively applied more stringent filtering and then estimated our reg-AI and CNA-AI 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 reg-AI 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[10], 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-AI, (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[68], (iv) the sum of somatic coding SNVs per gene-sample, and (v) the sum of somatic non-coding intragenic 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-AI 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 AI to be between 15.7% (PCAWG) and 17.3% (us). Crucially, this type of modeling can estimate the relative contribution of non-CNA-AI (reg-AI) only due to the known variables that were quantified and included in the model, but ignores the non-CNA-AI (reg-AI) stemming from unknown sources. This underestimates the relative contribution of the non-CNA-AI, 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 AI 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 noisier (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-AI in this type of analysis that quantifies explainable variance. In contrast, in our analyses we quantify all systematic AI, regardless of the source, and by a subtraction of the CNA-attributable AI arrive at the 54.6% estimate of AI from non-CNA sources. We note, nevertheless, that these estimates could be affected by various sources of noise.

Correlating CNA burden with CNA-AI

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-AI to mRNA-AI. Additionally, a Spearman correlation between ranks of cancer types by CNA-AI 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-AI.

AI-dNdES test for selection

To assess selection acting on cancer genes upon AI, we implemented a conditional dN/dS-like test, termed AI-dNdES. This analysis was conducted to determine whether pathogenic missense (N) variants exhibited higher enrichment levels of positive AI ($\log_2$(OR) >1, compared to negative AI and no AI) 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 AI—mRNA, reg, and CNA. The AI-dNdES was applied at two levels: across gene categories and at the individual gene-level.

Gene category-level enrichment: The AI-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 AI-dNdES test, that is, assessing the enrichment of negative AI ($\log_2$(OR) < -1) compared to positive and no AI categories, across all merged gene categories including essential genes.

Individual gene-level enrichment: We performed AI-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 AI differences, that is, evaluating whether the $\log_2$(OR) AI 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 mRNA-AI, reg-AI, and CNA-AI, respectively.

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

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

Following these AI-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[69,70] and the *survminer* package[71] 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 AI separately: mRNA-AI (9 genes), reg-AI (4 genes), and CNA-AI (9 genes). The patients whose tumors exhibited AI in these genes were compared against those without significant AI in these genes. The AI variable was stratified into seven bins for more detailed analysis: negative AI (< -0.5); neutral/reference ([-0.5, 0.5]); and positive AI 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[44], 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[45] were also used as covariates in the pan-cancer analysis only to simplify the models. Results were plotted using the *forestmodel* package for R[69].

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.

Allele-specific chromatin accessibility (ASCA) analysis

Similarly, as done with AI, 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 “AI quantifications”, here using ATAC-seq counts instead of RNA-seq counts as in the AI analysis. Hence, ASCA can be considered as analogous to reg-AI 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 reg-AI 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[72,73]; (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[74]), 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 reg-AI 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[48], which uses deep neural networks to assess the chromatin-altering effects of genetic variants, within our AI 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*, focusing 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 > 1e-04) or deactivated ("TN inactivating", *Sei* scores < -1e-04) separately for the transcriptional states TN1-TN4.

Variant effect predictions on transcription initiation by Puffin-D

We estimated the transcriptional impact of exonic mutations using Puffin-D[49], a deep learning model designed to predict transcription initiation signals across a 100 kb DNA sequence. This model allows us to assess promoter activity based on experimental data sources, including GRO/PRO-Seq, FANTOM, and ENCODE CAGE scores.
For our analysis, we focused on 100 kb-wide windows centered around each mutation ([-50 kb, +50 kb]). Both the *wild-type* and mutated DNA sequences were used as inputs (one sequence per mutation). Puffin-D generates a score for each nucleotide within this sequence, allowing us to measure the transcriptional impact of mutations. We calculated the delta Puffin-D score (difference between *wild-type* and mutated sequences) per nucleotide and focused specifically on a ±50 bp window centered on the gene's main TSS; The gene's main TSS was determined using the MANE Select transcripts, with TSS positions obtained from the GENCODE v26 .gff file. Summing the absolute Puffin-D scores within this narrow window yielded a Puffin-D score per somatic mutation in our dataset (separate for GRO-Seq, PRO-Seq, and FANTOM CAGE), indicating the mutation's potential influence on transcription initiation.

To interpret the magnitude of Puffin-D scores, we used established positive controls. These controls included known promoter mutations in cancer genes with known effects on transcriptional activity: the oncogenes *TERT* and *FOXA1*, and the tumor suppressor gene *CDC20*. We created artificial sequences with these promoter mutations and compared their Puffin-D scores to those of the corresponding *wild-type* sequences, as stated above. After calculating the single Puffin-D score for each positive control mutation, we compared these values to the Puffin-D scores of our somatic mutations to gauge their potential impact on transcription initiation.

Variant effect predictions on splicing by SpliceAI and enrichment analysis

We evaluated the effect of splicing alterations on AI in our dataset (792,303 unique somatic mutations), utilizing *SpliceAI*[52]. 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 AI-dNdES test, to explore the enrichment of positive and negative reg-AI ($\log_2$(OR) >1 or < -1, respectively, both compared to no reg-AI) among high-impact splicing versus neutral mutations, we conducted an enrichment analysis separately for each SNV type: effectively synonymous (ES) and high-protein impact missense mutations (N). The ES group was further divided into low-protein 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 AI-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 reg-AI 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[41]: (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. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820–823 (1971).
2. Park, S., Supek, F. & Lehner, B. Higher order genetic interactions switch cancer genes from two-hit to one-hit drivers. *Nat Commun* **12**, 7051 (2021).
3. Esteller, M. *et al.* DNA methylation patterns in hereditary human cancers mimic sporadic tumorigenesis. *Hum Mol Genet* **10**, 3001–3007 (2001).
4. Bielski, C. M. *et al.* Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer. *Cancer Cell* **34**, 852-862.e4 (2018).
5. Besedina, E. & Supek, F. Copy number losses of oncogenes and gains of tumor suppressor genes generate common driver mutations. *Nat Commun* **15**, 6139 (2024).
6. Yavuz, B. R., Tsai, C.-J., Nussinov, R. & Tuncbag, N. Pan-cancer clinical impact of latent drivers from double mutations. *Commun Biol* **6**, 1–14 (2023).
7. Ciani, Y. *et al.* Allele-specific genomic data elucidate the role of somatic gain and copy-number neutral loss of heterozygosity in cancer. *Cell Syst* **13**, 183-193.e7 (2022).
8. Correia, L. *et al.* Allelic expression imbalance of PIK3CA mutations is frequent in breast cancer and prognostically significant. *npj Breast Cancer* **8**, 71 (2022).
9. Sen, A., Huo, Y., Elster, J., Zage, P. E. & McVicker, G. Allele-specific expression reveals genes with recurrent cis-regulatory alterations in high-risk neuroblastoma. *Genome Biology* **23**, 71 (2022).
10. Calabrese, C. *et al.* Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).
11. Cleary, S. & Seoighe, C. Perspectives on Allele-Specific Expression. *Annual Review of Biomedical Data Science* **4**, 101–122 (2021).
12. Mayba, O. *et al.* MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol* **15**, 405 (2014).
13. Tuch, B. B. *et al.* Tumor Transcriptome Sequencing Reveals Allelic Expression Imbalances Associated with Copy Number Alterations. *PLOS ONE* **5**, e9317 (2010).
14. Castel, S. E. *et al.* A vast resource of allelic expression data spanning human tissues. *Genome Biol* **21**, 234 (2020).
15. Lo, H. S. *et al.* Allelic Variation in Gene Expression Is Common in the Human Genome. *Genome Res.* **13**, 1855–1862 (2003).
16. Pastinen, T. & Hudson, T. J. Cis-Acting Regulatory Variation in the Human Genome. *Science* **306**, 647–650 (2004).
17. Robles-Espinoza, C. D., Mohammadi, P., Bonilla, X. & Gutierrez-Arcelus, M. Allele-specific expression: applications in cancer and technical considerations. *Current Opinion in Genetics & Development* **66**, 10–19 (2021).
18. 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).
19. 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).

20. 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).
21. Ferguson-Smith, A. C. Genomic imprinting: the emergence of an epigenetic paradigm. *Nat Rev Genet* **12**, 565–575 (2011).
22. Gregg, C. Known unknowns for allele-specific expression and genomic imprinting effects. *F1000Prime Reports* **6**, 75 (2014).
23. Nag, A. *et al.* Chromatin signature of widespread monoallelic expression. *eLife* **2**, e01256 (2013).
24. Kravitz, S. N. *et al.* Random allelic expression in the adult human body. *Cell Reports* **42**, 111945 (2023).
25. Atak, Z. K. *et al.* Interpretation of allele-specific chromatin accessibility using cell state–aware deep learning. *Genome Res* **31**, 1082–1096 (2021).
26. 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).
27. Przytycki, P. F. & Singh, M. Differential Allele-Specific Expression Uncovers Breast Cancer Genes Dysregulated by Cis Noncoding Mutations. *Cell Systems* **10**, 193-203.e4 (2020).
28. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120 (2013).
29. Liu, Z., Dong, X. & Li, Y. A Genome-Wide Study of Allele-Specific Expression in Colorectal Cancer. *Front. Genet.* **9**, 570 (2018).
30. Martínez-Ruiz, C. *et al.* Genomic–transcriptomic evolution in lung cancer and metastasis. *Nature* **616**, 543–552 (2023).
31. Batcha, A. M. N. *et al.* Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia. *Sci Rep* **9**, 11796 (2019).
32. Wu, Y. *et al.* Improved pathogenicity prediction for rare human missense variants. *The American Journal of Human Genetics* **108**, 1891–1906 (2021).
33. 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.
34. 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).
35. Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat Genet* **50**, 591–602 (2018).
36. Du, Q. *et al.* Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat Commun* **10**, 416 (2019).
37. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696–705 (2018).
38. Hart, T. *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 (Bethesda)* **7**, 2719–2727 (2017).
39. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat Genet* **52**, 208–218 (2020).

40. MacArthur, D. G. *et al.* A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **335**, 823–828 (2012).
41. 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).
42. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat Genet* **47**, 1242–1248 (2015).
43. 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).
44. van Dijk, E. *et al.* Chromosomal copy number heterogeneity predicts survival rates across cancers. *Nat Commun* **12**, 3188 (2021).
45. 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).
46. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
47. Lorbeer, F. K. & Hockemeyer, D. *TERT* promoter mutations and telomeres during tumorigenesis. *Current Opinion in Genetics & Development* **60**, 56–62 (2020).
48. 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).
49. Dudnyk, K., Cai, D., Shi, C., Xu, J. & Zhou, J. Sequence basis of transcription initiation in human genome. *Science (New York, N.Y.)* **384**, eadj0116 (2024).
50. Godoy, P. M. *et al.* Functional analysis of recurrent CDC20 promoter variants in human melanoma. *Commun Biol* **6**, 1–17 (2023).
51. Rheinbay, E. *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55 (2017).
52. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).
53. 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).
54. Shiraishi, Y. *et al.* Systematic identification of intron retention associated variants from massive publicly available transcriptome sequencing data. *Nat Commun* **13**, 5357 (2022).
55. Tao, Z. *et al.* The repertoire of copy number alteration signatures in human cancer. *Briefings in Bioinformatics* **24**, bbad053 (2023).
56. Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. *Nature* **606**, 984–991 (2022).
57. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
58. Sakamoto, Y. *et al.* Phasing analysis of lung cancer genomes using a long read sequencer. *Nat Commun* **13**, 3464 (2022).
59. Lee, I. *et al.* Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nature methods* **17**, 1191 (2020).
60. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44**, e71 (2016).

61. Martínez-Jiménez, F. *et al.* Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618**, 333–341 (2023).
62. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
63. 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).
64. WASP/mapping at master · bmvdgeijn/WASP. *GitHub* https://github.com/bmvdgeijn/WASP (2023).
65. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591–594 (2018).
66. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164 (2010).
67. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
68. Morison, I. M., Ramsay, J. P. & Spencer, H. G. A census of mammalian imprinting. *Trends in Genetics* **21**, 457–465 (2005).
69. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. (Springer, New York, 2000).
70. Therneau, T. M., until 2009), T. L. (original S.->R port and R. maintainer, Elizabeth, A. & Cynthia, C. survival: Survival Analysis. (2023).
71. Kassambara, A., Kosinski, M., Biecek, P. & Fabian, S. survminer: Drawing Survival Curves using ‘ggplot2’. (2021).
72. Chen, Y. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
73. 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).
74. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).