



## Confounding factors in assessing the enriched expression of somatic mutant allele in bulk tumor samples

Kohei Hagiwara, Andrew Thrasher, Nadezhda V Terekhanova, et al.

*Genome Res.* published online March 4, 2026

Access the most recent version at doi:[10.1101/gr.281003.125](https://doi.org/10.1101/gr.281003.125)

---

|                                 |                                                                                                                                                                                                                                                                                                    |
|---------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>P&lt;P</b>                   | Published online March 4, 2026 in advance of the print journal.                                                                                                                                                                                                                                    |
| <b>Accepted Manuscript</b>      | Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.                                                                                                                                               |
| <b>Open Access</b>              | Freely available online through the <i>Genome Research</i> Open Access option.                                                                                                                                                                                                                     |
| <b>Creative Commons License</b> | This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution 4.0 International license), as described at <a href="http://creativecommons.org/licenses/by/4.0/">http://creativecommons.org/licenses/by/4.0/</a> . |
| <b>Email Alerting Service</b>   | Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or <a href="#">click here</a> .                                                                                                                                          |

---

**Comprehensive** immune receptor profiling.  
Discover the **DriverMap™ AIR Assay** difference.



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

Published by Cold Spring Harbor Laboratory Press

**Confounding factors in assessing the enriched expression of somatic mutant allele in bulk tumor samples**

**Running title:**

**Confounders in somatic mutation ASE analysis**

Kohei Hagiwara, Andrew Thrasher, Nadezhda V. Terekhanova and Jinghui Zhang\*

Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA

\* To whom correspondence should be addressed. Tel: +1(901) 595-7069; Fax: +1(901) 595-7100; Email: [jinghui.zhang@stjude.org](mailto:jinghui.zhang@stjude.org)

## ABSTRACT

Allele-specific expression (ASE) of somatic mutations can be caused by *cis*-activation of the mutant allele or silencing of the wild-type allele and has been investigated by examining the enrichment of mutant allele in RNA relative to DNA. Here we show that this mutation-based approach can be confounded by gene expression differences in tumor and normal cells that co-exist in most bulk tumor samples. We model mutant allele expression by incorporating tumor/normal expression difference, mutant allele dosage, tumor purity, and nonsense-mediated decay (NMD) efficiency, projecting that such enrichments can occur without ASE. This confounding effect is exacerbated with low tumor purity and is dependent on mutant allele dosage for NMD-triggering mutations. The model predictions are validated by a pan-cancer bulk tumor analysis with somatic insertions/deletions (indels) from 9,101 The Cancer Genome Atlas (TCGA) samples. A single-cell analysis in five cutaneous squamous cell carcinomas demonstrates the robustness of this model to intratumor heterogeneity. As a byproduct of this confounding effect, we evaluate whether the inverse relationship between mutant allele enrichment in RNA and tumor purity could be leveraged to complement DNA-based somatic mutation detection in low purity samples. Indeed, our *de novo* somatic indel calling from TCGA RNA-seq increases the TCGA driver indel repertoire by ~ 14%, especially in samples with purity < 0.4, including actionable *EGFR* indels in lung adenocarcinoma and *FLT3* in acute myeloid leukemia. Our study not only reveals confounders in somatic mutant ASE analysis but also demonstrates their utility in RNA-based mutation calling.

## INTRODUCTION

In a diploid genome, allele specific expression (ASE) refers to a phenomenon of preferential expression of one allele over the other in a cell, which can unveil *cis*-acting regulatory mechanisms of attenuating (including silencing) or enhancing (including activation) the transcription of one specific allele [Ge et al. 2009]. A conventional approach to characterize ASE is to test if the variant allele fraction (VAF) in RNA ( $\text{VAF}^{\text{RNA}}$ ) is significantly deviated from its expected VAF in DNA ( $\text{VAF}^{\text{DNA}}$ ). In non-cancerous tissues, the analysis of allelic imbalance of germline heterozygous single-nucleotide polymorphisms (SNPs) in RNA, which measures a significant deviation from the expected  $\text{VAF}^{\text{DNA}}$  of 0.5, has led to the identification of genes or regions that exhibit ASE [Castel et al. 2015]. However, when applying the same technique to cancerous tissues [PCAWG Transcriptome Core Group 2020; Mayba et al. 2014; Robles-Espinoza et al. 2020], these imbalances were mostly (84.3%) attributed to somatically acquired copy-number alterations (CNA) in the tumor DNA [PCAWG Transcriptome Core Group 2020].

More recent studies have attempted to characterize ASE events independent of somatic CNA by focusing on ASE of somatic mutations [Palou-Márquez et al. 2024; Black et al. 2025; Liu et al. 2018; Bollas et al. 2025]. This approach tests the biased expression of the mutant allele by directly comparing a mutation's  $\text{VAF}^{\text{RNA}}$  or comparable metrics to that in the tumor DNA genome, where the influence of CNA is accounted for, and a significant deviation is attributed to ASE. Such interpretation is valid under the condition that gene expression level is comparable in tumor cells and “contaminating” normal cells that co-exist in bulk tumor samples. However, it is currently unclear how the difference in expression levels between tumor and normal cells affects the assessment of somatic mutant allele expression relative to its  $\text{VAF}^{\text{DNA}}$ . Tumor and normal tissues are known to exhibit differential gene expressions [Zhang et al. 1997]. For example, the

tumor suppressor gene *CDKN2A* was noted to have significantly higher expression in tumor tissues compared with the matching normal tissues acquired from cancer patients [Chen et al. 2021]. Furthermore, tumor cells can undergo hypertranscription, a global change in transcription activity [Zatzman et al. 2022; Lin et al. 2012; Bowry et al. 2021], which can also result in expression difference.

Here, we present a theoretical framework to investigate the impact of expression difference on the somatic VAF imbalance in bulk RNA-seq samples. Our model shows that somatic mutant alleles can be enriched in RNA relative to DNA due to expression differences in tumor and normal cells without ASE, and that this relative enrichment becomes more pronounced with lower tumor purity. While this effect can confound the ASE analysis of somatic mutations, we propose that it can be leveraged to sensitively detect mutations from RNA-seq in low tumor purity samples.

## RESULTS

### *Study design*

Tumor purity, somatic CNA, expression difference per gene copy in tumor relative to normal tissues, and nonsense-mediated decay (NMD) efficiency were used as parameters to model the ratio of  $\text{VAF}^{\text{RNA}}$  to  $\text{VAF}^{\text{DNA}}$ , referred as allelic expression variation (AEV) (*Modelling* in **Fig. 1**). Theoretical and simulation analyses of this model were performed to investigate whether unbalanced AEV (i.e.,  $\text{AEV} \neq 1$ ) could be affected by these parameters. By focusing on mutations with elevated allelic expression (i.e.,  $\text{AEV} > 1$ ), we validated this theoretical model using 9,101 The Cancer Genome Atlas (TCGA) [Cancer Genome Atlas Research Network 2013]

samples with whole-exome sequencing (WES) and RNA-seq datasets available on Cancer Genomics Cloud (CGC) [Lau et al. 2017] (*Validation* workflow in **Fig. 1**). In this analysis, we used publicly available somatic insertions/deletions (indels) identified in WES by the Genomic Data Commons (GDC) [Zhang et al. 2021] multi-caller pipeline. The use of indels enables assessing the effect of NMD in an unbiased manner, whereas premature termination codons (PTCs) by nonsense single-nucleotide variants (SNV) can be constrained to certain genomic contexts (e.g., TCA>TGA and hotspots). While the model parameters can confound the detection of true ASE, elevated AEV can be exploited for RNA-based somatic mutation detection to complement DNA-based approach. Using the TCGA RNA-seq data, we performed *de novo* indel detection because DNA-based indel detection is known to have a limited sensitivity of 40–50% [Alioto et al. 2015; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020] (*Application* workflow in **Fig. 1**). We also evaluated the effectiveness of this approach for mutations rated actionable by the COSMIC Actionability database [Sondka et al. 2024] to explore the potential for clinical impact.

### ***Modelling allelic expression variation (AEV) for somatic mutation***

We considered a somatic mutation with a total copy number of  $c$  represented by  $m$  and  $w$  copies of mutant and wild-type alleles per cell, respectively (i.e., copy number  $c = m + w$ ) and modeled its DNA VAF in a bulk sample ( $\text{VAF}^{\text{DNA}}$ , Eq.1 in METHODS). We also modeled the mutation's VAF in RNA where the gene expression per copy is  $e$  times different between tumor and normal, and the mutant allele is degraded at NMD efficiency  $d$  ( $\text{VAF}^{\text{RNA}}$ , Eq.2). The ratio of  $\text{VAF}^{\text{RNA}}$  to  $\text{VAF}^{\text{DNA}}$ , defined as AEV (Eq.3) (illustrative examples in **Fig. 2A** and **Supplemental Fig. S1**), is used to show enriched (elevated AEV) or depleted (reduced AEV) mutant allele

representation in RNA-seq. Our model projects that higher expression in tumor per gene copy ( $e > 1$ ) is required for elevated AEV (Eq. 4) and we simulated the model over tumor purity in copy number  $c = 2$  (diploid) at various NMD efficiencies (**Fig. 2B**). For mutations that do not trigger NMD (i.e.,  $d = 0$ ), no mutant allele expression variation is expected (i.e., AEV = 1) when the gene expression levels per copy are identical in tumor and normal cells (i.e.,  $e = 1$ ) in contrast to the elevated and reduced AEV expected for  $e > 1$  and  $e < 1$ , respectively. Further, this effect on AEV is more prominent with lower purity due to a different rate of  $\text{VAF}^{\text{RNA}}$  reduction compared to that of  $\text{VAF}^{\text{DNA}}$  (**Fig. 2C**). In absence of NMD, the mutant allele dosage has no effect on AEV ( $d = 0$  in **Fig. 2B** and Eq.5). However, with  $d > 0$ , enriched mutant allele dosage due to amplification of the mutant allele or deletion of the wild-type allele (loss-of-heterozygosity or LOH) results in higher AEV (e.g., mut./mut. vs. mut./wt. in the three panels with  $d > 0$  in **Fig. 2B**). While higher NMD efficiency decreases AEV, elevated AEV is still detectable if the efficiency is smaller than  $(e - 1)/e$  (e.g.,  $d = 2/3$  in **Fig. 2B** and Eq.6). Similar results are obtained in other copy numbers (**Supplemental Fig. S2**) and also for subclonal mutation/CNA events (Eq.8 in METHODS and **Supplemental Fig. S3**).

In summary, under the condition where the gene expression per copy is higher in tumor than in normal cells (i.e.,  $e > 1$ ), elevated AEV (i.e., AEV > 1) can be detected in absence of ASE and the pattern is more pronounced with reduced tumor purity. While NMD decreases AEV, it can be retained at a higher level with enriched mutant allele dosage.

### ***Elevated AEV in somatic truncation indels in TCGA***

To empirically validate the AEV model, we used 106,157 somatic coding indels on the autosomal chromosomes identified by NCI Genome Data Commons (GDC). These indels were

annotated for amino acid changes and the underlying genes were categorized as tumor suppressor genes (TSGs), oncogenes, or nondriver genes (METHODS). Truncation indels (frameshifts and in-frame insertions of stop codons) accounted for the vast majority of all somatic indels (92.93% or 98,649/106,157) and those in driver genes primarily disrupted TSGs (93.05%, 4,724 truncations in TSGs of 5,077 in driver genes). In TSGs, truncations are more frequently accompanied by LOH, conforming to the two-hit hypothesis of tumorigenesis [Knudson 1971]. Therefore, this indel class is well suited for testing the predicted effects of LOH and NMD efficiency.

First, we quantified indel VAFs in both WES and RNA-seq using mutant and wild-type read counts extracted by indelPost [Hagiwara et al. 2022], as this tool outperformed those used by GDC in the benchmarking of indel VAF with serially diluted tumor DNA samples and expressed indels in RNA-seq (**Supplemental Fig. S4, Tables S1 and S2**, and METHODS). When NMD occurs at a high efficiency, the transcript harboring the truncation allele may be degraded and thus undetectable, but the wild-type transcript can be present in the transcriptome. To capture this scenario, we required the indel locus to be covered with  $\geq 1$  RNA-seq reads regardless of mutation status, resulting in inclusion of 4,450 truncations in TSGs and 69,840 in nondriver genes (**Supplemental Table S3**). Compared to nondriver genes, elevated AEV was four times as prevalent in TSGs (**Fig. 3A** and *left* in **3B**, 7.69% vs. 1.76%, Fisher's exact  $p < 2.2 \times 10^{-16}$ ), where LOH was eleven times more frequent (*right* in **Fig. 3B**, 19.78% in TSGs vs. 1.84% in nondrivers, Fisher's exact  $p < 2.2 \times 10^{-16}$ ).

We then examined whether the elevated AEV cases matched the predicted pattern of model parameters: purity, copy number, NMD efficiency as well as relative expression difference between tumor and normal cells. In the assessment with bulk data, we surrogated the  $e$  as the



TPM ratio between tumor and tissue-matched normal controls adjusted for copy number alterations to evaluate differential expression (**Fig. 3C** and METHODS). Consistent with our model,  $e$  was higher for genes carrying elevated AEV indels (median: 1.76 in elevated vs. 0.72 in overall for TSGs; 1.60 vs. 0.92 for nondriver genes, Mann–Whitney  $U p < 2.2 \times 10^{-16}$  for both gene classes). Alternatively, the  $e$  variable can be  $> 1$  under a global upregulation of transcription activity (hypertranscription), which was recently reported to be widespread in human cancers [Zatzman *et al.* 2022]. Using the degree of hypertranscription in the TCGA cohort by Zatzman *et al.* (METHODS), we found a higher degree of transcription activity in samples harboring the elevated AEV indels (**Fig. 3D**, median: 2.17 in elevated vs. 1.77 in overall for TSGs at Mann–Whitney  $U p = 3.1 \times 10^{-6}$ ; 2.17 vs. 1.63 for nondriver genes at  $p < 2.2 \times 10^{-16}$ ). AEV is predicted to be elevated by the  $e$  parameter regardless of mutation clonality (Eq.8 in METHODS and **Supplemental Fig. S3**). Thus, we stratified indels by clonality (METHODS) and inferred 35.9% of the TCGA indels as subclonal. When clonal and subclonal indels were analyzed separately, their consistent AEV patterns confirmed the robustness of our model to subclonality (**Supplemental Fig. S5A**). Elevated AEV indels tended to occur in lower purity samples (**Fig. 3E**, median purity: 0.53 in elevated vs. 0.66 in overall for TSGs; 0.49 vs. 0.64 for nondriver genes, Mann–Whitney  $U p < 2.2 \times 10^{-16}$  for both gene classes), which matched the predicted trend in our model. This inverse relationship between elevated AEV and tumor purity was also robust to subclonal mutation (**Supplemental Fig. S5B**), consistent with the model’s projection (Eq.8 in METHODS and **Supplemental Fig. S3**). Tumor purity varies greatly by cancer type; for example, tumor samples of pancreatic adenocarcinoma (PAAD) tend to have low purity due to high stromal content (50–80%) [Karamitopoulou 2019], whereas those of brain cancers often have high purity [Aran *et al.* 2016]. Indeed, there was a negative correlation between purity and

elevated AEV prevalence across the 33 TCGA tumor types (**Fig. 3F**, correlation coefficient  $\rho = -0.67$ ,  $p = 1.9 \times 10^{-5}$ ) with the highest elevated AEV prevalence (9.38%) detected in PAAD, the cancer type with the lowest tumor purity (median purity of 0.4).

We further examined conditions relevant to truncation indels. Consistent with the model, premature termination codons (PTCs) caused by elevated AEV indels occurred more frequently in NMD evasive regions (**Fig. 3G**,  $\chi^2 p < 2.2 \times 10^{-16}$  for both tumor suppressor genes and nondriver genes). Due to the higher occurrence of LOH in TSGs (*right* in **Fig. 3B**), we used truncation indels in TSGs to examine whether higher allele dosage of NMD-sensitive mutations caused by LOH is accompanied by elevated AEV as projected by our simulation (**Fig. 2B** and **Supplemental Fig. S2**). We found that indeed, LOH was enriched in elevated-AEV indels with NMD efficiency  $d > 0$  (**Fig. 3H**, 31.87% vs. 21.19%, Fisher's exact  $p = 1.2 \times 10^{-3}$ ). By contrast, LOH was not enriched when the truncation indels created PTCs in the last exon, where the NMD efficiency is close to 0 ( $d \approx 0$ ) [Lindeboom et al. 2019] (**Fig. 3H**, 14.43% vs. 14.31%,  $p = 1.0$ ). Therefore, this result and the higher prevalence of elevated AEV in TSGs (*left* in **Fig. 3B**) were consistent with the projection of our model showing an association between LOH and AEV only for  $d > 0$  (**Fig. 2B** and **Supplemental Fig. S2**).

### ***Elevated AEV in somatic in-frame indels in TCGA***

In-frame indels accounted for 7.07% (7,508/106,157) of the TCGA somatic indels and are not expected to trigger NMD. To evaluate the AEV of in-frame indels in driver genes, we focused on oncogenes as in-frame indels accounted for a much higher proportion of somatic indels than in TSGs (41.51% [247/595] vs. 7.86% [403/5,127], Fisher's exact  $p < 2.2 \times 10^{-16}$ ), and included

them for analysis if the locus was covered with  $\geq 1$  RNA-seq reads (238 in oncogenes and 5,673 in nondrivers) (**Supplemental Table S4**). Elevated AEV was more frequent in in-frame indels in oncogenes than in nondriver genes (**Fig. 4A**, 20.59% vs. 6.47% , Fisher's exact  $p = 2.7 \times 10^{-12}$ ) and was significantly enriched in samples with low tumor purity (**Fig. 4B**, median: 0.43 in elevated vs. 0.63 in overall for oncogenes at Mann–Whitney  $U p = 8.2 \times 10^{-7}$ ). Differential expression between tumor and normal tissues was not a significant contributor to the elevated AEV in oncogenes (*left* in **Fig. 4C**, median: 0.88 in overall vs. 0.93 in elevated at Mann–Whitney  $U p = 0.51$ ), whereas a higher degree of hyperactivation was associated with the elevated pattern (*right* in **Fig. 4C**, median: 1.52 in overall vs. 2.48 in elevated at  $p = 2.4 \times 10^{-3}$ ). Elevated AEV cases included known gain-of-function in-frame indels such as *EGFR* deletions in exon 19 and insertions in exon 20 [Lynch et al. 2004], *ERBB2* duplications of Y772\_A775 [Perera et al. 2009] and *CTNNB1* deletions at S45 [Morin et al. 1997]. Considering this, the lack of differential expression may suggest the possibility of bona fide ASE in elevated AEV cases.

### ***De novo driver indel detection in RNA-seq***

The inverse relationship between elevated AEV and purity (**Fig. 2**) may be exploited to recover somatic mutations missed by DNA-seq from RNA, especially in low tumor purity samples. To test this, we performed a *de novo* indel detection in TCGA RNA-seq focusing on truncations in TSGs and in-frame indels in oncogenes. A total of 3,259 somatic driver indels were called from tumor RNA-seq with  $\geq 1$  DNA supporting reads in the matched WES data (METHODS). When compared to 4,971 driver indels (4,724 truncations in TSGs and 247 in-frame indels in oncogenes) identified from WES data, we found that 2,581 were detected by both WES and RNA-seq (*Shared* in **Fig. 5A**), and 679 were detected only by RNAIndel (*RNA-only*)

(**Supplemental Table S5**). The remaining 2,390 were detected only by WES (*DNA-only*) due to low indel allele expression (median allele count: 1) with 834 unexpressed .

The increased sensitivity by *de novo* RNA-seq analysis was not limited to regions with low WES coverage — *de novo* indel detection by WES plateaued at a coverage of 20–40× and 40–60× for TCGA samples with medium-to-high (0.4–1.0) and low (< 0.4) purity, respectively (**Supplemental Fig. S6**), both below the 68× median WES coverage in the TCGA cohort. While all *RNA-only* indels were validated with WES, matched WGS data were available for 476 out of 679 *RNA-only* indels, and 457 of these were also validated by WGS (**Supplemental Table S5**). For the 19 indels that lacked WGS support, 18 had a high probability (median: 0.37, range 0.10–1.00) that the absence of indel reads in WGS was due to insufficient WGS coverage (METHODS). The exception event, a hotspot *PIK3R1* in-frame indel in breast cancer (TCGA-LD-A74U), has sufficient WGS coverage and the absence of indel reads was likely due to intratumor heterogeneity as the WGS and WES samples were obtained from different vials. The indels in *DNA-only* had a significantly lower  $\text{VAF}^{\text{RNA}}$  than in  $\text{VAF}^{\text{DNA}}$  (median: 0.11 in RNA vs. 0.33 in DNA, Mann–Whitney  $U$   $p < 2.2 \times 10^{-16}$ ). By contrast, the opposite pattern was found in *RNA-only* indels with much higher  $\text{VAF}^{\text{RNA}}$  (0.36 in RNA vs. 0.26 in DNA,  $p = 1.8 \times 10^{-12}$ ) and the *Shared* indels had a slightly elevated  $\text{VAF}^{\text{RNA}}$  (median: 0.43 in RNA vs. 0.39 in DNA,  $p = 1.3 \times 10^{-11}$ ) (**Fig. 5B**). When evaluating elevated AEV, which only occurs when copy-number adjusted expressions in tumor are higher than in normal, *DNA-only* indels had only a minimum presence (1.08%) in contrast to the significantly high frequency detected in *Shared* (13.8 %), and *RNA-only* (16.7%,  $\chi^2$   $p < 2.2 \times 10^{-16}$ ) (**Fig. 5C**). Consistent with our model, *RNA-only* indels were enriched in samples with low tumor purity. For example, samples with purity < 0.4 accounted for

25.04% *RNA-only* indels in contrast to 14.69% in other categories. (**Fig. 5D**, Kruskal–Wallis  $p = 2.1 \times 10^{-8}$ ).

The *de novo* indels detected by RNA-seq increased truncations in TSGs by 12.74% ( $n = 602$ ) compared to those known by WES analysis (4,724). The top 10 most frequently mutated TSGs had a varying degree of contribution by RNA-seq (**Fig. 5E**). Although not amongst the top 10, *CDKN2A* (ranked 11<sup>th</sup>) is a notable example of having the biggest gain of somatic indels contributed by RNA-seq analysis, resulting in a 35.71% increase due to the 24.6-fold expression difference in tumor versus normal. By contrast, *RB1* (ranked 9<sup>th</sup>), which has lower expression in tumor (0.48-fold in tumor), only had a marginal gain (5.0%) by RNA-seq. The differential expression patterns of these two genes in tumor vs. normal TCGA bulk samples were also replicated in single-cell RNA-seq data across multiple cancer types (**Supplemental Fig. S7**). Notably, *de novo* RNA-seq indel analysis reaped the biggest gain in in-frame indels in oncogenes with an overall 31.17% increase (77 added to 247 in-frame indels detected by DNA) across multiple oncogenes (**Fig. 5F**). For example, RNA-seq indels nearly doubled driver indels in *FOXA1*, an oncogene for prostate cancer and breast cancer [Bailey et al. 2018], with 13 RNA indels added to the 15 detected by WES. Of all 28 *FOXA1* indels, the vast majority (71.4%) are located between D249 and E269, which are known to be associated with favorable prognosis in prostate cancer [Hwang et al. 2025]. In addition to prognostic biomarkers, 32.46% (25/77) of the additional oncogene in-frame indels were rated actionable by COSMIC Actionability database [Sondka et al. 2024] (**Fig. 5G**). These included targetable indels by approved drugs (Rank1 in COSMIC Actionability classification) such as *EGFR* exon 19 deletions in lung adenocarcinoma ( $n = 5$ ) and *FLT3* deletion at I836 in acute myeloid leukemia ( $n = 2$ ).

## Single-cell analysis of AEV with subclonal CNA

To validate the AEV model's robustness to subclonal CNA at single-cell level, we analyzed five cutaneous squamous cell carcinomas (cSCC) samples with publicly available scRNA-seq, WES and cell type annotation [Ji et al. 2020]. Specifically, we examined the relationship between AEV and copy-number adjusted expression difference between tumor and normal cells (the  $e$  parameter in Eq. 8 in METHODS), which is not directly observable in bulk tumor sample. Using scRNA-seq, we derived cell-level gene expression and copy-number estimates accounting for subclonal CNAs (**Fig. 6** and METHODS). For 223 somatic single- and 36 di-nucleotide substitutions (e.g., CC>TT) identified from the paired WES with sufficient coverage in scRNA-seq, AEV was determined by the ratio of  $\text{VAF}^{\text{RNA}}$  in a pseudo-bulk of scRNA-seq to  $\text{VAF}^{\text{DNA}}$  in WES. Consistent with the prediction robustness (**Supplemental Fig. S3**), higher AEV was correlated with higher  $e$  values in both clonal (correlation coefficient  $\rho = 0.43$ ,  $p = 3.7 \times 10^{-10}$ ) and subclonal ( $\rho = 0.46$ ,  $p = 1.7 \times 10^{-4}$ ) events.

## DISCUSSION

In this study, we evaluated the effect of tumor/normal gene expression difference on somatic allele enrichment in bulk RNA-seq. While a recent study was conceptually aware of such effects [Black et al. 2025], our development of the AEV model has enabled a systematic evaluation of multiple factors contributing to the elevation of AEV. We demonstrated that elevated AEV in a bulk tumor sample can be caused by higher expression in tumor in absence of allele-specific expression. Furthermore, this confounding effect is more pronounced in samples with low tumor purity, which, at first glance, appeared to be paradoxical (**Figs. 2A and B, 3E and F**). This

phenomenon is due to a slower decline of  $\text{VAF}^{\text{RNA}}$  compared to that of  $\text{VAF}^{\text{DNA}}$  with lower tumor purity when gene expression is higher in tumor than normal (**Fig. 2C**). It is noteworthy that the same paradoxical trend can also arise from true ASE, where the mutant allele is preferentially expressed in tumor, when examining the enrichment of mutant allele (**Supplemental Fig. S8**). We demonstrated that this effect could enhance mutation detection from RNA-seq in low tumor purity samples (**Fig. 5D**), while DNA-seq was more effective in detecting lowly expressed mutations or those prone to reduced AEV (**Fig. 5B and C**). In addition to the improved sensitivity by these complementary advantages, integrating DNA-seq and RNA-seq in mutation calling would also reduce false positives by validating putative calls across platforms as demonstrated by the high WGS validation rate for *RNA-only* indels.

Instead of the mutant allele itself, the enrichment of haplotype harboring the mutant allele can be examined using the heterozygous germline SNPs. In general, confounding factors such as tumor/normal expression differences and tumor purity have no or minimum effects on AEV of SNPs (**Supplemental Fig. S9**). Furthermore, in contrast to AEV of somatic mutation, AEV of SNPs with true ASE would diminish with low tumor purity in the case of true ASE (**Supplemental Fig. S8**). This SNP-based approach can also be valuable for finding elevated AEV due to true ASE. For example, when tumor cells are diploid, the  $\text{VAF}^{\text{RNA}}$  of heterozygous SNPs in the bulk tumor sample remains 0.5 if the allele expression is balanced (no ASE), regardless of purity and tumor/normal expression differences (**Supplemental Fig. S9A**). To illustrate the contrasting patterns, we show allelic imbalance of SNPs associated with *FAT1* V1689fs (AEV: 1.94) in a head-and-neck squamous cell carcinoma, which significantly deviated from 0.5 (simulation-based  $p < 2.2 \times 10^{-16}$  in **Supplemental Fig. S10**) (METHODS), whereas those with *MAP3K1* P324\_N325fs (AEV: 2.23) in a breast cancer did not change ( $p = 0.08$ ).

When tumor cells are not diploid, SNP  $\text{VAF}^{\text{RNA}}$  also deviates from the expected value defined by CNA with expression differences, but the scale of such deviations is smaller than that in somatic mutations (**Supplemental Fig. S9B and C**). Thus, SNP-based AEV better controls for the confounders.

A systematic analysis to identify elevated AEV caused by ASE in the current TCGA RNA-seq data is challenging, as the SNP-based method is only applicable to a limited number of cases having a sufficient number of heterozygous exonic SNPs in TCGA RNA-seq data which were prepared with poly(A) enrichment protocol. Theoretically,  $\text{VAF}^{\text{RNA}}$  is upper bounded by the purity-adjusted  $\text{VAF}^{\text{DNA}}$  (adj.  $\text{VAF}^{\text{DNA}}$ , Eq. 9 in METHODS) in the absence of ASE. Thus, indels with  $\text{VAF}^{\text{RNA}} > \text{adj. VAF}^{\text{DNA}}$  may be characterized as true ASE without needing SNPs. When this approach was applied to the TCGA indels (METHODS), the estimated ASE prevalence was found to be very low: 0.86% in TSG and 0.19% in nondriver truncating indels, and 2.28% in oncogene and 0.37% in nondriver in-frame indels, respectively. Additionally, only highly elevated AEVs were inferred as ASE (average AEV: 3.30 vs. 2.66 for all elevated AEV cases) and moderate cases were missed, including the true ASE event of *FAT1* V1689fs (AEV: 1.94), which was discovered by the SNP-based approach. This limited sensitivity may be because ASE is only tested at a single mutation locus, whereas the SNP-based approach is more statistically powered by testing allele counts at multiple SNP loci.

RNA-seq has been known to be useful for detecting low VAF mutations missed by DNA-based analysis [Xu 2018]. Despite this empirical knowledge, it is currently not a common practice to use RNA-seq for mutation calling purposes, possibly because there is no formal theory to demonstrate the advantage of RNA-seq over DNA-seq. Our model projects that RNA-seq can be a more sensitive platform than DNA under the condition that the gene expression per



copy is higher in tumor than in normal, particularly in samples with low tumor purity. Encouragingly, such a condition can be met in a wide range of genes, not restricted to those exhibiting differential expression, owing to a global transcriptional activation in cancers (hypertranscription) [Zatzman *et al.* 2022] (**Fig. 3C** and **C**). We anticipate that this theoretical basis will motivate future studies to incorporate RNA-seq for mutation detection, especially when reporting therapy-relevant mutations at high sensitivity. For example, in our *de novo* RNA-seq indel analysis, in-frame indels in oncogenes, which are often actionable, increased by >30% compared to the DNA-based repertoire provided by GDC. Truncation indels can also be important for designing cancer immunotherapy because frameshifts may create potent neoantigens containing amino acid sequences distinct from the self [Turajlic *et al.* 2017]. Among the frameshift indels, only those evading NMD are known to be relevant to therapy [Lindeboom *et al.* 2019; Litchfield *et al.* 2020] and it is encouraging that such variants were the most prevalent in indels detected exclusively from RNA-seq (41.5% in *DNA-only* vs. 49.2% in *Shared* vs. 63.1% in *RNA-only*,  $\chi^2 p < 2.2 \times 10^{-16}$ ).

Our current model focuses on elevated AEV as the direction of change is consistent with preferential expression of mutant alleles, although AEV can be elevated ( $\text{VAF}^{\text{RNA}} > \text{VAF}^{\text{DNA}}$ ) or reduced ( $\text{VAF}^{\text{RNA}} < \text{VAF}^{\text{DNA}}$ ). Reduced AEV of truncation indels was less prevalent in TSGs than in the nondrivers (9.15% vs. 12.76%, Fisher's exact  $p = 2.5 \times 10^{-13}$ , **Supplemental Fig. S11A**), and, as expected, such indels are mostly located in the NMD-sensitive regions in the both gene categories (**Supplemental Fig. S11B**). The difference can be attributed to higher LOH frequency in TSGs (*right* in **Fig. 3B**), which results in higher mutant allele dosage. Our model projects that AEV is higher with enriched mutant allele dosage in presence of NMD (**Fig. 2B**). Therefore, the lower prevalence of reduced AEV in TSGs further supports the opposing effect of

NMD and LOH proposed by our model. Consistent with the absence of NMD in in-frame indels, reduced AEV in these variants were rare in both oncogenes (2.79%) and nondrivers (4.08%) (Fisher's exact  $p = 0.41$ , **Supplemental Fig. S11C**).

In conclusion, we identified tumor/normal expression difference as a confounding factor contributing to enriched somatic mutant allele expression in bulk tumor samples. The confounding effect is particularly pronounced in low-purity tumor samples. These findings serve to caution the interpretation of allelic imbalance characterized by somatic mutation alone, which suggests the inclusion of germline heterozygous SNPs for detecting bona fide ASE. A comprehensive assessment of ASE of somatic mutation may need to incorporate non-coding *cis*-regulatory variants from WGS, DNA methylation and miRNA-seq data along with RNA-seq data generated from total RNA extraction to capture more heterozygous SNPs. Of note, this study yielded a valuable byproduct — the confounding effect can be exploited to efficiently complement DNA-based mutation detection in samples with low tumor purity by calling somatic mutation from RNA-seq.

## METHODS

### *TCGA dataset*

This study used the TCGA dataset available as GDC v36 on Cancer Genomics Cloud (CGC) (<https://www.cancer-genomics-cloud.org>). We included TCGA cases for which tumor/normal-paired WES and tumor RNA-seq BAM files and mutation annotation format (MAF) files ( $n = 9,101$ ) were all available. The 9,101 data pairs of tumor WES and RNA-seq were from all primary tumor samples, and tumor DNA and RNA were extracted from the same vial for 9,058

samples (99.5%). The remaining 43 cases using different vials for DNA/RNA extraction were all acute myeloid leukemia, a liquid tumor, which are not expected to have extensive spatial heterogeneity. The MAF files were prepared by GDC by applying MuTect2 [Cibulskis et al. 2013], VarScan2 [Koboldt et al. 2012], and Pindel [Ye et al. 2009] to the paired WES for somatic indel calls. This multi-caller pipeline does not require concordance across callers to generate the somatic indel data set [Zhang et al. 2021]. As a result, 61.0% of the somatic indels were detected by all three callers, 1.4% were detected by a single caller, with the remaining 37.6% by two callers. Gene expression data by STAR Count [Dobin et al. 2013] and copy-number data by ASCAT [Van Loo et al. 2010] were also available on CGC. The ABSOLUTE-based tumor purity data for TCGA samples was obtained from TCGA PanCanAtlas (<https://gdc.cancer.gov/about-data/publications/pancanatlas>). The platform to generate this purity data was SNP6, and tumor samples using the same vial for SNP6, WES, and RNA-seq accounted for 8,481 of 9,101 cases. To ensure analytical consistency, we excluded 620 cases with unmatched vials from any analysis that involved the tumor purity data, and their purity value was set to “-1” in **Supplemental Tables S3–5** to indicate missing values.

#### ***Derivation of VAF ratio between RNA and DNA***

In a bulk tumor sample consisting of  $x$  tumor and  $y$  normal cells, we denote  $q$  as:

$$q = y/x = (1 - \text{purity})/\text{purity}, \text{ where } \text{purity} = x/(x + y).$$

For an autosomal gene harboring a somatic mutation in tumor cells, we expect  $m$  and  $w$  copies of mutant and wild-type alleles per cell, respectively.  $m$  is a positive integer and  $w$  is a non-negative integer, where  $w = 0$  represents loss of the wild-type allele due to loss-of-

heterozygosity (LOH). The normal cells are diploid and homozygous for the reference allele. The DNA VAF in the bulk tumor sample is given by:

$$VAF^{DNA} = mx/(mx + wx + 2y) = m/(m + w + 2q). \quad 1$$

Consider that each gene copy in tumor cells is expressed at  $e$ -fold relative to normal, regardless of the mutation status of the allele (i.e., no ASE). The mutant allele triggers NMD with decay efficiency  $d \in [0, 1]$ , where  $d = 0$  represents a complete evasion from NMD and  $d = 1$  represents a complete decay of the transcript. The RNA VAF is:

$$\begin{aligned} VAF^{RNA} &= e \cdot mx \cdot (1 - d) / \{e \cdot mx \cdot (1 - d) + e \cdot wx + 2y\} \\ &= e \cdot m \cdot (1 - d) / \{e \cdot (m + w) - e \cdot m \cdot d + 2q\}. \end{aligned} \quad 2$$

The VAF ratio of RNA to DNA (i.e., allelic expression variation or AEV) is:

$$VAF^{RNA}/VAF^{DNA} = e \cdot (1 - d) \cdot (m + w + 2q) / \{e \cdot (m + w) - e \cdot m \cdot d + 2q\}. \quad 3$$

When the VAF ratio is greater than one, solving for  $e$ , we have:

$$e \cdot \{2q \cdot (1 - d) - w \cdot d\} > 2q.$$

Because  $e$  and  $q$  are positive, it follows that  $2q \cdot (1 - d) - w \cdot d > 0$ . As  $w$  and  $d$  are non-negative,  $e$  must stratify the following for elevated AEV:

$$e > 2q / (2q - 2q \cdot d - w \cdot d) > 1. \quad 4$$

In the absence of NMD (i.e.,  $d = 0$ ), Eq.3 becomes:

$$\begin{aligned} VAF^{RNA}/VAF^{DNA} &= e \cdot (m + w + 2q) / \{e \cdot (m + w) + 2q\} \\ &= e \cdot (c + 2q) / (e \cdot c + 2q), \text{ where } c = m + w. \end{aligned} \quad 5$$

Thus, the ratio depends on the copy number  $c$ , but not the genotype. For example, when the copy number in the tumor cells is three (i.e.,  $c = 3$ ), AEV is identical in all genotypes:  $m:w = 1:2$ ,  $2:1$ , or  $3:0$ . In the presence of NMD, solving Eq.4 for  $d$ , we have the following as the efficiency upper bound for elevated AEV:

$$d < 2q/(2q + w) \cdot (e - 1)/e \leq (e - 1)/e. \quad 6$$

CNA in bulk tumor samples may be subclonal. This can be modelled by introducing a new parameter  $i \in [1, 0]$ , which represents  $i \cdot x$  of tumor cells harboring the CNA event. Somatic mutations may also be subclonal and their occurrence with CNA may be independent or be associated as seen in the case of deactivating TSG mutations with LOH. To capture any scenarios, the mutation subclonality is separately parameterized for tumor cells with and without CNA by  $j \in [0, 1]$  and  $k \in [0, 1]$ , where the mutation is found in  $i \cdot j \cdot x$  tumor cells with CNA and  $(1 - i) \cdot k \cdot x$  those without. Under this formulation, the subclonal DNA VAF is:

$$VAF^{DNA} = \{i \cdot j \cdot m + (1 - i) \cdot k\} / \{(c - 2) \cdot i + (1 + q) \cdot 2\}. \quad 7$$

The subclonal RNA VAF is:

$$VAF^{RNA} = e \cdot \{i \cdot j \cdot m + (1 - i) \cdot k\} \cdot (1 - d) / \{e \cdot i \cdot (c - m \cdot d \cdot j) + e \cdot (1 - i) \cdot (2 - d \cdot k) + 2q\}.$$

AEV model for subclonal case is derived as:

$$e \cdot \{(c - 2) \cdot i + (1 + q) \cdot 2\} \cdot (1 - d) / \{e \cdot i \cdot (c - m \cdot d \cdot j) + e \cdot (1 - i) \cdot (2 - d \cdot k) + 2q\}.$$

8

The model was studied by simulation with a custom script (*aeV\_sim.py*).

### ***Indel allele count benchmarking***

The Sequencing Quality Control Phase 2 (SEQC2) Consortium established a reference somatic call set from the HCC1395 breast cancer cell line and the HCC1395BL cell line, which was derived from the normal B cell of the same donor [Fang et al. 2021]. We downloaded the reference indel call set ([https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic\\_Mutation\\_WG/release/v1.2.1/](https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/release/v1.2.1/)). The consortium also generated whole-genome sequencing (WGS) data on a series of diluted samples by mixing  $i$  parts of DNA material from HCC1395 and  $j$  parts from HCC1395BL for  $i:j = 1:0$  (undiluted),  $3:1$ ,  $1:1$ ,  $1:4$ ,  $1:9$ , and  $1:19$  [Zhao et al. 2021] ([https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic\\_Mutation\\_WG/data/SPP/](https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/SPP/)). The true VAF of somatic mutations is unknown. Therefore, we used the dilution ratio as ground truth value. Specifically, the following equation holds for the amount of mutated DNA material between a sample at  $n:m$  and the undiluted sample:

$$k \times VAF^{i:j} = \{k \times i / (i + j)\} \times VAF^{1:0}, \text{ where } k \text{ is the total DNA amount.}$$

With this relation, the dilution ratio ideally matches the observed VAF ratio:

$$VAF_x^{i:j} / VAF_x^{1:0} = i / (i + j), \text{ for tool } x.$$

Using the serial dilution WGS dataset, we compared the variant callers used in the GDC pipeline [Zhang et al. 2021] against indelPost (v0.2.3) [Hagiwara et al. 2022] for the dilution ratio estimation. For VarScan2 (v2.3.9) [Koboldt et al. 2012], we used the *pileup2indel* subcommand with *--min-coverage 1 --min-reads2 1* and *--min-var-freq 0.0001* to increase sensitivity. Pindel (v0.2.4) [Ye et al. 2009] was used with *-insert-size 400* as estimated [Zhao et al. 2021]. MuTect2 (v4.1.8.0) [Cibulskis et al. 2013] was used with the default settings. indelPost's *count\_alleles()* function was applied with the *by\_fragment* option, which merges

forward and reverse reads if they are from the same fragment. At 1:0, indelPost detected all the high confidence indels ( $n = 1,467$ ), while 1,445 were found by MuTect2, 1,196 by VarScan2, and 1,432 by Pindel. We used the unfiltered outputs from these callers (MuTect2, VarScan2, and Pindel) to maintain their sensitivity. Compared to these tools, indelPost consistently generated the best match to the expected ratios of 3/4, 1/2, 1/5, 1/10 and 1/20 over the dilution series (**Supplemental Fig. S4A** and **Supplemental Table S1**).

In RNA-seq, the VAF from the diluted samples generally does not match the dilution ratio due to gene expression differences between the tumor and normal cells. Therefore, we used an RNA-seq data for the Genome in a Bottle (GIAB) reference germline indels in the NA12878 lymphoblastoid cell line [Zook et al. 2014] (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA389940>) to evaluate the consistency of indel VAF in RNA-seq with their genotype in DNA. We first selected expressed exonic indels covered with  $\geq 1$  RNA-seq reads in the autosomal chromosomes. To avoid bias from NMD, we retained indels annotated as in-frame or in untranslated regions (UTR) if there were no co-occurring nonsense SNPs, frameshifts, or variants affecting splicing in the target gene. To avoid bias from ASE, we excluded genes identified to have ASE based on a prior ASE study which included NA12878 [Chen et al. 2016]. Consequently, the indels retained for this analysis are expected to exhibit balanced bi-allelic expressions with  $\text{VAF}^{\text{RNA}}$  values of 0.5 and 1.0 for heterozygous and homozygous genotype, respectively. The indel allele count tools were run with the same parameter settings as in the DNA benchmarking. The  $\text{VAF}^{\text{RNA}}$  estimates by indelPost were the closest to the expected heterozygous and homozygous genotypes of NA12878, which were determined by a pedigree analysis by the GIAB consortium (**Supplemental Fig. S4B** and **Supplemental Table S2**).

## 497 ***Indel annotation***

498 Indels were annotated for amino acid changes by VEP (v100) [McLaren et al. 2016]. Those  
 499 annotated as frameshift or in-frame insertion of a stop codon were collectively defined as  
 500 truncation indels, while the other in-frame insertions/deletions were defined as in-frame indels.  
 501 Bailey *et al.* [Bailey et al. 2018] classified genes as drivers if their mutation frequencies were  
 502 significant in the entire TCGA dataset (i.e. pan-cancer drivers) or in individual TCGA tumor  
 503 types (i.e. cancer-specific drivers). Using this resource, we annotated genes as drivers if the  
 504 model predicted the gene to be a pan-cancer driver, or a cancer-specific driver in the sample's  
 505 tumor type. Microsatellite instability (MSI) can induce mutations in driver genes regardless of  
 506 their functional consequences, leading to over-annotation of driver status. To mitigate this  
 507 confounding factor, we only used cancer-specific driver genes in MSI-enriched tumor types  
 508 (COAD, ESCA, STAD and UCEC). Indels were annotated as drivers if they occurred in driver  
 509 genes. Bailey *et al.* also provides tumor suppressor/oncogene classification in two confidence  
 510 levels: “*tsg*” or “*possible tsg*”, and “*oncogene*” or “*possible oncogene*”. In the current study,  
 511 these levels were combined to annotate driver indels either of tumor suppressor genes or  
 512 oncogenes.

## 513 ***Allelic expression variation (AEV) analysis***

514 Empirically, AEV is estimated from allele count, a summary statistic contributed by all the  
 515 parameters in the AEV equation (Eq.3 and 8). To ensure accurate VAF estimation in RNA-seq  
 516 and WES, read counts for somatic indels in the TCGA dataset (*DNA somatic indels* in **Fig. 1**)  
 517 were analyzed by a custom script run on CGC (*cross\_platform\_checker.py*), in which indelPost  
 518 [Hagiwara et al. 2022] was used for realignment. This process also harmonizes the inconsistency  
 519 in indel representations across the variant callers used by GDC or correct inaccurate allele



reports (**Supplemental Fig. S12**). The mutant and wild-type allele counts in tumor WES and RNA-seq were compared by Fisher's exact test and were considered significantly different if the  $p$ -value was  $< 0.01$  after Benjamini–Hochberg correction [Benjamini and Hochberg 1995]. The effect size threshold was set at 20%, which required  $VAF^{RNA}/VAF^{DNA}$  to be  $> 1.2$  for elevated and  $< 0.8$  for reduced AEV.

To test if the model is robust to subclonal mutations, we characterized the clonality of somatic indels based on the allele-specific copy number estimate by ASCAT (see *TCGA dataset*). For genomic intervals, the copy number was separately estimated as major and minor alleles, where  $major\_copy\_num \geq minor\_copy\_num$ . The TCGA indels were annotated with the ASCAT results based on the genomic interval overlap. Assuming that CNAs detected by ASCAT were primarily clonal, the subclonal  $VAF^{DNA}$  equation (Eq. 7) with the CNA clonality parameter  $i = 1$  becomes:

$$VAF^{DNA} \times (c + 2q) = j \cdot m, \text{ where } q = (1 - \text{purity})/\text{purity}.$$

Because the number of mutant allele copies (i.e.,  $m$ ) is  $\leq major\_copy\_num$ , the cancer cell fraction with mutation (i.e.,  $j$ ) has a lower-bound estimate:

$$VAF^{DNA} \times (c + 2q) / major\_copy\_num \leq j.$$

Clonal indels are expected to have  $j = 1$  and  $j < 1$  would be observed for subclonal. Thus, indels were defined as clonal if the one-sided 95% confidence interval contains/exceeds one, and subclonal otherwise.

Genomic intervals with  $minor\_copy\_num = 0$  were regions with loss-of-heterozygosity (LOH) because one of the diploid alleles was lost and indels in these regions were potentially

coupled with LOH. For such indels, we adjusted the raw  $VAF^{DNA}$  value with purity and copy number to ensure the loss of the wild-type allele:

$$adj.VAF^{DNA} = VAF^{DNA} \times (c + 2q)/c = m/c. \quad 9$$

We used those with the adjusted  $VAF^{DNA} > 0.9$  to analyze the impact of NMD on AEV.

### ***Nonsense-mediated decay analysis***

Two related studies by Lindeboom *et al.* [Lindeboom et al. 2016; Lindeboom et al. 2019] estimated NMD efficiency by comparing expression levels of various truncated transcripts to the wild-type transcript and proposed a NMD rule. By this rule, an mRNA transcript containing a premature termination codon (PTC) is classified as NMD evasive if the PTC position is < 150-nt downstream of the start codon (estimated NMD efficiency = 0.12), in an exon spanning > 407-nt (0.41), < 55-nt upstream of the last exon junction (0.20), or within the last exon ( $\sim 0$ ), otherwise the PTC is considered sensitive to NMD (0.65). In this measurement, an NMD efficiency of 0.2 would result in a 10% expression reduction in a sample with a heterozygous truncating mutation in a diploid region. In our model, the NMD efficiency parameter  $d$  also simulates this expression change: a 10% reduction of the number of transcripts per cell at  $d = 0.2$  with  $m = 1$  and  $w = 1$ . For each frameshift, we searched for the first downstream occurrence of a PTC along RefSeq isoforms [Pruitt et al. 2005] using a custom script (*premature\_stop\_codon\_locator.py*) and annotated the frameshifts for the NMD features based on the PTC position.

### ***Relative expression difference surrogation***

The tumor versus normal expression ratio per gene copy (variable  $e$  in the AEV model) was surrogated in bulk sample under two scenarios: differential gene expression and hypertranscription. For differential expressions, we first divided the gene's transcript per million

(TPM) in the tumor sample by its copy number as estimated by ASCAT. As autosomal chromosomes in normal cells are diploid, we halved the median TPM from tissue-matched normal expression data also taken from TCGA for technical consistency. We used the copy-number adjusted ratio of TPM in tumor to normal median as an estimate for differential expression. Cancer types with < 5 normal matching RNA-seq data were excluded in this analysis. Zatzman *et al.* [Zatzman et al. 2022] estimated the degree of hypertranscription at sample level as a ratio of total RNA amount in tumor divided by that in normal with ploidy adjusted and we used their estimates for the TCGA dataset.

### ***Single-cell data analysis***

A publicly available single-cell RNA-seq (scRNA-seq) dataset was downloaded for cutaneous squamous cell carcinoma (cSCC) samples along with their matched tumor/normal-paired WES and cell type annotation with tumor cells ( $n = 5$ ) (SRA accession: GSE144240) [Ji et al. 2020]. The scRNA-seq reads were mapped to GRCh38 by 10× Genomics' Cell Ranger (v8.0.1), which also generated a unique molecular identifier (UMI) count matrix with each gene in each cell. We used inferCNV (v1.18.1) to estimate gene copy number in each tumor cell using the cell type annotations as reported by the original study. In a tumor cell barcoded by  $t$ , the copy-number adjusted expression of gene  $X$  is estimated by dividing the count of gene  $X$ 's UMIs in the cell ( $UMI_t^X$ ) by the copy number of gene  $X$  in the cell ( $c_t^X$ ). In the original study, Ji *et al.* identified tumor basal, cycling, and differentiating keratinocytes and tumor-specific keratinocyte as cell types uniquely localized to the tumor lesions by a spatial transcriptomics analysis. We used the set of cell barcodes annotated for these cell types ( $T$ ) to represent the tumor cell population, whereas the normal cell population was represented by the remaining set of barcodes ( $N$ ).

Assuming that normal cells were diploid, the tumor/normal expression difference per copy in a given sample was estimated for gene X as:

$$\frac{|N|}{|T|} \times \frac{\sum_{t \in T} UMI_t^X / c_t^X}{\sum_{n \in N} UMI_n^X / 2}, \text{ where } | \cdot | \text{ denotes the size of barcode set.}$$

The fraction of tumor cells with CNA at gene X was calculated based on the cell-level copy number estimate at gene X :

$$\text{Frac. of tumor cells with CNA} = |\{t | c_t^X \neq 2 \text{ for } t \in T\}| / |T|.$$

We defined the CNA clonality by this fraction: clonal CNA if  $> 0.95$ , diploid if  $< 0.05$ , and subclonal CNA otherwise. After mapping the downloaded WES reads to GRCh38 by BWA (v 0.7.12-r1039), MuTect2 (v4.2.6.1) was applied to the paired WES data using the default settings. Somatic mutations on autosomal chromosomes, which passed *FilterMutectCalls* at  $\text{VAF}^{\text{DNA}} > 0.05$ , were annotated by VEP (v100) to retain those in gene regions. Allele counts were derived from the alignment pileup of scRNA-seq reads, where the coverage was required  $\geq 5$  with 2 or more mutant supporting reads, to estimate a pseudo-bulk  $\text{VAF}^{\text{RNA}}$  and to estimate AEV by dividing with the  $\text{VAF}^{\text{DNA}}$ .

To analyze *CDKN2A* and *RB1* expression, single-cell/single-nucleus RNA-seq (sc/snRNA-seq data) and cell type annotation were obtained from the previous study [Terekhanova et al. 2023]. The methods for data normalization, feature selection, dimensionality reduction and data merging were the same as used in that study. Average gene expression visualization in the selected cell groups was performed using DotPlot function from the Seurat package (v5.1.0) [Butler et al. 2018].

### ***Somatic indel calling from RNA-seq***

*De novo* indel calling was carried out by running RNAIndel [Hagiwara et al. 2020] on CGC. Computing cost was a key consideration on this platform, so we refactored RNAIndel by replacing the original realignment module with the much more efficient indelPost. This refactored version ran ~ 10 times faster (median run time: 21 vs. 208 minutes/sample) and achieved an ~ 85% reduction in computing cost (median cost: 0.16 vs. 0.97 USD/sample) (**Supplemental Table S6**). We deployed this code (v.3.3.0) to CGC for the full analysis, and the cost ranged from 0.04 to 14.8 dollars/sample (median: 0.21). On CGC, RNAIndel (v.3.3.0) analysis on tumor RNA-seq BAM files used the default parameters except for the parallelism option which was set to 12. In the output VCF files, putative indels were classified as somatic, germline, or artifactual based on the machine-learning model implemented in RNAIndel. Those classified as somatic were extracted from the VCF file and annotated for driver status (*Driver annotation*).

For each sample, all somatic driver indels by RNAIndel and those by the GDC pipeline were realigned on the corresponding paired WES data by indelPost imported in a custom script (*cross\_platform\_checker.py*) to harmonize the indel representation ambiguity across callers and/or platforms (**Supplemental Fig. S12**). Indels called by GDC were defined as *DNA-only* (**Fig. 5A**) if they did not match any of indels called from RNA after harmonization within the sample, and as *Shared* if they matched. Using the read count from the realignment process, we validated harmonized indels unique to RNA by testing for the presence of mutant reads in the tumor WES and their absence in normal, or for a significantly higher VAF in tumor by proportion test if mutant reads were present in normal. WES-validated RNA findings (*RNA-only*) were further scrutinized by tumor/normal-paired WGS dataset if available. We included the WGS data at the sample-match level to maximize the dataset size available for similar validation. For

indels not validated by WGS, we calculated the probability that a mutation present at the  $VAF^{DNA}$  estimated in the tumor WES was not captured in the tumor WGS coverage using the binomial model:  $(1 - VAF^{DNA})^{WGS\ coverage}$ .

To evaluate the sensitivity of *de novo* indel detection in the TCGA WES dataset at varying coverages and tumor purities, we used the somatic indel set comprised of *DNA-only*, *Shared*, and *RNA-only* indels (**Fig. 5A**). The somatic origin of *RNA-only* indels were validated as described above. Samples were stratified into three categories of low (0.1–0.4), medium (0.4–0.7) and high (0.7–1.0) tumor purity. The sensitivity of *de novo* indel detection in WES was defined as the proportion of *DNA-only* and *Shared* indels out of all indels detected in each of coverage bins: 0–20×, 20–40×, 40–60×, 60–80×, 80–100×, and > 100×.

### ***Actionability annotation***

We focused on in-frame indels in oncogenes using the COSMIC Actionability database (v14) [Sondka et al. 2024] by querying gene name and tumor type for the fields of GENE and DISEASE in the database. If a match was found, we examined if the mutation curated in the MUTATION\_REMARK field matched the input in-frame indel. If a curated mutation was not specified, we considered the in-frame indel a match. For the matched in-frame indels, we retrieved the actionability category curated in the ACTIONABILITY\_RANK field: Rank1 – approved drug demonstrated efficacy at the mutation; Rank2 – phase2/3 clinical results meeting primary outcomes; Rank3 – drug in ongoing clinical trials; Rank4 – case studies only.

### ***Assessment of true ASE events***

We adopted a simulation-based *p*-value computation developed for somatic silencing of the *BARD1* gene in neuroblastoma [Cupit-Link et al. 2024]. Briefly, using heterozygous single-

nucleotide polymorphisms (SNP) annotated as common by dbSNP [Sherry et al. 2001], we calculated the sum of deviation from the heterozygous VAF (i.e.,0.5) in RNA-seq:

$$\sum_{SNPs} |VAF^{RNA} - 0.5| = \text{observed total deviation}.$$

We simulated the total deviation by a binomial model with RNA-seq coverage at each SNP locus:

$$\sum_{SNPs} |binom(cov, 0.5)/cov - 0.5|, \text{ where } cov \text{ is RNA coverage}.$$

We counted the number of simulations deviating more than the observed value. The  $p$ -value was calculated by dividing the count by the total number of simulations ( $n = 1,000$ ).

Putative ASE events were also assessed without using SNPs. In the absence of ASE,  $VAF^{RNA}$  (Eq.2) is upper-bounded by the purity-adjusted  $VAF^{DNA}$  (Eq. 9):

$$\lim_{c \rightarrow \infty} VAF^{RNA} = m/\{m + w/(1 - d)\} \leq m/(m + w) = m/c, \text{ where } c = m + w.$$

Assuming that the mutant allele is from tumor, the count of wild-type allele from tumor is adjusted from the total wild-type allele count:

$$adj. wt \text{ allele cnt} = (total - total \cdot adj VAF^{DNA})/adj VAF^{DNA}.$$

In the actual computation, we capped adj.  $VAF^{DNA}$  at 1.0 and the adjusted allele count was floored to integer. Similar to the AEV analysis, the mutant and adjusted wild-type allele count in WES was compared to the raw RNA-seq allele count by Fisher's exact test with Benjamini-Hochberg correction. Indels with  $VAF^{RNA} > adj. VAF^{DNA}$  at  $FDR < 0.01$  were defined as ASE events.

## Statistical analysis

For count data, Fisher's exact and  $\chi^2$  tests were used for pairwise and three-group comparisons, respectively. To compare distributions, Mann–Whitney  $U$  and Kruskal–Wallis tests were used for pairwise and three-group comparisons. All tests were two-sided and were taken significant at  $p$ -value  $< 0.05$  unless otherwise specified.

## CODE AVAILABILITY

Custom codes developed for this study are available at Zenodo (<https://zenodo.org/records/17602849>) and as Supplemental Code. The refactored version of RNAIndel has been integrated into the original code repository as version 3 (<https://github.com/stjude/RNAIndel>) and the v3.3.0 code, which was used in the current study, is also included in Supplemental Code.

## COMPETING INTEREST STATMENT

None declared.

## ACKNOWLEDGEMENTS

This study was supported by a Cancer Center Support Grant (P30 CA21765) to St. Jude Children's Research Hospital, specifically through a Supplement (3P30CA021765-41S3) to J.Z. The study was also supported by the American Lebanese Syrian Associated Charities. We thank Mr. Michael Edmonson for his help with proofreading the manuscript.



## AUTHOR CONTRIBUTIONS

K.H. developed and validated the AEV model, performed the analyses, and refactored RNAIndel. A.T. contributed to the cloud-based computation. N.V.T. contributed to the single-cell analysis. J.Z supervised the study. K.H. and J.Z. wrote the manuscript. All authors approved the final version of the manuscript.

## REFERENCES

- Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, Heisler LE, Beck TA, Simpson JT, Tonon L, et al. 2015. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun*. 6:10001. doi: 10.1038/ncomms10001.
- Aran D, Sirota M, Butte AJ. 2016. Systematic pan-cancer analysis of tumour purity. *Nat Commun* 6:8971. doi: 10.1038/ncomms9971.
- Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* 174:1034-1035.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Ser B* 57: 289-300.
- Black JRM, Jones TP, Martínez-Ruiz C, Litovchenko M, Puttick C, Swanton C, McGranahan N. 2025. Cancer gene identification from RNA variant allelic frequencies using RVdriver. *Genome Biol*. 26:165. doi: 10.1186/s13059-025-03557-y.
- Bollas A, Gaither J, Schieffer KM, White P, Mardis ER. 2025. Variant calling from RNA-Seq data reveals allele-specific differential expression of pathogenic cancer variants. *Commun Med* 5:202. doi: 10.1038/s43856-025-00901-y.
- Bowry A, Kelly RDW, Petermann E. 2021. Hypertranscription and replication stress in cancer. *Trends Cancer* 7:863-877.
- Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36:411-420.
- Cancer Genome Atlas Research Network. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45:1113-20.
- Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. 2015. Tools and best practices for data processing in allelic expression analysis. *Genome Biol* 16:195. doi: 10.1186/s13059-015-0762-6.

- 717 Chen J, Rozowsky J, Galeev TR, HarmanCI A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan  
718 L, Gerstein M. 2016. A uniform survey of allele-specific binding and expression over 1000-  
719 Genomes-Project individuals. *Nat Commun* 7:11101. doi: 10.1038/ncomms11101.
- 720 Chen Z, Guo Y, Zhao D, Zou Q, Yu F, Zhang L, Xu L. 2021. Comprehensive Analysis Revealed  
721 that *CDKN2A* is a Biomarker for Immune Infiltrates in Multiple Cancers. *Front Cell Dev Biol*  
722 9:808208. doi: 10.3389/fcell.2021.808208.
- 723 Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson  
724 M, Lander ES, Getz G. 2013. Sensitive detection of somatic point mutations in impure and  
725 heterogeneous cancer samples. *Nat Biotechnol* 31:213-9.
- 726 Cupit-Link M, Hagiwara K, Nagy M, Koo SC, Orr BA, Ruppini E, Easton J, Zhang J, Federico  
727 SM. 2024. Response to PARP Inhibition in *BARD1*-Mutated Refractory Neuroblastoma. *N Engl*  
728 *J Med* 391:659-661.
- 729 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras  
730 TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15-21.
- 731 Fang LT, Zhu B, Zhao Y, Chen W, Yang Z, Kerrigan L, Langenbach K, de Mars M, Lu C, Idler K,  
732 et al. 2021. Establishing community reference samples, data and call sets for benchmarking  
733 cancer mutation detection using whole-genome sequencing. *Nat Biotechnol* 39:1151-1160.
- 734 Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KC,  
735 Gagné V, et al. 2009. Global patterns of cis variation in human cells revealed by high-density  
736 allelic expression analysis. *Nat Genet* 41:1216-22.
- 737 Hagiwara K, Edmonson MN, Wheeler DA, Zhang J. 2022. indelPost: harmonizing ambiguities  
738 in simple and complex indel alignments. *Bioinformatics* 38:549-551.
- 739 Hagiwara K, Ding L, Edmonson MN, Rice SV, Newman S, Easton J, Dai J, Meshinchi S, Ries  
740 RE, Rusch M, et al. 2020. RNAIndel: discovering somatic coding indels from tumor RNA-Seq  
741 data. *Bioinformatics* 36:1382-1390.
- 742 Hwang J, Likasitwatanakul P, Deshmukh SK, Wu S, Kwon JJ, Toye E, Moline D, Evans MG,  
743 Elliott A, Passow R, et al. 2025. Structurally Oriented Classification of FOXA1 Alterations  
744 Identifies Prostate Cancers with Opposing Clinical Outcomes and Distinct Molecular and  
745 Immunologic Subtypes. *Clin Cancer Res* 31:936-948.
- 746 The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. Pan-cancer  
747 analysis of whole genomes. *Nature* 578:82-93. doi: 10.1038/s41586-020-1969-6.
- 748 Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, Guo MG, George BM,  
749 Mollbrink A, Bergenstråhle J, et al. 2020. Multimodal Analysis of Composition and Spatial  
750 Architecture in Human Squamous Cell Carcinoma. *Cell* 182:497-514.
- 751 Karamitopoulou E. 2019. Tumour microenvironment of pancreatic cancer: immune landscape is  
752 dictated by molecular and histopathological features. *Br J Cancer* 121:5-14.

- Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. 1971. *Proc Natl Acad Sci U S A* 68:820-3. doi: 10.1073/pnas.68.4.820.
- Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 22:568-76.
- Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, Groves-Kirkby N, Mihajlovic A, DiGiovanna J, Srdic M, et al. 2017. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. *Cancer Res* 77:e3-e6. doi: 10.1158/0008-5472.CAN-17-0387.
- Lin CY, Lovén J, Rahl PB, Paranal RM, Burge CB, Bradner JE, Lee TI, Young RA. 2012. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* 151:56-67.
- Lindeboom RG, Supek F, Lehner B. 2016. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat Genet* 48:1112-8.
- Lindeboom RGH, Vermeulen M, Lehner B, Supek F. 2019. The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. *Nat Genet* 51:1645-1651.
- Litchfield K, Reading JL, Lim EL, Xu H, Liu P, Al-Bakir M, Wong YNS, Rowan A, Funt SA, Merghoub T, et al. 2020. Escape from nonsense-mediated decay associates with anti-tumor immunogenicity. *Nat Commun* 11(1):3800. doi: 10.1038/s41467-020-17526-5.
- Liu Z, Dong X, Li Y. 2018. A Genome-Wide Study of Allele-Specific Expression in Colorectal Cancer. *Front Genet* doi: 10.3389/fgene.2018.00570.
- Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, et al. 2004. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med* 350:2129-39.
- Mayba O, Gilbert HN, Liu J, Haverty PM, Jhunjhunwala S, Jiang Z, Watanabe C, Zhang Z. 2014. MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol* 15(8):405. doi: 10.1186/s13059-014-0405-3.
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* 17:122. doi: 10.1186/s13059-016-0974-4.
- Morin PJ, Sparks AB, Korinek V, Barker N, Clevers H, Vogelstein B, Kinzler KW. 1997. Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science* 275:1787-90.
- Palou-Márquez G, Pericot-Masdevall P, Supek F. 2024. Allele-specific expression is selected in tumorigenesis, results from epigenetic changes and has prognostic relevance. *bioRxiv* 2024.09.07.611780; doi: <https://doi.org/10.1101/2024.09.07.611780>.

- 789 PCAWG Transcriptome Core Group. 2020. Genomic basis for RNA alterations in cancer. *Nature*  
790 578:129-136.
- 791 Perera SA, Li D, Shimamura T, Raso MG, Ji H, Chen L, Borgman CL, Zaghlul S, Brandstetter  
792 KA, Kubo S, et al. 2009. HER2YVMA drives rapid development of adenosquamous lung tumors  
793 in mice that are sensitive to BIBW2992 and rapamycin combination therapy. *Proc Natl Acad Sci*  
794 *U S A* 106:474-9.
- 795 Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-  
796 redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*  
797 33(Database issue):D501-4. doi: 10.1093/nar/gki025.
- 798 Robles-Espinoza CD, Mohammadi P, Bonilla X, Gutierrez-Arcelus M. Allele-specific  
799 expression: applications in cancer and technical considerations. 2020. *Curr Opin Genet Dev*  
800 66:10-19.
- 801 Sondka Z, Dhiri NB, Carvalho-Silva D, Jupe S, Madhumita, McLaren K, Starkey M, Ward S,  
802 Wilding J, Ahmed M, et al. 2024. COSMIC: a curated database of somatic variants and clinical  
803 data for cancer. *Nucleic Acids Res* 52(D1):D1210-D1217. doi: 10.1093/nar/gkad986.
- 804 Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP:  
805 the NCBI database of genetic variation. *Nucleic Acids Res* 29:308-11.
- 806 Terekhanova NV, Karpova A, Liang WW, Strzalkowski A, Chen S, Li Y, Southard-Smith AN,  
807 Iglesia MD, Wendl MC, Jayasinghe RG, et al. 2023. Epigenetic regulation during cancer  
808 transitions across 11 tumour types. *Nature* 623:432-441.
- 809 Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, Wong YNS, Rowan A,  
810 Kanu N, Al Bakir M, et al. 2017. Insertion-and-deletion-derived tumour-specific neoantigens and  
811 the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol* 18:1009-1021.
- 812 Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P,  
813 Zetterberg A, Naume B, et al. 2010. Allele-specific copy number analysis of tumors. *Proc Natl*  
814 *Acad Sci U S A* 107:16910-5.
- 815 Xu C. 2018. A review of somatic single nucleotide variant calling algorithms for next-generation  
816 sequencing data. *Comput Struct Biotechnol J* 16:15-24.
- 817 Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to  
818 detect break points of large deletions and medium sized insertions from paired-end short reads.  
819 *Bioinformatics* 25:2865-71.
- 820 Zatzman M, Fuligni F, Ripsman R, Suwal T, Comitani F, Edward LM, Denroche R, Jang GH,  
821 Notta F, Gallinger S, et al. 2022. Widespread hypertranscription in aggressive human cancers. *Sci*  
822 *Adv* 8:eabn0238. doi: 10.1126/sciadv.abn0238.
- 823 Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler  
824 KW. 1997. Gene expression profiles in normal and cancer cells. *Science* 276(5316):1268-72.

825 Zhang Z, Hernandez K, Savage J, Li S, Miller D, Agrawal S, Ortuno F, Staudt LM, Heath A,  
826 Grossman RL. 2021. Uniform genomic data analysis in the NCI Genomic Data Commons. *Nat*  
827 *Commun* 12:1226. doi: 10.1038/s41467-021-21254-9.

828 Zhao Y, Fang LT, Shen TW, Choudhari S, Talsania K, Chen X, Shetty J, Kriga Y, Tran B, Zhu B,  
829 et al. 2021. Whole genome and exome sequencing reference datasets from a multi-center and  
830 cross-platform benchmark study. *Sci Data* 8:296. doi: 10.1038/s41597-021-01077-5.

831 Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. Integrating human  
832 sequence data sets provides a resource of benchmark SNP and indel genotype calls. 2014. *Nat*  
833 *Biotechnol* 32:246-51

834

## FIGURE LEGENDS

**Figure 1. Study design with three main workflows.** **Modelling:** purity, copy number, expression level per copy in tumor relative to normal, and nonsense-mediated decay efficiency were used to model the VAF ratio of RNA to DNA (i.e.,  $VAF^{RNA}/VAF^{DNA}$ ), which is termed allelic expression variation (AEV). Theoretical and simulation studies were conducted to identify conditions for  $AEV > 1$  (elevated AEV). **Validation:** DNA somatic indels identified by Genomic Data Commons's (GDC) multi-caller platform in the TCGA dataset (purple workflow) were used to validate the model for elevated AEV (red dots). Reads supporting indels were realigned in both WES and RNA-seq to accurately calculate AEV. **Application:** Confounding effects leading to elevated AEV can be exploited to enhance mutation calling from RNA-seq. This was tested for somatic driver indel detection by calling from RNA-seq with WES read support to complement the TCGA indel set (Venn diagram). Indels in each Venn section were examined for AEV and purity along with potential clinical implications. Major computational analyses for **Validation** and **Application** were performed on Cancer Genomics Cloud.

**Figure 2. Simulation on AEV in absence of ASE.** **A)** Illustration of elevated AEV under the scenario that gene expression in tumor is 3-fold of that in normal ( $e = 3$ ) with absence of NMD ( $d = 0$ ). In tumor cells, three copies of the mutant and wild-type alleles are transcribed while a single copy of the alleles is transcribed in normal cells. The  $VAF^{RNA}$ ,  $VAF^{DNA}$ , and the ratio of  $VAF^{RNA}$  to  $VAF^{DNA}$  (AEV) projected from tumor purity at 1.0, 0.5 and 0.33 are presented below. **B)** Simulation of the AEV model for heterozygous- (blue) and homozygous (orange) mutation in a diploid region with varying degrees of tumor purity and NMD efficiencies. **C)** Simulated  $VAF^{DNA}$  vs.  $VAF^{RNA}$  under three expression difference scenarios for purity ranging from 0.1–1.0. The simulation is performed for NMD-insensitive, heterozygous mutation.

**Figure 3. Evaluation of the AEV model using somatic truncation indels in TCGA samples.** **A)** Comparison of elevated AEV indels ( $AEV > 1.2$  with  $FDR < 0.01$ ) in tumor suppressor genes (TSGs) and nondriver genes. Scatter plots showing  $VAF^{DNA}$  (x-axis) and  $VAF^{RNA}$  (y-axis) for indels in TSGs and nondrivers with elevated AEVs highlighted in red dots. Nondriver indels were down-sampled to match the TSG indel size for visualization. **B)** The prevalence of elevated AEV in TSGs and nondrivers was compared by Fisher's exact test using the full dataset, which contained 4,450 TSG and 69,840 nondriver indels, respectively (*left*). The involvement of LOH at indel loci was also compared by Fisher's exact test (*right*). Evaluation of expression difference in tumor vs. normal on elevated AEV is shown in **C)** for differential expression and in **D)** for global upregulation in transcription activity (hypertranscription). For measuring differential expressions in **C)**, relative expression per copy ( $e$ ) was surrogated by dividing each tumor sample's TPM with the median TPM in tissue-matched normal samples adjusted for copy number. Box plots compare the distribution of the TMP ratio for samples harboring elevated AEV indels (*Elevated*) vs. all indels in the corresponding gene category (*Overall*) with  $p$  values by Mann–Whitney  $U$  test. The expression ratio plotted in **D)** was based on the RNA amount ratio estimated by Zatzman *et al.* **E)** and **F)** present the effect of tumor purity on elevated AEV. **E)** Comparison of tumor purity distribution in samples harboring elevated indels (*Elevated*) vs. all indel-containing samples (*Overall*) by Mann–Whitney  $U$  test. **F)** Correlation of elevated AEV prevalence (y-axis) in a tumor type with the median purity of the tumor type (x-axis) by linear regression. PAAD: Pancreatic adenocarcinoma; HNSC: Head and neck squamous cell



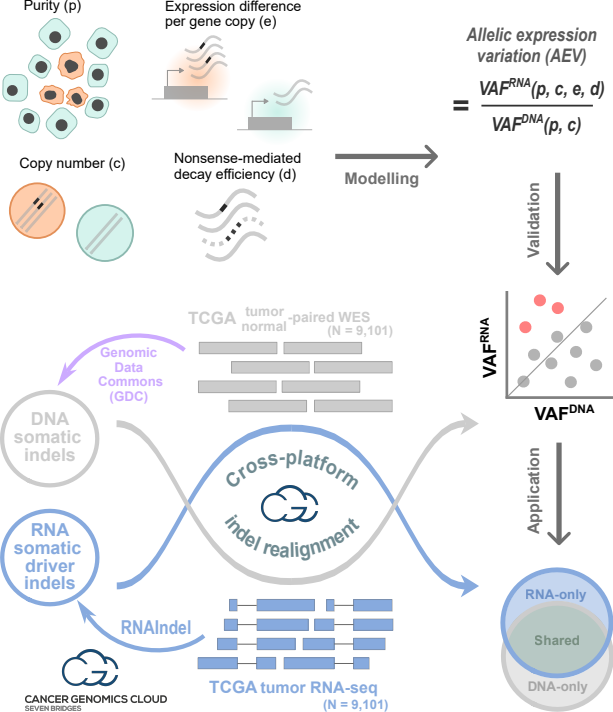
carcinoma; LUAD: Lung adenocarcinoma; LAML: Acute myeloid leukemia; GBM: Glioblastoma multiforme; LGG: Lower grade glioma; ACC: Adrenocortical carcinoma; UVM: Uveal Melanoma. The effect of NMD on elevated AEV is shown in **G**) for comparison of NMD features by  $\chi^2$  test and in **H**) for comparison of the prevalence of LOH by Fisher's exact test stratified by NMD efficiency ( $d > 0$  vs.  $d \approx 0$ ).

**Figure 4. Evaluation of the AEV model using somatic in-frame indels in TCGA samples.**

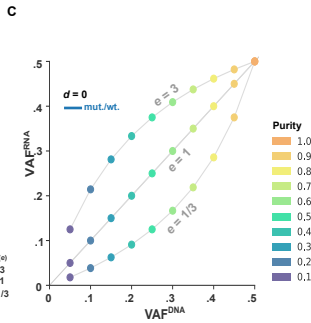
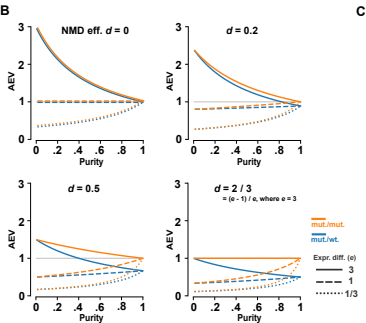
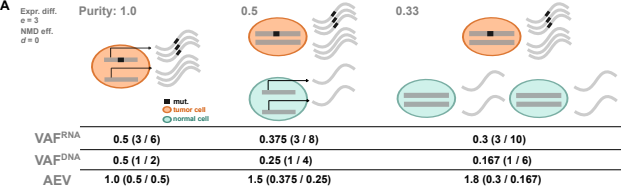
Elevated AEV prevalence, comparison of tumor purity and tumor expression ratio in elevated AEV vs. all indels are presented in **A**), **B**) and **C**), respectively, using the same style and statistical test as the corresponding analyses in **Fig. 3A**) Comparison of elevated AEV indels in oncogene and nondriver genes using down-sampled nondriver indels for a scatter plot (*left*) and the full data set (238 indels in oncogenes, 5,673 indels in nondrivers) for a bar plot (*right*). **B**) Comparison of tumor purity distribution in samples harboring elevated AEV indels (*Elevated*) vs. all indel-containing samples (*Overall*). **C**) Comparison of TPM ratio based on differential expression (*left*) and hypertranscription (*right*) between *Elevated* and *Overall* indels.

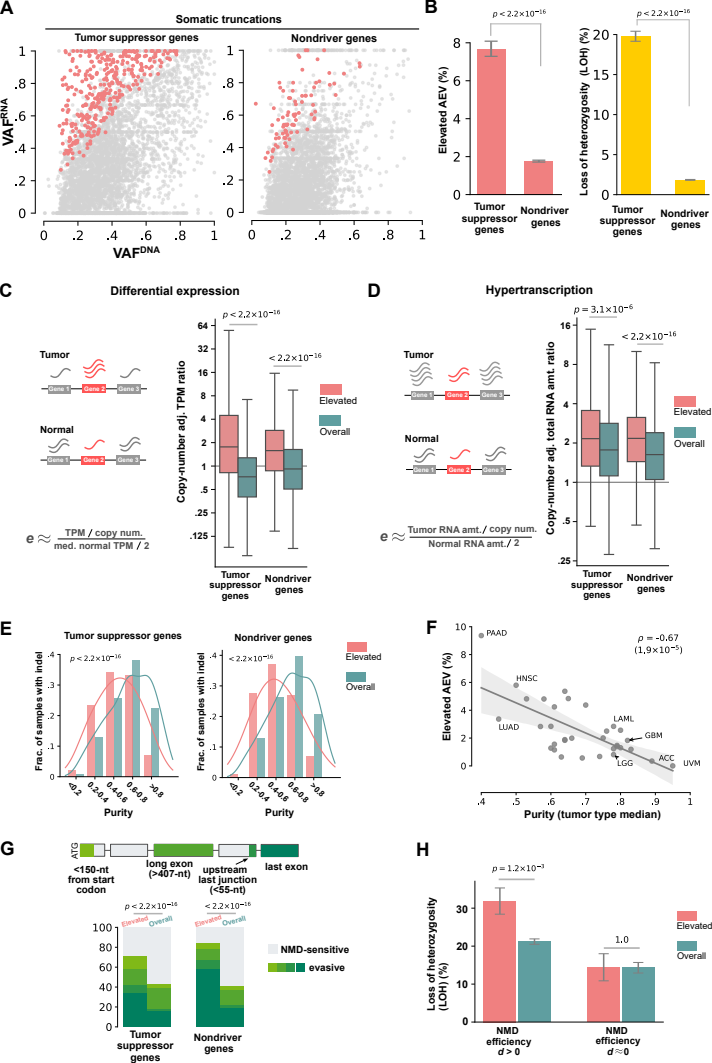
**Figure 5. Driver indels detected by RNA-seq and WES in TCGA samples.** Driver indels are defined as in-frame mutations in oncogene or truncations in TSGs. **A**) Venn diagram showing driver indels detected exclusively by GDC's WES analysis (*DNA-only*), by our *de novo* RNA-seq analysis (*RNA-only*), or by both approaches (*Shared*). **B**) Comparison of VAF distributions in WES (*DNA*) and RNA-seq (*RNA*) by Mann–Whitney *U* *p* value for driver indels in these three categories. Median VAF is shown by the red line. **C**) Comparison of prevalence of elevated AEV indels in these three categories by  $\chi^2$  test. **D**) Tumor purity distribution for samples containing indels in the three categories. Kruskal–Wallis test was performed to assess their differences. **E**) Top 10 most frequently mutated TSGs contributed by varying indel discovery approaches. *CDKN2A* (ranked top 11<sup>th</sup>) is also shown as an example for leveraging AEV for discovery due to its > 20-fold higher expression in tumor (**Supplemental Fig. S7**). **F**) Top 10 most frequently mutated oncogenes. \*: *TP53* was annotated as “possible oncogene” in glioblastoma multiform (GBM) and low-grade glioma (LGG) by Bailey *et al.* **G**) Actionability of oncogenic in-frame indels based on COSMIC Actionability database.

**Figure 6. Single-cell model evaluation.** Five cutaneous squamous cell carcinoma samples with scRNA-seq and WES were reanalyzed based on the cell-type annotation by Ji *et al.* In each sample, unique molecular identifier (UMI) counts per gene copy were averaged over cell populations annotated as tumor or normal. The ratio of the tumor average to normal was used as the copy-number adjusted expression difference between tumor and normal tissues (the *e* parameter). The clonality of CNA events was defined by a fraction of tumor cells carrying the alteration. For substitutions identified from WES, AEV and *e* values were plotted separately for 198 cases where the background ploidy is clonal and 61 subclonal counterparts (median CNA cell fraction: 0.31).

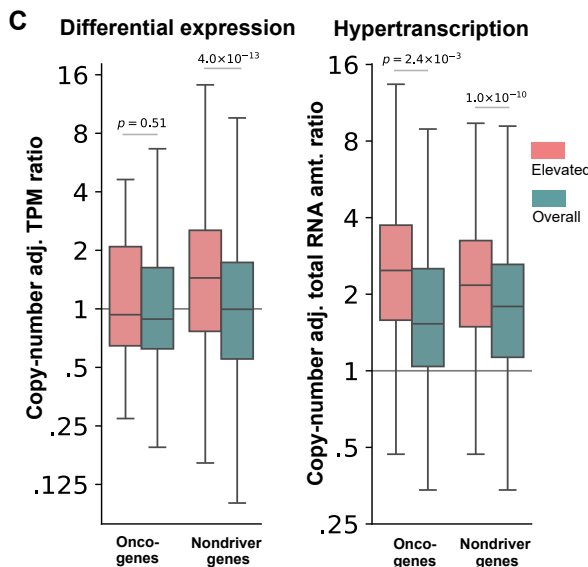
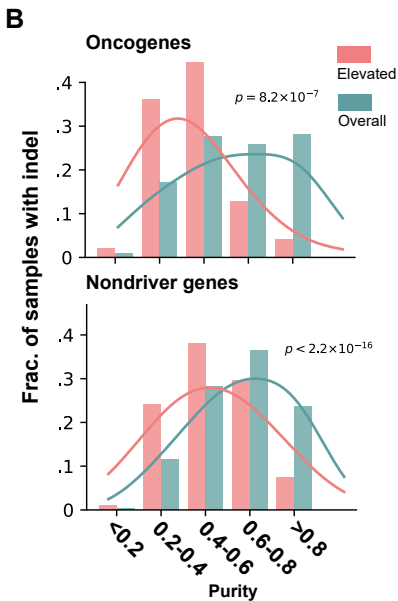
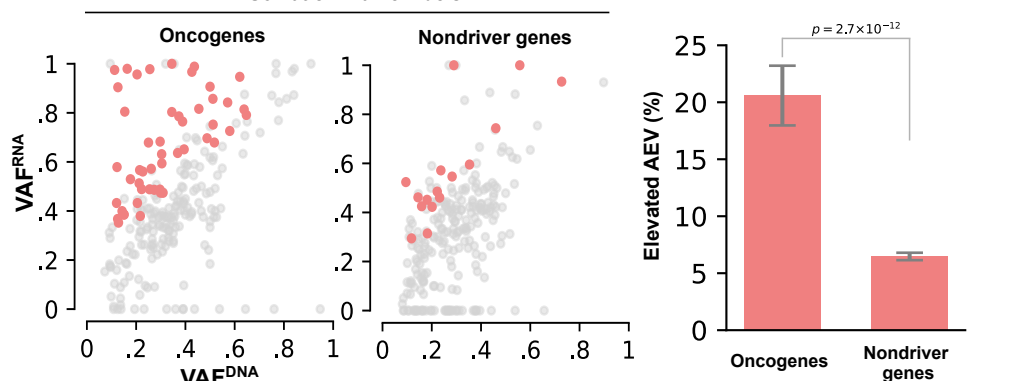


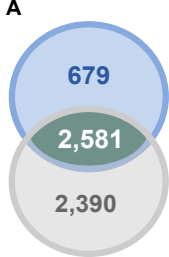




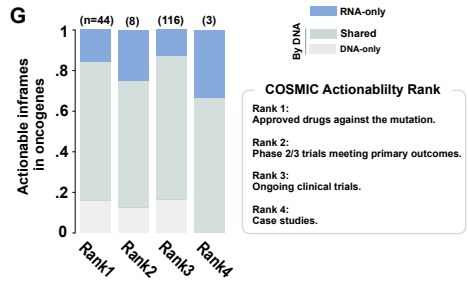
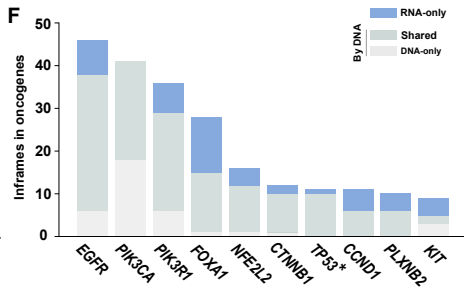
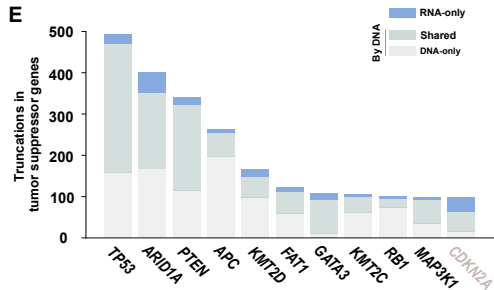
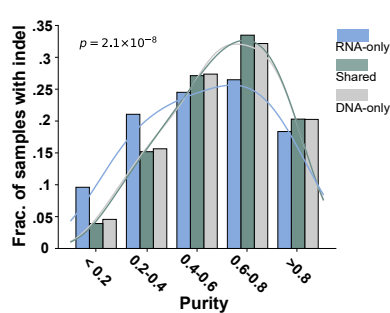
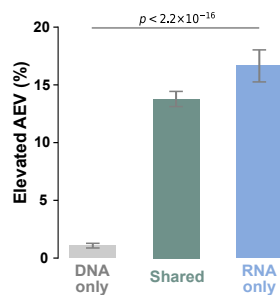
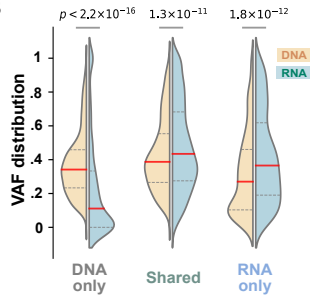


# Somatic inframe indels





RNA-only  
Shared  
DNA-only



## Cell-level measurement



$$e = \frac{\text{avg}_{\text{tumor}} \langle \text{UMI} / \text{copy num.} \rangle}{\text{avg}_{\text{normal}} \langle \text{UMI} / 2 \rangle}$$

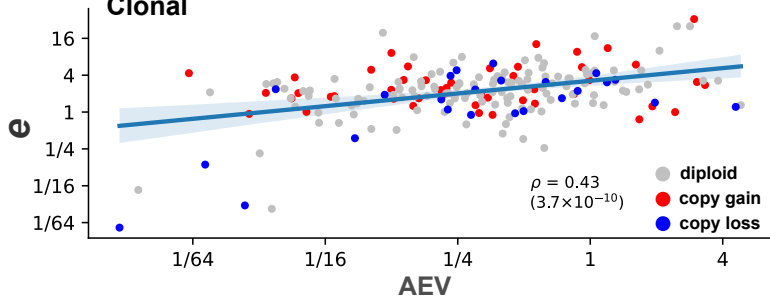
## CNA clonality



CNA cell frac. = 0.4 (2 / 5)

> 0.95: clonal CNA  
0.05 - 0.95: subclonal CNA  
< 0.05: clonal as diploid

## Clonal



## Subclonal

