

SCIENTIFIC REPORTS



OPEN

Systematic pan-cancer analysis of somatic allele frequency

Liam Spurr^{1,2}, Muzi Li^{2,3}, Nawaf Alomran^{2,3}, Qianqian Zhang^{1,4}, Paula Restrepo^{1,2}, Mercedesh Movassagh^{2,5}, Chris Trenkov², Nerissa Tunnessen², Tatiyana Apanasovich⁶, Keith A. Crandall⁷, Nathan Edwards^{2,3} & Anelia Horvath^{1,2,4,7}

Received: 29 December 2017

Accepted: 11 April 2018

Published online: 16 May 2018

Imbalanced expression of somatic alleles in cancer can suggest functional and selective features, and can therefore indicate possible driving potential of the underlying genetic variants. To explore the correlation between allele frequency of somatic variants and total gene expression of their harboring gene, we used the unique data set of matched tumor and normal RNA and DNA sequencing data of 5523 distinct single nucleotide variants in 381 individuals across 10 cancer types obtained from The Cancer Genome Atlas (TCGA). We analyzed the allele frequency in the context of the variant and gene functional features and linked it with changes in the total gene expression. We documented higher allele frequency of somatic variants in cancer-implicated genes (Cancer Gene Census, CGC). Furthermore, somatic alleles bearing premature terminating variants (PTVs), when positioned in CGC genes, appeared to be less frequently degraded via nonsense-mediated mRNA decay, indicating possible favoring of truncated proteins by the tumor transcriptome. Among the genes with multiple PTVs with high allele frequency, *ARID1*, *TP53* and *NSD1* were known key cancer genes. All together, our analyses suggest that high allele frequency of tumor somatic variants can indicate driving functionality and can serve to identify potential cancer-implicated genes.

Somatic genetic variants can express their effects both directly, by affecting the dynamics and efficiency of the transcript generation and degradation processes, and indirectly, through introduction of functional features that are subject to positive or negative selection. Both types of effects are likely to impact the abundance of the variant-bearing allele in the tumor transcriptome. As such, the tumor transcriptome is quickly emerging as an informative source for exploring the functionality of the somatic variants^{1–3}.

Single nucleotide variations (SNVs) comprise a major fraction of the somatic alterations in the tumor genome and are a significant contributor to the tumor cell phenotype and functionality. The functional consequences of the SNVs are inferred from their predicted effect on the protein, such as amino acid change, splicing alteration, or premature truncation of the protein. The later represents a special case expected to affect the allele prevalence through depletion of the variant-bearing transcripts via nonsense-mediated mRNA decay (NMD)⁴. In addition, SNVs can express their functionality through altering sequence motifs recognizable by other molecules, such as transcription or splicing factors, or stabilizing and supporting complexes^{5,6}. These effects commonly manifest in a cis fashion and directly impact, positively or negatively, the relative abundance of the variant bearing allele^{5–7}.

Imbalanced somatic allele prevalence can both cause and result from altered cellular functioning and can thereby play a substantial role in cancer initiation and progression. Indeed, asymmetrically expressed alleles are reported to play a role in variety of cancer types, including blood, breast, ovarian, and lung cancer^{8–12}. Importantly, several recent studies have suggested distinct patterns of allele expression for genes implicated in cancer^{1–3}. This, in turn, suggests that asymmetric alleles can be used to indicate potential cancer-implicated functionality.

¹Department of Pharmacology and Physiology, School of Medicine and Health Sciences, The George Washington University, Washington, DC, 20037, USA. ²McCormick Genomics and Proteomics Center, School of Medicine and Health Sciences, The George Washington University, Washington, DC, 20037, USA. ³Department of Biochemistry and Molecular and Cellular Biology, Georgetown University, School of Medicine, Washington, DC, 20057, USA. ⁴Department of Biochemistry and Molecular Medicine, School of Medicine and Health Sciences, The George Washington University, Washington, DC, 20037, USA. ⁵University of Massachusetts Medical School, Program in Bioinformatics and Integrative Biology, Worcester, MA, 01605, USA. ⁶Department of Statistics, The George Washington University, Washington, DC, 20037, USA. ⁷Computational Biology Institute, Milken Institute School of Public Health, The George Washington University, Washington, DC, 20052, USA. Correspondence and requests for materials should be addressed to A.H. (email: horvatha@gwu.edu)

Several factors are critical for the assessment of somatic allele abundance and its downstream effects on cellular function. Among them, of ultimate importance is the corresponding DNA's allele content, which reflects both copy number alterations (CNAs) and admixture with non-tumor genomes. The latter is commonly referred as genome (or sample) “purity”, and is acknowledged to impact the outcome of cancer genomic studies, with consequences on downstream analyses and results' interpretation¹³. Second, variant-independent imprinting effects need to be distinguished in order to identify variant-specific allele preferences^{14,15}. Finally, the effects of the asymmetrically expressed alleles on total gene expression are essential for assessing the downstream consequences of somatic variants. To account for all three factors in our analyses, we used the unique data set that consisted of matched tumor and normal RNA and DNA sequencing data from the same patient. By using this approach, we aimed to: (1) account for the contribution of DNA in the quantitation of the allele abundance, and (2) link allele frequency of somatic variants to total gene expression, as measured by direct assessment of the fold-change of the expression levels as compared to the matched normal tissue.

Herein, we report the results from the analysis on the allele-specific and total gene expression of somatic variants across 10 cancer types obtained through The Cancer Genome Atlas (TCGA)¹⁶: Urothelial Bladder Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Head and Neck Squamous Cell Carcinoma (HNSC), Kidney Renal Clear Cell Carcinoma (KIRC), Liver Hepatocellular Carcinoma (LIHC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Prostate Adenocarcinoma (PRAD), Thyroid Carcinoma (THCA), and Uterine Corpus Endometrial Carcinoma (UCEC). We estimate allele prevalence and test for correlation with functional features, including predicted effects on the protein, location in motifs involved in interaction with other molecules, and conservation of the genome position. Next, we link somatic allele frequency to gene expression levels, as compared to the corresponding normal tissue sample. We then compare the observations between genes known to be implicated in cancer (Cancer Gene Census, CGC)¹⁷ and the rest of the genome, and document several patterns that are frequently confined to the CGC genes. In summary, we present an expanded set of somatic allele features, whose integrated analysis suggests novel links between transcriptome allele preference and cancer-implicated functionality.

Results

Analytical pipeline and overall dataset characteristics. The pipeline of our analysis is presented on Fig. 1a. From the originally selected for the study samples (Supplementary Table 1), we excluded those with insufficient purity assessments, extensive number of somatic mutations, and technical variables that can affect the assessments (See Methods). This retained 5523 high-confidence exonic SNVs in 3983 genes, 230 of which listed in the CGC (Supplementary Table 2). For each SNV, we quantitated the somatic allele frequency as the expressed variant allele fraction in the tumor RNA (VAF_{tRNA}), including its extremes, where the somatic variant is either completely lost in the transcriptome (SOM-L, $VAF_{tRNA} \sim 0$) or over-expressed (SOM-E, $VAF_{tRNA} \sim 1$, Fig. 1b)¹². Of note, SOM-L and SOM-E are called in the presence of a bi-allelic signal in the DNA ($0 < VAF_{tDNA} < 1$)¹². The remaining somatic variants were designated as SOM. Across the ten cancer types, we called 350 SOM-E variants (6.3% of all variants) and 1044 (18.9%) as SOM-L (Fig. 1c and Supplementary Table 3). As compared to SOM-E, SOM-L SNVs were more frequent in all 10 cancers, with the difference ranging from 1.8-fold for HNSC, and 13-fold for PRAD and THCA. The highest proportion of SOM-E variants was observed in HNSC (consistent with another study exploring HNSC samples from TCGA³), LUSC, and BLCA. Next, we estimated the tumor expressed allele frequency (VAF_{tRNA}) relative to the variant allele fraction in the tumor DNA (VAF_{tDNA}), for which we used the previously introduced expression $V_{R,D} = VAF_{tRNA}/VAF_{tDNA[1]}$.

As expected, our analysis revealed positive correlation between VAF_{tRNA} and VAF_{tDNA} for the individual cancer types (Fig. 1d) and the pooled data across the ten cancer types (Spearman $r_s = 0.38$, $p < 2.2e-16$, Fig. 1e). The distribution of $V_{R,D}$ in regards to the SOM-E, SOM and SOM-L categorization of somatic mutations is presented on Fig. 1f. In the presented below results, we used $V_{R,D}$ to correlate with: (1) functional features of the mutation, including predicted effects on the protein, (2) position in genes that have been causally implicated in cancer, as defined by the CGC¹⁷, (3) total gene expression, and, (4) location in transcription factor binding sites (TFBS).

Somatic allele expression in the context of the predicted function on the protein. We first assessed the distribution of $V_{R,D}$ across the entire dataset, and compared variants located in CGC-genes, with those in the rest of the genome. As seen in Fig. 2a, the $V_{R,D}$ distribution between variants in CGC and non-CGC genes did not show a statistically significant difference ($p = 0.18$). We next assessed the distribution of the tumor $V_{R,D}$ across the different categories of somatic mutations with regards to their predicted effect on the protein: stop-codon generating (premature terminating variants, PTV), missense, and silent. Naturally, PTVs showed distinct allele expression as compared to the missense and noncoding variants (Kruskal-Wallis $p = 4.8e-16$, Fig. 2b). The difference presented both as lower average $V_{R,D}$, and higher proportion of mutations that are not expressed at all (SOM-L). This distribution of the variant allele fraction is suggestive for higher degradation rate of PTV-containing transcripts through nonsense-mediated mRNA decay (NMD)⁴ and concurs with the expected pattern and the observations from related studies¹⁻³. However, when we analyzed the variants in the CGC group separately, we observed a different shape of the PTV $V_{R,D}$ (Fig. 2c), with higher average allele frequency, and a lower proportion of SOM-L variants (See Fig. 3 and the corresponding section of the results). These effects were also apparent in the individual cancer types, although in some cancers the number of PTVs in CGC genes was low (Fig. 2d,e). Overall, in the CGC genes, the $V_{R,D}$ distribution for PTVs was more similar to the $V_{R,D}$ distributions of missense and silent variants.

PTV-containing transcripts. To further explore the above-noted difference in the CGC PTV-alleles' prevalence, we directly compared the $V_{R,D}$ of PTVs located in CGC and in non-CGC genes. Consistent with the above, higher PTV $V_{R,D}$ was seen in the CGC dataset ($p = 0.036$, Fig. 3a). Of note, a similar shape of the $V_{R,D}$

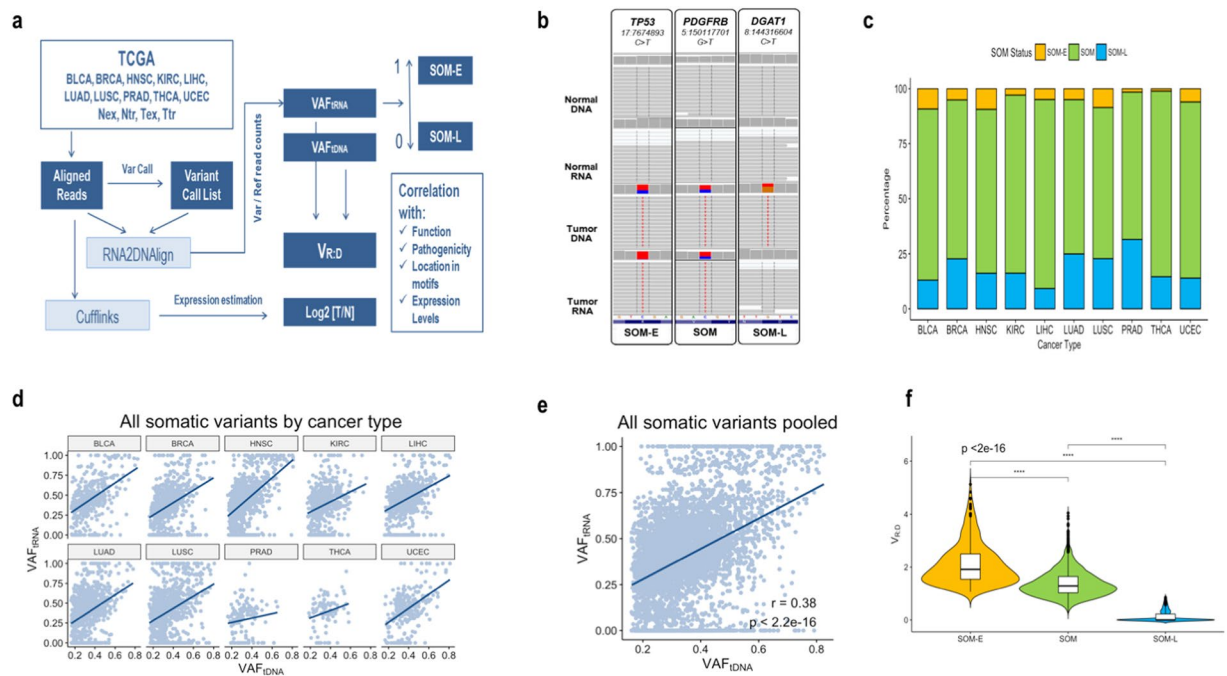


Figure 1. (a) Major steps of the analysis of allele distribution for somatic variants in the ten studied cancer types. Variant allele frequency ($V_{R,D}$) was analyzed for correlation with different functional features and expression levels in cancer implicated genes (CGC) and the rest of the genome. SOM-E and SOM-L represent the extremes of the VAF_{tRNA} . (b) Visualization of alignments (Integrative Genome Viewer, IGV) of examples of (from left to right) SOM-E, SOM and SOM-L variants. The bi-allelic position is reflected through color-coding of the summary flag on the top of each panel. The gray lines represent sequencing reads, and the colored letters show differences from the reference. (c) Proportion of SOM, SOM-E and SOM-L across BLCA, BRCA, HNSC, KIRC, LIHC, LUAD, LUSC, PRAD, THCA and UCEC. The distribution of the proportions was similar across the 10 cancer types, with HNSC showing the highest, and PRAD and THCA the lowest proportion of SOM-E variants. In all cancer types, the SOM-L variants represented a higher than SOM-E proportion of the total SOM variants. (d) Positive correlation (Spearman, r_s) between VAF_{tRNA} and VAF_{tDNA} in the ten cancer types, and, in the pooled across-cancer data (e) Values of 0 and 1 on the y-axis represent SOM-L and SOM-E variants, respectively. (f) Distribution of variant allele frequency ($V_{R,D}$) in the SOM-E, SOM, and SOM-L mutation categories.

distribution was seen in the individual cancer types with a number of PTVs in the CGC group sufficient for analysis (Supplementary Fig. 1a), where PTVs showed higher $V_{R,D}$ and lower proportion of SOM-L variants in the CGC genes.

The observed higher allele frequency of somatic PTVs in CGC genes (as compared to the non-CGC genes) suggests a lower NMD-degradation rate of PTV-containing transcripts. To assess if this observation is due to a higher proportion of PTVs residing in NMD-escaping regions/genes in the CGC data set, we set out to perform our analysis exclusively on PTV-alleles predicted to be NMD-degraded (i.e. NMD-elicited PTVs), as defined by their position in the gene and the gene's particular features. Accordingly, we removed from the analysis PTV-containing transcripts that are likely to escape NMD, combining escape rules defined by two major recent NMD studies^{18,19}. Based on the findings in those studies, we filtered out mutations located: (1) more than 50bp upstream of the last exon-exon junction, (2) in the first 200 nucleotides after the start codon where an alternative in-frame AUG is present, (3) in long exons, and positioned <250 nt from the closest exon boundary, (4) in short-living (<1 h half-life) transcripts^{20,21}, (5) in single-exon genes, and, (6) in NMD-insensitive genes. This filtering retained 154 exclusively NMD-elicited PTVs (Supplementary Table 4), for which we re-estimated the variant allele frequency in CGC and non-CGC genes. Notably, in this stringent NMD-elicited group, the allele frequency distribution preserved its shape and the significance of the observed higher $V_{R,D}$ in the CGC genes ($p = 0.011$, Fig. 3b). This result indicates that the high PTV allele expression in the CGC genes cannot be explained solely by NMD prediction based on mutation position, gene structure, and transcript processing dynamics. Instead, it suggests tolerance of cancer-implicated PTV-containing transcripts by the tumor transcriptional machinery.

We then set out to determine if the higher allele frequency of somatic PTVs in the tumor tissue correlates with total gene expression levels as compared to the matched normal tissue. To accomplish this, we estimated the expression changes of the genes bearing somatic mutations, measured as \log_2 of the fold-change between the expression levels in the tumor compared to the matched normal tissue ($\log_2[T/N]$) and correlated $\log_2[T/N]$ with the variant allele frequency $V_{R,D}$. Intriguingly, this analysis showed a positive correlation between $V_{R,D}$ and the total expression change of PTV-bearing genes ($r_s = 0.28$, $p = 1.2e-04$, Fig. 3c), which further increased in the group of the CGC genes ($r_s = 0.46$, $p = 3.5e-03$, Fig. 3d).

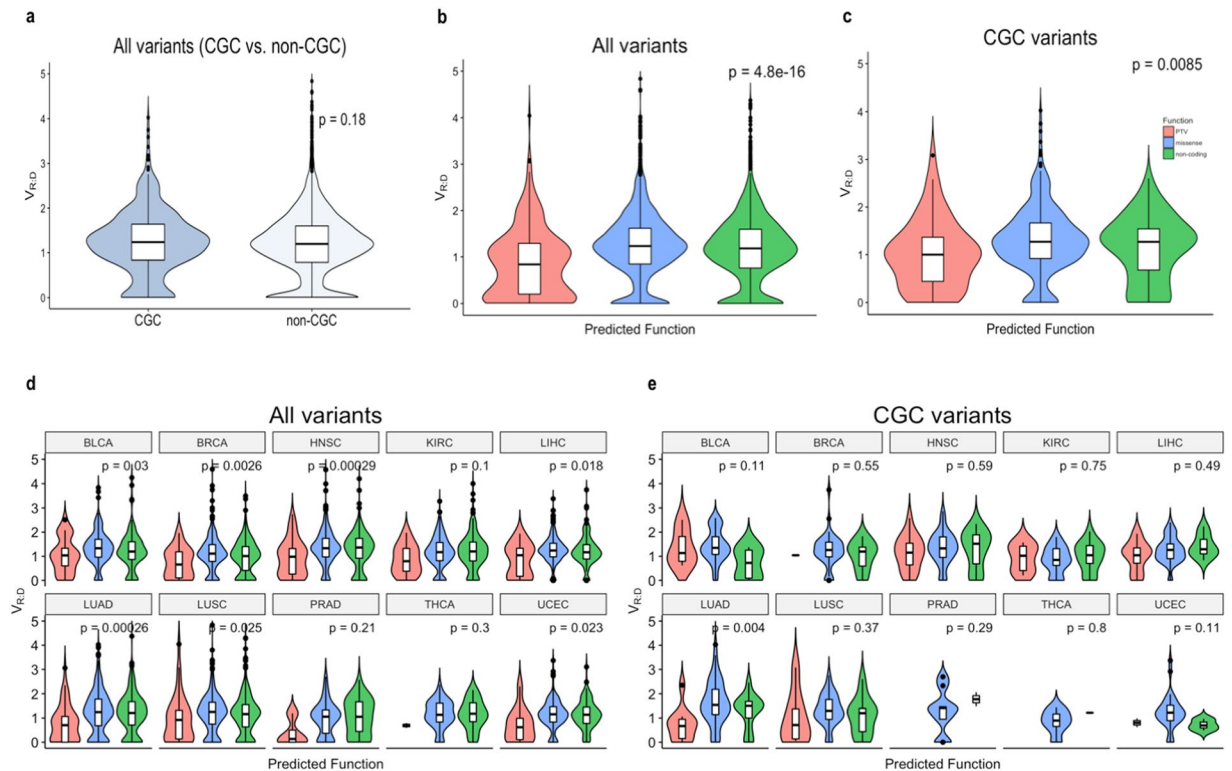


Figure 2. (a) Distribution of $V_{R,D}$ in somatic mutations between CGC and non-CGC genes. (b) Distribution of $V_{R,D}$ in somatic mutations categorized based on their predicted effect on the protein function in the entire dataset and in the CGC subset. (c) The mutation categories correspond to: generating a premature termination codon variant (PTV), substitution of an amino acid (missense) or not altering a coding sequence (silent). In both analyses the PTVs presented with significantly different distribution of $V_{R,D}$ ($p < 0.05$) expressed as lower average allele fraction and a higher proportion of SOM-L mutations ($V_{R,D} \sim 0$). However, while still significantly different, the PTV $V_{R,D}$ in the CGC genes appears more similar to the $V_{R,D}$ of missense and silent variants, as compared to the pooled data from all genes. (d) Distribution of $V_{R,D}$ in PTV, missense and silent variants in individual cancer types across the entire dataset, and (e) in CGC-genes.

Next, we assessed the allele frequency in CGC and non-CGC genes for the missense and silent somatic variants in our dataset. The difference between CGC- and non-CGC-located missense and silent variants did not reach statistical significance ($p = 0.12$, and $p = 0.81$, respectively, Fig. 4); analogous patterns were seen in most of the analyses in the individual cancer types (Supplementary Fig. 1b,c, respectively).

Correlation between allele-preferential and total gene expression in the entire dataset. We then set out to determine if there is a correlation between allele frequency and total gene expression of the variant-harboring gene relative to the matched normal tissue. This analysis revealed a positive correlation between $V_{R,D}$ and total gene expression change across the entire dataset ($r_s = 0.17$, $p < 2.2e-16$, for the pooled variants, Fig. 5a), as well as in the CGC subset ($r_s = 0.23$, $p = 1.9e-05$, for the pooled variants, Fig. 5b). In three of the individual cancer types, LUAD, THCA and UCEC, we observed a significant correlation between $V_{R,D}$ and $\log_2[T/N]$ in the CGC gene-set. In these three cancer types, the observed in the CGC gene-set correlation was substantially stronger as compared to the one seen in the entire gene-set (LUAD, $r_s = 0.53$, $p = 2.1e-04$, THCA: $r_s = 0.9$, $p = 0.037$, and UCEC: $r_s = 0.66$, $p = 5.8e-4$ (Fig. 5c,d).

We then outlined the extreme subsets of somatic variants with co-occurring high expressed allele frequency and increased or decreased total expression of the harboring gene. To do that, we selected variants with $V_{R,D} > 2$, (indicating at least 2-fold higher expressed allele frequency ($VA_{F_{IRNA}}$), as compared to the DNA allele frequency ($VA_{F_{DNA}}$)), and at least a two-fold increase or decrease in the total gene expression in the tumor as compared to the normal tissue. A total of 107 somatic variants had expressed allele frequency co-occurring with increased gene expression. In another 73, the increased allele frequency co-occurred with a more than 2-fold decrease in the total expression of the harboring gene (Supplementary Tables 5 and 6).

Correlation of allele frequency and predicted functionality. We then tested $V_{R,D}$ in correlation with mutation pathogenicity as modeled through several popular functionality predictive tools: Polyphen (PPH)²², Combined Annotation Dependent Depletion (CADD)²³, Genomic Evolutionary Rate Profiling (GERP)^{24–26}, and Functional Analysis Through Hidden Markov Models (FATHMM)^{27,28}. We did not observe a strong correlation between the $V_{R,D}$ and the pathogenicity value computed by FATHMM (Supplementary Fig. 2a,b). However, when

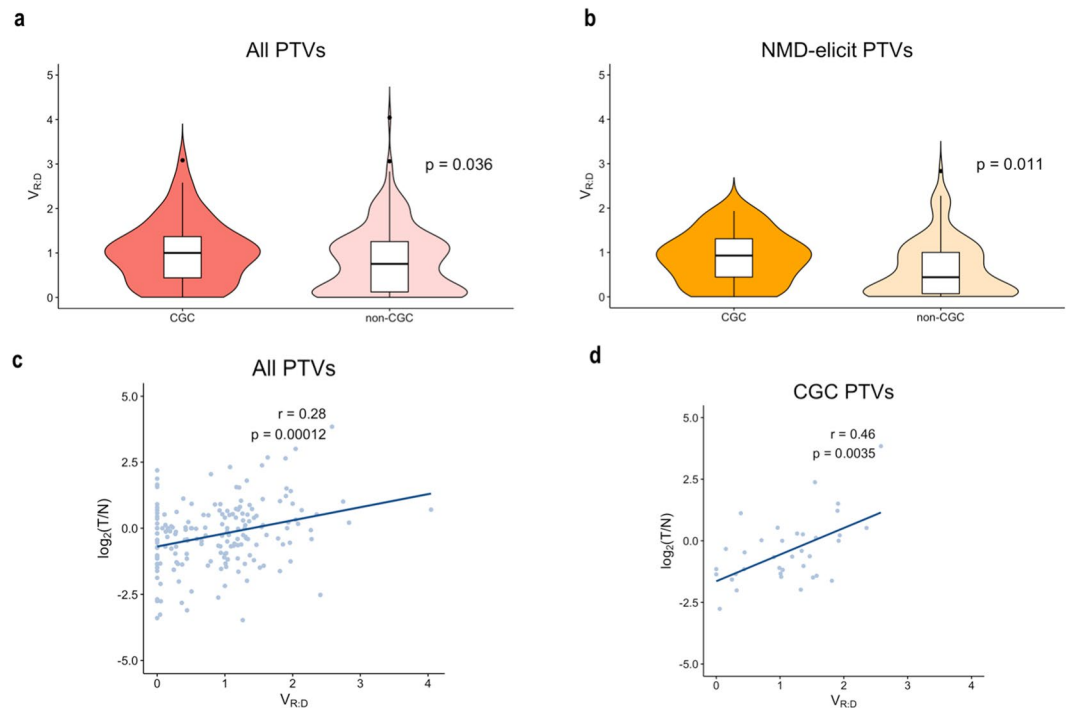


Figure 3. (a) Direct comparison of PTV $V_{R,D}$ distribution in CGC and non-CGC genes shows statistical significance ($p = 0.036$). The difference is more pronounced when exclusively NMD-elicited PTVs are analyzed ($p = 0.011$, b). (c) A positive correlation between $V_{R,D}$ of PTVs and $\log_2[T/N]$ (Spearman, r_s) is seen for the PTVs in all genes and (d) in the CGC genes only.

we assessed the categorical FATHMM estimations (categorized as either “pathogenic” or “neutral”), higher $V_{R,D}$ was seen in the pathogenic variants ($p = 0.023$, Supplementary Fig. 2c). No strong correlations between $V_{R,D}$ and CADD, GERP or PolyPhen scores of the variants in the CGC genes and non-CGC genes were observed (Supplementary Fig. 2d–i).

Somatic allele-frequency and TFBS. We next examined if somatic allele frequency correlates with residence of the mutation in motifs recognizable by transcription factors (i.e transcription factor binding sites, TFBS). To accomplish this, we assessed the variant and the reference motifs for TFBS using TRANSFAC²⁹, and then categorized the variants in the following three groups: generating a new TFBS (TFBS-gain, 2736 variants), destroying an existing TFBS (TFBS-loss, 1980 variants), and not changing a known motif (Supplementary Table 7). We then compared the allele-frequency and the total gene expression among the three categories. We did not observe a statistical difference between TFBS-gain and TFBS-loss variants in the CGC-subset ($p = 0.13$, Fig. 6a), however, in the TFBS-loss variants we detected a positive correlation with the gene expression levels ($r_s = 0.31$, $p = 0.03$). This observation is consistent with degradation (by the somatic mutations) of existing motifs for transcription suppressing factors (TFs). Because most of the TFs found to recognize somatically altered motifs are known to act both as activators and suppressors of the transcription, it was not possible to impartially determine if suppressing or activating activity is associated with allelic and total expression (See Supplementary Table 7).

Purity and CNA assessments. $V_{R,D}$ reflects both admixture of the tumor sample with non-tumor material and CNAs. To further delineate the effects of purity and CNAs on the observations, we obtained: (1) purity of each sample, as estimated by 5 different methods: Estimate, Absolute, LUMP, IHC, and the Consensus Purity Estimation (CPE)¹³, and, (2) the CNAs assessments measured as genomic segment mean values ($\log_2(\text{copy-number}/2)$) from TCGA.

Because purity affects both DNA and RNA content, we directly weighted the relative (to DNA) expressed allele frequency - $V_{R,D}$ - by the proportion of the sample estimated by each of the above methods to correspond to tumor tissue and used these adjusted values to perform the analyses described above. The results are presented in Supplementary Figs 3–7. Overall, most of the observations retained their significance. Notably, ESTIMATE- and IHC-weighted $V_{R,D}$ ($eV_{R,D}$ and $iV_{R,D}$, respectively) produced a higher number of significant observations as compared to the unweighted $V_{R,D}$ (Supplementary Figs 3 and 4). For example, the difference in $eV_{R,D}$ between CGC and non-CGC variants across the entire dataset reached statistical significance ($p = 0.05$, See Supplementary Fig. 3b).

To assess if CNAs contribute to the observations and more specifically to the differences between SNVs in CGC and non-CGC genes, we compared the absolute segment means between CGC and non-CGC genes in the context of their harbored SOM-E and SOM-L variants (Supplementary Fig. 8). This analysis showed similar distributions of the absolute segment means between variants in CGC and non-CGC genes in the entire dataset

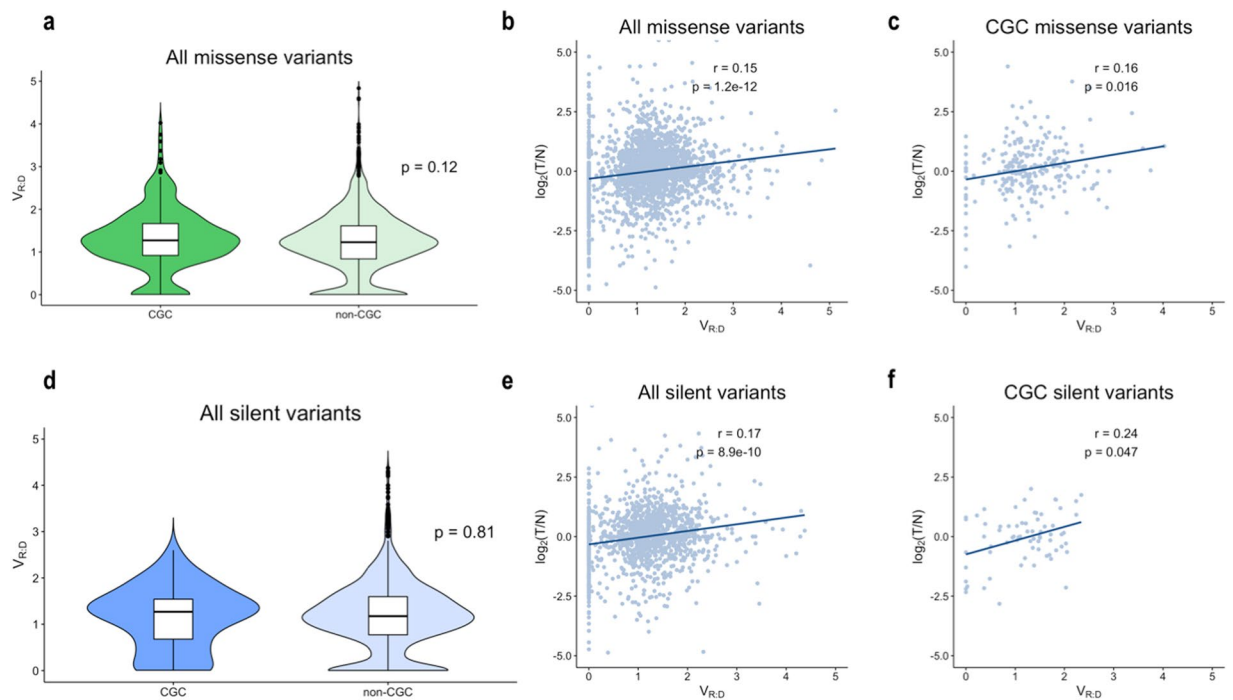


Figure 4. (a) $V_{R,D}$ distribution for missense variants in CGC and non-CGC genes. (b) Correlation between $V_{R,D}$ and expression change $\log_2(T/N)$ (Spearman, r_s), of missense variants in the entire dataset and in the CGC subset only (c). (d) $V_{R,D}$ distribution for silent variants between CGC and non-CGC genes was similar ($p = 0.81$). (e) Correlation between $V_{R,D}$ and $\log_2(T/N)$ in the entire set of silent variants, and for those positioned in CGC genes only (f).

(Supplementary Fig. 8a), in the groups of the SOM-E and SOM-L variants (Supplementary Fig. 8b,c, respectively), as well as in the groups of SOM-E PTVs and SOM-L PTVs (Supplementary Fig. 8d,e, respectively). These observations suggest that in the studied dataset, CNAs are unlikely to substantially contribute to the observed differences between CGC and non-CGC genes. Aligned with the above, the positive and negative segment means (corresponding, respectively, to amplifications and deletions) did not correlate with SOM-E or SOM-L estimations (Supplementary Fig. 9a,b).

We note that several factors that can affect the presented observations remain unaccounted for in our study. Technical variables related to differences in the RNA (and DNA) extraction, sequencing library generation and processing, depth of sequencing, and sequencing platform can affect the presented observations, including the proportion of transcripts containing PTVs. Importantly, if RNA and DNA are not extracted from the same tumor specimen, they may have different features, including purity and CNAs. Furthermore, “normal” tissue obtained from a tumor-adjacent site may confer expression patterns affected by exposure to tumor signals. While TCGA represents a major effort for uniformity of tumor collection and processing, it is important to consider possible effects of the above factors on the expressed variant allele frequencies. Finally, the threshold for minimal read counts per loci selected for analyses can impact the computations. Naturally, low thresholds include a higher number of variants but decrease the confidence of VAF estimations for under-expressed genes, while high thresholds support confident VAF measurements in a lower total number of variants and exclude genes with low expression. Our analyses with a range of different thresholds on subsets of the herein analyzed datasets showed comparative outcomes¹². However, the effects of the read count threshold on the VAF assessments are important to contemplate.

Genes with multiple somatic variants. In our dataset, 49 genes had 5 or more somatic mutations (Fig. 7a and Supplementary Table 8). To examine the $V_{R,D}$ in genes with multiple somatic mutations, we ranked the genes by their mean $V_{R,D}$.

Naturally, owing to the high somatic mutation frequency used for the selection, this gene-set was enriched in CGC genes ($p < 0.001$, chi-square test). From this gene-set, three NMD-sensitive genes had multiple NMD-elicited PTVs with $V_{R,D}$ values similar or higher to the $V_{R,D}$ values of missense and silent mutations in the same gene: *ARID1*, *TP53*, and *NSD1* (all of them CGC genes).

We then examined the set of 283 genes with 3 or more somatic mutations (Supplementary Table 9). While this dataset was also significantly enriched in CGC genes, no over-representation of CGC genes was seen in the 50 genes with the lowest mean $V_{R,D}$ (Fig. 7b). The CGC genes with lowest mean $V_{R,D}$ based on 3 and more mutations were the protein tyrosine phosphatases *PTPRB* and *PTPRC*, and *FGFR1*. Naturally, low $V_{R,D}$ confers features that imply intolerance by the transcriptional machinery of the harbored variant. Apart from CNAs and NMD, possible factors contributing to the SOM-L prevalence are (1) low DNA allele frequency and related technical variabilities

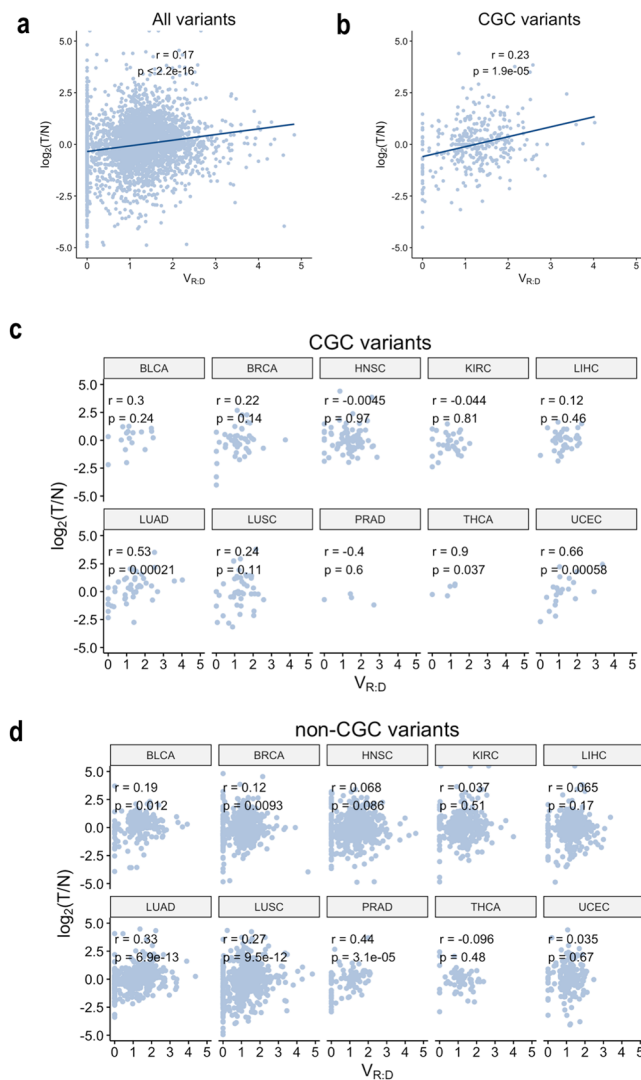


Figure 5. Correlation between $\log_2[T/N]$ and $V_{R,D}$ in the entire dataset (a) and the CGC genes subset (b). (c) Correlation between $\log_2[T/N]$ and $V_{R,D}$ in the individual cancer types in CGC genes, and, non-CGC genes. (d) In the group of the CGC genes, LUAD, THCA and UCEC showed statistically significant correlations.

and (2) general infidelity of the cancer transcriptional machinery. Additional factors, such as estrogen receptor (ER) expression levels, are also reported to correlate with the number of expressed somatic mutations in breast cancer². All these factors are likely to dilute functional annotations in the SOM-L mutations. In contrast, the set of the top 50 genes with high mean $V_{R,D}$ based on 3 or more somatic variants was significantly enriched with CGC genes ($p < 0.01$, Fig. 7c), including *CTCF*, *CDKN2A*, *CDK12*, *DNMT3A*, *PTEN*, *KMT2C*, and *KEAP1*³⁰.

Discussion

Despite the growing accessibility of human cancer sequencing data, the number of available normal and tumor DNA and RNA datasets from the same individual/sample is still limited. In contrast to tumor expression data alone, the matched normal transcriptome provides the opportunity to directly assess normal-to-tumor expression changes, and to link them to features of interest in the tumor sample, including somatic variants' allele frequency and functionality. In this study, we employed the unique set of matched tumor and normal RNA and DNA sequencing data obtained from TCGA to assess somatic allele frequency, analyze the correlation between allele expression and functionality of somatic SNVs, and connect to potentially mutation-instigated expression changes. Ultimately, the major outcome from our study is the demonstration of the vast value of the information provided by somatic allele frequency. In addition, several striking observations were apparent from our analyses.

First, somatic variants in CGC genes were more frequently expressed and had higher allele frequency in comparison to the somatic variants in the rest of the genes. Second, variants with high predicted functionality and/or pathogenicity generally had increased allele frequency as compared to neutral variants, and this observation was stronger in the subset of the CGC genes. High somatic allele frequency co-occurring with increased gene expression level is consistent with variant-mediated transcriptional up-regulation, acting in cis-fashion⁶. Alternatively, the mutant transcripts could be favored by the tumor cell due to a growth or a survival advantage provided by the mutation. Both scenarios infer functionality and/or tumorigenic potential, which are frequent features of

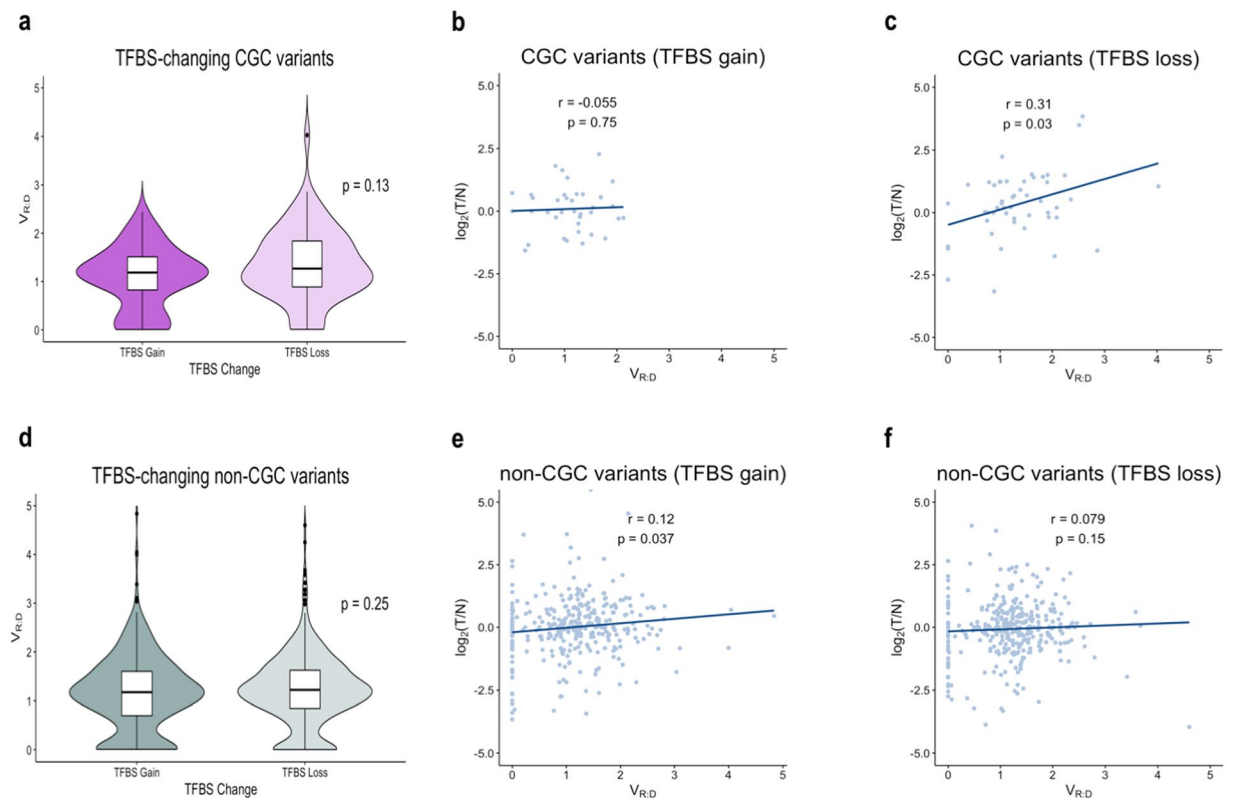


Figure 6. (a) $V_{R,D}$ distribution for somatic variants that generate a new TFBS (TFBS-gain) as compared to those that disturb an existing TFBS (TFBS-loss). (b) Correlation between $\log_2[T/N]$ and $V_{R,D}$ TFBS-gain variants, and in TFBS-loss variants (c). (d) Distribution of $V_{R,D}$ between TFBS-gain and TFBS-loss somatic variants in non-CGC genes; no difference was observed. (e) Correlation between $\log_2[T/N]$ and $V_{R,D}$ in TFBS-gain variants in the CGC genes, and in the non-CGC genes (f).

mutations in tumor suppressors and oncogenes, and hence concur with the observed allele behavior in the CGC subset. Comparatively, high somatic allele frequency co-occurring with decreased total expression of the gene can indicate more complex scenarios where the variant allele is selected in the tumor transcriptome possibly due to down-regulation of the produced protein. The variants outlined by our analysis represent examples that can be further explored for consistency with the herein proposed mechanisms.

The observations above are largely consistent across the ten individual cancer types, as well as with our previous findings on breast cancer, and related studies by others^{1–3}. All these studies illustrate the general tendency of the cancer transcriptome to favor advantage supplying variants, which are ultimately those with higher functionality and pathogenic potential¹. The new layer provided by this study is the linking of gene expression changes in the tumor as compared to the matched normal tissue, which further affirms the previous observations. Most importantly, it represents a step towards the interpretation of the downstream effects of the somatic variants. In most of our analysis, expression changes correlated positively with somatic allele frequency. Naturally, this is not surprising as the absolute allele quantity contributes both to the allele frequency and the total gene expression. However, an unexpected notable finding from the expression analysis is the high expression of the PTV-bearing transcripts in the CGC genes, which suggests that these genes are more frequently targets of NMD-escaping somatic variants.

Several explanations might account for the higher frequency of NMD-escaping PTVs in the CGC genes. Because CGC is a diverse group containing both oncogenes and tumor suppressor genes, it is reasonable to presume that the observation results at least partially from non-specific effects. For one, it is possible that PTV-transcripts and their encoded shorter proteins non-specifically (i.e. through depleting the cellular molecular sources) prevent the normal cellular processes from counteracting the tumorigenic activity, and are thus favored as advantage-supplying features. In line with the above, it is also possible that the dynamic growth and replication of the tumor cell does not allow enough time for the NMD machinery to allocate molecular resources and properly degrade newly arisen (somatic) PTV-containing transcripts. The above effects would be more pronounced in the CGC subset if the CGC genes were more actively transcribed (and degraded) in the tumor cell. The latter is an attractive possible explanation, especially given that short-living transcripts more frequently escape NMD¹⁹. Thus, by a wider definition of oncogenic action, the NMD-escaping transcripts might act as advantage-supplying molecules via general dysregulation of cellular functioning. It is not possible to predict from our analyses whether these PTV-containing transcripts are translated into shorter proteins. Nevertheless, NMD-escaping mutations appear to target CGC genes at a higher frequency than the rest of the genome, and this observation demands further studies, including at the protein level.

the high confidence curated alignments and variant calls directly from TCGA¹⁶. While hg38 arguably supplies higher annotation accuracy, it is still in the process of assimilation into the genome-studying research community, and most of the major NMD studies so far have used hg19^{18,19}, which can partially account for the differences between these studies and our own.

In summary, we believe that our analyses present novel and intriguing insights into the allele-preferential expression of somatic mutations in cancer. Importantly, ranking of genes based on somatic allele frequency scores key cancer genes at top positions, and thus suggests that high somatic allele frequency can be used to indicate potential carcinogenic variants, and possibly, to identify cancer-driving genes.

Methods

TCGA samples. To systematically quantify somatic allele prevalence, we compiled data on single nucleotide somatic variants from The Cancer Genome Atlas (TCGA) for patients for whom the following sequencing datasets were available: normal exome (Nex), normal transcriptome (Ntr), tumor exome (Tex), and tumor transcriptome (Ttr). We selected 12 cancer types: BLCA, BRCA, COAD (Colon Adenocarcinoma), HNSC, KIRC, LIHC, LUAD, LUSC, PRAD, STAD (Stomach Adenocarcinoma), THCA, and UCEC, that had more than 10 samples with all four sequencing datasets. In addition, we required each sample to have at least three of the following five purity estimates - Estimate, Absolute, LUMP, IHC, and the consensus purity estimate (CPE), (See Supplementary Table 1), as well as CNA estimation (genomic segment means based on Genome-Wide-SNPv6 hybridization array)^{13,33–35}. This initial set consisted of 416 individuals. From those, we excluded samples with an extensive number of somatic mutations (more than 1.5 interquartile ranges (IQR) above the third quartile, in our dataset 56), possibly due to clustered genomic rearrangements or other rare mechanisms of acquirement of somatic mutations^{36,37}. As a result, COAD and STAD sets retained less than 10 patients suitable for assessment and were removed from further analyses. On the remaining samples, we applied stringent filters to ensure compatible RNA and DNA frequency estimation and to minimize effects due to CNAs, technical variables, and admixture with non-tumor cells. Importantly, we filtered out variants residing in imprinted genes (See Methods)^{14,15,38,39}. These filtering steps yielded 5523 high-confidence exonic SNVs in 3983 genes, from which 230 were listed in the CGC (See Supplementary Table 2).

Allele count and expression level computation. All the datasets were generated through paired-end sequencing on an Illumina HiSeq platform. The human genome reference (hg38)-aligned sequencing reads (Binary Alignment Maps, .bams) and the Simple Nucleotide Variation mutation annotation file (SNV.maf) were downloaded from the Genomic Data Commons Data Portal (<https://portal.gdc.cancer.gov/>) and processed downstream through an in-house pipeline. The RNA and DNA alignments, together with the variant lists were processed through RNA2DNAlign¹². RNA2DNAlign produced variant and reference sequencing reads counts for all the variant positions in all four datasets (normal exome, normal transcriptome, tumor exome and tumor transcriptome). Selected read count assessments were visually examined using Integrative Genomics Viewer⁴⁰. We excluded from further analyses variants which (1) were covered with less than 10 sequencing reads in the tumor DNA or the RNA sequencing data; (2) reside in known imprinted regions, and (3) were present in the normal DNA or RNA, suggesting germline origin. Variants positioned in the X Chromosome and on stably imprinted autosomal genes^{14,15} were excluded from the analyses. For the NMD-analysis, short-living (<1 h half-life) transcripts were identified based on Tani *et al.*^{20,21}. The gene expression was quantified using the Cufflinks package from the Tuxedo suite, as we have previously described⁴¹.

Assessment for allele distribution. Allele frequencies within a sample were determined through estimation of the relative abundance of variant over total sequence read counts, expressed as Variant Allele Fraction (VAF). For each somatic mutation, we computed the $VAF = n_{var} / (n_{ref} + n_{var})$, for both tumor RNA (VAF_{tRNA}) and tumor DNA (VAF_{tDNA}), where n_{ref} and n_{var} are the counts of the reference and variant sequencing reads covering the position, respectively. To account for allele asymmetries related to DNA, we analyzed VAF_{tRNA} in the context of the corresponding VAF_{tDNA} . Over-expression of somatic mutations (SOM-E status) was defined as prevalence of variant sequencing reads in the transcriptome ($VAF_{tRNA} \sim 1$), while SOM-L was defined by complete loss of the mutant allele in the transcriptome ($VAF_{tRNA} \sim 0$)¹². VAF_{tRNA} relative to VAF_{tDNA} ($V_{R:D}$) was computed as previously described: $V_{R:D} = VAF_{tRNA} / VAF_{tDNA}$. Weighting for purity was performed by multiplying $V_{R:D}$ by the proportion of the sample assessed to correspond to the tumor tissue component (i.e. $aV_{R:D} = V_{R:D} \times ABSOLUTE$, $iV_{R:D} = V_{R:D} \times IHC$, $eV_{R:D} = V_{R:D} \times ESTIMATE$, $IV_{R:D} = V_{R:D} \times LUMP$, and $cV_{R:D} = V_{R:D} \times CPE$).

Functional and enrichment analyses. Functional annotations and conservation scores were extracted using the SeattleSeq annotation 147 (<http://snp.gs.washington.edu/SeattleSeqAnnotation147/index.jsp>). Pathogenicity was modeled using PolyPhen, CADD and FATHMM methods, and conservation was assessed based on GERP scores^{23–28}. Transcription factor binding sites were analyzed using TRANSFAC 7.0²⁹.

Statistics. SOM, SOM-E and SOM-L variants were called based on a binomial test for variant and reference sequencing read distribution, as previously described¹². The distributions of the variant allele frequency was assessed using Kruskal-Wallis rank sum test, and the Spearman rank correlation coefficient⁴². P-values below 0.05 were considered significant.

Availability of data and materials. The datasets supporting the conclusions of this article are included within the article and its additional files.

References

- Restrepo, P. *et al.* Overexpressed somatic alleles are enriched in functional elements in Breast Cancer. *Sci. Rep.* **7**, 8287 (2017).
- Shlien, A. *et al.* Direct Transcriptional Consequences of Somatic Mutation in Breast Cancer. *Cell Rep.* **16**, 2032–2046 (2016).
- 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).
- 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).
- Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. **13**, 59 (2011).
- Ding, J. *et al.* Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. **6**, 8554 (2015).
- Edsgård, D. *et al.* GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Sci. Rep.* **6**, 21134 (2016).
- Lin, W. *et al.* Allelic expression imbalance polymorphisms in susceptibility chromosome regions and the risk and survival of breast cancer. *Mol. Carcinog.* **56**, 300–311 (2017).
- French, J. D. & Edwards, S. L. Allelic imbalance in human breast cancer. *Oncotarget* **8**, 10763–10764 (2017).
- Halabi, N. M. *et al.* Preferential Allele Expression Analysis Identifies Shared Germline and Somatic Driver Genes in Advanced Ovarian Cancer. *PLoS Genet.* **12**, 1–22 (2016).
- Okudela, K. *et al.* Allelic imbalance in the miR-31 host gene locus in lung cancer—its potential role in carcinogenesis. *PLoS One* **9**, e100581 (2014).
- Movassagh, M. *et al.* RNA2DAlign: nucleotide resolution allele asymmetries through quantitative assessment of RNA and DNA paired sequencing data. *Nucleic Acids Res.* **44** (2016).
- Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).
- Glaser, R. L., Ramsay, J. P. & Morison, I. M. The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations. *Nucleic Acids Res.* **34**, D29–D31 (2006).
- Savova, V., Patsenker, J., Vigneau, S. & Gimelbrant, A. A. dbMAE: the database of autosomal monoallelic expression. *Nucleic Acids Res.* **44**, D753–6 (2016).
- Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* **45**, 1113–1120 (2013).
- Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177–183 (2004).
- Hu, Z., Yau, C. & Ahmed, A. A. A pan-cancer genome-wide analysis reveals tumour dependencies by induction of nonsense-mediated decay. *Nat. Commun.* **8**, 15943 (2017).
- 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).
- Tani, H. *et al.* Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* **22**, 947–956 (2012).
- Friedel, C. C., Dölken, L., Ruzsics, Z., Koszinowski, U. H. & Zimmer, R. Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic Acids Res.* **37**, e115–e115 (2009).
- Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **07**, Unit7.20–Unit7.20 (2013).
- Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
- Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
- Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
- Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
- Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).
- Shihab, H. A. *et al.* Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum. Genomics* **8**, 11 (2014).
- Matys, V. *et al.* TRANSFAC[®] and its module TRANSCompel[®]: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
- Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–11 (2015).
- Mudvari, P. *et al.* SNPllice: variants that modulate Intron retention from RNA-sequencing data. *Bioinformatics* **31**, 1191–1198 (2015).
- Caminsky, N., Mucaki, E. J. & Rogan, P. K. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. *F1000Research* **3**, 282 (2014).
- Pageès, F. *et al.* Immune infiltration in human tumors: a prognostic factor that should not be ignored. *Oncogene* **29**, 1093–1102 (2010).
- Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
- Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
- Storchova, Z. & Kloosterman, W. P. The genomic characteristics and cellular origin of chromothripsis. *Curr. Opin. Cell Biol.* **40**, 106–113 (2016).
- Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
- Baran, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* **25**, 927–36 (2015).
- Morison, I. M., Ramsay, J. P. & Spencer, H. G. A census of mammalian imprinting. *Trends Genet.* **21**, 457–465 (2005).
- Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
- Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7** (2012).
- Kruskal, W. H. & Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **47**, 583–621 (1952).

Author Contributions

L.S., M.L., N.A., P.R., Q.Z. and N.T. performed the raw large data analysis. C.T. prepared the organizational and management scripts. N.A., M.M., K.C. and N.E. implemented the read counting and statistical significance framework for the SOM-E and SOM-L models; N.E. supervised all the computational scripts. L.S., T.A. and Q.Z. performed the statistical analyses and created the figures. A.H. designed the project, devised the framework and the analytic approach, and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-25462-0>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018