Structural and functional analysis of somatic coding and UTR indels in breast and lung cancer genomes

Jing Chen and Jun-tao Guo*

Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC, 28223, USA

* To whom correspondence should be addressed. Tel: +1 704 687 7492; Fax: +1 704 687 8667; Email: jguo4@uncc.edu

ABSTRACT

Insertions and deletions (Indels) represent one of the major variation types in the human genome and have been implicated in diseases including cancer. To study the features of somatic indels in different cancer genomes, we investigated the indels from two large samples of cancer types: invasive breast carcinoma (BRCA) and lung adenocarcinoma (LUAD). Besides mapping somatic indels in both coding and untranslated regions (UTRs) from the cancer whole exome sequences, we investigated the overlap between these indels and transcription factor binding sites (TFBSs), the key elements for regulation of gene expression that have been found in both coding and non-coding sequences. Compared to the germline indels in healthy genomes, somatic indels contain more coding indels with higher than expected frame-shift (FS) indels in cancer genomes. LUAD has a higher ratio of deletions and a higher coding and FS indel rates than BRCA. More importantly, these somatic indels in cancer genomes tend to locate in sequences with important functions, which can affect the core secondary structures of proteins and have a bigger overlap with predicted TFBSs in coding regions than the germline indels. The somatic CDS indels are also enriched in highly conserved nucleotides when compared with germline CDS indels.

INTRODUCTION

Insertion and deletion (indel) is an important variation type in the human genome, second only to the single nucleotide variations (SNVs) [1-8]. Previous studies have estimated that indels contribute to 16% to 25% of sequence polymorphisms in human populations [2,9-11]. Like other types of variations, indels can alter human traits and lead to diseases including cancer [12-15]. Indel analyses have been carried out in both healthy and cancer genomes. In 2011, about 1.96 million small indels were identified in 79 human genomes, which was reported to have more than 97% validation rate [16]. The 1000 Genomes Project reported 1.48 million indels in 2010. There are 463,377 common indels between the above two studies, a result that reflect a combination of indel diversity and inaccurate indel annotations [1,16,17]. In addition to indels in healthy human genomes, efforts have been carried out to investigate indels in different cancer types. Recent pan-cancer analyses indicated the substantial variations

among different cancer types [18-20]. Similar to indel annotation in heathy genomes, different methods and algorithms may lead to different somatic indel annotations [17].

In coding regions, an indel can be frameshift (FS) or non-frameshift (NFS) depending on the length of an insertion or a deletion [2,16]. If the length of an indel is a multiple of three nucleotides, it is an NFS indel as it only affects the amino acid(s) of the indel while other coding indels that change the open reading frame are considered FS indels. For germline indels in healthy human genomes, the number of FS indels is much lower than expected, suggesting FS indels are potentially deleterious and less tolerated during evolution [21]. Several programs have been developed to predict the potential disease-causing NFS and FS indels [22-27]. To better understand the role of somatic indels in cancer genomes, studies have been done at both domain and protein level. Pagel *et al.* mapped somatic NFS indels from COSMIC onto protein structures and found that pathogenic variants tend to be enriched in helical and stand regions of protein structures [28]. Niu *et al.* developed a tool to identify 3-dimentional (3D) variants clusters on protein structures that can be used in variant-drug interaction analysis in cancer genomes [29]. Among the mutation-mutation and mutation-drug clusters from more than 4,400 samples across 19 cancer types, more than 6,000 clusters were identified at 3D structure level, including both intra-molecular and inter-molecular clusters. They reported that about 0.76% of the 553,496 somatic variants are indels [29].

Mutations in non-coding regions can also cause diseases [30-37]. Most analyses on non-coding variants in the regulatory regions in cancer genomes either focused on SNVs or did not differentiate SNVs from indels with relatively small sample sizes or a single cancer type [38,39]. Sakthikumar *et al.* investigated non-coding variants in Glioblastoma (GBM) genomes and demonstrated that the GBM somatic variants are enriched in non-coding regions of 78 GBM key genes [40]. Imielinski *et al.* showed that somatic non-coding indels in 79 lung adenocarcinoma genomes are exclusively enriched in surfactant protein genes [41]. Nakagomi *et al.* further analyzed 113 lung cancer samples and reported that other cancer types in lung also harbour non-coding indels and demonstrated the important role of those indels in lung cancer research [42].

While eukaryotic genomes generally are considered to have two major types of sequences: (1) coding-sequences (CDSs) that encode proteins or RNAs, and (2) non-coding sequences that include regulatory regions such as promoters and enhancers for regulation of gene expression, a number of studies have shown that sequences in CDSs and the untranslated regions (UTRs) can also function as enhancers [43-46]. Mutations in coding and UTR enhancers can cause diseases by changing their enhancer activities [47-49]. Recent large-scale studies have shown that transcription factor binding sites (TFBSs) exist in coding regions in both human and mouse genomes based on ChIP-seq data analyses [47,50]. About 15% of codons in the human genome were hypersensitive to DNase I treatment, suggesting the existence of likely dual-use sequences for both amino acid coding and transcriptional regulation [50]. These dual function sequences, termed as duons, were considered to be more conserved than non-duon coding sequences and mutations in these duons could lead to diseases [50,51].

The goal of this study is to investigate the potential role of somatic indels and the overlap between somatic indels and TFBSs in two of the most analyzed cancer types, invasive breast carcinoma (BRCA) and lung adenocarcinoma (LUAD). BRCA has the second largest proportion of indels among 19 cancer types [18,52]. LUAD has a high number of exonic somatic variants as reported in several studies [52,53]. Since the BRCA and LUAD sequences in TCGA data portal (https://portal.gdc.cancer.gov) are whole exome sequences, we focused our indel analysis on coding and the non-coding UTRs. In addition, while in principle the Whole Exome Sequencing (WXS) technology does not produce whole transcripts, studies have shown that 40-60% of the reads from exome sequencing are outside of the designed target regions including introns and these reads can be of high quality [54-57]. Therefore, besides the coding regions and UTRs, we also compared the indels in other regions of the transcripts as a side analysis.

Since somatic indel calling programs also predict germline indels found in healthy genomes [17], we first identified these types of indels and removed them for downstream somatic indel analyses, including structural analysis of the effect of somatic NFS indels on protein secondary structure types, and gene enrichment and conservation analyses. We also mapped the somatic indels on the significantly mutated genes (SMGs) across major cancer types identified by Kandoth *et al.* [58]. More importantly, we investigated the somatic indels on the coding regions and UTRs that overlap with TFBSs, the dual-use sequences. To the best of our knowledge, this is the first large scale comparative study of mapping somatic coding indels to TFBSs.

MATERIALS AND METHODS

Sequence data and somatic indel calling

The 436 BRCA and 564 LUAD whole exome sequencing data, TCGA-BRCA and TCGA-LUAD, were downloaded from TCGA data portal at https://portal.gdc.cancer.gov (dbGaP Study Accession: phs000178.v11.p8). Both tumor and normal blood/tissue sequencing data were used to call somatic indels using the human genome reference GRCh38.p13 and Strelka [59]. Previous studies have demonstrated that Strelka performed well for somatic variants calling [60-62]. The indel set from the GATK Resource bundle with 1,267,008 germline indels was used as the reference of germline indel annotations in healthy human genomes (https://storage.cloud.google.com/genomics-public-data/resources/broad/hg38/v0/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz). The transcript agreed by several references or the longest transcript for each gene was selected for annotating coding sequences and UTRs in both cancer exome sequences and germline sequences. In previous studies, a position $i \pm 5$ has been used to determine whether two indels are the same, without concerning the indel types (insertion or deletion) [63]. Here we used a more stringent approach to identify the germline indels predicted as somatic indels by considering the indel types and insertion/deletion sequences in addition to the indel positions. Two indels are considered the same only if both have the same positions, indel types and sequences when comparing the predicted

somatic cancer indels and the germline indels in the GATK Resource bundle. Since somatic indels are less conserved than the germline indels, we used the following two criteria: 1) the positions of two indels are within $i \pm 5$; and 2) same insertion/deletion type to check the overlap of somatic indels between BRCA and LUAD cancer genomes.

Protein secondary structure type analysis of coding indels

To locate the positions of coding somatic indels, we downloaded all protein coding gene annotations from Ensembl [64]. Each transcript with indel(s) was first searched against proteins with known structures in Protein Data Bank (PDB)[65] using BLAST[66,67]. If a protein has a known structure or highly homologous protein structure (with at least 50% coverage and 80% sequence identity), the secondary structure types of the protein or the template protein were used. The protein secondary structure types of the protein were assigned with DSSP [68]. Of the eight secondary structure types from DSSP, H (α-helix), G (3_{10}-helix) and I (π-helix) states were grouped as helix conformations; E (extended strand) and B (residue in isolated β-bridge) states were grouped as strand conformations and all the remaining states were considered as loop conformations [68]. If no known structures were found in PDB, RaptorX-Property was applied to predict secondary structure types with default settings [69]. RaptorX-Property uses conditional neural fields method to predict secondary structure types and achieves close to 84% Q3 prediction accuracy based on five different datasets [69]. The structural analysis of the germline indels from healthy genomes were performed with the 1,370 coding indels annotated by the GATK Resource bundle.

Overlap of indels with TFBSs and non-coding regions

To investigate the overlap between somatic/germline indels and TFBSs, we used the TFBS set predicted with dePCRM2, a recently developed program for genome scale TFBS prediction with a high sensitivity of more than 97% [70]. A total of 25,297,119 non-overlapping TFBSs were predicted using dePCRM2 with a p-value cutoff of 5×10^{-6}.

Gene enrichment analysis and assignment of conservation scores

To investigate the functional categories of the genes affected by somatic coding indels in BRCA and LUAD, we applied DAVID 6.8 (the Database for Annotation, Visualization and Integrated Discovery) to perform functional enrichment analysis [71]. A cutoff of 0.001 was set for the adjusted p-values with Bonferroni correction to identify the significantly enriched terms in biological process or molecular function.

The phyloP scores of each nucleotide position in the human genome were downloaded from the UCSC Genome Browser database [72,73]. The two flanking nucleotides for each insertion site and the deletion sequences were collected for phyloP distribution analysis as well as for finding genes with high phyloP conservation scores.

RESULTS

Comparison of somatic indels between BRCA and LUAD

We found 109,856 and 91,159 somatic indels from 436 BRCA samples and 564 LUAD samples respectively with 16,909 common indels between them (Table 1). As a reference, a total of 498,938 germline indels were mapped to transcripts in healthy genomes from the GATK Resource bundle. Since the predicted indels by somatic indel prediction algorithms include germline indels (false somatic indels), these germline indels need to be filtered out first for meaningful downstream analysis [17]. As described in the Materials and Methods section, two indels are considered the same only if both have the same positions, indel types and sequences when comparing the germline indels from the GATK Resource bundle and the predicted somatic indels from BRCA and LUAD. We found that 16.74% and 19.64% of indels in BRCA and LUAD respectively are the same as the germline indels (Table 1). After removing the germline and non-transcript indels, 61,543 and 43,684 somatic transcript indels for BRCA and LUAD respectively were used for further analysis. Not surprisingly, the overlapped indels between BRCA and LUAD have a higher percentage of germline indels (23.16%) since germline indels are more conserved than the somatic indels within populations of different cancer types [21].

Similar to germline indels in healthy genomes, relatively more deletions than insertions were found in BRCA and LUAD. The percentages of deletions and insertions in both cancer types are slightly higher than those in the GATK germline indel set (Table 1). The distributions of insertion/deletion in both BRCA and LUAD are significantly different from germline indels (chi-squared test, p-value = 4.532×10^{-16} for BRCA and p-value < 2.2×10^{-16} for LUAD). The number of transcripts that have somatic indels are 14,519 and 13,593 in BRCA and LUAD respectively. It should be noted that while the whole exome sequences from the BRCA and LUAD contain all the coding and UTRs, they do not have the whole transcript sequences as the healthy genomes do. Therefore, at the transcript level, the somatic indels are undercounted.

Somatic coding indels in BRCA and LUAD genomes

As shown in Table 2, the number of transcripts with somatic coding indels and the number of coding indels in both BRCA and LUAD are much higher than those of germline coding indels in healthy genomes (Table 2). The proportions of somatic coding indels are 8.64% in BRCA and 17.89% in LUAD while it is only 0.62% for the germline coding indels in healthy genomes. In terms of the deletion/insertion ratio in coding regions, LUAD has more deletion types (~70%) than that in the germline indels from healthy genomes (64.6%) while about 57.86% of somatic coding indels from BRCA samples are deletions. For the overlapping indels between BRCA and LUAD, the deletion and insertion are about 43.6% and 56.4% respectively.

Coding indels are typically divided into FS and NFS types based on the length of indels. FS indels cause a reading frame shift at the indel site, which are prone to be more deleterious [21,74,75]. Our previous analysis of healthy genomes from the 1000 Genomes Project revealed that the number of

germline FS indels is similar to that of NFS indels [21]. We also observed a similar pattern from germline coding indels in the GATK Resource bundle, 679 FS vs. 691 NFS (Table 2). These results indicate that healthy genomes tend to have much fewer FS coding indels than expected. However, for somatic coding indels in cancer genomes, the number of FS indels is about 2.8 (BRCA) to 4.5 (LUAD) times more than that of NFS indels. Over eighty percent of the somatic coding indels in LUAD are FS indels (Table 2). The overlapped coding somatic indels between BRCA and LUAD genomes have a relatively lower ratio of FS indels (59.4%), but it is still much higher than that in the germline (49.56%).

The somatic coding indels affect a total of 8,286 genes when BRCA and LUAD are combined. BRCA and LUAD have somatic coding indels in 3,979 and 5,458 genes respectively and 798 genes have somatic coding indels in both BRCA and LUAD. Among these genes, MAP3K1 has the most somatic coding indels in BRCA (45 indels), and TP53 has the most somatic coding indels in LUAD (37 indels) (Table 3). We compared the top genes with multiple somatic coding indels in our datasets to the 125 protein coding SMGs among the 127 total SMGs across 12 major cancer types (The other two are one lncRNA gene and one miRNA gene) [58]. Seven and five of the top 10 genes with multiple somatic coding indels in BRCA and LUAD respectively are in the list of 125 protein coding SMGs while none of the top 10 genes with germline coding indels are in the 125 SMGs (Table 3). Some of these genes have served as targets for drug development, such as EGFR [76]. Functional enrichment analysis revealed that the genes with somatic coding indels in LUAD are highly enriched in biological processes involved in cell adhesion while the indels in BRCA affect more chromatin remodelling and transcription (Supplementary Table S1).

Since NFS somatic coding indels only affect part of the protein while keeping the remaining sequence unchanged, we compared the distributions of secondary structure types of these indels with the germline indels in healthy genomes. Among the proteins with NFS somatic coding indels, 181 and 106 proteins in BRCA and LUAD respectively were found to have known or homologous structures in PDB. For proteins having NFS somatic coding indels without known structures, we used RaptorX-Property to predict the secondary structure types for each amino acid of the indels, as described in the Materials and Methods section. While the distributions between the two cancer types are slightly different (chi-square test, p-value= 0.003), both are significantly different from that in germline indels (Figure 1). Somatic coding NFS indels in cancer genomes have more helix and strand conformations with fewer loop types (chi-square test, p-values $<2.2 \times 10^{-16}$), suggesting the NFS coding indels in both BRCA and LUAD cancer genomes affect more core secondary structures in the encoded proteins and are potentially more deleterious than the germline NFS coding indels in healthy genomes.

Somatic non-coding UTR indels in BRCA and LUAD cancer genomes

For exonic non-coding somatic indels, we found 372 and 1,940 indels in 5'UTR and 3'UTR respectively in BRCA, and 375 and 1,187 indels in 5'UTR and 3'UTR respectively in LUAD (Supplementary Table S2). There are more somatic indels in 3'UTR than those in 5'UTR. In both BRCA and LUAD genomes, the indels are enriched in both 5'UTR (0.66% and 1.06% for BRCA and

LUAD, respectively) and 3'UTR (3.45% and 3.31% for BRAC and LUAD, respectively) when compared with those in germline indels of healthy genomes, with 0.18% and 2.55% in the 5'UTR and 3'UTR respectively (Supplementary Table S2). The majority of transcript somatic indels are located in the non-CDS, non-UTR regions. Therefore, even though the goal of whole exome sequencing is to get the exonic sequences, exome sequencing can generate high quality data and cover large non-target regions [54,55]. However, since the coverage of non-target regions in each cancer sample might be different from whole exome sequencing, it is difficult to draw conclusions when comparing the non-CDS, non-UTR noncoding transcript indels between two different cancer types and between cancer somatic indels and the germline indels.

Conservation analysis of somatic CDS and UTR indels

It is interesting to see how conserved the somatic CDS and UTR indel sequences are in BRCA and LUAD when compared with germline CDS and UTR indels in healthy genomes. To this end, we compared the phyloP scores for nucleotides at the indel positions [72]. Since the phyloP scores of nucleotides are based on the reference genome, we collected the nucleotides for insertions and deletions differently. For insertion cases, the phyloP scores of the two flanking nucleotides at the indel site were collected while phyloP scores for the whole deletion sequences were considered (see Materials and Methods). The larger a positive phyloP score of a nucleotide position in the genome, the more conserved of the position. Figure 2 shows the distributions of phyloP scores of insertions and deletions in CDS regions (Figure 2 a& b) and UTR (Figure 2 c&d) in BRCA (Figure 2 a&c) and LUAD (Figure 2 b&d) respectively. In CDS regions, the somatic insertions and deletions have more positions with higher phyloP scores when compared with the distribution of germline insertions and deletions. There seems no apparent differences between BRCA and LUAD as well as between insertion and deletion cases. As for UTR indels, there is a difference between the cancer indels and germline indels. However, the differences are very small, especially when compared with those in the CDS positions.

Overlap between somatic indels and TFBSs

The percent overlap between the somatic indels in cancer transcripts and TFBSs is larger than that between germline transcript indels and TFBSs (Table 4). The number of somatic CDS indels that overlap with TFBSs is much higher in cancer genomes, 2,367 and 3,140 in BRAC and LUAD respectively while there are only 520 for germline coding indels, suggesting cancer coding indels are enriched in these dual-functional regions. Somatic non-CDS transcript indels in cancer genomes are also enriched in the predicted TFBS sequences (25.4% and 27.01% for BRCA and LUAD, respectively) when compared to 17.41% in germline non-CDS transcript indels in healthy genomes (Table 4). A detailed look at these non-CDS transcript indels shows that there is a smaller percentage of overlap between 5'UTR and TFBSs in BRCA and LUAD than that in healthy genomes while the 3'UTR is the opposite (Supplementary Table S3). While we also found that other non-CDS, non-UTR transcript indels have a larger percent overlap with TFBSs in cancer exomes, unlike the CDS cases,

incomplete transcript sequences in the intron regions from exome sequencing makes it harder to make a fair comparison with the germline cases.

We also performed conservation enrichment analysis with a phyloP score cutoff of 5 for indel positions in CDS with TFBS overlap and indels positions in CDS without TFBS overlap. The genes were ranked by the number of indel positions with phyloP scores above the cutoff in each case. The top 10 genes in each case are listed in Table 5. More SMG genes were found in indels with CDS and TFBS overlap than those CDS indels without TFBS overlap in both cancer types (7 vs. 4 in BRAC and 4 vs. 1 in LUAD) (Table 5). Not surprisingly, none of the 125 SMG protein coding genes were found in the germline indels no matter if CDS overlaps TFBS or not.

Somatic indels in SMGs

We mapped the somatic indels to the annotated 125 protein coding SMGs and found that somatic indels in cancer genomes are enriched in SMGs, especially they are enriched in SMG's coding regions in both BRCA and LUAD cancer genomes when compared with the germline indels from healthy genomes (Table 6). In healthy genomes, there are only 11 (0.23%) SMG indels in coding regions, but in BRCA and LUAD cancer genomes, 349 (33.82%) and 267 (38.98%) of SMG somatic indels are located in the coding regions respectively (Table 6). Among the 125 SMGs, 70 and 71 of them have BRCA and LUAD somatic indels in CDS regions respectively while only 9 SMGs have germline coding indels. Twelve SMGs have somatic coding indels in both cancer types, suggesting different mutation/variant patterns in different cancer types while there are some commonalities between cancer types.

The overlap between SMG somatic coding indels with TFBSs is significantly more in BRCA and LUAD than that in germline indels in healthy genomes (Table 6). There are 172 (49.28%) and 132 (49.44%) somatic coding indels overlap with TFBSs in SMGs in BRCA and LUAD while there are only 5 (45.45%) such cases in healthy genomes (Table 6). The overlap between the non-CDS somatic indels and TFBSs in BRCA (22.55%) and LUAD (25.12%) is higher than that in healthy genomes (12.9%) as well.

DISCUSSION

With the advancement of biotechnology, especially the NGS technology, a large number of genomes have been sequenced for a variety of cancer types. Somatic variations in cancer genomes have been one of the main focuses in cancer studies, including variants in both coding and non-coding regions [77]. However, most of the studies in cancer genomes focused on SNVs [29,40,52]. In this study, we carried out a comparative study of the somatic indels in two major cancer types, BRCA and LUAD with their whole exome sequences and compared some of the features with germline indels from healthy genomes.

There are several novel aspects from this study. First, we removed the germline indels predicted from the somatic indels calling program before performing downstream analyses. We demonstrated

previously that some of the predicted somatic indels are exactly the same as the germline indels in healthy human genomes [17]. Therefore, these indels are considered as false somatic indels and represent "noise" when analysing features in cancer genomes, which need to be filtered out. Secondly, we investigated the overlap between cancer somatic indels, especially the coding indels with TFBSs. Previous case studies as well as large-scale analyses revealed the existence of the so-called duons that encode amino acids and also serve as TFBSs [43-47,50]. The percentage of such DNA sequences with dual functions varies in species and by different TFBSs annotations. Based on ChIP-seq data, Birnbaym *et al.* showed that there are 7% and 6% of binding peaks located in protein coding regions in human genomes and mouse genomes respectively [47]. Using DNase I footprinting method, Stergachis *et al.* found that at least 14% of coding regions in human genomes can bind transcription factors [50]. The prevalence of such sequences and their implication in diseases suggest their important roles in human genomes and diseases [47-49,51]. Third, we compared the conservation score distributions of CDS and UTR indels between cancer genomes and healthy genomes. Finally, we assessed the structural effects of the coding somatic NFS indels and investigated somatic indels in the 125 SMGs identified from different cancer types.

The somatic indels from different cancer types vary greatly. As shown in Table 1, only 9,988 somatic indels appear in both cancer types, which account for 16.23% of BRCA and 22.86% of LUAD somatic indels respectively. The somatic transcript indels in two cancer types have different proportion of indel types. LUAD has more deletions, more indels in coding regions and more FS indels, than the BRCA cancer type. In our datasets, we did not find any complex indels, which are formed by simultaneously deleting and inserting DNA fragments of different sizes at a common genomic location [78]. Our data on somatic coding indels revealed a number of top SMGs with most indel variations in BRCA and LUAD. Except for TP53, other top 10 mutated genes are different between BRAC and LUAD, suggestion involvement of different gene mutations in different cancer types (Table 3). Functional enrichment analysis also shows different biological processes involved in each type of cancer (Tables 1 and 2, Supplementary Table S1).

Compared with germline transcript indels in healthy genomes, somatic transcript indels in cancer genomes have higher proportions involved in the CDS regions. Coding somatic indels also have a higher rate of FS types, especially in SMGs (Tables 2 and 6). This phenomenon is not surprising since FS indels are prone to be deleterious [21,74,75]. More importantly, the somatic coding indels are more likely to be enriched in the structurally and functionally important regions of proteins than the germline indels in the heathy genomes. First of all, we found that the NFS somatic indels in BRCA and LUAD are enriched in helical and strand secondary structure types (Figure 1). Helices and strands represent the core of protein structures. Changes in the core would more likely affect the stability of the protein and disrupt the structure, which in turn affect the function of the protein. Secondly, the somatic coding indels are enriched in coding regions that are also predicted as TFBSs, or duons (Tables 4 and 6). Therefore, these indels not only affect the protein sequences, they can also change the regulation of gene expression. In addition, compared to germline indels, somatic

CDS indels are enriched in positions that have high conservation score based on phyloP analyses, suggesting these indels are more deleterious.

While the cancer whole exome sequencing data have all the coding and UTR sequences that can be compared directly with the germline coding and UTR sequences in healthy genomes, one of the limitations of the whole exome sequences is that they only have partial non-coding sequences for the transcripts. It would be interesting to see the differences in all the non-coding regions among different cancer types and between germline indels and cancer somatic indels from a large-scale comparative analysis. More detailed analyses on structural and functional effect can be carried out in the future to investigate the structural basis for better understanding these somatic indels as previous work done on point mutations[79-82] and if a somatic indel is deleterious [27].

FUNDING

This work was supported by the National Science Foundation [DBI-2051491 to J.G]; and National Institutes of Health [R15GM132846 to J.G]. Funding for open access charge: National Science Foundation.

REFERENCES

1 Consortium, G. P. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010).
2 Mills, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res* **16**, 1182-1190, doi:10.1101/gr.4565806 (2006).
3 Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-934 (2001).
4 Gibbs, R. A. *et al.* The international HapMap project. (2003).
5 Consortium, I. H. A haplotype map of the human genome. *Nature* **437**, 1299 (2005).
6 Weber, J. L. *et al.* Human diallelic insertion/deletion polymorphisms. *The American Journal of Human Genetics* **71**, 854-862 (2002).
7 Bhangale, T. R., Rieder, M. J., Livingston, R. J. & Nickerson, D. A. Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. *Human molecular genetics* **14**, 59-69 (2005).
8 Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature genetics* **36**, 949-951 (2004).
9 Berger, J. *et al.* Genetic mapping with SNP markers in Drosophila. *Nature genetics* **29**, 475-481 (2001).
10 Wicks, S. R., Yeh, R. T., Gish, W. R., Waterston, R. H. & Plasterk, R. H. Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map. *Nature genetics* **28**, 160-164 (2001).
11 Dawson, E. *et al.* A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. *Genome research* **11**, 170-178 (2001).
12 Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Human molecular genetics* **19**, R131-R136 (2010).
13 Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome research* **23**, 749-761 (2013).
14 Chuzhanova, N. A., Anassis, E. J., Ball, E. V., Krawczak, M. & Cooper, D. N. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. *Human mutation* **21**, 28-44 (2003).

15 Collins, F. S. *et al.* Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* **235**, 1046-1049 (1987).
16 Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome research* **21**, 830-839 (2011).
17 Chen, J. & Guo, J.-t. Comparative assessments of indel annotations in healthy and cancer genomes with next-generation sequencing data. *BMC medical genomics* **13**, 1-11 (2020).
18 Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *The lancet oncology* **18**, 1009-1021 (2017).
19 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101, doi:10.1038/s41586-020-1943-3 (2020).
20 Consortium, I. T. P.-C. A. o. W. G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93, doi:10.1038/s41586-020-1969-6 (2020).
21 Lin, M. *et al.* Effects of short indels on protein structure and function in human genomes. *Scientific reports* **7**, 1-9 (2017).
22 Folkman, L. *et al.* DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. *Bioinformatics* **31**, 1599-1606, doi:10.1093/bioinformatics/btu862 (2015).
23 Zhao, H. *et al.* DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biology* **14**, R23, doi:10.1186/gb-2013-14-3-r23 (2013).
24 Hu, J. & Ng, P. C. Predicting the effects of frameshifting indels. *Genome Biology* **13**, R9, doi:10.1186/gb-2012-13-2-r9 (2012).
25 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).
26 Yue, Z., Zhao, L., Cheng, N., Yan, H. & Xia, J. dbCID: a manually curated resource for exploring the driver indels in human cancer. *Brief Bioinform* **20**, 1925-1933, doi:10.1093/bib/bby059 (2019).
27 Yue, Z., Chu, X. & Xia, J. PredCID: prediction of driver frameshift indels in human cancer. *Brief Bioinform* **22**, doi:10.1093/bib/bbaa119 (2021).
28 Pagel, K. A. *et al.* Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. *PLoS computational biology* **15**, e1007112 (2019).
29 Niu, B. *et al.* Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nature genetics* **48**, 827-837 (2016).
30 Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Human molecular genetics* **24**, R102-R110 (2015).
31 Arking, D. E. *et al.* A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nature genetics* **38**, 644-651 (2006).
32 Kapoor, A. *et al.* An enhancer polymorphism at the cardiomyocyte intercalated disc protein NOS1AP locus is a major regulator of the QT interval. *The American Journal of Human Genetics* **94**, 854-869 (2014).
33 Spieler, D. *et al.* Restless legs syndrome-associated intronic common variant in Meis1 alters enhancer function in the developing telencephalon. *Genome research* **24**, 592-603 (2014).
34 Bauer, D. E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253-257 (2013).
35 Stadhouders, R. *et al.* HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *The Journal of clinical investigation* **124**, 1699-1710 (2014).
36 Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108, doi:10.1038/nrg.2015.17 (2016).
37 Fujimoto, A. *et al.* Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet* **48**, 500-509, doi:10.1038/ng.3547 (2016).
38 Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat Genet* **47**, 710-716, doi:10.1038/ng.3332 (2015).
39 Capasso, M. *et al.* Transcription Factors Involved in Tumorigenesis Are Over-Represented in Mutated Active DNA-Binding Sites in Neuroblastoma. *Cancer Res* **80**, 382-393, doi:10.1158/0008-5472.CAN-19-2883 (2020).
40 Sakthikumar, S. *et al.* Whole-genome sequencing of glioblastoma reveals enrichment of non-coding constraint mutations in known and novel genes. *Genome Biology* **21**, 1-22 (2020).
41 Imielinski, M., Guo, G. & Meyerson, M. Insertions and deletions target lineage-defining genes in human cancers. *Cell* **168**, 460-472. e414 (2017).
42 Nakagomi, T. *et al.* Clinical implications of noncoding indels in the surfactant-encoding genes

in lung cancer. *Cancers* **11**, 552 (2019).
43 Neznanov, N., Umezawa, A. & Oshima, R. G. A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice. *J Biol Chem* **272**, 27549-27557, doi:10.1074/jbc.272.44.27549 (1997).
44 Tumpel, S., Cambronero, F., Sims, C., Krumlauf, R. & Wiedemann, L. M. A regulatory module embedded in the coding region of Hoxa2 controls expression in rhombomere 2. *Proc Natl Acad Sci U S A* **105**, 20077-20082, doi:10.1073/pnas.0806360105 (2008).
45 Dong, X. *et al.* Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. *Nucleic Acids Res* **38**, 1071-1085, doi:10.1093/nar/gkp1124 (2010).
46 Birnbaum, K., Benfey, P. N. & Shasha, D. E. cis element/transcription factor analysis (cis/TF): a method for discovering transcription factor/cis element relationships. *Genome Res* **11**, 1567-1573 (2001).
47 Birnbaum, R. Y. *et al.* Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation. *PLoS Genet* **10**, e1004592 (2014).
48 Bhatia, S. & Kleinjan, D. A. Disruption of long-range gene regulation in human genetic disease: a kaleidoscope of general principles, diverse mechanisms and unique phenotypic consequences. *Hum Genet* **133**, 815-845, doi:10.1007/s00439-014-1424-6 (2014).
49 Lango Allen, H. *et al.* Next generation sequencing of chromosomal rearrangements in patients with split-hand/split-foot malformation provides evidence for DYNC1I1 exonic enhancers of DLX5/6 expression in humans. *J Med Genet* **51**, 264-267, doi:10.1136/jmedgenet-2013-102142 (2014).
50 Stergachis, A. B. *et al.* Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342**, 1367-1372 (2013).
51 Yadav, V. K., Smith, K. S., Flinders, C., Mumenthaler, S. M. & De, S. Significance of duon mutations in cancer genomes. *Sci Rep* **6**, 27437, doi:10.1038/srep27437 (2016).
52 Peterson, T. A., Gauran, I. I. M., Park, J., Park, D. & Kann, M. G. Oncodomains: A protein domain-centric framework for analyzing rare variants in tumor samples. *PLoS computational biology* **13**, e1005428 (2017).
53 Yang, F. *et al.* Protein domain-level landscape of cancer-type-specific somatic mutations. *PLoS Comput Biol* **11**, e1004147 (2015).
54 Guo, Y. *et al.* Exome sequencing generates high quality data in non-target regions. *BMC Genomics* **13**, 194, doi:10.1186/1471-2164-13-194 (2012).
55 Samuels, D. C. *et al.* Finding the lost treasures in exome sequencing data. *Trends Genet* **29**, 593-599, doi:10.1016/j.tig.2013.07.006 (2013).
56 Sulonen, A. M. *et al.* Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* **12**, R94, doi:10.1186/gb-2011-12-9-r94 (2011).
57 Asan *et al.* Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol* **12**, R95, doi:10.1186/gb-2011-12-9-r95 (2011).
58 Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339 (2013).
59 Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811-1817, doi:10.1093/bioinformatics/bts271 (2012).
60 Cai, L., Yuan, W., Zhang, Z., He, L. & Chou, K. C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci Rep* **6**, 36540, doi:10.1038/srep36540 (2016).
61 Roberts, N. D. *et al.* A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics* **29**, 2223-2230, doi:10.1093/bioinformatics/btt375 (2013).
62 Xu, H., DiCarlo, J., Satya, R. V., Peng, Q. & Wang, Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* **15**, 244, doi:10.1186/1471-2164-15-244 (2014).
63 Hasan, M. S., Wu, X. & Zhang, L. Performance evaluation of indel calling tools using real short-read data. *Human genomics* **9**, 20, doi:10.1186/s40246-015-0042-2 (2015).
64 Yates, A. D. *et al.* Ensembl 2020. *Nucleic acids research* **48**, D682-D688 (2020).
65 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).
66 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
67 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
68 Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of

hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* **22**, 2577-2637 (1983).
69 Wang, S., Peng, J., Ma, J. & Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep* **6**, 18962, doi:10.1038/srep18962 (2016).
70 Ni, P. & Su, Z. Accurate prediction of cis-regulatory modules reveals a prevalent regulatory genome of humans. *NAR Genom Bioinform* **3**, lqab052, doi:10.1093/nargab/lqab052 (2021).
71 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).
72 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121, doi:10.1101/gr.097857.109 (2010).
73 Casper, J. *et al.* The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res* **46**, D762-D769, doi:10.1093/nar/gkx1020 (2018).
74 de la Chaux, N., Messer, P. W. & Arndt, P. F. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC evolutionary biology* **7**, 191 (2007).
75 Taylor, M. S., Ponting, C. P. & Copley, R. R. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome Research* **14**, 555-566 (2004).
76 Gerber, D. E. EGFR Inhibition in the Treatment of Non-Small Cell Lung Cancer. *Drug Dev Res* **69**, 359-372, doi:10.1002/ddr.20268 (2008).
77 Vogelstein, B. *et al.* Cancer genome landscapes. *science* **339**, 1546-1558 (2013).
78 Ye, K. *et al.* Systematic discovery of complex insertions and deletions in human cancers. *Nature medicine* **22**, 97-104, doi:10.1038/nm.4002 (2016).
79 Rajendran, V., Gopalakrishnan, C. & Purohit, R. Impact of point mutation P29S in RAC1 on tumorigenesis. *Tumour Biol* **37**, 15293-15304, doi:10.1007/s13277-016-5329-y (2016).
80 Joerger, A. C., Ang, H. C. & Fersht, A. R. Structural basis for understanding oncogenic p53 mutations and designing rescue drugs. *Proc Natl Acad Sci U S A* **103**, 15056-15061, doi:10.1073/pnas.0607286103 (2006).
81 Shi, Z. & Moult, J. Structural and functional impact of cancer-related missense somatic mutations. *J Mol Biol* **413**, 495-512, doi:10.1016/j.jmb.2011.06.046 (2011).
82 Malhotra, S. *et al.* Understanding the impacts of missense mutations on structures and functions of human cancer-related genes: A preliminary computational analysis of the COSMIC Cancer Gene Census. *PLoS One* **14**, e0219935, doi:10.1371/journal.pone.0219935 (2019).

Author contributions

JTG conceived the study and designed the experiment. JC carried out the experiments and performed data analysis. JTG and JC wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Data availability

The data used in this study were downloaded from the TCGA data portal at https://portal.gdc.cancer.gov (dbGaP Study Accession: phs000178.v11.p8).

Figure Legends:

Figure 1. Distribution of secondary structure types of somatic NFS indels and germline NFS indels.

Figure 2. Distributions of the phyloP scores in somatic and germline CDS and UTR indels. A: BRCA CDS indels; B: LUAD CDS indels; C: BRCA UTR indels; and D: LUAD UTR indels. Blue is for somatic indels and red is for germline indels. The dashed line represents insertions and the solid line represents deletions.

Tables

Table 1: Somatic transcript indels in BRCA and LUAD

Cancer type	# of total indels	Overlap with germline indels	# of somatic transcript indels*	Deletions	Insertions	# of transcripts with indels
BRCA	109,856	18,391 (16.74%)	61,543	36,109 (58.67%)	25,434 (41.33%)	14,519
LUAD	91,159	17,900 (19.64%)	43,684	27,148 (62.15%)	16,536 (37.85%)	13,593
BRCA ∩ LUAD	16,909	3,916 (23.16%)	9,988	5,330 (53.36%)	4,658 (46.64%)	6,600
Germline	1,267,008	-	498,938	284,597 (57.04%)	214,341 (42.96%)	17,278

*The numbers of somatic transcript indels in BRCA and LUAD are indels on transcripts after removing the ones that overlap with germline indels.

Table 2: Somatic coding (CDS) indels in BRCA and LUAD

Cancer type	# of transcripts with CDS indels	CDS indels*	Deletions	Insertions	FS indels	NFS indels
BRCA	3,979	5,320 (8.64%)	3,078 (57.86%)	2,242 (42.14%)	3,947 (74.19%)	1,373 (25.81%)
LUAD	5,458	7,813 (17.89%)	5,526 (70.73%)	2,287 (29.27%)	6,387 (81.75%)	1,426 (18.25%)
BRCA ∩ LUAD	798	835 (8.36%)	364 (43.59%)	471 (56.41%)	496 (59.40%)	339 (40.60%)
Germline	1,180	1,370 (0.62%)	885 (64.60%)	485 (35.40%)	679 (49.56%)	691 (50.44%)

*The percentages are calculated against the number of transcript indels.

Table 3: Top ten genes with multiple somatic CDS indels in BRCA and LUAD

BRCA*	LUAD*	BRCA ∩ LUAD*	Germline
MAP3K1	**TP53**	**TP53**	SSPOP
GATA3	**STK11**	PABPC3	HLA-DRB1
TP53	TTN	MUC5B	TEKT4
CDH1	MUC16	HAVCR1	OR4C5
DSPP	**KEAP1**	PABPC1	SCYGR8
KMT2C	RBM10	ACIN1	MAML3
PIK3R1	RYR2	ZFHX4	ZFPM1
SPEN	CSMD3	**EPHB6**	ABCA10
TBX3	**NF1**	ZAN	KRT14
TTN	**EGFR**	FAM71D	MYO15B

*The genes are ranked by the number of somatic CDS indels and the genes in bold are the ones in the 125 SMG list.

Table 4: Somatic transcript indels overlapping with TFBSs

Cancer type	Transcript indels	Overlapping with TFBSs	CDS indels	CDS indels overlapping with TFBSs	Non-CDS Transcript indels	Non-CDS indels overlapping with TFBSs
BRCA	61,543	16,646 (27.05%)	5,320	2,367 (44.49%)	56,223	14,279 (25.40%)
LUAD	43,684	12,830 (29.37%)	7,813	3,140 (40.19%)	35,871	9,690 (27.01%)
BRCA ∩ LUAD	9,988	2,977 (29.81%)	835	332 (39.76%)	9,153	2,645 (28.90%)
Germline	498,938	87,156 (17.47%)	1,370	520 (37.96%)	497,568	85,465 (17.18%)

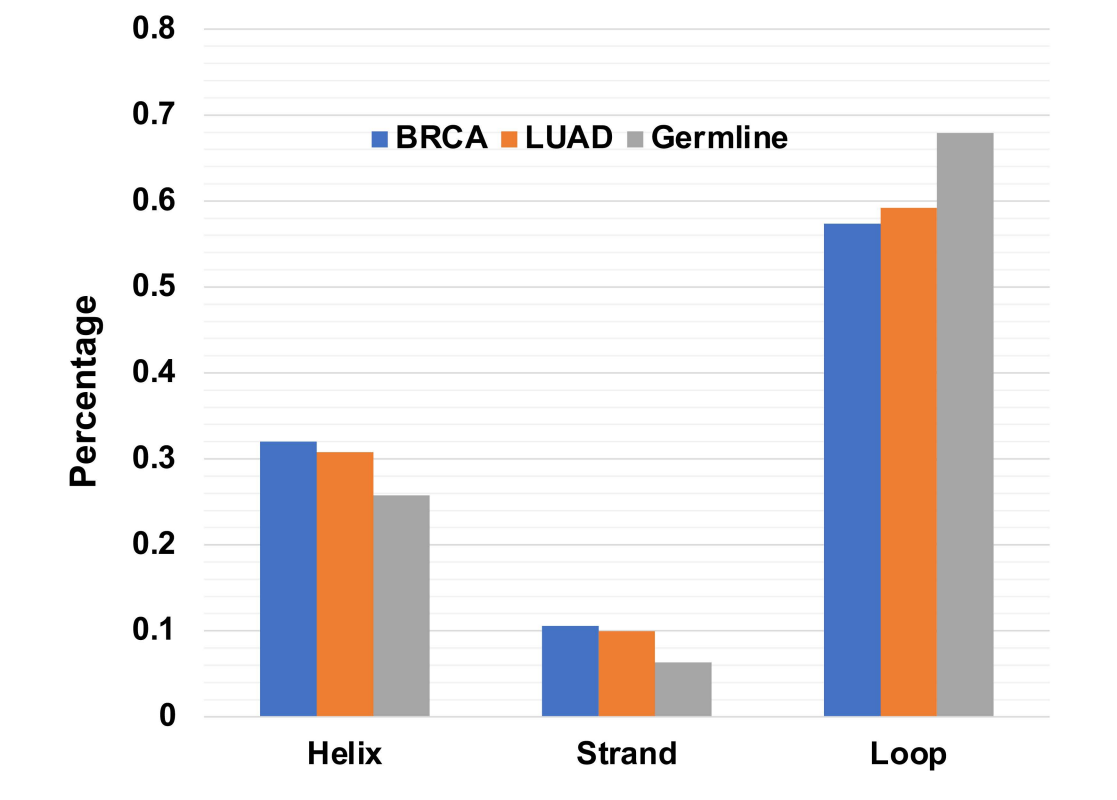
Table 5: Top ten genes with multiple high phyloP scores (>5) in somatic CDS indels

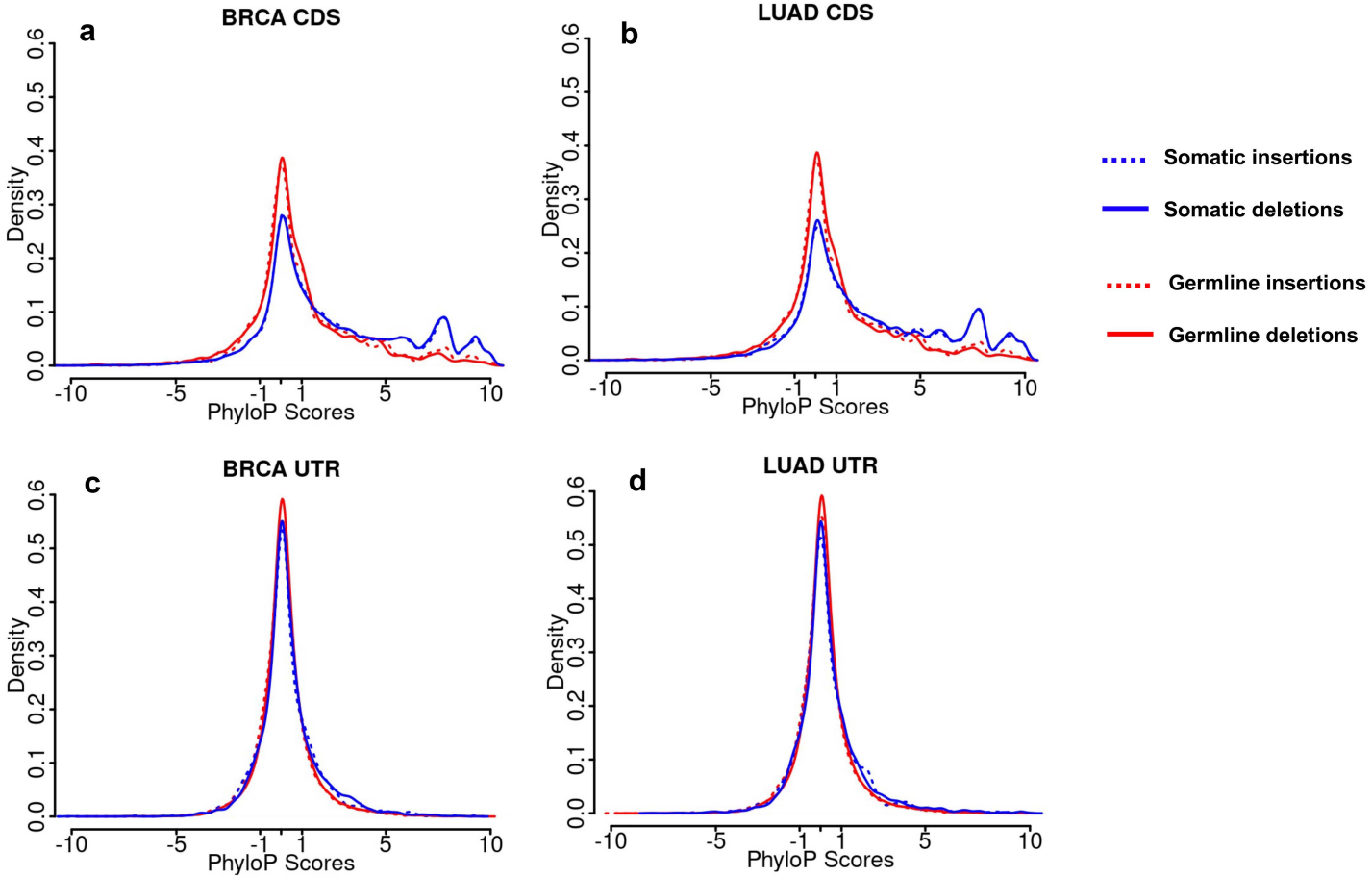
BRCA		LUAD		Germline	
CDS overlap with TFBS	CDS not overlap with TFBS	CDS overlap with TFBS	CDS not overlap with TFBS	CDS overlap with TFBS	CDS not overlap with TFBS
GATA3	**PIK3CA**	**STK11**	ADGRL3	LZTR1	RFX7
MAP3K1	**PIK3R1**	**TP53**	CDH8	ZEB2	CLTCL1
TP53	**MAP3K1**	**EGFR**	LRFN5	CDK8	RBBP6
PTEN	**PTEN**	TAF2	ATP2B1	GJB7	CHD9
YTHDF2	TTN	**APC**	**RPL5**	DBX1	NEDD4
CDH1	TM9SF2	PCDH9	KCNH7	GMNC	RERE
TM9SF4	PLCE1	MEAF6	CNOT1	OR5AU1	CARD11
TBX3	EPHB3	TUBB8B	DHX9	SRRM3	HYDIN
TTN	ADCYAP1R1	DOCK5	PSEN2	ZNF730	TMCC1
RUNX1	PDE11A	TAPT1	COG2	SPON1	DNAJC28

*The genes in bold are among the125 SMG list.

Table 6: Somatic transcript indels in 125 SMGs of BRCA and LUAD

Cancer type	Indels in SMGs	SMGs with indels in CDS regions	Indels in SMGs' CDS	SMG CDS indels overlapping with TFBSs	Indels in SMGs' non-CDS regions	SMG non-CDS indels overlap with TFBSs
BRCA	1,032	70 (56.00%)	349 (33.82%)	172 (49.28%)	683	154 (22.55%)
LUAD	685	71 (56.80%)	267 (38.98%)	132 (49.44%)	418	105 (25.12%)
BRCA ∩ LUAD	129	12 (9.6%)	19 (14.73%)	9 (47.37%)	110	35 (31.82%)
Germline	4,818	9 (7.2%)	11 (0.23%)	5 (45.45%)	4,807	620 (12.90%)


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Percentage
BRCA
LUAD
Germline
Helix
Strand
Loop


a
BRCA CDS
b
LUAD CDS
c
BRCA UTR
d
LUAD UTR
Density
PhyloP Scores
Somatic insertions
Somatic deletions
Germline insertions
Germline deletions

Supplementary Information

Structural and functional analysis of somatic coding and UTR indels in breast and lung cancer genomes

Jing Chen and Jun-tao Guo*

Table S1. Functional enrichment analysis for genes with CDS indels

A: LUAD

Biological Process		**Molecular Function**	
homophilic cell adhesion via plasma membrane adhesion molecules	9.50E-12	calcium ion binding	5.20E-06
cell adhesion	5.50E-09	ATP binding	3.50E-05
detection of chemical stimulus involved in sensory perception of smell	2.90E-03	RNA polymerase II core promoter proximal region sequence-specific DNA binding	2.90E-04
membrane depolarization during action potential	2.90E-03	metal ion binding	2.90E-04
extracellular matrix organization	3.20E-03	chromatin binding	2.90E-04
regulation of ion transmembrane transport	8.40E-03	G-protein coupled receptor activity	9.30E-04
synapse assembly	8.40E-03	olfactory receptor activity	9.40E-04
transcription, DNA-templated	8.40E-03	extracellular-glutamate-gated ion channel activity	3.40E-03
axon guidance	8.40E-03	extracellular matrix structural constituent	3.40E-03
		ionotropic glutamate receptor activity	7.60E-03

B: BRCA

Biological Process		**Molecular Function**	
chromatin remodeling	2.30E-03	chromatin binding	4.60E-09
transcription, DNA-templated	2.70E-03	DNA binding	3.70E-04
		ATPase activity	5.40E-04
		transcription regulatory region DNA binding	3.30E-03
		actin binding	3.30E-03
		calcium ion binding	5.70E-03

Table S2: Somatic non-CDS transcript indels in BRCA and LUAD

Cancer type	# of non-CDS transcript indels	Indels in 5'UTR	Indels in 3'UTR	Indels in other non-CDS transcript regions
BRCA	56,223	372 (0.66%)	1,940 (3.45%)	53,912 (95.89%)
LUAD	35,871	375 (1.06%)	1,187 (3.31%)	34,309 (95.65%)
BRCA ∩ LUAD	9,153	89 (0.97%)	345 (3.76%)	8,719 (95.26%)
Germline	497,568	879 (0.18%)	12,674 (2.55%)	484,016 (97.28%)

Table S3: Somatic non-CDS transcript indels overlapping with TFBSs

Cancer type	Indels in 5'UTR	5'UTR indels overlapping with TFBSs	Indels in 3'UTR	3'UTR indels overlapping with TFBSs	Indels in other non-CDS transcript regions	Other non-CDS transcript indels overlapping with TFBSs
BRCA	372	175 (47.04%)	1,940	713 (36.75%)	53,912	13,391 (24.84%)
LUAD	375	184 (49.07%)	1,187	465 (39.17%)	34,309	9,041 (26.35%)
BRCA ∩ LUAD	89	41 (46.07%)	345	134 (38.84%)	8,719	2,470 (28.33%)
Germline	879	463 (52.67%)	12,674	4,164 (32.85%)	484,016	82,009 (16.94%)