


Article

Inferring Potential Cancer Driving Synonymous Variants

Zishuo Zeng ^{1,*}  and Yana Bromberg ^{1,2,*}¹ Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08873, USA² Department of Genetics, Rutgers University, Piscataway, NJ 08854, USA

* Correspondence: zzeng@bromberglab.org (Z.Z.); yana@bromberglab.org (Y.B.)

Abstract: Synonymous single nucleotide variants (sSNVs) are often considered functionally silent, but a few cases of cancer-causing sSNVs have been reported. From available databases, we collected four categories of sSNVs: germline, somatic in normal tissues, somatic in cancerous tissues, and putative cancer drivers. We found that screening sSNVs for recurrence among patients, conservation of the affected genomic position, and synVep prediction (synVep is a machine learning-based sSNV effect predictor) recovers cancer driver variants (termed *proposed drivers*) and previously unknown putative cancer genes. Of the 2.9 million somatic sSNVs found in the COSMIC database, we identified 2111 proposed cancer driver sSNVs. Of these, 326 sSNVs could be further tagged for possible RNA splicing effects, RNA structural changes, and affected RBP motifs. This list of proposed cancer driver sSNVs provides computational guidance in prioritizing the experimental evaluation of synonymous mutations found in cancers. Furthermore, our list of novel potential cancer genes, galvanized by synonymous mutations, may highlight yet unexplored cancer mechanisms.

Keywords: synonymous variants; sSNV; cancer drivers; somatic variants; variant functional impact



Citation: Zeng, Z.; Bromberg, Y. Inferring Potential Cancer Driving Synonymous Variants. *Genes* **2022**, *13*, 778. <https://doi.org/10.3390/genes13050778>

Academic Editors: Stefania Bortoluzzi, Piero Fariselli and Anelia D. Horvath

Received: 25 March 2022

Accepted: 26 April 2022

Published: 27 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Despite many years of concerted research efforts, cancer remains a major public health challenge with 19.3 million new cases and 10 million deaths worldwide in 2020 alone [1]. On the molecular level, cancer is caused by genetic variation, whether inherited or acquired via chance mutation, infection, or environmental exposure to toxins or ionizing radiation [2–4]. These changes result in aberrant and uncontrolled cell growth—a cancer hallmark [5].

Genetic mutations found in cancerous tissues can be designated as drivers or passengers [6]. Driver mutations are selectively advantageous to cancer development and growth (carcinogenesis), whereas passenger mutations are “by-products” of the carcinogenesis process. It is estimated that each tumor contains four or five driver mutations [7], while the vast majority of the remaining variants are passengers [8]. Differentiating driver mutations from passenger mutations remains an unsolved problem in cancer biology [9]. Identification of drivers typically involves multiple steps: identifying variants recurrent in different cancer samples, predicting the functional impact of these variants, and inspecting the variants’ underlying pathways and interaction networks—all in addition to experimental validation [10].

Cancer drivers range in size and effect from SNVs and small InDels (insertion or deletion of a few nucleotides) to genome rearrangement and copy number variation [11]. According to the International Cancer Genome Consortium (ICGC) data portal (<https://dcc.icgc.org/> (accessed on 2 March 2022)) [12], the vast majority (>91%) of mutations found in cancer tumor samples are SNVs. OncoVar (ONCOgenic driver VARIants, <https://oncovar.org/> (accessed on 2 March 2022)) is a recently developed database containing 20,162 cancer driver (missense, stop-gain, and stop-loss) mutations spanning 814 genes and 33 cancer types [13]. Note that SNVs located in the protein coding region may have different consequences: mutation that change the corresponding protein sequence are

known as missense (or non-synonymous—nsSNV) variants, while those that, due to codon degeneracy, do not affect the protein sequence are synonymous (sSNV). The role of sSNVs in cancer is often overlooked [14] as for OncoVar and other databases (e.g., ICGC data portal [12]). However, sSNVs can have a variety of functional impacts on biological functionality (e.g., transcription, splicing, cotranslational folding) [15] and thus may also be cancer drivers. Supek et al. estimated that synonymous variants account for 6–8% of all SNV driver mutations in oncogenes [16]. In fact, multiple sSNVs in various genes and cancer types have been recognized as drivers, e.g., variants in BCL2L12/melanoma [17], VHL/hemangioblastoma [18], and BAP1/clear-cell renal cell carcinoma [19].

Here, we evaluated the effects of sSNVs from four categories: germline mutations, somatic mutations found in normal tissues, cancer somatic mutations, and putative cancer driver mutations. Based on the comparisons of variant effect predictions in these four categories of sSNVs, we demonstrated the utility of synVep [20], a machine learning-based method for sSNV effect prediction, in prioritizing putative cancer drivers. We then identified a list of putative cancer driver sSNVs and filtered this list via functional analysis to select 72 sSNVs, which are highly likely drivers in multiple cancer types, such as skin, large intestine, and liver, and should be among the priority candidates for experimental evaluation.

2. Materials and Methods

sSNV collection. We consider four categories of sSNVs: germline sSNVs (denoted as *germline*), somatic sSNVs in normal tissue (*somatic normal*), somatic sSNVs in cancerous tissues (*somatic cancer*), and putative cancer driver sSNVs (*putative drivers*). *Germline*, *somatic normal*, and *somatic cancer* variants are obtained from the gnomAD project [21], SomaMutDB [22], and COSMIC [23] databases, respectively. gnomAD (Genome Aggregation Database, <https://gnomad.broadinstitute.org/> (accessed on 9 November 2021)) houses data from large-scale sequencing efforts, identifying genomic variants from 16,708 genomes and 125,748 exomes; for the purposed of this paper, we only considered gnomAD exomes data, curated as described in our previous work [20]. SomaMutDB [22] (<https://vijglab.einsteinmed.org/SomaMutDB/> (accessed on 7 December 2021)) contains 2.42 million SNVs and 0.12 million INDELS (insertions or deletions) identified from 19 normal human tissue samples or cell line types (e.g., brain, blood, breast, heart, lung, liver, skin) of 374 individuals. The Catalogue of Somatic Mutations In Cancer (COSMIC) [23] houses a collection of somatic mutations found in cancerous tissues. The latest release of COSMIC (v95) includes 41 million confirmed somatic coding point mutations (SNVs—single nucleotide polymorphisms) from genome wide screenings of 1.4 million cancer tissue samples from 37 cancer primary sites. To be consistent with *somatic cancer* sSNVs, we only selected tissues, but not cell lines, from SomaMutDB to create the *somatic normal* set of sSNVs. To compile *somatic cancer* sSNVs, we downloaded the “CosmicGenomeScreensMutantExport.tsv.gz” file from COSMIC (GRCh37, <https://cancer.sanger.ac.uk/cosmic/download> (accessed on 9 November 2021)) and filtered the data to be “Confirmed somatic variant” and “Substitution—coding silent”. We mapped the genomic positions of COSMIC and SomaMutDB variants to all possible human transcript-based positions of sSNVs from the synVep database [20].

The *putative drivers* were sSNVs selected from the SynMICdb database (Synonymous Mutations in Cancer database, <http://synmicdb.dkfz.de/rsynmicdb/> (accessed on 7 December 2021)) [24]. SynMICdb houses 659,194 somatic sSNVs from COSMIC annotating their multiple aspects: whether the variant is in a cancer gene; variant frequency among healthy populations and in tumor samples; conservation of the affected genomic position (PhastCons [25]); pathogenicity/deleteriousness of the variant predicted by FATHMM-MKL [26] and CADD [27]; and the associated mRNA structural change predicted by remuRNA [28]. SynMICdb also provides SynMICdb scores, which are a heuristic combination of these annotations and are informative of the functional impact of sSNVs found

in cancer; we selected variants with SynMICdb scores in or above the 95th percentile of all scores.

Gene-set enrichment analysis. We conducted gene-set enrichment analysis (GSEA) for gene ontology (GO) terms [29,30] on *germline*, *somatic normal*, *somatic cancer*, and *putative driver* sSNVs using clusterProfile [31] R package. The GSEA was performed for the top 10% of the genes with the highest normalized sSNV rate, i.e., the number of sSNVs in a gene divided by the coding length of that gene. Note that, to reduce GO term redundancy, we used the GO terms semantic similarity analysis [32] of the top identified GO terms removing lower-ranked terms that were >0.5 similar to higher ranked ones.

Cancer-associated genes. We downloaded 576 Cancer Gene Census [33] tier 1 genes from the COSMIC database (<https://cancer.sanger.ac.uk/census> (accessed on 7 December 2021)); tier 2, according to COSMIC, lacks extensive evidence and is thus not included. We extracted 51 cancer pathways from KEGG [34] (<https://www.genome.jp/pathway/hsa05200> (accessed on 7 December 2021)) and identified 2210 corresponding genes using the clusterProfile [31] R package. We also obtained 217 disease ontology (DO) cancer terms from the supplementary data of Wu et al. [35] and identified 2895 corresponding genes using clusterProfile [31] R package. We term this collection of genes *cancer-associated*. In addition, we obtained from the literature [36] a set of 54 known oncogenes and 71 tumor suppressor genes.

Proposing a novel list of cancer driver sSNVs. We applied the following criteria to all *somatic cancer* variants to propose a novel list of potential cancer driver sSNVs (denoted as *proposed driver*): (1) synVep score > 0.81, i.e., the median of synVep predictions for *putative driver* set; (2) GERP++ score [37] (from <http://mendel.stanford.edu/SidowLab/downloads/gerp/> (accessed on 7 December 2021)) > 2.31, i.e., the median of GERP++ scores for the *putative driver* set; (3) located in a *cancer-associated* gene as defined above; (4) recurrent among cancer patients; here, we adopted the Sharma et al.'s approach [24] to define recurrence, i.e., mutations occurring more than once among different patients.

Functional impact prediction for annotation of proposed driver variants. We used the CADD online server (<https://cadd.gs.washington.edu/score> (accessed on 3 March 2022)) annotations for GRCh37-v1.6 to retrieve CADD-splice (CADD v1.6) and spliceAI predictions for the *proposed driver* variants. For CADD-splice predictions, we considered sSNVs scoring > 15 to be splicing-disruptive (recommended cutoff at <https://cadd.gs.washington.edu/info> (accessed on 3 March 2022)). For spliceAI, we considered an sSNV to be splicing-disruptive if one of the four predictions generated (acceptor gain, acceptor loss, donor gain, and donor loss) was greater than 0.5.

We used the RNAsnp [38] package for the prediction of changes to sSNV-affected RNA structures. As per the instructions (<https://rth.dk/resources/rnasnp/software.php> (accessed on 3 March 2022)), mode 1 was used for transcripts less than 200 nucleotides long, and mode 2 otherwise. Other parameters were set as default.

We further predicted all putative RBP motifs in all human protein coding transcripts (extracted from Ensembl BioMart assembly GRCh37 [39], https://figshare.com/articles/dataset/transcript_sequences_zip/19407530 (accessed on 3 March 2022)) using the online interface of the FIMO (Find Individual Motif Occurrences) [40] method from the MEME (Multiple Em for Motif Elicitation) suite [41] (<https://meme-suite.org/meme/tools/fimo> (accessed on 3 March 2022)); all parameters set as default).

The human RBP motifs file ("Ray2013_rbp_Homo_sapiens.dna_encoded.meme" [42]) was obtained from the MEME motif database (<https://meme-suite.org/meme/doc/download.html> (accessed on 3 March 2022)). We then examined whether these extracted motifs overlapped with our *proposed driver* sSNV locations.

We also mapped sSNVs to potential transcription factor binding sites (TFBS) via the SNP2TFBS [43] web server (<https://ccg.epfl.ch/snp2tfbs/snpselect.php> (accessed on 3 March 2022)).

Statistical analysis. Kruskal–Wallis test [44] was used as a non-parametric alternative to ANOVA to test whether the mean ranks of multiple groups are the same; post hoc

pairwise comparison was performed with Dunn test [45] using FSA package [46] (<https://cran.r-project.org/web/packages/FSA/index.html> (accessed on 10 March 2022)) in R [47] (<https://www.r-project.org/> (accessed on 10 March 2018)).

3. Results and Discussion

3.1. sSNV-Affected Molecular Functions Differ by Variant Class

We evaluated the per-gene sSNV burden, i.e., the number of sSNVs per gene normalized by the length of the corresponding coding region (Methods), for all genes of all cancer patients in the COSMIC database. We found that sSNV burden of oncogenes and tumor suppressor genes (from [36]) does not differ significantly. However, both oncogenes and tumor suppressors have lower sSNV burden than either the Cancer Gene Census (CGC) [33] cancer genes or non-cancer genes (Supplementary Figure S1). This unexpected observation may be due to the necessity of maintaining the specific (high or low) levels of functionality of oncogenes and TSGs in cancer development, while no such limitations/selection pressures are imposed on other genes.

We also evaluated gene mutability overall by evaluating occurrence of other genetic variants. The numbers of nsSNVs per gene highly correlated with numbers of sSNVs (Pearson correlation = 0.86). However, the per gene nsSNV/sSNV ratio was also somewhat indicative of oncogenes. That is, the top 100 genes with highest nsSNV/sSNV ratio had more cancer genes (18%; oncogenes, tumor suppressors, and CGC) compared to the 100 genes with the lowest nsSNV/sSNV ratio (2%). The per gene nsSNV/sSNV derived from COSMIC database can be found in Supplementary Table S2.

We further collected 4,221,244 *germline*, 54,368 *somatic normal*, 2,894,289 *somatic cancer*, and 27,878 *putative driver* sSNVs (Methods). For each of the sSNV categories, we calculated the normalized sSNV burden (highest ranked genes in Supplementary Table S1). As expected, genes containing the putative cancer drivers were heavily enriched in cancer association (61 of 100 were cancer genes). Cancer-associated genes were also found among germline and somatic cancer variant-enriched genes (3 of 100 genes each), but not in the somatic normal set. Curiously, the gene overlap among the four categories was minimal, indicating that somatic sSNVs affect different genes than germline sSNVs, as well as that mutation and selection mechanisms in cancer and normal tissues are also different.

To compare the sSNVs across the four categories, we performed gene-set enrichment analyses (GSEA) for gene ontology (GO) [29,30] terms of the genes most-enriched in *germline*, *somatic normal*, *somatic cancer*, and *putative driver* sSNVs. Nine of the top ten *putative driver*-gene GO terms were unique to this set of variants, i.e., they were not in the top ten GO terms of *germline*, *somatic normal*, or *somatic cancer* genes (Figure 1), indicating that the biological functions of the *putative driver*-enriched genes are different from those of genes enriched in the other three categories of sSNVs. To evaluate the consistency of this observation, we conducted additional analyses with varying number of input genes (top 10%, 20%, and 30% genes with highest sSNV density), as well as varying number of GO terms (top 10, 20, and 30). The GO terms' overlaps between *putative driver* and other groups were consistently low, ranging from 0 (e.g., overlap with *germline*, top 30% genes, top 10 GO terms) to 0.1 (e.g., overlap with *somatic normal*, top 20% genes, top 20 GO terms). Curiously, we also note that *germline* GO terms were very different from *somatic* ones, while *somatic cancer* and *normal*-enriched terms were somewhat similar.

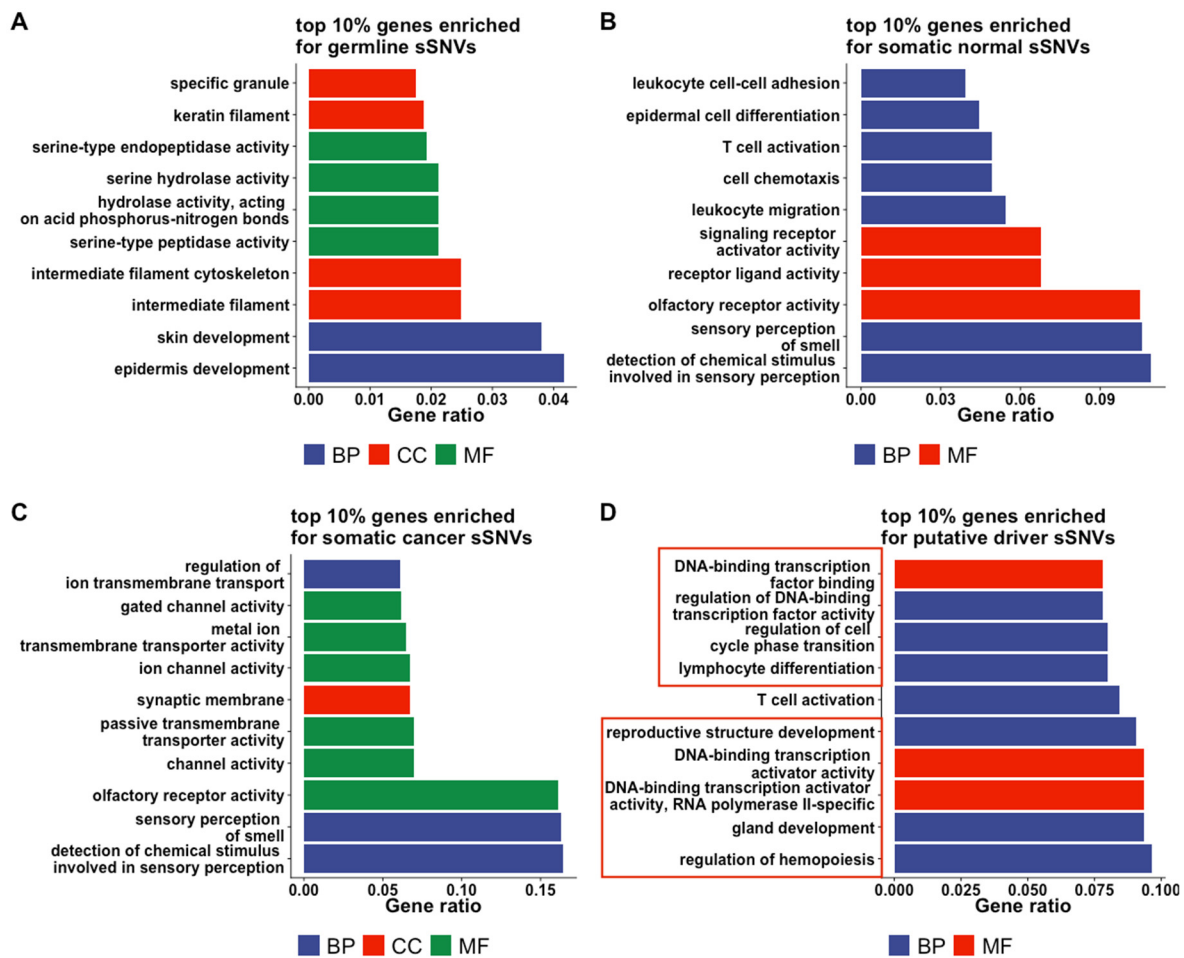


Figure 1. Genes enriched for different categories of sSNVs differ in GO terms. (GO) terms from resulting from gene-set enrichment analysis (GSEA) are shown for (A) germline, (B) somatic normal, (C) somatic cancer, and (D) putative driver sSNVs. The X-axis (gene ratio) is the percentage of the input genes that are associated with the specific GO term. The bars are colored by GO term groups, i.e., BP: biological processes in blue, CC: cellular component in green, MF: molecular function in red. Only the top 10 GO terms are shown for each category of sSNVs. The GO terms that are specific to the putative driver category are shown in red boxes in panel (D).

3.2. SynVep Variant Effect Scores Are Higher for Putative Drivers

Putative driver sSNVs are usually not observed in the general population, as was reflected by gnomAD, (Figure 2A). Furthermore, they were often localized to more conserved regions than the other three categories of sSNVs (Figure 2B). While reassuring, we note that this observation is trivial as mutation population frequency and conservation (PhastCons [25]) are included in the calculation of SynMICdb scores, which define putative driver variants.

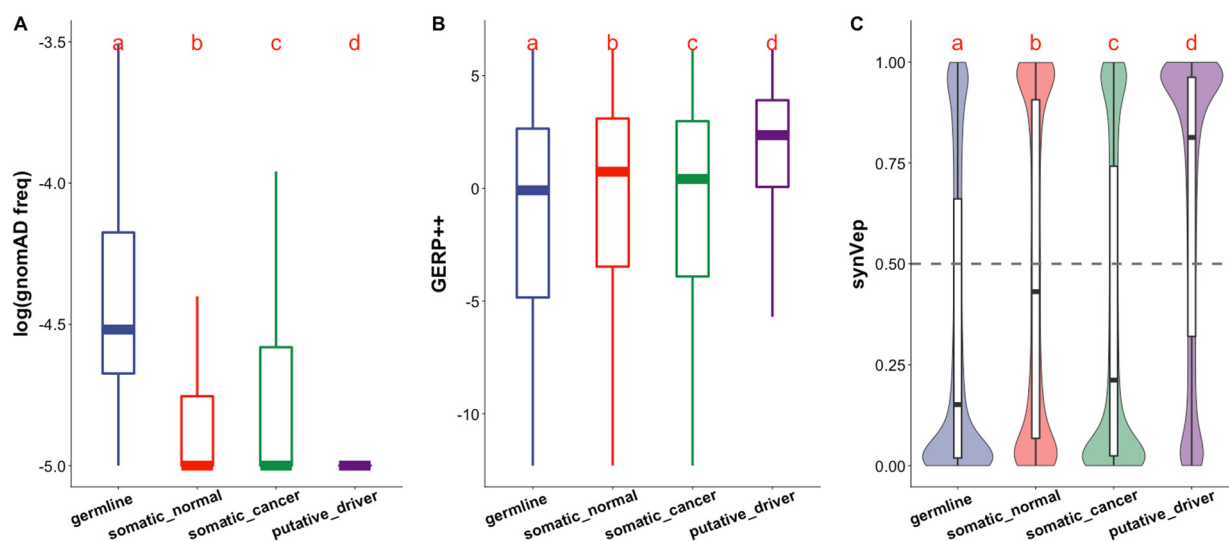


Figure 2. Variation in population frequency, conservation, and synVep predictions of the *germline*, *somatic normal*, *somatic cancer*, and *putative driver* sSNVs. Variant types are indicated by color: germline is blue, somatic normal is red, somatic cancer is green, and putative driver is purple. Variants differ by (A) population frequency (gnomAD frequencies; sSNVs with 0 frequency are set to $\log(\text{freq}) = -5$ for display purposes), (B) conservation (GERP++ scores), and (C) effects (synVep predictions; scores > 0.5 , i.e., above the gray dashed line, indicate effect). For each panel, the Kruskal–Wallis test rejected (p -value $< 2 \times 10^{-16}$) the null hypothesis that all groups follow the same distribution; as did the Dunn test pairwise comparisons (different letters indicate statistically different distributions).

SynMICdb scores also includes functional prediction by CADD [27] (deleteriousness) and FATHMM-MKL [26] (pathogenicity). We previously developed synVep (synonymous Variant effect predictor) [20]—a machine learning-based method for predicting the likelihood of a human sSNV having an effect on the function of the corresponding gene product. synVep training relies on observation of sSNVs in human population, with a fundamental assumption that the unobserved sSNVs are enriched in functional effects. We demonstrated earlier that synVep can identify experimentally validated sSNV effects, pathogenic sSNVs, and splicing-disruptive sSNVs [20]. Importantly, synVep does not use conservation as a feature, thus providing information orthogonal to that of other functional impact predictors and to the SynMICdb score as a whole. From the synVep functional effect perspective, the *germline*, *somatic normal*, and *somatic cancer* variants may have an effect or not, but cancer driver mutations must have an effect. Thus, *putative driver* sSNVs were expected to have higher synVep scores, indicating variants that are more likely to have an effect, than variants of the other three categories of sSNVs (Figure 2C).

We note that higher synVep scores of the *putative driver* sSNVs may in part be due to their absence from the general population (Figure 2A)—a feature of most of the effect variants in synVep’s training set. To evaluate the effect of this potential bias, we removed all sSNVs labeled as observed in gnomAD from the somatic categories of variants and re-evaluated the synVep scores for the remaining data. The synVep predictions of *putative driver* sSNVs were still substantially higher than those of both *somatic normal* and *somatic cancer* sSNVs (Supplementary Figure S2B).

Curiously, the *somatic normal* sSNVs, on average, scored higher than the *somatic cancer* sSNVs. This finding is in line with the fact that positive selection rules the likelihood of somatic variants [48,49] propagating throughout cells that make up individual tissues and, in order to be selected for, the variants need to have a molecular effect. In contrast, the vast majority of somatic mutations in cancerous tissues are passengers [8], as opposed to very few driver mutations, and are thus selectively neutral [50] having no or weak effect.

3.3. Screening sSNVs to Recover Cancer-Underlying Genes

Cancer genes are defined as those that can harbor mutations conferring growth advantage of tumor cells [51]. CGC [33] collects cancer genes with extensive evidence, but the discovery of all cancer genes is not yet close to completion [48]. For example, CGC genes only account for 54% and 70% of genes in KEGG cancer pathways [34] and in cancer ontologies [35], respectively. It is thus possible that mutations in non-CGC genes may be indirectly involved in causing cancer, by, e.g., contributing to initiation or progression of cancer or by enhancing the effects of cancer drivers [8,52,53]. With the increase in large-scale tumor sequencing, more data for analysis has become available and may identify additional cancer genes. However, different cancer gene identification methods produce different results and often fail to recover the previously identified cancer genes [54]. In other words, mutations labeled as non-drivers due to their localization to non-CGC genes may be incorrectly labeled, i.e., false negative.

Given that most somatic mutations are random, recurrent mutations (same mutation in different cancer patients) are unlikely to occur by chance and are thus likely carcinogenic [10]. Evolutionary conservation is informative for prioritizing cancer drivers [55]. Furthermore, synVep, as we demonstrated earlier [20], is precise in differentiating sSNV molecular effects. Importantly, as conservation is not one of synVep's feature, these two sSNV features are orthogonal. Following these observations, we identified four groups of genes based on whether they harbor certain types of sSNVs (Methods): (1) genes with non-recurrent sSNVs only, i.e., genes harboring recurrent sSNVs are excluded; (2) genes with recurrent sSNVs; (3) genes with recurrent sSNVs that are located at conserved positions; and (4) genes with recurrent sSNVs that are located at conserved positions and are scored high by synVep. We found that incorporation of recurrence, conservation, and synVep prediction filters identified genes that are more likely to be involved in cancer (Figure 3). For example, our most rigorous filtering identified 40% (229) of the 576 CGC genes in addition to another set of 4819 genes that are possibly cancer associated. In fact, 26% ($n = 1329$) of our genes were present in CGC, KEGG cancer pathways, or in the DO cancer gene list.

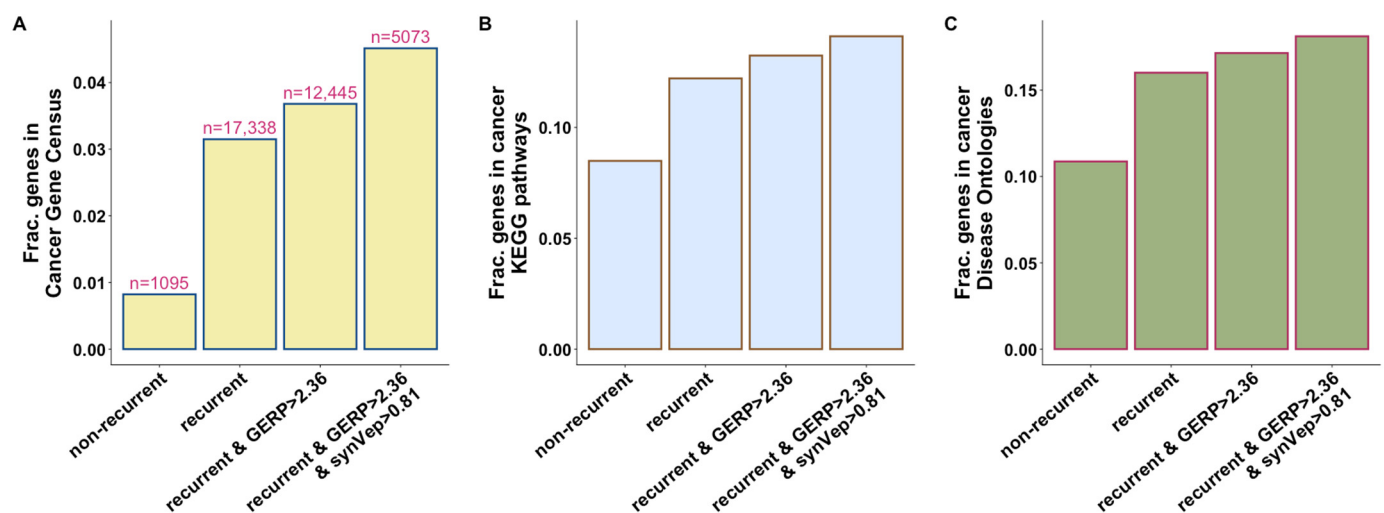


Figure 3. Genes identified by recurrence, conservation, and synVep prediction are more likely to be involved in cancer. The four categories of genes are identified by the filters (X-axis) as described in text. Y-axis represents the fraction of the selected category genes that are found in (A) Cancer Gene Census, (B) KEGG cancer pathway genes, and (C) DO cancer genes. Numbers on top of each bar in panel A show the number genes of each category.

We also found that narrowing the lists to genes with more sSNVs that pass the above filters identifies more likely cancer genes (Figure 4). Note that, since different filters result in different distributions of variant counts, we use “>x percentile” to represent the top-ranking genes. For example, if the recurrence filter identifies 10 genes with 1, 1, 2, 2, 3, 3, 3, 3, 8, and 9 sSNVs, respectively, then the “>70-percentile” of the counts would include 8 and 9 sSNVs. The observation that genes with more sSNVs passing the filter are more likely cancer-associated is especially true for genes with recurrent variants. However, known cancer genes tend to have more non-recurrent sSNVs as well (Figure 4). One possible explanation is that the normal activity of cancer genes may also be disrupted by an accumulation of variants within the functional domains, whether the variants are recurrent or not [56,57].

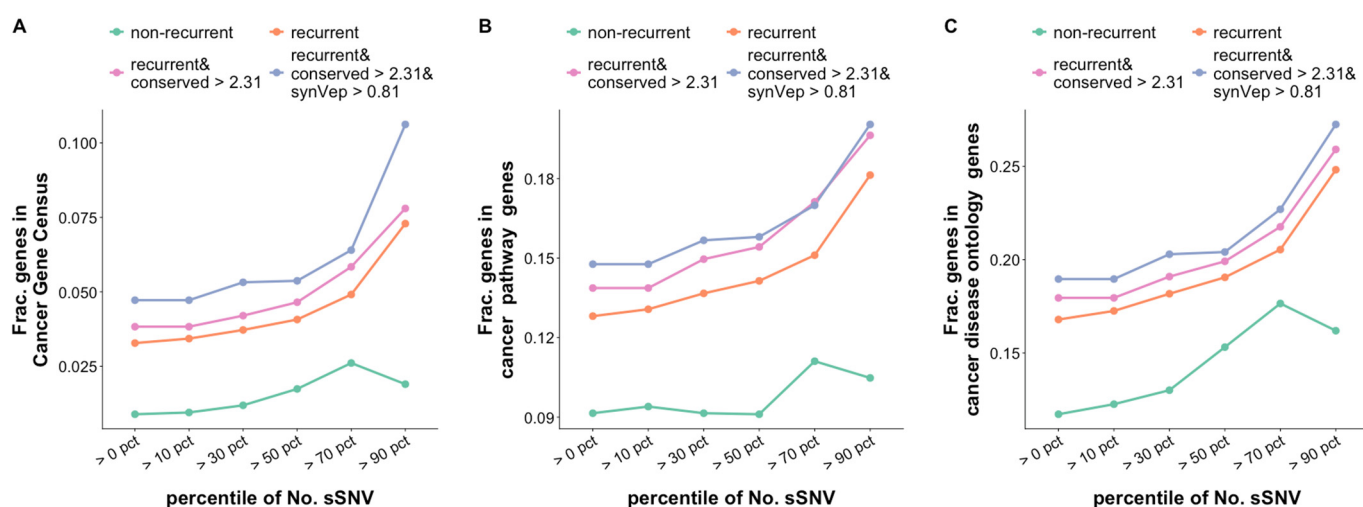


Figure 4. Genes with more sSNVs are more likely to be involved in cancer. The four categories of genes are identified by the above-described filters (Methods). The X-axis indicates that genes with more than the corresponding amount of sSNVs (represented as percentile) are selected. Y-axis represents the fraction of the selected category genes that are found in (A) CGC, (B) KEGG cancer pathway genes, and (C) DO cancer genes. Numbers on top of each bar in panel A show the number genes of each category.

Our results show that a high number of variants per gene that pass the recurrence, variant position conservation, and synVep filters can much better identify potential cancer genes than sSNV recurrence alone. There are 417 genes (Supplementary Table S3) containing > 17 sSNVs (> 90-percentile) that pass all of these filters. Among these genes, 40% ($n = 166$) are known to be cancer-associated, according to CGC, KEGG cancer pathways, or cancer DO. We expect that many of the remaining 251 genes may also be cancer-associated, although their mechanisms are yet not understood. As an example, consider three with the most sSNVs: *PCDH15*, *CEL4*, and *MYBPC1*.

- *PCDH15* encodes protocadherins, a group of calcium-dependent cell–cell adhesion protein [58]. It has been noted in earlier work as a potential marker for NK (natural killer)/T cell lymphomas [59]. Mutations in *PCDH15* have been identified in a whole-genome sequencing study [60] and an exome sequencing study [61] of prostate cancer. Another whole-exome sequencing study revealed that *PCDH15* harbored mutations associated with metastasis in ocular adnexal sebaceous carcinoma [62]. Furthermore, a genome-wide association study (GWAS) identified multiple loci in *PCDH15* to be significantly associated with acute myeloid leukemia [63].
- *CEL4* is one of the CELF proteins (CUGBP, ELAV-like family of proteins), which are a type of RNA-binding protein (RBP) with various roles in RNA regulation [64]. An earlier study identified an intronic *CEL4* germline variant associated with colorectal cancer risk [65]. Multiple other analyses found that *CEL4* can be used to

prognose colorectal cancer [66–68]. Additionally, methylation of *CELF4* was proposed as a detection method for endometrial cancer [69].

- *MYBPC1* encodes a member of myosin-binding protein C family with a role in muscle contraction [70]. Significant differential expression of *MYBPC1* has been observed in tongue cancer [71], breast cancer [72,73], and prostate cancer [74]. Additionally, *MYBPC1* expression level was found to positively correlate with NK cell content [73].

Concordance between our findings and literature evidence for the likely involvement of our top ranked genes in cancer highlights the utility of our prioritization strategy, suggesting which unknown cancer-associated genes remain to be explored.

3.4. Selecting Novel Potential Cancer Driver sSNVs

As described above, we assume that cancer driver sSNVs can be identified by three filters: recurrence among cancer patients, affected genome position conservation, and synVep prediction on the sSNV impact. To identify a list of potential cancer driving sSNVs, we performed the following filtering: starting from 2,894,289 *somatic cancer* sSNVs from the COSMIC database, we applied four filters (recurrent variant, GERP++ score > 2.31, synVep prediction > 0.81, cancer-associated genes). We thus obtained 2111 (genomic position-based variants; mapping to 5021 transcript-based) sSNV candidates (Supplementary Table S4). These were evaluated for functional impact mechanism from three perspectives: mRNA alternative slicing, mRNA structural changes, as well as localization to RNA-binding protein (RBP) binding motifs; functional impacts of 326 sSNVs (genomic position-based; 609 transcript-based) were thus identified. A brief flowchart describing these processes is shown in Supplementary Figure S3. We describe more detailed results of the functional impact evaluations below:

Splicing changes: After transcription of a gene, splicing removes intronic sequences from the pre-mRNA molecule and/or joins exonic sequences. A primary transcript can be spliced into multiple mature mRNAs (known as alternative splicing) corresponding to different protein isoforms with varying functionalities [75]. Mutations can disrupt the splicing regulatory elements, resulting in aberrant splicing [76]. sSNV-induced aberrant splicing is common in multiple diseases [77,78], including cancers [16], and has been observed in many cancer genes, such as BRCA1 [79], BRCA2 [80], APC [81], and BAP1 [19]. CADD-splice [82] and spliceAI [83] are two state-of-the-art tools to predict splicing disruption induced by mutations. Of the 2111 *proposed driver* sSNVs, 136 (genomic coordinate-based; mapping to 222 transcript-based) sSNVs were predicted to be splicing-disruptive by CADD-splice or spliceAI (Supplementary Table S5) to be associated with aberrant splicing. In our set of variants, these putatively splicing-disrupting sSNVs affect multiple cancer types, including liver, large intestine, ovary, central nervous system, etc.

mRNA structural changes: sSNVs can alter mRNA structure (Figure 5), stability [84–86], and translational speed [87], potentially causing disease [88–90]. RNAsnp [38] is a computational tool to predict whether an sSNV induces significant mRNA structural changes. Of our set of 2111 *proposed driver* sSNVs, 104 (Supplementary Table S6) were predicted by RNAsnp to cause significant mRNA structural changes. These predicted mRNA structure-changing sSNVs are found in multiple cancer types in our set, e.g., breast, skin, urinary tract, and liver.

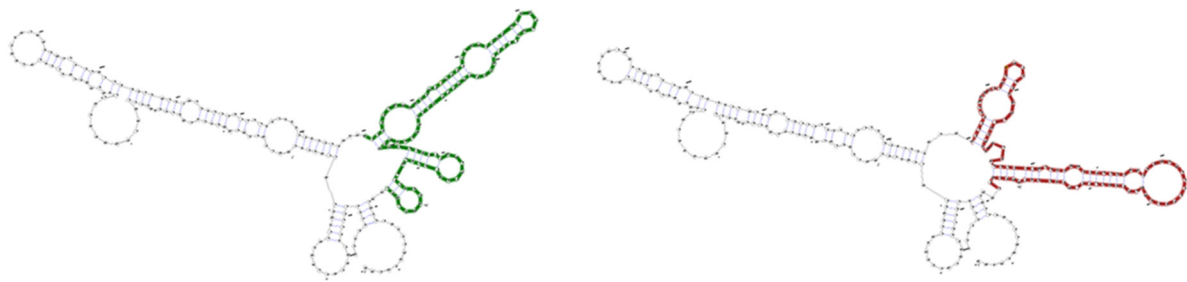


Figure 5. An example of RNA structural change due to an sSNV. A transcript (ENST00000371081) experiences structural change from wildtype (green) to mutant (red) due to a proposed driver sSNV (C165A). The illustration is generated by RNAsnp [38] web server (<https://rth.dk/resources/rnasnp/> (accessed on 12 March 2022)).

Changes to binding of proteins: RNA Binding Proteins (RBPs) bind to a specific RNA sequence motif or secondary structure to regulate multiple post-transcriptional events, including mRNA splicing, polyadenylation, localization, and degradation [91,92]. To date, over 1500 RBPs have been identified [93], which bind to a motif that is 3–7 nucleotide long [42]. Oncogenic effects of mutations in RBP-coding genes have been well documented [92]. Mutations in cancer-associated RBP binding sites can alter RNA expression and splicing [94]. Notably, Teng et al. experimentally demonstrated that sSNVs can disrupt the binding between RBP and the transcripts of cancer genes (e.g., *DAB2* and *PCBP3*, *ZFH3* and *PTBP1*) [95]. Here, we extracted all putative RBP motifs in all human protein coding sequences using the FIMO (Find Individual Motif Occurrences) [40] and examined whether these motifs overlap with our *proposed driver* sSNVs. We identified 107 genomic-based *proposed driver* sSNVs that overlap with RBP binding motifs (Supplementary Table S7).

Changes to transcription factor binding: Another possible oncogenic effect of cancer driver mutations is alteration of transcription factor binding sites (TFBS) [96,97]. We used the SNP2TFBS tool [43] to find *proposed driver* sSNVs mapping to TFBS. However, none of our variants were labeled as TFBS-affecting.

Of the 2111 genomic position-based *proposed driver* variants, our functional analysis identified 326 sSNVs of specific impact mechanisms (Figure 6; 136 sSNVs affecting splice sites, 104 sSNVs inducing RNA structural changes, and 107 sSNVs affecting RBP motifs; some variants with multiple impacts). These 326 sSNVs (genomic position-based; 609 transcript-based) are primarily found in skin, large intestine, lung, and liver cancers (Supplementary Figure S4) in our set. The functional impacts of other *proposed driver* sSNVs require further investigation. Note that our pipeline for putative driver sSNV selection and evaluation of results are inherently limited by the accuracy of the computational tools (synVep, RNAsnp, spliceAI, and CADD-splice, and FIMO) used in the analysis. Additionally, some driver mutations may fail to pass our recurrence filters due to low frequency or high tumor heterogeneity [98]. Finally, it is also possible that the *proposed driver* sSNVs do not individually act as cancer driver mutations, but collectively contribute to cancer progression [8,52,53].

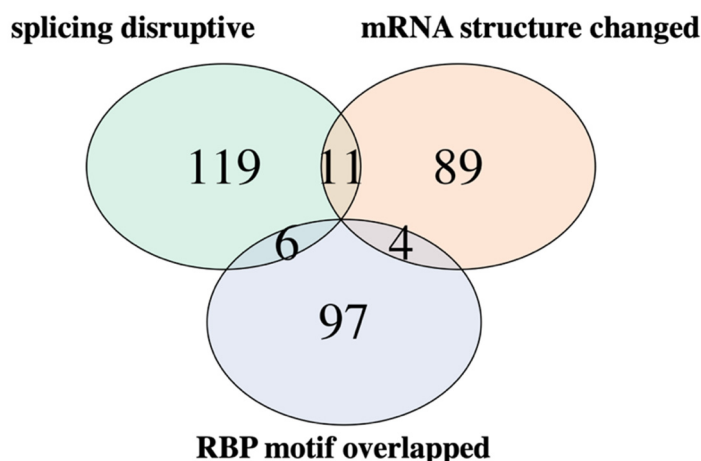


Figure 6. Venn diagram of the sSNV functional impacts. Count of sSNVs affecting splicing (predicted by CADD-splice and spliceAI), RBP motifs, or changing mRNA structure. Pairwise hypergeometric tests suggest that the overlaps between variant sets are not statistically significant (p -value > 0.05).

4. Conclusions

Here, we developed and evaluated a new way to identify sSNV cancer drivers and proposed a means of tagging cancer genes (for a graphical overview, see Supplementary Figure S3). To identify drivers, we used variant recurrence in cancer data, together with synVep functional impact scores and variant position conservation. We showed that our *proposed driver* variants selected in this manner are enriched in known cancer genes and pathways. However, they also identify genes that have not previously been deemed relevant to cancer. We further found that a higher number of putative drivers per gene is likely an indication of that gene's involvement in cancer appearance and/or progression. Finally, we showed that at least 15% of our putative driver variants likely disrupt cellular mechanisms known to be cancer associated. Our results highlight the potential importance of synonymous variants in causing cancer. Our methods may also be used in prioritizing experimental validation of cancer driver sSNVs and novel cancer genes in the future.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/genes13050778/s1>, Figure S1: Gene sSNV burden by gene type; Figure S2: Variation in conservation and synVep predictions of the *somatic normal*, *somatic cancer*, and *putative driver* sSNVs; Figure S3: Overview of the analysis and procedures of identifying potential cancer driving sSNVs; Figure S4: Distribution of *proposed driver* sSNVs with identified functional impacts by cancer primary site; Table S1: top_ranking_genes.csv; Table S2: nsSNV_sSNV_ratio.csv; Table S3: genes_all_filters.csv; Table S4: proposed_drivers.csv; Table S5: splicing_disrupted.csv; Table S6: RNA_structure_changed.csv; Table S7: RBP_motif_overlapped.csv.

Author Contributions: Z.Z. and Y.B. designed the study, evaluated the results, and wrote the manuscript; Z.Z. conducted the study. All authors have read and agreed to the published version of the manuscript.

Funding: Z.Z. and Y.B. were supported by the NIH/NIGMS grant R01 (GM115486); Y.B. was also supported by NIH grant R01 (MH115958).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Transcript sequences used for RBP motif extraction can be found in figshare https://figshare.com/articles/dataset/transcript_sequences_zip/19407530 (accessed on 24 March 2022).

Acknowledgments: We thank Yannick Mahlich and Ariel Aptekmann (both Rutgers) for their constructive discussion and Rutgers Amarel compute cluster for the computational resource. We also thank all scientific researchers who made the data and tools relevant to this study available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [[CrossRef](#)] [[PubMed](#)]
2. Danaei, G.; Vander Hoorn, S.; Lopez, A.D.; Murray, C.J.; Ezzati, M.; Comparative Risk Assessment Collaborating Group (Cancers). Causes of cancer in the world: Comparative risk assessment of nine behavioural and environmental risk factors. *Lancet* **2005**, *366*, 1784–1793. [[CrossRef](#)]
3. Jiang, X.; Finucane, H.K.; Schumacher, F.R.; Schmit, S.L.; Tyrer, J.P.; Han, Y.; Michailidou, K.; Lesueur, C.; Kuchenbaecker, K.B.; Dennis, J. Shared heritability and functional enrichment across six solid cancers. *Nat. Commun.* **2019**, *10*, 431. [[CrossRef](#)] [[PubMed](#)]
4. Bromberg, Y. Chapter 15: Disease Gene Prioritization. *PLoS Comput. Biol.* **2013**, *9*, e1002902. [[CrossRef](#)]
5. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [[CrossRef](#)]
6. Pon, J.R.; Marra, M.A. Driver and passenger mutations in cancer. *Annu. Rev. Pathol. Mech. Dis.* **2015**, *10*, 25–50. [[CrossRef](#)]
7. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **2020**, *578*, 82–93. [[CrossRef](#)]
8. Kumar, S.; Warrell, J.; Li, S.; McGillivray, P.D.; Meyerson, W.; Salichos, L.; Harmanci, A.; Martinez-Fundichely, A.; Chan, C.W.; Nielsen, M.M. Passenger mutations in more than 2500 cancer genomes: Overall molecular functional impact and consequences. *Cell* **2020**, *180*, 915–927.e16. [[CrossRef](#)]
9. Cheng, F.; Zhao, J.; Zhao, Z. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Brief. Bioinform.* **2016**, *17*, 642–656. [[CrossRef](#)]
10. Raphael, B.J.; Dobson, J.R.; Oesper, L.; Vandin, F. Identifying driver mutations in sequenced cancer genomes: Computational approaches to enable precision medicine. *Genome Med.* **2014**, *6*, 5. [[CrossRef](#)]
11. Alexandrov, L.B.; Kim, J.; Haradhvala, N.J.; Huang, M.N.; Ng, A.W.T.; Wu, Y.; Boot, A.; Covington, K.R.; Gordenin, D.A.; Bergstrom, E.N. The repertoire of mutational signatures in human cancer. *Nature* **2020**, *578*, 94–101. [[CrossRef](#)] [[PubMed](#)]
12. Zhang, J.; Bajari, R.; Andric, D.; Gerthoffert, F.; Lepsa, A.; Nahal-Bose, H.; Stein, L.D.; Ferretti, V. The international cancer genome consortium data portal. *Nat. Biotechnol.* **2019**, *37*, 367–369. [[CrossRef](#)] [[PubMed](#)]
13. Wang, T.; Ruan, S.; Zhao, X.; Shi, X.; Teng, H.; Zhong, J.; You, M.; Xia, K.; Sun, Z.; Mao, F. OncoVar: An integrated database and analysis platform for oncogenic driver variants in cancers. *Nucleic Acids Res.* **2021**, *49*, D1289–D1301. [[CrossRef](#)] [[PubMed](#)]
14. Soussi, T.; Taschner, P.E.; Samuels, Y. Synonymous somatic variants in human cancer are not infamous: A plea for full disclosure in databases and publications. *Hum. Mutat.* **2017**, *38*, 339–342. [[CrossRef](#)] [[PubMed](#)]
15. Zeng, Z.; Bromberg, Y. Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. *Front. Genet.* **2019**, *10*, 914. [[CrossRef](#)]
16. Supek, F.; Miñana, B.; Valcárcel, J.; Gabaldón, T.; Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **2014**, *156*, 1324–1335. [[CrossRef](#)]
17. Gartner, J.J.; Parker, S.C.; Prickett, T.D.; Dutton-Regester, K.; Stitzel, M.L.; Lin, J.C.; Davis, S.; Simhadri, V.L.; Jha, S.; Katagiri, N. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 13481–13486. [[CrossRef](#)]
18. Liu, F.; Calhoun, B.; Alam, M.S.; Sun, M.; Wang, X.; Zhang, C.; Haldar, K.; Lu, X. Case report: A synonymous VHL mutation (c. 414A>G, p. Pro138Pro) causes pathogenic familial hemangioblastoma through dysregulated splicing. *BMC Med. Genet.* **2020**, *21*, 42. [[CrossRef](#)]
19. Niersch, J.; Vega-Rubín-de-Celis, S.; Bazarna, A.; Mergener, S.; Jendrossek, V.; Siveke, J.T.; Peña-Llopis, S. A BAP1 synonymous mutation results in exon skipping, loss of function and worse patient prognosis. *IScience* **2021**, *24*, 102173. [[CrossRef](#)]
20. Zeng, Z.; Aptekmann, A.A.; Bromberg, Y. Decoding the effects of synonymous variants. *Nucleic Acids Res.* **2021**, *49*, 12673–12691. [[CrossRef](#)]
21. Karczewski, K.J.; Francioli, L.C.; Tiao, G.; Cummings, B.B.; Alfoldi, J.; Wang, Q.; Collins, R.L.; Laricchia, K.M.; Ganna, A.; Birnbaum, D.P. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **2020**, *581*, 434–443. [[CrossRef](#)] [[PubMed](#)]
22. Sun, S.; Wang, Y.; Maslov, A.Y.; Dong, X.; Vijg, J. SomaMutDB: A database of somatic mutations in normal human tissues. *Nucleic Acids Res.* **2022**, *50*, D1100–D1108. [[CrossRef](#)] [[PubMed](#)]
23. Tate, J.G.; Bamford, S.; Jubb, H.C.; Sondka, Z.; Beare, D.M.; Bindal, N.; Boutselakis, H.; Cole, C.G.; Creatore, C.; Dawson, E. COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **2019**, *47*, D941–D947. [[CrossRef](#)]
24. Sharma, Y.; Miladi, M.; Dukare, S.; Boulay, K.; Caudron-Herger, M.; Groß, M.; Backofen, R.; Diederichs, S. A pan-cancer analysis of synonymous mutations. *Nat. Commun.* **2019**, *10*, 2569. [[CrossRef](#)]

25. Siepel, A.; Bejerano, G.; Pedersen, J.S.; Hinrichs, A.S.; Hou, M.; Rosenbloom, K.; Clawson, H.; Spieth, J.; Hillier, L.W.; Richards, S. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **2005**, *15*, 1034–1050. [[CrossRef](#)] [[PubMed](#)]
26. Shihab, H.A.; Rogers, M.F.; Gough, J.; Mort, M.; Cooper, D.N.; Day, I.N.; Gaunt, T.R.; Campbell, C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **2015**, *31*, 1536–1543. [[CrossRef](#)]
27. Kircher, M.; Witten, D.M.; Jain, P.; O’roak, B.J.; Cooper, G.M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **2014**, *46*, 310. [[CrossRef](#)]
28. Salari, R.; Kimchi-Sarfaty, C.; Gottesman, M.M.; Przytycka, T.M. Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: Application to disease studies. *Nucleic Acids Res.* **2013**, *41*, 44–53. [[CrossRef](#)]
29. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [[CrossRef](#)]
30. Consortium, T.G.O. The Gene Ontology resource: Enriching a Gold mine. *Nucleic Acids Res.* **2021**, *49*, D325–D334. [[CrossRef](#)]
31. Yu, G.; Wang, L.-G.; Han, Y.; He, Q.-Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *Omic J. Integr. Biol.* **2012**, *16*, 284–287. [[CrossRef](#)] [[PubMed](#)]
32. Yu, G. Gene ontology semantic similarity analysis using GOsemSim. In *Stem Cell Transcriptional Networks*; Humana: New York, NY, USA, 2020; pp. 207–215.
33. Sondka, Z.; Bamford, S.; Cole, C.G.; Ward, S.A.; Dunham, I.; Forbes, S.A. The COSMIC Cancer Gene Census: Describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **2018**, *18*, 696–705. [[CrossRef](#)] [[PubMed](#)]
34. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [[CrossRef](#)] [[PubMed](#)]
35. Wu, T.-J.; Schriml, L.M.; Chen, Q.-R.; Colbert, M.; Crichton, D.J.; Finney, R.; Hu, Y.; Kibbe, W.A.; Kincaid, H.; Meerzaman, D. Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis. *Database* **2015**, *2015*, bav032. [[CrossRef](#)] [[PubMed](#)]
36. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A., Jr.; Kinzler, K.W. Cancer genome landscapes. *Science* **2013**, *339*, 1546–1558. [[CrossRef](#)]
37. Davydov, E.V.; Goode, D.L.; Sirota, M.; Cooper, G.M.; Sidow, A.; Batzoglou, S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **2010**, *6*, e1001025. [[CrossRef](#)]
38. Sabarinathan, R.; Tafer, H.; Seemann, S.E.; Hofacker, I.L.; Stadler, P.F.; Gorodkin, J. RNA snp: Efficient detection of local RNA secondary structure changes induced by SNP s. *Hum. Mutat.* **2013**, *34*, 546–556. [[CrossRef](#)]
39. Kinsella, R.J.; Kahari, A.; Haider, S.; Zamora, J.; Proctor, G.; Spudich, G.; Almeida-King, J.; Staines, D.; Derwent, P.; Kerhornou, A.; et al. Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database* **2011**, *2011*, bar030. [[CrossRef](#)]
40. Grant, C.E.; Bailey, T.L.; Noble, W.S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **2011**, *27*, 1017–1018. [[CrossRef](#)]
41. Bailey, T.L.; Johnson, J.; Grant, C.E.; Noble, W.S. The MEME suite. *Nucleic Acids Res.* **2015**, *43*, W39–W49. [[CrossRef](#)]
42. Ray, D.; Kazan, H.; Cook, K.B.; Weirauch, M.T.; Najafabadi, H.S.; Li, X.; Gueroussov, S.; Albu, M.; Zheng, H.; Yang, A. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **2013**, *499*, 172–177. [[CrossRef](#)] [[PubMed](#)]
43. Kumar, S.; Ambrosini, G.; Bucher, P. SNP2TFBS—A database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **2017**, *45*, D139–D144. [[CrossRef](#)] [[PubMed](#)]
44. Kruskal, W.H.; Wallis, W.A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [[CrossRef](#)]
45. Dunn, O.J. Multiple comparisons using rank sums. *Technometrics* **1964**, *6*, 241–252. [[CrossRef](#)]
46. Ogle, D.H.; Doll, J.; Wheeler, P.; Dinno, A. FSA: Fisheries Stock Analysis. (2021). Available online: <https://cran.r-project.org/web/packages/FSA/index.html> (accessed on 10 March 2022).
47. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020; Available online: <https://www.R-project.org/> (accessed on 10 March 2018).
48. Martincorena, I.; Raine, K.M.; Gerstung, M.; Dawson, K.J.; Haase, K.; Van Loo, P.; Davies, H.; Stratton, M.R.; Campbell, P.J. Universal patterns of selection in cancer and somatic tissues. *Cell* **2017**, *171*, 1029–1041.e21. [[CrossRef](#)]
49. Olafsson, S.; Anderson, C.A. Somatic mutations provide important and unique insights into the biology of complex diseases. *Trends Genet.* **2021**, *37*, 872–881. [[CrossRef](#)]
50. Bozic, I.; Gerold, J.M.; Nowak, M.A. Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLoS Comput. Biol.* **2016**, *12*, e1004731. [[CrossRef](#)]
51. Stratton, M.R.; Campbell, P.J.; Futreal, P.A. The cancer genome. *Nature* **2009**, *458*, 719–724. [[CrossRef](#)]
52. McFarland, C.D.; Korolev, K.S.; Kryukov, G.V.; Sunyaev, S.R.; Mirny, L.A. Impact of deleterious passenger mutations on cancer progression. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 2910–2915. [[CrossRef](#)]
53. Wodarz, D.; Newell, A.C.; Komarova, N.L. Passenger mutations can accelerate tumour suppressor gene inactivation in cancer evolution. *J. R. Soc. Interface* **2018**, *15*, 20170967. [[CrossRef](#)]
54. Hofree, M.; Carter, H.; Kreisberg, J.F.; Bandyopadhyay, S.; Mischel, P.S.; Friend, S.; Ideker, T. Challenges in identifying cancer genes by analysis of exome sequencing data. *Nat. Commun.* **2016**, *7*, 12096. [[CrossRef](#)] [[PubMed](#)]
55. Reva, B.; Antipin, Y.; Sander, C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res.* **2011**, *39*, e118. [[CrossRef](#)] [[PubMed](#)]

56. Ghalamkari, S.; Alavi, S.; Mianesaz, H.; Khosravian, F.; Bahreini, A.; Salehi, M. A novel carcinogenic PI3K α mutation suggesting the role of helical domain in transmitting nSH2 regulatory signals to kinase domain. *Life Sci.* **2021**, *269*, 118759. [[CrossRef](#)] [[PubMed](#)]
57. Rayner, E.; Van Gool, I.C.; Palles, C.; Kearsey, S.E.; Bosse, T.; Tomlinson, I.; Church, D.N. A panoply of errors: Polymerase proofreading domain mutations in cancer. *Nat. Rev. Cancer* **2016**, *16*, 71–81. [[CrossRef](#)]
58. Frank, M.; Kemler, R. Protocadherins. *Curr. Opin. Cell Biol.* **2002**, *14*, 557–562. [[CrossRef](#)]
59. Rouget-Quermalet, V.; Giustiniani, J.; Marie-Cardine, A.; Beaud, G.; Besnard, F.; Loyaux, D.; Ferrara, P.; Leroy, K.; Shimizu, N.; Gaulard, P. Protocadherin 15 (PCDH15): A new secreted isoform and a potential marker for NK/T cell lymphomas. *Oncogene* **2006**, *25*, 2807–2811. [[CrossRef](#)]
60. Berger, M.F.; Lawrence, M.S.; Demichelis, F.; Drier, Y.; Cibulskis, K.; Sivachenko, A.Y.; Sboner, A.; Esgueva, R.; Pflueger, D.; Sougnez, C. The genomic complexity of primary human prostate cancer. *Nature* **2011**, *470*, 214–220. [[CrossRef](#)]
61. Kumar, A.; White, T.A.; MacKenzie, A.P.; Clegg, N.; Lee, C.; Dumpit, R.F.; Coleman, I.; Ng, S.B.; Salipante, S.J.; Rieder, M.J. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 17087–17092. [[CrossRef](#)]
62. Xu, S.; Moss, T.J.; Laura Rubin, M.; Ning, J.; Eterovic, K.; Yu, H.; Jia, R.; Fan, X.; Tetzlaff, M.T.; Esmaeli, B. Whole-exome sequencing for ocular adnexal sebaceous carcinoma suggests PCDH15 as a novel mutation associated with metastasis. *Mod. Pathol.* **2020**, *33*, 1256–1263. [[CrossRef](#)]
63. Lv, H.; Zhang, M.; Shang, Z.; Li, J.; Zhang, S.; Lian, D.; Zhang, R. Genome-wide haplotype association study identify the FGFR2 gene as a risk gene for acute myeloid leukemia. *Oncotarget* **2017**, *8*, 7891. [[CrossRef](#)]
64. Nasiri-Aghdam, M.; Garcia-Garduño, T.C.; Jave-Suárez, L.F. CELF Family Proteins in Cancer: Highlights on the RNA-Binding Protein/Noncoding RNA Regulatory Axis. *Int. J. Mol. Sci.* **2021**, *22*, 11056. [[CrossRef](#)] [[PubMed](#)]
65. Teerlink, C.C.; Stevens, J.; Hernandez, R.; Facelli, J.C.; Cannon-Albright, L.A. An intronic variant in the CELF4 gene is associated with risk for colorectal cancer. *Cancer Epidemiol.* **2021**, *72*, 101941. [[CrossRef](#)] [[PubMed](#)]
66. Chang, K.; Yuan, C.; Liu, X. A new RBPs-related signature predicts the prognosis of colon adenocarcinoma patients. *Front. Oncol.* **2021**, *11*, 354. [[CrossRef](#)] [[PubMed](#)]
67. Fan, X.; Liu, L.; Shi, Y.; Guo, F.; Wang, H.; Zhao, X.; Zhong, D.; Li, G. Integrated analysis of RNA-binding proteins in human colorectal cancer. *World J. Surg. Oncol.* **2020**, *18*, 222. [[CrossRef](#)] [[PubMed](#)]
68. Li, T.; Hui, W.; Halike, H.; Gao, F. RNA Binding Protein-Based Model for Prognostic Prediction of Colorectal Cancer. *Technol. Cancer Res. Treat.* **2021**, *20*, 15330338211019504. [[CrossRef](#)] [[PubMed](#)]
69. Huang, R.-L.; Su, P.-H.; Liao, Y.-P.; Wu, T.-I.; Hsu, Y.-T.; Lin, W.-Y.; Wang, H.-C.; Weng, Y.-C.; Ou, Y.-C.; Huang, T.H.-M. Integrated epigenomics analysis reveals a DNA methylation panel for endometrial cancer detection using cervical scrapings. *Clin. Cancer Res.* **2017**, *23*, 263–272. [[CrossRef](#)]
70. Geist, J.; Kontrogianni-Konstantopoulos, A. MYBPC1, an emerging myopathic gene: What we know and what we need to learn. *Front. Physiol.* **2016**, *7*, 410. [[CrossRef](#)]
71. Lee, D.Y.; Kang, Y.; Im, N.R.; Kim, B.; Kwon, T.K.; Jung, K.Y.; Baek, S.K. Actin-Associated Gene Expression is Associated with Early Regional Metastasis of Tongue Cancer. *Laryngoscope* **2021**, *131*, 813–819. [[CrossRef](#)]
72. Hu, H.; Wang, J.; Gupta, A.; Shidfar, A.; Branstetter, D.; Lee, O.; Ivancic, D.; Sullivan, M.; Chatterton, R.T.; Dougall, W.C. RANKL expression in normal and malignant breast tissue responds to progesterone and is up-regulated during the luteal phase. *Breast Cancer Res. Treat.* **2014**, *146*, 515–523. [[CrossRef](#)]
73. Zhang, H.; Wang, X.; Hou, C.; Yang, Z. Identification of Driver Genes and Interaction Networks Related to Brain Metastasis in Breast Cancer Patients. *Dis. Markers* **2022**, *2022*. [[CrossRef](#)]
74. Pudova, E.A.; Lukyanova, E.N.; Nyushko, K.M.; Mikhaylenko, D.S.; Zaretsky, A.R.; Snezhkina, A.V.; Savvateeva, M.V.; Kobelyatskaya, A.A.; Melnikova, N.V.; Volchenko, N.N. Differentially expressed genes associated with prognosis in locally advanced lymph node-negative prostate cancer. *Front. Genet.* **2019**, *10*, 730. [[CrossRef](#)] [[PubMed](#)]
75. Lee, Y.; Rio, D.C. Mechanisms and regulation of alternative pre-mRNA splicing. *Annu. Rev. Biochem.* **2015**, *84*, 291–323. [[CrossRef](#)] [[PubMed](#)]
76. Singh, R.K.; Cooper, T.A. Pre-mRNA splicing in disease and therapeutics. *Trends Mol. Med.* **2012**, *18*, 472–482. [[CrossRef](#)] [[PubMed](#)]
77. Cartegni, L.; Chew, S.L.; Krainer, A.R. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nat. Rev. Genet.* **2002**, *3*, 285–298. [[CrossRef](#)]
78. Wang, G.-S.; Cooper, T.A. Splicing in disease: Disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* **2007**, *8*, 749–761. [[CrossRef](#)]
79. Anczuków, O.; Buisson, M.; Salles, M.J.; Triboulet, S.; Longy, M.; Lidereau, R.; Sinilnikova, O.M.; Mazoyer, S. Unclassified variants identified in BRCA1 exon 11: Consequences on splicing. *Genes Chromosomes Cancer* **2008**, *47*, 418–426. [[CrossRef](#)]
80. Hansen, T.V.; Steffensen, A.Y.; Jønson, L.; Andersen, M.K.; Ejlertsen, B.; Nielsen, F.C. The silent mutation nucleotide 744 G→A, Lys172Lys, in exon 6 of BRCA2 results in exon skipping. *Breast Cancer Res. Treat.* **2010**, *119*, 547–550. [[CrossRef](#)]
81. Montera, M.; Piaggio, F.; Marchese, C.; Gismondi, V.; Stella, A.; Resta, N.; Varesco, L.; Guanti, G.; Mareni, C. A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. *J. Med. Genet.* **2001**, *38*, 863–867. [[CrossRef](#)]

82. Rentzsch, P.; Schubach, M.; Shendure, J.; Kircher, M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **2021**, *13*, 31. [[CrossRef](#)]
83. Jaganathan, K.; Panagiotopoulou, S.K.; McRae, J.F.; Darbandi, S.F.; Knowles, D.; Li, Y.I.; Kosmicki, J.A.; Arbelaez, J.; Cui, W.; Schwartz, G.B. Predicting splicing from primary sequence with deep learning. *Cell* **2019**, *176*, 535–548.e24. [[CrossRef](#)]
84. Chamary, J.-V.; Hurst, L.D. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* **2005**, *6*, R75. [[CrossRef](#)] [[PubMed](#)]
85. Presnyak, V.; Alhusaini, N.; Chen, Y.-H.; Martin, S.; Morris, N.; Kline, N.; Olson, S.; Weinberg, D.; Baker, K.E.; Graveley, B.R. Codon optimality is a major determinant of mRNA stability. *Cell* **2015**, *160*, 1111–1124. [[CrossRef](#)] [[PubMed](#)]
86. Ritz, J.; Martin, J.S.; Laederach, A. Evaluating our ability to predict the structural disruption of RNA by SNPs. *BMC Genom.* **2012**, *13*, S6. [[CrossRef](#)] [[PubMed](#)]
87. Gorochoowski, T.E.; Ignatova, Z.; Bovenberg, R.A.; Roubos, J.A. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Res.* **2015**, *43*, 3022–3032. [[CrossRef](#)] [[PubMed](#)]
88. Duan, J.; Wainwright, M.S.; Comeron, J.M.; Saitou, N.; Sanders, A.R.; Gelernter, J.; Gejman, P.V. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* **2003**, *12*, 205–216. [[CrossRef](#)] [[PubMed](#)]
89. Halvorsen, M.; Martin, J.S.; Broadaway, S.; Laederach, A. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.* **2010**, *6*, e1001074. [[CrossRef](#)]
90. Solem, A.C.; Halvorsen, M.; Ramos, S.B.; Laederach, A. The potential of the riboSNitch in personalized medicine. *Wiley Interdiscip. Rev. RNA* **2015**, *6*, 517–532. [[CrossRef](#)]
91. Pereira, B.; Billaud, M.; Almeida, R. RNA-binding proteins in cancer: Old players and new actors. *Trends Cancer* **2017**, *3*, 506–528. [[CrossRef](#)]
92. Wang, Z.-L.; Li, B.; Luo, Y.-X.; Lin, Q.; Liu, S.-R.; Zhang, X.-Q.; Zhou, H.; Yang, J.-H.; Qu, L.-H. Comprehensive genomic characterization of RNA-binding proteins across human cancers. *Cell Rep.* **2018**, *22*, 286–298. [[CrossRef](#)]
93. Gerstberger, S.; Hafner, M.; Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **2014**, *15*, 829–845. [[CrossRef](#)]
94. Singh, B.; Trincado, J.L.; Tatlow, P.; Piccolo, S.R.; Eyra, E. Genome sequencing and RNA-motif analysis reveal novel damaging noncoding mutations in human tumors. *Mol. Cancer Res.* **2018**, *16*, 1112–1124. [[CrossRef](#)] [[PubMed](#)]
95. Teng, H.; Wei, W.; Li, Q.; Xue, M.; Shi, X.; Li, X.; Mao, F.; Sun, Z. Prevalence and architecture of posttranscriptionally impaired synonymous mutations in 8,320 genomes across 22 cancer types. *Nucleic Acids Res.* **2020**, *48*, 1192–1205. [[CrossRef](#)] [[PubMed](#)]
96. Huang, F.W.; Hodis, E.; Xu, M.J.; Kryukov, G.V.; Chin, L.; Garraway, L.A. Highly recurrent TERT promoter mutations in human melanoma. *Science* **2013**, *339*, 957–959. [[CrossRef](#)]
97. Mansour, M.R.; Abraham, B.J.; Anders, L.; Berezovskaya, A.; Gutierrez, A.; Durbin, A.D.; Etchin, J.; Lawton, L.; Sallan, S.E.; Silverman, L.B. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **2014**, *346*, 1373–1377. [[CrossRef](#)] [[PubMed](#)]
98. Cibulskis, K.; Lawrence, M.S.; Carter, S.L.; Sivachenko, A.; Jaffe, D.; Sougnez, C.; Gabriel, S.; Meyerson, M.; Lander, E.S.; Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **2013**, *31*, 213–219. [[CrossRef](#)] [[PubMed](#)]