BMC Bioinformatics

RESEARCH

Open Access

Incorporating mutational heterogeneity to identify genes that are enriched for synonymous mutations in cancer

Yiyun Rao[1], Nabeel Ahmed[1,5], Justin Pritchard[2*] and Edward P. O'Brien[3,4*]

*Correspondence:
jrp94@psu.edu; epo2@psu.edu

[1] Huck Institute of the Life Sciences, Pennsylvania State University, University Park, State College, PA 16802, USA
[2] Department of Biomedical Engineering, Pennsylvania State University, University Park, State College, PA 16802, USA
[3] Department of Chemistry, Pennsylvania State University, University Park, State College, PA 16802, USA
[4] Institute for Computational and Data Sciences, Pennsylvania State University, University Park, State College, PA 16802, USA
[5] Moderna, Inc., Cambridge, USA

Abstract

Background: Synonymous mutations, which change the DNA sequence but not the encoded protein sequence, can affect protein structure and function, mRNA maturation, and mRNA half-lives. The possibility that synonymous mutations might be enriched in cancer has been explored in several recent studies. However, none of these studies control for all three types of mutational heterogeneity (patient, histology, and gene) that are known to affect the accurate identification of non-synonymous cancer-associated genes. Our goal is to adopt the current standard for non-synonymous mutations in an investigation of synonymous mutations.

Results: Here, we create an algorithm, MutSigCVsyn, an adaptation of MutSigCV, to identify cancer-associated genes that are enriched for synonymous mutations based on a non-coding background model that takes into account the mutational heterogeneity across these levels. Using MutSigCVsyn, we first analyzed 2572 cancer whole-genome samples from the Pan-cancer Analysis of Whole Genomes (PCAWG) to identify non-synonymous cancer drivers as a quality control. Indicative of the algorithm accuracy we find that 58.6% of these candidate genes were also found in Cancer Census Gene (CGC) list, and 66.2% were found within the PCAWG cancer driver list. We then applied it to identify 30 putative cancer-associated genes that are enriched for synonymous mutations within the same samples. One of the promising gene candidates is the B cell lymphoma 2 (BCL-2) gene. BCL-2 regulates apoptosis by antagonizing the action of proapoptotic BCL-2 family member proteins. The synonymous mutations in BCL2 are enriched in its anti-apoptotic domain and likely play a role in cancer cell proliferation.

Conclusion: Our study introduces MutSigCVsyn, an algorithm that accounts for mutational heterogeneity at patient, histology, and gene levels, to identify cancer-associated genes that are enriched for synonymous mutations using whole genome sequencing data. We identified 30 putative candidate genes that will benefit from future experimental studies on the role of synonymous mutations in cancer biology.

Keywords: MutSigCV, Cancer driver, Synonymous mutations

Background

'Driver' mutagenic events confer a selective growth advantage to cells and contribute to tumorigenesis [1, 2]. Discovering and characterizing these cancer-driver genes using large-scale cancer genome sequencing data is a major component of modern cancer research [2, 3]. These drivers are typically identified through aberrantly high mutation rates in specific genes relative to an estimate of the background mutation rate [4–6]. Classic efforts have identified a "long-tail" distribution of cancer driver mutations, where some mutations (*e.g.*, KRAS G12D [7]) are highly prevalent, and other mutations are extraordinarily rare [8, 9]. However, many tumors do not harbor any known cancer drivers. A reasonable assumption is that these tumors harbor driver mutations that are rare enough to be undetectable in existing cohorts [10]. The unambiguous detection of these novel long-tail drivers is a challenge because of the underpowered sample size of many cohorts [11]. However, it may also be a challenge because research labs have primarily looked for cancer drivers involving non-synonymous mutations or non-coding mutations in promoters and other regulatory regions [12, 13].

Synonymous mutations are one class of historically disregarded mutations that might be long-tail drivers. Synonymous mutations alter the mRNA coding sequence but not the encoded protein's primary structure. In the past, these mutations were assumed to be phenotypically "silent" [14, 15]. Nonetheless, synonymous codons encode information beyond amino acids. Protein structure and function can be altered by introducing synonymous mutations that change the rate of protein translation [16–18]. Such variation had been found to affect co-translational folding [19], translational accuracy [20], and posttranslational modifications [21]. Additionally, synonymous mutations also play a regulatory role in transcription by altering mRNA structure [22], and in some cases affecting the mRNA splicing process [23]. Both of these translational and transcriptional effects had been found to impact cell fitness in bacteria [18, 24], and are linked to a number of human diseases [25]. It is now generally accepted that synonymous mutations can affect subcellular processes and phenotypes [26, 27].

Two sets of evidence indicate that selective constraints act at synonymous mutation positions in cancer, suggesting a functional role. First, bioinformatic analyses indicate a global selection for synonymous mutations in oncogenes. Supek et al. [28] found that the synonymous mutation rate is elevated in oncogenes, especially near exon–intron boundaries, regardless of local mutation rates. Analyses from Chu et al. [29] on single nucleotide polymorphisms (SNPs) in healthy patients suggested synonymous SNP sites in cancer-related genes may undergo a selection constraint, and are more conservative in oncogenes than in other cancer-related genes. In addition, results from Benisty et al. [30] suggest that the frequently mutated oncogene in oncogene families (e.g., KRAS) may adapt codon usage to promote cancer cell proliferation. Second, circumstantial evidence connects synonymous mutations and cancer. For example, synonymous mutations in the MDR1 gene, which encodes the efflux pump Pgp, contribute to chemotherapy resistance [31]. In cancer cells, synonymous SNPs in MDR1 affect P-glycoprotein substrate specificity. And synonymous mutations in BAP1 were found to cause exon11 skipping, generating a premature stop codon, and thus a complete loss of function for BAP1 [32]. These findings suggest that synonymous mutations are possibly cancer-associated.

To identify cancer-associated genes that are enriched for synonymous mutations, one of the key aspects is a comprehensive model to estimate background synonymous mutation rates. Several studies have used a variety of computational approaches [28, 33–35]. The background models in these studies have ranged in complexity and sophistication. For example, in the seminal study by Supek et al. [28], thirteen covariates at the gene level controlled for regional mutation variation between non-cancer genes and oncogenes of interest, but patient-level biases were not accounted for. In another study, Sharma et al. [33] examined and ranked common synonymous mutations in COSMIC [36] (a curated database of somatic mutations in cancer) and combined this with orthogonal data including mRNA secondary structural change predictions as well as evolutionary conservation score. However, this approach did not have a formal estimate of the background synonymous mutation frequency. No approach to date has accounted for all three levels of patient-, gene- and disease-specific mutational heterogeneity that are known to lead to inaccurate results in non-synonymous cancer identification [4, 37], and are certain to affect the identification of cancer-associated genes that are enriched for synonymous mutations.

Controlling for patient-, gene- and disease-specific mutation biases is exemplified by the MutSigCV [4] algorithm, which is the community standard for driver identification in non-synonymous mutations. Here, we bring this same level of background mutational modeling to synonymous mutations by developing an algorithm we refer to as MutSigCVsyn, which allows us to detect putative synonymous candidates while controlling for confounding mutational biases. This approach is enabled by The Pan-Cancer Analysis of Whole Genomes (PCAWG) sequencing data [38] from which we use the non-coding mutations within genic regions to adjust for triplet nucleotide mutation biases across diverse patients, tumor histologies, and genes. With this approach, we identify 30 putative genes that are enriched for synonymous mutations across 18 histology cohorts.

Results

Synonymous mutation rate varies across patients, tumor types, and genes, impeding cancer-associated gene discovery

The accurate identification of non-synonymous drivers requires explicit corrections for background mutation biases across patients, genes, and diseases. We first examined if the same should be done when identifying genes that are enriched for synonymous mutations because it is highly likely that there are distinct synonymous mutation rates across these categories.

To demonstrate this synonymous mutation heterogeneity, we collected synonymous mutations in 18,638 protein-coding genes across 2,572 PCAWG patients. We calculated the rate of synonymous substitutions per 1 Mbp synonymous site for each indication (see Methods Section). As expected, the synonymous mutation rate was lower than the total mutation rate (Fig. 1a, top). We observed that the synonymous mutation frequencies vary widely across patients and histology indications. Across the indications, Skin-Melanoma has the highest median synonymous mutation frequencies across patients at 21.7 per Mbp. Towards the other extreme, the lowest median frequency is observed in CNS-PiloAstro (0 per Mbp, due to patients having no synonymous mutations) which is over 20 times smaller than Skin-Melanoma.

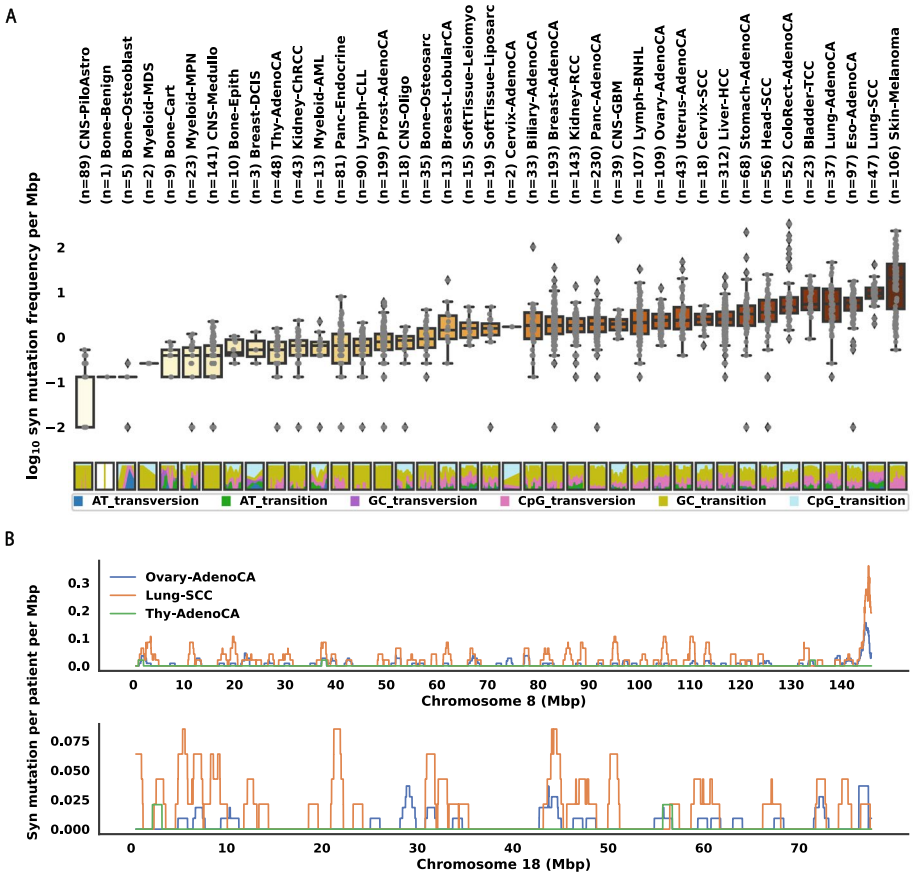

Fig. 1 The synonymous mutation rate in cancer varies across patients, histology types, and genes. **A** Box plot (Top) of patient synonymous mutation frequency across all histology types. Mutation frequency is shown as a logarithmically transformed mutation number per mega base pair. Patients that don't have any synonymous mutations are set to have -2 transformed mutation frequency per mega base pair. Each dot represents a patient. Histology types are ordered by their median somatic mutation frequency. The relative percentage (bottom) of mutations falls into 6 mutation categories (see Methods Section) for all individual patients across the histology types. **B** Synonymous mutation number averaged by number of patients in Ovary-AdenoCA (blue), Lung-SCC (orange), and Thy-AdenoCA (green), respectively, illustrated on the entire chromosome 8 (top) and chromosome 18 (bottom)

We also observe large variations in mutation frequency within individual cancer indications. Except for some of the extremely small cohorts (e.g., Bone-Benign (n = 1), Bone-Osteoblast (n = 5), Myeloid-MDS (n = 2), Cervix-AdenoCA (n = 2)), the maximum mutation frequency is at least 1 order of magnitude larger than the minimum in each indication. The largest such variation occurs in ColoRect-AdenoCA, where the highest synonymous mutation frequency is 329 per Mbp, while the lowest is 0.917 per Mbp. This is consistent with the existence of a hypermutated microsatellite instability subpopulation [39].

These variations are partly explained by mutational etiology (Fig. 1a, bottom). A typical example is Skin-Melanoma, which exhibits an enrichment of GC transition mutations, consistent with the known mutational signature due to UV radiation [40]. In addition, the high content of GC transition in Bladder-TCC patients is likely caused by APOBEC protein family activity, which is a prominent mutational signature pattern in

TCGA bladder tumors [41]. In Lung-SCC, we also observe signs of signatures related to tobacco smoke, which is characterized by G to T transversion caused by lesions when polycyclic aromatic hydrocarbons enter the human body [42]. Thus, as expected, known mutational signatures contribute to synonymous mutation heterogeneity as well.

Next, in order to illustrate the heterogeneity of mutation rate across genomes for a given cancer indication, we plotted the average synonymous mutation number per patient across chromosome 8 and chromosome 18 for 3 histology cohorts(Ovary-AdenoCA, Lung-SCC, Thy-AdenoCA). As shown in Fig. 1b, variation of local mutation numbers is observed across all 3 histology types.

These results demonstrate that there is substantial variability in the synonymous mutation burden at the histology, patient, and gene levels. Therefore, the assumption of a constant mutation rate and completely independent mutation events is not appropriate for the identification of genes that are enriched for synonymous mutations in cancer. To accurately identify such genes, driver predictions must explicitly correct for these covariates.

MutSigCVsyn detects differences between observed and expected synonymous mutation frequencies in cancer cohorts

In order to correct for these covariates, especially the gene-specific differences in mutation rate, we adopted and modified MutSigCV [4] (Fig. 2a), which corrects for variation by using patient-specific mutation frequencies and the 192-triplet nucleotide mutation context (e.g., A(A->C)A), and gene-specific background mutation rates through the incorporation of expression level and chromosome replication position.

MutSigCV was originally designed for the identification of non-synonymous drivers in the context of exome sequencing data. To convert MutSigCV into an algorithm detecting synonymous mutation enrichment, we made several modifications (Fig. 2b). The biggest modification is using only the non-coding mutations in our background mutation model. The original MutSigCV's background model is composed of synonymous mutations and non-coding mutations found in the untranslated regions of transcripts but with limited coverage in non-coding regions. This is because it was originally designed for cancer exome re-sequencing datasets. However, the high data quality and coverage in PCAWG Whole Genome sequencing datasets allow us to use the mutations in the complete intronic region and untranslated regions for the mutational background. The two major reasons for using such a background are: (1) we adopted a simplifying assumption that on average, non-coding mutations are 'more neutral' than the synonymous mutations. The lower rate in the intronic region than in exonic regions across species [43] suggests that non-coding regions of genes are under weaker selection than the coding region. (2) By restricting the non-coding mutations to the mutations occurring in transcribed regions, we prevent bias caused by different mutation frequencies in transcribed versus non-transcribed regions. Specifically, the non-coding mutations in our analysis only include (a) intronic mutations and (b) mutations in untranslated regions.

In MutSigCVsyn, protein-coding gene coverage information for every patient in PCAWG is calculated. In addition, we re-annotated the gene covariate file to adapt the gene name annotation in PCAWG. To benefit the community, the files and scripts are

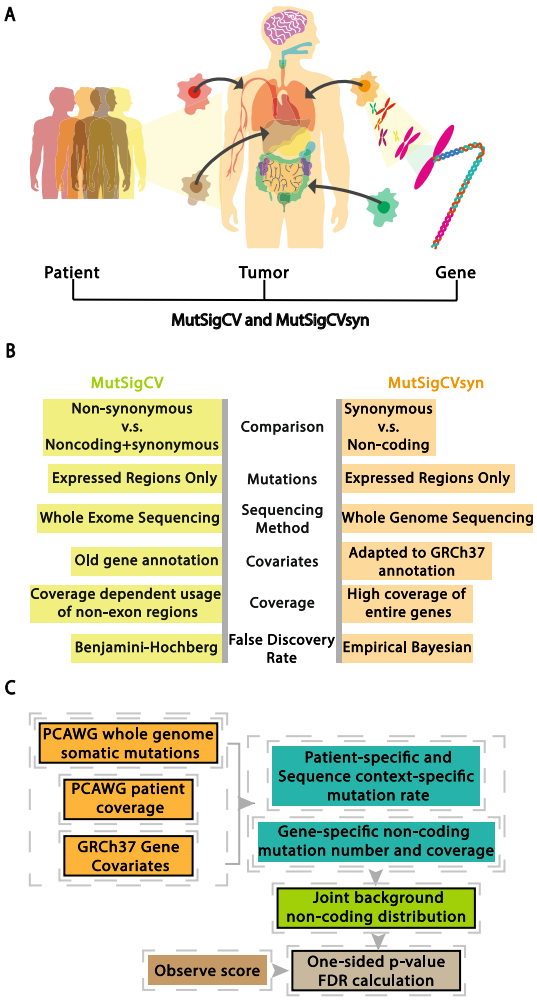

Fig. 2 Changing MutSigCV to MutSigCVsyn to identify genes that are enriched for synonymous mutations. **A** MutSigCV accounts for mutation heterogeneity across patients, diseases, and genes. **B** Comparison between MutSigCV and MutSigCVsyn: (1) MutSigCVsyn uses only non-coding mutations instead of a background comprised of both non-coding and synonymous mutations adopted by a majority of driver mutation detection algorithms. (2) MutSigCVsyn utilizes whole genome sequencing input data instead of whole-exome sequencing. (3) Both MutSigCVsyn and MutSigCV only utilize mutations in transcriptionally expressed regions. (4) MutSigCVsyn utilizes a re-annotated covariate file that was adapted to the PCAWG Gencode v19 annotation. (5) MutSigCVsyn patients have high-quality coverage data over non-coding regions, compared to limited coverage in the original MutsigCV. (6) MutSigCVsyn utilizes a non-parametric empirical Bayesian method to calculate the local FDR value. **C** The outline of MutSigCVsyn. Boxes with solid lines show MutSigCVsyn exclusive input/steps (see Methods section for detailed description)

available publicly on GitHub. The workflow of MutSigCVsyn is shown in Fig. 2c. A more detailed description of MutSigCVsyn can be found in the Methods Section.

Quality control: MutSigCVsyn identifies non-synonymous drivers with high sensitivity. MutSigCVsyn is designed for the identification of synonymous drivers. However, if MutSigCVsyn builds a valid non-coding background, MutSigCVsyn should be able to identify non-synonymous drivers as well. Therefore, as quality check for our approach, we applied MutSigCVsyn to 2572 donors in 39 PCAWG histology types to identify non-synonymous drivers, using non-coding mutations as background.

We identified a total of 133 significant genes (Additional file 1: Figure S1) across 29 cohorts. As expected, most of the genes in the candidate gene list have been reported before. As the most frequently altered gene in human cancer, TP53 is the most frequently significant driver across all indications. It is called significant in 21 out of 39 histology types, including ColoRect-AdenoCA, Lymph-BNHL, Liver-HCC, and Panc-AdenoCA. Furthermore, our significant driver list for each indication overlaps the known cancer drivers in that indication. We identified candidate genes CDKN2A in HCSCC (Head and Neck Cancer), which is a known tumor suppressor and whose inactivation has been well studied in HCSCC [44]. In CRC (Colorectal Cancer), APC and SMAD4 are also identified as the candidates. APC constitutively activates the canonical WNT signaling in most colorectal cancer cases, leading to cell proliferation and tumor formation [45]. Another known gene, SMAD4 [46], which negatively regulates TGF-beta, is also frequently found in CRC patients. Finally, our results in Breast-AdenoCA also highlighted some genes that are specifically known to be frequently mutated in breast cancer [47], including PIK3CA, CDH1, GATA3, and MAP2K4.

As a further test of our result, we compared our output to CGC (Cancer Gene Census) and PCAWG driver list (see Methods Section) (Fig. 3a). We observed 58.6% (78 out of 133) of our non-synonymous list overlaps with the CGC genes. The high overlap rate may be due to the nearly full coverage of non-coding regions and the accurate calculation of the coverage file for the analysis. We also observe 66.2% (88 out of 133) of our candidate genes overlap with PCAWG drivers. Additionally, we successfully identified 6 genes out of the 15 PCAWG exclusive drivers (Additional file 1: Figure S3), which are the genes identified in the PCAWG cohort for the first time. In conclusion, these results indicate that our modifications to MutSigCV do not dramatically affect the ability of MutSigCVsyn to reproduce previously known results.

The landscape of genes that are enriched for synonymous mutations in cancer

Given our ability to identify non-synonymous drivers with high sensitivity, we used MutSigCVsyn to identify genes that are enriched for synonymous mutations in all 39 histology types in PCAWG. We identified 30 putative synonymous candidates in total (Fig. 3b; Additional file 3: Table S6). As expected, this list is parsimonious and smaller than the non-synonymous driver list. Lymph-BNHL has the most significant synonymous candidates (n = 5), followed by Panc-AdenoCA(n = 4). In total, there are 18 distinct indications having significant genes. The variety of indications implies that MutSigCVsyn is not biased by histology-wise mutation frequencies. Among all candidates, 11 genes across 7 indications have the smallest p-values (p-value $< 1.0 \times 10^{-7}$), including BCL2 and SRSF2 (Lymph-BNHL), ITLN1 (CNS-PiloAstro), PPWD2 (Head-SCC), PURA and MAGEC1 (Breast-AdenoCA), SIGLEC15 and TP53I3 (Panc-AdenoCA).

Two of the top candidates, BCL-2 and SRSF2, are known to be non-synonymous drivers of cancer as cataloged in the Cancer Gene Census. Both genes were identified in the Lymph-BNHL cohort. The t(14;18) translocation in BCL-2 is critical in follicular lymphoma progression [48] and SRSF2 is a global splicing regulator that binds to exonic splicing motifs. It is associated with hematopoietic diseases (i.e., myelodysplastic syndrome [49]) but hadn't been specifically characterized in Non-Hodgkin Lymphoma. In PCAWG, 3 unique Lymph-BNHL patients have 3 distinct synonymous

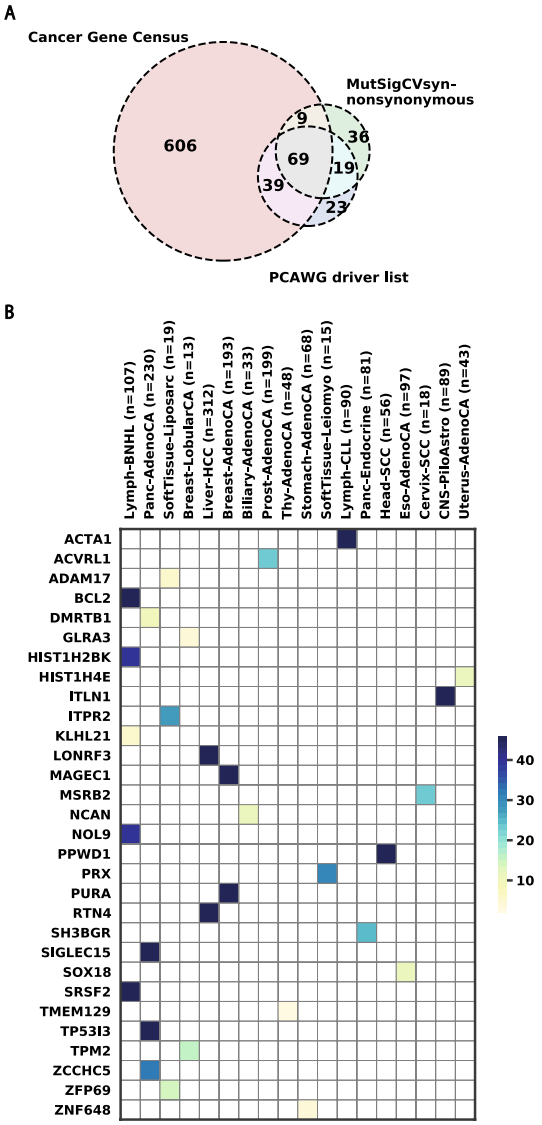

Fig. 3 MutSigCVsyn identifies non-synonymous and synonymous cancer-associated genes. **A** Venn Diagram displaying overlapped gene numbers of MutSigCVsyn significant non-synonymous drivers with Cancer Gene Census and PCAWG driver lists. **B** Heatmap shows significant synonymous candidate genes (Bayesian FDR $< 1 \times 10^{-2}$) identified by MutSigCVsyn. Genes are colored by the -log10 transformed FDR value from high (dark blue) to low (light yellow). A table containing the exact FDR values can be found in Additional file 3: Table S6

mutations in SRSF2: p.Y3Y (DO27764), p.V79V (DO52664), and p.G82G (DO52672), the latter 2 reside in the RNA recognition motif (RRM) of SRSF2. Though one of the patients (DO52664) carried a missense mutation at the Proline95 position that is known to alter mRNA binding affinity [50], 2 other patients only harbor SRSF2 synonymous mutations. As synonymous mutations can perturb mRNA translation initiation and elongation processes [51], it is possible that SRSF2 synonymous mutations alter RRM binding affinity and contribute to a global transcriptional profile change in cancer cells.

While many of the candidate genes are poorly studied in cancer, there is evidence to suggest some of them could be required for tumor growth. For example, PURA, which encodes the nucleic acid-binding proteins Purα, is one of the significant candidates in Breast-AdenoCA. Studies have found that overexpression of PURA inhibits proliferation and anchorage-independent colony formation of Ras-transformed NIH3T3 Fibroblast cells, suggesting PURA acts as a potential tumor suppressor gene [52]. In our analysis, the PURA expression level is significantly lower (Mann–Whitney U-test p-value $= 5 \times 10^{-3}$) in tumor samples (n = 85) than in normal samples (n = 6) (Additional file 1: Figure S2a). This low expression suggests a plausible contribution to breast cancer cell proliferation. Another example is the immune checkpoint gene SIGLEC15, the top significant gene in Panc-AdenoCA. SIGLEC15 is a well-conserved member of the immunoglobulin superfamily of receptor Siglecs that bind to sialic acid. In a recent study [53], upregulated SIGLEC15 has been widely found across different cancer types and had been related to a worse patient survival rate. Moreover, SIGLEC15, rather than other immune checkpoint genes, was found to have a positive expression correlation with upregulated genes in pancreatic cancer [54]. We observe a significantly higher expression of SIGLEC15 mRNA expression in Pancreas exocrine lineage cancer cell lines than in other cell lines in DepMap [82] (Mann–Whitney U-test p-value $= 5 \times 10^{-3}$) (Additional file 1: Figure S2b). We used DepMap data because the transcriptome data of PCAWG pancreatic patient normal specimens is unavailable. Combining this with the observation of SIGLEC15's mutually exclusive expression with B7-H1(PD-L1) [55] suggests that SIGLEC15 levels may play a role in pancreatic cancer immune evasion. The role of synonymous mutations in both cases may be a fruitful area of future study.

MutSigCVsyn exclusive synonymous candidates might contribute to cancer

We expect that the significant candidates called by MutSigCVsyn have the potential to contribute to a cancer phenotype. We focus on one of our particular candidates, BCL-2, that has compelling cancer associations. BCL-2 (B cell lymphoma 2) regulates apoptosis by antagonizing the action of proapoptotic BCL-2 family members [56]. It was originally identified as the proto-oncogene involved in the t(14;18) translocation in follicular lymphoma [57]. Among the BCL-2 protein motifs, the BH4 motif is essential for the anti-apoptotic activity of BCL-2. The deletion of the BH4 region in a human fibroblast cell line largely impairs cell viability under IL-3 deprivation [58] and melanoma growth in vitro and in vivo [59].

In our analysis, we observe 41 synonymous mutations in 26 unique patients, and 9 of the mutations in 9 different patients reside in the BH4 motif (Fig. 4a). Combining this observation with the known anti-apoptotic effect of the BH4 motif, we hypothesize that there may be an enrichment for synonymous mutations in the BH4 motif that might enhance its function and thus promote cancer cell survival. If this is true, we would expect to see a significant enrichment of synonymous mutations in the BH4 motif versus the BH1, 2, or 3 motifs. To test for enrichment, we conducted a permutation test for the observed number of mutations in the BH4 motif by comparing it to the 10,000 permutations where all 39 mutations are randomly assigned across all the BCL-2 coding positions. The results show that BCL-2 synonymous mutations are significantly enriched in

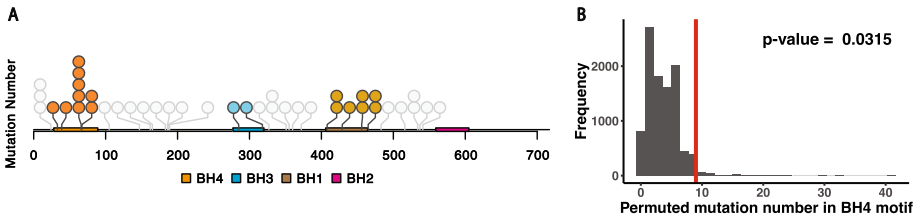

Fig. 4 Synonymous mutations are significantly enriched in BCL-2's BH4 motif. **A** Illustration and distribution of all BCL-2 synonymous mutations identified in PCAWG Lymph-BNHL patients across the BCL-2 coding sequence. Circles represent the occurrence of each synonymous mutation. BCL-2 motifs and the synonymous mutation in those motifs are colored: BH4 (Orange), BH3 (Blue), BH1 (Brown), and BH2 (Bright Pink). No synonymous mutations were observed in BH2 motif. Synonymous mutations that fall outside of the motifs are colored grey. **B** BH4 synonymous mutation number distribution from permutation test (10,000 permutations). The red line shows the observed number of synonymous mutations (n = 9)

the BH4 motif (p-value = 0.032) (Fig. 4b). This indicates that synonymous mutations in the BCL-2 BH4 motif might be positively selected for in lymphomas.

Discussion

MutSigCVsyn controls for the patient-, histology-, and gene-specific mutation rate variations to identify genes that are enriched for synonymous mutations in cancer. What is novel about this approach compared to previous ones is that the background mutation model we constructed accounts for the covariates that are standards in identifying non-synonymous drivers. Without adjusting for these covariates, there is a high likelihood of misidentifying synonymous candidates. To test this approach, we reasoned that our new background mutation model should still be able to identify known non-synonymous drivers. And indeed, we find MutSigCVsyn identifies above 60% of the drivers reported in the CGC. 60% is a reasonably high success rate, given that an evaluation of eight different driver-gene-detection algorithms [60] found that they identified between ~ 10% and 50% of the drivers in CGC.

By applying MutSigCVsyn to the PCAWG database, we identified 30 putative genes that are enriched for synonymous mutations. Among them, BCL-2 appears to be the most promising candidate due to the extensive literature concerning its role in follicular lymphoma [48, 61, 62] and the significant clustering of synonymous mutations in BCL-2's BH4 regulatory motif. Thus, we hypothesize that synonymous mutations in the BH4 motif might contribute to BCL-2's gain-of-function role in oncogenesis. One potential argument against this hypothesis is that the enrichment of mutations in BCL-2 is the result of somatic hypermutation caused by activation-induced cytidine deaminase, which is frequent in immunoglobulin variable regions [63]. However, we find that only 4 of the 26 patients that have synonymous mutations in BCL-2 harbor an IgG translocation in this gene. Further, the breakpoints of the BCL2 translocation within these 4 patients are at least 100kbp away from the observed synonymous mutations—a distance that is not consistent with hypermutations in immunoglobulin variable regions. And most importantly, the background estimate of the activation-induced-cytidine-deaminase signature is already accounted for in the MutSigCVsyn analysis by integrating mutational context, as well as other candidate hypermutations in lymphomas, meaning they are statistically excluded from our candidate list. Thus, these results suggest that the

positively selected signal of BCL-2 in our analysis is probably not confounded, and the synonymous mutations might be one of the significant causes that result in a gain-of-function effect in lymphoma patients.

The divergence at nonsynonymous and synonymous sites in cancer cohorts, known as the dN/dS ratio, is a conventional measure of evolutionary selection pressure [64]. It has been applied in many somatic evolution studies [65–68] under the assumption that nearly all synonymous mutations are neutral [14]. A small dN/dS ratio is usually interpreted as a global signal of negative selection on non-synonymous mutations. However, the signal we observe in the analysis opens the possibility that the synonymous mutations in BCL-2 challenge this interpretation. In a study by Lohr et al. [69], a small dN/dS ratio was found across the entire BCL-2 gene in a 50 diffuse large B-Cell lymphoma patient cohort. It was thus concluded that BCL-2 undergoes strong negative selection. Contrary to this, our study suggests that an increase in dS creates a robust positive selection signal of synonymous mutations in the BH4 motif of BCL-2. Thus, it may not entirely be that evolution is selecting negatively on the numerator dN, but rather, positively on the denominator dS. Therefore, the possibility exists that the negative selection pressures on BCL-2 are overestimated when only using the dN/dS ratio across the entire gene. More broadly, this indicates that the interpretation of the dN/dS ratio may not be straightforward when synonymous mutations are not neutral.

Except for BCL-2, most of the other candidate genes identified by MutSigCVsyn (Fig. 3b) have not been identified previously. To evaluate their functional relevance, we carried out a mutual exclusivity analysis, a Combined Annotation Dependent Depletion (CADD) [70] score prediction analysis, and a gProfiler [71] Gene Ontology Enrichment Analysis. Detailed rationale, methods, and results are described in Additional file 2. As a result, we didn't identify any statistically significant mutual exclusive gene pairs due to the lack of power in the PCAWG dataset. However, we observed that the CADD score (p-value$=3\times10^{-14}$, n$=137$) is significantly higher for synonymous mutations in candidate genes compared to the ones in other genes (Additional file 1: Figure S2C), which indicates that the synonymous mutations in candidate genes are more likely to be functional.

We did a comprehensive analysis comparing our findings to previous literature on cancer-associated synonymous mutations, including Sharma et al. [33], Zeng et al. [72], and Bin et al. [34]. Upon analysis, we discovered several overlapping genes of which variants were listed as potential cancer-associated genes in the other papers: Variants in BCL2 (Lymph-BNHL), SRSF2 (Lymph-BNHL) and TMEM129 (Thy-AdenoCA) were listed among top 1000 synMICdb variants, and we observed 3 exact BCL2 mutations in 5 unique patients in our dataset were within the top 1% in synMICdb (Additional file 3: Table S4). Furthermore, in Zeng et al., besides BCL2, variants in our candidate gene PURA (Breast-AdenoCA) were also identified as proposed cancer-associated mutations. Although these variants might not have been observed within the same cohort as in the PCAWG dataset, they provide evidence supporting that the synonymous mutations are likely being positively selected in these candidate genes.

Differences in datasets and methodology are two reasons differences in the published lists of synonymous 'drivers' can arise. For example, PCAWG, which we used in this study, is less comprehensive than COSMIC in terms of the number and source of

identified synonymous mutations. However, PCAWG uses a uniform mutation calling standard that ensures variant calling accuracy by validating calls across multiple variants calling pipelines, whereas COSMIC uses human curation of publications reporting somatic mutation results based on heterogeneous analysis standards (e.g., differences in alignments, variant callers, manual annotations) – which are known to affect accuracy [73].

By not accounting for covariates in a background mutation model, studies can identify spurious synonymous mutation candidates. In one study [34], multiple mutations in two extremely long human genes, which encode the muscle protein titin and neuronal synaptic vesicle protein piccolo, were identified as synonymous candidates. However, these genes are commonly observed as false positives in non-synonymous driver identification studies that don't account for gene-specific mutational biases [4]. In another study [33], two top synonymous candidates are present in highly mutable microsatellite regions: MLLT3 (c.501T > C) has the 5th highest SynMICdb score, and ARID1B (c.768C > A) has the 13th highest SynMICdb score. Creating an appropriate background mutation model minimizes such microsatellite biases. Therefore, these putative false-positive results highlight the importance of methodologies that utilize comprehensive background mutation models, especially when identifying weak and rare signals like synonymous mutations.

Our study has limitations. Firstly, due to the rarity and weaker signals of cancer-contributing synonymous mutations compared to missense or nonsense mutations, we lack statistical power to detect synonymous candidates in smaller cohorts like bone neoplasm subtype cohorts ($n < 10$) and myeloid cohorts ($n < 30$). Secondly, the PCAWG dataset has limited transcriptome data, with only 1188 patients from 27 cohorts out of 2572 high-quality patient samples across 39 cohorts having available transcriptome data. This restricts our ability to perform in-depth bioinformatic analysis on significant gene expression and splicing patterns. These limitations emphasize the need for larger sample sizes and complete RNA-seq and WGS datasets for comprehensive analysis. Moreover, we acknowledge the drawbacks of using non-coding mutations in the UTR and intronic regions to build the mutational background used in our approach. Sequences in some non-coding regions are under evolutionary constraints, especially regulatory elements, such as intron–exon junctions [74, 75]. A positively selected non-coding background may diminish the synonymous mutation signal and decrease the number of synonymous candidates. For these reasons, it would be useful in future studies to exclude specific background regions that are already known to be under evolutionary selection. However, principled exclusion criteria will require much larger cohorts and more complete knowledge of positive selection in non-coding regions of the genome.

Conclusion

MutSigCVsyn attempts to identify genes that are enriched for synonymous mutations in cancer using the paradigm that is commonly found in algorithms for non-synonymous cancer drivers. We have identified a list of 30 putative synonymous candidates that provide opportunities for future experimental research to understand how these synonymous mutations within the candidate genes can contribute to cancer.

Methods

Dataset

The patient MAF (Mutation Annotation Format) files, wig coverage files, RNA-seq data, and cancer driver data were retrieved from the PCAWG portal (https://dcc.icgc.org/releases/PCAWG). Driver genes in Cancer Gene Census were retrieved from the COSMIC website (https://cancer.sanger.ac.uk/cosmic). Gene sequence and annotation data were downloaded from the Gencode Release19 website (https://www.gencodegenes.org/human/release_19.html). CERES scores were retrieved from the DepMap web portal (https://depmap.org/portal/).

Patient and geneset

2572 PCAWG [38] patients whose SNV mutation information and wig coverage files both exist were selected. The selected patients are 'white-listed' in the PCAWG dataset. This means that they have met strict criteria for data quality. The patients were divided into 39 histology cohorts based on the PCAWG annotation. 139 patients who have a total mutation number > 50,000 [13] were defined as hypermutators and were excluded from MutSigCVsyn analysis.

Only protein-coding genes were selected for MutSigCVsyn analysis. As the PCAWG SNVs (Single Nucleotide Variants) were annotated based on Gencode v19, known protein-coding genes in Gencode v19 were selected based on filter "KNOWN" and "protein_coding" in the Gencode v19 gene annotation file. The 'principal' [76] transcript, if exists, was used. Otherwise, the longest transcript was used. To make sure all mutations were correctly accounted for in the coverage file as in the MAF file, genes of which the coding/intron/UTR SNV positions don't match between the MAF files and the coverage files were excluded. This left a final gene set of size 18,638 for analysis.

Preprocess of MutSigCVsyn inputs

MAF file preparation

Mutations of PCAWG patients were annotated via customized script (available on GitHub) into 7 mutation categories based upon the mutational context as in MutSigCV [4]. The categories are:

1. Transition mutations at CpG dinucleotides
2. Transversion mutations at CpG dinucleotides
3. Transition mutations at C: G base pairs not in CpG dinucleotides
4. Transversion mutations at C: G base pairs not in CpG dinucleotides
5. Transition mutations at A: T base pairs
6. Transversion mutations at A:T base pairs
7. Null and Indel mutations

Coverage file preparation

The coverage for every single patient at every genomic position in the geneset was calculated based on the wig file to ensure accurate coverage, instead of a simple full coverage model. The calculation process was re-engineered as in the original MutSigCV.

One covered genomic position was counted as 1. It was equally divided into 3 parts because the nucleotide has $\frac{1}{3}$ chance to mutate to any of the rest nucleotides (i.e. A could be mutated to C/G/T). Each possible mutation has its consequence, which consists of 3 mutation zones:

1. Synonymous
2. Nonsynonymous
3. Non-coding (Defined as intronic and untranslated regions)

and mutation categories 1 to 6 are defined above. The coverage for category 7, the null and indel mutation, was the coverage of the entire gene, which was the sum across categories 1 to 6. These consequences constituted 21 bins in total for each gene. For every position in a gene, the $\frac{1}{3}$ mutation counts were assigned to the corresponding bins and the summed counts were the category-specific coverage for the gene. Full coverage was assumed for unreported positions in the wig file.

Covariate file re-annotation

The covariate file provided for MutSigCV [4] was adopted. However, to avoid the inconsistency of gene naming between the BROAD Institute and PCAWG, the gene names in the covariate file were re-annotated in MutSigCVsyn. All synonyms of the PCAWG gene names were identified using the R package *BiomaRt*. 862 synonym names were mapped to the BROAD original covariate file and replaced by the new name to generate a new gene covariate file, while the expression, replication timing, and chromatin status data remained the same.

Gene dictionary file

MutSigCVsyn only takes mutations in intron and UTR (Untranslated region) into account to avoid transcription-associated mutation bias. Therefore, mutations in the regions that are not transcribed, such as intergenic, promoter, and up-/downstream regions, were excluded by removing the variant classification in the gene dictionary file and weren't recognized in MutSigCVsyn.

MutSigCVsyn workflow

MutSigCVsyn is adopted from MutSigCV [4]. Several key changes were made to identify the synonymous mutations. A more detailed and technical overview of changes can be found in the GitHub repository (https://github.com/ryy1221/MutSigCVsyn)The workflow of MutSigCVsyn is as follows:

The number of synonymous, non-coding, and non-synonymous mutations for each gene (g), patient (p) and mutation category (c) were defined as $n_{g,c,p}^{synonymous}$, $n_{g,c,p}^{noncoding}$ and $n_{g,c,p}^{nonsynonymous}$. Similarly, the coverage was defined as $N_{g,c,p}^{synonymous}$, $N_{g,c,p}^{noncoding}$, $N_{g,c,p}^{nonsynonymous}$. The total count of mutation/coverage across all categories was defined as $c+1$ as in MutSigCV, whereas for mutations, it meant the sum of all mutations, but in coverage, it meant the sum across categories 1 to 6.

To account for the gene-specific covariates in BMR (background mutation rate), MutSigCVsyn finds the nearest neighbor genes, which share the closest mutational

property based on the covariates (expression level, DNA replication timing, and chromatin compartment), for each target gene.

First, as in MutSigCV, the pairwise Euclidean Distance between every gene pair was calculated according to the gene covariate information. For each gene g, the raw background mutation number and coverage were defined as the non-coding mutation number and coverage(Eq. 1.1) across all patients (p)and mutation categories (c).

$$\begin{aligned} n_g^{bkgd} &= \sum_{p=1}^{n_p} n_{g,c+1,p}^{noncoding} \\ N_g^{bkgd} &= \sum_{p=1}^{n_p} N_{g,c+1,p}^{noncoding} \end{aligned} \tag{1.1}$$

Then, MutSigCVsyn evaluates the non-coding mutation and coverage similarity between pairs of the closest neighbor genes (i) and the target genes (g) using beta-binomial distribution as in MutSigCV. All qualified neighbor genes ($i = 0, 1, 2, \dots$) composed a 'Bagel' for the target gene ($\forall i \in B_g$). The gene's background mutations and coverage were calculated by summing the mutation count and coverage (Eq. 1.2) across the gene itself and the other qualified genes in its 'Bagel'.

$$\begin{aligned} x_g &= n_g^{bkgd} + \sum_{i \in B_g} n_i^{bkgd} \\ X_g &= N_g^{bkgd} + \sum_{i \in B_g} N_i^{bkgd} \end{aligned} \tag{1.2}$$

Then, MutSigCVsyn incorporated the marginal relative rate of patient-specific and mutation-category-specific mutation rate calculated within each histology cohort. The category and patient-specific mutation rate were calculated based upon all mutations (synonymous, non-synonymous, and non-coding) to obtain an accurate estimation of mutational load for each gene. They were then combined with the background mutation count and coverage for the gene of interest to obtain the gene, patient, mutation category level background mutation rate ($x_{g,c,p}$) and coverage ($X_{g,c,p}$).

After that, for the gene of interest, the probability of observing 0, 1, or more synonymous mutations in each mutation context and patient was calculated (Eq. 1.3). Here, the $N_{g,c,p}^{synonymous}$ indicates that only the possible mutations that happen in the synonymous positions were considered.

$$\begin{aligned} P_{g,c,p}^{(0)} &= H\left(0, N_{g,c,p}^{synonymous}, x_{g,c,p}, X_{g,c,p}\right) \\ P_{g,c,p}^{(1)} &= H\left(1, N_{g,c,p}^{synonymous}, x_{g,c,p}, X_{g,c,p}\right) \\ P_{g,c,p}^{(2+)} &= 1 - P_{g,c,p}^{(0)} - P_{g,c,p}^{(1)} \end{aligned} \tag{1.3}$$

The mutational categories were rank ordered from high to low based on the probability of having 0, 1, or more mutations in that category. The probabilities were combined and projected for each 2D combination of the mutation category of the 0, 1st, and 2nd mutations and then log-transformed into the scores as in MutSigCV. In addition, the 'null score boost', an additional score for deletion and insertion mutations,

was set to 0 as synonymous mutations do not fall into this category. A background null distribution was then built by convoluting the mutation probabilities across all 2D projected categories. Finally, the observed score was obtained by summing the scores across observed 2D projected categories of each patient. The p-value for the gene was obtained as the probability of observing a score at least as extreme as the observed score in the null distribution.

The last step was FDR calculation for multiple hypothesis testing. During the identification of genes of which synonymous mutations are significantly mutated, we were identifying signals of substitutions that are commonly known as 'passenger' mutations. Therefore, the false discovery rate control would be much more difficult as most of the genes will accept the null hypothesis, leaving a much smaller number of potentially interesting genes for more intensive investigation. Thus, instead of the original Benjamini–Hochberg FDR method, a nonparametric, empirical Bayes FDR method was employed.

Significant synonymous candidate discovery by Bayesian FDR

The Bayesian false discovery rate as described in Efron et al. [77] was adopted. Two classes of genes were defined: genes of which the synonymous mutations are significantly mutated, and genes of which the synonymous mutations are not significantly observed. The p-values for each gene are $S_0, S_1, S_2, \ldots, S_N$ to avoid confusion with the probability p.

Let the prior probabilities and the hypotheses be:

$$\begin{aligned} &p_0 = Prob\{Not\ significantly\ mutated\} \\ &\text{H0}: \text{ The gene is not a significantly mutated gene} \\ &p_1 = Prob\{Significantly\ mutated\} \\ &\text{H1}: \text{ The gene is a significantly mutated gene} \end{aligned} \tag{2.1}$$

The prior probability has corresponding density $f_0(s)$ and $f_1(s)$ for the S_i of the gene. Therefore, the mixture density of the 2 populations is.

$$f(s) = p_0 f_0(s) + p_1 f_1(s) \tag{2.2}$$

Define $F_0(s)$ and $F(s)$ be the cumulative distribution functions corresponding to $f_0(s)$ and $f(s)$ in (Eq. 2.2). According to the definition of Bayesian FDR, The FDR value for $\{S \leq s\}$ is defined as:

$$Fdr(s) \equiv \frac{p_0 F_0(s)}{F(s)}\ For\ S \leq s \tag{2.3}$$

which is the probability of identifying genes coming from the null hypothesis, given p-values equal or less than s.

In MutSigCVsyn FDR calculation, a nonparametric estimate for $Fdr(s_i)$ was calculated using the empirical CDF of S:

$$\widehat{Fdr}(\mathrm{s_i}) = \frac{\mathrm{p_0}\widehat{F_0}(\mathrm{s_i})}{\widehat{F}(\mathrm{s_i})}\ For\ S \leq s_i \tag{2.4}$$

where the *Fdr* value was calculated for every gene i with p-value < 0.05.

Note, (1) Both $\overline{F_0}$ and $\overline{F}$ [78]were estimations. To estimate the null distribution, non-expressed genes (FPKM < 1) across all tumor types were used as they are usually regarded to have no role in cancer. 1048 genes in total were used to build the empirical null distribution. The kCDF function in R package *sROC* [79] was used for estimating the cumulative distributions. The package gives asymptotically unbiased and consistent estimates for $F(s)$ and $F_0(s)$ given a large number of genes [80]. (2) The conservative assumption that $p_0 = 0.99$ was adopted because significant candidate genes are expected to occur at a very low chance. (3) As a final step of determining significant candidates, the candidate genes (*i.e.*, protocadherin gene families) of which coding and intronic regions are highly clustered in the same genomic regions were excluded [81] to avoid ambiguity of mutation annotation in overlapped gene regions.

MutSigCVsyn non-synonymous result analysis

For the drivers in PCAWG, only drivers identified in protein-coding regions were collected ('element_type' is 'cds'). We collected in total 150 PCAWG coding drivers, including drivers discovered previously and drivers discovered exclusively by PCAWG. The 15 PCAWG exclusive drivers were identified by the 'discovery_unique' flag in PCAWG.

Synonymous mutational heterogeneity analysis

The synonymous mutation rate was defined as the rate of synonymous substitutions per 1Mbp synonymous site. The synonymous sites were defined as genome positions where synonymous mutations were likely to occur. For every nucleotide in protein-coding gene sequences, there is $\frac{1}{3}$ chance for it to mutate into each of the rest nucleotides. Each nucleotide change that caused a synonymous mutation was counted as $\frac{1}{3}$ bp. For each patient, the number of total synonymous mutations across all synonymous positions were calculated and the synonymous mutation rate was then calculated as

$$Synonymous\ mutation\ rate = \frac{Number\ of\ total\ synonymous\ mutations}{Mega\ base\ pairs\ of\ synonymous\ positions\ sequenced}$$

Patients who have 0 synonymous mutations were set to have 0.01 synonymous mutations per mega base pair. The number of synonymous mutations that fell into the mutation category 1–6 was collected and scaled into fractions by the total number of synonymous mutations for each patient.

To show local mutation rate variation, chromosome 8 and chromosome 18 were selected and the mutation rate of 3 histology cohorts (Ovary-AdenoCA, Lung-SCC, Thy-AdenoCA) across the entire chromosome were examined. Mutation number in a 1Mbp window sliding over each base pair was collected and averaged across the patient number in that cohort.

BCL-2 mutation enrichment analysis

The BCL-2 synonymous mutations were extracted from PCAWG Lymph-BNHL maf files. In the permutation analysis, each mutation was randomly assigned to a BCL-2

coding position in one permutation and the number of mutations that fall into the BH4 motif was recorded. After 10,000 permutations, the observed BH4 mutation number and the permuted distribution were compared. The p-value is calculated as

$$p\text{-}value = \frac{Number\ of\ permutations\ (mutation\ number > observed\ mutation\ number)}{Number\ of\ total\ permutation}$$

Gene mRNA expression analysis and CERES score analysis

Gene mRNA expression data in patient tumor sample and normal sample(if exists) were collected. For DepMap cell line expression analysis, the cell line lineage that matches the corresponding histology cohort was first retrieved. The expression of the gene in the cell line lineage was then extracted and compared to all other cell lines. The Mann–Whitney U test was then performed to determine the significance of the difference in gene mRNA expression.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05521-8.

Additional file 1. Figure S1. MutSigCVsyn non-synonymous cancer driver landscape: Heatmap displaying 133 significant nonsynonymous candidate genes (Benjamini-Hochberg FDR < 1× 10-2) identified by MutSigCVsyn. Candidate genes are divided into two columns and are ranked from most frequent across all histology cohorts (left top) to the least frequent ones (right bottom). Candidate genes are colored by negative logarithmic transformed FDR value from high (dark blue) to low (light yellow). **Figure S2**. Potential functional role of MutSigCVsyn candidate genes: (A) Boxplot of Breast-AdenoCA patient PURA mRNA expression level of normal samples and tumor samples. The P-value is calculated by the Mann-Whitney U test. (B) Boxplot of SIGLEC15 expression data from DepMap Pancreas exocrine cell lines and all other tested cell lines. The P-values are calculated by the Mann-Whitney U test. (C) CADD analysis of synonymous mutations in synonymous candidate genes against the ones in all other genes. The P-values are calculated by the Mann-Whitney U test. **Figure S3**. MutSigCVsyn identifies PCAWG-exclusive drivers in non-synonymous analysis: PCAWG-exclusive drivers are the cancer driver genes that were first identified by the PCAWG working group. There are in total 15 exclusive non-synonymous protein-coding drivers in PCAWG and they are shown in the table. The 'gene' column shows the gene name. 'cds' in the 'Element_type' column shows that the coding region of the gene is identified as a cancer driver. 'discovery_unique' in the 'category' column shows that the gene is first identified by PCAWG. 6 of them (highlighted yellow) were identified by MutSigCVsyn in non-synonymous mutation analysis.

Additional file 2. Detailed methods and results of mutual exclusivity and co-occurrence analysis, CADD analysis, and gProfiler analysis.

Additional file 3. Table S1 Synonymous candidates v.s. Non-synonymous Drivers(PCAWG) mutual exclusivity analysis: Columns are cohort tested, synonymous candidate gene, PCAWG driver gene, number of patients have mutations in both genes, number of patients only have mutations in synonymous candidate gene, number of patients only have mutations in PCAWG driver gene, number of patients don't have mutations in either of the genes, Fisher's exact test statistics(odds ratio), Fisher's exact test p-value(left-tail), Benjamini-Hochberg corrected q value. **Table S2**: Non-synonymous Drivers(PCAWG) vs. Non-synonymous Drivers(PCAWG) mutual exclusivity analysis: Columns are cohort tested, first PCAWG driver gene, second PCAWG driver gene, number of patients have mutations in both genes, number of patients only have mutations in first PCAWG driver gene, number of patients only have mutations in second PCAWG driver gene, number of patients don't have mutations in either of the genes, Fisher's exact test statistics(odds ratio), Fisher's exact test p-value(left-tail), Benjamini-Hochberg corrected q value. **Table S3**: Synonymous candidates vs. Non-synonymous Drivers(PCAWG) co-occurrence analysis: Columns are cohort tested, synonymous candidate gene, PCAWG driver gene, number of patients have mutations in both genes, number of patients only have mutations in a synonymous candidate gene, number of patients only have mutations in PCAWG driver gene, number of patients don't have mutations in either of the genes, Fisher's exact test statistics (odds ratio), Fisher's exact test p-value(right-tail), Benjamini-Hochberg corrected q value. **Table S4**: Cander drivers identified in Tokheim et al. using different algorithms: Columns A-I are driver genes detected by Tokheim et al. with different algorithms. Column J is candidate genes identified by MutSigCVsyn. For each column, genes that are overlapped with MutSigCVsyn candidates are highlighted in red. **Table S5**: BCL2 synonymous mutations in PCAWG & overlap with synMICdb. Table S6: gProfiler analysis on MutSigCVsyn candidate genes: output of gProfiler analysis result using all synonymous candidate genes. Table S7: Synonymous candidate genes and FDR q values identified by MutSigCVsyn in PCAWG: Detailed q-value data for figure 3C heatmap.

Acknowledgements
Not applicable.

Author contributions
YR analyzed and interpreted the data. NA assisted early-stage analysis. YR, JRP and EPO wrote the manuscript. All authors read and approved the manuscript.

Funding
JRP acknowledges funding from NIBIB (5R21EB026617). EPO acknowledges funding from NIH (R35-GM124818) and NSF (ABI-1759860).

Availability of data and materials
The data, Pan-Cancer Analysis of Whole Genomes, that support the findings of this study are available from ICGC Data Portal (https://dcc.icgc.org/releases/PCAWG/consensus_snv_indel) but restrictions apply to the availability of TCGA portion of the PCAWG data, which were used under license for the current study, and so are not publicly available. Data are however available from the corresponding author upon reasonable request and with permission of TCGA.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
YR has no competing interest. NA has no competing interest. EPO has no competing interest. JRP has ownership interests in Theseus Pharmaceutical and MOMA therapeutics. JRP is a co-founder of Theseus Pharmaceuticals. JRP is a paid consultant for Takeda Pharmaceuticals, Theseus Pharmaceuticals, MOMA therapeutics, and third rock ventures. JRP receives research funding from Theseus Pharmaceuticals.

Received: 31 January 2023 Accepted: 5 October 2023
Published online: 07 December 2023

References

1. Nowell PC. The clonal evolution of tumor cell populations. Science. 1979;194(4260):23 LP – 28.
2. Elliott K, Larsson E. Non-coding driver mutations in human cancer. Nat Rev Cancer. 2021;21(8):500–9.
3. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371-385.e18.
4. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8. https://doi.org/10.1038/nature12213.
5. Shuai S, Abascal F, Amin SB, Bader GD, Bandopadhayay P, Barenboim J, et al. Combined burden and functional impact tests for cancer driver discovery using DriverPower. Nat Commun. 2020;11(1):734.
6. Lochovsky L, Zhang J, Fu Y, Khurana E, Gerstein M. LARVA: An integrative framework for large-scale analysis of recurrent variants in noncoding annotations. Nucleic Acids Res. 2015;43(17):8123–34.
7. Vatansever S, Erman B, Gümüş ZH. Oncogenic G12D mutation alters local conformations and dynamics of K-Ras. Sci Rep. 2019;9(1):11730.
8. Armenia J, Wankowicz SAM, Liu D, Gao J, Kundra R, Reznik E, et al. The long tail of oncogenic drivers in prostate cancer. Nat Genet. 2018;50(5):645–51.
9. Elman JS, Ni TK, Mengwasser KE, Jin D, Wronski A, Elledge SJ, et al. Identification of FUBP1as a long tail cancer driver and widespread regulator of tumor suppressor and oncogene alternative splicing. Cell Rep. 2019;28(13):3435-3449.e5. https://doi.org/10.1016/j.celrep.2019.08.060.
10. Kumar S, Warrell J, Li S, McGillivray PD, Meyerson W, Salichos L, et al. Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences. Cell. 2020;180(5):915-927.e16.
11. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505(7484):495–501.
12. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502(7471):333–9.
13. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature. 2020;578(7793):102–11.
14. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977;267(5608):275–6.

15. Tomoko O. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J Mol Evol. 1995;40(1):56–63.
16. Sharma AK, O'Brien EP. Non-equilibrium coupling of protein structure and function to translation–elongation kinetics. Curr Opin Struct Biol. 2018;49:94–103.
17. Liu Y. A code within the genetic code: codon usage regulates co-translational protein folding. Cell Commun Signal. 2020;18(1):145.
18. Walsh IM, Bowman MA, Soto Santarriaga IF, Rodriguez A, Clark PL. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. Proc Natl Acad Sci. 2020;117(7):3528.
19. Komar AA. A pause for thought along the co-translational folding pathway. Trends Biochem Sci. 2009;34(1):16–24.
20. Akashi H. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics. 1994;136(3):927–35.
21. Zhang F, Saha S, Shabalina SA, Kashina A. Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. Science. 2010;329(5998):1534–7.
22. Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res. 2013;41(4):2073–94.
23. Mueller WF, Larsen LSZ, Garibaldi A, Hatfield GW, Hertel KJ. The silent sway of splicing by synonymous substitutions. J Biol Chem. 2015;290(46):27700–11.
24. Lebeuf-Taylor E, McCloskey N, Bailey SF, Hinz A, Kassen R. The distribution of fitness effects among synonymous mutations in a gene under directional selection. Landry CR, Wittkopp PJ, Venkataram S, editors. Elife. 2019;8:e45952.
25. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet. 2011;12(10):683–91.
26. Vedula P, Kurosaka S, MacTaggart B, Ni Q, Papoian G, Jiang Y, et al. Different translation dynamics of β- and γ-actin regulates cell migration. Elife. 2021;24(10): e68712.
27. Zhou M, Guo J, Cha J, Chae M, Chen S, Barral JM, et al. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature. 2013;495:111. https://doi.org/10.1038/nature11833.
28. Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell. 2014;156(6):1324–35.
29. Chu D, Wei L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. BMC Cancer. 2019;19(1):359.
30. Benisty H, Weber M, Hernandez-Alias X, Schaefer MH, Serrano L. Mutation bias within oncogene families is related to proliferation-specific codon usage. Proc Natl Acad Sci. 2020;117(48):30848 LP – 30856.
31. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, et al. A silent polymorphism in the MDR1 gene changes substrate specificity. Science. 2007;315(5811):525.
32. Niersch J, Vega-Rubín-de-Celis S, Bazarna A, Mergener S, Jendrossek V, Siveke JT, et al. A BAP1 synonymous mutation results in exon skipping, loss of function and worse patient prognosis. iScience. 2021;24(3):102173.
33. Sharma Y, Miladi M, Dukare S, Boulay K, Caudron-Herger M, Groß M, et al. A pan-cancer analysis of synonymous mutations. Nat Commun. 2019;10(1):2569. https://doi.org/10.1038/s41467-019-10489-2.
34. Bin Y, Wang X, Zhao L, Wen P, Xia J. An analysis of mutational signatures of synonymous mutations across 15 cancer types. BMC Med Genet. 2019;20(2):190. https://doi.org/10.1186/s12881-019-0926-4.
35. Teng H, Wei W, Li Q, Xue M, Shi X, Li X, et al. Prevalence and architecture of posttranscriptionally impaired synonymous mutations in 8320 genomes across 22 cancer types. Nucleic Acids Res. 2020;48(3):1192–205. https://doi.org/10.1093/nar/gkaa019.
36. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47(D1):D941–7. https://doi.org/10.1093/nar/gky1015.
37. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012;22(8):1589–98.
38. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93.
39. Sammalkorpi H, Alhopuro P, Lehtonen R, Tuimala J, Mecklin JP, Järvinen HJ, et al. Background mutation frequency in microsatellite-unstable colorectal cancer. Cancer Res. 2007;67(12):5691–8.
40. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94–101.
41. Glaser AP, Fantini D, Wang Y, Yu Y, Rimar KJ, Podojil JR, et al. APOBEC-mediated mutagenesis in urothelial carcinoma is associated with improved survival, mutations in DNA damage response genes, and immune response. Oncotarget. 2017;9(4):4537–48.
42. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415–21.
43. Wolfe KH, Sharp PM, Li WH. Mutation rates differ among regions of the mammalian genome. Nature. 1989;337(6204):283–5.
44. Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. Nat Rev Dis Primers. 2020;6(1):92.
45. Hankey W, Frankel WL, Groden J. Functions of the APC tumor suppressor protein dependent and independent of canonical WNT signaling: implications for therapeutic targeting. Cancer Metastasis Rev. 2018;37(1):159–72.
46. Zhou S, Buckhaults P, Zawel L, Bunz F, Riggins G, le Dai J, et al. Targeted deletion of Smad4 shows it is required for transforming growth factor β and activin signaling in colorectal cancer cells. Proc Natl Acad Sci. 1998;95(5):2412 LP – 2416.
47. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.

48. Küppers R. Mechanisms of B-cell lymphoma pathogenesis. Nat Rev Cancer. 2005;5(4):251–62. https://doi.org/10.1038/nrc1589.
49. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. 2011;478(7367):64–9.
50. Liang Y, Tebaldi T, Rejeski K, Joshi P, Stefani G, Taylor A, et al. SRSF2 mutations drive oncogenesis by activating a global program of aberrant alternative splicing in hematopoietic cells. Leukemia. 2018;32(12):2659–71.
51. Bentele K, Saffert P, Rauscher R, Ignatova Z, Blüthgen N. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol. 2013;18(9):675.
52. Daniel DC, Johnson EM. PURA, the gene encoding Pur-alpha, member of an ancient nucleic acid-binding protein family with mammalian neurological functions. Gene. 2018;643:133–43.
53. Li QT, Huang ZZ, Chen YB, Yao HY, Ke ZH, He XX, et al. Integrative analysis of siglec-15 mRNA in human cancers based on data mining. J Cancer. 2020;11(9):2453–64.
54. Xu JL, Guo Y. A comprehensive analysis of different gene classes in pancreatic cancer: SIGLEC15 may be a promising immunotherapeutic target. Invest New Drugs. 2021;40(1):58–67.
55. Wang J, Sun J, Liu LN, Flies DB, Nie X, Toki M, et al. Siglec-15 as an immune suppressor and potential target for normalization cancer immunotherapy. Nat Med. 2019;25(4):656–66.
56. Akl H, Vervloessem T, Kiviluoto S, Bittremieux M, Parys JB, De Smedt H, et al. A dual role for the anti-apoptotic Bcl-2 protein in cancer: Mitochondria versus endoplasmic reticulum. Biochim Biophys Acta (BBA) Mol Cell Res. 2014;1843(10):2240–52.
57. Bakhshi A, Jensen JP, Goldman P, Wright JJ, McBride OW, Epstein AL, et al. Cloning the chromosomal breakpoint of t(14;18) human lymphomas: clustering around JH on chromosome 14 and near a transcriptional unit on 18. Cell. 1985;41(3):899–906.
58. Huang DC, Adams JM, Cory S. The conserved N-terminal BH4 domain of Bcl-2 homologues is essential for inhibition of apoptosis and interaction with CED-4. EMBO J. 1998;17(4):1029–39.
59. Trisciuoglio D, De Luca T, Desideri M, Passeri D, Gabellini C, Scarpino S, et al. Removal of the BH4 domain from Bcl-2 protein triggers an autophagic process that impairs tumor growth. Neoplasia. 2013;15(3):315–27.
60. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci U S A. 2016;113(50):14330–5. https://doi.org/10.1073/pnas.1616440113.
61. Kridel R, Sehn LH, Gascoyne RD. Pathogenesis of follicular lymphoma. J Clin Invest. 2012;122(10):3424–31. https://doi.org/10.1172/JCI63186.
62. Tsujimoto Y, Cossman J, Jaffe E, Croce CM. Involvement of the bcl-2 gene in human follicular lymphoma. Science (1979). 1985;228(4706):1440.
63. Singh K, Briggs JM. Functional Implications of the spectrum of BCL2 mutations in Lymphoma. Mutat Res Rev Mut Res. 2016;769:1–18.
64. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994;11(5):725–36.
65. Yang Z, Ro S, Rannala B. Likelihood models of somatic mutation and codon substitution in cancer genes. Genetics. 2003;165(2):695–705.
66. Williams MJ, Zapata L, Werner B, Barnes CP, Sottoriva A, Graham TA. Measuring the distribution of fitness effects in somatic evolution by combining clonal dynamics with dN/dS ratios. Elife. 2020;30(9): e48714.
67. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal patterns of selection in cancer and somatic tissues. Cell. 2017;171(5):1029-1041.e21.
68. Zapata L, Pich O, Serrano L, Kondrashov FA, Ossowski S, Schaefer MH. Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. Genome Biol. 2018;19(1):67.
69. Lohr JG, Stojanov P, Lawrence MS, Auclair D, Chapuy B, Sougnez C, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc Natl Acad Sci. 2012;109(10):3879.
70. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5. https://doi.org/10.1038/ng.2892.
71. Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. 2023;5:gkad347. https://doi.org/10.1093/nar/gkad347.
72. Zeng Z, Bromberg Y. Inferring potential cancer driving synonymous variants. Genes (Basel). 2022;13(5):778.
73. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med. 2013;5(10):91. https://doi.org/10.1186/gm495.
74. Hare MP, Palumbi SR. High intron sequence conservation across three mammalian orders suggests functional constraints. Mol Biol Evol. 2003;20(6):969–78.
75. Maurano MT, Richard H, Eric R, Thurman RE, Eric H, Hao W, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190–5.
76. Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink JJ, Lopez G, et al. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. 2013;41(Database issue):110–7.
77. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol. 2002;23(1):70–86.
78. Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc. 2004;99(465):96–104.
79. Wang XF. sROC: Nonparametric Smooth ROC Curves for Continuous Data. 2012.
80. Nadaraya EA. Some new estimates for distribution functions. Theory Probab Appl. 1964;9(3):497–500.
81. Drees F, Nelson WJ. Cadherin-Mediated Cell–Cell Adhesion. In: Lennarz WJ, Lane MDBTE of BC, editors. New York: Elsevier; 2004. p. 205–11.

82. Shimada K, Bachman JA, Muhlich JL, Mitchison TJ. shinyDepMap, a tool to identify targetable cancer genes and their functional connections from Cancer Dependency Map data. Elife [Internet]. 2021;10. Available from: https://doi.org/10.7554/eLife.57116

Publisher's Note