Published online 12 November 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D1289–D1301***
doi: 10.1093/nar/gkaa1033

OncoVar: an integrated database and analysis platform for oncogenic driver variants in cancers

Tao Wang[1,2,†], Shasha Ruan[3,†], Xiaolu Zhao[4], Xiaohui Shi[2], Huajing Teng[2], Jianing Zhong[5], Mingcong You[6], Kun Xia[1,7,8,*], Zhongsheng Sun[2,9,10,*] and Fengbiao Mao[11,*]

[1]Center for Medical Genetics & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan 410083, China, [2]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China, [3]Department of Clinical Oncology, Renmin Hospital of Wuhan University, Wuhan, Hubei 430072, China, [4]Center for Reproductive Medicine, Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing 100191, China, [5]Key Laboratory of Prevention and Treatment of Cardiovascular and Cerebrovascular Diseases of Ministry of Education, Gannan Medical University, Ganzhou 341000, China, [6]Baiyining Medicine, Beijing 102200, China, [7]CAS Center for Excellence in Brain Science and Intelligences Technology (CEBSIT), Shanghai 200031, China, [8]School of Basic Medical Science, Central South University, Changsha, Hunan 410078, China, [9]CAS Center for Excellence in Biotic Interactions, University of Chinese Academy of Sciences, Beijing 100049, China, [10]State Key Laboratory of Integrated Management of Pest Insects and Rodents, Chinese Academy of Sciences, Beijing 100101, China and [11]Center of Basic Medical Research, Institute of Medical Innovation and Research, Peking University Third Hospital, Beijing 100191, China

Received August 28, 2020; Revised October 15, 2020; Editorial Decision October 16, 2020; Accepted October 19, 2020

ABSTRACT

The prevalence of neutral mutations in cancer cell population impedes the distinguishing of cancer-causing driver mutations from passenger mutations. To systematically prioritize the oncogenic ability of somatic mutations and cancer genes, we constructed a useful platform, OncoVar (https://oncovar.org/), which employed published bioinformatics algorithms and incorporated known driver events to identify driver mutations and driver genes. We identified 20 162 cancer driver mutations, 814 driver genes and 2360 pathogenic pathways with high-confidence by reanalyzing 10 769 exomes from 33 cancer types in The Cancer Genome Atlas (TCGA) and 1942 genomes from 18 cancer types in International Cancer Genome Consortium (ICGC). OncoVar provides four points of view, 'Mutation', 'Gene', 'Pathway' and 'Cancer', to help researchers to visualize the relationships between cancers and driver variants. Importantly, identification of actionable driver alterations provides promising druggable targets and repurposing opportunities of combinational therapies. OncoVar provides a user-friendly interface for browsing, searching and downloading somatic driver mutations, driver genes and pathogenic pathways in various cancer types. This platform will facilitate the identification of cancer drivers across individual cancer cohorts and helps to rank mutations or genes for better decision-making among clinical oncologists, cancer researchers and the broad scientific community interested in cancer precision medicine.

INTRODUCTION

The exome constitutes <2% of the human genome but contains ∼85% of known disease-causing variants (1). Early somatic mutations in coding regions can cause developmental disorders (2), whereas progressive accumulation of somatic mutations throughout life can lead to cancer (3). Understanding genetic events that lead to cancer initiation and progression remains one of the biggest challenges in cancer biology. Large cancer sequencing projects, such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC), provide unprecedented opportunities to identify causative variants underlying human cancers (4). However, the majority of the somatic missense mutations do not have a noticeable effect (5), and the prevalence of neutral mutations in a cancer cell population impede the distinguishing of cancer-causing driver mutations (6). Dozens of computational algorithms have been developed to predict whether a missense mutation is deleteri-

*To whom correspondence should be addressed. Tel: +86 0731 84805357; Email: xiakun@sklmg.edu.cn
Correspondence may also be addressed to Zhongsheng Sun. Tel: +86 10 64864959; Email: sunzs@biols.ac.cn
Correspondence may also be addressed to Fengbiao Mao. Tel: +86 10 82266115; Email: maofengbiao08@163.com
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

ous or pathogenic based on concepts including evolutionary conservation, structural constraints and the physicochemical attributes of amino acids (7). But cancer driver mutations predicted by these computational methods are lack of consistencies and are prone to false positives (8). Even though we recently have developed a machine learning method to specifically predict cancer-driving deleterious mutations with high accuracy (9), there is still no convenient database to access driver mutations for cancer therapeutic targets.

On the other hand, many approaches have been developed to prioritize cancer driver genes with the advance of next-generation sequencing technologies (10). Most of these tools can be classified into three categories based on three basic principles: (i) frequency-based methods, which consist of identifying genes that are more frequently mutated than the background mutation rate (BMR) (11–16); (ii) subnetwork methods, which identify groups of driver genes based on prior knowledge of networks, such as protein–protein interactions (17–23); (iii) hotspot-based methods (24–26), which are driven by positive selection and are particularly located in functional domains or important residues for 3D protein structures (27,28). However, driver genes predicted from these computational tools also lack consistency since many of these tools are not optimally balanced between precision and sensitivity (10,29). Some of them are overly conservative and missing many true drivers while the others are over-relaxed and yield too many false-positive calls (10). Therefore, discovering a complete catalog of driver genes truly associated with cancer is far from being achieved.

Recently, many databases have been developed to deposit cancer driver genes with the rapid development of prediction methods (30) and the expansion of experimental validations (30,31). For example, the Cancer Gene Census (CGC) is an ongoing project within COSMIC database to catalogue all genes that are causally implicated in cancer through somatic and germline mutations (32). OncoKB is a comprehensive and curated oncology knowledge database with oncologists detailed and evidence-based information about individual somatic mutations and structural alterations present in patient tumors (33). MutPanning provides a resource of driver genes across 28 tumor types with additional driver genes identified according to mutations in unusual nucleotide context (34). Sleeping Beauty Cancer Driver Database (SBCDDB) provides information of cancer driver genes identified in tumor models generated by Sleeping Beauty insertional mutagenesis (35). More and more databases, such as DriverDBv3 (36) and IntOGen (37), are incorporating published prediction approaches to identify driver genes from large-scale cancer projects. Using combinational strategies, consensus-based methods like CTAT (30), ConsensusDriver (38), IntOGen (37) and C3 (39) promise to harness the strengths of different driver prediction methods and provide the best trade-off between sensitivity and specificity. However, there is no available integrated database for convenient search of annotations of driver genes from TCGA and ICGC cancer genomics projects by incorporating cancer driver predictions and prior oncology knowledge.

To satisfy these demands, we developed a database, called ONCOVAR (Figure 1), to discover ONCOgenic driver VARiants from large cancer sequencing projects by employing published bioinformatics algorithms and incorporating known driver events. OncoVar provides four points of view, 'Mutation', 'Gene', 'Pathway' and 'Cancer', to help researchers to visualize the relationships between cancers and driver variants. Our database provides a valuable resource for cancer studies by presenting cancer-causing mutations, mutated driver genes and oncogenic signaling pathways. The findings based on our database highlight the importance of combination of algorithm predictions and prior knowledge for the interpretation of pathogenic variants in human cancers and complex diseases.

MATERIALS AND METHODS

Somatic mutation collection

OncoVar includes unbiased interpretation of 2 605 700 somatic mutations from the entire 10 769 tumor sample datasets of 33 cancer types by harmonizing the results of seven algorithms, yielded by the uniform analysis of all TCGA exome data by the Multi-Center Mutation-Calling in Multiple Cancers (MC3) network (40) (https://api.gdc.cancer.gov/data/1c8cfe5f-e52d-41ba-94da-f15ea1337efc). To reduce the false-positive rate for driver gene discovery, we implemented three strategies to optimize driver detection and data quality. Briefly, we excluded 344 hypermutated tumors because of artifact sensitivity to high background mutation rates. All mutations that passed the MC3 filter criteria were included. Finally, samples marked with inconsistent pathology were excluded. Clinical information on TCGA was downloaded from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). Moreover, we curated 23 159 591 somatic mutations from 1942 samples of 18 cancer types from the ICGC Data Portal (https://dcc.icgc.org/api/v1/download?fn=/PCAWG/consensus_snv_indel/final_consensus_passonly.snv_mnv_indel.icgc.public.maf.gz).

Germline mutations from normal population

Over 270 million variants in the gnomAD v2.1 dataset (https://gnomad.broadinstitute.org) have been widely used as a resource for allele frequency estimates in the context of rare disease, which can improve power for disease gene discovery and exploring the biological impact of genetic variation (41,42). We aggregated 14 967 411 mutations from 125 748 control-only reference individuals after removing mutations without 'PASS' in the filtering criterion. Gene-level pLI scores of all genes were downloaded from gnomAD database. Meanwhile, we collected gene-level values of loss-of-function observed/expected upper bound fraction (LOEUF) calculated by using 141 456 human genomes from gnomAD database.

Identification of extreme and driver mutations

Similar to our previous studies (2,4,43–48), we performed ANNOVAR (49) to annotate all somatic mutations with respect to variant-level data sources, including the following information: (i) functional effects of variants; (ii)

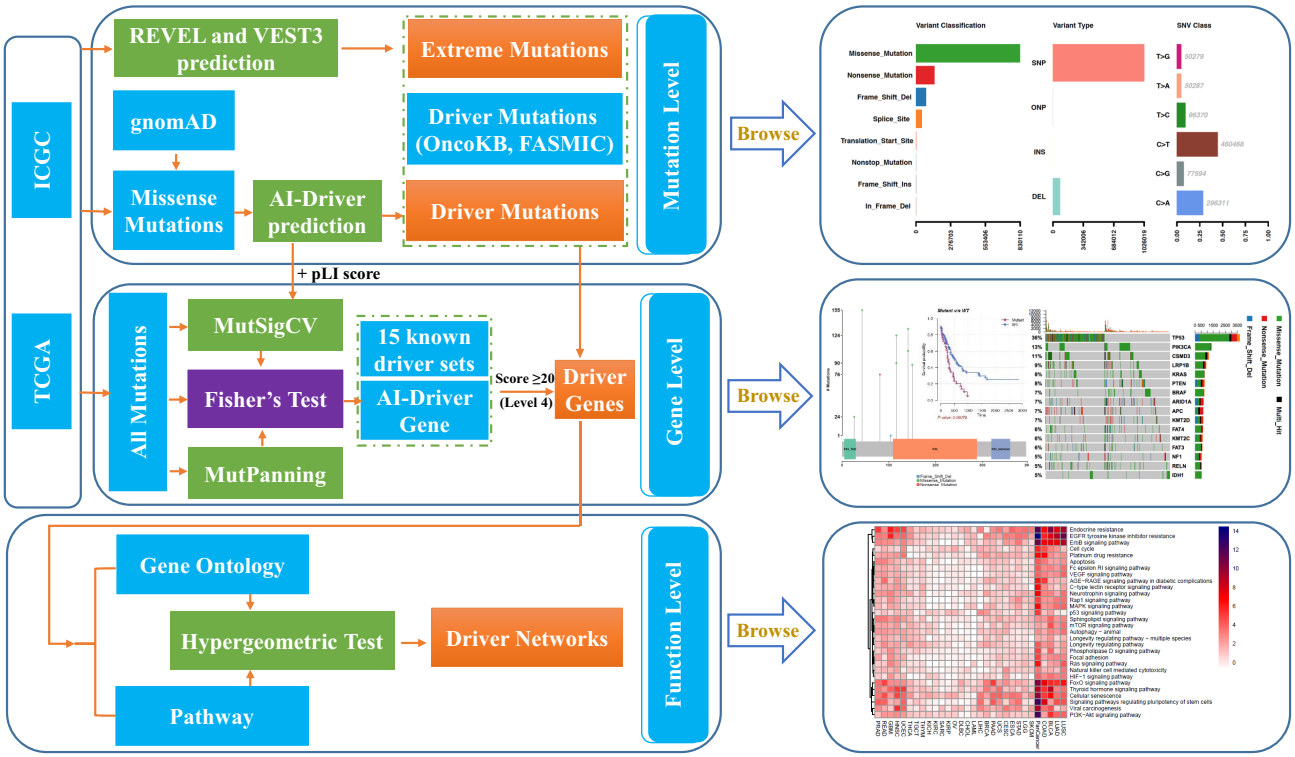

Figure 1. Workflow of OncoVar pipeline. The somatic missense mutations identified from TCGA and ICGC cohorts were used for driver mutation prediction with our recently developed AI-Driver method. Somatic missense mutations with AI-Driver score ≥0.95 and occurred in at least two patients were defined as driver mutations. The MutsigCV, with two additional gene-level covariates – pLI score and maximum AI-Driver score, and MutPanning methods were used for ranking the genes based on the generated P values with all somatic mutations identified from TCGA and ICGC cohorts, respectively. Then, two *P* values were combined by using Fisher method for each gene and further corrected by Benjamini–Hochberg method. Genes with a corrected *P* value ≤0.05 were considered as AI-DriverGenes. AI-DriverGenes combined with other 15 known driver sets were used for gene classification and the pathogenic genes with score ≥20 (Level 4) were considered as driver genes. Only the driver genes with driver or extreme mutations were used for gene-level exploration. Finally, driver genes were used to identify driver pathways (FDR < 0.05) by using the hypergeometric test.

functional prediction of missense mutations by 23 predictive algorithms; (iii) allele frequencies in different populations; (iv) reported variants in different disease- and phenotype-related databases and (v) some other genome features, such as CytoBand. LoF variants, including stop-gain, stop-loss, splicing site SNVs, frameshift indels and deleterious missense somatic mutations were regarded as potential extreme variants. Similar to our previous study (31), we obtained predictive scores and the pathogenicity consequences of missense variants from 23 in silico algorithms or tools, including SIFT (50), PolyPhen2-HDIV (51), PolyPhen2-HVAR (51), LRT (52), MutationTaster (53), MutationAssessor (54), FATHMM (55), PROVEAN (56), MetaSVM (57), MetaLR (57), VEST3 (58), M-CAP (59), CADD (60), GERP++ (61), DANN (62), fathmm-MKL (63), Eigen (64), GenoCanyon (65), fitCons (66), PhyloP (pP100way) (67), PhastCons (pC100way) (68), SiPhy (69) and REVEL (70) (Supplementary Table S1). To facilitate interpretation, we presented the final predictive scores of ReVe as percentiles that reflect the relative rank of pathogenicity for missense variants according to our recently published study (71), with the lowest score (i.e., 0.00) being the most benign variant and the highest score (i.e. 1.00) being the most deleterious variant. Extreme mutations were defined by loss of function mutations and missense mutations with a ReVe score ≥0.5. Then, we employed our recently developed machine-learning method, AI-Driver , to determine candidate driver mutations by testing missense mutations with the model trained by using the features of known driver mutations. The missense mutations with AI-Driver score ≥0.95 and occurred in at least two patients were defined as driver mutations in each cancer cohort.

Identification of AI-DriverGene

AI-DriverGene is prioritized by combining predictions of MutSigCV (72) and MutPanning (34) methods. First, MutSigCV was employed to generate the *P* value of each gene with two additional covariates including gene-level pLI score and maximum AI-Driver score of mutations within each gene. Then, the MutSigCV *P*-value was combined with the MutPanning *P*-value using the Fisher method. Finally, the joint *P* values were corrected by Benjamini-Hochberg and genes with a corrected *P* threshold of 0.05 were considered as AI-DriverGenes.

Five-tiered consensus-based classification of driver genes

We used a consensus-score for the classification of driver genes. Consensus-score is based on an improved Borda approach, where each gene was given a score equal to the sum across all driver sets of its weight (38). Totally, 16 driver sets (SupplementaryTable S2) were used for consensus-score

calculation and these driver sets were further classified into two groups including Group1 involving four gold standard driver sets (CGC, OncoKB, AI-DriverGene and MutPanning) and Group2 containing the remaining twelve driver sets (https://oncovar.org/welcome/links).

$$Score\ (gene_i) = \sum_{j=0,\ \ over\ 16\ driver\ sets}^{j=16} \text{weight}_j \begin{cases} 10, & j \in Group1 \\ 1, & j \in Group2 \end{cases} \quad (1)$$

Genes covered by neither of these 16 driver sets were assigned 0 and then all other genes were ranked according to this score. Furthermore, to make the gene ranking to a standardized and easily interpretable format, we introduced the following procedures to classify all genes into four grades from non-pathogenic to pathogenic): Level 0 (non-pathogenic, score = 0), Level 1 (possible pathogenic, score = 1), Level 2 (likely pathogenic, 1 < score ≤ 10), Level 3 (probable pathogenic, 10< score < 20) and Level 4 (pathogenic, score ≥ 20). Only the pathogenic genes in Level 4 were considered as driver genes and used for downstream analysis.

OncoVar scoring system based on Gaussian model

To systematically compare the 'driverness' of each mutation, gene and pathway, we calculated the OncoVar score using Gaussian model (2) based on $x = -\log_2(\text{FDR})$, in which σ represents the standard deviation of x while μ represents the expected value of x. An OncoVar score of x is the cumulative probability of $X \leq x$ following Z-score transformation. A higher OncoVar score (0–1) represents a greater ability of cancer initiation and progression by the mutation, gene or pathway.

$$f(\chi;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\chi} \exp\left(-\frac{(\chi-\mu)^2}{2\sigma^2}\right) d\chi \quad (2)$$

Essential and nonessential genes in human cancers

We curated 7168 essential and 21 373 nonessential genes from the OGEE v2 database (73). Then, we calculated the mean expression of these genes in 33 tumor types from TCGA cohorts. We filtered the essential genes with a mean TPM (Transcripts Per Kilobase Million) value < 10, while we filtered the nonessential genes with a mean TPM value >1 in 33 human cancer types. Finally, we obtained 6197 and 903 high-confidence essential and nonessential genes in human cancers, respectively (Supplementary Table S3).

Gene Ontology enrichment analyses

Gene enrichment was performed using the R package clusterProfiler version 3.2.14 (74). Each cluster from the driver genes was compared with the background of all other genes sequenced at sufficient depth in our study, with a Benjamini–Hochberg FDR threshold of 0.05 as significant enrichment. Enrichment of KEGG pathways was analyzed with the enrichKEGG function.

Profile and visualization of cancer driver events

We employed Maftools to summarize, analyze and annotate MAF files in an efficient manner from either TCGA or ICGC sources (75). In detail, we used 'plotmafSummary' to plot the summary of the MAF file, which displays the number of variants in each sample as a stacked barplot and variant types as a boxplot summarized by Variant_Classification. Better representation of the MAF file was shown as oncoplots. Side barplot and top barplots were controlled by drawRowBar and drawColBar arguments, respectively. The titv function classifies SNPs into transitions and transversions and returns a list of summarized tables. Summarized data were visualized as a boxplot showing overall distribution of six different conversions and as a stacked barplot showing the fraction of conversions in each sample. Lollipop plots are a simple and effective way to show mutation spots on protein structure. Many oncogenes have preferential sites which are mutated more often than any other locus. These spots are considered to be mutational hot-spots and lollipop plots are used to display them along with the rest of the mutations. We drew such plots using the function lollipopPlot in Maftools. In addition, we plotted word cloud plots for mutated genes with the function geneCloud. The size of each gene is proportional to the total number of samples in which it is mutated.

Mutually exclusive and co-occurring driver genes

Many disease-causing genes in cancer are co-occurring or show strong exclusiveness in their mutation pattern. Such mutually exclusive or co-occurring sets of genes can be detected using the somaticInteractions function in Maftools, which performs a pairwise Fisher's exact test to detect such a significant pair of genes. The somaticInteractions function also uses cometExactTest to identify potentially altered gene sets involving > two genes. The top 50 driver genes were used to perform exclusive/co-occurrence analysis and labeled with a P value threshold of 0.05 and 0.01.

Survival analysis

The function mafSurvive in Maftools was used to perform survival analysis and draw a Kaplan–Meier curve by grouping samples based on the mutation status of driver genes or manually provided samples that make up a group.

Drug–gene interactions

Numerous drugs that are already approved for specific diseases have known protein targets, which may be relevant for other disease types as well. A systematic identifying druggable genes in various diseases would help streamline the process of developing new drugs for these targets, even if no specific drugs are available for them yet (76). Therefore, we integrated the data for drug–gene interactions and gene druggability from the Drug–gene Interaction Database (DGIdb 3.0) (77) to assist with precision medicine in cancer treatment.

RESULTS

Website interface and search results

OncoVar provides four points of browse, 'Mutation', 'Gene', 'Pathway' and 'Cancer', to help researchers to vi-

sualize the relationships between cancers and driver variants for each cancer type and pan-cancer cohort. Firstly, users could search driver/extreme mutations by inputting variants, dbSNP ids or genomic regions based on human GRCh37/hg19 genome. Outputs of mutation search include OncoVar scores, six driver mutation annotations and 23 pathogenicity scores. Secondly, users could search driver genes by inputting gene symbols, Ensembl gene ids or Entrez ids. Outputs of gene search include consensus score, driver level, OncoVar scores, targeting drugs and 18 additional driver gene annotations coupled with lollipop plot of associated mutations and survival plot of clinical data. Thirdly, users could search oncogenic pathways by inputting GO and KEGG ids or GO and KEGG names. Outputs of pathway search include OncoVar scores and associated genes. Finally, OncoVar provides comprehensive summaries of driver/extreme mutations, driver genes and oncogenic pathways for each cancer type and pan-cancer cohort. The summaries of driver/extreme mutations include plot of mutation classification and plot of transition and transversions. The summaries of driver genes include waterfall plot of top 30 mutated genes, interaction plot and GeneCloud plot. The summaries of oncogenic pathways include KEGG dotplot, biological process dotplot, cellular component dotplot, molecular function dotplot and biological process GOgraph (Figure 1, Supplementary Figure S1). In order to improve the robustness of our database, we developed four convenient functions including 'Batch mutation annotation', 'Cancer specific pathway enrichment', 'Hg19ToHg38 conversion' and 'Hg38ToHg19 conversion' in drop down box of 'Analysis' column. Especially, the function of 'Cancer specific pathway enrichment' enables to perform GO and KEGG enrichment analysis based on the uploaded input genes by users. This enrichment analysis implements a hypergeometric test restricted to the identified driver genes in each GO/KEGG item for each cancer type as well as pan-cancer cohort. To our knowledge, our platform is the first webserver which could perform enrichment analysis of cancer-specific driver pathway in individual cancer types as well as pan-cancer cohort.

The features of cancer driver mutations

By employing our method AI-Driver, we identified 16 923 and 3409 driver missense mutations from TCGA and ICGC cohorts, respectively. We then classified the driver mutations into five groups according to consensus score based on four known driver sets and AI-Driver prediction. We found the proportions of driver mutations of CGC increased in groups with higher consensus score (Supplementary Figure S2A, B). To symmetrically evaluate the performance of predictions by AI-Driver, we employed receiver operating characteristic (ROC) analysis benchmarked by four known driver sets. We found driver predictions by AI-Driver had a superior and stable performance with high area under the ROC curve (AUC) value (Supplementary Figure S2C, D). The AUC values for TCGA cohorts were 0.972, 0.959, 0.969 and 0.966 while the AUC values for ICGC cohorts were 0.968, 0.958, 0.967 and 0.969, benchmarked by CGI (78), FASMIC (31), OncoKB (33) and PMID25348012 (79), respectively. All of these newly detected driver mutations by AI-Driver were integrated into our OncoVar database. More than 46% patients in TCGA cohorts had at least one driver mutations in driver gene TP53, PIK3CA, KRAS, BRAF, IDH1, PTEN, CTNNB1 and NRAS (Figure 2A, B) while more than 29% patients in ICGC cohorts suffered from at least one driver mutations in driver gene TP53, KRAS, CTNNB1 and PIK3CA (Figure 2C, D). Top 30 driver genes and all genes harbored driver mutations were showed in Figure 2 and supplemental table 4, respectively. Therefore, OncoVar provides access to search all newly detected and known driver mutations in each cancer type and pan-cancer cohort from TCGA and ICGC projects.

The landscape of cancer driver genes

By employing the OncoVar pipeline, we identified 713 and 686 driver genes from TCGA and ICGC pan-cancer cohorts, respectively. Specially, we identified 806 and 696 driver genes from TCGA and ICGC individual cancer types, respectively. The number of cancer driver genes varies among cancer types, with Uveal Melanoma (UVM having the fewest (6 genes) and Acute Myeloid Leukemia (LAML) having the most (134 genes) from TCGA cohorts (Supplementary Table S5). Our results revealed that the roles of driver genes across cancer types is much more widespread than previously documented (Figure 3). For example, the pattern of somatic mutations in ataxia telangiectasia mutated (ATM) shows signals of positive selection across 11 and 5 tumor types in TCGA and ICGC cohorts, respectively (Figure 3). However, it is annotated in the CGC only as a driver of T-cell-prolymphocytic leukemia. In addition, we observed a moderate positive correlation (Pearson's R = 0.35, P value = 0.04) between mean mutation burden in a cancer type and the number of identified driver genes (Figure 4A; Supplementary Table S6). Moreover, OncoVar identified 23 and 7 novel driver genes from TCGA and ICGC pan-cancer cohorts, respectively, compared with four well-known databases including CGC (80), OncoKB (33), IntOGen (37) and CTAT (30) (Figure 4B, C). Furthermore, we collected the LOEUF value for each gene calculated by using 141,456 human genomes from gnomAD (41). High LOEUF scores suggest a relatively higher tolerance to inactivation while low LOEUF scores indicate strong selection against predicted loss-of-function variation in a given gene. The distribution of LOEUF of driver genes from both TCGA (mean LOEUF = 0.5062) and ICGC (mean LOEUF = 0.5028) pan-cancer cohorts maintain the same trend with that of cancer essential genes (mean LOEUF = 0.7533) but not cancer nonessential genes (mean LOEUF = 1.5739) (Figure 4D). The mean LOEUF values of TCGA driver genes, ICGC driver genes and cancer essential genes are significantly lower than that of cancer nonessential genes (Mann–Whitney U test, $P < 2.2e–16$). Our results indicated that cancer driver genes are much more constraint than cancer nonessential genes in the human population and these driver genes are essential for carcinogenesis in human cancers. To demonstrate the functions of the novel driver genes identified by OncoVar, we employed the DepMap database (https://depmap.org/portal/) to further investigate whether these novel driver genes are cancer dependency genes. Interestingly, we found 19 out of

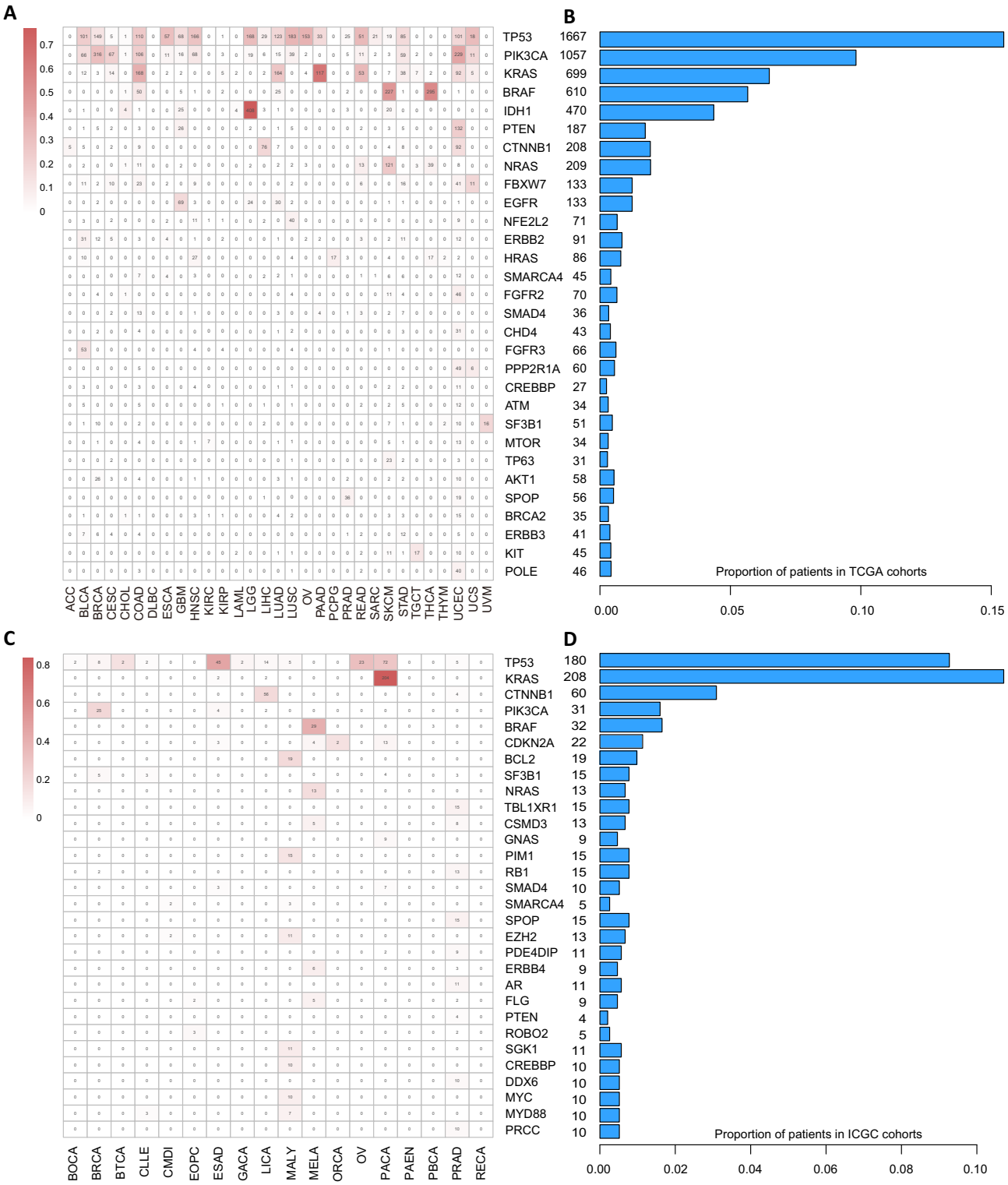

Figure 2. A snapshot of driver mutations identified from TCGA and ICGC cohorts. (A, C). Top 30 enriched driver genes by the driver mutations across each cancer type from TCGA cohorts (**A**) and ICGC cohorts (**C**). (B, D) TP53 is the top one mutated gene across all cancers and detected with driver mutations in ∼23% and ∼17% patients from TCGA (**B**) and ICGC (**D**) cohorts, respectively.

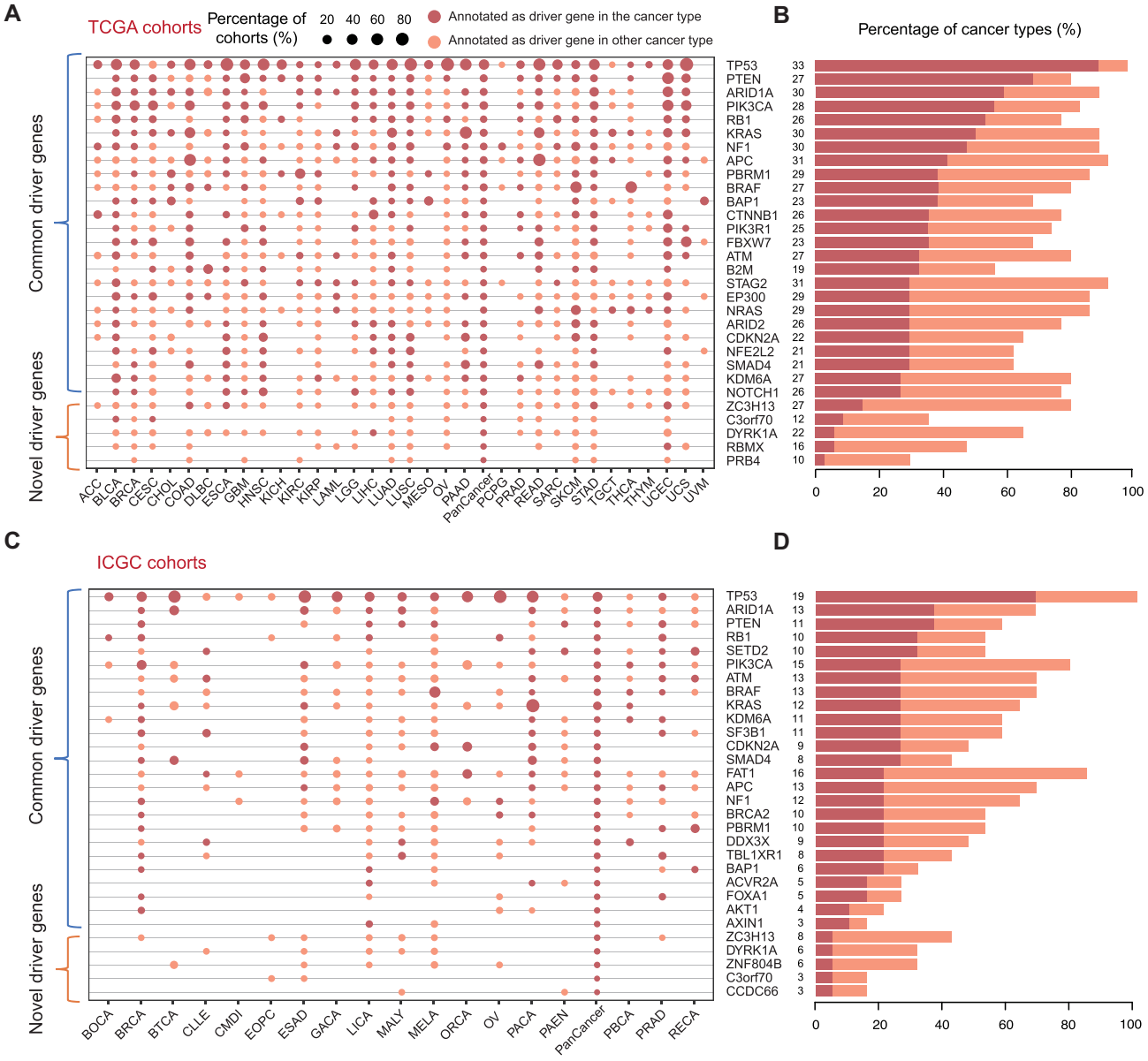

Figure 3. A snapshot of driver genes identified in TCGA and ICGC cohorts. (**A, C**). The range of cancer types with 25 exemplary common driver genes and five exemplary novel driver genes represented as dots. The size of the dots represents the percentage of all cohorts of the cancer type in which the gene is identified as a driver and with driver or extreme mutation. (**B, D**). The number and percentage of cancer types in which each gene appears as a driver in all cancer types is represented in the bars.

23 (82.6%) and 6 out of 7 (85.7%) novel driver genes from TCGA and ICGC pan-cancer cohorts are related to cancer vulnerabilities of at least one kind of cancers which were validated by CRISPR or RNAi knockout libraries, respectively (Supplementary Table S7).

The panorama of oncogenic pathways

Genetic alterations in signaling pathways that control cell-cycle progression, apoptosis, and cell growth are common hallmarks of cancers (81). However, other kinds of signaling pathways and gene ontologies (GOs) are rarely investigated. Briefly, we identified 1941 and 1919 GOs and KEGG pathways from TCGA and ICGC pan-cancer cohorts (Supplementary Table S8). Consistent with a previous study (82), we found that enriched GOs and pathways are shared across anatomical origins and cell types in both TCGA and ICGC cohorts (Supplementary Figure S3). For instance, the top 30 enriched pathways ranked by FDR in TCGA pan-cancer include well-known cancer pathways such as the ErbB signaling pathway, cell-cycle regulation, apoptosis regulation, p53 signaling pathway, MAPK signaling pathway, mTOR signaling pathway, Ras signaling pathway, PI3K-Akt signaling pathway, regulation of mitotic cell cycle and epithelial cell proliferation (Figure 5A, Supplementary Figure S4A). Similarly, top 30 functional terms in ICGC pan-cancer contain common cancer pathways such as PI3K–Akt signaling pathway, Wnt signaling pathway, mTOR signaling path-

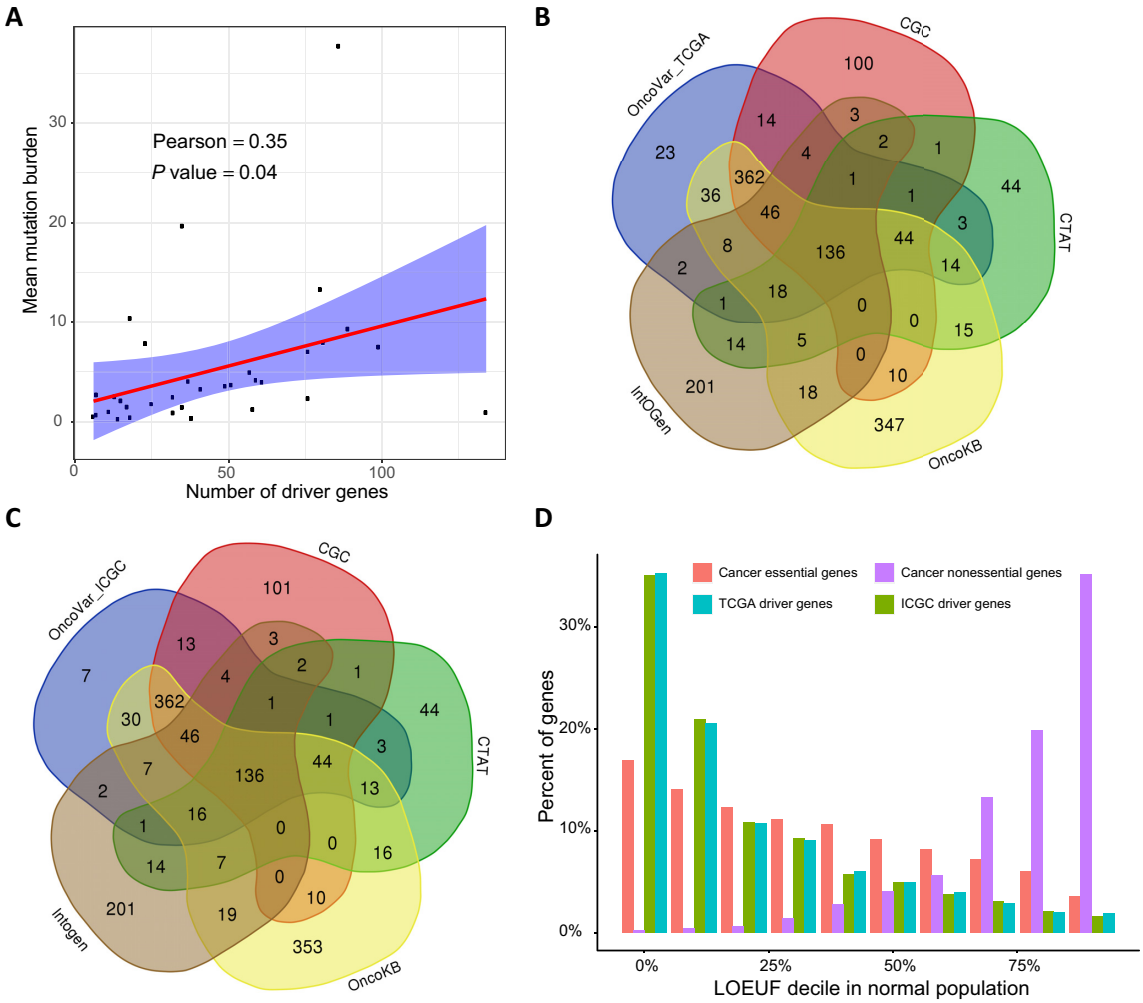

Figure 4. Features of identified driver genes. (**A**). A significant positive correlation between average mutation burden in a cancer type and the number of identified driver genes. (**B, C**). OncoVar identified many known driver genes which are consistent with results by other methods including CGC (49), CTAT (31), OncoKB (50), IntOGen (37) and several novel driver genes. (**D**). Percent of genes among different loss-of-function genes observed/expected upper bound fraction (LOEUF) (28) in a normal population from gnomAD.

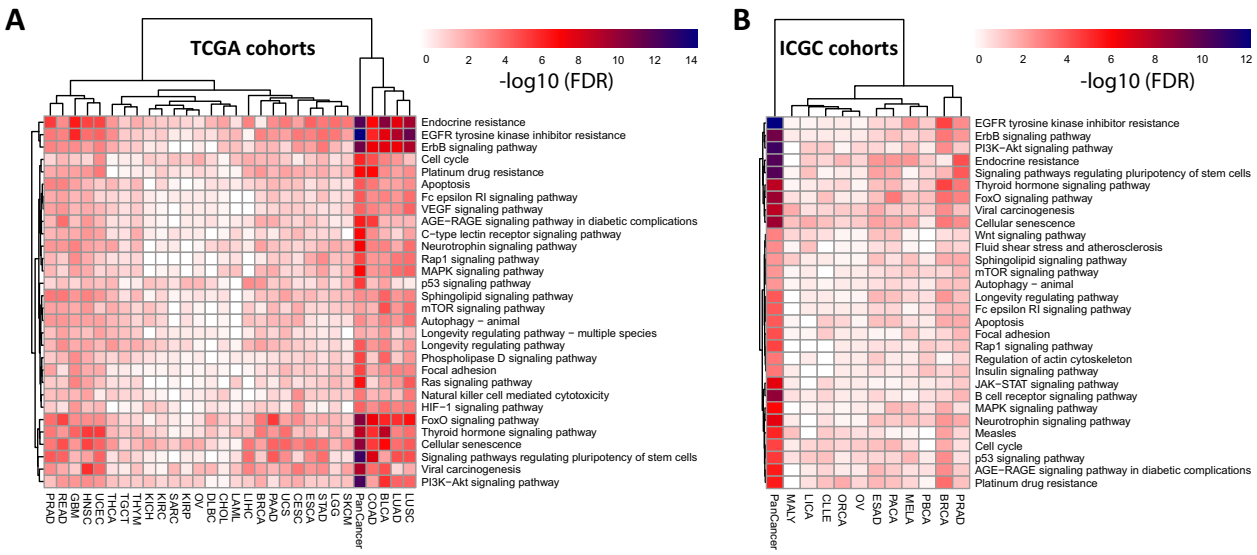

Figure 5. Top 30 enriched KEGG pathways in TCGA (**A**) and ICGC (**B**).

way, B cell receptor signaling pathway, JAK-STAT signaling pathway, G1/S transition of mitotic cell cycle, lymphocyte differentiation and histone modification (Figure 5B, Supplementary Figure S4B). Some pathways were mutated across most of the cancer types while other pathways were more specific to specific tumor types. Our platform OncoVar provides enrichment summary and search of GO and pathways in each cancer type and pan-cancer cohort from TCGA and ICGC projects.

Mutually exclusive and co-occurring driver genes

Mutually exclusive patterns between alterations across large patient cohorts have been associated with functional redundancy, indicating that once one occurred, the second will not provide a further selective advantage, or alternatively indicating that cells cannot survive with both alterations with synthetic lethality. On the other hand, co-occurrence patterns of alterations in many tumor samples indicate functional synergies and may reflect therapeutic resistance targeting one of the alterations (81). We employed Maftools to explore significantly mutually exclusive and co-occurring driver genes for pan-cancer and each cancer cohort. Among the top 50 cancer driver genes in TCGA pan-cancer cohorts, we found that the TP53, IDH1 and BRAF gene is most likely to be mutually exclusive with other altered driver genes (Supplementary Figure S5 and S6), indicating that oncogenic alteration in one of these three genes is sufficient to carcinogenesis. In terms of co-occurrence patterns, we found that the rest of driver genes, such as NOTCH1, EGFR, PIK3CA, MTOR and EP300 gene are most likely to be co-mutated with other driver genes. Our results indicated that drug combinations may improve efficient treatments based on the occurrence of actionable alterations across different tumor types. Our platform OncoVar provides mutually exclusive and co-occurring patterns of top 50 driver genes in each cancer type and pan-cancer cohort from TCGA and ICGC projects.

The landscape of targeted therapies

The development of therapies targeting altered driver proteins provides the promise of precision cancer medicine (83). We identified 30 driver genes predicted to be clinically actionable genes by OncoKB (33) in 29 cancer types from TCGA cohorts (Figure 6A, B). These driver genes harbored different numbers of driver mutations in different cancer types. For instance, PIK3CA, KRAS and BRAF harbored 1457, 805 and 707 driver mutations from pan-cancer cohorts, respectively. Specially, PIK3CA gene had 365 and 283 driver mutations in breast invasive carcinoma (BRCA) and thyroid cancer (THCA), respectively. In addition, BRAF gene had 270 and 311 driver mutations in skin cutaneous melanoma (SKCM) and THCA, respectively. The KRAS gene had 181 and 180 driver mutations in colon adenocarcinoma (COAD) and lung adenocarcinoma (LUAD), respectively. PTEN gene has 162 driver mutations in uterine corpus endometrial carcinoma (UCEC). Moreover, we identified 28 driver genes that harbor potentially actionable mutations annotated by OncoKB database. 28.78% of tumors had multiple targetable driver genes in TCGA pan-cancer cohorts (Figure 6A), indicating opportunities for combination therapy, which is consistent with a previous study (81). We then explored combination therapeutic opportunities based on the actionable mutations which were currently approved for clinical therapies or investigational therapies. Intriguingly, we revealed 148 kinds of potential combination therapies within all actionable driver genes in 19 cancer types (Supplementary Table S9). Specially, we revealed 52 kinds of potential combination therapies within top 10 actionable driver genes, including 22 tri-combinations and 30 bi-combinations in 16 cancer types (Figure 6C, D). We found that the COAD and UCEC cohorts have the most possibilities for combination therapies within 19.17% and 25.00% patients, respectively (Figure 6C). COAD has 13 possible therapy combinations such as inhibition of PIK3CA+KRAS and PIK3CA+BRAF, while UCEC has 18 possible therapy combinations such as PIK3CA + KRAS and PIK3CA + PTEN (Figure 6D). In addition, we validated these 30 actionable driver genes in 11 cancer types from ICGC cohorts (Supplementary Figure S7). Continuous update of drug–gene interaction in OncoVar will improve the *in-silico* prescription based on actionable cancer drivers, thus increasing additional targeting opportunities in personalized cancer medicine.

DISCUSSION

Recent studies began to pay efforts to systematically investigate potential driver mutations (84), cancer driver genes (30) and oncogenic signaling pathways (81) in different cancer types. Several studies evaluated the performance of existing tools for predicting cancer-causing mutations, but the results showed that identifying oncogenic driving mutations remains a significant challenge (8,79). Herein, we employed our recently developed method AI-Driver, which integrates 23 pathogenicity scores using machine learning algorithm, to determine the 'driverness' level of a somatic mutation. Rapid progress has been made in computational approaches to prioritize cancer driver genes. Nevertheless, driver gene lists predicted from these computational tools lack consistency and are prone to false positives (10). Current research is far from achieving the ultimate goal of discovering a complete catalog of driver mutations and genes truly associated with human cancers.

Currently, several databases and frameworks have been developed to integrate driver genes from large-scale genomic data (2), such as DriverDBv3 (36) and IntOGen (85). DriverDBv3 is a multi-omics database for cancer driver gene research which applies published bioinformatics algorithms to determine driver genes along with molecular features and provides an informative visualization of integrative cancer omics data (36). IntOGen is a framework for systematic and automatic identification of mutational driver genes across tumor types (85). Nevertheless, these platforms are not convenient to simultaneously annotate and prioritize tens of thousands of somatic mutations and mutated genes detected by large-scale genomic sequencing. DriverDBv3 predicted driver genes based on the individual criteria of various algorithms but did not combine the driver gene predictions into a consensus ranking (36). The IntOGen framework employed a weighted method to combine

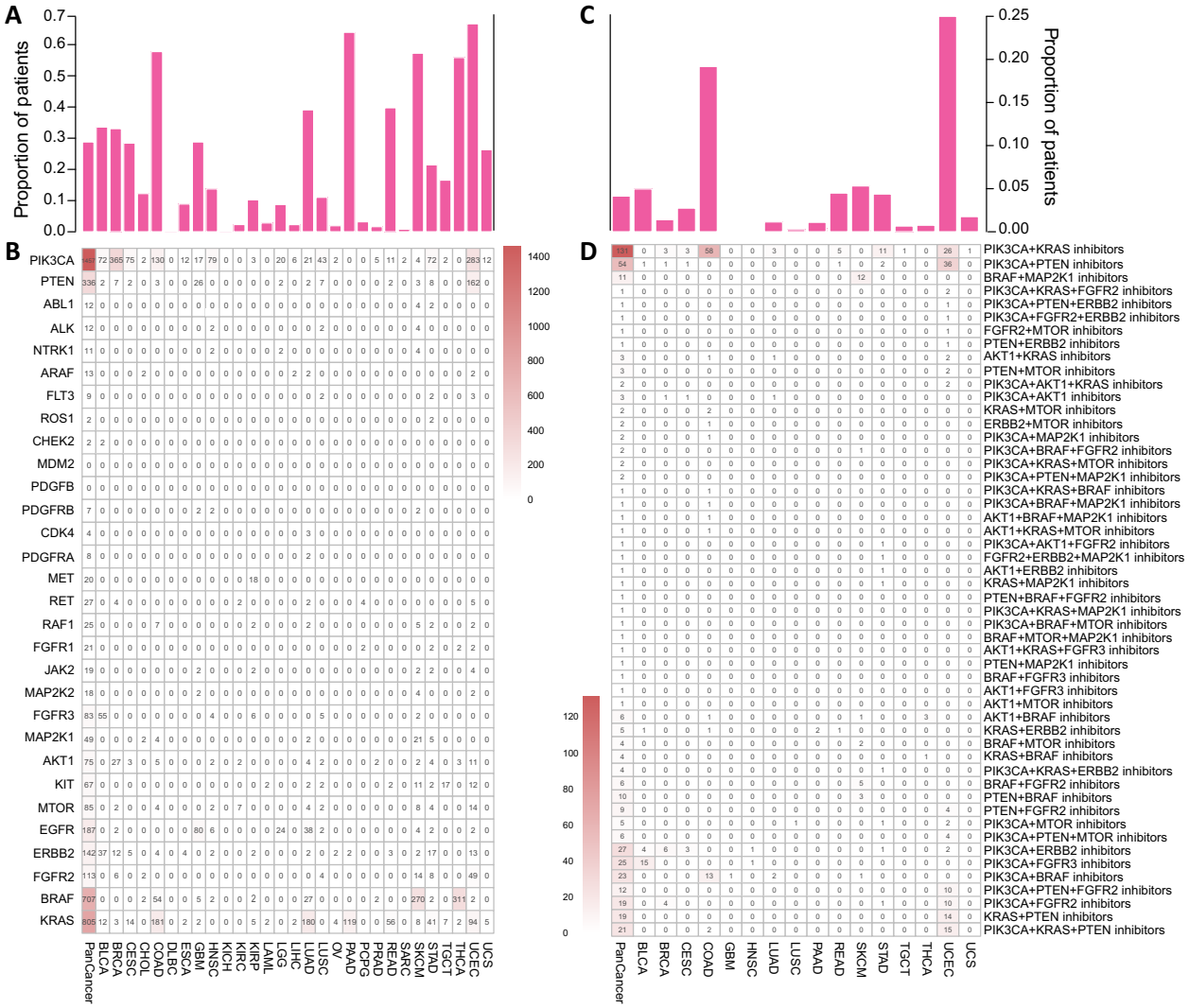

Figure 6. Driver genes and their candidate actionable genes in TCGA. (**A**). Cumulated proportion of patients of all the thirty driver genes identified by OncoVar and annotated as the candidate actionable genes by OncoKB for each cancer type. (**B**). The number of patients for each cancer type in which the gene is identified as a driver and with driver or extreme mutation. (**C**). Cumulated proportion of patients with multiple candidate actionable gene combinations for each cancer type. (**D**). The number of patients for each cancer type with multiple candidate actionable gene combinations.

driver predictions of seven methods to render a consensus ranking of genes while IntOGen did not integrate known driver mutations and genes from other resources. Whereas, OncoVar employed a new strategy to combine driver predictions as well as known driver genes and predicted driver mutations by our recently developed method AI-Driver. OncoVar is an integrated database to systematically prioritize the oncogenic ability of somatic mutations and mutated genes detected from large cancer sequencing projects. Herein, we have analyzed driver events from TCGA and ICGC sequencing projects in the current release while the collection of driver events from other cancer projects/literatures will be updated regularly in the future.

Apart from mutational driver events by coding point mutations, copy-number alterations (CNAs) and rearrangements can also act as cancer drivers in cancer development though nearly 80% of cancer patients harbored at least one mutational driver events (86). In this database, we purposefully focused on missense driver mutations affecting protein-coding genes instead of CNAs or rearrangements. On the other hand, large number of non-coding somatic mutations are linked to regulatory networks (4) though non-coding *cis*-regulatory driver mutations in known cancer genes are much less frequent than protein-coding ones (87). Moreover, our previous studies demonstrated that synonymous mutations involved in posttranscriptional dysregulation and enhancer regions involved in epigenome remodeling could contribute to the etiology of human cancers (88,89). Therefore, we need to pay more attention to driver events occurred in not only coding regions but also non-coding regions in the future research.

In summary, OncoVar is a useful platform which employs published bioinformatics algorithms to describe mutational driver patterns of large-scale genomic sequencing datasets. First, we identified 16 923 and 3409 highly confident driver mutations from the TCGA and ICGC cohorts by using our

machine learning method AI-Driver, respectively. Second, we identified 806 and 696 driver genes from TCGA and ICGC individual cancer types, respectively. Third, we determined 1941 and 1919 GO/KEGG pathways from TCGA and ICGC individual cancer types, respectively. Finally, drug annotations of driver genes provide opportunities of bi-combination and tri-combination therapy in some kinds of cancer types. To our knowledge, OncoVar is the first integrated database which was designed to explore the driver events and interpret their putative mechanism of carcinogenesis across tumor types by incorporating cancer driver predictions and prior oncology knowledge. OncoVar aids the identification of drivers across tumor types and helps rank mutations or genes for better decision-making for the clinical and scientific community interested in cancer precision medicine. We are dedicated to maintaining and improving OncoVar since it is a useful resource for both research and clinical community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author contributions: F.B.M. and Z.S.S. conceived and designed the project; T.W. and M.C.Y. collected datasets from public resources. T.W. established and maintained the full functional database; S.S.R. and J.N.Z. tested and debugged the database. T.W. and X.H.S. drew the figures. F.B.M., S.S.R. and X.L.Z. wrote the manuscript. S.S.R. and H.J.T. revised the manuscript.

FUNDING

National Natural Science Foundation of China [31872237, 81730036 and 81525007]; National Key R&D Program of China [2016YFC0900400]; National High-tech R&D Program of China [2012AA02A210]. Funding for open access charge: National Natural Science Foundation of China [31872237]; National Key R&D Program of China [2016YFC0900400]; National High-tech R&D Program of China [2012AA02A210].
Conflict of interest statement. None declared.

REFERENCES

1. Bamshad,M.J., Ng,S.B., Bigham,A.W., Tabor,H.K., Emond,M.J., Nickerson,D.A. and Shendure,J. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, **12**, 745–755.
2. Mao,F., Liu,Q., Zhao,X., Yang,H., Guo,S., Xiao,L., Li,X., Teng,H., Sun,Z. and Dou,Y. (2018) EpiDenovo: a platform for linking regulatory de novo mutations to developmental epigenetics and diseases. *Nucleic Acids Res.*, **46**, D92–D99.
3. Martincorena,I. and Campbell,P.J. (2015) Somatic mutation in cancer and normal cells. *Science*, **349**, 1483–1489.
4. Li,X., Shi,L., Wang,Y., Zhong,J., Zhao,X., Teng,H., Shi,X., Yang,H., Ruan,S., Li,M. *et al.* (2019) OncoBase: a platform for decoding regulatory somatic mutations in human cancers. *Nucleic Acids Res.*, **47**, D1044–D1055.
5. Wang,Y., Li,G., Mao,F., Li,X., Liu,Q., Chen,L., Lv,L., Wang,X., Wu,J., Dai,W. *et al.* (2014) Ras-induced epigenetic inactivation of the RRAD (Ras-related associated with diabetes) gene promotes glucose uptake in a human ovarian cancer model. *J. Biol. Chem.*, **289**, 14225–14238.
6. Cho,A., Shim,J.E., Kim,E., Supek,F., Lehner,B. and Lee,I. (2016) MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biol.*, **17**, 129.
7. Li,J.C., Zhao,T.T., Zhang,Y., Zhang,K., Shi,L.S., Chen,Y., Wang,X.X. and Sun,Z.S. (2018) Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res.*, **46**, 7793–7804.
8. Chen,H., Li,J., Wang,Y., Ng,P.K., Tsang,Y.H., Shaw,K.R., Mills,G.B. and Liang,H. (2020) Comprehensive assessment of computational algorithms in predicting cancer driver mutations. *Genome Biol.*, **21**, 43.
9. Wang,H., Wang,T., Zhao,X., Wu,H., You,M., Sun,Z. and Mao,F. (2020) AI-Driver: an ensemble method for identifying driver mutations in personal cancer genomes. *NAR Genomics Bioinformatics*, **2**, doi:10.1093/nargab/lqaa84.
10. Han,Y., Yang,J., Qian,X., Cheng,W.C., Liu,S.H., Hua,X., Zhou,L., Yang,Y., Wu,Q., Liu,P. *et al.* (2019) DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. *NucleicAcids Res.*, **8**, e45.
11. Youn,A. and Simon,R. (2011) Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, **27**, 175–181.
12. Gonzalez-Perez,A. and Lopez-Bigas,N. (2012) Functional impact bias reveals cancer drivers. *Cancer Res.*, **40**, e169.
13. Reimand,J. and Bader,G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.*, **9**, 637.
14. Lawrence,M.S., Stojanov,P., Polak,P., Kryukov,G.V., Cibulskis,K., Sivachenko,A., Carter,S.L., Stewart,C., Mermel,C.H., Roberts,S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
15. Lu,Y., Hua,X., Xu,H.M. and Liu,P.Y. (2012) DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *Cancer Res.*, **93**, 439-451.
16. Mularoni,L., Sabarinathan,R., Deu-Pons,J., Gonzalez-Perez,A. and Lopez-Bigas,N. (2016) OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.*, **17**, 128.
17. Ciriello,G., Cerami,E., Sander,C. and Schultz,N. (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, **22**, 398–406.
18. Vandin,F., Upfal,E. and Raphael,B.J. (2012) De novo discovery of mutated driver pathways in cancer. *Genome Res.*, **22**, 375–385.
19. Zhao,J., Zhang,S., Wu,L.Y. and Zhang,X.S. (2012) Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics*, **28**, 2940–2947.
20. Bashashati,A., Haffari,G., Ding,J., Ha,G., Lui,K., Rosner,J., Huntsman,D.G., Caldas,C., Aparicio,S.A. and Shah,S.P. (2012) DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.*, **13**, R124.
21. Hou,J.P. and Ma,J. (2014) DawnRank: discovering personalized driver genes in cancer. *Genome Med*, **6**, 56.
22. Guo,W.F., Zhang,S.W., Liu,L.L., Liu,F., Shi,Q.Q., Zhang,L., Tang,Y., Zeng,T. and Chen,L. (2018) Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics*, **34**, 1893–1903.
23. Hou,Y., Gao,B., Li,G. and Su,Z. (2018) MaxMIF: a new method for identifying cancer driver genes through effective data integration. *Adv Sci (Weinh)*, **5**, 1800640.
24. Tamborero,D., Gonzalez-Perez,A. and Lopez-Bigas,N. (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, **29**, 2238–2244.
25. Porta-Pardo,E. and Godzik,A. (2014) e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics*, **30**, 3109–3114.
26. Jia,P.L., Wang,Q., Chen,Q.X., Hutchinson,K.E., Pao,W. and Zhao,Z.M. (2014) MSEA: detection and quantification of mutation hotspots through mutation set enrichment analysis. *Genome Biol.*, **15**, 489.
27. Chung,I.F., Chen,C.Y., Su,S.C., Li,C.Y., Wu,K.J., Wang,H.W. and Cheng,W.C. (2016) DriverDBv2: a database for human cancer driver gene research. *Nucleic Acids Res.*, **44**, D975–D979.
28. Watson,I.R., Takahashi,K., Futreal,P.A. and Chin,L. (2013) Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.*, **14**, 703–718.

29. Tokheim,C.J., Papadopoulos,N., Kinzler,K.W., Vogelstein,B. and Karchin,R. (2016) Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 14330–14335.
30. Bailey,M.H., Tokheim,C., Porta-Pardo,E., Sengupta,S., Bertrand,D., Weerasinghe,A., Colaprico,A., Wendl,M.C., Kim,J., Reardon,B. *et al.* (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell*, **173**, 371–385.
31. Li,J.C., Shi,L.S., Zhang,K., Zhang,Y., Hu,S.S., Zhao,T.T., Teng,H.J., Li,X.F., Jiang,Y., Ji,L.Y. *et al.* (2018) VarCards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic Acids Res.*, **46**, D1039–D1048.
32. Tate,J.G., Bamford,S., Jubb,H.C., Sondka,Z., Beare,D.M., Bindal,N., Boutselakis,H., Cole,C.G., Creatore,C., Dawson,E. *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **47**, D941–D947.
33. Chakravarty,D., Gao,J., Phillips,S.M., Kundra,R., Zhang,H., Wang,J., Rudolph,J.E., Yaeger,R., Soumerai,T., Nissan,M.H. *et al.* (2017) OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.*, **2017**, PO.17.00011.
34. Dietlein,F., Weghorn,D., Taylor-Weiner,A., Richters,A., Reardon,B., Liu,D., Lander,E.S., Van Allen,E.M. and Sunyaev,S.R. (2020) Identification of cancer driver genes based on nucleotide context. *Nat. Genet.*, **52**, 208–218.
35. Newberg,J.Y., Mann,K.M., Mann,M.B., Jenkins,N.A. and Copeland,N.G. (2018) SBCDDB: Sleeping Beauty Cancer Driver Database for gene discovery in mouse models of human cancers. *Nucleic Acids Res.*, **46**, D1011–D1017.
36. Liu,S.H., Shen,P.C., Chen,C.Y., Hsu,A.N., Cho,Y.C., Lai,Y.L., Chen,F.H., Li,C.Y., Wang,S.C., Chen,M. *et al.* (2020) DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Res.*, **48**, D863–D870.
37. Gonzalez-Perez,A., Perez-Llamas,C., Deu-Pons,J., Tamborero,D., Schroeder,M.P., Jene-Sanz,A., Santos,A. and Lopez-Bigas,N. (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods*, **10**, 1081–1082.
38. Bertrand,D., Drissler,S., Chia,B.K., Koh,J.Y., Li,C., Suphavilai,C., Tan,I.B. and Nagarajan,N. (2018) ConsensusDriver improves upon individual algorithms for predicting driver alterations in different cancer types and individual patients. *Cancer Res.*, **78**, 290–301.
39. Zhu,C.Y., Zhou,C., Chen,Y.Q., Shen,A.Z., Guo,Z.M., Yang,Z.Y., Ye,X.Y., Qu,S., Wei,J. and Liu,Q. (2019) C(3): Consensus cancer driver gene caller. *Genomics Proteomics Bioinformatics*, **17**, 311–318.
40. Ellrott,K., Bailey,M.H., Saksena,G., Covington,K.R., Kandoth,C., Stewart,C., Hess,J., Ma,S., Chiotti,K.E., McLellan,M. *et al.* (2018) Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.*, **6**, 271–281.
41. Karczewski,K.J., Francioli,L.C., Tiao,G., Cummings,B.B., Alföldi,J., Wang,Q., Collins,R.L., Laricchia,K.M., Ganna,A., Birnbaum,D.P. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. **581**, 434–443.
42. Kosmicki,J.A., Samocha,K.E., Howrigan,D.P., Sanders,S.J., Slowikowski,K., Lek,M., Karczewski,K.J., Cutler,D.J., Devlin,B., Roeder,K. *et al.* (2017) Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.*, **49**, 504–510.
43. Wang,Z.J., Cai,W.S., Cui,F., Cai,T., Chen,Z.H., Mao,F.B., Teng,H.J., Chen,L., Wang,J.S., Sun,Z.S. *et al.* (2014) Identification of a novel missense (C7W) mutation of SOD1 in a large familial amyotrophic lateral sclerosis pedigree. *Neurobiol. Aging*, **35**, 725.
44. Li,J., Cai,T., Jiang,Y., Chen,H., He,X., Chen,C., Li,X., Shao,Q., Ran,X., Li,Z. *et al.* (2016) Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol. Psychiatry*, **21**, 290–297.
45. Chen,S.Y., Li,M., Zhu,W.J., Mao,F.B., Wang,J.S., Sun,Z.S. and Huang,X.S. (2016) A novel 10-base pair insertion mutation in exon 5 of the SOD1 gene in a Chinese family with amyotrophic lateral sclerosis. *Neurobiol. Aging*, **45**, 212.
46. Zhu,Y., Zhu,M.X., Zhang,X.D., Xu,X.E., Wu,Z.Y., Liao,L.D., Li,L.Y., Xie,Y.M., Wu,J.Y. and Zou,H.Y. (2016) SMYD3 stimulates EZR and LOXL2 transcription to enhance proliferation, migration, and invasion in esophageal squamous cell carcinoma. *Hum. Pathol.*, **52**, 153–163.
47. Jia,Z., Mao,F.B., Wang,L., Li,M.Z., Shi,Y.Y., Zhang,B.R. and Gao,G.L. (2017) Whole-exome sequencing identifies a de novo mutation in TRPM4 involved in pleiotropic ventricular septal defect. *Int. J. Clin. Exp. Pathol.*, **10**, 5092–5104.
48. Liang,J.L., Cai,W.S., Feng,D.D., Teng,H.J., Mao,F.B., Jiang,Y., Hu,S.S., Li,X.F., Zhang,Y.J., Liu,B.G. *et al.* (2018) Genetic landscape of papillary thyroid carcinoma in the Chinese population. *J. Pathol.*, **244**, 215–226.
49. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
50. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1082.
51. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
52. Chun,S. and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.*, **19**, 1553–1561.
53. Jian,X.Q., Boerwinkle,E. and Liu,X.M. (2014) In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet. Med.*, **16**, 497–503.
54. Reva,B., Antipin,Y. and Sander,C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, E118.
55. Shihab,H.A., Gough,J., Cooper,D.N., Stenson,P.D., Barker,G.L.A., Edwards,K.J., Day,I.N.M. and Gaunt,T.R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, **34**, 57–65.
56. Choi,Y. and Chan,A.P. (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, **31**, 2745–2747.
57. Dong,C.L., Wei,P., Jian,X.Q., Gibbs,R., Boerwinkle,E., Wang,K. and Liu,X.M. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*, **24**, 2125–2137.
58. Carter,H., Douville,C., Stenson,P.D., Cooper,D.N. and Karchin,R. (2013) Identifying mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, **14**, S3.
59. Jagadeesh,K.A., Wenger,A.M., Berger,M.J., Guturu,H., Stenson,P.D., Cooper,D.N., Bernstein,J.A. and Bejerano,G. (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.*, **48**, 1581–1586.
60. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
61. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using gerp plus. *PLoS Comput. Biol.*, **6**, e1001025.
62. Quang,D., Chen,Y.F. and Xie,X.H. (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, **31**, 761–763.
63. Shihab,H.A., Rogers,M.F., Gough,J., Mort,M., Cooper,D.N., Day,I.N.M., Gaunt,T.R. and Campbell,C. (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, **31**, 1536–1543.
64. Ionita-Laza,I., McCallum,K., Xu,B. and Buxbaum,J.D. (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.*, **48**, 214–220.
65. Lu,Q.S., Hu,Y.M., Sun,J.H., Cheng,Y.W., Cheung,K.H. and Zhao,H.Y. (2015) A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep-Uk*, **5**, 10576.
66. Gulko,B., Hubisz,M.J., Gronau,I. and Siepel,A. (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.*, **47**, 276–283.
67. Siepel,A., Pollard,K.S. and Haussler,D. (2006) New methods for detecting lineage-specific selection. *Lect. Notes Comput. Sci.*, **3909**, 190–205.
68. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M.M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

69. Garber,M., Guttman,M., Clamp,M., Zody,M.C., Friedman,N. and Xie,X.H. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**, I54–I62.
70. Ioannidis,N.M., Rothstein,J.H., Pejaver,V., Middha,S., McDonnell,S.K., Baheti,S., Musolf,A., Li,Q., Holzinger,E., Karyadi,D. *et al.* (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.*, **99**, 877–885.
71. Li,J., Zhao,T., Zhang,Y., Zhang,K., Shi,L., Chen,Y., Wang,X. and Sun,Z. (2018) Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res.*, **46**, 7793–7804.
72. Lawrence,M.S., Stojanov,P., Mermel,C.H., Robinson,J.T., Garraway,L.A., Golub,T.R., Meyerson,M., Gabriel,S.B., Lander,E.S. and Getz,G. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.
73. Chen,W.H., Lu,G., Chen,X., Zhao,X.M. and Bork,P. (2017) OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res.*, **45**, D940–D944.
74. Yu,G.C., Wang,L.G., Han,Y.Y. and He,Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.
75. Mayakonda,A., Lin,D.C., Assenov,Y., Plass,C. and Koeffler,H.P. (2018) Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.*, **28**, 1747–1756.
76. Finan,C., Gaulton,A., Kruger,F.A., Lumbers,R.T., Shah,T., Engmann,J., Galver,L., Kelley,R., Karlsson,A., Santos,R. *et al.* (2017) The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.*, **9**, eaag1166.
77. Cotto,K.C., Wagner,A.H., Feng,Y.Y., Kiwala,S., Coffman,A.C., Spies,G., Wollam,A., Spies,N.C., Griffith,O.L. and Griffith,M. (2018) DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res.*, **46**, D1068–D1073.
78. Tamborero,D., Rubio-Perez,C., Deu-Pons,J., Schroeder,M.P., Vivancos,A., Rovira,A., Tusquets,I., Albanell,J., Rodon,J., Tabernero,J. *et al.* (2018) Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med*, **10**, 25.
79. Martelotto,L.G., Ng,C.K., De Filippo,M.R., Zhang,Y., Piscuoglio,S., Lim,R.S., Shen,R., Norton,L., Reis-Filho,J.S. and Weigelt,B. (2014) Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol.*, **15**, 484.
80. Forbes,S.A., Beare,D., Boutselakis,H., Bamford,S., Bindal,N., Tate,J., Cole,C.G., Ward,S., Dawson,E., Ponting,L. *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.
81. Sanchez-Vega,F., Mina,M., Armenia,J., Chatila,W.K., Luna,A., La,K.C., Dimitriadoy,S., Liu,D.L., Kantheti,H.S., Saghafinia,S. *et al.* (2018) Oncogenic signaling pathways in the cancer genome atlas. *Cell*, **173**, 321–337.
82. Hoadley,K.A., Yau,C., Hinoue,T., Wolf,D.M., Lazar,A.J., Drill,E., Shen,R., Taylor,A.M., Cherniack,A.D., Thorsson,V. *et al.* (2018) Cell-of-Origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, **173**, 291–304.
83. Rubio-Perez,C., Tamborero,D., Schroeder,M.P., Antolin,A.A., Deu-Pons,J., Perez-Llamas,C., Mestres,J., Gonzalez-Perez,A. and Lopez-Bigas,N. (2015) In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell*, **27**, 382–396.
84. Ng,P.K., Li,J., Jeong,K.J., Shao,S., Chen,H., Tsang,Y.H., Sengupta,S., Wang,Z., Bhavana,V.H., Tran,R. *et al.* (2018) Systematic functional annotation of somatic mutations in cancer. *Cancer Cell*, **33**, 450–462.
85. Martinez-Jimenez,F., Muinos,F., Sentis,I., Deu-Pons,J., Reyes-Salazar,I., Arnedo-Pac,C., Mularoni,L., Pich,O., Bonet,J., Kranas,H. *et al.* (2020) A compendium of mutational cancer driver genes. *Nat. Rev. Cancer*, **20**, 555–572.
86. Consortium, I.T.P.-C.A.o.W.G. (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93.
87. Rheinbay,E., Nielsen,M.M., Abascal,F., Wala,J.A., Shapira,O., Tiao,G., Hornshoj,H., Hess,J.M., Juul,R.I., Lin,Z. *et al.* (2020) Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, **578**, 102–111.
88. Teng,H., Wei,W., Li,Q., Xue,M., Shi,X., Li,X., Mao,F. and Sun,Z. (2020) Prevalence and architecture of posttranscriptionally impaired synonymous mutations in 8,320 genomes across 22 cancer types. *Nucleic Acids Res.*, **48**, 1192–1205.
89. Sun,Y., Zhou,B., Mao,F., Xu,J., Miao,H., Zou,Z., Phuc Khoa,L.T., Jang,Y., Cai,S., Witkin,M. *et al.* (2018) HOXA9 reprograms the enhancer landscape to promote leukemogenesis. *Cancer Cell*, **34**, 643–658.