CANCER

CancerVar: An artificial intelligence–empowered platform for clinical interpretation of somatic mutations in cancer

Quan Li[1,2], Zilin Ren[2], Kajia Cao[3], Marilyn M. Li[3,4], Kai Wang[2,4]*, Yunyun Zhou[2]*

Several knowledgebases are manually curated to support clinical interpretations of thousands of hotspot somatic mutations in cancer. However, discrepancies or even conflicting interpretations are observed among these databases. Furthermore, many previously undocumented mutations may have clinical or functional impacts on cancer but are not systematically interpreted by existing knowledgebases. To address these challenges, we developed CancerVar to facilitate automated and standardized interpretations for 13 million somatic mutations based on the AMP/ASCO/CAP 2017 guidelines. We further introduced a deep learning framework to predict oncogenicity for these variants using both functional and clinical features. CancerVar achieved satisfactory performance when compared to several independent knowledgebases and, using clinically curated datasets, demonstrated practical utility in classifying somatic variants. In summary, by integrating clinical guidelines with a deep learning framework, CancerVar facilitates clinical interpretation of somatic variants, reduces manual work, improves consistency in variant classification, and promotes implementation of the guidelines.

INTRODUCTION

A large number of somatic variants have been identified by next-generation sequencing (NGS) during clinical oncology practice to facilitate precision medicine (*1*, *2*). To better understand the clinical impacts of somatic variants in cancer, several knowledgebases have been curated, including OncoKB (*1*), My Cancer Genome (*3*), CIViC (*4*), Precision Medicine Knowledge Base (*5*), JAX Clinical Knowledgebase (CKB) (*6*), and Cancer Genome Interpreter (*7*). Although clinically relevant, interpretation of somatic variants is not a standardized practice, and different clinical groups often generate different or even conflicting results. To standardize clinical interpretation of somatic variants in cancer and support clinical decision-making, the Association for Molecular Pathology (AMP), American Society of Clinical Oncology (ASCO), and College of American Pathologists (CAP) jointly proposed standards and guidelines for the interpretation and reporting of somatic variants, classifying somatic variants into four tiers: strong clinical significance (tier I), potential clinical significance (tier II), uncertain significance (tier III), and benign (tier IV) (*8*). These AMP/ASCO/CAP 2017 guidelines incorporate 12 pieces of evidence, including diagnostic, prognostic, and therapeutic clinical evidence; mutation type; variant allele fraction [mosaic variant frequency (likely somatic) and nonmosaic variant frequency (potential germline)]; population databases; germline databases; somatic databases; predictive results of different computational algorithms; pathway involvement; and publications (*8*, *9*).

However, as the AMP/ASCO/CAP classification scheme heavily relies on published clinical evidence for a given variant, ambiguous assignments among human curators frequently occur when using the same evidence for a given variant. For example, Sirohi *et al.* (*10*) compared human classifications for 51 variants by 20 randomly selected molecular pathologists from 10 institutions. The original overall observed agreement was only 58%; when providing the same evidential data for variants to the pathologists, the agreement rate of reclassification increased to 70%. The reasons for such discordance are as follows: (i) gathering information/evidence is complicated and may not be reproducible by the same interpreter at different time points; (ii) different researchers may use different algorithms, cutoffs, and parameters, rendering interpretation less reproducible; and (iii) newly published evidence for certain variants may not be incorporated into the evaluation system instantly and systematically, which is especially relevant for variants of unknown significance (VUSs).

To standardize interpretation of somatic variants across multiple knowledgebases, a more recently published knowledgebase, MetaKB from The Variant Interpretation for Cancer Consortium, has aggregated evidence based on the AMP/ASCO/CAP 2017 guidelines (*11*). However, this MetaKB knowledgebase also has the following limitations: (i) it only focuses on consensus interpretations for a limited number of known hotspot mutations, such that a large number of variants are now classified as unknown clinical significance, but they may be oncogenic through "loss of function" or "gain of function" in cancer; (ii) it only provides a summarized classification for each variant, without demonstrating itemized evidence in detail when mapping to the 12 criteria of the AMP/ASCO/CAP 2017 guidelines, and therefore, users cannot conduct customized evaluations based on their own protocols, experiences, and updated clinical knowledge; and (iii) it uses a simple scoring system to rank driver mutations without considering heterogeneity of functional consequences (such as predictions of deleteriousness) of variants, especially for newly identified variants reported in the literature.

In clinical practice, when a somatic mutation is considered to have strong confidence in causing a functional impact on protein changes, clinicians likely interpret it with clinical significance or likely clinical significance (*12*, *13*). Although a number of useful software tools (*2*, *14*–*22*), especially sorting intolerant from tolerant

[1] Princess Margaret Cancer Centre, University Health Network, University of Toronto, Toronto, ON M5G2C1, Canada. [2] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [3] Division of Genomic Diagnostics, Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [4] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
*Corresponding author. Email: wangk@chop.edu (K.W.); zhouy6@chop.edu (Y.Z.)

(SIFT) (*23*), PolyPhen-2 (*24*), and functional analysis through hidden Markov models (FATHMM) (*21*, *25*), have been developed to predict functional impacts, disagreements regarding certain mutations are consistently observed. Some of these tools use similar background information based on alignment, evolutionary conservation, and homology, such as MutationAssessor (*20*), FATHMM (*21*, *25*), cancer-specific high-throughput annotation of somatic mutations (CHASM) based on random forest (RF) classifier (*15*), and CanDrA based on support vector machine (SVM) (*16*); others use consensus information by integrating multiple sources of information from many computational tools, such as combined tool adjusted total CTAT-cancer (*2*). Although some meta-analysis tools, such as deleterious annotation of genetic variants using neural networks (DANN) (*26*) and DriverPower (*27*), were later developed to prioritize functionally important variants using more comprehensive functional scoring features as input, they have limitations in jointly modeling clinical impact features based on the AMP/ASCO/CAP guidelines. Because the guidelines tend to be conservative ("negative diagnosis" is preferred to "wrong diagnosis"), more variants than expected were misinterpreted as VUSs (*28*–*36*). In addition, the AMP/ASCO/CAP guidelines only designate seven functional impact prediction tools, such as MutationAssessor (*20*), as the official recommended tools, and only the variant from majority voting (more than four from seven tools) can be considered "with clinical significance," which oversimplifies the heterogeneity of predictions of functional consequence in cancer progression. Although using existing tools may be useful in the prediction of the overall impacts of cancer driver genes, it may not be optimal for prioritizing novel mutations in these genes. In addition to the in silico predictive methods above, interactome network approaches have attracted much attention for identifying oncogenic variants in cancer through genotype-phenotype studies (*37*, *38*). In interactome networks, certain perturbed mutations (network nodes) can disrupt certain signaling pathways and protein-protein interactions (PPIs), resulting in similar cancer phenotypes in different patients. These perturbed mutations, termed "edgetic" mutations, are functionally important but are understudied with existing cancer variant interpretation tools (*39*, *40*). To address these challenges and improve automated clinical interpretations of somatic variants in cancer, there is a strong need for reliable and accurate computational methods using both clinical evidence and functional impact score features.

We previously developed the standalone software VIC written in Java, which was among the first tools for interpreting clinical impacts of somatic variants using a rule-based scoring system based on 12 criteria of the AMP/ASCO/CAP 2017 guidelines (*9*). In the current study, we developed an improved somatic variant interpretation tool called CancerVar (cancer variant interpretation) implemented in Python with an accompanying web server. Compared to VIC, CancerVar is a markedly improved tool providing more options to users: (i) Python implementation provides more flexibility to incorporate CancerVar into custom command-line workflows; (ii) CancerVar involves a user-friendly web server with precomputed clinical evidence for 13 million variants from 1911 cancer census genes through literature mining and database aggregations; (iii) we use a flexible AMP/ASCO/CAP rule-based score system and a deep learning–based scoring system that allows for improved interpretations; and (iv) RESTful application programming interface (API) is used to enable developers to freely access complied knowledge. CancerVar allows users to query clinical interpretations for variants using the chromosome position, cDNA change, or protein change and interactively fine-tune weights of scoring features based on prior knowledge or additional user-specified criteria. The CancerVar web server generates automated text with summarized descriptive interpretations, such as diagnostic, prognostic, targeted drug responses, and clinical trial information for many hotspot mutations, significantly reducing the workload of human reviewers in drafting clinical reports in the practice of precision oncology.

RESULTS

Summary of the functionality of the CancerVar web server

The CancerVar web server provides multiple query options at variant, gene, and copy number alteration (CNA) levels across 30 cancer types and two versions of reference genomes: hg19 (GRCh37) and hg38 (GRCh38). With user-supplied input, CancerVar generates an output web page, with information organized as cards, including interpretation summary, gene overview, mutation information, evidence overview, pathways, clinical publications, protein domains, in silico predictions, and exchangeable information from other knowledgebases. The CancerVar web server provides full details for variants, including all automatically generated criteria, most of the supporting evidence, and predictive scores for clinical significance. In the CancerVar webserver, we compiled a variant database from a list of 1911 cancer censuses or driver genes with 13 million exonic variants from seven existing cancer knowledgebases, including catalogue of somatic mutations in cancer (COSMIC), CIViC, and OncoKB, and two datasets collected from the literature about driver gene predictions (table S1). CancerVar can be accessed at https://cancervar.wglab.org, and the Python command-line program can be downloaded at https://github.com/WGLab/CancerVar.

By implementing a rule-based approach, users have the ability to manually adjust the criteria and perform reinterpretation based on their prior knowledge or experience. If the user already knows the information of each of the scoring criteria for the variant (possibly inferred by using other software tools), they can alternatively compute the clinical significance of the variant from the "Interpret by Criteria" service. Each variant is provided with a prediction score and clinical classification categories, which are strong clinical significance, potential clinical significance, uncertain significance, and likely benign/benign, based on the 12 criteria of the AMP/ASCO/CAP 2017 guidelines. Figure 1 provides descriptions of 12 types of evidence and summarizes the functionality of the CancerVar server. Descriptions of the complete scoring system for each piece of evidence can be found in table S2. In addition, we developed a scoring method for oncogenic prioritization by artificial intelligence (OPAI). Using a deep learning–based approach, CancerVar provides a probability score predicted by OPAI to determine the oncogenicity of a variant using 12 evidence features from the AMP/ASCO/CAP guidelines and 23 functional features with scoring metrics predicted by various computational tools. However, the absolute value of the predictive score cannot provide meaningful information, and the classification greatly depends on the cutoff of the predictive score. Unlike typical deep learning algorithms that train a predictive model from a set of positive and negative training samples, OPAI is based on a semisupervised generative adversarial network that includes observed but unclassified somatic mutations from real-world sequencing data for patients with cancer. Figure 2 illustrates the workflow and architecture of the generator and discriminator/classifier used in

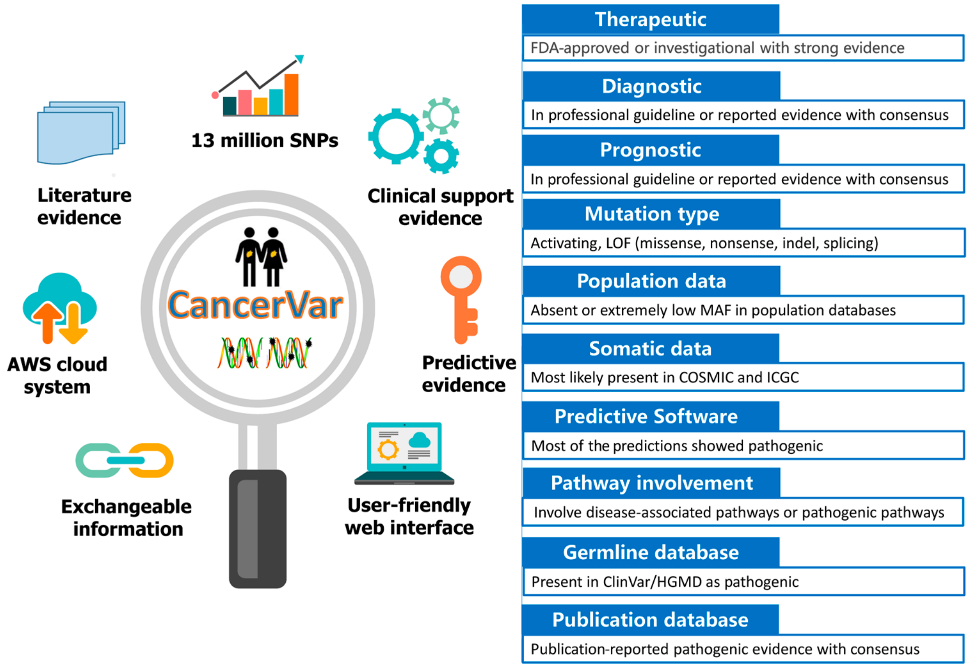

13 million SNPs
Literature evidence
Clinical support evidence
CancerVar
AWS cloud system
Predictive evidence
Exchangeable information
User-friendly web interface
Therapeutic
FDA-approved or investigational with strong evidence
Diagnostic
In professional guideline or reported evidence with consensus
Prognostic
In professional guideline or reported evidence with consensus
Mutation type
Activating, LOF (missense, nonsense, indel, splicing)
Population data
Absent or extremely low MAF in population databases
Somatic data
Most likely present in COSMIC and ICGC
Predictive Software
Most of the predictions showed pathogenic
Pathway involvement
Involve disease-associated pathways or pathogenic pathways
Germline database
Present in ClinVar/HGMD as pathogenic
Publication database
Publication-reported pathogenic evidence with consensus

Fig. 1. Summary of the functionality of CancerVar and descriptions of 12 types of evidence. AWS, Amazon Web Services; LOF, Loss of Function; MAF, Minor allele frequency; HGMD, Human Gene Mutation Database.

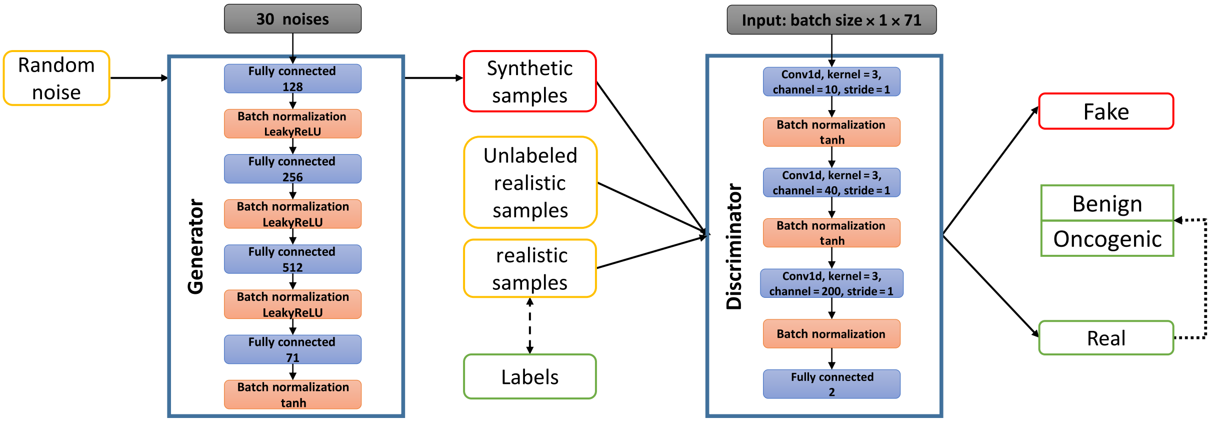


Fig. 2. Workflow and architecture of the generator and discriminator/classifier used in OPAI. The generator contains three linear layers with batch normalization, LeakyReLu as the activation layer, and a 60% dropout rate in each layer. The final layer is a linear layer with batch normalization and tanh as the activation layer. For the discriminator we implemented three Convolutional Neural Network (CNN) layers with tanh as the activation layer.

the generative adversarial network of OPAI. The OPAI's functionality and user instructions are already integrated within CancerVar software, allowing users to predict their own variants in the future.

Comparative evaluations of CancerVar with human interpreters

Sirohi *et al.* (*10*) measured the reliability of the 2017 AMP/ASCO/CAP guidelines using 51 variants [31 single-nucleotide variants (SNVs), 14 insertions and deletions (indels), 5 CNAs, and 1 fusion] based on a literature review. Among these variants, we selected 43, including all 31 SNVs and 12 indel variants (we did not find alternative allele information for two indels in the *CHEK1* and *MET* genes). CancerVar interpreted the clinical impacts of these 43 variants with respect to the specified cancer types. As these 43 variants do not have solid/consistent clinical interpretation, we compared 20 pathologists' opinions from 10 institutions with CancerVar's predictions. As shown in Table 1, CancerVar assigned 21 variants as tier I/II (strong or potential clinical significance). Among these 21 variants, the pathologists classified 17 variants (17 of 21, approximately 81%) as tier I/II, in agreement with CancerVar. Moreover, CancerVar assigned 21 variants as VUS; among these 21 variants, 9 variants (9 of 21, approximately 43%) were also classified as VUS by pathologists. In total, the clinical significance of 26 variants (approximately 61%) matched between human interpreters and CancerVar.

Table 1. Comparison of classification of 43 variants between 20 pathologists and CancerVar.

Annotators	Classifications	20 Pathologists			
		I/II*	III	IV	Total
CancerVar	I/II	17	4	0	21
	III	12	9	0	21
	IV	0	1	0	1
	Total	29	14	0	43

*Tier I, strong clinical significance; tier II, potential clinical significance; tier III, variants of unknown clinical significance (VUS); and tier IV, benign/likely benign.

Details regarding the interpretation of these 43 variants can be found in table S3 and Fig. 3. Compared to human interpreters, the advantage of CancerVar is clear, in that it can automatically generate clinical interpretations with standardized, consistent, and reproducible workflows, with evidence-based support for each of the 12 mentioned criteria. Therefore, CancerVar can greatly reduce the workload of human reviewers and facilitate the generation of precise and reproducible clinical interpretations.

Benchmark studies on OncoKB annotations

OncoKB (*1*), a manually curated database of somatic mutations with oncogenic effects, is widely used in the cancer research community. OncoKB provides an evidence-based classification system to interpret somatic variants and classifies them as inconclusive, likely neutral, predicted oncogenic, likely oncogenic, or oncogenic. In total, 3455 SNVs involving 245 genes were downloaded from the OncoKB annotation database (downloaded 1 March 2020). This version contains 2582 oncogenic/likely oncogenic (O/LO) mutations, 587 likely neutral mutations, and 286 mutations annotated as inconclusive for this study. CancerVar evidence-based and deep learning–based prediction methods were applied to classify the mutations and compared them with OncoKB classifications. For the O/LO group in OncoKB, the CancerVar rule-based method classified 1839 (1839 of 2582, 71.2% consistent with OncoKB classification) variants as having strong or potential clinical significance; the CancerVar deep learning–based method (OPAI) classified 2319 variants (2319 of 2582, 90% consistent with OncoKB classification). The details are given in Table 2, Fig. 4A, and fig. S1. An UpSet plot shows the intersections of predictions between OncoKB, the CancerVar rule-based method, and the deep learning–based method (OPAI). Although most variants were consistently interpreted by the three methods, OPAI appears to be more concordant with OncoKB than rule-based interpretations.

Benchmark studies on CIViC annotations

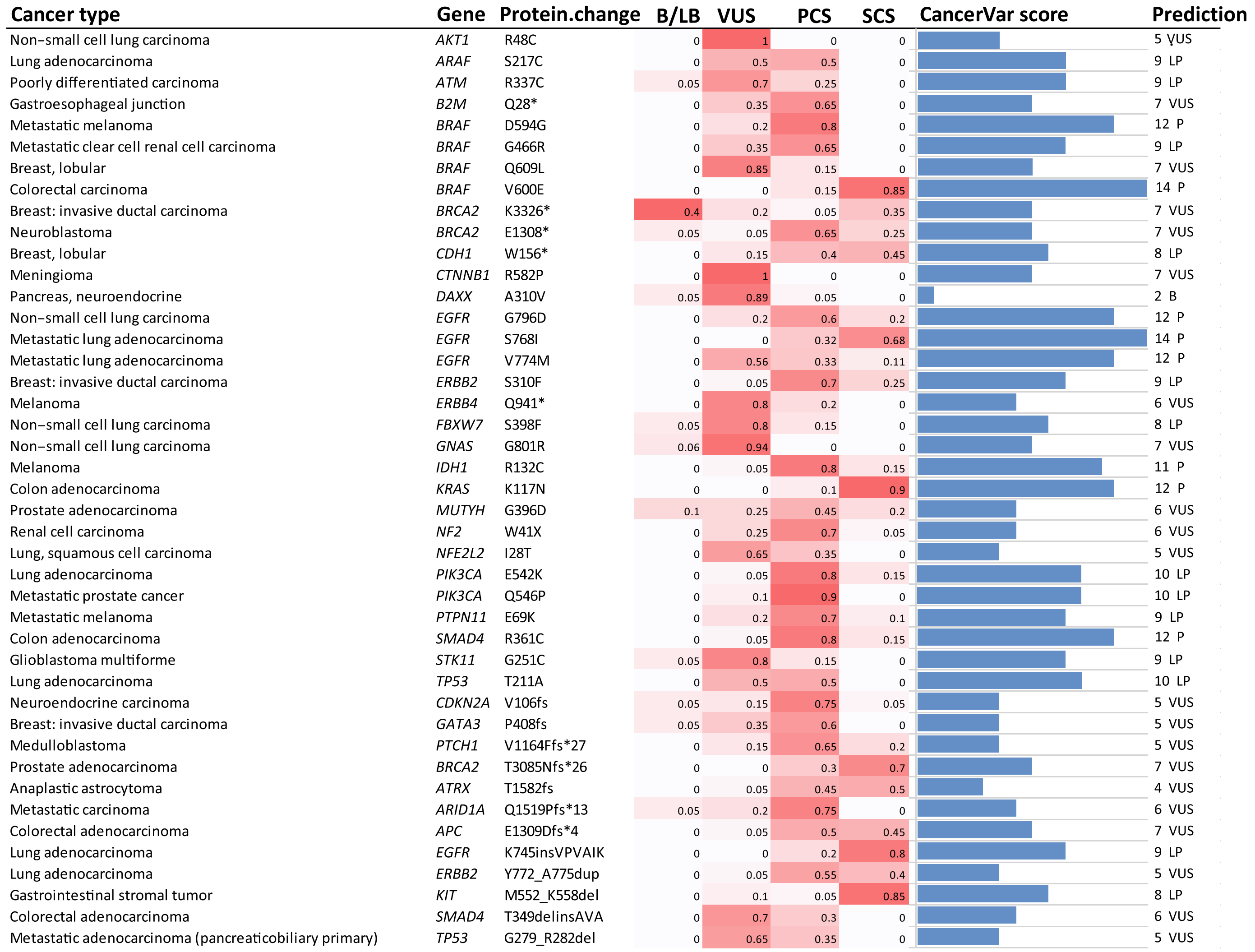
CIViC is a crowdsourced and expert-moderated public resource for somatic variants in cancer (*4*) with five evidence levels to differentiate reported mutations, namely, A: validated, B: clinical, C: case study, D: preclinical, and E: inferential. In total, 1681 unique SNVs/indels of 113 unique genes were retrieved from the CIViC website (https://civicdb.org/releases, accessed 1 May 2020) and assessed by CancerVar. The CancerVar rule-based method predicted 1230 (1230 of 1681, 73.2% consistent with CIViC classification) variants as having strong or potential clinical significance, whereas the CancerVar deep learning–based method OPAI predicted 1581 (94.1% consistent with CIViC classification) variants. Table 3 and Fig. 4B show the details of the CancerVar predictions. Similar to what we report above, OPAI appears to be more concordant with CIViC than rule-based interpretations.

Benchmark studies on IARC TP53 transactivation mutations

TP53 is the most frequently mutated gene in human cancer, and many of its mutants have been functionally assessed based on median transactivation levels and compiled in the International Agency for Research on Cancer (IARC) TP53 database (*41*). On the basis of the median of eight different yeast functional assays (WAF1, MDM2, BAX, h1433s, AIP1, GADD45, noxa, and P53R2), *TP53* mutations can be classified as oncogenic, resulting in lower transactivation (a median transactivation level ≤25% wild type), or neutral, resulting in higher transactivation (level ≥25% wild type). We retrieved 1915 missense mutations (532 mutations used as oncogenic cases and 1383 mutations used as neutral cases) from the IARC TP53 database. For 532 oncogenic mutations, the CancerVar rule-based method predicted 522 [true-positive rate (TPR) = 98%] variants, and the model-based method predicted 512 (TPR = 96.2%) variants as having strong/potential clinical significance. Compared to OncoKB, which predicted 489 (489 of 532 = 91.9%) oncogenic variants, CancerVar rule-based and deep learning–based methods have a higher TPR. The details of CancerVar and OncoKB prediction are shown in Fig. 4C, table S4, and fig. S2.

Benchmark studies on cell viability in vitro assays

The oncogenic effects of somatic mutations can be directly assessed by preferential growth or survival advantage to cells using cellular assays. Ng *et al.* (*42*) recently developed a medium-throughput in vitro system to assess the functional effects of mutations using two growth factor–dependent cell lines: Ba/F3 (a sensitive leukemia cell line frequently used in drug screening) and MCF10A (a breast epithelial cell line). Cell viability data of mutations in these two cell lines were used to generate consensus functional annotation to interpret the functional impacts of somatic mutations. The mutations were considered oncogenic when cell viability was labeled as activating and neutral when labeled as neutral from the consensus functional annotation. Although we recognize that results from in vitro assays cannot be directly translated to the clinic, they offer a comprehensive and unbiased way of assessing the potential functional significance of somatic variants. From the published study, we retrieved 717 missense mutations (253 oncogenic and 464 neutral) in 44 genes. For 253 oncogenic variants, the CancerVar rule-based method predicted 217 (TP = 85.7%) and the deep learning–based



Cancer type	Gene	Protein.change	B/LB	VUS	PCS	SCS	CancerVar score	Prediction
Non–small cell lung carcinoma	*AKT1*	R48C	0	1	0	0		5 VUS
Lung adenocarcinoma	*ARAF*	S217C	0	0.5	0.5	0		9 LP
Poorly differentiated carcinoma	*ATM*	R337C	0.05	0.7	0.25	0		9 LP
Gastroesophageal junction	*B2M*	Q28*	0	0.35	0.65	0		7 VUS
Metastatic melanoma	*BRAF*	D594G	0	0.2	0.8	0		12 P
Metastatic clear cell renal cell carcinoma	*BRAF*	G466R	0	0.35	0.65	0		9 LP
Breast, lobular	*BRAF*	Q609L	0	0.85	0.15	0		7 VUS
Colorectal carcinoma	*BRAF*	V600E	0	0	0.15	0.85		14 P
Breast: invasive ductal carcinoma	*BRCA2*	K3326*	0.4	0.2	0.05	0.35		7 VUS
Neuroblastoma	*BRCA2*	E1308*	0.05	0.05	0.65	0.25		7 VUS
Breast, lobular	*CDH1*	W156*	0	0.15	0.4	0.45		8 LP
Meningioma	*CTNNB1*	R582P	0	1	0	0		7 VUS
Pancreas, neuroendocrine	*DAXX*	A310V	0.05	0.89	0.05	0		2 B
Non–small cell lung carcinoma	*EGFR*	G796D	0	0.2	0.6	0.2		12 P
Metastatic lung adenocarcinoma	*EGFR*	S768I	0	0	0.32	0.68		14 P
Metastatic lung adenocarcinoma	*EGFR*	V774M	0	0.56	0.33	0.11		12 P
Breast: invasive ductal carcinoma	*ERBB2*	S310F	0	0.05	0.7	0.25		9 LP
Melanoma	*ERBB4*	Q941*	0	0.8	0.2	0		6 VUS
Non–small cell lung carcinoma	*FBXW7*	S398F	0.05	0.8	0.15	0		8 LP
Non–small cell lung carcinoma	*GNAS*	G801R	0.06	0.94	0	0		7 VUS
Melanoma	*IDH1*	R132C	0	0.05	0.8	0.15		11 P
Colon adenocarcinoma	*KRAS*	K117N	0	0	0.1	0.9		12 P
Prostate adenocarcinoma	*MUTYH*	G396D	0.1	0.25	0.45	0.2		6 VUS
Renal cell carcinoma	*NF2*	W41X	0	0.25	0.7	0.05		6 VUS
Lung, squamous cell carcinoma	*NFE2L2*	I28T	0	0.65	0.35	0		5 VUS
Lung adenocarcinoma	*PIK3CA*	E542K	0	0.05	0.8	0.15		10 LP
Metastatic prostate cancer	*PIK3CA*	Q546P	0	0.1	0.9	0		10 LP
Metastatic melanoma	*PTPN11*	E69K	0	0.2	0.7	0.1		9 LP
Colon adenocarcinoma	*SMAD4*	R361C	0	0.05	0.8	0.15		12 P
Glioblastoma multiforme	*STK11*	G251C	0.05	0.8	0.15	0		9 LP
Lung adenocarcinoma	*TP53*	T211A	0	0.5	0.5	0		10 LP
Neuroendocrine carcinoma	*CDKN2A*	V106fs	0.05	0.15	0.75	0.05		5 VUS
Breast: invasive ductal carcinoma	*GATA3*	P408fs	0.05	0.35	0.6	0		5 VUS
Medulloblastoma	*PTCH1*	V1164Ffs*27	0	0.15	0.65	0.2		5 VUS
Prostate adenocarcinoma	*BRCA2*	T3085Nfs*26	0	0	0.3	0.7		7 VUS
Anaplastic astrocytoma	*ATRX*	T1582fs	0	0.05	0.45	0.5		4 VUS
Metastatic carcinoma	*ARID1A*	Q1519Pfs*13	0.05	0.2	0.75	0		6 VUS
Colorectal adenocarcinoma	*APC*	E1309Dfs*4	0	0.05	0.5	0.45		7 VUS
Lung adenocarcinoma	*EGFR*	K745insVPVAIK	0	0	0.2	0.8		9 LP
Lung adenocarcinoma	*ERBB2*	Y772_A775dup	0	0.05	0.55	0.4		5 VUS
Gastrointestinal stromal tumor	*KIT*	M552_K558del	0	0.1	0.05	0.85		8 LP
Colorectal adenocarcinoma	*SMAD4*	T349delinsAVA	0	0.7	0.3	0		6 VUS
Metastatic adenocarcinoma (pancreaticobiliary primary)	*TP53*	G279_R282del	0	0.65	0.35	0		5 VUS

Fig. 3. Comparison of the interpretation of 43 variants between 20 pathologists and CancerVar. The heatmap shows the ratio of 20 pathologists voting for the four tiers: tier I, strong clinical significance (SCS); tier II, potential clinical significance (PCS); tier III, variant of uncertain clinical significance (VUS); and tier IV, benign/likely benign (B/LB). The last two columns are CancerVar-predicted scores and classifications. CancerVar showed an 81% (17 of 21) agreement rate with pathologists' majority voting for tier I/II and a 60.5% (26 of 43) agreement rate for all tiers. This agreement rate is comparable to the 58% agreement rate among the 20 pathologists, but CancerVar can automate the interpretation process. P, Pathogenic/strong clinical significance; LP:Likely Pathogenic/potential clinical significance; B:(Likely)Benign.

method predicted 208 (82.2%) as having strong/potential clinical significance; OncoKB predicted 204 (TP = 80.6%) variants as oncogenic, likely, or predicted oncogenic (Fig. 4D, table S5, and fig. S3). Therefore, CancerVar rule-based and deep learning–based methods perform slightly better than OncoKB.

Performance of OPAI in predicting oncogenic variants

One unique feature of CancerVar is the inclusion of the OPAI approach, which allows prediction of oncogenic variants that have never been reported in public databases. Unlike conventional models trained on labeled data, OPAI can learn the hidden distribution of unlabeled mutations collected from clinical data. In the current study, training the OPAI models required ~100 hours with 1000 epochs on an Nvidia Tesla M40 GPU. Next, we assessed the ability to predict novel oncogenic variants. OPAI was compared with five other machine learning algorithms, including gradient boosting tree, SVM, AdaBoost, RF, and XGBoost, using the Python package scikit-learn (*43*). We further compared performance with the other five cancer-specific driver mutation analysis tools, including CanDrA, CHASM, CTAT-cancer, and MutationAssessor, using the area under the curve (AUC) score from receiver operating characteristic (ROC) plots and the true-negative rate (or specificity) as performance measurements. To evaluate imbalanced classes, we also calculated the area under precision-recall curve (AUPRC) to assess the predictive performance of various models. On the basis of an independent testing set of 6226 somatic variants, Fig. 5 (A and B) shows that when using the AUC measure, the OPAI method in CancerVar (AUC-ROC = 0.854) performed the best compared to cancer-specific driver predicting methods and any individual functional prediction tool. Figure 5 (C and D) highlights that OPAI performed the best (AUPRC = 0.686) among these methods. Because of the imbalanced classes, that is, the fraction of positives (oncogenic) was less than that of negatives (benign), it was expected that OPAI's AUPRC value would be lower than the AUC-ROC value.

Table 2. Summary of CancerVar prediction on OncoKB mutations.

OncoKB	Model-based CancerVar (OPAI)		Rule-based CancerVar		
	Oncogenic	Neutral	I/II (S/P*)	III (VUS*)	IV (Benign)
Oncogenic/likely oncogenic	2319	263	1839	690	53
Neutral/likely neutral	348	239	281	279	27
Inconclusive	152	134	132	150	4
Total	2819	636	2252	1119	84

*S, strong clinical significance (tier I); P, potential clinical significance (tier II); VUS, variant of unknown clinical significance (tier III).

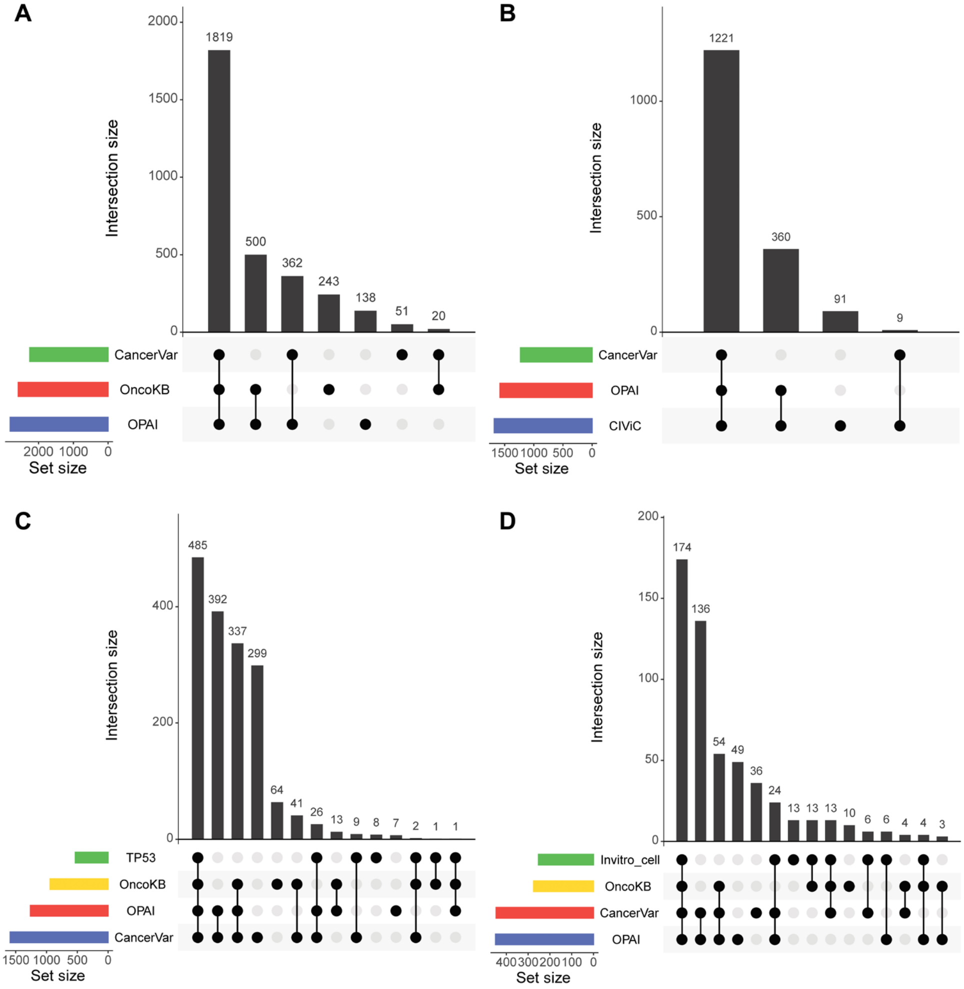


Fig. 4. UpSet plot highlighting the intersection of multiple methods with oncogenic prediction from different datasets. (**A**) Mutations were taken from the OncoKB dataset. (**B**) Mutations were taken from CIViC. (**C**) Mutations were taken from the IARC TP53 transactivation dataset. (**D**) Mutations were taken from in vitro cell viability by Ng *et al.* (*42*).

Evaluation of FDA-approved or -recognized cancer biomarkers for therapeutic, diagnostic, and prognostic purposes

To further evaluate the performance and reliability of CancerVar in clinical applications, we collected 22 cancer biomarkers specified by the U.S. Food and Drug Administration (FDA) and interpreted these biomarkers and predicted their oncogenicity. Among them, 9 were classified as tier I strong clinical significance when using only the evidence-based method; the other 13 biomarkers were classified as tier II potential clinical significance. For the OPAI model, most of the biomarkers (19 of 22) were predicted to have scores ≥0.95, suggesting a very high probability of having clinical significance. The interpretation of these biomarkers is shown in Table 4.

Table 3. Summary of CancerVar prediction of CIViC mutations.

CIViC	Model-based CancerVar (OPAI)		Rule-based CancerVar			
	Oncogenic	Neutral	I (Strong)	II (Potential)	III (VUS)	IV (Benign)
A: Validated	17	2	2	7	0	10
B: Clinical	259	40	111	109	61	18
C: Case study	792	30	178	439	198	7
D: Preclinical	466	22	91	277	119	1
E: Inferential	47	6	5	11	33	4
Total	1581	100	387	843	411	40

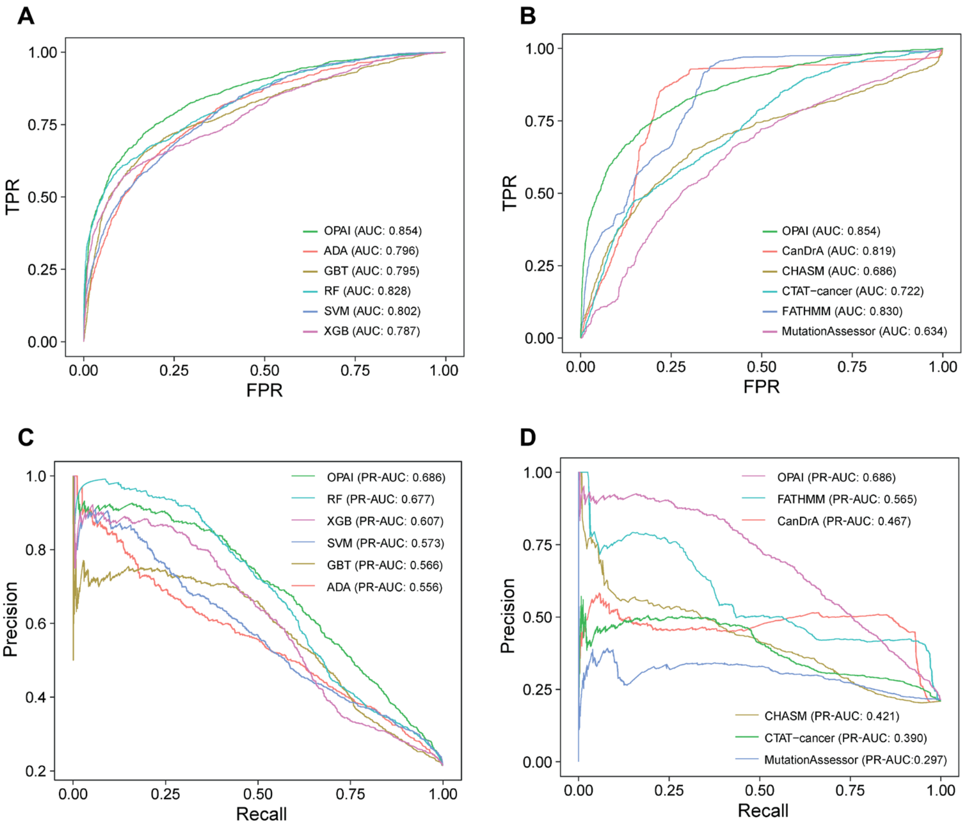


Fig. 5. Performance comparisons. (**A** and **B**) Receiver operating characteristic (ROC) curves for performance comparison between OPAI and five other machine learning algorithms, including gradient boosting tree (GBT), support vector machine (SVM), AdaBoost (ADA), random forest (RF), and XGBoost (XGB), and five other in silico predictive tools using 6226 somatic mutations as the testing set. (**C** and **D**) Area under the precision-recall curve (AUPRC) comparison between OPAI and five other machine learning tools and in silico predictive tools. OPAI outperformed any individual tool in the prediction of somatic driver mutations in cancer. TPR, true-positive rate; FPR, false-positive rate.

Oncogenicity prediction of edgetic mutations

Gene products can interact with each other in PPI networks or cellular signaling networks. Network rewiring is crucial for understanding the complex genotype-phenotype relationships in cancer (*37*, *39*). Some cancer mutations have been found to broadly or specifically affect these interactome networks or pathways via perturbation (*38*). These mutations in cancer are called edgetic mutations, and they may cause loss of molecular function, representing novel oncogenic candidates (*44*).

Li *et al.* (*45*) built an e-MutPath (edgetic mutation–mediated pathway perturbations) database to report cancer somatic edgetic mutations based on genome-wide somatic mutation profiles with gene expression and PPI networks. We downloaded 2541 edgetic mutations ($P < 0.05$) related to non–small cell lung cancer from the e-MutPath database and predicted their oncogenicity using CancerVar. Among these 2541 variants, CancerVar predicted 987 (39%) variants to be oncogenic using clinical features only; it predicted 2047 (81%) variants to be oncogenic using ensemble features, including both

Table 4. FDA-approved or -recognized biomarkers (therapeutic, diagnostic, and prognostic) from the rule-based model and the deep learning model (OPAI) in CancerVar.

Gene	Alternation	Cancers	Levels	Rule-based CancerVar	OPAI
ABL1	T315I	B-lymphoblastic leukemia/lymphoma/chronic myelogenous leukemia	Therapeutic (ponatinib)	Tier II (score 8)	0.93
AKT1	E17K	Breast cancer/ovarian cancer/endometrial cancer	Therapeutic (AZD5363)	Tier II (score 10)	0.98
BRAF	V600E	Melanoma/non–small cell lung cancer	Therapeutic (dabrafenib + trametinib; vemurafenib)	Tier I (score 11)	0.98
EGFR	T790M	Non–small cell lung cancer	Therapeutic (osimertinib)	Tier II (score 9)	0.98
EGFR	L861Q	Non–small cell lung cancer	Therapeutic (afatinib)	Tier II (score 10)	0.99
EGFR	S768I	Non–small cell lung cancer	Therapeutic (afatinib)	Tier I (score 11)	0.99
EZH2	Y646F	Follicular lymphoma	Therapeutic (tazemetostat)	Tier I (score 11)	0.99
	Y646H			Tier I (score 11)	0.99
	Y646N			Tier I (score 11)	0.99
	Y646S			Tier I (score 11)	0.99
FGFR3	R248C	Bladder cancer	Therapeutic (erdafitinib)	Tier II (score 9)	0.99
	S249C			Tier II (score 9)	0.99
	Y373C			Tier II (score 10)	0.99
JAK2	V617F	Primary myelofibrosis	Prognostic	Tier I (score 11)	0.87
KIT	A829P	Gastrointestinal stromal tumor	Therapeutic (imatinib, regorafenib, ripretinib, and sunitinib)	Tier II (score 9)	0.83
	T670I			Tier II (score 9)	0.98
	V654A			Tier I (score 11)	0.99
	Y823D			Tier II (score 10)	0.99
	D816V	Systemic mastocytosis	Diagnostic	Tier II (score 9)	0.99
KRAS	G12C	Non–small cell lung cancer	Therapeutic (AMG-510)	Tier I (score 11)	0.99
PDGFRA	D842V	Gastrointestinal stromal tumor	Therapeutic (avapritinib)	Tier II (score 9)	0.97
	D842Y			Tier II (score 10)	0.99

clinical and functional evidence. This result is consistent with the expectation that the functional impact of edgetic variants contributes more weight than clinical features in the prediction model for oncogenicity.

Enrichment analysis for CancerVar-predicted driver mutations in well-known ONCs and TSGs

We also performed enrichment analysis on CancerVar-predicted driver mutations in 10 well-known tumor suppressor genes (TSGs) and 10 oncogenes (ONCs) reported in Bayley's driver gene list (*46*). In total, ~223,000 exome mutations (147,141 in TSGs and 76,362 in ONCs) were observed. CancerVar predicted 1641 driver variants in TSGs (approximately 1641 of 147,141 = 1.1%) and 875 in ONCs (approximately 875 of 76,362 = 1.1%). According to the theoretical estimation of Bozic *et al.* (*47*), on average, ~114 driver mutations are expected in one TSG, and ~14 driver mutations are expected in one ONC. CancerVar predicted ~164 (1641/10) driver mutations in each TSG and ~88 (875/10) in each ONC, greater than the theoretical expectations. This observation may be partially attributed to the fact that although ONCs tend to have hotspot (recurring) somatic mutations as gain-of-function mutations, this is not taken into account in the prediction model. The numbers and frequencies of predicted oncogenic variants are shown in table S6.

For comparison, we used CancerVar to predict oncogenic variants in 10 randomly selected nonessential genes ("human knockout" genes) from two studies [Narasimhan *et al.* (*48*) and Saleheen *et al.* (*49*)] with 250 overlapping genes. We scanned 86,094 possible exome mutations in these 10 genes, including *AKR1E2*, *BROX*, *BTN3A3*, *EFCAB13*, *KLHL25*, *LRRC69*, *MGST1*, *MROH2A*, *TTLL2*, and *ZSCAN16*. CancerVar did not detect any oncogenic variants (0%) in these genes but predicted 84,445 (98%) as benign and 1649 (2%) as VUSs. These results further suggest that oncogenic variants

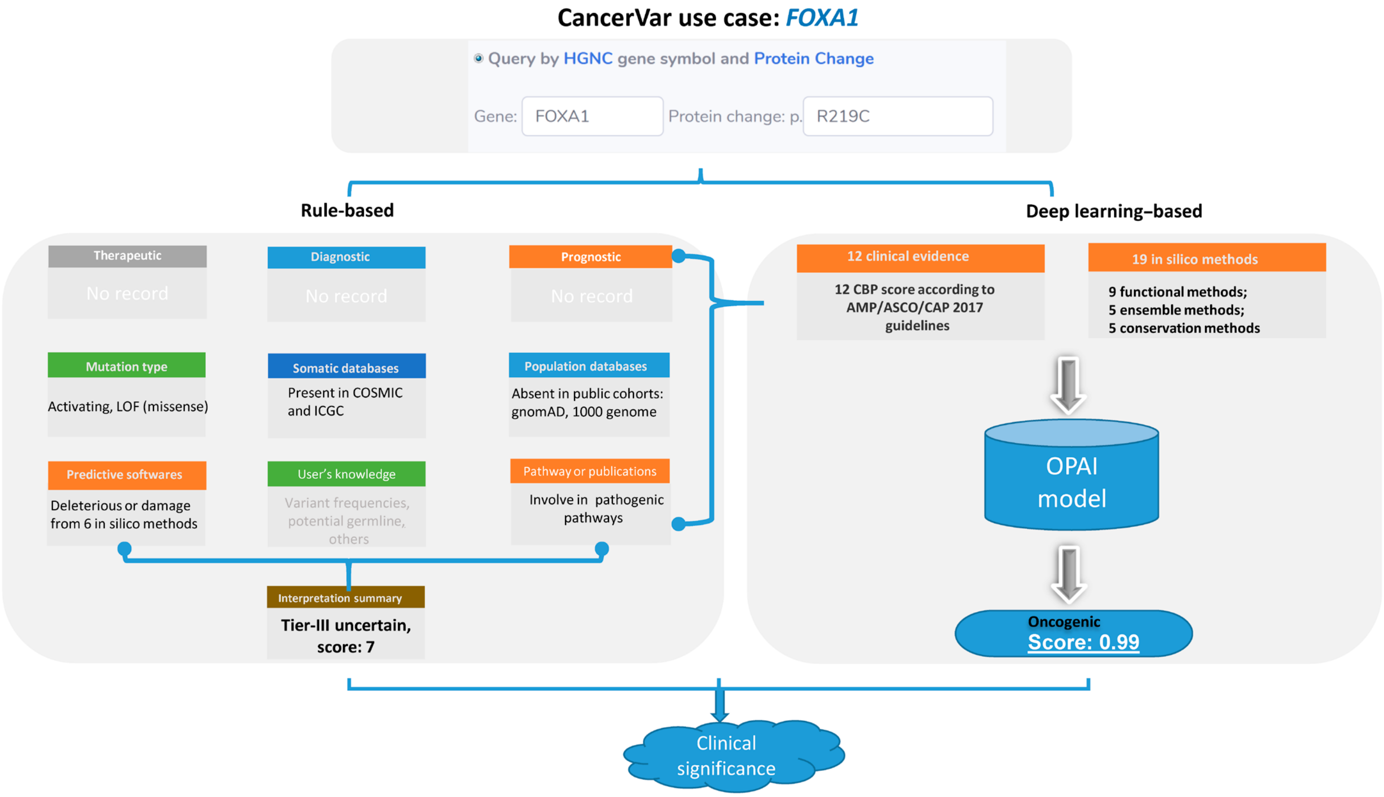

CancerVar use case: FOXA1
Query by HGNC gene symbol and Protein Change
Gene: FOXA1
Protein change: p. R219C
Rule-based
Deep learning–based
Therapeutic
No record
Diagnostic
No record
Prognostic
No record
Mutation type
Activating, LOF (missense)
Somatic databases
Present in COSMIC and ICGC
Population databases
Absent in public cohorts: gnomAD, 1000 genome
Predictive softwares
Deleterious or damage from 6 in silico methods
User's knowledge
Variant frequencies, potential germline, others
Pathway or publications
Involve in pathogenic pathways
Interpretation summary
Tier-III uncertain, score: 7
12 clinical evidence
12 CBP score according to AMP/ASCO/CAP 2017 guidelines
19 in silico methods
9 functional methods; 5 ensemble methods; 5 conservation methods
OPAI model
Oncogenic
Score: 0.99
Clinical significance

Fig. 6. A use case of using rule-based and deep learning-based models in CancerVar for interpretation of *FOXA1* variants. – We queried the *FOXA1* mutation R219C in prostate cancer. The rule-based prediction of this variant was tier III (uncertain significance), with a score of 7, which is very close to tier II. However, the OPAI model predicted this variant to be oncogenic, with a score of 0.99. On the basis of a manual review of the results, we suggest that this variant has clinical significance.

predicted by CancerVar are enriched (1.1%) in putative ONCs and TSGs compared to the lack of findings (0%) for the human knockout genes.

Use case: Comprehensive interpretation of *FOXA1* somatic mutations in prostate cancer

In this use case, we show the clinical interpretation of the *FOXA1*'s mutation in prostate cancer (Fig. 6) from rule-based and deep learning–based models in CancerVar. Prostate cancer is the most commonly diagnosed cancer in men worldwide (*50*). The FOXA1 protein (forkhead Box A1, previously known as *HNF3a*) is essential for normal development of the prostate (*51*). *FOXA1* somatic mutations are frequently observed in prostate cancer (*52*) and associated with poor outcomes. In 2019, two studies demonstrated that *FOXA1* acts as an ONC in prostate cancer (*53*, *54*), whereby the hotspot mutation at R219 (R219S and R219C) drives a proluminal phenotype in prostate cancer exclusive to other fusions or mutations (*53*, *54*). We interpreted these two mutations, but, here, we only illustrate the clinical interpretation of R219C, as that of R219S is very similar. We searched for the missense mutation R219C using the protein change and gene name "*FOXA1*" in the CancerVar web server. CancerVar did not find any therapeutic, diagnostic, or prognostic evidence for this mutation. Because this mutation has recently been incorporated into somatic databases, including COSMIC (ID: COSM3738526) and international cancer genome consortium (ICGC) (ID: MU67448716), CBP_9 was applied as moderate evidence. Recently, two publications reported its biological functions in prostate cancer, and CBP_12 was applied. According to CBP_7, this mutation is absent or has an extremely low minor allele frequency in the public allele frequency database. As all seven in silico methods predicted this mutation as (likely) pathogenic, CBP_10 was applied. According to the AMP/ASCO/CAP/CGC guidelines, this variant falls into the class "tier III uncertain significance," with a score of 7, but very close to the class "tier II potential." On the other hand, on the basis of the OPAI model, the score on this variant is 0.99, suggesting that it is very likely oncogenic. This use case demonstrates that a semiautomated interpretation approach can greatly improve prediction accuracy for each variant given existing knowledge and domain expertise. In addition, this user case shows that a model-based approach involving machine intelligence can be applied as additional evidence to support rule-based methods.

DISCUSSION

Clinical interpretation of cancer somatic variants is important for clinicians and researchers working in the field of precision oncology, especially given the transition from panel sequencing to whole-exome/genome sequencing in cancer genomics studies. To build a standardized, rapid, and user-friendly interpretation tool, we developed command-line software tools together with a web server to assess the clinical impacts of somatic variants using the AMP/ASCO/CAP 2017 guidelines. CancerVar is an enhanced version of the cancer variant knowledgebase incorporated from our previously developed tools for variant annotations and prioritizations, including

ANNOVAR (*55*), InterVar (*56*), VIC (*9*), and iCAGES (*57*), as well as assembling existing variant annotation databases such as CIViC (*4*), CKB (*6*), and OncoKB (*1*). We stress here that CancerVar will not replace human acumen in clinical interpretation but rather generate evidence to help human reviewers by providing standardized, reproducible, and precise output for interpreting somatic variants.

With CancerVar, we did not reconcile the well-known "conflicting interpretation" issues across knowledgebases; instead, we documented and harmonized all types of clinical evidence (i.e., drug information and publications) for both hotspot and non-hotspot mutations in detail to allow users to make informed clinical decisions based on their own domain knowledge and expertise. Compared to existing knowledgebases such as OncoKB, CIViC, and metaKB, CancerVar provides an improved platform in four areas: (i) comprehensive, evidence-based annotations with rigorous quality control for ~13 million somatic variants, which is not limited to the small number of known hotspot mutations; (ii) a flexible scoring system allowing users to fine-tune the importance of clinical evidence criteria according to their own domain knowledge; (iii) improved prioritization for cancer driver mutations using the novel semisupervised deep learning method OPAI; and (iv) automatically summarized interpretation text such that users do not need to compile evidence from multiple knowledgebases manually. We expect CancerVar to become a useful web service for the interpretation of somatic variants in clinical cancer research.

We also need to acknowledge several limitations in CancerVar. First, the scoring weight system is not very robust and may be considered ad hoc by some users. We note that existing clinical guidelines did not provide recommendations for weighting different evidence types, and therefore, we treat all weights as equal by default. Nevertheless, with increasing amounts of clinical knowledge regarding somatic mutations, we expect that we may build a weighted model in the future to enhance prediction accuracy. Second, a small number of CNAs (similar to hotspot mutations) have emerged as important biomarkers for disease characterization and therapeutic decision-making; however, there is a lack of a specific database for clinically actionable somatic CNAs. Although AMP/ASCO/CAP recently published a CNA guideline, CNAs are very heterogeneous in size, and their significance is much harder to score in practice. Therefore, in the future, we will design and implement the scoring system for CNAs based on the platform used to discover CNAs (which determines the resolution of calls), the reliability of the CNA calls, the genes covered by the CNAs, and additional cancer type–specific information from existing databases (given that different cancer types have different CNA profiles). Third, CancerVar currently cannot interpret inversions or gene fusions and cannot interpret gene expression alterations, although these genomic alterations may play important roles in cancer development/progression. Before a specific guideline for these types of mutations becomes available, we suggest that users treat them as CNAs (gene inversions/fusions as deletions and gene expression down-regulation or up-regulation as deletions or duplications). Last, CancerVar web servers cannot process indels because there are many more possible indels than SNVs in cancer to be precomputed en masse. However, users have the flexibility to use the command-line version to predict the clinical significance of indels, including those in genes that are not included in the cancer gene census.

Overall, accurate interpretation of clinical significance depends greatly on evidence harmonization, which should be precisely derived and standardized from multiple databases and annotations. Compared to existing knowledgebases that document a limited number of hotspot mutations, CancerVar provides polished, comprehensive, and semiautomated clinical interpretations for somatic variants with clinical evidence, and it greatly facilitates human reviewers' drafts of clinical reports for panel sequencing, exome sequencing, or whole-genome sequencing in cancer. Although some commercial software tools also use the AMP/ASCO/CAP rules to standardize variant interpretation, they require a high license fee that many academic researchers may not afford. In addition to interpretation based on the AMP/ASCO/CAP rules specified by consensus from human experts, the CancerVar deep learning–based approach jointly models both rule-based clinical features and functional prediction features to support oncogenic predictions for novel unreported mutations. We believe that CancerVar allows for comprehensive clinical interpretations and prioritizations for both hotspot and non-hotspot variants, which may facilitate the implementation of precision oncology.

In summary, CancerVar is both a web server and a command-line software tool that provides polished and semiautomated clinical interpretations for somatic variants in cancer. Moreover, CancerVar facilitates drafting clinical reports semiautomatically for panel sequencing, exome sequencing, or genome sequencing in cancer. We expect to continuously improve CancerVar and incorporate new functionalities in the future, similar to the wInterVar server (*56*) and wANNOVAR server (*58*).

MATERIALS AND METHODS

Overview of mapping of clinical evidence to the AMP/ASCO/CAP 2017 guidelines

According to the AMP/ASCO/CAP 2017 guidelines, there are a total of 12 types of clinical-based evidence to predict the clinical significance of somatic variants, including therapies, mutation types, variant allele fraction [mosaic variant frequency (likely somatic) and non-mosaic variant frequency (potential germline)], population databases, germline databases, somatic databases, predictive results of different computational algorithms, pathway involvement, and publications (*8*, *9*). As shown in Fig. 1, CancerVar contains all the above 12 pieces of evidence, among which 10 are automatically generated; the other two, including the variant allele fraction and potential germline, require user input [possibly included in variant call format (VCF) files] for manual adjustment.

Cancer variant collection and preprocessing

The cancer gene census list and potential driver gene list are essential for somatic variant annotation. We curated a list of 1911 cancer census or driver genes with 13 million exonic variants from seven existing cancer knowledgebases, including COSMIC, CIViC, and OncoKB, and two datasets collected from the literature on driver gene predictions (table S1). For each exon position in these 1911 genes, we generated all three possible nucleotide changes. Unlike other knowledgebases, which only compile variants reported or documented previously, we precomputed annotations for all possible variants for CancerVar interpretation. We compiled clinical evidence based on the AMP/ASCO/CAP 2017 guidelines, which makes the variant searching in the CancerVar web server very fast. In CancerVar, we document all types of clinical evidence, such as in silico prediction, drug information, and publications, in detail to help users adjust criteria and make their own decisions using prior knowledge.

The evidence-based scoring method to prioritize the clinical significance of somatic variants

CancerVar evaluates each set of evidence and then scores each clinical-based prediction (CBP). The variant evidence receives 2 points for strong clinical significance or oncogenic, 1 point for supporting clinical significance or oncogenic, 0 for no support, and −1 for benign or neutral. The CancerVar score is the sum of all evidence scores. The complete scoring system for each CBP can be found in table S2. Let CBP[i] be the ith evidence score, and let weight [i] be the score for the ith evidence; the CancerVar score is calculated using Eq. 1. The weight is 1 by default, but users can adjust it based on its importance from prior knowledge. On the basis of the score range in Eq. 2, we classify each variant into one of the four tiers: strong clinical significance, potential clinical significance, variants of unknown clinical significance (VUS), and benign/likely benign (neutral)

$$\text{CancerVar score (CS)} = \textstyle\sum_{i=1}^{12} \text{Weight}\,[\,i\,] * \text{CBP}\,[\,i\,] \tag{1}$$

$$\text{Interpretation} = \begin{cases} \text{Strong clinical significance CS} \ \geq \ 11 \\ \text{Potential clinical sinigicance } 8 \ \leq \ \text{CS} \ \leq \ 10 \\ \text{VUS } 3 \ \leq \ \text{CS} \ < \ 8 \\ \text{(Likely) Benign CS} \ \leq \ 2 \end{cases} \tag{2}$$

The deep learning–based scoring method to predict oncogenic variants

We developed a scoring method for OPAI using 12 clinical evidence prediction scores and 23 precomputed scores predicted by other computational tools, such as SIFT (*23*), DANN (*26*), and SiPhy (*59*), by improving a semisupervised generative adversarial network model that we previously built. As the score ranges are diverse among the predictive tools, we used their categorical outputs as prediction features. Some variants have missing values on certain features, and we thus excluded variants with more than two missing features. After filtering, 12.9 million variants were used for downstream analysis.

OPAI uses one type of deep learning architecture, the semisupervised generative adversarial network method, to predict the probability of oncogenic variants. As depicted in Fig. 2, this architecture consists of two parts: the generator (G) and the discriminator (D). The generator generates synthetic samples (fake) by random noise from the normal distribution, and the discriminator differentiates realistic samples and synthetic data. In our model, the input data consist of labeled samples, unlabeled samples, and random noise from a normal distribution as the synthetic data. Then, the discriminator/classifier classifies the sample into three classes: (i) neutral, (ii) nonneutral (oncogenic mutations), and (iii) fake synthetic data, in which the unlabeled real samples can be identified as 1 or 2 and the synthetic samples are 3. Therefore, loss function L can be written as two parts

$$L \ = \ L_{\text{supervised}} + L_{\text{unsupervised}}$$

$$L_{\text{supervised}} \ = \ -\mathbb{E}_{X,y\sim p_{\text{data}}(X,y)} \log\left[p_{\text{model}}(y \ = \ l_i \,|\, X, y \ < \ 3)\right]$$

$$\begin{aligned} L_{\text{unsupervised}} \ = \ -\big\{ & \mathbb{E}_{X\sim p_{\text{data}}(X)} \log\left[1 - p_{\text{model}}\left(y \ = \ 3 \,|\, X\right)\right] + \\ & \mathbb{E}_{X\sim p_G} \log\left[p_{\text{model}}\left(y \ = \ 3 \,|\, X\right)\right]\big\} \end{aligned} \tag{3}$$

where p_{data} is the underlying distribution of real samples and p_G is the distribution of the output from the generator. For the loss of the generator, we used feature matching as our loss function: $\| \mathbb{E}_{x\sim p_{\text{data}}} D(x) - \mathbb{E}_{z\sim p_G} D(G(z)) \|$.

OPAI training and testing

Labeled data were obtained from expert curation on diagnostic reports of patients with cancer in our in-house database; we used 4000 variants (1000 positive) as the training set and 1234 variants (669 positive) as the validation set. We tested our model on 6226 variants (1335 positive), which were compiled from a literature review. For unlabeled data, we randomly selected 60,000 variants from 12.9 million samples with non-missing features and repeated this process multiple times.

For synthetic samples, the generator produces random noise from a standard normal distribution in each batch step and outputs the synthetic samples. In each minibatch, the model calculates 2000 labeled samples, 10,000 unlabeled samples, and 10,000 synthetic samples from the generator. The discriminator/classifier is trained by calculating the loss from supervised learning and unsupervised training separately. Then, the generator is trained by minimizing the feature matching in each batch.

To improve analysis interpretability and evaluate the feature contribution in our study, RF analysis was used to evaluate the importance of 35 features (23 in silico functional features and 12 clinical evidence scores) in the above expert manually labeled variant database. The importance score of each feature was calculated and ranked, as provided in the Supplementary Materials (fig. S4).

Pan-cancer benchmarks from public datasets

Complementary and comprehensive benchmark datasets are needed for systematic evaluation of the performance of CancerVar. In the current study, we used several different benchmark datasets with references: (i) a multi-institutional evaluation study with 51 variants from Sirohi *et al.* (*10*); (ii) a literature annotation database from OncoKB (*1*) and CIViC (*4*); (iii) *TP53* mutations and their target transcription activity from the IARC database (*41*); and (iv) functional annotation based on in vitro cell viability assays from the study of Ng *et al.* (*42*).

High-quality expert-labeled variants from in-house clinical reports

In addition to datasets from public resources, we also have an in-house dataset with 7967 somatic mutations from deidentified patients at the Children's Hospital of Philadelphia (CHOP). Each variant has been manually annotated and classified by human experts in the diagnostic labs. Using the AMP/ASCO/CAP guidelines, the four-tier classification assignment for each somatic mutation was agreed upon by at least two cancer experts from the Division of Genomic Diagnostics Laboratory at CHOP. Furthermore, to train the deep learning model, we used variants from strong clinical significance (tier I) and potential clinical significance (tier II) categories as positive samples and used benign/likely benign variants (tier IV) as negative samples. There were 5234 variants, 1668 positive samples and 3566 negative samples, for training and validation.

CancerVar and OPAI software accessibility

Users can access CancerVar and OPAI in three ways, including a web server that is free and open to all users without login requirements (https://cancervar.wglab.org), a command-line software written in Python that is freely available from GitHub (https://github.com/wglab/CancerVar) for noncommercial users, and a RESTful API service to facilitate other web developers in accessing our precomputed evidence and OPAI scores for ~13 million coding variants. Users can use the command-line version to predict the clinical significance of indels, including indels in genes not included in the cancer gene list precomputed by CancerVar.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abj1624

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. D. Chakravarty, J. Gao, S. M. Phillips, R. Kundra, H. Zhang, J. Wang, J. E. Rudolph, R. Yaeger, T. Soumerai, M. H. Nissan, M. T. Chang, S. Chandarlapaty, T. A. Traina, P. K. Paik, A. L. Ho, F. M. Hantash, A. Grupe, S. S. Baxi, M. K. Callahan, A. Snyder, P. Chi, D. Danila, M. Gounder, J. J. Harding, M. D. Hellmann, G. Iyer, Y. Janjigian, T. Kaley, D. A. Levine, M. Lowery, A. Omuro, M. A. Postow, D. Rathkopf, A. N. Shoushtari, N. Shukla, M. Voss, E. Paraiso, A. Zehir, M. F. Berger, B. S. Taylor, L. B. Saltz, G. J. Riely, M. Ladanyi, D. M. Hyman, J. Baselga, P. Sabbatini, D. B. Solit, N. Schultz, OncoKB: A precision oncology knowledge base. *JCO Precis. Oncol.* **2017**, (2017).
2. M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, P. K.-S. Ng, K. J. Jeong, S. Cao, Z. Wang, J. Gao, Q. Gao, F. Wang, E. M. Liu, L. Mularoni, C. Rubio-Perez, N. Nagarajan, I. Cortes-Ciriano, D. C. Zhou, W. W. Liang, J. M. Hess, V. D. Yellapantula, D. Tamborero, A. Gonzalez-Perez, C. Suphavilai, J. Y. Ko, E. Khurana, P. J. Park, E. M. Van Allen, H. Liang; MC3 Working Group; Cancer Genome Atlas Research Network, M. S. Lawrence, M. S. Lawrence, A. Godzik, N. Lopez-Bigas, J. Stuart, D. Wheeler, G. Getz, K. Chen, A. J. Lazar, G. B. Mills, R. Karchin, L. Ding, Comprehensive characterization of cancer driver genes and mutations. *Cell* **174**, 1034–1035 (2018).
3. C. M. Micheel, S. M. Sweeney, M. L. LeNoue-Newton, F. Andre, P. L. Bedard, J. Guinney, G. A. Meijer, B. J. Rollins, C. L. Sawyers, N. Schultz, K. R. M. Shaw, V. E. Velculescu, M. A. Levy; AACR Project GENIE Consortium, American Association for Cancer Research Project Genomics Evidence Neoplasia Information Exchange: From inception to first data release and beyond-lessons learned and member institutions' perspectives. *JCO Clin. Cancer Inform.* **2**, 1–14 (2018).
4. M. Griffith, N. C. Spies, K. Krysiak, J. F. McMichael, A. C. Coffman, A. M. Danos, B. J. Ainscough, C. A. Ramirez, D. T. Rieke, L. Kujan, E. K. Barnell, A. H. Wagner, Z. L. Skidmore, A. Wollam, C. J. Liu, M. R. Jones, R. L. Bilski, R. Lesurf, Y. Y. Feng, N. M. Shah, M. Bonakdar, L. Trani, M. Matlock, A. Ramu, K. M. Campbell, G. C. Spies, A. P. Graubert, K. Gangavarapu, J. M. Eldred, D. E. Larson, J. R. Walker, B. M. Good, C. Wu, A. I. Su, R. Dienstmann, A. A. Margolin, D. Tamborero, N. Lopez-Bigas, S. J. Jones, R. Bose, D. H. Spencer, L. D. Wartman, R. K. Wilson, E. R. Mardis, O. L. Griffith, CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **49**, 170–174 (2017).
5. L. Huang, H. Fernandes, H. Zia, P. Tavassoli, H. Rennert, D. Pisapia, M. Imielinski, A. Sboner, M. A. Rubin, M. Kluk, O. Elemento, The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. *J. Am. Med. Inform. Assoc.* **24**, 513–519 (2017).
6. S. E. Patterson, R. Liu, C. M. Statz, D. Durkin, A. Lakshminarayana, S. M. Mockus, The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Hum. Genomics* **10**, 4 (2016).
7. D. Tamborero, C. Rubio-Perez, J. Deu-Pons, M. P. Schroeder, A. Vivancos, A. Rovira, I. Tusquets, J. Albanell, J. Rodon, J. Tabernero, C. de Torres, R. Dienstmann, A. Gonzalez-Perez, N. Lopez-Bigas, Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
8. M. M. Li, M. Datto, E. J. Duncavage, S. Kulkarni, N. I. Lindeman, S. Roy, A. M. Tsimberidou, C. L. Vnencak-Jones, D. J. Wolff, A. Younes, M. N. Nikiforova, Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J. Mol. Diagn.* **19**, 4–23 (2017).
9. M. M. He, Q. Li, M. Yan, H. Cao, Y. Hu, K. Y. He, K. Cao, M. M. Li, K. Wang, Variant Interpretation for Cancer (VIC): A computational tool for assessing clinical impacts of somatic variants. *Genome Med.* **11**, 53 (2019).
10. D. Sirohi, R. L. Schmidt, D. L. Aisner, A. Behdad, B. L. Betz, N. Brown, J. F. Coleman, C. L. Corless, G. Deftereos, M. D. Ewalt, H. Fernandes, S. J. Hsiao, M. M. Mansukhani, S. S. Murray, N. Niu, L. L. Ritterhouse, C. J. Suarez, L. J. Tafe, J. A. Thorson, J. P. Segal, L. V. Furtado, Multi-institutional evaluation of interrater agreement of variant classification based on the 2017 Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists standards and guidelines for the interpretation and reporting of sequence variants in cancer. *J. Mol. Diagn.* **22**, 284–293 (2020).
11. A. H. Wagner, B. Walsh, G. Mayfield, D. Tamborero, D. Sonkin, K. Krysiak, J. Deu-Pons, R. P. Duren, J. Gao, J. McMurry, A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. *Nat. Genet.* **52**, 448–457 (2020).
12. S. S. Kalia, K. Adelman, S. J. Bale, W. K. Chung, C. Eng, J. P. Evans, G. E. Herman, S. B. Hufnagel, T. E. Klein, B. R. Korf, Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2. 0): A policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).
13. S. M. Harrison, H. L. Rehm, Is 'likely pathogenic' really 90% likely? reclassification data in ClinVar. *Genome Med.* **11**, 72 (2019).
14. D. Bertrand, K. R. Chng, F. G. Sherbaf, A. Kiesel, B. K. Chia, Y. Y. Sia, S. K. Huang, D. S. Hoon, E. T. Liu, A. Hillmer, N. Nagarajan, Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res.* **43**, e44 (2015).
15. H. Carter, S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, R. Karchin, Cancer-specific high-throughput annotation of somatic mutations: Computational prediction of driver missense mutations. *Cancer Res.* **69**, 6660–6667 (2009).
16. Y. Mao, H. Chen, H. Liang, F. Meric-Bernstam, G. B. Mills, K. Chen, CanDrA: Cancer-specific driver missense mutation annotation with optimized features. *PLOS ONE* **8**, e77945 (2013).
17. L. Mularoni, R. Sabarinathan, J. Deu-Pons, A. Gonzalez-Perez, N. Lopez-Bigas, OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
18. E. Porta-Pardo, A. Godzik, e-Driver: A novel method to identify protein regions driving cancer. *Bioinformatics* **30**, 3109–3114 (2014).
19. J. Reimand, G. D. Bader, Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9**, 637 (2013).
20. B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
21. H. A. Shihab, J. Gough, D. N. Cooper, I. N. Day, T. R. Gaunt, Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* **29**, 1504–1510 (2013).
22. D. Tamborero, A. Gonzalez-Perez, N. Lopez-Bigas, OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).
23. P. C. Ng, S. Henikoff, SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
24. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, Chapter 7:Unit7.20, (2013).
25. H. A. Shihab, J. Gough, D. N. Cooper, P. D. Stenson, G. L. Barker, K. J. Edwards, I. N. Day, T. R. Gaunt, Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).
26. D. Quang, Y. Chen, X. Xie, DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
27. S. Shuai, S. Gallinger, L. Stein, Combined burden and functional impact tests for cancer driver discovery using DriverPower. *Nat. Commun.* **11**, 1–12 (2020).
28. J. E. Posey, J. A. Rosenfeld, R. A. James, M. Bainbridge, Z. Niu, X. Wang, S. Dhar, W. Wiszniewski, Z. H. Akdemir, T. Gambin, F. Xia, R. E. Person, M. Walkiewicz, C. A. Shaw, V. R. Sutton, A. L. Beaudet, D. Muzny, C. M. Eng, Y. Yang, R. A. Gibbs, J. R. Lupski, E. Boerwinkle, S. E. Plon, Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genet. Med.* **18**, 678–685 (2016).
29. Y. Yang, D. M. Muzny, J. G. Reid, M. N. Bainbridge, A. Willis, P. A. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu, M. Hardison, R. Person, M. R. Bekheirnia, M. S. Leduc, A. Kirby, P. Pham, J. Scull, M. Wang, Y. Ding, S. E. Plon, J. R. Lupski, A. L. Beaudet, R. A. Gibbs, C. M. Eng, Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
30. K. Retterer, J. Juusola, M. T. Cho, P. Vitazka, F. Millan, F. Gibellini, A. Vertino-Bell, N. Smaoui, J. Neidich, K. G. Monaghan, D. McKnight, R. Bai, S. Suchy, B. Friedman, J. Tahiliani, D. Pineda-Alvarez, G. Richard, T. Brandt, E. Haverfield, W. K. Chung, S. Bale, Clinical application of whole-exome sequencing across clinical indications. *Genet. Med.* **18**, 696–704 (2016).
31. D. Trujillano, A. M. Bertoli-Avella, K. Kumar Kandaswamy, M. E. Weiss, J. Koster, A. Marais, O. Paknia, R. Schroder, J. M. Garcia-Aznar, M. Werber, O. Brandau, M. Calvo Del Castillo, C. Baldi, K. Wessel, S. Kishore, N. Nahavandi, W. Eyaid, M. T. Al Rifai, A. Al-Rumayyan, W. Al-Twaijri, A. Alothaim, A. Alhashem, N. Al-Sannaa, M. Al-Balwi, M. Alfadhel, A. Rolfs,

R. Abou Jamra, Clinical exome sequencing: Results from 2819 samples reflecting 1000 families. *Eur. J. Hum. Genet.* **25**, 176–182 (2017).

32. H. Lee, J. L. Deignan, N. Dorrani, S. P. Strom, S. Kantarci, F. Quintero-Rivera, K. Das, T. Toy, B. Harry, M. Yourshaw, M. Fox, B. L. Fogel, J. A. Martinez-Agosto, D. A. Wong, V. Y. Chang, P. B. Shieh, C. G. Palmer, K. M. Dipple, W. W. Grody, E. Vilain, S. F. Nelson, Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* **312**, 1880–1887 (2014).
33. M. M. Clark, Z. Stark, L. Farnaes, T. Y. Tan, S. M. White, D. Dimmock, S. F. Kingsmore, Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom. Med.* **3**, 16 (2018).
34. E. L. Salfati, E. G. Spencer, S. E. Topol, E. D. Muse, M. Rueda, J. R. Lucas, G. N. Wagner, S. Campman, E. J. Topol, A. Torkamani, Re-analysis of whole-exome sequencing data uncovers novel diagnostic variants and improves molecular diagnostic yields for sudden death and idiopathic diseases. *Genome Med.* **11**, 83 (2019).
35. J. Ramchand, M. Wallis, I. Macciocca, E. Lynch, O. Farouque, M. Martyn, D. Phelan, B. Chong, S. Lockwood, R. Weintraub, T. Thompson, A. Trainer, D. Zentner, J. Vohra, M. Chetrit, D. L. Hare, P. James, Prospective evaluation of the utility of whole exome sequencing in dilated cardiomyopathy. *J. Am. Heart Assoc.* **9**, e013346 (2020).
36. X. Dong, B. Liu, L. Yang, H. Wang, B. Wu, R. Liu, H. Chen, X. Chen, S. Yu, B. Chen, S. Wang, X. Xu, W. Zhou, Y. Lu, Clinical exome sequencing as the first-tier test for diagnosing developmental disorders covering both CNV and SNV: A Chinese cohort. *J. Med. Genet.* **57**, 558–566 (2020).
37. S. Yi, S. Lin, Y. Li, W. Zhao, G. B. Mills, N. Sahni, Functional variomics and network perturbation: Connecting genotype to phenotype in cancer. *Nat. Rev. Genet.* **18**, 395–410 (2017).
38. M. Bouhaddou, M. Eckhardt, Z. Z. Chi Naing, M. Kim, T. Ideker, N. J. Krogan, Mapping the protein-protein and genetic interactions of cancer to guide precision medicine. *Curr. Opin. Genet. Dev.* **54**, 110–117 (2019).
39. N. Sahni, S. Yi, Q. Zhong, N. Jailkhani, B. Charloteaux, M. E. Cusick, M. Vidal, Edgotype: A fundamental link between genotype and phenotype. *Curr. Opin. Genet. Dev.* **23**, 649–657 (2013).
40. B. M. Kuenzi, T. Ideker, A census of pathway maps in cancer systems biology. *Nat. Rev. Cancer* **20**, 233–246 (2020).
41. L. Bouaoun, D. Sonkin, M. Ardin, M. Hollstein, G. Byrnes, J. Zavadil, M. Olivier, TP53 variations in human cancers: New lessons from the IARC TP53 database and genomics data. *Hum. Mutat.* **37**, 865–876 (2016).
42. P. K. Ng, J. Li, K. J. Jeong, S. Shao, H. Chen, Y. H. Tsang, S. Sengupta, Z. Wang, V. H. Bhavana, R. Tran, S. Soewito, D. C. Minussi, D. Moreno, K. Kong, T. Dogruluk, H. Lu, J. Gao, C. Tokheim, D. C. Zhou, A. M. Johnson, J. Zeng, C. K. M. Ip, Z. Ju, M. Wester, S. Yu, Y. Li, C. P. Vellano, N. Schultz, R. Karchin, L. Ding, Y. Lu, L. W. T. Cheung, K. Chen, K. R. Shaw, F. Meric-Bernstam, K. L. Scott, S. Yi, N. Sahni, H. Liang, G. B. Mills, Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* **33**, 450–462.e10 (2018).
43. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
44. N. Sahni, S. Yi, M. Taipale, J. I. F. Bass, J. Coulombe-Huntington, F. Yang, J. Peng, J. Weile, G. I. Karras, Y. Wang, I. A. Kovacs, A. Kamburov, I. Krykbaeva, M. H. Lam, G. Tucker, V. Khurana, A. Sharma, Y. Y. Liu, N. Yachie, Q. Zhong, Y. Shen, A. Palagi, A. San-Miguel, C. Fan, D. Balcha, A. Dricot, D. M. Jordan, J. M. Walsh, A. A. Shah, X. Yang, A. K. Stoyanova, A. Leighton, M. A. Calderwood, Y. Jacob, M. E. Cusick, K. Salehi-Ashtiani, L. J. Whitesell, S. Sunyaev, B. Berger, A. L. Barabasi, B. Charloteaux, D. E. Hill, T. Hao, F. P. Roth, Y. Xia, A. J. M. Walhout, S. Lindquist, M. Vidal, Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
45. Y. Li, B. Burgman, I. S. Khatri, S. R. Pentaparthi, Z. Su, D. J. McGrail, Y. Li, E. Wu, S. G. Eckhardt, N. Sahni, S. S. Yi, e-MutPath: Computational modeling reveals the functional landscape of genetic mutations rewiring interactome networks. *Nucleic Acids Res.* **49**, e2 (2021).
46. M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
47. I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein, M. A. Nowak, Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci.* **107**, 18545–18550 (2010).
48. V. M. Narasimhan, K. A. Hunt, D. Mason, C. L. Baker, K. J. Karczewski, M. R. Barnes, A. H. Barnett, C. Bates, S. Bellary, N. A. Bockett, Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474–477 (2016).
49. D. Saleheen, P. Natarajan, I. M. Armean, W. Zhao, A. Rasheed, S. A. Khetarpal, H.-H. Won, K. J. Karczewski, A. H. O'Donnell-Luria, K. E. Samocha, Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239 (2017).
50. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
51. N. Gao, J. Zhang, M. A. Rao, T. C. Case, J. Mirosevich, Y. Wang, R. Jin, A. Gupta, P. S. Rennie, R. J. Matusik, The role of hepatocyte nuclear factor-3 alpha (Forkhead Box A1) and androgen receptor in transcriptional regulation of prostatic genes. *Mol. Endocrinol.* **17**, 1484–1507 (2003).
52. Cancer Genome Atlas Research Network, The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
53. E. J. Adams, W. R. Karthaus, E. Hoover, D. Liu, A. Gruet, Z. Zhang, H. Cho, R. DiLoreto, S. Chhangawala, Y. Liu, P. A. Watson, E. Davicioni, A. Sboner, C. E. Barbieri, R. Bose, C. S. Leslie, C. L. Sawyers, FOXA1 mutations alter pioneering activity, differentiation and prostate cancer phenotypes. *Nature* **571**, 408–412 (2019).
54. A. Parolia, M. Cieslik, S. C. Chu, L. Xiao, T. Ouchi, Y. Zhang, X. Wang, P. Vats, X. Cao, S. Pitchiaya, F. Su, R. Wang, F. Y. Feng, Y. M. Wu, R. J. Lonigro, D. R. Robinson, A. M. Chinnaiyan, Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. *Nature* **571**, 413–418 (2019).
55. K. Wang, M. Li, H. Hakonarson, ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
56. Q. Li, K. Wang, InterVar: Clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am. J. Hum. Genet.* **100**, 267–280 (2017).
57. C. Dong, Y. Guo, H. Yang, Z. He, X. Liu, K. Wang, iCAGES: Integrated CAncer GEnome Score for comprehensively prioritizing driver genes in personal cancer genomes. *Genome Med.* **8**, 135 (2016).
58. X. Chang, K. Wang, wANNOVAR: Annotating genetic variants for personal genomes via the web. *J. Med. Genet.* **49**, 433–436 (2012).
59. M. Garber, M. Guttman, M. Clamp, M. C. Zody, N. Friedman, X. Xie, Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–i62 (2009).

Acknowledgments: We would like to thank S. N. Martins-Filho and N.-A. Pham for constructive comments. We also thank members of the Wang laboratory for helpful comments on the user interface of the CancerVar web server and for testing the CancerVar web server. Q.L. dedicates this study to the memory of his beloved father, Hong Li. **Funding:** This work was supported by the National Institutes of Health (NIH)/National Library of Medicine (NLM)/ National Human Genome Research Institute (NHGRI) (grant number LM012895), NIH/National Institute of General Medical Sciences (NIGMS) (grant number GM120609 and GM132713), CHOP Pathology Diagnostic Innovation Award, and CHOP Research Institute. **Author contributions:** Q.L., Y.Z., and K.W. conceived the study. Q.L. and Z.R. performed the coding and generated and interpreted the results. K.C. and M.M.L. provided manual annotation data and advised on the study. Q.L., Y.Z., and K.W. wrote the original manuscript with input from all authors. All authors reviewed and approved the final manuscript. **Competing interests:** K.W. indirectly owned shares in PierianDx. All other authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 24 April 2021
Accepted 21 March 2022
Published 6 May 2022
10.1126/sciadv.abj1624