CancerVar: a web server for improved evidence-based clinical interpretation of cancer somatic mutations and copy number abnormalities

Quan Li[1,2*], Zilin Ren[2], Yunyun Zhou[2], and Kai Wang[2,3*]

[1]Princess Margaret Cancer Centre, University Health Network, University of Toronto, Toronto, Ontario, M5G2C1, Canada

[2]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[3]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

* To whom correspondence should be addressed: Dr. Kai Wang, Tel: +1 267 425 9573; Fax: +1 215 590 3660; Email: wangk@email.chop.edu Correspondence may also be addressed to Dr. Quan Li, Tel: +1 416 634 8721; Email: leequan@gmail.com;

ABSTRACT

Several knowledgebases, such as CIViC, CGI and OncoKB, have been manually curated to support clinical interpretations of somatic mutations and copy number abnormalities (CNAs) in cancer. However, these resources focus on known hotspot mutations, and discrepancies or even conflicting interpretations have been observed between these knowledgebases. To standardize clinical interpretation, AMP/ASCO/CAP/ACMG/CGC jointly published consensus guidelines for the interpretations of somatic mutations and CNAs in 2017 and 2019, respectively. Based on these guidelines, we developed a standardized, semi-automated interpretation tool called CancerVar (Cancer Variants interpretation), with a user-friendly web interface to assess the clinical impacts of somatic variants. Using a semi-supervised method, CancerVar interpret the clinical impacts of cancer variants as four tiers: strong clinical significance, potential clinical significance, unknown clinical significance, benign/likely benign. CancerVar also allows users to specify criteria or adjust scoring weights as a customized interpretation strategy, and allows phenotype-driven scoring for specific types of cancer. Importantly, CancerVar generates automated texts to summarize clinical evidence on somatic variants, which greatly reduces manual workload to write interpretations that include relevant information from harmonized knowledgebases. CancerVar can be accessed at http://cancervar.wglab.org and it is open to all users without login requirements. The command line tool is also available at https://github.com/WGLab/CancerVar.

Keywords: somatic variants, copy number alteration, clinical interpretation, precision oncology, CancerVar

INTRODUCTION

A large number of somatic variants have been identified by next-generation sequencing (NGS) during the practice of clinical oncology to facilitate precision medicine (1,2). In order 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(PMKB) (5), the JAX-Clinical Knowledgebase (CKB) (6), and Cancer Genome Interpreter (CGI) (7). However, the interpretation of cancer variants is still not a standardized practice, and different clinical groups often generate different or even conflicting results. To standardize clinical interpretation of cancer variants, the Association for Molecular Pathology (AMP), American Society of Clinical Oncology (ASCO), College of American Pathologists (CAP), American College of Medical Genetics and Genomics (ACMG) and the Cancer Genomics Consortium (CGC) jointly proposed standardized guidelines for reporting and interpretation of single nucleotide variants (SNVs), indels (8) and copy number abnormalities (CNAs) (9) in 2017 and 2019. These guidelines categorize somatic variants CNAs into a four-tiered system, namely strong clinical significance or oncogenic, potential clinical significance or likely oncogenic, unknown clinical significance, and benign/likely benign.

Accurate interpretation of clinical significance depends greatly on the harmonization of evidence, which should be precisely derived and standardized from multiple databases and annotations. To evaluate the reliability of the 2017 AMP/ASCO/CAP guidelines, Sirohi et al., compared human classifications for fifty-one variants by randomly selected 20 molecular pathologists from 10 institutions (10). The original overall observed agreement was only 58%. When providing the same evidential data of variants to the pathologists, the agreement rate of re-classification can increase to 70%. However, there are still some interpretation discordance in intra- and inter-laboratory settings. The reasons for discordance are: (i) gathering information/evidence is quite complicated, and may not even be reproducible by the same interpreter at different time points; (ii) different researchers may prefer to use different algorithms, cutoffs and parameters, making the interpretation less comparable between laboratories.

To address these issues and improve automated clinical interpretations of cancer variants, we developed a new web application called CancerVar (Cancer Variants interpretation), with a user-friendly web interface to assess the clinical impacts of somatic variants and CNAs. CancerVar is a standardized, semi-automated interpretation tool using 12 clinical-based

criteria from AMP/ASCO/CAP guidelines. It can generate automated texts with summarized descriptive interpretations, such as diagnostic, prognostic, targeted drug responses and clinical trial information for many hotspot mutations, which will significantly reduce the workload of human reviewers and advance the precision medicine in clinical oncology. CancerVar also allows users to query clinical interpretations for variants using chromosome position, cDNA change or protein change, and interactively fine-tune weights of scoring features based on prior knowledge or additional user-specified criteria. Compared to existing knowledgebases that document specific hotspot mutations, CancerVar is an improved web server providing polished and semi-automated clinical interpretations for somatic variants in cancer, and it greatly facilitates human reviewers draft clinical reports for panel sequencing, exome sequencing or whole genome sequencing on cancer.

MATERIAL AND METHODS

Collection of clinical evidence criteria

According to the AMP/ASCO/CAP 2017 guidelines, there are a total of 12 types of clinically derived evidence to predict the clinical significance for somatic variants, including therapies, mutation types, variant allele fraction (mosaic variant frequency (likely somatic), non-mosaic variant frequency (potential germline)), population databases, germline databases, somatic databases, predictive results of different computational algorithms, pathway involvement, and publications (8,11). As shown in Figure 1, CancerVar contains all the above 12 types of evidence, among which 10 of them are automatically generated, while the other two, including variant allele fraction and potential germline, require user input for manual adjustment.

Prioritization of clinical significance for variants from the evidence-based scoring system

CancerVar evaluates each set of evidence as clinical evidence based prediction (CBP). The variant evidence will get 2 points for moderate evidence for clinical significance,1 point for supporting evidence for clinical significance or oncogenic, 0 for no support of clinical significance or neutral, -1 for evidence supporting benign impact. The CancerVar score will be the sum of all the evidence. The complete score system for each CBP can be found in Supplementary Table 1. Let the $CBP[i]$ be the ith evidence score, weight $[i]$ is the score for i^{th} evidence. The CancerVar score can be calculated in Equation 1. The weight is 1 by default, but users can adjust it based on its importance from prior knowledge. Based on the score range in Equation 2, we classify each variant into one of the four Tiers: strong clinical

significance (oncogenic), potential clinical significance (likely oncogenic), unknown clinical significance (VUS), and benign/likely benign.

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

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

Clinical interpretations for somatic CNAs in cancer

The interpretation of somatic CNAs is slightly different from the above approach since it does not have a scoring system. The ACMG/CGC guideline (9) proposed four Tier evidence-based categorization system for CNAs as:

1. Variants with strong clinical significance (Tier 1/ Strong clinical significance): the CNA had diagnostic, prognostic, and/or therapeutic evidence, and included in professional guidelines, and/or can be treated with FDA approved drugs.
2. Variants with some clinical significance (Tier 2/ Potential clinical significance): Recurrent CNAs observed in different neoplasm but not specific to a particular tumor type, or shown average quality of diagnostic, prognostic or therapeutic evidence in specific neoplasm.
3. Variants with no documented cancer association (Tier 3/Uncertain Significance): any CNAs that do not meet the criteria for Tiers 1 and 2 and cannot be classified as benign or likely benign.
4. Variants as benign or likely benign (Tier 4/(Likely) Benign): CNA listed in the ClinGen with curated benign variants and/or in the Database of Genomic Variants (DGV) with >=1% population frequency.

We curated the CNA list from the ACMG/CGC guideline with the above criteria and complied the final CNA list for oncogenicity prediction.

Database sources and pre-processing

The cancer gene census list of potential cancer driver genes were important to many somatic variants annotation tools. To focus on established cancer genes, we curated 1,911 cancer genes with 13 million exonic variants and 1,063 CNAs from existing cancer databases, including COSMIC, CIViC, OncoKB, CGI and others. For each exon position in these 1,911 genes, we generated all three possible nucleotide changes. All other knowledge-based annotation datasets were compiling the variants which were only previously reported or documented, while CancerVar fully scanned all the potential variants of significance. This is one of different aspects when comparing CancerVar with other knowledge-based annotators.

We pre-compiled 10 criteria for all the possible variant changes. This makes the variant searching in CancerVar very fast. In CancerVar, we documented all types of clinical evidence such as *in-silico* prediction, drug information, and publications in detail to help users making their own clinical decisions according to their prior knowledge.

Web server implementation

CancerVar is a web server that is free and open to all users without login requirements. CancerVar is implemented in PHP framework (version 7.2.15, http://www.php.net), MongoDB (version 1.3.4, https://www.mongodb.com), and Apache (version 2.4.29, https://httpd.apache.org), with support for all major web browsers. The front client-side interface was implemented using HTML5, Bootstrap (https://getbootstrap.com/) and JavaScript libraries, including jQuery (http://jquery.com). The data of pre-compiled mutations and CNAs are stored in MongoDB tables. To enhance security and protect privacy, the interpretation is calculated on the fly, and no user data are stored on server-side except access logs.

RESULTS

Summary of input and output features

The illustration of CancerVar web interface is shown in Figure 2. CancerVar provides multiple query options at variants-, gene-, and CNA levels across 30 cancer types and two versions of reference genomes: hg19 (GRCh37) and hg38 (GRCh38). Given user-supplied input, CancerVar generates an output web page, with information organized as cards including free text interpretation summary, gene overview, mutation information, evidence overview, pathways, clinical publications, protein domains, *in silico* predictions, exchangeable information from other knowledgebases.

Inputs

In the CancerVar web portal, users can search their exonic cancer variants by genomic coordinate positions, dbSNP ID, and HGNC gene symbol with cDNA or protein change. If the user already know the information of each of the scoring criteria for the variant (possibly inferred by themselves using other software tools), they can alternatively compute the clinical significance of the variant from the “Interpret by Criteria” service instead. In addition to the web server, we also provide local version of CancerVar for batch processing of variants. For the local command-line version of CancerVar, the input file could be pre-annotated files in tab-delimited format (generated by ANNOVAR or other annotation tools), or unannotated input files in VCF (CancerVar will call ANNOVAR to generate necessary annotations).

Outputs

The CancerVar web server provides full details on the variants, including all the automatically generated criteria, most of the supporting evidence and other necessary information. Users then have the ability to manually adjust these criteria and perform re-interpretation based on their prior knowledge or experience. For the local command-line version of CancerVar, the output will be 12 sets of criteria that are either automatically generated or manually supplied by the user, each variant will be provided with a prediction score and clinical interpretation as strong clinical significance, potential clinical significance, uncertain significance, likely benign/benign based on the AMP/ASCO/CAP/ACMG/CGC guidelines.

Web application programming interface (API) and standalone software for programmatic access

The CancerVar RESTful service provides other web applications access to CancerVar with a straightforward protocol. With the support of RESTful services, other variant annotation applications which are built on various programming languages and platforms can easily access CancerVar interpretation information. The API of CancerVar is implemented to provide data in two ways, JavaScript Object Notation (JSON) or direct web page as HyperText Markup Language (HTML). Full details and code samples are provided at the CancerVar online tutorial. Additionally, the local version of CancerVar is command-line driven program implemented in Python, and can be used on a variety of popular operating systems where Python is installed. The source code of the local version of CancerVar and step-by-step instructions are freely available from GitHub (https://github.com/wglab/CancerVar) for non-commercial users.

Performance assessment and comparative evaluations of CancerVar

Sirohi et. al measured the reliability of the 2017 AMP/ASCO/CAP guidelines (10) using fifty-one variants (31 SNVs, 14 indels, 5 CNAs, one fusion) based on literature review. In their study, they found an agreement rate of only around 58% between different groups. When provided with detailed information on evidence for variants, the agreement rate can increase to 70%. Among these variants, we selected 48 variants including all 31 SNVs, all 5 CNAs, 12 insertion-deletion variants (we did not find alternative alleles information for two indels in gene *CHEK1* and *MET*). CancerVar interpreted these 48 variants with the specified cancer types. Since these 48 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 24 variants as strong or potential clinical significance. Among these 24 variants, the pathologists classified 20 variants as strong or potential clinical significance in agreement. Moreover, CancerVar assigned 23 variants as VUS; among these 23 variants, 11 variants also be classified as VUS by pathologist reporters. In total, 31 variants (around 65%)

have a match of clinical significance between human reporters and CancerVar. The interpretation details of these 48 variants can be found in the Supplementary Table 2 and Supplementary Figure 1. Compared to human interpreters, the advantages of CancerVar is clear, in that it can automatically generate clinical interpretations with standardized, consistent and reproducible workflow, with evidence-based support for each of the 12 criteria. Therefore, CancerVar will greatly reduce the workload of human reviewers and facilitate the generation of precise and reproducible clinical interpretation.

USE CASES

In this section, we provide three examples of the possible user cases for CancerVar and demonstrate the advantages of CancerVar.

Use Case 1: Comprehensive interpretation of *AKT1* somatic mutations in Breast Cancer

In this use case, we showed the clinical interpretation of the E17K mutation in *AKT1* for breast cancer. Breast cancer is one of the most common cancer diagnosed in women in the world (12), and a number of significantly mutated genes were previously implicated in breast cancer, such as *PIK3CA*, *PTEN*, *AKT1*, *TP53*, *PTPN22*, *PTPRD*, *NF1*, *SF3B1* and *CDKN1B* (13). The *AKT1* (encoding protein kinase B) is a member of the serine-threonine kinase class and a known oncogene in breast cancer, which plays a key role in breast cancer onset (14,15). One hotspot mutation in *AKT1* [c.G49A:p.E17K] has been observed in the highest incidence in breast cancer, acting as oncogenic and a therapeutic target (15,16).

We queried this missense mutation using protein change and selected cancer type as “Breast”. CancerVar quickly returned the information displayed as several cards in the web page, including interpretation summary, gene overview, mutation information, evidence overview, pathways, clinical publications, protein domains, other *in silico* predictions, exchangeable information from other databases. CancerVar also automatically generated the free text interpretation summary of the variant’s clinical impact. For this mutation, CancerVar assigned the automated clinical significance as “Tier II/Potential clinical significance”. Additionally, it provided a one-stop shop from variants to genes and to drugs under specific cancer types. From CBP_1, CancerVar found some therapeutic evidence in breast cancer, and showed that known treatments mostly involved the drug Capivasertib and AZD5363, and most of the drug responses are sensitive (each of them >38%). However, from the clinical publication card and therapeutic evidence detail table, we also found some reports mentioning that the drug MK-2206(PI3K pathway inhibitor) did not show sensitivity or show no benefit (15,17) after the treatments. CancerVar did not find more evidence from the diagnostic and prognostic aspects. From CBP_7, this mutation is absent or extremely low minor allele frequency in the public

cohorts such as gnomAD, ExAC, ESP6500 and 1000 Genomes Project. ClinVar also noted this mutation as pathogenic. This mutation were also found in the COSMIC database and ICGC database. Four out of 7 *in silico* methods (including SIFT (18), PolyPhen2 (19), MutationAssessor (20,21), MetaLR (22), GERP++ (23), MetaSVM (22), FATHMM (24)) predicted this mutation as (likely) pathogenic. Finally, this E17K mutation got a score of 10 and then was assigned as Potential clinical significance by CancerVar. The detail of this use case can be viewed in Figure 3.

Use Case 2: Re-interpretation of *FOXA1* somatic mutation in Prostate Cancer

This user case illustrates how to use automated interpretation and manual adjustment to derive a final interpretation for somatic mutations. Prostate cancer is the most commonly diagnosed cancer in men in the world (12). The *FOXA1* protein (Forkhead box A1, previously known as HNF3a) is essential for the normal development of the prostate (25). The *FOXA1* somatic mutations have been observed frequently in prostate cancer(26) and are associated with poor outcome. However, the mechanism of driving prostate cancer by mutations in *FOXA1* was still not clear. Recently, two papers published in Nature demonstrated that *FOXA1* acts as an oncogene in prostate cancer (27,28). They found that the hotspot mutation at R219 (R219S and R219C) drove a pro-luminal phenotype in prostate cancer and exclusive with other fusions or mutations (27,28). We interpreted these two mutations, but here we only illustrated the clinical interpretation for R219S since the interpretation result of R219C was very similar to R219S. We searched this missense mutation R219S using protein change and gene name as "*FOXA1*" in the CancerVar web server. CancerVar gave automated clinical significance as "Uncertain Significance". CancerVar did not find any therapeutic, diagnostic and prognostic evidence for this mutation. In addition, from CBP_7, this mutation is absent or has extremely low minor allele frequency in the public allele frequency database. All seven *in silico* methods predicted this mutation as (likely) pathogenic. According the AMP/ASCO/CAP guidelines, this variant falls into the class of "uncertain significance" with a score of 5. However, we need to manually adjust the weight of CBP_12 since recently two publications reported its biological functions in prostate cancer. We also need to adjust the weight of CBP_9 as moderate evidence, since this mutation has been recently incorporated in somatic databases including COSMIC (ID: COSM3738526) and MSK-IMPACT(29). The new prediction of clinical significance was therefore changed to "Potential clinical significance" with a score of 8. This semi-automated interpretation approach will greatly improve the prediction accuracy for each variant, given existing knowledge and domain expertise. We acknowledge that a model-based approach involving machine intelligence can be as another option and may be explored in our future work. The detail of this use case can be viewed in Figure 4.

Use Case 3: CNA interpretation for *ERBB2* in Lung Cancer

CancerVar also provides clinical impact interpretation for CNAs in specific cancer types. Lung cancer is the most common cause of global cancer-related mortality worldwide. *ERBB2/HER2* amplification has been classified as one of the oncogenic drivers for lung cancer (30,31). We checked "Query by HGNC gene symbol or Alternations" in the CancerVar web server, using *ERBB2* as gene name with option as Copy numbers, cancer type as "Lung". CancerVar returned a table with 11 entries. We found that CancerVar listed all the *ERBB2* with CNAs as amplification in lung cancer and the clinical significance as "Potential clinical significance". The table also provided the drug and response for this CNA, with publications as PubMed ID and hyperlinked URL. The details of this use case can be viewed in Figure 5. Other comprehensive genomic alternations (such as gene fusions) may be incorporated in an expanded version of CancerVar in the future.

DISCUSSION

Clinical interpretation of cancer somatic variants remains an urgent need for clinicians and researchers in precision oncology, especially given the transition from panel sequencing to whole exome/genome sequencing in cancer genomics. To build a standardized, rapid and user-friendly interpretation tool, we developed a web server to assess the clinical impacts of somatic variants using the AMP/ASCO/CAP/ACMG/CGC guidelines. CancerVar is an enhanced version of cancer variants knowledgebase incorporated from our previously developed tools for variant annotations and prioritizations including InterVar (32), VIC (11), iCAGES (33), as well as assembling existing variants 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 to generate automatic evidence to facilitate/enhance human reviewers by providing a standardized, reproducible, and precise output for interpreting somatic variants.

In CancerVar, we did not reconcile the well-known "conflicting interpretation" issues across knowledgebases, but we documented and harmonized all types of clinical evidence (i.e. drug information, publications, etc) in detail to allow users make their own clinical decisions based on their own domain knowledge and expertise. Compared to existing knowledgebases such as OncoKB, CIViC and CBK, CancerVar provides an improved platform in four areas: (i) comprehensive, evidence-based annotations with rigorous quality control for several types of somatic variants including SNPs, INDELs, and CNAs; (ii) well-designed, flexible scoring system allowing users to fine-tune the importance of evidence criteria according to their own prior beliefs; (iii) improved automation workflow for faster querying of variants of interests by

genomic positions, SNP ids, official gene symbols; (iv) automatically summarized interpretation text so that users do no need to query evidence from multiple knowledgebase 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 sufficiently robust. We note that the existing clinical guidelines did not provide the recommendations for weighting different evidence types, and therefore treated all weights as equal by default; however, with the increasing amounts of clinical knowledge on somatic mutations, we expect that we may build a weighted model in the future to enhance the prediction accuracy. Second, there are no scoring and weighting systems for CNA interpretation in the existing interpretation guidelines, so we did not design such a scoring system currently; in the future, we will design and implement the scoring system for CNAs based on the platform used to discover CNAs, 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 cannot interpret inversions and gene fusions, and cannot interpret gene expression alterations, even though these genomic alterations may also play important roles in cancer progression. Before a specific guideline for these types of mutations become 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). Fourth, our treatment of specific types or subtypes of cancer can be improved. Although CancerVar interprets “All cancer types” by default, we currently support the use of different cancer classification systems (such as OncoTree(http://oncotree.mskcc.org/)) to re-weight and fine-tune the interpretations (see Figure 3 and 4 as examples on breast and prostate cancer). Additionally, we are using phenotype-driven gene scoring approaches (such as Phenolyzer(34) and Phen2Gene(35)) to re-weight genes based on the cancer types in the interpretation procedure. Finally, artificial intelligence or machine learning may be used to integrate with larger knowledgebase, replace part of the rule-based decisions in the current guidelines. In user case 2, we specifically demonstrated that new literature based knowledge may be incorporated in the interpretation process to perform adjustment, and this procedure to update database information may be partially performed by artificial intelligence. In summary, CancerVar is a web server providing polished and semi-automated clinical interpretations for somatic variants in cancer, and it greatly facilitates human reviewers to draft clinical reports for panel sequencing, exome sequencing,whole genome sequencing and copy number assays on cancer. We expect to

continuously improve CancerVar and incorporate new functionalities in the future, similar to what we have done on the wInterVar server and wANNOVAR server.

AVAILABILITY

CancerVar is a web server, and it can be accessed at http://cancervar.wglab.org. The local command-line version of CancerVar is available on GitHub (https://github.com/wglab/CancerVar) for users who wish to perform batch analysis of somatic variants.

ACKNOWLEDGEMENT

We would like to thank Drs. Sebastiao N. Martins-Filho and Nhu-An Pham for constructive comments. We would like to thank Dr. Marilyn Li and Kajia Cao at Division of Genomic Diagnostics at Children's Hospital of Philadelphia for testing the web server. We also thank members of the Wang lab for helpful comments on the user interface of the CancerVar web server and for testing the CancerVar web server.

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] and National Institutes of Health (NIH)/National Institute of General Medical Sciences (NIGMS) [grant number GM120609 and GM132713] and CHOP Research Institute.

CONFLICT OF INTEREST

KW indirectly own shares but is not involved in the operation of PierianDx, which develops cloud-based solution for clinical interpretation of somatic mutations.

REFERENCES

1. 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 precision oncology*, **2017**.

2. 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*, **174**, 1034-1035.
3. Micheel, C.M., Sweeney, S.M., LeNoue-Newton, M.L., Andre, F., Bedard, P.L., Guinney, J., Meijer, G.A., Rollins, B.J., Sawyers, C.L., Schultz, N. *et al.* (2018) 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 clinical cancer informatics*, **2**, 1-14.
4. Griffith, M., Spies, N.C., Krysiak, K., McMichael, J.F., Coffman, A.C., Danos, A.M., Ainscough, B.J., Ramirez, C.A., Rieke, D.T., Kujan, L. *et al.* (2017) CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature genetics*, **49**, 170-174.
5. Huang, L., Fernandes, H., Zia, H., Tavassoli, P., Rennert, H., Pisapia, D., Imielinski, M., Sboner, A., Rubin, M.A., Kluk, M. *et al.* (2017) The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. *Journal of the American Medical Informatics Association : JAMIA*, **24**, 513-519.
6. Patterson, S.E., Liu, R., Statz, C.M., Durkin, D., Lakshminarayana, A. and Mockus, S.M. (2016) The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Human genomics*, **10**, 4.
7. 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 medicine*, **10**, 25.
8. Li, M.M., Datto, M., Duncavage, E.J., Kulkarni, S., Lindeman, N.I., Roy, S., Tsimberidou, A.M., Vnencak-Jones, C.L., Wolff, D.J., Younes, A. *et al.* (2017) 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. *The Journal of molecular diagnostics : JMD*, **19**, 4-23.
9. Mikhail, F.M., Biegel, J.A., Cooley, L.D., Dubuc, A.M., Hirsch, B., Horner, V.L., Newman, S., Shao, L., Wolff, D.J. and Raca, G. (2019) Technical laboratory standards for interpretation and reporting of acquired copy-number abnormalities and copy-neutral loss of heterozygosity in neoplastic disorders: a joint consensus recommendation from the American College of Medical Genetics and Genomics (ACMG) and the Cancer Genomics Consortium (CGC). *Genetics in medicine : official journal of the American College of Medical Genetics*, **21**, 1903-1916.
10. Sirohi, D., Schmidt, R.L., Aisner, D.L., Behdad, A., Betz, B.L., Brown, N., Coleman, J.F., Corless, C.L., Deftereos, G., Ewalt, M.D. *et al.* (2020) 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. *The Journal of molecular diagnostics : JMD*, **22**, 284-293.
11. He, M.M., Li, Q., Yan, M., Cao, H., Hu, Y., He, K.Y., Cao, K., Li, M.M. and Wang, K. (2019) Variant Interpretation for Cancer (VIC): a computational tool for assessing clinical impacts of somatic variants. *Genome medicine*, **11**, 53.
12. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A. (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, **68**, 394-424.
13. Cancer Genome Atlas, N. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61-70.

14. Ju, X., Katiyar, S., Wang, C., Liu, M., Jiao, X., Li, S., Zhou, J., Turner, J., Lisanti, M.P., Russell, R.G. *et al.* (2007) Akt1 governs breast cancer progression in vivo. *Proc Natl Acad Sci U S A*, **104**, 7438-7443.
15. Beaver, J.A., Gustin, J.P., Yi, K.H., Rajpurohit, A., Thomas, M., Gilbert, S.F., Rosen, D.M., Ho Park, B. and Lauring, J. (2013) PIK3CA and AKT1 mutations have distinct effects on sensitivity to targeted pathway inhibitors in an isogenic luminal breast cancer model system. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **19**, 5413-5422.
16. Hyman, D.M., Smyth, L.M., Donoghue, M.T.A., Westin, S.N., Bedard, P.L., Dean, E.J., Bando, H., El-Khoueiry, A.B., Perez-Fidalgo, J.A., Mita, A. *et al.* (2017) AKT Inhibition in Solid Tumors With AKT1 Mutations. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **35**, 2251-2259.
17. Xing, Y., Lin, N.U., Maurer, M.A., Chen, H., Mahvash, A., Sahin, A., Akcakanat, A., Li, Y., Abramson, V., Litton, J. *et al.* (2019) Phase II trial of AKT inhibitor MK-2206 in patients with advanced breast cancer who have tumors with PIK3CA or AKT mutations, and/or PTEN loss/PTEN mutation. *Breast cancer research : BCR*, **21**, 78.
18. Ng, P.C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, **31**, 3812-3814.
19. Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen‐2. *Current protocols in human genetics*, **76**, 7.20. 21-27.20. 41.
20. Schwarz, J.M., Rödelsperger, C., Schuelke, M. and Seelow, D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods*, **7**, 575.
21. Reva, B., Antipin, Y. and Sander, C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*, **39**, e118.
22. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K. and Liu, X. (2014) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics*, **24**, 2125-2137.
23. 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++. *PLoS computational biology*, **6**, e1001025.
24. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L., Edwards, K.J., Day, I.N. and Gaunt, T.R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation*, **34**, 57-65.
25. Gao, N., Zhang, J., Rao, M.A., Case, T.C., Mirosevich, J., Wang, Y., Jin, R., Gupta, A., Rennie, P.S. and Matusik, R.J. (2003) The role of hepatocyte nuclear factor-3 alpha (Forkhead Box A1) and androgen receptor in transcriptional regulation of prostatic genes. *Molecular endocrinology*, **17**, 1484-1507.
26. Cancer Genome Atlas Research, N. (2015) The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, **163**, 1011-1025.
27. Adams, E.J., Karthaus, W.R., Hoover, E., Liu, D., Gruet, A., Zhang, Z., Cho, H., DiLoreto, R., Chhangawala, S., Liu, Y. *et al.* (2019) FOXA1 mutations alter pioneering activity, differentiation and prostate cancer phenotypes. *Nature*, **571**, 408-412.
28. Parolia, A., Cieslik, M., Chu, S.C., Xiao, L., Ouchi, T., Zhang, Y., Wang, X., Vats, P., Cao, X., Pitchiaya, S. *et al.* (2019) Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. *Nature*, **571**, 413-418.
29. Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M. *et al.* (2017) Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med*, **23**, 703-713.

30. Li, B.T., Ross, D.S., Aisner, D.L., Chaft, J.E., Hsu, M., Kako, S.L., Kris, M.G., Varella-Garcia, M. and Arcila, M.E. (2016) HER2 Amplification and HER2 Mutation Are Distinct Molecular Targets in Lung Cancers. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, **11**, 414-419.
31. Kris, M.G., Camidge, D.R., Giaccone, G., Hida, T., Li, B.T., O'Connell, J., Taylor, I., Zhang, H., Arcila, M.E., Goldberg, Z. *et al.* (2015) Targeting HER2 aberrations as actionable drivers in lung cancers: phase II trial of the pan-HER tyrosine kinase inhibitor dacomitinib in patients with HER2-mutant or amplified tumors. *Annals of oncology : official journal of the European Society for Medical Oncology*, **26**, 1421-1427.
32. Li, Q. and Wang, K. (2017) InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet*, **100**, 267-280.
33. Dong, C., Guo, Y., Yang, H., He, Z., Liu, X. and Wang, K. (2016) iCAGES: integrated CAncer GEnome Score for comprehensively prioritizing driver genes in personal cancer genomes. *Genome medicine*, **8**, 135.
34. Yang, H., Robinson, P.N. and Wang, K. (2015) Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods*, **12**, 841-843.
35. Zhao, M., Havrilla, J.M., Fang, L., Chen, Y., Peng, J., Liu, C., Wu, C., Sarmady, M., Botas, P., Isla, J. *et al.* (2020) Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR genomics and bioinformatics*, **2**, lqaa032.

TABLE AND FIGURES LEGENDS

Table 1. Comparison of the classification on 48 variants between 20 pathologists and CancerVar.

CancerVar interpretation	Interpretation from 20 pathologists			
	S/P*	VUS	B	Total
S/P	20	4	0	24
VUS	11	11	1	23
B	0	1	0	1
Total	31	16	1	48

*S stands for Strong clinical significance, P stands for Potential clinical significance, VUS stands for variant uncertain significance, B stands for (likely) benign.

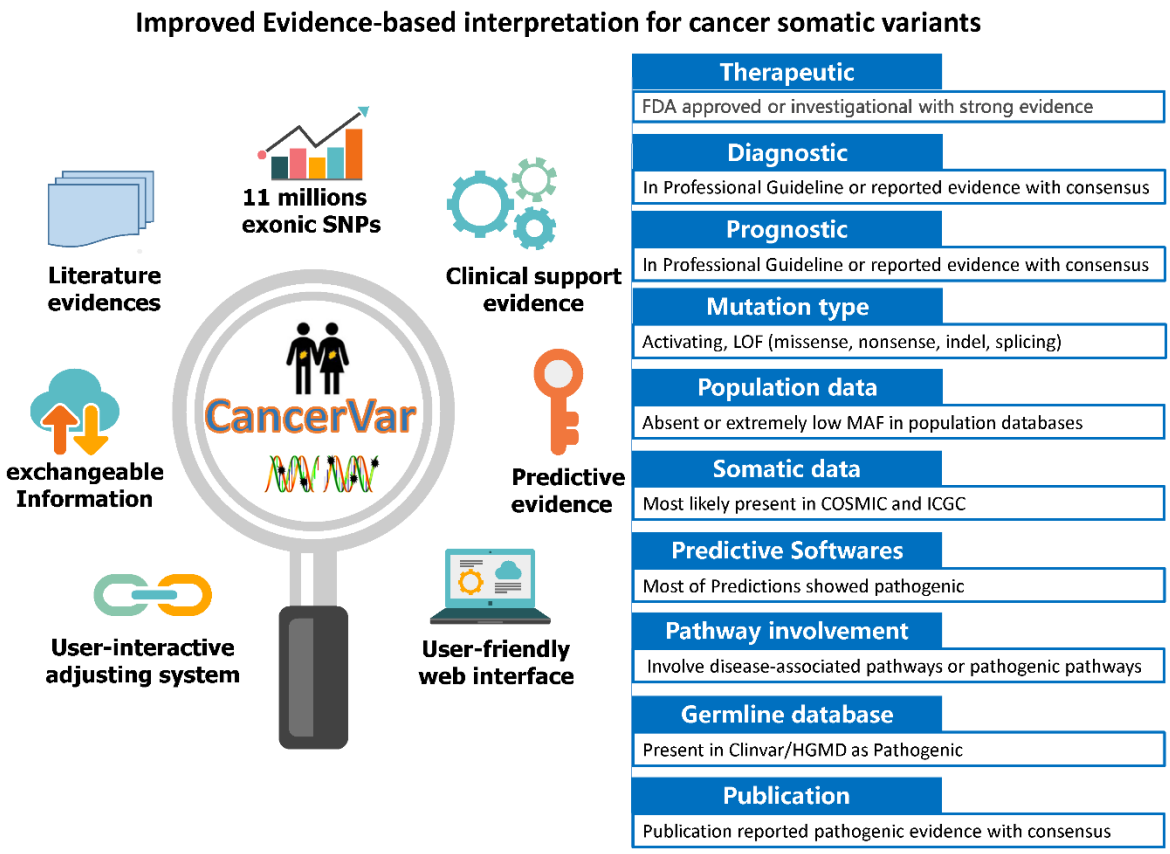


Figure 1. The functions of CancerVar and the descriptions of 12 types of evidence.

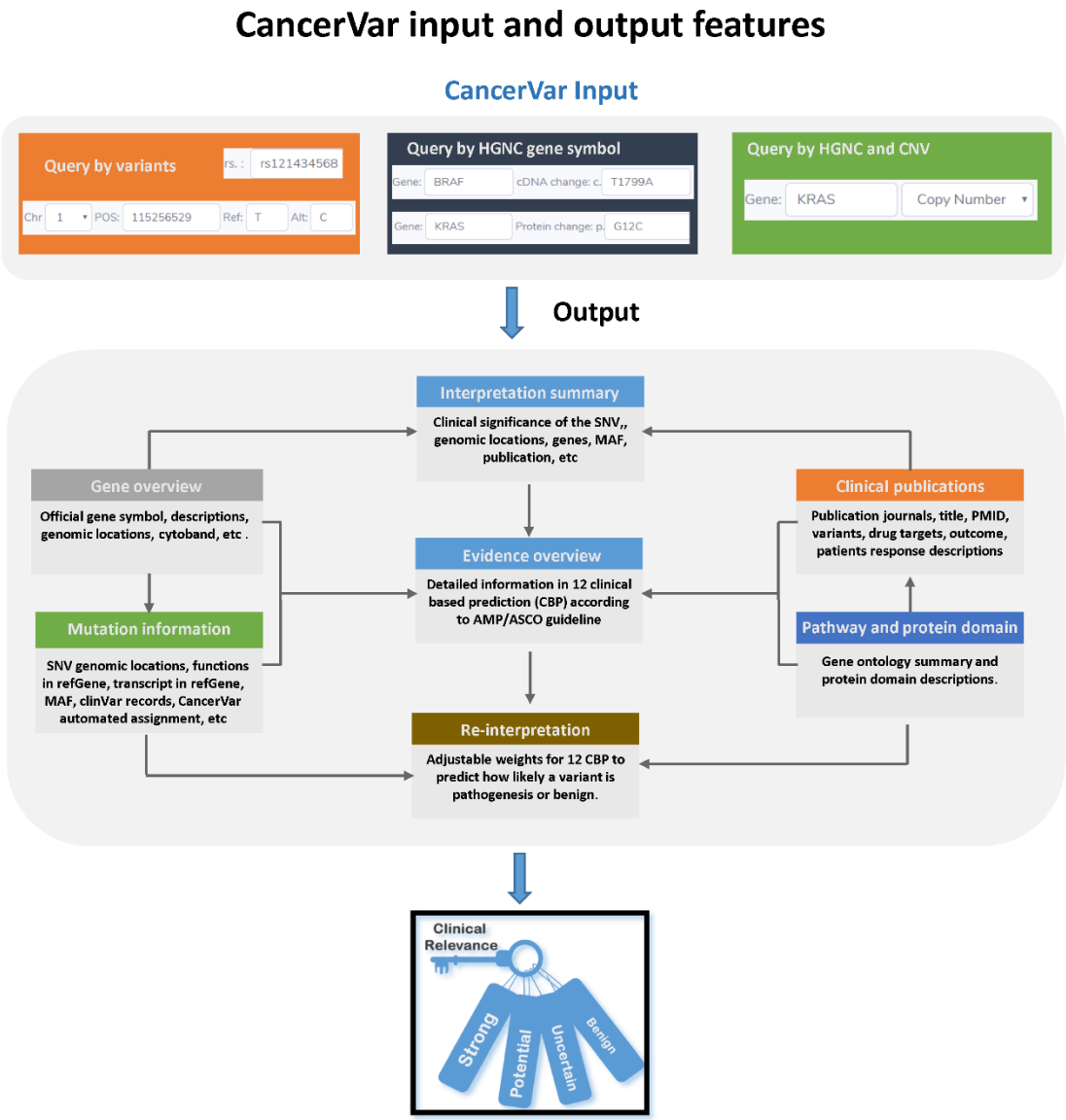
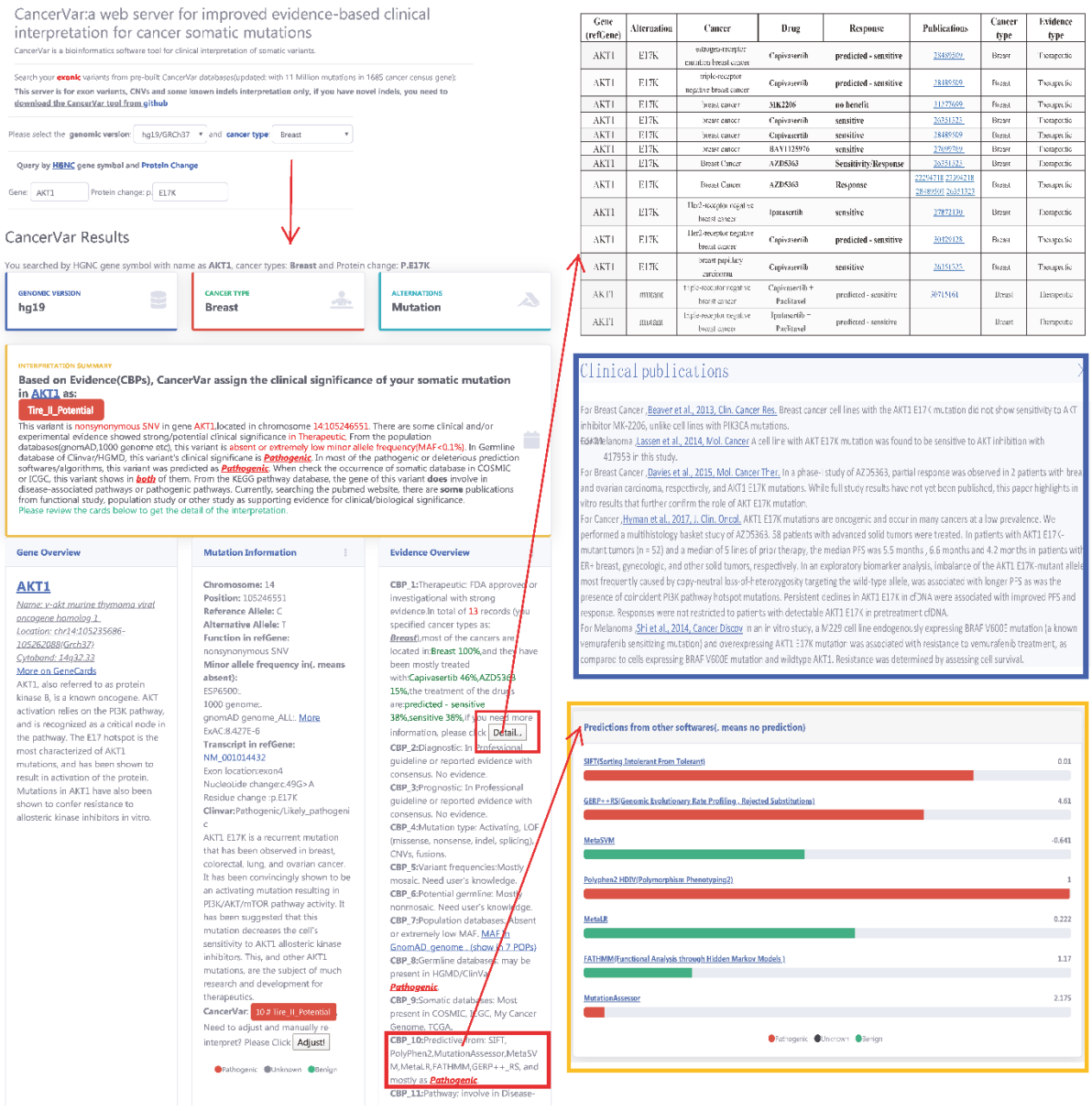


Figure 2. Illustration of input and output features in CancerVar web interface. Users can query variants by genomic locations, gene symbols with cDNA or protein change, and copy number variations. CancerVar classifies each variant into one of the four categories and displays all evidence in nine cards. CancerVar is also available in command-line version to facilitate batch analysis by users.



Gene (refGene)	Alternation	Cancer	Drug	Response	Publications	Cancer type	Evidence type
AKT1	E17K	estrogen-receptor mutation breast cancer	Capivasertib	predicted - sensitive	28489509	Breast	Therapeutic
AKT1	E17K	triple-receptor negative breast cancer	Capivasertib	predicted - sensitive	28489509	Breast	Therapeutic
AKT1	E17K	breast cancer	MK2206	no benefit	31127669	Breast	Therapeutic
AKT1	E17K	breast cancer	Capivasertib	sensitive	26351323	Breast	Therapeutic
AKT1	E17K	breast cancer	Capivasertib	sensitive	28489509	Breast	Therapeutic
AKT1	E17K	breast cancer	BAY1125976	sensitive	27689769	Breast	Therapeutic
AKT1	E17K	Breast Cancer	AZD5363	Sensitivity/Response	26351323	Breast	Therapeutic
AKT1	E17K	Breast Cancer	AZD5363	Response	22294718 27399218 28489509 26351323	Breast	Therapeutic
AKT1	E17K	Her2-receptor negative breast cancer	Ipatasertib	sensitive	27872139	Breast	Therapeutic
AKT1	E17K	Her2-receptor negative breast cancer	Capivasertib	predicted - sensitive	30429128	Breast	Therapeutic
AKT1	E17K	breast papillary carcinoma	Capivasertib	sensitive	26351323	Breast	Therapeutic
AKT1	mutant	triple-receptor negative breast cancer	Capivasertib + Paclitaxel	predicted - sensitive	36715161	Breast	Therapeutic
AKT1	mutant	triple-receptor negative breast cancer	Ipatasertib + Paclitaxel	predicted - sensitive		Breast	Therapeutic

Figure 3. Example outputs in use case 1: we queried AKT1 with E17K protein change in breast cancer. CancerVar provided detailed information for this variant, genes, summaries of 12 CBPs, and automatic classification as “Potential clinical significance”. The detailed information for each CBP is clickable. For example, when clicking CBP_1, specific therapeutic evidence such as target drugs and responses linking to PubMed IDs.

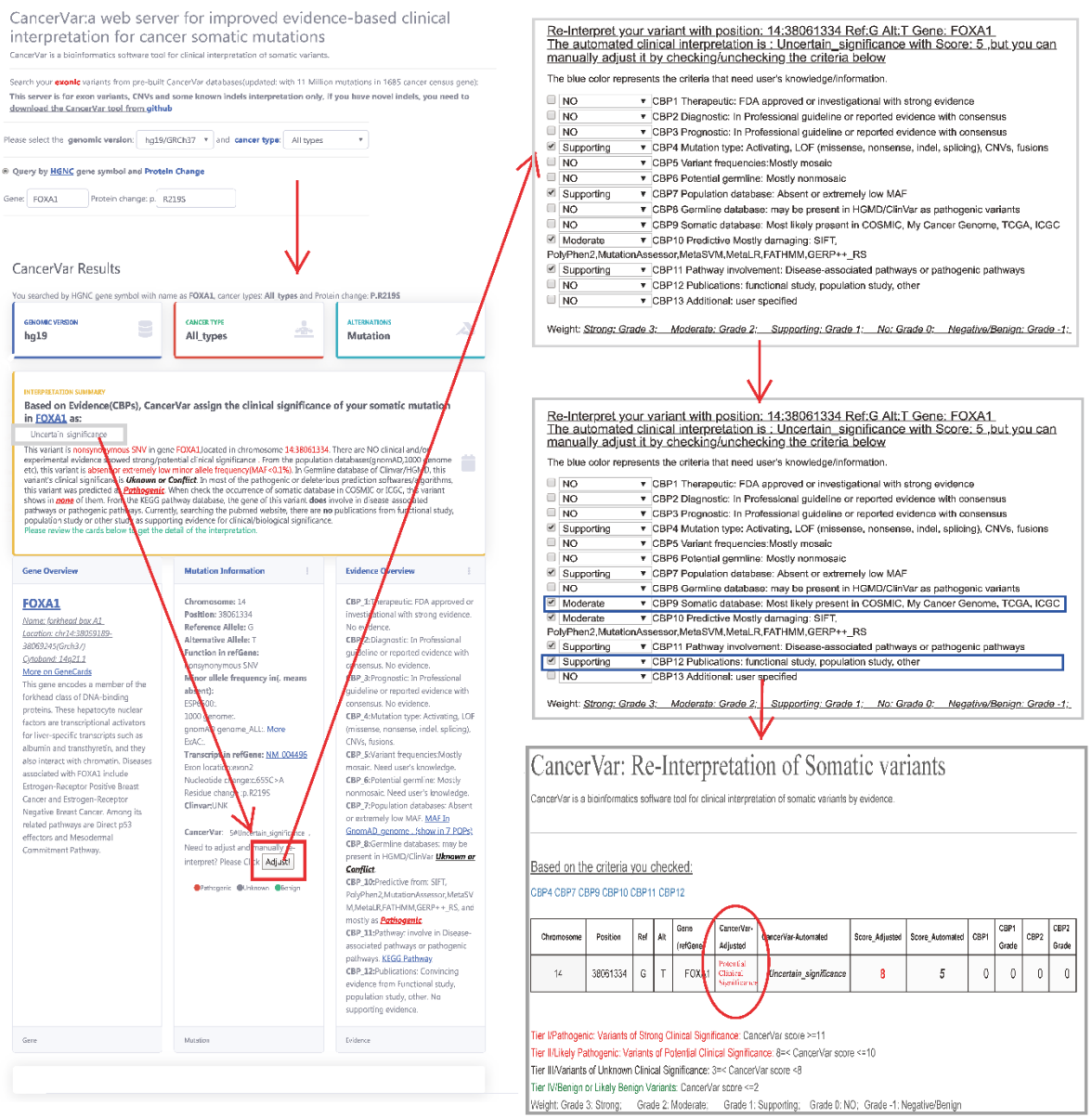


Figure 4. Adjustable weights for scoring variants in use case 2: we queried FOXA1 mutation R219S in prostate cancer. The original prediction of this variant was uncertain-significance. However, after adjusting the importance of CBP9 and CBP12 based on users' prior knowledge, CancerVar predicted this variant as Potential clinical significance.

You searched by other alternation with **ERBB2** , cancer types: **Lung** in **CNA**

GENOMIC VERSION	CANCER TYPE	ALTERNATIONS
hg19	Lung	CNA

Gene (refGene)	Alternation	CancerVar	Cancer	Drug	Response	Publications	Cancer type	Evidence type
ERBB2	Amplification	Potential Clinical Significance	Lung Adenocarcinoma	Erlotinib,Gefitinib	Resistance	23470965	Lung	Therapeutic
ERBB2	Amplification	Potential Clinical Significance	Lung Non-small Cell Carcinoma	Trastuzumab Emtansine	Sensitivity/Response	24898067	Lung	Therapeutic
ERBB2	Amplification	Potential Clinical Significance	Lung Small Cell Carcinoma	Irinotecan,Trastuzumab	Sensitivity/Response	25601188	Lung	Therapeutic
ERBB2	Amplification	Potential Clinical Significance	lung adenocarcinoma	Dacomitinib	no benefit	25899785	Lung	Therapeutic
ERBB2	Amplification	Potential Clinical Significance	lung cancer	TAS0728	predicted - sensitive	30787176	Lung	Therapeutic
ERBB2	Amplification	Potential Clinical Significance	lung non-small cell carcinoma	Afatinib	sensitive	26545934	Lung	Therapeutic
ERBB2	Amplification	Potential Clinical Significance	lung non-small cell carcinoma	Pertuzumab + Trastuzumab	predicted - sensitive	29320312	Lung	Therapeutic
ERBB2	Amplification	Potential Clinical Significance	lung squamous cell carcinoma	Gefitinib	resistant	26545934	Lung	Therapeutic
ERBB2	Amplification	Potential Clinical Significance	Lung adenocarcinoma	[EGFR inhibitor]	Resistant	22956644	Lung	Therapeutic
ERBB2	Amplification	Potential Clinical Significance	Lung adenocarcinoma	Osimertinib	Resistant	27252416	Lung	Therapeutic
ERBB2	Amplification	Potential Clinical Significance	Lung adenocarcinoma	Rociletinib	Resistant	27252416	Lung	Therapeutic

Figure 5. Clinical interpretations for CNAs in use case 3: we queried ERBB2 CNAs in Lung cancer. CancerVar displayed evidence that the majority of ERBB2 CNAs are amplification, predicted as Potential clinical significance. CancerVar also provided detailed therapeutic evidence of target drugs and patients' treatment responses for each CNAs.

Supp.Figure-1 Comparison of interpretation on 43 variants between 20 pathologists and CancerVar.

Cancer type	gene	protein.chang	B	VUS	LP	P	CancerVar Score	Pred.
Non-small cell lung carcinoma	AKT1	R48C	0	1	0	0	4	VUS
Lung adenocarcinoma	ARAF	S217C	0	0.5	0.5	0	8	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	6	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	5	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	6	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	8	LP
Pancreas, neuroendocrine	DAXX	A310V	0.05	0.89	0.05	0	1	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	13	P
Metastatic lung Adenocarcinoma	EGFR	V774M	0	0.56	0.33	0.11	13	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	7	VUS
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	12	P
Colon Adenocarcinoma	KRAS	K117N	0	0	0.1	0.9	13	P
Prostate adenocarcinoma	MUTYH	G396D	0.1	0.25	0.45	0.2	5	VUS
Renal cell carcinoma	NF2	W41X	0	0.25	0.7	0.05	5	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	8	LP
Metastatic prostate cancer	PIK3CA	Q546P	0	0.1	0.9	0	8	LP
Metastatic Melanoma	PTPN11	E69K	0	0.2	0.7	0.1	10	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	7	VUS
Lung adenocarcinoma	TP53	T211A	0	0.5	0.5	0	9	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	6	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	8	LP
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

*P stands for Strong clinical significance, LP stands for Potential clinical significance, VUS stands for variant uncertain significance, B stands for (likely) benign.