Focus on Computer Resources

Cancer Research

CRAVAT 4: Cancer-Related Analysis of Variants Toolkit

David L. Masica[1,2], Christopher Douville[1,2], Collin Tokheim[1,2], Rohit Bhattacharya[2,3], RyangGuk Kim[4], Kyle Moad[4], Michael C. Ryan[4], and Rachel Karchin[1,2,5]

Abstract

Cancer sequencing studies are increasingly comprehensive and well powered, returning long lists of somatic mutations that can be difficult to sort and interpret. Diligent analysis and quality control can require multiple computational tools of distinct utility and producing disparate output, creating additional challenges for the investigator. The Cancer-Related Analysis of Variants Toolkit (CRAVAT) is an evolving suite of informatics tools for mutation interpretation that includes mutation mapping and quality control, impact prediction and extensive annotation, gene- and mutation-level interpretation, including joint prioritization of all nonsilent mutation consequence types, and structural and mechanistic visualization. Results from CRAVAT submissions are explored in an interactive, user-friendly web environment with dynamic filtering and sorting designed to highlight the most informative mutations, even in the context of very large studies. CRAVAT can be run on a public web portal, in the cloud, or downloaded for local use, and is easily integrated with other methods for cancer omics analysis. *Cancer Res; 77(21); e35–38. ©2017 AACR.*

Background

An investigator's work is far from over when results are returned from the sequencing center. Depending on the service, genetic mutation calls can require additional mapping to include all relevant RNA transcripts or correct protein sequences. Assignment of mutation consequence type or sequence ontology (e.g., missense, splice, or indel) might be incomplete. Once mapping is complete and sequence ontology assigned, the task of interpreting mutation impact remains. This interpretation can entail annotation from multiple sources, employing one or more bioinformatics classifiers, testing for functional or pathway enrichment, and referencing each mutation against databases of known cancer drivers or common polymorphisms. Making full use of these diverse resources can be a daunting task. The individual utilities often assess a limited number of consequence types, require different formatting of input data, and return disparate output not amenable to direct comparison. Once the investigator has managed to assemble a suitable *ad hoc* pipeline to process and interpret each mutation, the task of effectively sorting and filtering large lists of mutations may be difficult.

The Cancer-Related Analysis of Variants Toolkit (CRAVAT; ref. 1) is designed to streamline the many steps outlined above, quickly returning mutation interpretations in an interactive and user-friendly web environment for sorting, visualizing, and inferring mechanism. CRAVAT (see Supplementary Video) is suitable for large studies (e.g., full-exome and large cohorts) and small studies (e.g., gene panel or single patient), performs all mapping and assigns sequence ontology, predicts mutation impact using multiple bioinformatics classifiers normalized to provide comparable output, allows for joint prioritization of all nonsilent mutation types, organizes annotation from many sources on graphical displays of protein sequence and 3D structure, and facilitates dynamic filtering and sorting of results so that the most important mutations can be quickly retrieved from large submissions. In addition to running on our webserver, CRAVAT is available as a Docker (https://www.docker.com/) image to run in a self-contained environment locally or in the cloud and is easily integrated with other software packages. The following describes the details of the CRAVAT interactive results explorer; machine-readable text and Excel spreadsheets are also provided with each job submission.

CRAVAT 4.x

The CRAVAT webserver (cravat.us) prompts the user to submit sequencing data, select impact prediction methods, and provide an email address for communicating job status. Once sequencing data are submitted, CRAVAT performs quality control and maps all mutations from genome to transcript, mapping each mutation to all relevant transcripts, and then to protein sequence; additionally, mutations are mapped to available protein structures and homology models. Next, all mutation consequence types (sequence ontologies) are assigned, including missense, stop gains and losses, in-frame and frame-shifting insertions and deletions (indel), splice, and silent mutations.

Mutation impact prediction

CRAVAT employs two approaches for predicting mutation impact, namely Cancer-Specific High-Throughput Annotation of Somatic Mutations (CHASM; ref. 2) and Variant Effect Scoring

[1]Department of Biomedical Engineering, The Johns Hopkins University, Baltimore, Maryland. [2]The Institute for Computational Medicine, The Johns Hopkins University, Baltimore, Maryland. [3]Department of Computer Science, The Johns Hopkins University, Baltimore, Maryland. [4]In Silico Solutions, Falls Church, Virginia. [5]Department of Oncology, The Johns Hopkins University School of Medicine, Baltimore, Maryland.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corresponding Author: Rachel Karchin, Johns Hopkins University, 316 Hackerman Hall, 3400 North Charles St., Baltimore, MD 21218. Phone: 410-516-5578; Fax: 410-516-5294; E-mail: rkarchi1@jhmi.edu

doi: 10.1158/0008-5472.CAN-17-0338

Tool (VEST; refs. 3, 4). CHASM uses a random Forest classifier that is trained from a positive class of cancer driver mutations from COSMIC (5) and a putative set of passenger mutations; thus, CHASM classifies mutations as cancer drivers or passengers (random Forests are ensembles of decision trees derived via machine learning). VEST uses a random Forest classifier that is trained from a positive class of disease-associated germline variants from the Human Gene Mutation Database (6) and a negative class of common variants from the exome sequencing map (7); thus, VEST classifies mutations as pathogenic or benign. CHASM can score missense mutations. VEST scores all nonsilent mutation consequence types and facilitates joint prioritization (i.e., scores are directly comparable among all mutations, regardless of sequence ontology). Both CHASM and VEST provide composite gene-level *P* values and false discovery rates (FDR), as well as mutation-specific *P* values and FDRs.

Results summary

Figure 1A shows selected displays from the CRAVAT Summary tab. This summary includes aggregate results for all mutations from a single submission, at the patient and cohort level. Distribution of submitted mutations by gene function and sequence ontology are displayed as pie charts (Fig. 1A); a pie chart showing distribution by Cancer Genome Landscape definitions (8) is also provided. Chromosome-specific distribution of nonsilent, missense, and inactivating mutations is displayed on a circos plot (Fig. 1A). The results summary also shows the most frequently mutated genes (corrected for gene length) as a bar chart, the top 10 most significant genes as determined by VEST or CHASM composite *P* values, and sample-level consequence-type distributions displayed as a series of stacked histograms (Fig. 1A). Gene set enrichment of collections of genes with significant composite VEST or CHASM *P* values is embedded as biological networks from the Network Data Exchange (NDEx; ref. 9); here, the user can toggle through each network detected for their submission and view significant genes in the context of larger biological networks (Fig. 1A).

Gene- and mutation-level analysis

Figure 1B shows selected displays from the CRAVAT Gene and Variant tabs. The Gene tab displays an interactive spreadsheet, including every gene that harbored a mutation from the submission (truncated example shown at top of Fig. 1B). This table includes gene-specific mutation frequency, the most severe consequence type among submitted mutations for each gene, presence of statistically significant mutation clustering in the related protein [inferred from The Cancer Genome Atlas (TCGA) data, see below], composite VEST and CHASM *P* values and FDRs, classification as a tumor suppressor or oncogene if applicable, associated drugs, COSMIC occurrences, disease status from ClinVar (a database of clinically annotated variants), and Gene Ontology terms. The table can be sorted by any category and filtered by CHASM and VEST significance (see filtering widget, Fig. 1B); any instance of this table can be exported in Excel format. Clicking any row (i.e., gene) in this table retrieves a drill-down table and a lollipop chart. The drill-down table displays information for each mutation from the selected gene, including chromosome position, sequence ontology, protein change, mutation-specific VEST and CHASM *P* values, and the reference and alternative DNA base (s). The lollipop diagram plots all submitted mutations for the selected gene onto the protein sequence, with domain labels and mutation-specific information accessible through the tooltip (mouse-over); the user can optionally choose to plot tissue-specific mutations from TCGA data.

The CRAVAT interactive Variant tab serves information that will be familiar to users after exploring the Gene tab, but here, the user can drill down into the details even further (see for instance, the second spreadsheet in Fig. 1B). This tab includes *P* values and FDRs for each mutation in the context of all relevant transcripts. Frequencies from population databases (1000 Genomes, ref. 10; ESP6500, ref. 7; and ExAC, ref. 11) are graphically displayed and can be further partitioned by ethnicity. Figure 1B shows an example lollipop diagram, where the submitted R230I missense substitution (top side of lollipop diagram) is compared with TCGA kidney cancer data (bottom of lollipop diagram), revealing the nearby D228N substitution in the *ACVR1C* gene.

MuPIT interactive

Figure 1C shows selected displays from Mutation Position Imaging Toolbox (MuPIT) Interactive (12). MuPIT maps mutation positions to annotated, interactive 3D protein structures, and links directly from the gene and mutation entries in the above-mentioned widgets and tables (see for instance, the red pill–shaped button in the bottom right-hand corner of Fig. 1B). MuPIT achieves high coverage by including experimental structures from the protein data bank and homology models filtered using standard measures for model quality. MuPIT was designed for investigators without expertise in protein structure/function and can be used to develop mechanistic hypotheses regarding mutations of interest.

For each structure, the MuPIT Protein tab allows manipulation of graphical parameters, such as color, outline, opacity, and style (cartoon, stick, space fill, etc.), as well as options for displaying specific domains, chains, and ligands; additionally, publication-ready images can be exported from this tab. The Mutations tab allows the user to toggle through submitted mutations alongside statistically significant 3D mutation clusters from TCGA data; statistically significant mutation clusters are precomputed for each of 31 TCGA cancer subtypes using the HotMAPS algorithm (13). In Fig. 1C, MuPIT is located at the Annotations tab, and user-submitted mutations (shown in green) are spatially proximal to known TGFBR1 binding and active sites (cyan and blue), suggesting a potential mechanistic role for these mutations.

CRAVAT 5

The next update of CRAVAT is scheduled for release in 2017 and will include annotations for noncoding mutations, detection of statistically significant sequence hotspots for mutations, updated VEST and CHASM classifiers, inclusion of top-performing impact prediction algorithms from other groups, migration to the current reference human genome (GRCh38), and integration with the Broad Integrated Genomics Viewer. The corresponding MuPIT update will include pathogenic germline mutations, and rare and common germline mutations from healthy populations for optional mapping and display onto protein structure.

In Context with Other Omics Analysis

CRAVAT is developed for easy integration with other software, including programmatic interfaces for web services. This flexibility

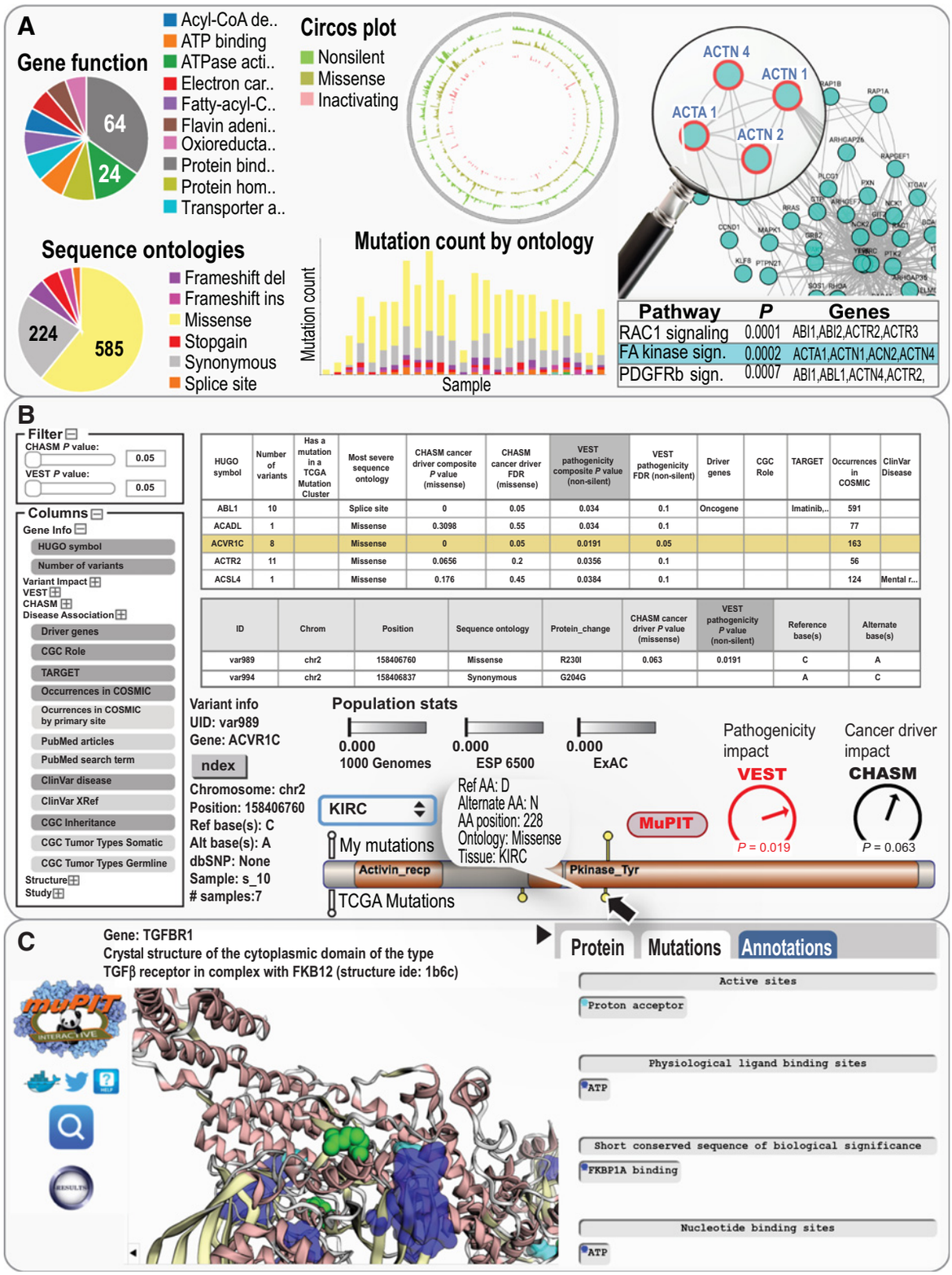

Figure 1.
Selected widgets and displays from the CRAVAT Summary tab (**A**), the Gene and Variant tabs (**B**), and from MuPIT (**C**).

affords the combination of CRAVAT's mutation interpretation capabilities with software considering other omics datatypes (e.g., copy number alterations, methylation, expression). For instance, other methods currently integrating CRAVAT include Trinity, which assembles Illumina RNA-seq data (14); UCSC Xena, which can combine many omics and clinical datatypes (15); and the NDEx, an online commons for biological network data (9). CRAVAT can also be used with Galaxy tools.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: C. Douville, M.C. Ryan, R. Karchin
Development of methodology: C. Douville, M.C. Ryan, R. Karchin
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): C. Tokheim
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): C. Tokheim, R. Bhattacharya, R. Kim, K. Moad, M.C. Ryan
Writing, review, and/or revision of the manuscript: D.L. Masica, C. Douville, C. Tokheim, R. Karchin
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): C. Douville, K. Moad
Study supervision: R. Karchin

Grant Support

This work was supported by NIH, NCI grant U24CA204817-01 to R. Karchin.

Received February 2, 2017; revised March 18, 2017; accepted June 14, 2017; published online November 1, 2017.

References

1. Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, et al. CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics 2013;29:647–8.
2. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res 2009;69:6660–7.
3. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 2013;14:S3.
4. Douville C, Masica DL, Stenson PD, Cooper DN, Gygax DM, Kim R, et al. Assessing the pathogenicity of insertion and deletion variants with the variant effect scoring tool (VEST-Indel). Hum Mutat 2016;37:28–35.
5. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2014;43:D805–11.
6. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, et al. The Human Gene Mutation Database: 2008 update. Genome Med 2009;1:13.
7. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 2013;493:216–20.
8. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science 2013;339:1546–58.
9. Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, et al. NDEx, the network data exchange. Cell Syst 2015;1:302–5.
10. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.
11. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285–91.
12. Niknafs N, Kim D, Kim R, Diekhans M, Ryan M, Stenson PD, et al. MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. Hum Genet 2013;132:1235–43.
13. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. Cancer Res 2016;76:3719–31.
14. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013;8:1494–512.
15. Goldman M, Craft B, Zhu J, Swatloski T, Cline M, Haussler D. Abstract 5270: The UCSC Xena system for integrating and visualizing functional genomics. In: Proceedings of the 107th Annual Meeting of the American Association for Cancer Research; 2016 Apr 16–20; New Orleans, LA. Philadelphia, PA: AACR; 2016. Abstract nr 5270.