Journal of proteome research

Technical Note

Bridging the Chromosome-Centric and Biology and Disease Human Proteome Projects: Accessible and automated tools for interpreting biological and pathological impact of protein sequence variants detected via proteogenomics

Ray Sajulga, Subina Mehta, Praveen Kumar, James E. Johnson, Candace R. Guerrero, Michael C. Ryan, Rachel Karchin, Pratik D. Jagtap, and Timothy J. Griffin

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.jproteome.8b00404 • Publication Date (Web): 21 Aug 2018

ACS Publications

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Bridging the Chromosome-Centric and Biology and Disease Human Proteome Projects: Accessible and automated tools for interpreting biological and pathological impact of protein sequence variants detected via proteogenomics

Ray Sajulga[1], Subina Mehta[1], Praveen Kumar[1,2], James E. Johnson[3], Candace R. Guerrero[1], Michael C. Ryan[4], Rachel Karchin[5,6,7], Pratik D. Jagtap[1], and Timothy J. Griffin[1*]

Affiliations

[1]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN, 55455

[2]Bioinformatics and Computational Biology Program, University of Minnesota-Rochester, Rochester, MN, 55904

[3]Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN, 55455

[4]In-Silico Solutions, Falls Church, VA, 22043

[5]Department of Biomedical Engineering, The Johns Hopkins University, Baltimore, MD, 21218

[6]The Institute for Computational Medicine, The Johns Hopkins University, Baltimore, MD, 21218

[7]Department of Oncology, The Johns Hopkins University School of Medicine, Baltimore, MD, 21217

*To whom correspondence should be addressed:

University of Minnesota
6-155 Jackson Hall
321 Church St SE
Minneapolis, MN 55455
Email: tgriffin@umn.edu
Tel: 612-624-5249

ABSTRACT

The chromosome-centric human proteome project (C-HPP) seeks to comprehensively characterize all protein products coded by the genome, including those expressed sequence variants confirmed via proteogenomics methods. The closely related biology and disease human proteome project (B/D-HPP) seeks to understand the biological and pathological associations of expressed protein products, especially those carrying sequence variants which may be drivers of disease. To achieve these objectives, informatics tools are required that interpret potential functional or disease implications of variant protein sequence detected via proteogenomics. Towards this end, we have developed an automated workflow within the Galaxy for proteomics (Galaxy-P) platform which leverages the Cancer-Related Analysis of Variants Toolkit (CRAVAT) and makes it interoperable with proteogenomic results. Protein sequence variants confirmed by proteogenomics are assessed for potential structure-function effects, as well as associations with cancer using CRAVAT's rich suite of functionalities, including visualization of results directly within the Galaxy user interface. We demonstrate the effectiveness of this workflow on proteogenomic results generated from an MCF7 breast cancer cell line. Our free and open software should enable improved interpretation of functional and pathological effects of protein sequence variants detected via proteogenomics, acting as a bridge between the C-HPP and B/D-HPP.

KEYWORDS

proteogenomics

bioinformatics

cancer

multi-omics

Galaxy-P

CRAVAT

Chromosome-centric human proteome project

Biology and disease human proteome project

INTRODUCTION

The Human Proteome Project (HPP)[1], is an international effort to characterize the protein products expressed from all genes in the human genome, providing a resource to better understand the molecules responsible for function and disease. Underlying the HPP, are two related projects: the chromosome centric HPP (C-HPP)[2] and the biology and disease HPP (B/D-HPP)[3]. The C-HPP seeks to systematically characterize all expressed proteins products and map these to their genomic coding locations on the chromosomes. Key to this effort, is the use of proteogenomics[4-5]. Specifically, "peptide sequence-centric" proteogenomics is a key approach, where genomic, transcriptomic (e.g. RNA-Seq) and mass spectrometry (MS)-based proteomics data are integrated to verify the expression of proteins, including those carrying amino acid sequence variations (e.g. single amino-acid substitutions, small insertions-deletions, frameshifts etc). Once verified as expressed protein products, the B/D-HPP seeks to provide an understanding of their possible functional role, and potential ties to disease.

Informatics tools are required to achieving the goals of the HPP -- from analysis of the raw 'omics data generated for proteogenomics studies, to the interpretation of possible biological and disease effects. As described recently in a thorough review of the proteogenomics field[5], numerous software tools and platforms have been developed for carrying out proteogenomic analyses and verifying sequence variants from MS-based proteomics data. Most of these tools have focused on the challenges related to the integrated processing of raw multi-omics data for proteogenomics (e.g. protein sequence database generation from RNA-Seq data, matching mass spectrometry data to variant sequences, mapping to genomic coordinates, etc). Indeed our group has been extending the open and accessible Galaxy platform[6] to facilitate integration of genomic, transcriptomic and MS-based proteomic data for proteogenomics[7-9], under an initiative called the Galaxy for proteomics (Galaxy-P) project.

Despite these many available software tools for raw data processing in proteogenomics, a relatively small number of choices exist for automated interpretation of functional effects and/or disease relationships of expressed protein sequence variants. Whether or not the variant protein sequence detected is completely novel, or is derived from a known genomic variant, researchers seek answers to many questions about these variants within the context of their biological investigation, such as: How might the variant impact the structure of the protein? Is there prior evidence on this variants being implicated in pathologies, such as cancer? Does the detection of this variant peptide confirm expression of a cancer driver mutation? Answers to these and other related questions are critical for prioritizing these variants, generating hypotheses, and further testing for better understanding the biological roles of these expressed proteins. Some notable examples of available tools for functional impact analysis include computational resources aimed at assessing potential effects of variants at known sites of post-translational modification[10-11], which may help in prioritizing those sequence changes with the most significant effects on function. However, for the most part, researchers are left with the intimidating prospect of manually interpreting their proteogenomics results one-by-one, using existing knowledgebase and literature searches -- a prospect which does not scale well with large datasets.

Fortunately, a possible solution exists for improving results interpretation in proteogenomics. Tailored towards results from next generation sequencing data, the Cancer-Related Analysis of Variants Toolkit (CRAVAT)[12-13] provides a rich suite of tools for interpretation of genomic sequence variants. Using customized algorithms and leveraging existing knowledgebases, CRAVAT predicts the impact of sequence variants, including structure-function consequences at the protein level, as well as known cancer relationships of non-silent variants. CRAVAT also offers a powerful selection of visualizations for

exploring results, including visualization of amino acid variant positions on 3D protein structures using the Mutation position imaging toolbox (MuPIT)[14], and the recent addition of the Network Data Exchange (NDEx) resource[15] for visualizing interaction networks of variant genes and their products. A Galaxy implementation of CRAVAT has also been developed for analysis of genomic and transcriptomic data[13].

Given these capabilities, CRAVAT offers an ideal platform for interpreting amino acid sequence variants detected by proteogenomics. However, to-date CRAVAT lacks compatibility with standard outputs from proteogenomics workflows. To this end, we have leveraged and extended the Galaxy implementation of CRAVAT, enabling interoperability of proteogenomic workflows and outputted results with this powerful suite of interpretation and visualization tools. Deployed within the overarching umbrella of the Galaxy-P project, our workflow extends CRAVAT to recognize results produced in standard formats from proteogenomics analysis. We have also developed a plugin which enables display of CRAVAT visualizations within the Galaxy user-interface, for facile viewing and interpretation of results. We have made all tools available within the Galaxy Tool Shed, providing free and open access to the research community. We also have made available the workflow on a cloud-hosted, publicly available Galaxy instance for demonstration and training purposes. Our work will add valuable components to the tool kit available to researchers seeking to achieve the goals of the HPP, specifically in the impact analysis of protein variants detected via proteogenomic workflows.

EXPERIMENTAL SECTION

Extending the CRAVAT Query Galaxy tool for proteogenomic compatibility

The CRAVAT Query tool was implemented previously in the Galaxy framework. CRAVAT Query utilizes CRAVAT's Representational State Transfer (RESTful) Application Programming Interface (API) to query jobs onto the main server. Only two inputs are required: (1) a tabular input file containing variant

information (e.g. in the variant call format, VCF) and (2) the analysis program to use (Variant Effect Score Toolkit (VEST)[16], and/or Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM)[17]). After analysis on the server, the Galaxy tool retrieves a tabular results file containing the CRAVAT annotations for the coding variants.

In our extension of this CRAVAT Query tool, several features were added to make it compatible with proteogenomics outputs. For example, a new CRAVAT version (5.0) was released that switched Genome Reference Consortium Human (GRCh) builds over to the newer build, hg38 (GRCh38/hg38). To accommodate the older version (GRCh37/hg19), a dropdown for GRCh selection was added within the Galaxy interface for this tool. Furthermore, the tool was improved to retrieve three additional tabular outputs from the CRAVAT server: error feedback, gene annotations, and non-coding variant annotations, in addition to the already available coding variant annotations.

Critical to proteogenomics, we defined the standard output format proBED[18] for proteogenomics as an input for CRAVAT Query in Galaxy. This allows for a focused analysis of variant proteins sequences detected by proteogenomics, by intersecting the VCF input file and the proBED file. The user simply selects the option to intersect the proBED file and the VCF file, and CRAVAT will then return results only for those variants in the VCF file with peptide-level confirmation by proteogenomics. Users can also choose to analyze the entire VCF file, which will then include all variants from the genomic and/or transcriptomic data, including those with peptide-level confirmation.

Once generated, the intersected VCF file containing information on selected variants detected by proteogenomics is submitted to CRAVAT's server for analysis. After CRAVAT performs its server-side analysis and annotation, the genomic locations of variants within the results are initially matched back to genomic locations for peptide sequences contained in the proBED input file. The genomic locations for variants (e.g., chr12:6,561,055) in the CRAVAT output are compared to the genomic ranges of the peptides in the proBED file (e.g. chr12:6,561,014-6,561,056, coding the peptide STGVILANDANAER). Peptides found to contain a CRAVAT-annotated variant within its range are verified via sequence matching. This is performed for each peptide sequence by using Ensembl's REST API service to obtain the protein sequence of the most severe sequence ontology (S.O.) annotation to the transcript annotated by CRAVAT (e.g., ENST00000616948.4). Then, the Ensembl peptide sequence for the transcript is compared to the CRAVAT-annotated variants information, to verify the assigned peptide from proteogenomics matches to the variant analyzed by CRAVAT. In the case of the CRAVAT variant within a gene location not matching the peptide sequences from the proBED file, the process is repeated for other S.O. annotated transcripts within that same region, until a match to the detected peptide sequence is made. Once matched and verified, the peptide sequences and variant amino acid positions are inserted as another column in CRAVAT's tabular variant output file.

A Galaxy plugin for CRAVAT visualizations

In our Galaxy-based extension of CRAVAT for proteogenomic data analysis, we also sought to incorporate the rich variety of visualizations available in CRAVAT[13] directly into the Galaxy interface. To this end, we capitalized on Galaxy's amenability to customized visualizations, thereby avoiding dependence on CRAVAT's server, by creating a visualization plugin that replicates the CRAVAT viewer's functionalities within the Galaxy interface. In order to expand the viewing options within Galaxy for large tabular data, such as the results from CRAVAT analysis of proteogenomic data, we leveraged the freely available jQuery plugin, DataTables.

After initialization, the viewer uses the Galaxy API to first obtain the ID of the selected dataset collection containing results returned from the CRAVAT server, and subsequently extracts the IDs of the tabular CRAVAT outputs. These IDs are used to directly connect the DataTables to the datasets through AJAX (asynchronous JavaScript and XML) loading, which supports display of large datasets. To expedite loading, the input tabular dataset (e.g., 'Variant', 'Genes' etc) for launching the viewer from the Galaxy history will be loaded first, since its ID is available immediately. Once finished, the variants are sorted and displayed by default from most impactful to least via VEST/CHASM p-values.

Like the original CRAVAT viewer, a collapsible sidebar is included with these DataTables that contains all the column headers available for each dataset. Here, users can toggle column visibility. Due to the large amount of information annotated by CRAVAT (the variant table has around 80 columns), default columns are pre-selected for each table.

For more sophisticated visualizations of results, the Galaxy-based viewer takes advantage of a new feature developed by the CRAVAT team: single-variant pages. They are accessed from the CRAVAT server by including genomic position and variant information within their template URL extension (e.g. including text such as the following in the URL "...?variant=chr22_40418496_-_A_G"). In the Galaxy-CRAVAT viewer, users can select a variant from the Variant DataTable to fetch and display its associated page below the table. These pages each contain a protein sequence diagram (using UniProt information), lollipop visualization of the variant in the context of somatic mutations from 33 cancer types (The Cancer Genome Atlas, TCGA[19-20] within that sequence, a three-dimensional interactive protein structure (MuPIT) and a network interaction diagram (NDEx) of the selected variant, if available

for that gene and protein. Additionally, any important annotations are highlighted in red, such as predicted disease phenotypes, population statistics, etc. For the summary visualizations, specific columns from the Variant and Gene datasets are retrieved from the datasets once their IDs are obtained. The data are manipulated for input into NVD3 charts (i.e., pie, bar) and a BioCircos.js plot[21].

Generation of demonstration proteogenomic input data

For demonstration and testing purposes, data from an MCF-7 breast cancer-derived cell line was used. The MCF-7 whole cell lysate was divided and a portion processed for transcriptome analysis by RNA-Seq (Paired End, 220 million reads); another portion was processed for MS-based proteomic analysis, where isolated proteins were separated using a denaturing polyacrylamide gel, followed by excision of molecular weight regions and digestion of proteins with trypsin. Selected molecular weight regions were analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS) using an Orbitrap Fusion mass spectrometer.

The RNA-Seq and MS-based proteomics data were analyzed using an adapted version of an established proteogenomics workflow in Galaxy-P[8]. This workflow generates as outputs standard formatted results which can act as input for the extended CRAVAT Query tool. **Supporting Material S1** provides more details on the generation of the RNA-Seq and MS-based proteomics data used for demonstration purposes. Briefly, the workflow is divided into three parts: protein sequence database generation using RNA-Seq data, database searching (matching MS/MS spectra to peptide sequences) and verification of detected peptide sequence variants. Inputs for this workflow include raw RNA-Seq paired-end data (.FASTQ) along with a genomic annotation file (.GTF), which are analyzed by a series of tools to identify and assemble potential sequence variants from these data. As part of this workflow, the Freebayes tools is used which generates a standard variant call format (.VCF) file, providing a summary of all

potential variants identified from the starting RNA-Seq data. CustomProDB[22] utilizes the outputs from the RNA-Seq analysis to create a protein sequence database (.FASTA format), including proteins containing sequence variants such as single amino acid variants (SAVs) and Insertion/Deletions (InDels). MS/MS spectra are matched against sequences within this FASTA database, assessed for quality and matches to peptide sequence variants are verified. The variant peptide sequences matched to MS/MS spectra are outputted, and along with genomic coordinates, used to create a Browser Extensible Data (BED) file, which also contains peptide sequence information along with genomic coordinates for these expressed gene products (proBED format)[18]. The proBED file, along with the VCF file created in this workflow, act as inputs for the extended CRAVAT Query Galaxy tool.

RESULTS AND DISCUSSION

Figure 1 provides an overview of the Galaxy-based workflow developed here, capable of taking peptide sequence variants detected using proteogenomics, performing impact analysis of these variants using the suite of CRAVAT tools, and visualizing results within the Galaxy interface. To create this workflow, we leveraged the already existing Galaxy-based CRAVAT Query tool. The initial implementation of this tool read gene or transcript variant data presented in a variety of specified formats, including the community-standard variant call format (VCF). We extended the CRAVAT Query tool in two important ways, to enable compatibility with data outputted from a proteogenomics workflow. First, we defined a new data type as input for CRAVAT, specifically proBED formatted data[18]. The proBED file is generated from results from upstream proteogenomics analysis[8]. This file contains verified peptide sequence variants (e.g. single amino acid variants, InDels) confirmed via matching of tandem mass spectrometry (MS/MS) data to a protein sequence database generated from RNA-Seq transcriptome assembly. Genomic coordinates for the coding regions of each peptide is also provided in the proBED file. ProBED has been recommended by the Human Proteome Organization Standards Initiative as the preferred format for results reporting from proteogenomic analyses[18].

Our second extension of the CRAVAT Query tool merges proteomic and genomic data to selectively analyze only those RNA-Seq variants with peptide-level confirmation of their expression. Here, CRAVAT Query was extended to carry-out an intersection between genomic coordinates for detected peptide variants contained in the proBED file, and the VCF file generated from the RNA-Seq assembly and analysis. For this intersection analysis, CRAVAT Query reads both the proBED and VCF input files to the Galaxy tool (**Figure 2,** red shaded region). After this intersection mapping, CRAVAT only analyzes variants which have been detected at the peptide-level. Users also have the option of analyzing the entire VCF file, where all variants will be analyzed by CRAVAT including those with peptide-level confirmation. The sortable outputs described below allow the user to selectively sort and view data for those variants with peptide detection if desired.

With peptide variants of interest now defined, executing the CRAVAT Query tool relays the variants to the main CRAVAT server for analysis, making use of its RESTful API. Choices for analysis are given to the user (**Figure 2**), including analysis using Variant Effect Score Toolkit (VEST)[16] for predicting pathogenic effects of a variant, and/or Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM)[17] for predicting functional significance of somatic missense variants. Once analyzed, a set of standard CRAVAT output files are returned, appearing as a dataset collection of results within the active Galaxy History (bottom of Figure 2). These results can be downloaded as tabular files, if desired.

Visualization of results is the final critical piece of the overall workflow. Fortunately, CRAVAT has been designed with web-based visualizations, providing users an interactive environment to explore and interpret results. We sought to automatically display these visualizations within the Galaxy interface,

providing users a single portal for generation and viewing of results. To facilitate this visualization functionality, we developed a Galaxy visualization plugin tool. This plugin can be started by simply selecting the visualization button which appears within the "Variants" output item within the active History. Once initiated, the visualization opens the CRAVAT interactive results viewer directly within the center pane of the Galaxy interface.

Figure 3 provides an overview of the rich set of CRAVAT visualizations available through the Galaxy-plugin. Here, we highlight a portion of the available CRAVAT visualizations. A more in-depth description of these visualizations can be found in the online documentation for CRAVAT http://cravat.us/CRAVAT/help.jsp?chapter=help_report&article=top). Focusing on the view enabled by selecting the "Variants" tab, the plugin provides an overview of the analysis results for peptide sequence variants detected via the proteogenomics workflow. A sortable table of variants is presented, which lists all of the variant sequences analyzed, along with annotation and metrics from the CRAVAT analysis. We have extended this table to include the variant peptide sequences confirmed using proteogenomics and mapping to genomic variants, to complement the other information already available within this output.

Clicking on the row for any peptide sequence in the variant table displays an additional panel of information on the nature of that variant. The full detailed view for an example variant is shown in **Figure 3**. Depending on the analysis carried out, information on pathogenecity impact (VEST analysis) or cancer driver impact (CHASM) is displayed, along with population stats. Information on frequency of occurrence for known variants contained in databases such as dbSNP or Catalogue of Somatic Mutations in Cancer (COSMIC) is given for the selected variant; if the variant has not been characterized previously CRAVAT returns a zero value for known occurrences. Additionally, and highly relevant to

proteogenomic studies, a "lollipop" diagram of the full protein sequence for the reference genomic coding sequence is shown, with the option of displaying the location of the detected sequence variant within the overall sequence. The protein sequence view also shows location of other variants -- either those contained within the active data being analyzed or known variants from TCGA. For those protein products with known structures archived in the Protein Database (PDB), users can click on the MuPIT[14] button to display a three-dimensional structure, along with color coded display of the location of the sequence variant within the known structure (see bottom of **Figure 3**). Lastly, CRAVAT also queries the NDEx resource[15], returning a network diagram of known interactions for the coding gene in question.

To guide the implementation of these Galaxy-based extensions of the CRAVAT software, we utilized data and results from a proteogenomic analysis of an MCF7 breast cancer cell line, using a modification of a previously described workflow[8]. **Figure 3** shows representative results from the analysis of a selected portion of this proteogenomics data, where a small number of example variant peptide sequences (in proBED format), along with the genomic variant information from the RNA-Seq data (in VCF format) were used as a small input dataset for demonstration purposes. As an example, results for the CRAVAT analysis of one of these variant peptide sequences is shown, a missense mutation from the heat shock 90 (HSP90) protein.

For demonstration and training purposes, we have made available example input data and the data analysis workflow for integrative proteogenomic-CRAVAT analysis on a publicly available Galaxy instance. **Supporting Material S1** provides instructions on how to access and use this instance for demonstration purposes. Here, users can run the workflow and explore the rich array of visualization options available within the Galaxy plugin of the CRAVAT viewer.

Collectively, the informatic resources described here provide a bridge between the objectives of the C-HPP and the B/D-HPP. Proteogenomics, commonly conducted through integration of RNA-Seq and MS-based proteomics, provides a powerful means to confirm the expression of stable protein products carrying sequence variants derived from genomic mutation and/or unexpected RNA-level events (e.g. splice isoforms). These confirmed variant peptide sequences provide valuable new information towards the C-HPP goal of cataloging all known protein products expressed across chromosomes. By making CRAVAT interoperable with outputs from proteogenomic analyses, we provide a means for the automated interpretation of potential functional and pathological effects of detected protein variants -- aligning directly with the goals of the B/D-HPP. Researchers will able to better ascertain whether a peptide sequence variant is derived from a well-annotated genomic mutation, with known pathological correlation, or if it is highly novel. Viewing the variant within the three dimensional protein structure (if available) using MuPIT, as well as known interaction networks of the gene and its products using NDEx, will provide researchers further clues into possible functional effects due to the amino acid sequence change.

The Galaxy-based implementation of this platform provides a number of advantages as well. First, the Galaxy-P team, along with others, have demonstrated the value of the Galaxy platform for developing and disseminating sophisticated proteogenomic workflows[7-9, 23-25]. Galaxy provides a workbench well-suited for such applications requiring integration of many disparate software programs from across 'omic domains. Importantly, our extension of CRAVAT utilizes the recommended community standard output for proteogenomics, proBED[18], which can be produced from a Galaxy-based proteogenomics workflow (e.g. see[8]), or could be uploaded from an alternate workflow of the users choice. As such, the

CRAVAT implementation described here is agnostic to the upstream source of the proteogenomic data that acts as input. The MS-based proteomics data could come from any instrument vendor, most commonly in the form of MS/MS data collected in data-dependent mode; however, data-independent acquisition could also be used in conjunction with specific software[26] which can re-construct peptide MS/MS spectra for matching against a sequence database. The CRAVAT tool assumes that upstream software for matching MS/MS data to variant sequences has been employed in a manner such that the peptide spectral matches (PSMs) have been rigorously filtered to ensure confident sequence matches at acceptable false discovery rate. We have previously described Galaxy tools for this upstream analysis, including tools to view and assess PSM quality and also verify the novel nature of putative peptide sequence variants[9].

A second advantage is the scalability and automation offered by Galaxy for the analysis of a large number of variants contained within the proBED, and VCF inputs. Lastly, Galaxy implementation provides a means for disseminating the tools and workflows, in a transparent and complete manner. Tools are accessible via the Galaxy Tool Shed, for implementation on any local Galaxy instance. For demonstration and training purposes, the workflow and example input data are also available on a publicly available instance (see **Supporting Material S1** for instructions).

Although this initial integration of proteogenomics results and CRAVAT analysis provides powerful functionalities which researchers will find immediately useful, there are a number of possibilities for further development. For example, we are currently developing a Galaxy plugin for “protein-centric” visualization of MS-based proteogenomic results from Galaxy-based workflows, enabling users to view quality of MS/MS matches to variant sequences, assess protein coverage, and map peptide sequences

to genomic locations. One could envision using such a tool to further filter and select variant peptides of highest quality and interest, sending these sequences back to the Galaxy instance for submission to CRAVAT for further analysis. An additional protein level analysis could also involve retrieval of known sites of post-translational modification[10-11] within the CRAVAT software, enabling users to assess and visualize the possible effect of protein sequence changes on functionally important sites of modification.

CONCLUSIONS

In conclusion, we have enabled a Galaxy-based workflow for integrating peptide sequence variants, detected via proteogenomics, with the CRAVAT analysis suite. This integration provides an accessible platform for interpreting the impact, both functional and pathological, of expressed protein variants resulting from genomic mutation or other gene expression regulatory events. The open and freely available software are flexible and extensible, and offer potential for further extensions and customization of functionalities as new requirements and analysis tools emerge. These tools should provide the research community a valuable informatics resource, bridging the goals of the C-HPP and B/D-HPP.

SOFTWARE AVAILABILITY

The extended CRAVAT Query tool and CRAVAT visualization plugin is available in the Galaxy Tool Shed (https://toolshed.g2.bx.psu.edu/view/galaxyp/cravatool/83181dabeb90). These tools are also available on Github (https://github.com/galaxyproteomics/tools-galaxyp/tree/master/tools/cravatool). **Supporting Material S1** also provides instructions on accessing demonstration input data and the workflow on a publicly available Galaxy instance. A Docker container containing a Galaxy instance, tools, workflow and demonstration data has also been created. Instructions for access and install can be found at z.umn.edu/gpcravatdocker.

ACKNOWLEDGEMENTS

The authors acknowledge funding for this work from the Informatics Technologies for Cancer Research (ITCR) program at the NIH/NCI, from grant U24CA204817 to R. Karchin and grant U24CA199347 to T. Griffin. The authors also acknowledge support from the Center for Mass Spectrometry and Proteomics and the University of Minnesota Genome Center for assistance generating demonstration proteogenomic data. The authors also acknowledge use of the Jetstream cloud-based computing resource for scientific computing (https://jetstream-cloud.org/) maintained at Indiana University for assistance in maintaining the publicly available Galaxy instance used for demonstration purposes.

SUPPORTING INFORMATION

The following supporting information is available free of charge at ACS website http://pubs.acs.org

Supporting Material S1: Instructions for accessing and utilizing the demonstration data and workflow for proteogenomic-CRAVAT analysis on a publicly available Galaxy instance.

REFERENCES CITED

1. Omenn, G. S.; Lane, L.; Lundberg, E. K.; Overall, C. M.; Deutsch, E. W., Progress on the HUPO Draft Human Proteome: 2017 Metrics of the Human Proteome Project. *J Proteome Res* **2017,** *16* (12), 4281-4287.
2. Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S., The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat Biotechnol* **2012,** *30* (3), 221-3.
3. Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J. E.; Kussmann, M.; Qin, J.; Omenn, G. S., The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J Proteome Res* **2013,** *12* (1), 23-7.
4. Nesvizhskii, A. I., Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **2014,** *11* (11), 1114-25.
5. Ruggles, K. V.; Krug, K.; Wang, X.; Clauser, K. R.; Wang, J.; Payne, S. H.; Fenyo, D.; Zhang, B.; Mani, D. R., Methods, Tools and Current Perspectives in Proteogenomics. *Mol Cell Proteomics* **2017,** *16* (6), 959-981.
6. Afgan, E.; Baker, D.; van den Beek, M.; Blankenberg, D.; Bouvier, D.; Cech, M.; Chilton, J.; Clements, D.; Coraor, N.; Eberhard, C.; Gruning, B.; Guerler, A.; Hillman-Jackson, J.; Von Kuster, G.; Rasche, E.; Soranzo, N.; Turaga, N.; Taylor, J.; Nekrutenko, A.; Goecks, J., The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* **2016,** *44* (W1), W3-W10.
7. Boekel, J.; Chilton, J. M.; Cooke, I. R.; Horvatovich, P. L.; Jagtap, P. D.; Kall, L.; Lehtio, J.; Lukasse, P.; Moerland, P. D.; Griffin, T. J., Multi-omic data analysis using Galaxy. *Nat Biotechnol* **2015,** *33* (2), 137-9.
8. Chambers, M. C.; Jagtap, P. D.; Johnson, J. E.; McGowan, T.; Kumar, P.; Onsongo, G.; Guerrero, C. R.; Barsnes, H.; Vaudel, M.; Martens, L.; Gruning, B.; Cooke, I. R.; Heydarian, M.; Reddy, K. L.; Griffin, T. J., An Accessible Proteogenomics Informatics Resource for Cancer Researchers. *Cancer Res* **2017,** *77* (21), e43-e46.
9. Jagtap, P. D.; Johnson, J. E.; Onsongo, G.; Sadler, F. W.; Murray, K.; Wang, Y.; Shenykman, G. M.; Bandhakavi, S.; Smith, L. M.; Griffin, T. J., Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework. *J Proteome Res* **2014,** *13* (12), 5898-908.
10. Hornbeck, P. V.; Zhang, B.; Murray, B.; Kornhauser, J. M.; Latham, V.; Skrzypek, E., PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **2015,** *43* (Database issue), D512-20.
11. Keegan, S.; Cortens, J. P.; Beavis, R. C.; Fenyo, D., g2pDB: A Database Mapping Protein Post-Translational Modifications to Genomic Coordinates. *J Proteome Res* **2016,** *15* (3), 983-90.
12. Douville, C.; Carter, H.; Kim, R.; Niknafs, N.; Diekhans, M.; Stenson, P. D.; Cooper, D. N.; Ryan, M.; Karchin, R., CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics* **2013,** *29* (5), 647-8.
13. Masica, D. L.; Douville, C.; Tokheim, C.; Bhattacharya, R.; Kim, R.; Moad, K.; Ryan, M. C.; Karchin, R., CRAVAT 4: Cancer-Related Analysis of Variants Toolkit. *Cancer Res* **2017,** *77* (21), e35-e38.
14. Niknafs, N.; Kim, D.; Kim, R.; Diekhans, M.; Ryan, M.; Stenson, P. D.; Cooper, D. N.; Karchin, R., MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. *Hum Genet* **2013,** *132* (11), 1235-43.
15. Pratt, D.; Chen, J.; Welker, D.; Rivas, R.; Pillich, R.; Rynkov, V.; Ono, K.; Miello, C.; Hicks, L.; Szalma, S.; Stojmirovic, A.; Dobrin, R.; Braxenthaler, M.; Kuentzer, J.; Demchak, B.; Ideker, T., NDEx, the Network Data Exchange. *Cell Syst* **2015,** *1* (4), 302-305.

16. Carter, H.; Douville, C.; Stenson, P. D.; Cooper, D. N.; Karchin, R., Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **2013,** *14 Suppl 3*, S3.
17. Carter, H.; Chen, S.; Isik, L.; Tyekucheva, S.; Velculescu, V. E.; Kinzler, K. W.; Vogelstein, B.; Karchin, R., Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* **2009,** *69* (16), 6660-7.
18. Menschaert, G.; Wang, X.; Jones, A. R.; Ghali, F.; Fenyo, D.; Olexiouk, V.; Zhang, B.; Deutsch, E. W.; Ternent, T.; Vizcaino, J. A., The proBAM and proBed standard formats: enabling a seamless integration of genomics and proteomics data. *Genome Biol* **2018,** *19* (1), 12.
19. Bailey, M. H.; Tokheim, C.; Porta-Pardo, E.; Sengupta, S.; Bertrand, D.; Weerasinghe, A.; Colaprico, A.; Wendl, M. C.; Kim, J.; Reardon, B.; Ng, P. K.; Jeong, K. J.; Cao, S.; Wang, Z.; Gao, J.; Gao, Q.; Wang, F.; Liu, E. M.; Mularoni, L.; Rubio-Perez, C.; Nagarajan, N.; Cortes-Ciriano, I.; Zhou, D. C.; Liang, W. W.; Hess, J. M.; Yellapantula, V. D.; Tamborero, D.; Gonzalez-Perez, A.; Suphavilai, C.; Ko, J. Y.; Khurana, E.; Park, P. J.; Van Allen, E. M.; Liang, H.; Group, M. C. W.; Cancer Genome Atlas Research, N.; Lawrence, M. S.; Godzik, A.; Lopez-Bigas, N.; Stuart, J.; Wheeler, D.; Getz, G.; Chen, K.; Lazar, A. J.; Mills, G. B.; Karchin, R.; Ding, L., Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **2018,** *173* (2), 371-385 e18.
20. Ellrott, K.; Bailey, M. H.; Saksena, G.; Covington, K. R.; Kandoth, C.; Stewart, C.; Hess, J.; Ma, S.; Chiotti, K. E.; McLellan, M.; Sofia, H. J.; Hutter, C.; Getz, G.; Wheeler, D.; Ding, L.; Group, M. C. W.; Cancer Genome Atlas Research, N., Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **2018,** *6* (3), 271-281 e7.
21. Cui, Y.; Chen, X.; Luo, H.; Fan, Z.; Luo, J.; He, S.; Yue, H.; Zhang, P.; Chen, R., BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. *Bioinformatics* **2016,** *32* (11), 1740-2.
22. Wang, X.; Zhang, B., customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **2013,** *29* (24), 3235-7.
23. Crappe, J.; Ndah, E.; Koch, A.; Steyaert, S.; Gawron, D.; De Keulenaer, S.; De Meester, E.; De Meyer, T.; Van Criekinge, W.; Van Damme, P.; Menschaert, G., PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res* **2015,** *43* (5), e29.
24. Fan, J.; Saha, S.; Barker, G.; Heesom, K. J.; Ghali, F.; Jones, A. R.; Matthews, D. A.; Bessant, C., Galaxy Integrated Omics: Web-based Standards-Compliant Workflows for Proteomics Informed by Transcriptomics. *Mol Cell Proteomics* **2015,** *14* (11), 3087-93.
25. Pang, C. N.; Tay, A. P.; Aya, C.; Twine, N. A.; Harkness, L.; Hart-Smith, G.; Chia, S. Z.; Chen, Z.; Deshpande, N. P.; Kaakoush, N. O.; Mitchell, H. M.; Kassem, M.; Wilkins, M. R., Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. *J Proteome Res* **2014,** *13* (1), 84-98.
26. Tsou, C. C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A. C.; Nesvizhskii, A. I., DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* **2015,** *12* (3), 258-64, 7 p following 264.

FIGURE LEGENDS

Figure 1. Overview of Galaxy-based integration of proteogenomics results with the CRAVAT Query Tool and the Galaxy-based plugin for visualization of results.

Figure 2. Screenshot of Galaxy user interface for the extended CRAVAT Query Tool. Red shaded region highlights the extensions developed to make the tool compatible with proteogenomics results.

Figure 3. Screenshot of example CRAVAT results visualization as viewed through the Galaxy plugin tool. **A)** The Results Viewing Window which opens in the main viewing pane of the Galaxy user interface. The left column (**i**) provides options for the user to select information to include in the results table. The middle portion of the window (**ii**) provides basic information on any selected variant, such as known occurrences from databases such as dbSNP and COSMIC, as well as information on disease associations; **B)** Magnified view of the interactive results table. Here, any given variant can be selected for further visualization and information. In this example, a variant from the HSP90 gene is selected. The results table also includes information on the nature of the variant (Protein Sequence Change column), which classifies the variant as an SAV or InDel. The Reference base(s) and Alternate base(s) columns also provide information on the change to the nucleic acid sequence driving the sequence change in the protein. The Variant Peptide column shows the variant peptide sequence detected and confirmed by mass spectrometry-based proteogenomics (if available). The CHASM and/or VEST p-values for each variant are also provided, if available; **C)** The Protein Diagram shows a linear diagram of the protein sequence for the selected gene, creating a lollipop diagram with locations of sequence variants found in the current dataset, as well as other variant sites, if available, from the TCGA database. Mousing over these variant positions opens up information windows. Information on known protein domains and also visualization of known variants from different tissue types are also available in this view; **D)** The interactive NDex network visualization for the selected gene is shown, providing information on

interactions with other genes and gene products, as well as options for selecting related networks containing the gene of interest, based on function; **E)** The three-dimensional (3D) view of the selected protein structure (a magnified portion of the HSP90 structure is shown for this example). The variant amino acid residues are colored green in this 3D structure. Mousing over any location in the structure provides information on the amino acid identities within that region of the protein.

Figure 1

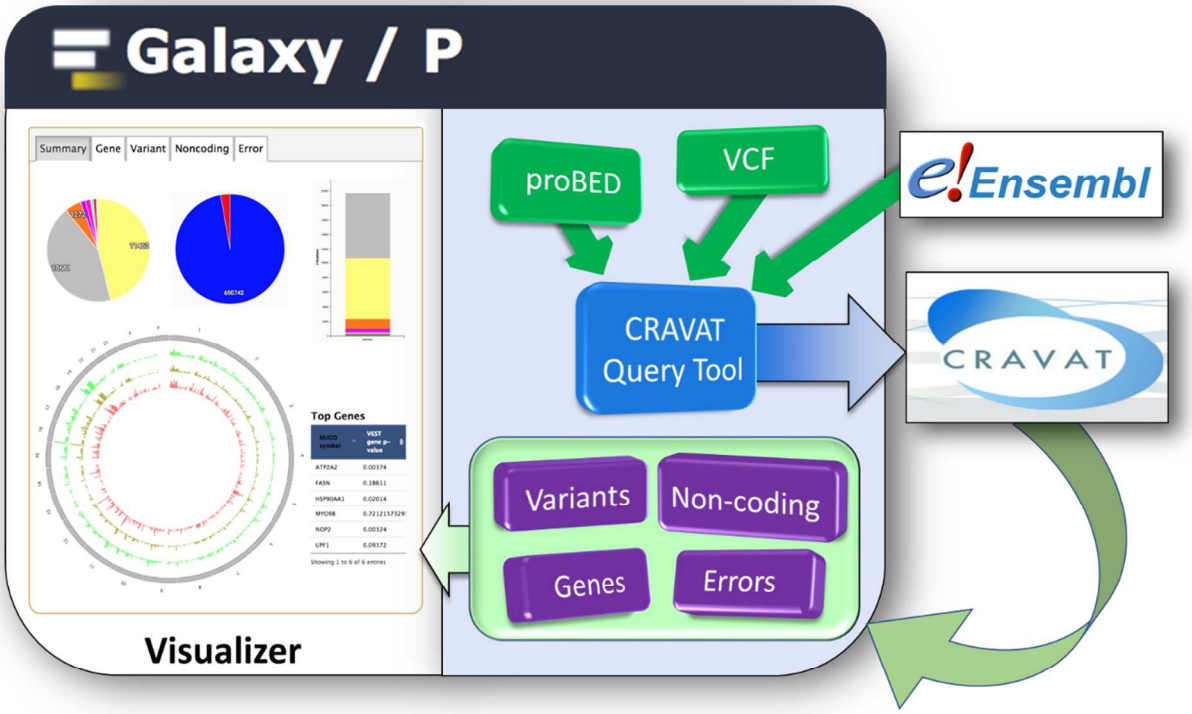

Galaxy / P
Summary
Gene
Variant
Noncoding
Error
Top Genes
Visualizer
proBED
VCF
Ensembl
CRAVAT Query Tool
CRAVAT
Variants
Non-coding
Genes
Errors

Figure 2

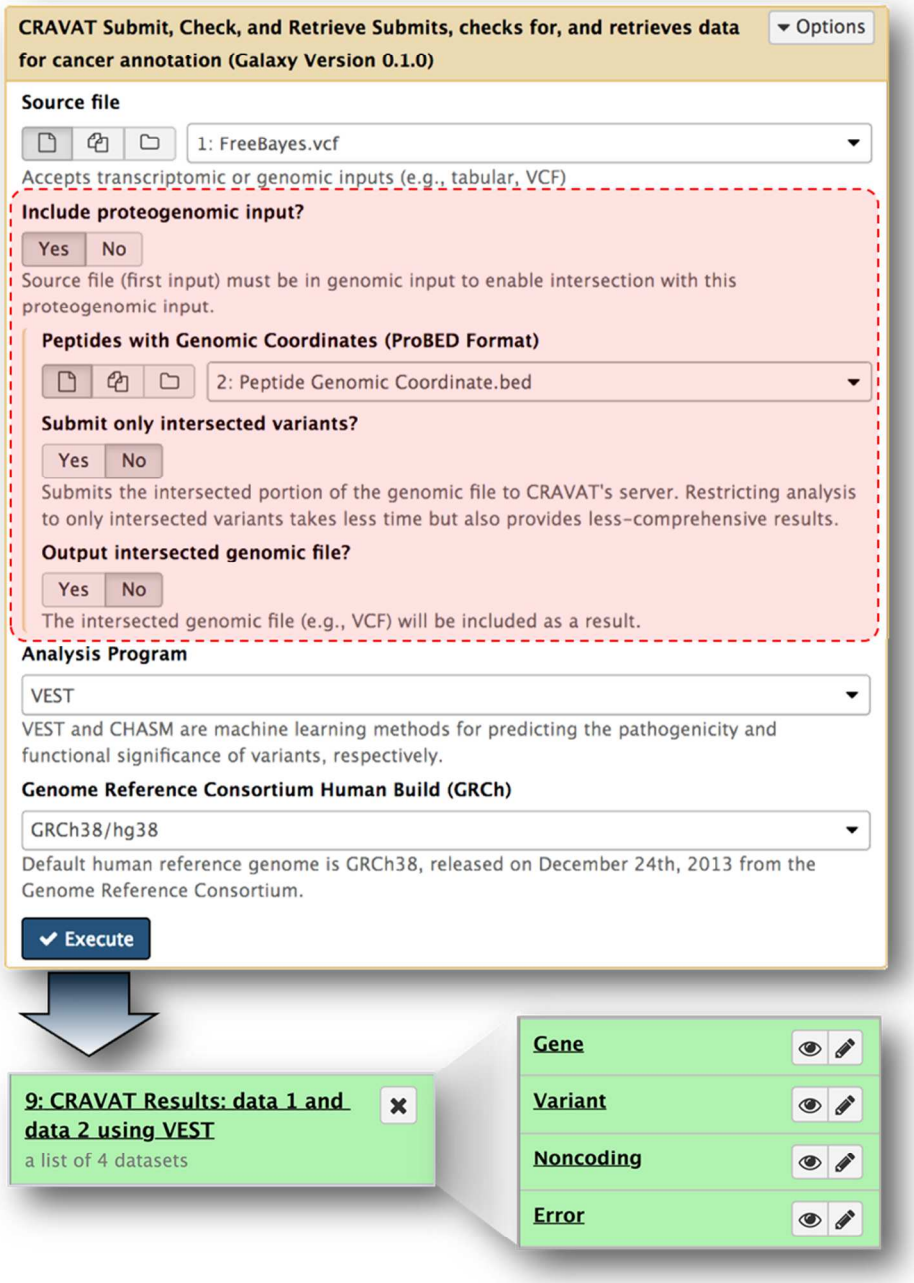
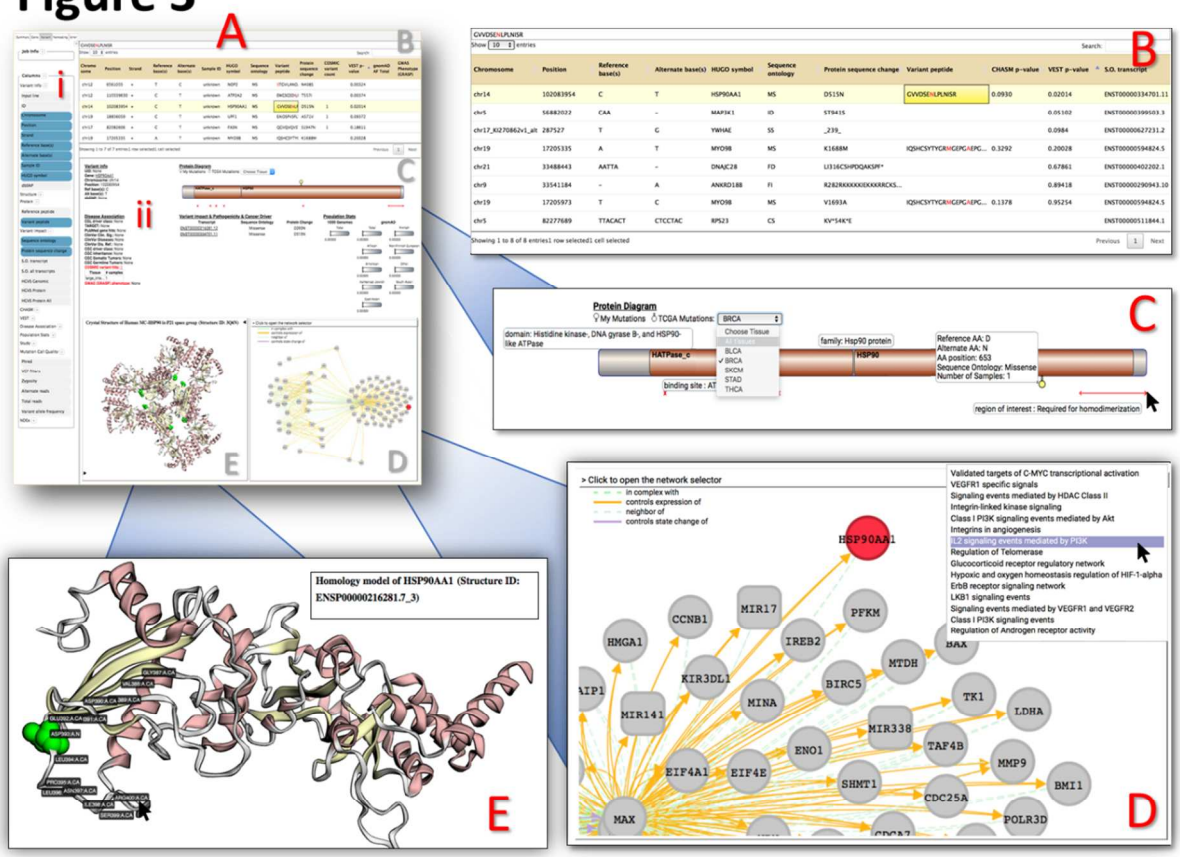
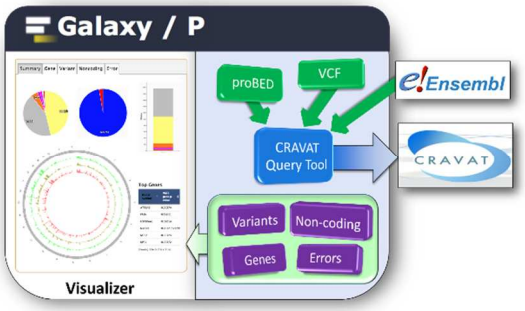

CRAVAT Submit, Check, and Retrieve Submits, checks for, and retrieves data for cancer annotation (Galaxy Version 0.1.0)
Options
Source file
1: FreeBayes.vcf
Accepts transcriptomic or genomic inputs (e.g., tabular, VCF)
Include proteogenomic input?
Yes
No
Source file (first input) must be in genomic input to enable intersection with this proteogenomic input.
Peptides with Genomic Coordinates (ProBED Format)
2: Peptide Genomic Coordinate.bed
Submit only intersected variants?
Yes
No
Submits the intersected portion of the genomic file to CRAVAT's server. Restricting analysis to only intersected variants takes less time but also provides less-comprehensive results.
Output intersected genomic file?
Yes
No
The intersected genomic file (e.g., VCF) will be included as a result.
Analysis Program
VEST
VEST and CHASM are machine learning methods for predicting the pathogenicity and functional significance of variants, respectively.
Genome Reference Consortium Human Build (GRCh)
GRCh38/hg38
Default human reference genome is GRCh38, released on December 24th, 2013 from the Genome Reference Consortium.
Execute
9: CRAVAT Results: data 1 and data 2 using VEST
a list of 4 datasets
Gene
Variant
Noncoding
Error

Figure 3


A
B
C
D
E
i
ii
Protein Diagram
My Mutations
TCGA Mutations:
BRCA
Choose Tissue
BLCA
BRCA
BXCM
STAD
THCA
domain: Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase
HATPase_c
HSP90
family: Hsp90 protein
binding site : AT
Reference AA: D
Alternate AA: N
AA position: 653
Sequence Ontology: Missense
Number of Samples: 1
region of interest : Required for homodimerization
Homology model of HSP90AA1 (Structure ID: ENSP00000216281.7_3)
> Click to open the network selector
in complex with
controls expression of
neighbor of
controls state change of
Validated targets of C-MYC transcriptional activation
VEGFR1 specific signals
Signaling events mediated by HDAC Class II
Integrin-linked kinase signaling
Class I PI3K signaling events mediated by Akt
Integrins in angiogenesis
IL2 signaling events mediated by PI3K
Regulation of Telomerase
Glucocorticoid receptor regulatory network
Hypoxic and oxygen homeostasis regulation of HIF-1-alpha
ErbB receptor signaling network
LKB1 signaling events
Signaling events mediated by VEGFR1 and VEGFR2
Class I PI3K signaling events
Regulation of Androgen receptor activity
HSP90AA1
CCNB1
MIR17
PFKM
HMGA1
IREB2
BAX
KIR3DL1
MTDH
AIP1
BIRC5
TK1
MIR141
MINA
LDHA
MIR338
ENO1
TAF4B
EIF4A1
EIF4E
MMP9
SHMT1
BMI1
CDC25A
MAX
POLR3D

For TOC only


Galaxy / P
proBED
VCF
Ensembl
CRAVAT Query Tool
CRAVAT
Variants
Non-coding
Genes
Errors
Visualizer