CRAVAT (Cancer-Related Analysis of Variants Toolkit)

Integration into Galaxy-P and extension towards proteogenomic studies

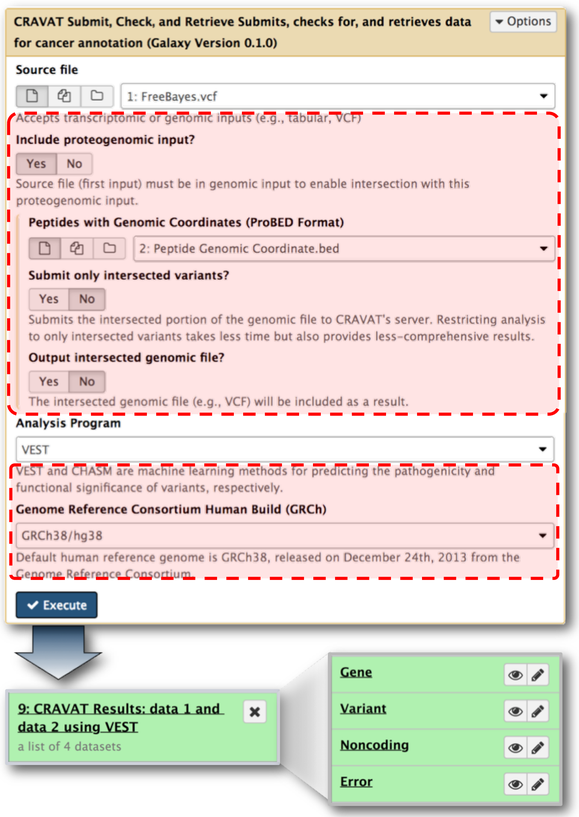
Ray Sajulga[1], Subina Mehta[1], Praveen Kumar[1,2], James E. Johnson[3], Candace Guerrero[1], Michael Ryan[4], Rachel Karchin[5,6,7], Pratik D. Jagtap[1], and Timothy J. Griffin[1]

[1]University of Minnesota, Biochemistry, Molecular Biology and Biophysics; [2]University of Minnesota-Rochester, Bioinformatics and Computational Biology Program; [3]University of Minnesota, Minnesota Supercomputing Institute; [4]In Silico Solutions, LLC; [5]Johns Hopkins University, Department of Biomedical Engineering; [6]Johns Hopkins University, The Institute for Computational Medicine; [7]Johns Hopkins University School of Medicine, Department of Oncology

Introduction

- CRAVAT - an existing tool suite developed by the Karchin lab at Johns Hopkins University and *In Silico Solutions*
- Compiles variant information:
  - Literature/knowledgebase annotation:
    - PubMed, GeneCards, ClinVar, COSMIC, gnomAD, CGC, etc…
  - Harnesses machine learning algorithms for predicting cancer impact and pathogenicity
  - Visualizes variant location on 3-D protein models
  - Displays interaction gene networks (NDeX)
  - Offers extensibility for other –omics analyses:
    - Docker image
    - RESTful API

Galaxy Tool



Parameters:
1. Input file (CRAVAT or VCF format)
2. Proteogenomic input
2a. Submit only intersected variants to CRAVAT
2b. Output the intersected genomic file
3. Analysis Programs (machine learning predicting algorithms):
- VEST (Variant Effect Scoring Tool) – evaluates variants based on pathogenicity
- CHASM (Cancer-specific High-throughput Annotation of Somatic Mutations)
4. Genome Reference Consortium Human Build

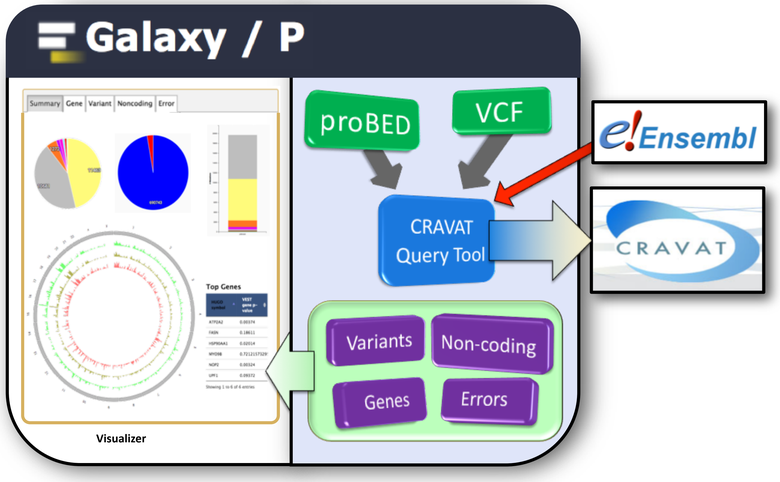
** New parameters in red*

Methodology

ID	Chr.	Position	Strand	Ref. base	Alt. base
TR1	chr14	94079142	-	G	T
TR2	chr10	121520166	-	G	A
TR3	chr13	48459831	+	C	A
TR4	chr7	116777451	+	G	T
TR5	chr7	140753336	-	T	A
TR6	chr17	39724745	+	G	T
Ins1	chr17	39724745	+	-	T
Del1	chr17	39724745	+	A	-
CSub1	chr2	39644095	+	ATGCT	GA

Chr.	Start	End	Peptide	Strand
chr14	94079127	94079178	ADVSAWKDLFVPGPVLR	-
chr10	121520124	121520166	EAIDSPVSFLVLHNQJR	-
chr13	48459831		EWGSGSDILR	+
chr7	116777451		GVVDSENLPLNISR	+
chr7	140753336		IQSHCSYTYGRMGEPGAEP GHFGVCVDSLTSDK	-

- The updated Galaxy tool performs an intersection between the genomic (VCF) and proteogenomic (proBED) file to find which variants were shown to be expressed (via peptides).
- Utilizes the S.O. (sequence ontology) transcript IDs annotated by CRAVAT to obtain protein sequences from Ensembl for amino acid sequence confirmation (and alters the reference sequence based on the input mutations)



Future Directions

- Integrate peptides into the circos plot as a "pircos" plot
- Support interpretation of more complex variants (e.g., alternative splicing)
- Include the user's e-mail within the CRAVAT API submission (may allow for notification of job completion and a unique job ID
- Allow for exporting selected variants back into the Galaxy workspace
- Establish interactivity with our Multi-omics Visualization Plugin (MVP)
- Explore using a Galaxy Interactive Environment through CRAVAT's Docker image

Visualization Plugin

A.) Sortable and filterable variants table with column visibility toggled in the sidebar

B.) Gene, positional, and disease-related information about the variant.

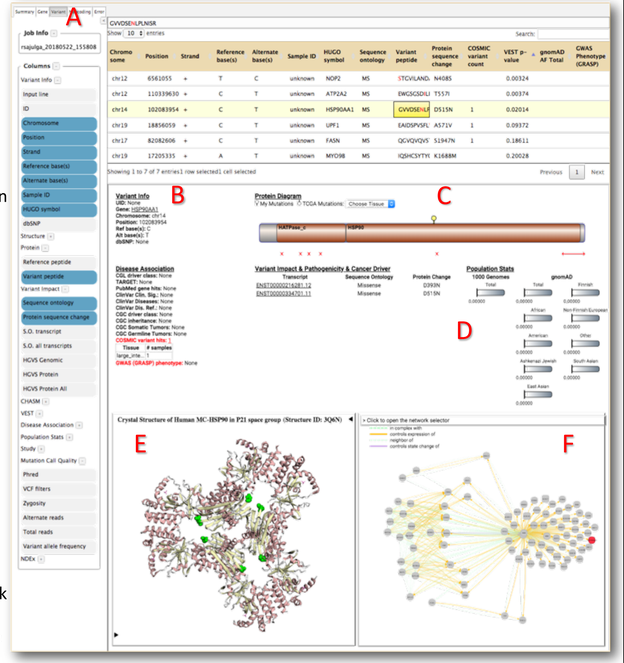
C.) Protein diagram that shows domains and known variants from TCGA (The Cancer Genome Atlas) displayed as a lollipop diagram.

D.) Bar meters displaying allele frequencies from 1000 Genomes, ESP6500, and gnomAD.

E.) Mapping of variant amino acid positions onto three-dimensional structure of protein using MuPIT (Mutational Position Imaging Toolbox)

F.) Visualization of network of known interactions for the variant gene and gene product using NDEx (Network Data Exchange)

CRAVAT Viewer:



Conclusions

- We have enabled a Galaxy-based workflow for integrating peptide sequence variants (detected via proteogenomics) with the CRAVAT analysis suite.
- This extension demonstrates the value of the Galaxy platform for developing and disseminating sophisticated proteogenomic workflows.
- Also, this extension uses the recommended community standard output for proteogenomic, proBED, which can be generated from a Galaxy-based proteogenomics workflow.

Acknowledgements

- This project is supported by the Informatics Technologies for Cancer Research (ITCR) program at the NIH/NCI, from grant U24CA204817 to R. Karchin and grant U24CA199347 to T. Griffin
- Demonstration proteogenomic data were generated with the assistance of the University of Minnesota Center for Mass Spectrometry and Proteomics and Genome Center
- We acknowledge use of the Jetstream cloud-based computing resource for scientific computing maintained at Indiana University for assistance in maintaining the publicly available Galaxy instance for demonstration purposes.