An accessible proteogenomics informatics resource for cancer researchers

Matthew C. Chambers[1#]; Pratik D. Jagtap[2#]; James E. Johnson[3#]; Thomas McGowan[3#]; Praveen Kumar[2,4]; Getiria Onsongo [3]; Candace R. Guerrero[2]; Harald Barsnes [5,6]; Marc Vaudel [7,8]; Martens Lennart [9,10,11]; Grüning Björn [12,13]; Ira R. Cooke [14]; Mohammad Heydarian [15]; Karen L. Reddy [16]; Timothy J. Griffin [2*]

[1]Department of Biochemistry, Vanderbilt University, Nashville, TN, USA

[2]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN, USA

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

[4]Bioinformatics and Computational Biology Program, University of Minnesota-Rochester, Rochester, MN, USA

[5]Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway

[6]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway

[7]KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway

[8]Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway

[9]VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium

[10]Department of Biochemistry, Ghent University, Ghent, Belgium

[11]Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium

[12]Department of Computer Science, Albert-Ludwigs-University, Freiburg, Freiburg, Germany

[13]Center for Biological Systems Analysis (ZBSA), University of Freiburg, Freiburg, Germany.

[14]Comparative Genomics Centre and Department of Molecular and Cell Biology, James Cook University, Queensland, Australia

[15]Department of Biology, Johns Hopkins University, Baltimore, MD

[16]Department of Biological Chemistry, Center for Epigenetics and Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD

*Corresponding author:
tgriffin@umn.edu

[#]These authors contributed equally to this work

CONFLICT OF INTEREST STATEMENT:
The authors declare no potential conflicts of interest.

Abstract (142 words/150 max)

Proteogenomics has emerged as a valuable approach in cancer research, which integrates genomic and transcriptomic data with mass spectrometry-based proteomics data to directly identify expressed, variant protein sequences that may have functional roles in cancer. This approach is computationally intensive, requiring integration of disparate software tools into sophisticated workflows, challenging its adoption by non-expert, bench scientists. To address this need, we have developed an extensible, Galaxy-based resource aimed at providing more researchers access to, and training in, proteogenomic informatics. Our resource brings together software from several leading research groups to address two foundational aspects of proteogenomics: 1) generation of customized, annotated protein sequence databases from RNA-Seq data; and 2) accurate matching of tandem mass spectrometry data to putative variants followed by filtering to confirm their novelty. Directions for accessing software tools and workflows, along with instructional documentation, can be found at z.umn.edu/canresgithub.

Main text (1489 words/1500 max)

Introduction

Proteogenomics integrates genomic, transcriptomic and mass spectrometry (MS)-based proteomics data to verify the expression of protein sequence variants resulting from sequence variations at the DNA or RNA level.(1) Most commonly, assembled RNA-Seq data containing potential sequence variants are translated *in-silico*, generating possible expressed protein variants. Tandem mass spectrometry (MS/MS) spectra of peptides are acquired from proteolytic digestion of proteins isolated from the same sample. These MS/MS spectra are matched to the database containing the protein variants, as well as reference, known protein sequences. Peptide spectrum matches (PSMs) of MS/MS spectra to variant sequences within the database confirm the expression of novel protein sequences, helping to distinguish important variants and improve genomic annotation.(1) Recently, high profile studies demonstrated the value of proteogenomics for discovery of protein variants that may be drivers of cancer.(2-5)

Despite its value, using the proteogenomics approach by the wider research community remains a challenge. This is primarily due to its intensive informatics requirements.(6) Proteogenomics requires integration of disparate software from different 'omic domains, optimally within a single, user-friendly environment. The software and supporting hardware must scale to accommodate memory and compute-intensive needs presented by the large-scale datasets encountered. Finally, the workflow must address possible pitfalls, such as false positives and the need to confirm the novelty of putative variants identified.(1) Although some described software platforms meet at least some of these requirements,(7,8) most proteogenomics studies to-date have been accomplished using relatively inaccessible, in-house informatics solutions.

Here, we present an informatics resource aimed at expanding the use of proteogenomics in cancer research. The resource is built upon the Galaxy bioinformatics platform(6). Galaxy enables integration of

disparate, multi-omic tools in a single, user-friendly environment, as required for proteogenomics.(6,9,10) The new resource described here provides workflows and training in the most critical aspects of proteogenomics: generation of customized protein sequence databases from RNA-Seq data, matching of MS/MS data to putative variant peptide sequences, and confirmation of the novelty of these identified sequences.

Description of resource

Figure 1 describes the workflows that make up this resource. Each workflow is also detailed in-turn below. The page z.umn.edu/canresgithub provides directions to access workflows and related instructional material, including on-screen, interactive Galaxy Tours tutorials.

Customized database generation workflow. This workflow, in part, takes advantage of well-documented, mature software for RNA-Seq data analysis that are long-standing, core tools in the Galaxy platform. The workflow's input is raw RNA-Seq 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. The current workflow focuses on insertion-deletion (Indel) variants and single amino acid variants (SAVs). These tools generate a variant call format (.VCF) file that provides a summary of all potential variants identified from the starting RNA-Seq data. Along with a .BAM file (RNA sequence alignment information), the .VCF file acts as an input to the tool CustomProDB.(11) CustomProDB creates a customized protein sequence database in the common .FASTA format, which contains potential variant protein sequences, and annotation for the type of variant (e.g. SAV, Indel). The possible variant sequences are merged with reference protein sequences for the organism being studied to create a comprehensive sequence database for the sample being studied. We have developed workflows (accessed through z.umn.edu/canresgithub) for analyzing single-end RNA-Seq data (from a mouse sample) and also for paired-end RNA-Seq data (from human MCF7 cells).

Sequence database searching and variant confirmation workflow. We have deployed the software SearchGUI (compomics.github.io/projects/searchgui.html),(12) which bundles several of the most popular sequence database searching programs to match MS/MS spectra to peptide sequences contained in the sequence database. The use of complementary searching programs provides more comprehensive and higher confidence PSM identification.(13)

Inputs for the workflow are the customized protein sequence database and also Mascot generic format (.MGF) files, which contain peaklists from the raw MS/MS data. Often, MS-based proteomics data is generated from the fractionation of a single sample, with each fraction generating a separate MS raw file (and .MGF file). Galaxy can define a group of such files as a "Dataset Collection" (See z.umn.edu/canresgithub for a Dataset Collection Galaxy Tour). For this workflow, a Dataset Collection of .MGF files acts as an input to SearchGUI, where each separate .MGF file is analyzed in-turn using the same parameters. A single output from SearchGUI is produced, aggregating the results from each sequence database search program on each .MGF file.

The results file from SearchGUI acts as the input to the companion program PeptideShaker.(14) PeptideShaker further processes PSM information from SearchGUI. This processing includes PSM quality control, statistical analysis and false-discovery rate (FDR) estimation, PTM localization scoring, protein inference from PSMs, as well as organization and annotation for viewing of the output. In its Galaxy implementation, users are offered a number of output options for PeptideShaker, including a PSM report, inferred protein identities and a zipped .cpsx file. The zipped .cpsx file contains all results and can be downloaded from the Galaxy web-interface and viewed using the free PeptideShaker viewer (compomics.github.io/projects/peptide-shaker.html).

The final part of the workflow acts on the PSM report from PeptideShaker to confirm novel variants. PSMs to putatively novel peptide sequences are selected via their annotation from the .FASTA protein sequence

database, and submitted for BLAST-P analysis. BLAST-P compares the sequences to known sequences of the organism being studied. Putative variant sequences that do not perfectly match to known sequences after BLAST-P analysis are selected and outputted as confirmed, novel peptide sequences, ready for further analysis. The output of this workflow is a tabular list of confirmed, novel peptide sequence present in the sample. Instructions on workflow operation and results interpretation are at z.umn.edu/canresgithub.

Accessibility

We have made this proteogenomics informatics resource available in multiple ways. Our public Galaxy instance (usegalaxyp.org) is a training site for use of these workflows, including small-scale data for users to access and use with published workflows. These workflows are also available on a larger capacity instance housed on the cloud-based Jetstream infrastructure.(15) Instructions on accessing usegalaxyp.org and Jetstream are provided at z.umn.edu/canresgithub. Additionally, our workflows and software have been published in the Galaxy Tool Shed. Galaxy users can directly import and use these on their own instance. The archived workflows track and store all operating parameters and version information for the software employed in the analysis pipeline.

Conclusions

The resource described here provides foundational tools and workflows for proteogenomics analysis, implemented in the extensible Galaxy platform to facilitate further enhancements. For example, customized workflows for multi-stage database searching to facilitate variant-specific FDR estimates(1) are being developed. We are also working on a Galaxy plugin for visualizing proteogenomic results, enabling further viewing of PSM and protein identifications. Adding functionality for converting PSM information to a SAM file(7) for downstream viewing in the Integrated Genomics Viewer (software.broadinstitute.org/software/igv) are also in progress. Although not the focus here, Galaxy-

based tools for quantifying RNA-Seq and MS-based proteomics data are available for quantitative proteogenomic analysis. In addition, we expect the active and collaborative community of Galaxy users and developers will continue to add to the proteogenomic resource described.

Acknowledgements

We thank Jeremy Fischer and Tom Doak for Jetstream assistance. We also thank John Chilton for his initial assistance wrapping tools in Galaxy. We acknowledge support from Ghent University Concerted Research Action BOF12/GOA/014 to L.M., Bergen Research Foundation and the Research Council of Norway to H.B., BMBF grant 031 A538A RBC (de.NBI) to B.G., and NCI-ITCR grant 1U24CA199347, as well as NSF (U.S.) grant 1458524 to T.J.G. and the University of Minnesota research team.

References Cited (15 maximum)

1. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods **2014**;11:1114-25
2. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P*, et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. Nature **2016**;534:55-62
3. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z*, et al.* Proteogenomic characterization of human colon and rectal cancer. Nature **2014**;513:382-7
4. Alfaro JA, Sinha A, Kislinger T, Boutros PC. Onco-proteogenomics: cancer proteomics joins forces with genomics. Nat Methods **2014**;11:1107-13
5. Helmy M, Sugiyama N, Tomita M, Ishihama Y. Onco-proteogenomics: a novel approach to identify cancer-specific mutations combining proteomics and transcriptome deep sequencing. Genome Biology **2010**;11(Suppl 1):17
6. Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, Kall L*, et al.* Multi-omic data analysis using Galaxy. Nat Biotechnol **2015**;33:137-9
7. Wang X, Slebos RJ, Chambers MC, Tabb DL, Liebler DC, Zhang B. proBAMsuite, a Bioinformatics Framework for Genome-Based Representation and Analysis of Proteomics Data. Mol Cell Proteomics **2016**;15:1164-75
8. Risk BA, Spitzer WJ, Giddings MC. Peppy: proteogenomic search software. J Proteome Res **2013**;12:3019-25
9. Jagtap PD, Johnson JE, Onsongo G, Sadler FW, Murray K, Wang Y*, et al.* Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework. J Proteome Res **2014**;13:5898-908
10. Fan J, Saha S, Barker G, Heesom KJ, Ghali F, Jones AR*, et al.* Galaxy Integrated Omics: Web-based Standards-Compliant Workflows for Proteomics Informed by Transcriptomics. Mol Cell Proteomics **2015**;14:3087-93
11. Wang X, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics **2013**;29:3235-7
12. Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics **2011**;11:996-9
13. Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW. Combining results of multiple search engines in proteomics. Mol Cell Proteomics **2013**;12:2383-93
14. Vaudel M, Burkhart JM, Zahedi RP, Oveland E, Berven FS, Sickmann A*, et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nat Biotechnol **2015**;33:22-4
15. Stewart CA, Cockerill TM, Foster I, Hancock D, Merchant N, Skidmore E*, et al.* Jetstream: a self-provisioned, scalable science and engineering cloud environment. Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure. St. Louis, Missouri: ACM; 2015. p 1-8.

Figure 1.

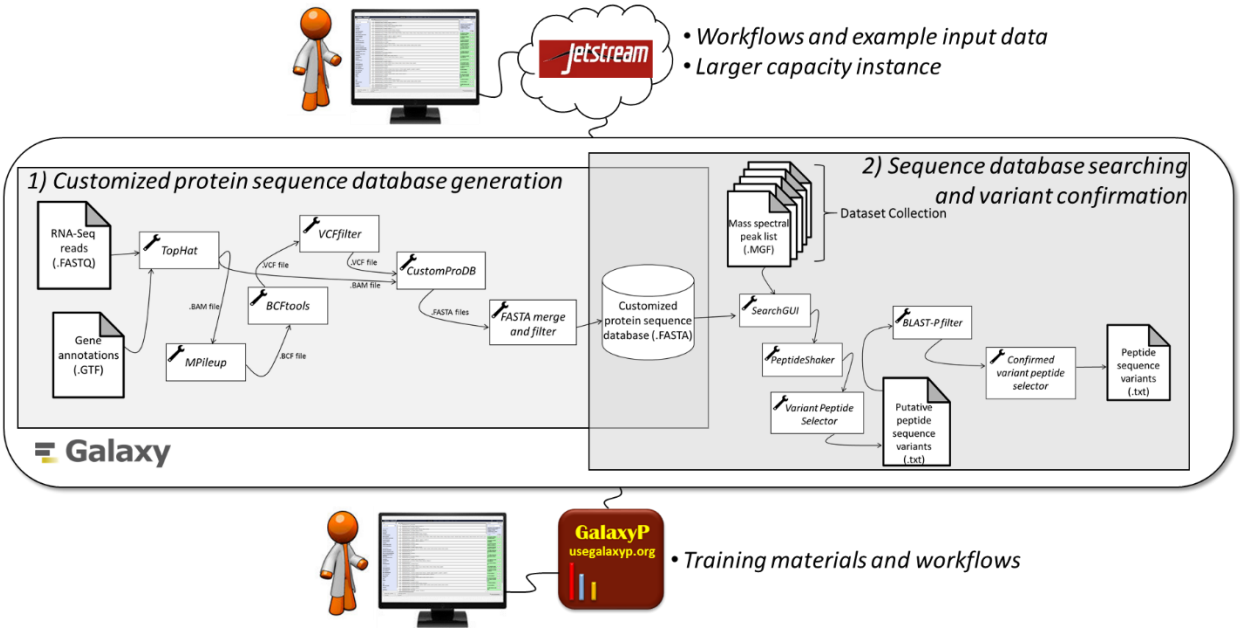

Figure 1. Overview of the proteogenomic informatics resource. The main steps are shown comprising the two core workflows making up this resource: 1) Customized protein sequence database generation from RNA-Seq data; and 2) Sequence database searching using MS/MS data and the customized protein database, followed by variant peptide confirmation.