Journal of proteome research

Article

pubs.acs.org/jpr

Flexible and Accessible Workflows for Improved Proteogenomic Analysis Using the Galaxy Framework

Pratik D. Jagtap,*[,†,‡] James E. Johnson,[§] Getiria Onsongo,[§] Fredrik W. Sadler,[∥] Kevin Murray,[‡] Yuanbo Wang,[⊥] Gloria M. Shenykman,[#] Sricharan Bandhakavi,[▽] Lloyd M. Smith,[#] and Timothy J. Griffin*[,‡]

[†]Center for Mass Spectrometry and Proteomics, University of Minnesota, 43 Gortner Laboratory, 1479 Gortner Avenue, St. Paul, Minnesota 55108, United States

[‡]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, 6-155 Jackson Hall, 321 Church Street South East, Minneapolis, Minnesota 55455, United States

[§]Minnesota Supercomputing Institute, University of Minnesota, 117 Pleasant Street South East, Minneapolis, Minnesota 55455, United States

[∥]St. Olaf College, 1500 St. Olaf Avenue, Northfield, Minnesota 55057, United States

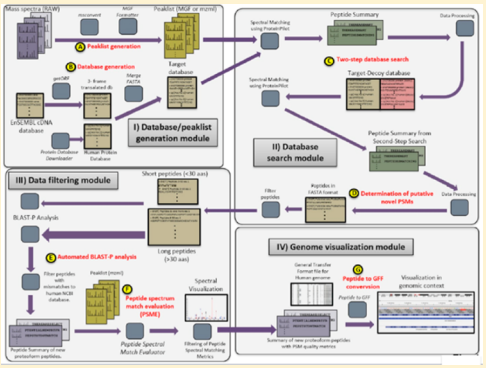
[⊥]Department of Computer Science, Carleton College, 300 North College Street, Northfield, Minnesota 55057, United States

[#]Chemistry Department, University of Wisconsin-Madison, 111 University Avenue, Madison, Wisconsin 53705, United States

[▽]Bio-Rad Laboratories, 6000 James Watson Drive, Hercules, California 94547, United States

Supporting Information

ABSTRACT: Proteogenomics combines large-scale genomic and transcriptomic data with mass-spectrometry-based proteomic data to discover novel protein sequence variants and improve genome annotation. In contrast with conventional proteomic applications, proteogenomic analysis requires a number of additional data processing steps. Ideally, these required steps would be integrated and automated via a single software platform offering accessibility for wet-bench researchers as well as flexibility for user-specific customization and integration of new software tools as they emerge. Toward this end, we have extended the Galaxy bioinformatics framework to facilitate proteogenomic analysis. Using analysis of whole human saliva as an example, we demonstrate Galaxy's flexibility through the creation of a modular workflow incorporating both established and customized software tools that improve depth and quality of proteogenomic results. Our customized Galaxy-based software includes automated, batch-mode BLASTP searching and a Peptide Sequence Match Evaluator tool, both useful for evaluating the veracity of putative novel peptide identifications. Our complex workflow (approximately 140 steps) can be easily shared using built-in Galaxy functions, enabling their use and customization by others. Our results provide a blueprint for the establishment of the Galaxy framework as an ideal solution for the emerging field of proteogenomics.



KEYWORDS: *proteogenomics, workflows, salivary proteins, customized database generation, peptide corresponding to a novel proteoform, peptide-spectral match evaluation*

INTRODUCTION

The rapidly emerging field of proteogenomics utilizes large-scale genomic or transcriptomic data combined with mass-spectrometry (MS)-based proteomics data to identify peptides corresponding to novel proteoforms[1] arising from genome reorganization, mutations, or transcriptional splicing. Proteogenomic discoveries also lead to new insights into genome biology and improved annotation of genes.[2−7] The convergence of high-throughput genomic/transcriptomic sequencing with more comprehensive proteome characterization via advances in MS instrumentation allows researchers to gather the large-scale data necessary for effective proteogenomic analysis. Consequently, the numbers of proteogenomic studies are increasing; a search of PubMed publications with the term "proteogenomics" shows an increase of over 400% between years 2010 and 2013. Applications of proteogenomics cover many fields. The discovery of novel peptides corresponding to proteoforms in bacteria[2,8−11] and nonmodel organisms for which there is no available protein reference database[12] is an emerging application area. Although only a modest number of reports exist on clinically relevant proteogenomic studies,[13−15]

Received: August 6, 2014
Published: October 10, 2014

efforts such as the Clinical Proteomics Tumor Analysis Consortium (CPTAC)[16] and the Chromosome-centric Human Proteome Project[17] show increasing momentum for proteogenomic analysis as a means to gain new insights into disease mechanisms and biomarker discovery. Additionally, the recently completed drafts of the human proteome[18,19] both employed an MS-based proteogenomics approach.

At the core of proteogenomics is integrated analysis of large-scale data sets, demanding complex workflows. A proteogenomics workflow can be generically broken down into the following modules, with each of these containing one or more subworkflows composed of multiple processing steps using one or more software programs: (1) Peaklist generation and protein sequence database generation from assembled DNA or RNA sequences; (2) sequence database searching; (3) data filtering and confidence assignment; and (4) visualization and interpretation of novel protein products with respect to genomic organization.

A number of laboratories have described algorithms and software for accomplishing at least a portion of these procedures to make up a proteogenomics workflow.[7,16,20−24] For example, various algorithms have been used for generating protein sequence databases from genomic/transcriptomic data. These include six-frame translation from genomic DNA sequence data[3,6,25,26] and three-frame translation from transcriptomic sequence data derived from cDNA[27,28] or RNA-Seq data.[7,22,29−31] Because the generated protein sequence databases are large, methods have been developed to reduce the database size and improve results when matching tandem mass spectrometry (MS/MS) to these sequences.[3,27,32] A number of different programs have been used for matching of MS/MS data to these databases, including both publicly available programs such as X! Tandem,[10,28,33] Myrimatch,[10,28,33] and MS-GF+[3] and commercial programs such as Mascot,[10] Sequest,[33] and ProteinPilot.[27] After identification of possible novel peptide sequences, a number of steps for filtering and assessing confidence have been used, including BLAST analysis to verify putatively novel peptide spectral matches (PSMs).[27,28] Lastly, there are multiple tools to visualize peptide positions on the genome.[23,24,34]

On the basis of this examination of the current practices, it is clear that there are a number of options available to accomplish many of the important aspects of proteogenomic data analysis. However, coupling these software programs together to create a comprehensive proteogenomic analysis workflow challenges even those with expertise in computer science and software development. Furthermore, these workflows require sample- and experiment-specific tuning in response to different experimental approaches,[3,27] different input data types (e.g., genomic DNA data versus RNA-Seq data), and desired analysis outcomes (e.g., qualitative versus quantitative studies). Ideally, the platform would also contain functionality for assembling DNA or RNA sequencing data from which the protein database can be constructed. Although some effective software exists expressly for proteogenomic analysis,[35] no platform currently available offers the flexibility to meet all of the above requirements.

How can this need for flexibility in proteogenomic analysis be met? Seeking an answer, we have turned to the Galaxy framework for bioinformatic workflow management and development. Designed initially to address computational bottlenecks in genomic and transcriptomic data analysis,[36] the open-source, web-based Galaxy framework offers multiple benefits for complex analytical workflow development. We have extended the Galaxy framework (called Galaxy for Proteomics, or Galaxy-P) to run MS-based proteomics software (usegalaxyp.org), seeking to expand the use of the framework into new "omic" analysis areas.

Using the analysis of whole human saliva as a representative application, we demonstrate how Galaxy enables the generation of an integrated, modular proteogenomics workflow that can be customized to sample-specific needs, thereby improving results. We also highlight the additional benefits offered by proteogenomic analysis using the Galaxy framework, the ability to share complex workflows in their entirety, promoting access and use by others, and, if desired, customization to meet their own needs. Additionally, Galaxy already offers a suite of tools for assembling DNA and RNA sequencing data, making it a platform capable of all data-processing steps necessary for proteogenomics. Collectively, our results prove the utility of the Galaxy framework as an effective solution for proteogenomic analysis.

MATERIALS AND METHODS

Input and output files used and generated from Galaxy-P workflow modules (encapsulated in shareable "histories"; see "database generation" below as an example) as well as the analytical workflows themselves within each module have been provided as links (See Supplementary Table S1 in the Supporting Information and z.umn.edu/proteinpilotpage and z.umn.edu/xtandempage).

Salivary Supernatant Data Set

Supernatant from saliva that was collected and pooled from six healthy subjects is used for this analysis. Proteins were treated with Proteominer (Bio-Rad Laboratories, Hercules, CA) for dynamic range compression and were subjected to multidimensional peptide fractionation after trypsin digestion. The data were generated using an LTQ-Orbitrap XL mass spectrometer as previously described.[37] Additionally, 45 Thermo RAW files were generated from ProteoMiner Library-2-treated saliva and also analyzed.

The RAW files were grouped into four categories and used in Galaxy-P workflows: (a) 2D fractionated salivary supernatant with and without ProteoMiner treatment (40 RAW files); (b) 3D fractionated salivary supernatant without ProteoMiner treatment (41 RAW files); (c) 3D fractionated salivary supernatant with ProteoMiner treatment (52 RAW files); and (d) 3D fractionated salivary supernatant with ProteoMiner Lib-2 treatment (58 RAW files).

Generation of Three-Frame Translated cDNA and Microbial Database

EnSEMBL cDNA database (version GRCh37.72; 192 628 cDNA sequences; http://z.umn.edu/ensembldb) was downloaded and converted into a three-frame translated cDNA database (5 459 808 protein sequences; http://z.umn.edu/getorfensembl) using getORF, a tool from the EMBOSS suite of software. The translated cDNA database was merged with the Human UniProt database (88 378 protein sequences) and contaminant proteins (115 protein sequences) to eliminate redundant sequences and build a database (2 768 639 sequences) that was used for the first-step proteogenomic search. The data input (http://z.umn.edu/ensembldb), workflow (http://z.umn.edu/3framecdnadb), and output (http://z.umn.edu/step1output) for database generation have been provided on the Galaxy-P public Web site (usegalaxyp.org).

The history for generation of the database can be accessed at http://z.umn.edu/dbgenhistory.

Additionally, the Human Oral Microbiome Database (HOMD) was used for searching the data set for microbial peptides. HOMD was employed to reduce the possibility of a microbial-peptide-associated spectrum being erroneously assigned to peptide derived from a novel human proteoform.

The oral microbiome dynamic protein database (dated 08/09/2013; 4 317 054 protein sequences) was downloaded from the HOMD Web site (http://www.homd.org/index.php?&name=seqDownload&type=G). The HOMD database was merged with the Human UniProt database (88 378 protein sequences) and contaminant proteins (115 protein sequences) to eliminate redundant sequences and generate a database (4 302 210 protein sequences) that was used for the first-step metaproteomic search. The input (http://z.umn.edu/inputhomd), workflow (http://z.umn.edu/homdwf), and output (http://z.umn.edu/outputhomd) for the database generation have been provided on the Galaxy-P public Web site (usegalaxyp.org). The history for generation of the database can be accessed at http://z.umn.edu/homddbgenhistory.

Peaklist Generation

msconvert and MGF formatter were used to convert RAW files into intermediate mzml files and MGF files for ProteinPilot search using a multifile data set approach[38] (recently modified to "data set collection" method). In brief, multiple RAW files associated with the set of fractions were merged into a single file that was used for MS database searching within ProteinPilot and subsequent steps such as PSM evaluation. (See Peptide Spectrum Match Evaluation section.)

Database Search

The linked MGF files generated in the step above were searched using ProteinPilot (4.5.0.0, 1654 revision: 1656) using a modified version of the "Minnesota two-step" method, a strategy for identifying peptides from large databases.[27] The MS searches were conducted as follows using several different "first-step" database searches:

(a) The first search was carried out against the target version of merged Human UniProt database, contaminant database, and three-frame translated cDNA database; (b) the second search was carried out against the target-decoy version of merged Human UniProt database, contaminant database, and three-frame translated cDNA database; (c) the third search was carried out against the target version of merged Human UniProt database, contaminant database, and HOMD database; and (d) the fourth search was carried out against the target-decoy version of merged Human UniProt database, contaminant database, and HOMD database.

Each search generated a peptide summary file, which listed an accession number associated with either the cDNA (ENST accession associated with the three-frame translated Ensembl database) or HOMD sequences to which the identified peptide matched. Accession numbers associated with the three-frame EnSEMBL database or HOMD database were parsed out.

All protein entries from the original three-frame translated EnSEMBL (cDNA) (a and b searches above) and HOMD database (c and d searches above), which contained one or more peptide identifications, were used to generate a subset FASTA file. The resulting subset FASTA file was merged with the Human UniProt database (88 378 protein sequences) and contaminant proteins (115 protein sequences). This eliminated redundant sequences and produced a smaller database. This "Human UniProt + contaminant + subset cDNA + subset HOMD database" was used for the "second step" for the combined metaproteomic and proteogenomic MS search using ProteinPilot.

The workflow for this step (http://z.umn.edu/mn2stepms) is available in Supplementary Table S1 in the Supporting Information, and outputs for all of the fractions mentioned in the Salivary Supernatant Data Set section are in Supplementary Sections S2, S3, S4, and S5 in the Supporting Information under the "Second-Step Peptide Summary" tab.

Identifying Peptides from Three-Frame Translated cDNA Database

The peptides identified from the modified two-step search were used to identify novel peptide sequences, those not present in the Human UniProt, HOMD (microbial), or contaminant database. Only peptides with a Conf score more than 95% were used for further analysis. Various text-formatting tools within Galaxy-P were used to generate a list of peptides corresponding to potential alternatively splice proteins.

The workflow for this step (http://z.umn.edu/pepsum23frame) is available in Supplementary Table S1 in the Supporting Information, and outputs for all the fractions mentioned in the Salivary Supernatant Data Set section are in the Supplementary Sections S2, S3, S4, and S5 in the Supporting Information under the "Peptides with ENST acc no" tab.

BLAST-P Search

The distinct peptides that were identified exclusively with the three-frame translated cDNA database were searched against the human NCBI nr remote database (current database that enlists all proteins from human proteome) using a custom-built Galaxy tool. Using this tool, the peptides were divided into two groups depending on their peptide length. Shorter peptides (<30 aas) were searched using the short BLAST-P tool (version 2.2.28+), wherein the following parameters were used: set expectation value cutoff: 200 000; scoring matrix: PAM30; gap costs: Existence 9, Extension 1; word size for wordfinder algorithm: 2; multiple hits window size: 15; threshold: minimum score to add a word to the BLAST lookup table: 16; and use composition-based statistics: 0 or F. Longer peptides with length 31 aas and more were searched using BLAST-P tool (version 2.2.28+), wherein the following parameters were used : set expectation value cutoff: 10; scoring matrix: BLOSUM62; gap costs: existence 11, extension 1; word size for wordfinder algorithm: 3; multiple hits window size: 40; threshold: minimum score to add a word to the BLAST lookup table: 11; use composition-based statistics: 2,T or D. Both the short BLAST-P and BLAST-P searches were used to generate a BLAST XML output with maximum one hit per peptide as an output. The BLAST XML outputs were converted into tabular format with 25 various metric outputs including peptide sequence, percentage of identical matches (pIdent), alignment length, query sequence length, and total number of gaps.

Peptide matches with BLAST pIdent score less than 100, with at least one gap and a ratio of alignment length to query sequence length of <1, were selected along with those peptides that did not show any matches to the NCBI nr human database. This set of unmatched peptides was used for peptide−spectrum match evaluation.

The workflow for this step (http://z.umn.edu/blastpms) is available in Supplementary Table S1in the Supporting Information, and outputs for all the fractions mentioned in

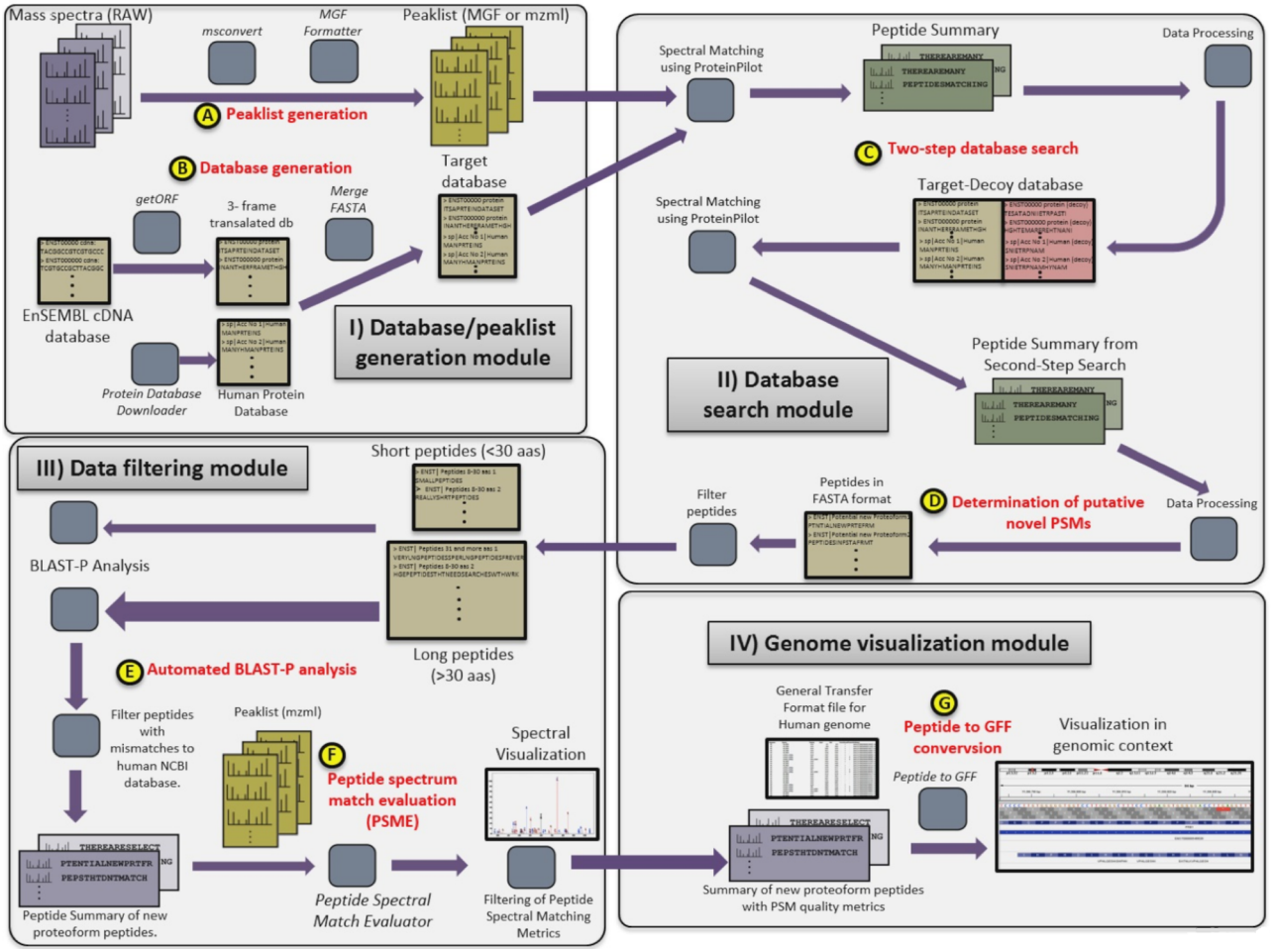


Figure 1. Overview of modules and subworkflows comprising the Galaxy-based proteogenomic analysis workflow.

the Salivary Supernatant Data Set section are in Supplementary Sections S2, S3, S4, and S5 in the Supporting Information under "BLAST-P Search Output" and "Peptides with mismatches" tab.

Peptide Spectrum Match Evaluation

To manually verify the quality of peptide spectrum matches (PSMs), we developed a Galaxy tool called the Peptide Spectrum Match Evaluator (PSME). PSME enables the evaluation of spectral features underlying each peptide identification, which allows more information to be used to discriminate true from false-peptide identifications. Spectra with the highest associated score for each such peptide sequence were used to generate a peptide summary. Given the list of peptide identifications and the original raw mass spectrometry files (mzml format), PSME generates a tabular file that lists, for each PSM, several spectral properties such as number of continuous b ions, number of continuous y ions, percent of peaks matched, number of peaks unmatched above 10% intensity of maximum intensity peak, total ion current, and other custom features as required. Furthermore, the tool identifies b and y ions matched, monoisotopic peaks, and losses of water and ammonium and internal ions within a predefined mass error based on rules defined by the ProteinProspector tool (http://prospector.ucsf.edu/prospector/mshome.htm). PSME allows filtering of peptide identifications using thresholds for spectral features. For this study, we filtered for spectral matches that had at least four continuous b and y ions; at least 50% of peaks matched with predicted ions had no more than one missed matching peak above 10% of highest intensity peak and had a total ion current of at least 20 000.

PSME also generates HTML links that are used to visualize spectral assignments. (See Figure 2.) Here interactive controls can be used to change ion assignments and other parameters. The filtered spectra in this study were manually examined using PSME-generated HTML links, and peptides with acceptable quality matches were subsequently mapped to the genome using the "Peptides to GFF" tool (see below).

The workflow for this step (http://z.umn.edu/psmems) is available in Supplementary Table S1 in the Supporting Information, and outputs for all the fractions mentioned in the Salivary Supernatant Data Set section are in Supplementary Sections S2, S3, S4, and S5 in the Supporting Information under the "PSME Metrics" and "Quality PSME peptides" tabs.

Genome Context Analysis

To visualize the location of identified peptides on the genome, we created a Galaxy tool, "Peptides to GFF", that finds where on the genome each peptide maps to and then generates a GTF (general transfer format) file containing coordinates of the peptide alignments. The "Peptides to GFF" uses a peptide summary file, the Ensembl GTF file, and raw cDNA sequences as an input to generate a GFF3 file. The GFF3 file can be viewed within most genome browsers, such as the Integrated Genomics Viewer (http://www.broadinstitute.org/igv/home) and displays a peptide "track". (See Figure 3.)

dx.doi.org/10.1021/pr500812t | J. Proteome Res. 2014, 13, 5898−5908

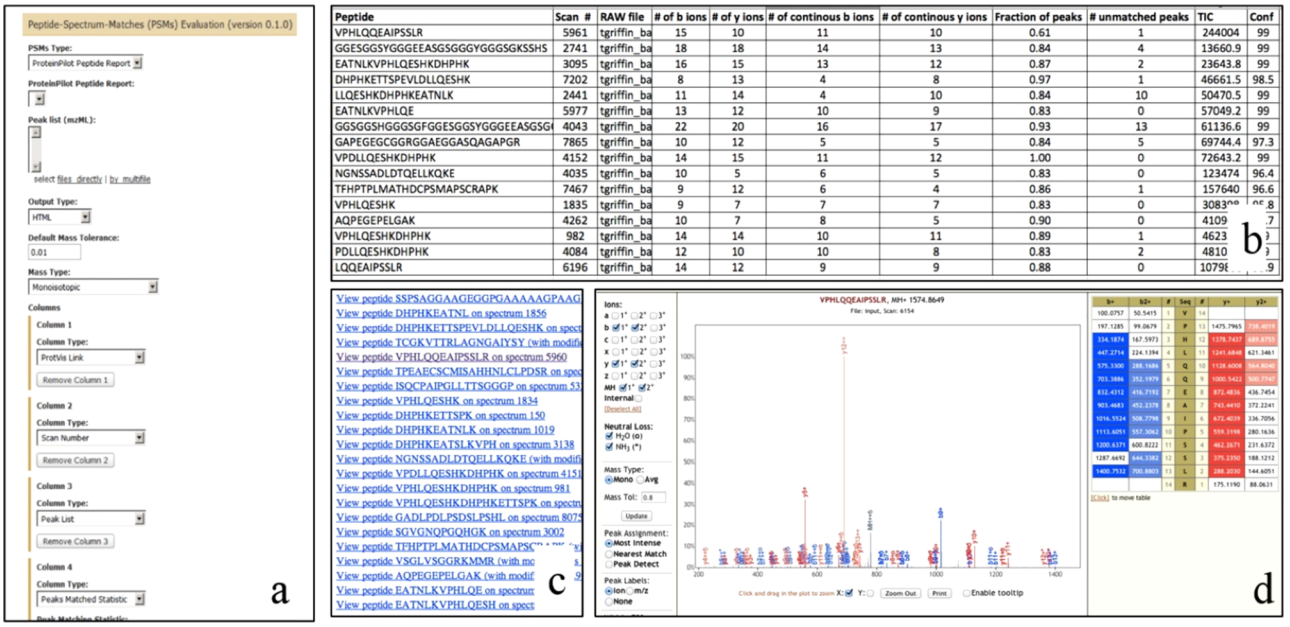


Figure 2. Overview of components of the Peptide Spectrum Match Evaluation tool. Screenshot of the PSME tool within Galaxy-P showing (a) user interface for setting parameters for PSM evaluation, (b) Tabular format output from the PSME tool, (c) HTML output from the PSME tool, and (d) interactive spectral annotation that can be used to visualize PSMs before further evaluation.

The workflow for this step (http://z.umn.edu/peptides2gtf) is available in Supplementary Table S1 in the Supporting Information, and outputs for all fractions mentioned in the Salivary Supernatant Data Set section are in Supplementary Sections S2, S3, S4, and S5 in the Supporting Information under the "GTF file for IGV" tab.

The methods, inputs, workflows, and outputs from each of the above modules have been shared and documented via an accessible Galaxy "page" using ProteinPilot (http://z.umn.edu/proteinpilotpage). Workflows using freely available software (e.g., X! Tandem) can be run on a public Galaxy instance at usegalaxyp.org after registering via the "User" tab. The raw data for the analyses of this representative data set have been deposited in PRIDE (Supplementary Section S6 in the Supporting Information).

RESULTS

Data derived from whole human saliva were analyzed using Galaxy-based proteogenomics workflow, organized in four modules, with each module containing subworkflows (Figure 1). Supplementary Table S1 in the Supporting Information summarizes the subworkflows and provides Web links to each, as well as the entire workflow, available through our public Galaxy instance (usegalaxyp.org). We applied the analytical workflow to the analysis of large-scale, high-resolution MS-based proteomics data from whole saliva from a previously described study[36] coupled to archived, publicly available large-scale human cDNA and metagenomic data. The Materials and Methods details the steps within the analytical workflows as well as analyzes the raw data sets. Here we highlight how aspects of each module enable and improve proteogenomic analysis.

The first module generates peaklists from raw tandem mass spectrometry (MS/MS) data and a protein sequence database from assembled nucleic acid sequences by using two separate subworkflows (Workflows A and B in Figure 1 and Supplementary Table S1 in the Supporting Information). We translated human cDNA, derived from expressed RNA, contained in the ENSEMBL database in three different frames (i.e., three-frame translation), to account for all amino acid coding possibilities. Unique cDNA-derived sequences were added to a database of known human protein sequences. Because whole saliva is known to contain a diverse microbiome contributing to expressed proteins in this fluid,[39] we also included proteins translated from metagenomics data contained in the Human Oral Microbiome Database (HOMD).[40] Using Galaxy's tools for FASTA file manipulation and customized database generation, we added the microbial proteins to the human proteins database to generate a database of over 4 million protein sequences.

The next module contains a subworkflow (Workflow C in Figure 1) for matching MS/MS spectra to peptide sequences via database searching. Currently, the Galaxy-P framework contains a selection of popular sequence database searching programs (seeusegalaxyp.org), either freely available or commercial. For our proteogenomic workflow, we chose the popular and powerful program ProteinPilot[41,42] as the main database search engine. We also took advantage of the diverse suite of software available in Galaxy-P by an accompanying search of a portion of the MS/MS data using the freely available X! Tandem program,[43] seeking to investigate the effectiveness of dual database searching to increase confidence in results. We incorporated a modified version of our previously described "Minnesota Two-Step" method[27] to address the inherent challenge presented by large protein sequence databases in proteogenomic analysis.[32] This method employs a first-step sequence database search using relaxed stringency to identify a smaller subset of proteins that are most likely to be present in the sample. A refined, smaller sequence database is generated from this first step and MS/MS was matched to this database in a second step, applying high stringency to these results to generate PSMs at acceptable FDR levels. We automated the two-step database searching method in our Galaxy-based workflow, enabling a first-step low-stringency sequence data-

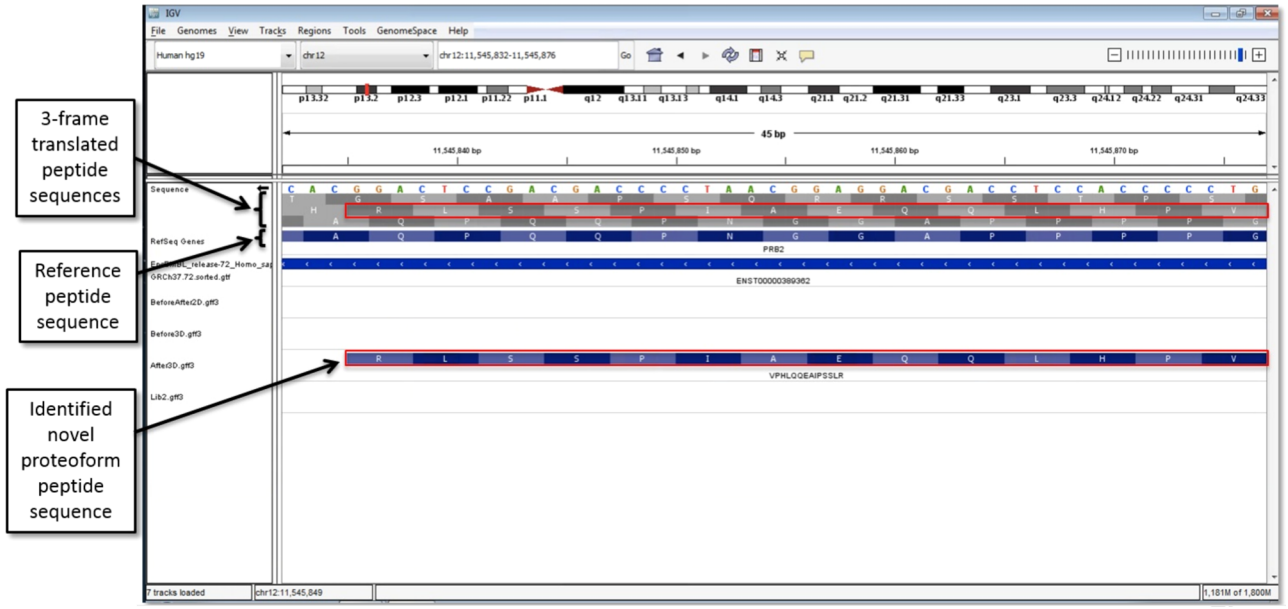


Figure 3. Screenshot of a peptide corresponding to a novel proteoform within Integrated Genomic Viewer. View is a zoomed-in screenshot of chromosome 12, which shows the orientation of expression, amino acid sequences within three frames of translation, and reference files in the tracks and amino acid sequence of the identified peptide corresponding to a novel proteoform.

base search to identify both possible human and microbial proteins present, and created a much smaller sequence database for the second-step high stringency sequence database search.

Our two-step database search resulted in 9333 PSMs to putatively novel human cDNA-derived peptide sequences along with 16 370 PSMs to microbial peptides. The PSMs to cDNA-derived sequences were further analyzed via subworkflows contained in the filtering module (Workflows E and F in Figure 1 and Supplementary Table S1 in the Supporting Information). This module provides critical filtering steps to ensure only PSMs to novel sequences of highest confidence are outputted from the workflow for further consideration. As a first filtering step, we developed a process in Galaxy-P to automatically submit large numbers of peptide sequences to the BLASTP software for sequence characterization. Analysis in BLASTP identifies those PSMs that do not perfectly match to any known sequences within the large NCBI database and are of most interest for further characterization as peptide sequences corresponding to novel proteoforms. The BLASTP analysis identified 1630 PSMs from the original 9333 PSMs as not matching known sequences using relatively strict criteria. (See the Materials and Methods.)

As a further filtering step to ensure high confidence, the PSMs to novel sequences determined via the BLASTP search were subjected to custom software implemented in Galaxy-P called the Peptide Spectrum Match Evaluation (PSME) tool. (See the description in the Materials and Methods.) Figure 2 shows a screen shot of the PSME tool. PSME provides a means to not only visualize MS/MS spectra against their putative sequence match but also to filter large numbers of spectra based on a variety of user-defined parameters relating to PSM quality. Using high-stringency PSM quality criteria (see the Materials and Methods) reduced the number of matches to novel peptide sequences to a total of 55.

The final module of the workflow incorporates steps for visualization and interpretation of the peptide sequences surviving the filtering steps. To achieve this, we developed a tool called "Peptides to GFF" for conversion of peptide amino acid sequences to a format compatible with the popular Integrated Genome Viewer (IGV).[44] Once converted, novel peptide sequences can be viewed and characterized in the context of known genome structure. Figure 3 shows a representative screenshot of a novel peptide visualized in IGV. Using this visualization tool, we categorized the nature of the 55 novel peptides identified from our workflow. (See Supplementary Table S7 in the Supporting Information.) From this analysis, it was determined that two of the peptides matched to genomic regions already known to be translated, which were missed by the BLASTP filtering. Another peptide was identified as translated from a novel DNA junction, but the junction erroneously combined different DNA frames. This peptide was discarded. Table 1 summarizes the nature of remaining novel peptides mapping to 18 different human chromosomes. Supplementary Table S8 in the Supporting Information summarizes the results from the proteogenomic workflow.

Table 1. Summary of Genomic Organization of Peptides Corresponding to Novel Proteoforms

genomic rearrangements	peptides	chromosome location(s)
alternate frame	26	1,3,5,7, 8, 9, 11, 12, 14, 16, and 19
untranslated region	15	2, 4, 6, 7, 8, 11, 12, 13, 14, and 19
pseudogenes	6	1, 3, 6, 14, 19, and X
intronic region	2	12 and 16
novel exon junctions	2	15 and 17
antisense	1	8

With the analytical workflow in hand, we investigated how workflow customization in Galaxy improved proteogenomic results. First, we investigated the effect of including nonhost microbial proteins on the number of high-confidence matches to novel peptide sequences. Using a representative subset of the MS-based proteomics data and applying stringent filtering previously described, we found that including microbial proteins with the three-frame translated human proteins

dx.doi.org/10.1021/pr500812t | J. Proteome Res. 2014, 13, 5898−5908

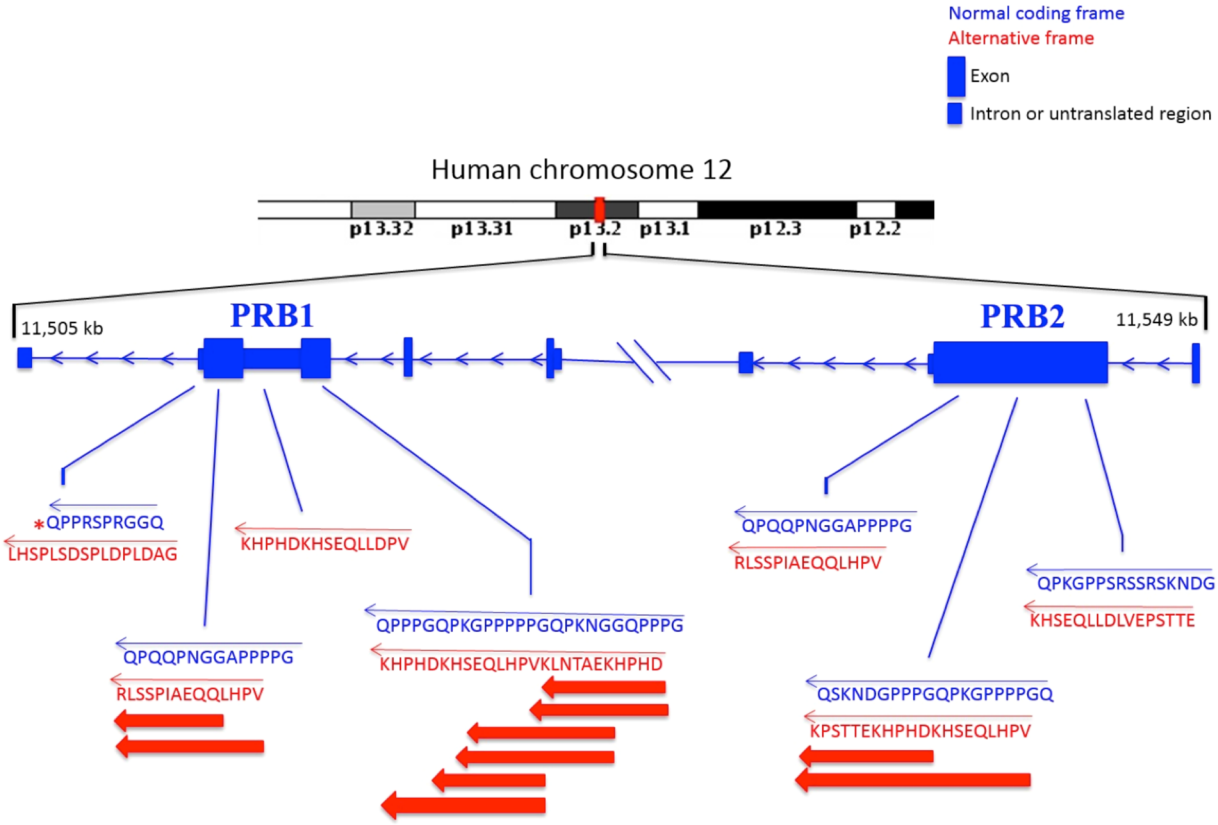


Figure 4. Representation of organization of identified peptides corresponding to a novel proteoform from PRB1 and PRB2 genes on chromosome 12. View is a zoomed-in screenshot of chromosome 12, which shows the orientation of expression, amino acid sequences within three frames of translation, reference files in the tracks, and amino acid sequence of the identified peptide corresponding to a novel proteoform. The red arrows indicate the direction and amino acid sequence (from amino-terminal to carboxy-terminal) of the identified peptides. A red asterisk indicates a stop codon in the normal coding frame. Block arrows in red indicate multiple distinct peptides identified during the proteogenomic analysis.

increased novel peptide identifications by almost two-fold compared with using only the human proteins (35 versus 19 novel peptides, respectively). Supplemental Table S9 in the Supporting Information summarizes the results from this comparison.

We also investigated improvements offered by the use of a second sequence database-searching program (X! Tandem), available through the suite of tools incorporated into Galaxy-P. Using a representative subset of the large-scale MS/MS data and applying the filtering steps of our workflow, we confirmed the identification of a number of the novel peptides initially identified via the ProteinPilot analysis (z.umn.edu/xtandempage and Supplementary Tables S7 and S8 in the Supporting Information).

Examining high-confidence PSMs identified by both sequence database searching programs revealed interesting matches to the Proline Rich Protein (PRP) gene-coding region on chromosome 12. (See Figure 4.) This region codes for the expression of the basic proline-rich proteins (PRB1, PRB2, PRB3, and PRB4),[45] which are highly abundant in human saliva.[46] In total, 10 PSMs mapped to sequences expressed via a frameshift to the PRB1 (proline-rich protein BstNI subfamily 1) coding region. A number of the identified peptide sequences overlapped. One of these frame-shifted peptide sequences (VPDLLQESHKDHPHK) mapped to an intronic region, while another (GADLPDLPSDSLPSH) mapped to a putatively untranslated region of PRB1. Four additional PSMs mapped to a frameshifted protein expressed from the PRB2 (proline-rich protein BstNI subfamily 2) coding region on chromosome 12 (Figure 4). Two of the novel peptide sequences from PRB1 also map to the PRB2 region, reflecting the close sequence similarity between this protein family. Notably, we also identified several peptides matching to the canonical sequences of the PRB1 and PRB2 proteins.

DISCUSSION

Proteogenomic analysis is gaining momentum, driven by the convergence of high throughput nucleic acid sequencing technologies and high-resolution mass spectrometry-based proteomics technologies, making generation of the necessary large-scale data feasible. With the advent of new instrumentation, data outputs yield information with increasing depth. However, accessibility to effective informatics tools for its analysis has lagged behind, especially in analysis of outputs from "omics" technologies.[47] The complexity presented by merging multiomic data heightens this challenge for proteogenomic analysis.[48] In addition to handling of large-scale and complex data, proteogenomics analysis also requires flexibility for tuning analysis to the specifics of any given sample or experimental context. The tools also must be adaptable to new algorithms and data types, which arise as new data-generating technologies emerge.

We address these current needs in proteogenomic analysis via extension of the Galaxy workflow framework. Our proteogenomics workflow consists of about 140 processing steps, grouped into four modules defined in Figure 1. Despite

dx.doi.org/10.1021/pr500812t | J. Proteome Res. 2014, 13, 5898−5908

this complexity, the entire workflow can be run in "one-click" fashion after the parameters have been optimized, with little intervention needed from the user. Importantly, the subworkflows contained within each module can be run on their own if so desired. These workflows can be easily shared, through a weblink or through a saved Galaxy workflow file. Paired with the workflow file, a history file can also be shared, which contains all software and intermediate data inputs and outputs necessary for reproducing the analytical workflow.

Our results also demonstrate a key feature of Galaxy—its flexibility. We show how this flexibility enables development of workflows tuned to sample-specific characteristics, leading to improved results for proteogenomic analysis. Specifically, Galaxy enabled easy generation of a customized protein sequence database merging both human cDNA three-frame translated protein sequences and microbial proteins derived from the HOMD database. The use of this more comprehensive database increased the number of quality PSMs to novel sequence variants in our whole saliva data by two-fold. An explanation for this is that omission of the microbial sequences means MS/MS spectra from nonhost peptides are forced to match against host proteins, increasing false-positives and forcing higher scoring thresholds for PSMs to achieve acceptable FDR, thereby decreasing the number of confident matches to novel peptide sequences. These findings are significant beyond the goals of this work, suggesting that proteogenomics in samples that may contain proteins expressed by nonhost organisms will suffer if using only the host protein sequences for database searching.

Galaxy also enables implementation of database reduction methods,[3,7] which tackle the challenge of large sequence databases inherent to proteogenomic methods.[3,27] We were able to automate our previously described "Minnesota Two-Step" method for database reduction, which increases confident PSMs from large-databases.[27,39] To maximize the identification of variant peptide sequences, we used Paragon algorithm[41] (within ProteinPilot software), which has the ability to automatically detect multiple modifications. We could identify some variant peptides with common post-translational modifications (PTMs), many of which are introduced during sample handling (e.g., deamidation, oxidation). We have not elaborated on these PTMs and have focused only on the identified peptide sequences. We anticipate that for other studies, wherein researchers are seeking to detect biologically relevant PTMs (e.g., phosphorylation, acetylation, etc.), a comprehensive PTM analysis on variant peptides could add further depth to proteogenomic studies.

Galaxy's flexibility to customized software implementation enabled the development of steps for rigorous filtering of data, a key to effective proteogenomic analysis. Unlike conventional MS-based proteomics where PSMs are used to infer protein level information, proteogenomic analysis results are usually based on single PSMs to putatively novel sequences. Given the potential for false-positives when considering single PSMs,[49] a cautious approach employing rigorous filtering is warranted. Our Galaxy-based workflow supports such an approach, providing several levels of quality control and filtering.

For one, we complemented our main sequence database search via ProteinPilot with a search using X! Tandem. This second database search program provided confirmation of many of the matches to novel peptides identified by ProteinPilot, notably the peptides mapping to the PRP locus. Galaxy's amenability to diverse software provides a platform for workflows utilizing multiple database search programs to improve confidence in results.

In addition, our automated BLASTP tool provides an essential filtering step postdatabase search, revealing both small and large peptides that are truly novel based on comparison with known sequences in the current NCBI database. Our PSME tool provides another valuable layer of evaluation of PSM quality on top of the scoring provided by the initial sequence database search program. We designed this tool with flexibility in mind, allowing the user to select desired parameters and level of stringency for filtering PSMs. The PSME tool provides users the ability to discern between PSMs identified with noisy spectra or due to chimeric spectra (although identified using FDR controlled threshold scores) by offering customizable parameters and visualization of PSMS. Collectively, these tools provide rigorous filtering of PSMs, such that PSMs outputted are of only the highest confidence for further consideration. Compatibility of these sequences with the IGV software via our "Peptide to GFF" tool provides the final, necessary ability of the user to view these results in the context of a reference genome. This output is also compatible with the recently described, Galaxy-based CAPER software for peptide to genome mapping.[24]

Application of our workflow to the proteogenomic analysis of whole human saliva yielded interesting results. Notably, we matched multiple, novel frameshifted peptide sequences that mapped to the basic proline-rich proteins (PRB1 and PRB2) located within the proline-rich protein (PRP) gene locus on chromosome 12[50,51] (Figure 4). The PRB proteins are thought to play a protective role in whole saliva, binding tannin toxins,[52] and, when glycosylated, play a role in bacterial adherence and clearance within the oral cavity.[53] The biological meaning of these frameshift proteins is unclear at this point. However, these genes are known to be susceptible to frequent mutation.[45] Frameshifted variants may enable diversification of expressed PRBs, increasing the number of expressed forms of these structurally disordered proteins that some have speculated provide increased defensive functions in saliva.[51] Because our sample consisted of saliva pooled from six different individuals, it is also unclear if these PRB variants are expressed in every individual. More investigation will be necessary to answer these questions.

We readily acknowledge that our Galaxy-based approach is not the only way to accomplish proteogenomic data analysis. A number of software programs[7,16,20−24] have emerged recently to support such analyses, offering tools for steps within one or more of the modules defined in Figure 1. Platforms expressly designed for proteogenomic analysis, such as the Peppy software,[35] are also emerging. Although none of these software tools offer the combined flexibility, accessibility, and completeness of our Galaxy-based workflow, many of these programs are well-designed, offering innovative solutions for meeting challenges presented by proteogenomic analysis. We see these existing and emerging programs as opportunities for further implementation in Galaxy, enhancing its value for proteogenomic analysis. Some proteogenomic tools are already available in Galaxy, for example, a peptide mapping and viewer program,[23] and workflows for the recently described HiREF proteogenomic method[7] are being implemented in Galaxy (J. Lehtio, personal communication). Indeed, we are currently exploring the implementation of alternative approaches to proteogenomics analysis, and we have recently demonstrated the use of Galaxy for generating protein sequence databases

from RNA-seq data,[54] which, in part, utilizes Galaxy's well-established suite of tools for sequence assembly and analysis using high-throughput DNA or RNA sequencing data.[55−58] We have also recently started working on Galaxy implementation of SearchGUI[59] that allows for the use of multiple search engines and PeptideShaker software (http://code.google.com/p/peptide-shaker/), which offers statistical validation along with a vibrant community of developers.[60] As such, Galaxy's amenability to new software implementation should only enhance its value for proteogenomic analysis, offering a centralized resource wherein the most valuable software can be accessed, combined, and evaluated.

To conclude, we have demonstrated the value of the Galaxy framework as a powerful and unique solution for proteogenomic analysis. Its flexibility improves proteogenomic results, via tuning of the proteogenomic workflow to different sample-type and experimental characteristics and implementation of processing steps and customized software ensuring high confidence results. Galaxy's flexibility promotes new proteogenomic software developments and serves as a central resource to make these programs available to more researchers. Galaxy also simplifies complex proteogenomic analysis, creating workflows that are accessible and usable by noncomputer scientists as well as easily shared with others in their completeness, promoting reproducibility and transparency. We provide a blueprint for the continued development of Galaxy as an enabling tool for proteogenomics, whose value should only continue to grow as more researchers turn to proteogenomic analysis in their work. The upshot will be more impactful discoveries made via proteogenomics, achieving a better understanding of biological processes and diseases.

■ ASSOCIATED CONTENT

Supporting Information

Supplementary Table S1: Overview of Modules and analytical workflows for the proteogenomic analysis. The table provides inputs for the workflow, hyperlinks for the workflows, software tools used and the outputs that are generated from each analytical workflow. Supplementary Section S2: Workflow outputs for the 2D-fractionated salivary supernatant with or without Proteominer treatment at various stages of the proteogenomic workflow. The tabs include (a) the output from Second-Step Peptide Summary (Workflow C); (b) Peptides with three-frame translated cDNA accession numbers (output from workflow D); (c) BLAST-P search output and peptides with mismatches (outputs from Workflow E); (d) Output and filtered output from PSME metrics (outputs from Workflow F); and (e) GTF file for Integrated Genomics Viewer (output from Workflow G). Supplementary Section S3: Workflow outputs for the 3D-fractionated salivary supernatant without Proteominer treatment at various stages of the proteogenomic workflow. The tabs are as described for Supplementary Section S2. Supplementary Section S4: Workflow outputs for the 3D-fractionated salivary supernatant with Proteominer treatment at various stages of the proteogenomic workflow. The tabs are as described for Supplementary Section S2. Supplementary Section S5: Workflow outputs for the 3D-fractionated salivary supernatant with Lib-2 library Proteominer treatment at various stages of the proteogenomic workflow. The tabs are as described for Supplementary Section S2. Supplementary Section S6: The raw data for representative data set analyses have been deposited in PRIDE (Project accession: PXD000991). Supplementary Table S7: Characteristics of peptides corresponding to novel proteoforms. Supplementary Table S8: Summary of identification results for all fractions at various analytical steps within the proteogenomics workflow. Supplementary Table S9: Summary of identification results for comparison of search results with and without use of microbial database. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*P.D.J.: Phone: 612-624-9275. Fax: 612-625-5780. E-mail: pjagtap@umn.edu.
*T.J.G.: Phone: 612-624-5249. Fax: 612-624-0432. E-mail: tgriffin@umn.edu.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by NSF grant 11476079 to T.J.G. and P.D.J., and The Wisconsin Center of Excellence in Genomic Sciences NIH grant 1P50HG004952 to L.M.S. G.M.S. was supported by the NIH Genomic Sciences Training Program 5T32HG002760.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Galaxy-P is maintained by the Minnesota Supercomputing Center at the University of Minnesota. Software development and implementation work was done by John Chilton (Penn State). Data were acquired at the Center for Mass Spectrometry and Proteomics (CMSP) at University of Minnesota. We greatly appreciate help from Dr. Sean Seymour from AB Sciex regarding use of ProteinPilot software. We greatly appreciate discussion with LeeAnn Higgins (CMSP) during the development of PSM Evaluator tool.

■ REFERENCES

(1) Smith, L. M.; Kelleher, N. L. Proteoform: a single term describing protein complexity. *Nat. Methods* **2013**, *10* (3), 186−187.
(2) Armengaud, J.; Hartmann, E. M.; Bland, C. Proteogenomics for environmental microbiology. *Proteomics* **2013**, *13* (18−19), 2731−2742.
(3) Branca, R. M.; Orre, L. M.; Johansson, H. J.; Granholm, V.; Huss, M.; Perez-Bercoff, A.; Forshed, J.; Kall, L.; Lehtio, J. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* **2014**, *11* (1), 59−62.
(4) Castellana, N.; Bafna, V. Proteogenomics to discover the full coding content of genomes: a computational perspective. *J. Proteomics* **2010**, *73* (11), 2124−2135.
(5) Renuse, S.; Chaerkady, R.; Pandey, A. Proteogenomics. *Proteomics* **2011**, *11* (4), 620−630.
(6) Volkening, J. D.; Bailey, D. J.; Rose, C. M.; Grimsrud, P. A.; Howes-Podoll, M.; Venkateshwaran, M.; Westphall, M. S.; Ane, J. M.; Coon, J. J.; Sussman, M. R. A proteogenomic survey of the Medicago truncatula genome. *Mol. Cell. Proteomics* **2012**, *11* (10), 933−944.
(7) Woo, S.; Cha, S. W.; Merrihew, G.; He, Y.; Castellana, N.; Guest, C.; MacCoss, M.; Bafna, V. Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.* **2014**, *13* (1), 21−28.

(8) Christie-Oleza, J. A.; Miotello, G.; Armengaud, J. Proteogenomic definition of biomarkers for the large roseobacter clade and application for a quick screening of new environmental isolates. *J. Proteome Res.* **2013**, *12* (11), 5331−5339.
(9) Christie-Oleza, J. A.; Pina-Villalonga, J. M.; Guerin, P.; Miotello, G.; Bosch, R.; Nogales, B.; Armengaud, J. Shotgun nanoLC-MS/MS proteogenomics to document MALDI-TOF biomarkers for screening new members of the Ruegeria genus. *Environ. Microbiol.* **2013**, *15* (1), 133−147.
(10) Muller, S. A.; Findeiss, S.; Pernitzsch, S. R.; Wissenbach, D. K.; Stadler, P. F.; Hofacker, I. L.; von Bergen, M.; Kalkhof, S. Identification of new protein coding sequences and signal peptidase cleavage sites of Helicobacter pylori strain 26695 by proteogenomics. *J. Proteomics* **2013**, *86*, 27−42.
(11) Venter, E.; Smith, R. D.; Payne, S. H. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *PLoS One* **2011**, *6* (11), e27587.
(12) Evans, V. C.; Barker, G.; Heesom, K. J.; Fan, J.; Bessant, C.; Matthews, D. A. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat. Methods* **2012**, *9* (12), 1207−1211.
(13) Flynn, J. M.; Czerwieniec, G. A.; Choi, S. W.; Day, N. U.; Gibson, B. W.; Hubbard, A.; Melov, S. Proteogenomics of synaptosomal mitochondrial oxidative stress. *Free Radic. Biol. Med.* **2012**, *53* (5), 1048−1060.
(14) Jacob, F.; Goldstein, D. R.; Fink, D.; Heinzelmann-Schwarz, V. Proteogenomic studies in epithelial ovarian cancer: established knowledge and future needs. *Biomarkers Med.* **2009**, *3* (6), 743−756.
(15) Vergara, D.; Tinelli, A.; Martignago, R.; Malvasi, A.; Chiuri, V. E.; Leo, G. Biomolecular pathogenesis of borderline ovarian tumors: focusing target discovery through proteogenomics. *Curr. Cancer Drug Targets* **2010**, *10* (1), 107−116.
(16) Ellis, M. J.; Gillette, M.; Carr, S. A.; Paulovich, A. G.; Smith, R. D.; Rodland, K. K.; Townsend, R. R.; Kinsinger, C.; Mesri, M.; Rodriguez, H.; Liebler, D. C. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discovery* **2013**, *3* (10), 1108−1112.
(17) 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), 221−223.
(18) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature* **2014**, *509* (7502), 575−581.
(19) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509* (7502), 582−587.
(20) Ivankov, D. N.; Payne, S. H.; Galperin, M. Y.; Bonissone, S.; Pevzner, P. A.; Frishman, D. How many signal peptides are there in bacteria? *Environ. Microbiol* **2013**, *15* (4), 983−90.
(21) Krug, K.; Carpy, A.; Behrends, G.; Matic, K.; Soares, N. C.; Macek, B. Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol. Cell. Proteomics* **2013**, *12* (11), 3420−3430.
(22) Liu, S.; Im, H.; Bairoch, A.; Cristofanilli, M.; Chen, R.; Deutsch, E. W.; Dalton, S.; Fenyo, D.; Fanayan, S.; Gates, C.; Gaudet, P.; Hincapie, M.; Hanash, S.; Kim, H.; Jeong, S. K.; Lundberg, E.; Mias, G.; Menon, R.; Mu, Z.; Nice, E.; Paik, Y. K.; Uhlen, M.; Wells, L.; Wu, S. L.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Omenn, G. S.; Beavis, R. C.; Hancock, W. S. A chromosome-centric human proteome project (C-HPP) to characterize the sets of proteins encoded in chromosome 17. *J. Proteome Res.* **2013**, *12* (1), 45−57.
(23) 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.
(24) Wang, D.; Liu, Z.; Guo, F.; Diao, L.; Li, Y.; Zhang, X.; Huang, Z.; Li, D.; He, F. CAPER 2.0: an interactive, configurable, and extensible workflow-based platform to analyze data sets from the Chromosome-centric Human Proteome Project. *J. Proteome Res.* **2014**, *13* (1), 99−106.
(25) Fermin, D.; Allen, B. B.; Blackwell, T. W.; Menon, R.; Adamski, M.; Xu, Y.; Ulintz, P.; Omenn, G. S.; States, D. J. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.* **2006**, *7* (4), R35.
(26) Pawar, H.; Sahasrabuddhe, N. A.; Renuse, S.; Keerthikumar, S.; Sharma, J.; Kumar, G. S.; Venugopal, A.; Sekhar, N. R.; Kelkar, D. S.; Nemade, H.; Khobragade, S. N.; Muthusamy, B.; Kandasamy, K.; Harsha, H. C.; Chaerkady, R.; Patole, M. S.; Pandey, A. A proteogenomic approach to map the proteome of an unsequenced pathogen - Leishmania donovani. *Proteomics* **2012**, *12* (6), 832−844.
(27) Jagtap, P.; Goslinga, J.; Kooren, J. A.; McGowan, T.; Wroblewski, M. S.; Seymour, S. L.; Griffin, T. J. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **2013**, *13* (8), 1352−1357.
(28) Menon, R.; Omenn, G. S. Identification of alternatively spliced transcripts using a proteomic informatics approach. *Methods Mol. Biol.* **2011**, *696*, 319−326.
(29) Halvey, P. J.; Wang, X.; Wang, J.; Bhat, A. A.; Dhawan, P.; Li, M.; Zhang, B.; Liebler, D. C.; Slebos, R. J. Proteogenomic analysis reveals unanticipated adaptations of colorectal tumor cells to deficiencies in DNA mismatch repair. *Cancer Res.* **2014**, *74* (1), 387−397.
(30) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Scalf, M.; Smith, L. M. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J. Proteome Res.* **2014**, *13* (1), 228−240.
(31) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol. Cell. Proteomics* **2013**, *12* (8), 2341−2353.
(32) Blakeley, P.; Overton, I. M.; Hubbard, S. J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.* **2012**, *11* (11), 5221−5234.
(33) Wang, X.; Slebos, R. J.; Wang, D.; Halvey, P. J.; Tabb, D. L.; Liebler, D. C.; Zhang, B. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **2012**, *11* (2), 1009−1017.
(34) Kuhring, M.; Renard, B. Y. iPiG: integrating peptide spectrum matches into genome browser visualizations. *PLoS One* **2012**, *7* (12), e50246.

(35) Risk, B. A.; Spitzer, W. J.; Giddings, M. C. Peppy: proteogenomic search software. *J. Proteome Res.* **2013**, *12* (6), 3019−3025.
(36) Goecks, J.; Nekrutenko, A.; Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **2010**, *11* (8), R86.
(37) Bandhakavi, S.; Stone, M. D.; Onsongo, G.; Van Riper, S. K.; Griffin, T. J. A dynamic range compression and three-dimensional peptide fractionation analysis platform expands proteome coverage and the diagnostic potential of whole saliva. *J. Proteome Res.* **2009**, *8* (12), 5590−600.
(38) Johnson, J.; Chilton, J.; Jagtap, P.; Lynch, B.; Griffin, T. In *Reproducible Proteomic Workflows Using Extensions to the Galaxy Framework*, 61st ASMS Conference on Mass Spectrometry and Allied Topics, Minneapolis, MN, 2013.
(39) Jagtap, P.; McGowan, T.; Bandhakavi, S.; Tu, Z. J.; Seymour, S.; Griffin, T. J.; Rudney, J. D. Deep metaproteomic analysis of human salivary supernatant. *Proteomics* **2012**, *12* (7), 992−1001.
(40) Chen, T.; Yu, W. H.; Izard, J.; Baranova, O. V.; Lakshmanan, A.; Dewhirst, F. E. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* **2010**, *2010*, baq013.
(41) Shilov, I. V.; Seymour, S. L.; Patel, A. A.; Loboda, A.; Tang, W. H.; Keating, S. P.; Hunter, C. L.; Nuwaysir, L. M.; Schaeffer, D. A. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics* **2007**, *6* (9), 1638−1655.
(42) Tang, W. H.; Shilov, I. V.; Seymour, S. L. Nonlinear fitting method for determining local false discovery rates from decoy database searches. *J. Proteome Res.* **2008**, *7* (9), 3661−3667.
(43) MacLean, B.; Eng, J. K.; Beavis, R. C.; McIntosh, M. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* **2006**, *22* (22), 2830−2832.
(44) Robinson, J. T.; Thorvaldsdottir, H.; Winckler, W.; Guttman, M.; Lander, E. S.; Getz, G.; Mesirov, J. P. Integrative genomics viewer. *Nat. Biotechnol.* **2011**, *29* (1), 24−26.
(45) Kim, H. S.; Lyons, K. M.; Saitoh, E.; Azen, E. A.; Smithies, O.; Maeda, N. The structure and evolution of the human salivary proline-rich protein gene family. *Mamm. Genome* **1993**, *4* (1), 3−14.
(46) Carlson, D. M. Salivary proline-rich proteins: biochemistry, molecular biology, and regulation of expression. *Crit Rev. Oral Biol. Med.* **1993**, *4* (3−4), 495−502.
(47) Chicurel, M. Bioinformatics: bringing it all together. *Nature* **2002**, *419* (6908), 751, 753,755 passim.
(48) Palsson, B.; Zengler, K. The challenges of integrating multi-omic data sets. *Nat. Chem. Biol.* **2010**, *6* (11), 787−789.
(49) Chen, Y.; Zhang, J.; Xing, G.; Zhao, Y. Mascot-derived false positive peptide identifications revealed by manual analysis of tandem mass spectra. *J. Proteome Res.* **2009**, *8* (6), 3141−3147.
(50) Azen, E. A. Genetics of salivary protein polymorphisms. *Crit Rev. Oral Biol. Med.* **1993**, *4* (3−4), 479−485.
(51) Kim, H. S.; Smithies, O.; Maeda, N. A physical map of the human salivary proline-rich protein gene cluster covers over 700 kbp of DNA. *Genomics* **1990**, *6* (2), 260−267.
(52) Canon, F.; Giuliani, A.; Pate, F.; Sarni-Manchado, P. Ability of a salivary intrinsically unstructured protein to bind different tannin targets revealed by mass spectrometry. *Anal. Bioanal. Chem.* **2010**, *398* (2), 815−822.
(53) Murray, P. A.; Prakobphol, A.; Lee, T.; Hoover, C. I.; Fisher, S. J. Adherence of oral streptococci to salivary glycoproteins. *Infect. Immun.* **1992**, *60* (1), 31−38.
(54) Sheynkman, G. M.; Johnson, J. E.; Jagtap, P. D.; Shortreed, M. R.; Onsongo, G.; Frey, B. L.; Griffin, T. J.; Smith, L. M. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics* **2014**, *15*, 703.
(55) Blankenberg, D.; Hillman-Jackson, J. Analysis of next-generation sequencing data using Galaxy. *Methods Mol. Biol.* **2014**, *1150*, 21−43.
(56) Keller, O.; Kollmar, M.; Stanke, M.; Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **2011**, *27* (6), 757−763.
(57) Barash, Y.; Vaquero-Garcia, J.; Gonzalez-Vallinas, J.; Xiong, H. Y.; Gao, W.; Lee, L. J.; Frey, B. J. AVISPA: a web tool for the prediction and analysis of alternative splicing. *Genome Biol.* **2013**, *14* (10), R114.
(58) Blankenberg, D.; Johnson, J. E.; Taylor, J.; Nekrutenko, A. Wrangling Galaxy's reference data. *Bioinformatics* **2014**, *30* (13), 1917−1919.
(59) Vaudel, M.; Barsnes, H.; Berven, F. S.; Sickmann, A.; Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **2011**, *11* (5), 996−999.
(60) Gottschalk, B.; Jagtap, P.; Barsnes, H.; Vaudel, M.; Gruening, B.; Cooke, I.; Johnson, J.; Chilton, J.; Higgins, L.; Markowski, T.; Wennblom, T.; Lamblin, A.; Chen, Y.; Kim, S.; Martens, L.; Griffin, T. In *Community-Based Development and Evaluation of Biological Mass Spectrometry Software via the Galaxy Tool Shed*, 62nd ASMS Conference on Mass Spectrometry and Allied Topics, Baltimore, MD, 2014.