Journal of proteome research

Technical Note

pubs.acs.org/jpr

PGx: Putting Peptides to BED

Manor Askenazi*

Biomedical Hosting LLC, 33 Lewis Avenue, Arlington, Massachusetts 02474, United States

Kelly V. Ruggles

NYU Langone Medical Center, 227 East 30th Street, New York, New York 10016, United States

David Fenyö*

NYU Langone Medical Center, 227 East 30th Street, New York, New York 10016, United States

ABSTRACT: Every molecular player in the cast of biology's central dogma is being sequenced and quantified with increasing ease and coverage. To bring the resulting genomic, transcriptomic, and proteomic data sets into coherence, tools must be developed that do not constrain data acquisition and analytics in any way but rather provide simple links across previously acquired data sets with minimal preprocessing and hassle. Here we present such a tool: PGx, which supports proteogenomic integration of mass spectrometry proteomics data with next-generation sequencing by mapping identified peptides onto their putative genomic coordinates.

KEYWORDS: *proteogenomics, proteogenomic mapping, proteomics*

1. INTRODUCTION

Systems biology is premised on the ability to integrate data sets covering all aspects of cellular biochemistry. One such integrative approach is termed proteogenomics[1] and is defined as the integration of proteomic and genomic information, usually referring to the use of mass spectrometry (MS)-based proteomics to improve gene annotation. The field has taken off thanks to recent improvements in both next-generation sequencing (NGS) and proteomic methodologies. It has become feasible both in terms of cost and time to sequence the DNA and RNA of every sample set being studied by MS proteomics. Additionally, modern mass spectrometers are able to sequence peptides at such a depth of coverage that they are now becoming useful in the very identification and validation of genes (whereas historically, proteomics depended entirely on a complete predicted proteome). The integration of proteomics and genomics can therefore improve our understanding of both genomic annotation and of course the functional characterization of protein products in their biological context.

As the results of proteogenomics research accumulate, be they in the form of genome annotation, splice isoform prediction, or novel protein discovery, there arises a pressing need to map and visualize all data types onto the same unified coordinate system. There currently exist many tools for the analysis and display of genomic features where the coordinate system of choice is naturally the underlying reference sequence for the organism being studied. This is true even for the most advanced and challenging forms of next generation sequencing data. It follows naturally that the ideal unified coordinate system for proteogenomics should remain genomic in nature. Indeed, effective tools that can map MS-based proteomics results onto genomic coordinates have recently become available (Peppy,[2] Proteogenomic Mapping Tool,[3] Pepline,[4] MS-Dictionary,[5] GappedDictionary,[6] IggyPep,[7] MSProGene,[8] ProteoAnnotator,[9] PGNexus,[10] and GalaxyP[11]); however, these tools are usually couched in a relatively involved and comprehensive pipeline (e.g., the GalaxyP pipeline consists of up to 140 steps) and typically impose a specific mass-informatic[12] workflow on the practitioner, by, for example, requiring the generation of short peptide sequence tags (PSTs) or some complex form of de novo peptide sequencing followed by a lookup against the full six-frame translation of the genomic sequence. Our experience suggests that a more common scenario involves the production, by the genomic arm of the workflow, of a (liberally) predicted proteome (containing what is assumed to be a superset of the observable proteome) so as to leverage existing PSM search engines (such as Mascot,[13] Sequest,[14] X!Tandem[15]) that require a straightforward representation of the predicted proteome (in the form of a FASTA file). We have thus identified the need for an exceedingly focused tool following the Unix tradition ("do one thing and do it well"[16]) that simply leverages the analysis done at the genomic level (represented as a BED file to accompany the FASTA file provided to the search engine), thereby enabling the efficient mapping of proteogenomics results onto the common sequence map. The coupling between the proteomic workflow and the genomic arms of the research

Special Issue: Large-Scale Computational Mass Spectrometry and Multi-Omics

Received: September 16, 2015
Published: December 7, 2015

DOI: 10.1021/acs.jproteome.5b00870
J. Proteome Res. 2016, 15, 795−799

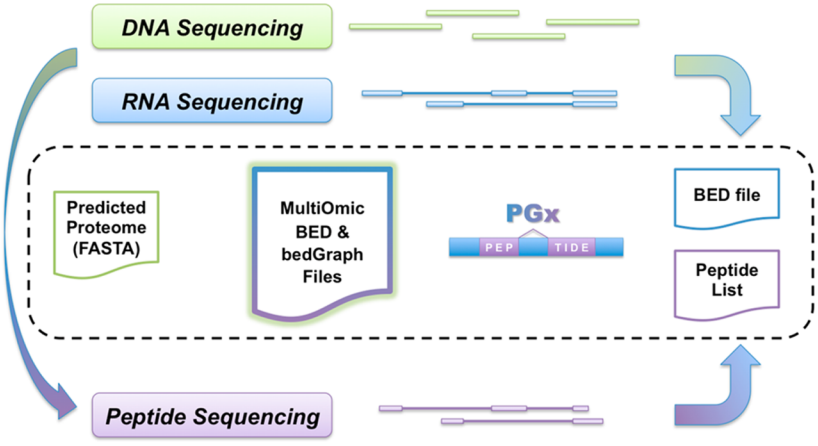

Figure 1. PGx integrates all "ome" data sets using only a BED file, FASTA file, and a peptide list as input.

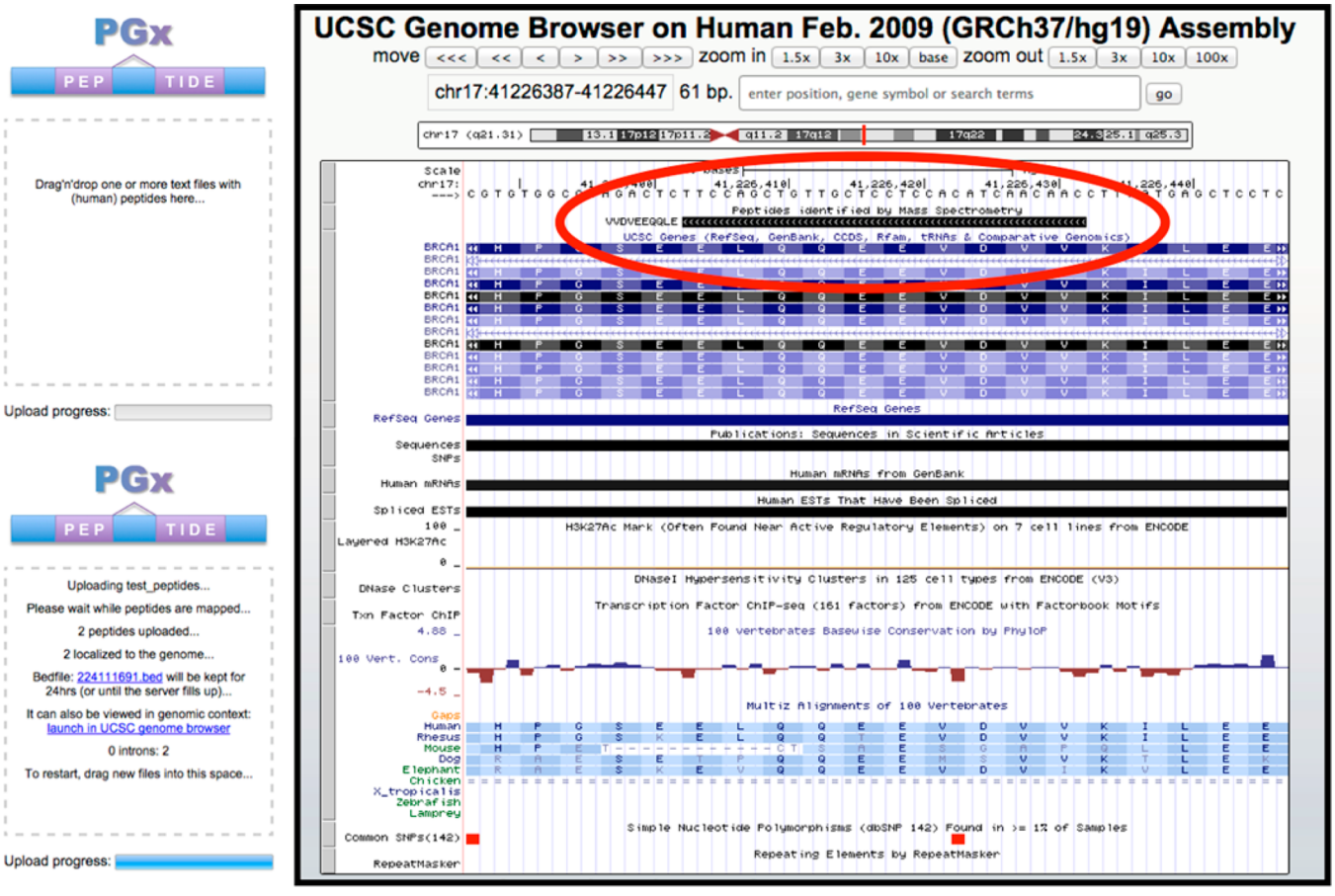

Figure 2. Typical interaction with the PGx Web site: The user simply drags a file containing query peptides onto the dashed rectangle. The example text file yielding this visualization is provided on the Web site itself.

project is minimized, which allows the proteomic analysis to proceed using standard proteomic software tools. Our solution is implemented as a Python framework called PGx, which allows for sensitive, relevant, and rapid proteogenomic data integration either at the command line or through a user-friendly web-accessible interface. The key distinguishing property of PGx is that it relies solely on three standard files that succinctly summarize the contribution of the three main arms of the proteogenomics effort: a BED file integrating the results of DNA and RNA sequencing, a FASTA file representing the complete predicted proteome and a peptide list representing the results of peptide sequencing (Figure 1).

Our choice of BED files as input stems from the fact that nearly every genome browser supports the visualization of this file format, including: the Broad Institute's Interactive Genome Browser,[17] UCSC's Genome Browser,[18] the WashU Epigenome Browser,[19] and the Ensembl Genome Browser.[20] Additionally, many alignment and genomic tools output data in the form of a BED file such as the RNA-Seq alignment tool TopHat,[21] which outputs a splice junction file in the form of a BED file and the commonly used bedtools software,[22] which is

DOI: 10.1021/acs.jproteome.5b00870
J. Proteome Res. 2016, 15, 795−799

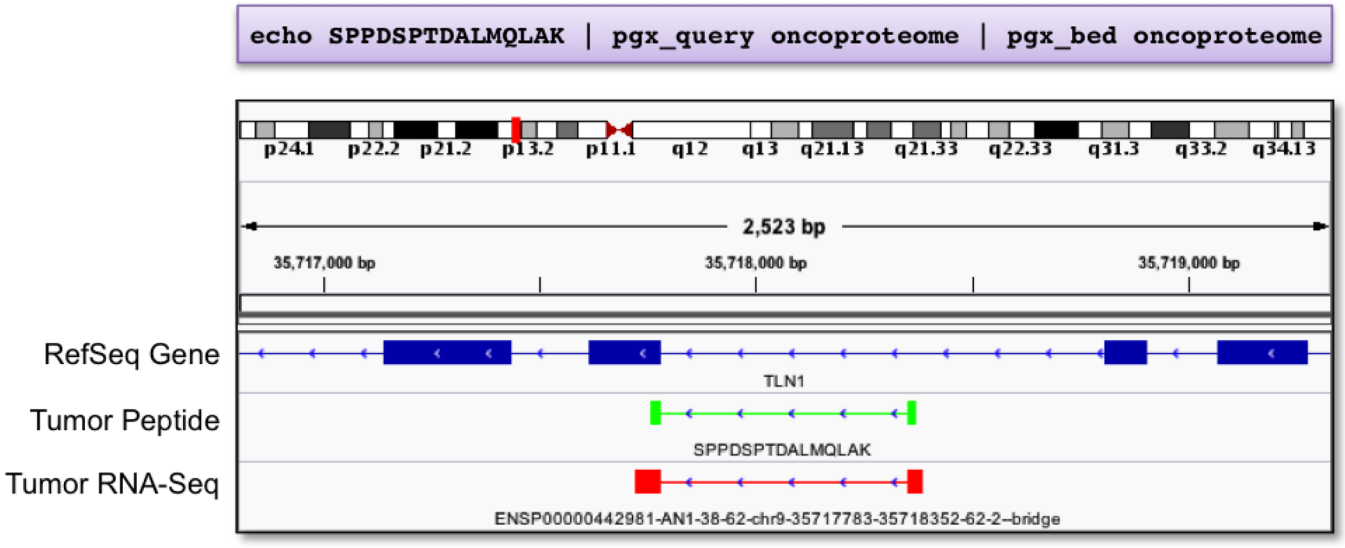

Figure 3. Example of a novel peptide resulting from intronic expression is mapped using PGx framework. (The exact command-line required to generate the final bed file is shown in the purple inset; for more details see the tutorial included with the source code.)

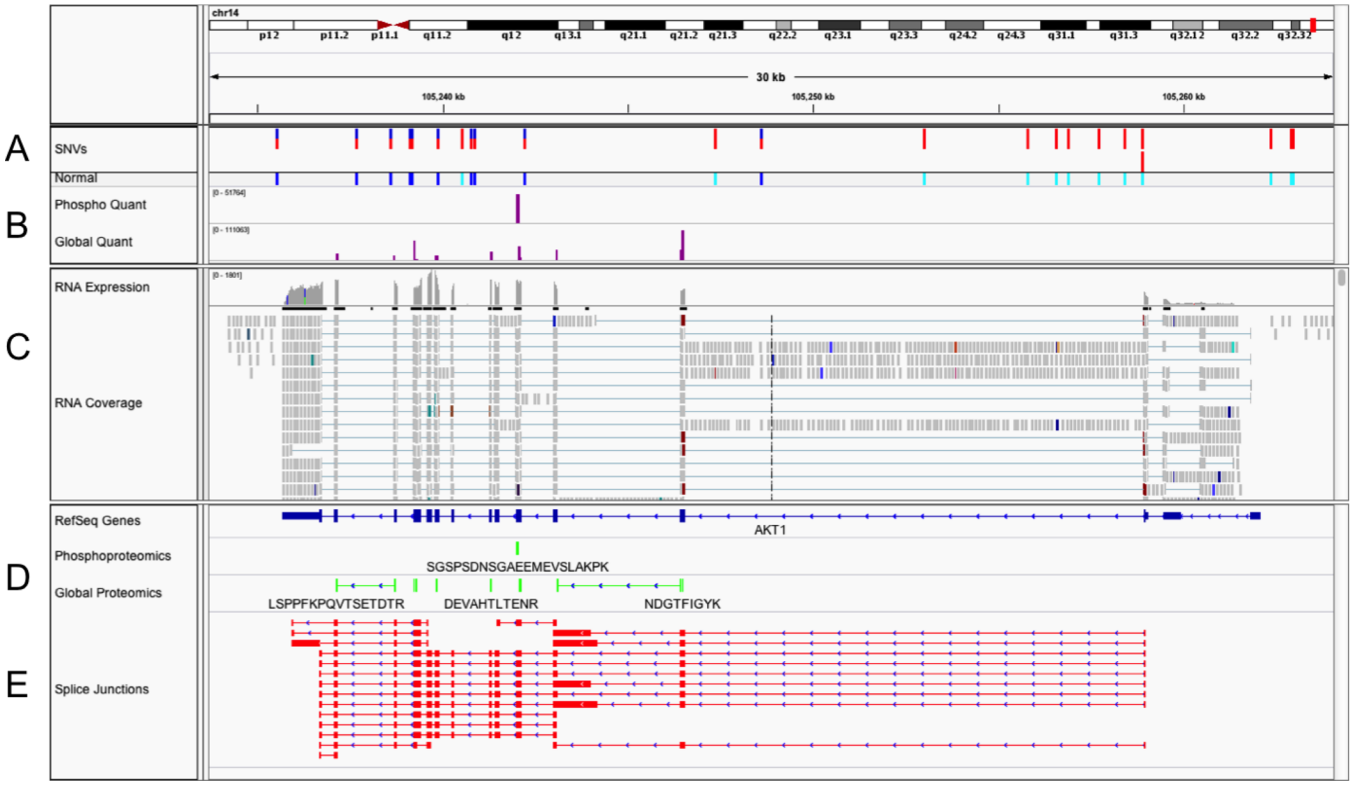

Figure 4. Multi-omic integration: The quantitative peptide track is provided by a PGx bedGraph file. Contains data from a tumor sample for (A) single nucleotide variants (SNVs) (from VCF files), (B) global and phosphoproteomic quantitative data (from PGx derived bedGraph files), (C) RNA expression and coverage data (BAM file), (D) global and phosphoproteomic peptide mapping (from PGx derived bed files), and (E) RNA splice junction predictions (from junction bed file).

able to complete a wide range of genomic analysis methods using the BED format. Because of its frequent use by the genomic and transcriptomic communities, simple conversion of proteomics data to the familiar BED format allows for seamless inclusion of proteomics data in already existing genome-based tools.

The FASTA file, on the contrary embodies a prior interaction between the sequencing efforts in that it is usually a key enabler in sample-specific proteomics: Given the diversity of protein isoforms in different cell types and the growing affordability of next generation sequencing (NGS) technology, it has become advantageous to create sample-specific protein sequence databases for comprehensive peptide identification. RNA-Seq and genome sequencing information can be used to create these databases, incorporating variant proteins, alternatively spliced isoforms, and novel expression, as coded within the

DOI: 10.1021/acs.jproteome.5b00870
J. Proteome Res. 2016, 15, 795−799

genome and transcriptome, allowing for the identification of sample specific peptides from the tandem MS analysis.[23−25] In the Clinical Proteomics Tumor Analysis Consortium (CPTAC) we have combined patient-specific protein databases and used PGx to map identified peptide sequences and a relative estimate of their abundance (via Spectral Counting, Figure 4) onto sample-specific genomic coordinates, providing easy-to-use proteogenomic integration techniques for these patient-centric studies.

Finally, the choice of a simple peptide list as the third PGx input file minimizes any formatting requirement by the proteomic software. PGx simply researches the protein sequence space provided in the FASTA file, thereby maximizing the decoupling between the various tools in the proteogenomic workflow.

2. IMPLEMENTATION

PGx is an open-source project released under the MIT license and is also publicly accessible via a web-based API supporting access by researchers using a web browser (Figure 2) and programmers using http-based API calls to a simple RESTful interface. It is implemented in pure Python and is therefore extremely portable and easy to customize. PGx leverages a memory resident indexing scheme[26] to perform a very fast (essentially interactive) mapping of peptides to genomic sequence. PGx performs this mapping by using two indexes for each protein sequence database (e.g., RefSeq, Ensembl, or a sample-specific database based on RNA-Seq and whole genome sequencing or exome sequencing data). The first index is a peptide dictionary that contains all four amino acid peptides in the protein sequence database. The dictionary is designed to consider leucine and isoleucine as equivalent because they cannot be distinguished by typical mass spectrometry workflows. The dictionary is used to rapidly lookup and to retrieve all proteins that might contain an experimentally observed peptide based on the occurrence of its constituent 4-mers. The presence of the peptide is then validated in every candidate protein. The second index is a mapping of each protein sequence in the database onto the genome, and this index is used in the second step to map each peptide onto its genomic coordinates. PGx supports the mapping of many peptides at the same time, and the submission of a list with peptide sequences and their quantities will return a BED (qualitative information) and a bedGraph (quantitative information) that can be used to visualize the proteomics data using a broad range of genome browsers such as the UCSC browser (Figure 2) or IGV (Figures 3 and 4).

PGx is available for testing against the standard Refseq build at the following Web site http://pgx.fenyolab.org. The site simply expects a file containing peptides to be "drag and dropped" onto it. The results are then available for download and visualization on the UCSC Genome Browser. (The whole process is shown in Figure 2.) A test file is made available on the site, which is the exact input used to generate the Figure.

While the Web site is useful in gaining an understanding of PGx's functionality, the framework is implemented and distributed first and foremost as a collection of Python-based command line tools. In addition to the core-indexing query and formatting scripts, the framework provides some support functionality such as the automatic downloading of genomic resources (e.g., RefSeq[27] gpff files) or the ability to query for the position of peptides that might be present only as nsSNPs[28] relative to the existing sequence base. The complete set of scripts is hosted on github, along with end-user documentation in the form of a tutorial and a test data set capable of regenerating Figure 3. In brief, custom proteomes are stored in directories containing two files called: "proteome.fasta" and "proteome.bed" referring, respectively, to the sequence space and its genomic mapping. All pgx commands take a proteome (directory path) as a first input and a stream or named file as second argument. As a result, the full power and succinctness of command line streaming can be leveraged, resulting in a simple one-liner capable of generating Figure 3 (purple inset).

3. DISCUSSION

PGx allows for seamless integration of proteomic mapping and quantitation data into pre-existing multi-omic pipelines. Two examples of this are shown. Figure 3 demonstrates the ability of PGx to map peptides to the genome in cases where a peptide is not contained within the reference protein database. Here a novel splice junction was identified by RNA-seq within the intronic region of the mitochondrial cochaperone HSCB in a tumor sample. Using a proteogenomic-based method of peptide searching in which these novel junction sites were included in the search database,[23−25] one is able to identify the peptide "SPPSDPTDALMQLAK" corresponding to the same novel intronic expression and subsequent PGx processing allows for the visualization of this peptide within a genomic context.

Furthermore, Figure 4 demonstrates how PGx can be used to easily obtain a comprehensive visualization of genomic, transcriptomic, and proteomic data. PGx files can be directly uploaded into the IGV or UCSC genome browser to display peptide mapping (Figure 4D) or peptide quantitation (Figure 4B) alongside RNA expression and coverage data (Figure 4C), RNA-seq splice junction mapping (Figure 4E), and genomic single nucleotide variants (SNVs) (Figure 4A). In this example, data from whole genome sequencing, RNA-seq, and quantitative MS/MS of a tumor sample were mapped for the serine/threonine kinase, AKT1.

4. CONCLUSIONS

We believe that PGx represents a useful contribution to the software toolset of any proteogenomics practitioner because it does not impose any mass-informatic on the proteomics branch of the workflow but simply relies on three files summarizing the results of the intermediate sequencing efforts to establish full data integration: a BED file, a FASTA file, and a set of peptides. A live instance of PGx supporting peptide mapping onto the standard RefSeq build is freely available for public use at http://pgx.fenyolab.org, and the Python scripts licensed under the MIT license are hosted at https://github.com/FenyoLab/PGx.

AUTHOR INFORMATION

Corresponding Authors

*M.A.: E-mail: manor@biomedical.hosting.
*D.F.: E-mail: david@fenyolab.org.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by funding provided by the National Institutes of Health through grant CA160035 and contract S13-068 from Leidos. This work has utilized computing resources at the High Performance Computing Facility of the Center for

DOI: 10.1021/acs.jproteome.5b00870
J. Proteome Res. 2016, 15, 795−799

Health Informatics and Bioinformatics at the New York University Langone Medical Center.

REFERENCES

(1) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **2014**, *11* (11), 1114−1125.
(2) Risk, B. A.; Spitzer, W. J.; Giddings, M. C. Peppy: proteogenomic search software. *J. Proteome Res.* **2013**, *12* (6), 3019−3025.
(3) Sanders, W. S.; Wang, N.; Bridges, S. M.; Malone, B. M.; Dandass, Y. S.; McCarthy, F. M.; Nanduri, B.; Lawrence, M. L.; Burgess, S. C. The proteogenomic mapping tool. *BMC Bioinf.* **2011**, *12*, 115.
(4) Ferro, M.; Tardif, M.; Reguer, E.; Cahuzac, R.; Bruley, C.; Vermat, T.; Nugues, E.; Vigouroux, M.; Vandenbrouck, Y.; Garin, J.; et al. PepLine: a software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences. *J. Proteome Res.* **2008**, *7* (5), 1873−1883.
(5) Kim, S.; Gupta, N.; Bandeira, N.; Pevzner, P. A. Spectral Dictionaries Integrating de novo Peptide Sequencing with Database Search of Tandem Mass Spectra. *Mol. Cell. Proteomics* **2009**, *8* (1), 53−69.
(6) Jeong, K.; Kim, S.; Bandeira, N.; Pevzner, P. A. Gapped Spectral Dictionaries and Their Applications for Database Searches of Tandem Mass Spectra. *Mol. Cell. Proteomics* **2011**, *10* (6), M110.002220.
(7) Menschaert, G.; Vandekerckhove, T. T. M.; Baggerman, G.; Landuyt, B.; Sweedler, J. V.; Schoofs, L.; Luyten, W.; Van Criekinge, W. A hybrid, de novo based, genome-wide database search approach applied to the sea urchin neuropeptidome. *J. Proteome Res.* **2010**, *9* (2), 990−996.
(8) Zickmann, F.; Renard, B. Y. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. Bioinforma. *Bioinformatics* **2015**, *31* (12), i106−i115.
(9) Ghali, F.; Krishna, R.; Perkins, S.; Collins, A.; Xia, D.; Wastling, J.; Jones, A. R. ProteoAnnotator−open source proteogenomics annotation software supporting PSI standards. *Proteomics* **2014**, *14* (23−24), 2731−2741.
(10) Pang, C. N. I.; Tay, A. P.; Aya, C.; Twine, N. A.; Harkness, L.; Hart-Smith, G.; Chia, S. Z.; Chen, Z.; Deshpande, N. P.; Kaakoush, N. O.; et al. 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.
(11) 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), 5898−5908.
(12) Askenazi, M.; Linial, M. Mass Informatics: From Mass Spectrometry Peaks to Biological Pathways. *Isr. J. Chem.* **2013**, *53* (3−4), 157−165.
(13) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−3567.
(14) Eng, J. K.; McCormack, A. L.; Yates, J. R., III An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−989.
(15) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466−1467.
(16) Raymond, E. S. *The Art of UNIX Programming*; Addison-Wesley Professional, 2003.
(17) Thorvaldsdóttir, H.; Robinson, J. T.; Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinf.* **2013**, *14* (2), 178−192.
(18) Rosenbloom, K. R.; Armstrong, J.; Barber, G. P.; Casper, J.; Clawson, H.; Diekhans, M.; Dreszer, T. R.; Fujita, P. A.; Guruvadoo, L.; Haeussler, M.; et al. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **2015**, *43* (D1), D670−D681.
(19) Zhou, X.; Maricque, B.; Xie, M.; Li, D.; Sundaram, V.; Martin, E. A.; Koebbe, B. C.; Nielsen, C.; Hirst, M.; Farnham, P.; et al. The Human Epigenome Browser at Washington University. *Nat. Methods* **2011**, *8* (12), 989−990.
(20) Stalker, J.; Gibbins, B.; Meidl, P.; Smith, J.; Spooner, W.; Hotz, H.-R.; Cox, A. V. The Ensembl Web Site: Mechanics of a Genome Browser. *Genome Res.* **2004**, *14* (5), 951−955.
(21) Kim, D.; Pertea, G.; Trapnell, C.; Pimentel, H.; Kelley, R.; Salzberg, S. L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **2013**, *14* (4), R36.
(22) Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* **2014**, *47*, 11.12.1−11.12.34.
(23) Li, J.; Su, Z.; Ma, Z.-Q.; Slebos, R. J. C.; Halvey, P.; Tabb, D. L.; Liebler, D. C.; Pao, W.; Zhang, B. A Bioinformatics Workflow for Variant Peptide Detection in Shotgun Proteomics. *Mol. Cell. Proteomics* **2011**, *10* (5), M110.006536.
(24) Wang, X.; Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **2013**, *29* (24), 3235−3237.
(25) Castellana, N. E.; Shen, Z.; He, Y.; Walley, J. W.; Cassidy, C. J.; Briggs, S. P.; Bafna, V. An Automated Proteogenomic Method Uses Mass Spectrometry to Reveal Novel Genes in Zea mays. *Mol. Cell. Proteomics* **2014**, *13* (1), 157−167.
(26) Askenazi, M.; Marto, J. A.; Linial, M. The complete peptide dictionary - A meta-proteomics resource. *Proteomics* **2010**, *10* (23), 4306−4310.
(27) Pruitt, K. D.; Brown, G. R.; Hiatt, S. M.; Thibaud-Nissen, F.; Astashyn, A.; Ermolaeva, O.; Farrell, C. M.; Hart, J.; Landrum, M. J.; McGarvey, K. M.; et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **2014**, *42* (1), D756−D763.
(28) Schaefer, C.; Meier, A.; Rost, B.; Bromberg, Y. SNPdbe: constructing an nsSNP functional impacts database. *Bioinformatics* **2012**, *28* (4), 601−602.

DOI: 10.1021/acs.jproteome.5b00870
J. Proteome Res. 2016, 15, 795−799