Published in final edited form as:
J Proteome Res. 2013 June 7; 12(6): 3019–3025. doi:10.1021/pr400208w.

Peppy: Proteogenomic Search Software

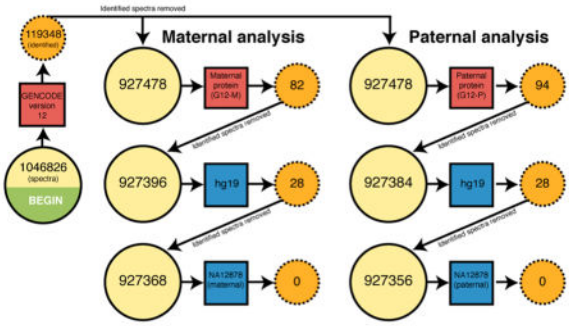
Brian A. Risk[†,‡,*], **Wendy J. Spitzer**[‡], and **Morgan C. Giddings**[†,‡]

[†]Department of Biochemistry & Biophysics, UNC School of Medicine, Chapel Hill, North Carolina 27599, United States

[‡]College of Arts and Sciences, Boise State University, Boise, Idaho 83725, United States

Abstract

Proteogenomic searching is a useful method for identifying novel proteins, annotating genes and detecting peptides unique to an individual genome. The approach, however, can be laborious, as it often requires search segmentation and the use of several unintegrated tools. Furthermore, many proteogenomic efforts have been limited to small genomes, as large genomes can prove impractical due to the required amount of computer memory and computation time. We present Peppy, a software tool designed to perform every necessary task of proteogenomic searches quickly, accurately and automatically. The software generates a peptide database from a genome, tracks peptide loci, matches peptides to MS/MS spectra and assigns confidence values to those matches. Peppy automatically performs a decoy database generation, search and analysis to return identifications at the desired false discovery rate threshold. Written in Java for cross-platform execution, the software is fully multithreaded for enhanced speed. The program can run on regular desktop computers, opening the doors of proteogenomic searching to a wider audience of proteomics and genomics researchers. Peppy is available at http://geneffects.com/peppy.



Keywords

MS/MS; tandem mass spectrometry; protein identification; proteogenomics; gene annotation

[*]Corresponding Author: brian@geneffects.com. Phone: +1 (919) 260 3063. Fax: (919) 843-3328.

Notes

The authors declare no competing financial interest.

INTRODUCTION

Searching MS/MS data against a protein database is an established method for determining the protein composition of a biological sample. However, as a protein database contains only known proteins, the ability to discover new or altered proteins is limited. In the analysis of diseases such as cancer, aberrant forms of expression may produce disease-related proteins not found in the reference protein data sets;[1] therefore, searching standard protein databases (e.g., Uni-ProtKB/Swiss-Prot) will miss these cancer-related, novel proteins.

Proteogenomic searches have led to the discovery of novel coding regions, the determination of correct start and stop positions of a gene, and the verification of splice variants. When Castellana et al. performed a proteogenomic search of *Arabidopsis thaliana*,[2] they found 778 new protein-coding genes, refined 695 existing gene annotations, and estimated that 13% of the *A. thaliana* annotated proteome was incomplete. In another large-scale proteogenomic analysis, Payne et al. found nonannotated genes, corrected protein boundaries and removed incorrect annotations of open reading frames (ORFs) for *Yersinia pestis KIM* and 21 other *Yersinia* genomes, ultimately altering 141 incorrectly annotated gene models of the bacteria.[3]

Further, as the speed of genome sequencing and its cost-effectiveness rapidly improve, the resulting explosion of genomic data can vastly benefit the world of proteomics.[4] Proteogenomic analysis using the subject's genome that produced the biological sample can potentially reveal phenotypic variants of peptides or entire proteins that are not normally coded (e.g., their expression is the result of a drastic genomic alteration).

The Steps of a Proteogenomic Search

A proteogenomic search is largely similar to typical protein searches with a basic workflow illustrated in Figure 1. The main difference from searching a reference protein database is that the theoretical peptide database to which the spectra are compared is derived from an *in silico* digestion of a six-frame translation of the full genome sequence and that the genomic locus of every peptide must be tracked. A six-frame translation of a nucleotide sequence is necessary because transcripts may start at any index on either strand of a chromosome. The peptide-spectrum match (PSM) software compares the spectra against the theoretical peptide database and returns a score for each match; there exist a plurality of options for this step.[5] If this score is not a global confidence value (e.g., *E*-value, *P*-value), then such a value can be assessed after the fact.[6]

False discovery rate (FDR) estimation is a critical final step in any proteogenomic (or proteomic) search.[7] False positive identifications are unavoidable but employing FDR thresholds allows a researcher to control the true positive saturation of the resulting set of PSMs. Several methods exist for estimating FDRs;[8] we have elected to implement the target-decoy strategy set forth by Elias and Gygi,[9] which necessitates the creation of a decoy of the six-frame translation.

The final output of a proteogenomic search is a table of PSMs that typically contains the peptide's amino acid sequence, identifying information for the spectrum (e.g., file name,

locus within the file), a confidence score conveying the quality of the PSM and information regarding the coding location of the given peptide (e.g., chromosome, strand, start and stop loci). These results are often cross-referenced with a search on a standard protein database so that the novel peptides can be immediately identified.[10] The peptides not present within the reference gene set suggest candidates of unannotated or incorrectly annotated genes.

Challenges for Implementation of Proteogenomic Searching

The primary hurdles of proteogenomic searches, as compared to protein database searches, have little to do with the scientific complexities, but are a product of the logistics of the task. Some of these logistics include issues of task segmentation to accommodate limited computer memory, the need for different tools to perform the six-frame translation, peptide loci mapping and FDR analysis.

As the protein-coding region of the human genome comprises ~1% of the nucleotide content,[11] it stands to reason that a peptide database derived from a six-frame translation will be on the order of 600 times larger than one produced from a reference protein database. This increased search space has presented challenges for implementation due to insufficient computer memory. The solution we have implemented is to segment the genome into manageable sections, search the full set of spectra within each set of peptides and combine the individual results into one master results set. Peppy has been optimized to automatically segment the genome and then reaggregate those segments, allowing for once-impossible searching of genomes to be performed on regular desktop computers, in reasonable amounts of time.

Furthermore, the generation of a six-frame translation of a genome is not always an automated task. For example, the generation of a six-frame translation can be done via an online tool such as the BCM Search Launcher;[12] however, online tools are only useful for short genomic sequences. Alternately, the translation can be done via a script such as the Perl script included with the InsPecT software package.[13] Though tools are available to handle the translation of the genome, they require human operation, and the output of one step may not easily integrate into the input of the next step. The lack of integration can cause problems of human error, and in getting the data from disparate systems to ‘speak to’ one another.

Reporting the genomic loci of a peptide typically entails a separate process wherein the amino acid sequence must be analyzed by a tool to find the probable locations of its encoding.[14] Peppy negates the need for this extra step by retaining a peptide’s genomic location as it is translated from the genome. This genomic coding location is associated with the peptide throughout the proteogenomic search process, so the peptide’s location never needs to be ‘found’. Mapping back to the peptide’s encoding locus is therefore an unnecessary step.

Limited computer memory presents the primary challenge estimating false discovery rates for six-frame searches. Common target/decoy FDR techniques call for concatenating the decoy database with the reference database wherein the decoy proteins contain header information identifying them as decoys.[15] As a six-frame target/decoy database will be ~600

times that of a reference protein target/decoy, measures for task segmentation are typically required. However, this segmentation would then require extra steps to unify all results, resolve instances of multiple peptide matches between result sets to a given spectrum, and potentially complicate deriving peptide coding locations.

With each separate tool requiring human intervention, there is added time, complexity and an increased chance for human error. In their survey of proteogenomics, Renuse et al. addressed the lack of an integrated solution as a major hurdle for proteogenomic analysis. Peppy is a single, integrated tool and the automation of these steps creates a streamlined pipeline for data, from processing of genome files to final FDR analysis, negating the need for separate tools to perform those functions and thereby reducing required effort and the possibility of error.

MATERIALS AND METHODS

To use Peppy, the user must provide three input sources: (i) a directory containing all spectral data, (ii) a directory containing genomic or proteomic sequences, and (iii) a user-defined properties file. The properties that can be specified in this file are those familiar to users of protein identification software, such as fragment tolerance, precursor tolerance, the maximum number of missed cleavages in a generated peptide, etc.

Generating the Peptide Database

The first stage of Peppy's execution is generating the collection of peptides from the genome, i.e., creating the "peptide database". This is accomplished via a six-frame translation and *in silico* digestion of the sequence files. Whole genome proteogenomic searching tasks often involve a peptide database so large that common desktop computers do not have sufficient RAM to process it all at once. Peppy accommodates memory limitation by performing the translation and digestion on segments of each chromosome. The number of nucleotides in a given segment is user-adjustable to allow for a variety of memory requirements. This segmentation produces manageable sets of peptides against which the full set of spectra are searched. Results from each search are aggregated into a master results list. A small number of nucleotides (120) overlap between the segments to account for peptides that span the segment boundaries; redundant peptides created by this overlapping process are eliminated in the final stage of PSM evaluation. Both the translation and digestion processes are multithreaded to take advantage of the speed afforded by multicore processors. Peptide loci with respect to the segment locus is resolved to global genomic start and stop coordinates that are stored along with the peptides during the digestion stage. This locus tracking renders the mapping of PSMs back to their genomic loci an automated process.

Loading the Spectra

As the spectral files are loaded, measures are taken to track and filter them. Each spectrum is assigned an MD5 hash algorithm identification string based on the peak mass and abundance data of the spectrum. The MD5 can be used as an identifier that is independent of spectrum name or file locus and can facilitate spectrum tracking when search results are merged into a

unified database. After a spectrum is loaded and its MD5 hash has been generated, spectral cleaning is applied. Intelligently reducing the number of peaks in a spectrum can significantly reduce the time required to process the quality of a PSM, without negatively impacting sensitivity. Spectral cleaning can also reduce false positives as noise peaks increase the chances of erroneous alignments with theoretical ions.[16] The built-in spectrum cleaning method is similar to a technique employed by Ascore[17] wherein a spectrum is broken into 100 Da windows and the 10 most intense peaks for each window are retained. We tested several mass window sizes and number of peaks to retain per window and found that 100/10 increases positive identifications for many data sets. These values can be altered within the properties file should the user require different settings.

Matching and Scoring Peptides to Spectra

Peppy contains a novel PSM scoring system[18] that combines estimated *P*-values for ion matches, above-median matched ion abundances and appropriate relative b- and y-ion abundances. Also included with Peppy is an implementation of the PSM scoring system found in Morpheus.[19] Users may select which PSM scoring system is used for analysis by designating this in the properties file. The Peppy source code uses a generic constructor when instantiating PSM objects; this framework facilitates the efforts of developers interested in crafting their own scoring system.

Peptide/spectrum matching and scoring is accomplished with parallel programming techniques that fully exploit the power of multicore computers. Comparing a peptide to a spectrum and assigning a score for that PSM is an event independent of all other peptide/spectrum comparisons and scores. This process is easily parallelizable as it is not prone to data synchronization issues. Parallelization is implemented via several worker threads all controlled by a thread server. The number of available computer processors that are detected by the software determines the number of worker threads. Thus, if the program is running on an 8-core machine, 8 worker threads will be created. Figure 2 shows Peppy's speed performance vs the number of processor cores, using three separate data sets.

The function of the thread server is to marshal the division of labor among the worker threads. The thread server gives each worker thread one spectrum and the entire collection of peptides. A worker will search for the best scoring peptides for its given spectrum by comparing that spectrum to every peptide within a given precursor tolerance window of a spectrum's precursor mass. As the peptide list is sorted by mass, this process is optimized by using binary search to find the lowest and highest mass peptides within the desired precursor tolerance then iteratively comparing the spectrum to every peptide between those indices within the peptide list. The worker thread reports its findings back to the server and the server assigns the worker thread a new spectrum to investigate. This process repeats until all spectra have been processed.

False Discovery Rates

The central question for every peptide identification is: is this true? Did a particular amino acid sequence produce the observed spectrum, or is the peptide/spectrum match merely a product of coincidence? As previously mentioned, integrated into Peppy is the functionality

to calculate false discovery rates for a given set of spectra using the target-decoy method described by Elias and Gygi. The first step in FDR analysis is the creation of the decoy; an ideal decoy database contains no true positives and mimics the characteristics of the target database (e.g., cardinality, mass distribution, acid content and peptide length distribution). The decoy peptide database is generated in a manner similar to that of the target database: a six-frame translation of the genome is split into theoretical proteins along the stop codons. The amino acid sequences of each theoretical protein are then reversed and *in silico* digested according to the rules for the given protease. This creates a peptide database that is similar in peptide count, amino acid content and peptide length distribution as that of the "target" database, but has a low probability for containing a significant numbers of true peptides. When a search is performed, Elias et al. stress that "competition" must exist between target and decoy results such that only the best scoring match for any given spectrum is used for the FDR calculation. With typical protein database searches, this is most easily accomplished by appending the decoy to the target database. However, computer memory constraints can be a limiting factor for this approach with a proteogenomic database; therefore, database segmentation and results combination must be performed in a way that produces the desired competition. To that end, the search results are reduced to only the top match for every spectrum, which are then ordered by confidence score. If a decoy peptide provides the best match for a given spectrum, then it is the one PSM entered into the list for that spectrum. The FDR at each confidence threshold is calculated as two times the number of passing decoy matches divided by the full set of passing matches.

To speed calculation, one may set the parameter of size of spectra subset to be used for FDR calculation. If, for example, the body of spectra contains 1 000 000 files, a researcher may elect to use a random subset of 10 000 so that FDR calculation will take about 1/100th of the time, yet still remain statistically significant.[20]

EXPERIMENTAL PROCEDURE

Peppy was used for the creation of the ENCODE proteogenomic results;[21] here we revisit the proteogenomic analysis of the ENCODE Tier 1 cell line GM12878 with the aim of demonstrating how proteogenomic searches can be used in conjunction with proteomic searches as a first step to reveal instances of undocumented proteins/isoforms. While the previous analysis mapped spectra to their probable coding locations in order to ascertain protein expression, this analysis is focused on identifying subject-specific peptides and potential instances of novel coding. This will be accomplished with two primary differences in approach: (i) The previous genomic results were produced by searching only a six-frame translation of hg19 (GRCh37). For this analysis, we incorporated a personalized protein database and personalized genome by using the sequenced diploid genome NA12878.[22] Searches were performed according to the process displayed in Figure 3 to incorporate both maternal and paternal sequences so as to reveal which peptides are unique to, or that are shared by the strands. (ii) To lend high confidence to the novel results, we employed separate FDR analysis to the nonexonic matches.

The GM12878 spectral data set (available from http://giddingslab.org/data/encode/) contains ~1 million spectra; it was created using trypsin as the digestive enzyme and reversed-phase

liquid chromatography MS/MS analysis performed on a nanoLC-Ultra system (Eksigent, Dublin, CA) coupled with an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, San Jose, CA). FDR analysis of the spectra using a liberal precursor tolerance (500 ppm) revealed that 99% of the results at the 1% FDR fell within 40.41 ppm of the predicted peptide mass. This precursor tolerance value was used for all successive searches.

We used the GENCODE v. 12 gene annotations[23] as an instruction set for transcription/translation of the NA12878 diploid genome to create personalized maternal and paternal protein databases ("G12-M" and "G12-P", respectively). The resulting custom databases were very similar to the GENCODE reference database (G12) except that nsSNPs within the maternal and paternal genomic sequences had been translated to amino acid variations in the protein sequences. Creating these databases required producing a version of NA12878 aligned to the hg19 assembly. Using the hg18-based chain files included with NA12878, the NA12878 sequences were reconstructed to conform to an hg18 assembly. Nucleotide differences between the hg18-based NA12878 and hg18 were saved in a SNP file. The genomic indices of the SNPs were lifted over to hg19 indices using the UCSC liftOver tool (http://genome.ucsc.edu). The hg19-based SNPs were then overlaid onto hg19 to create a version of NA12878 that conformed to the hg19 assembly.

Six-frame genome searches present a major challenge that we addressed in our analysis pipeline. The challenge is that the vast majority of peptides populating six-frame databases do not exist in nature, which increases the chances for false positive spectrum matches. These false positives may "ride the coattails" of known peptides when estimating FDR, leading to spurious inflation in the number of "novel" identifications. Put another way, if a 1% FDR threshold is set for proteogenomic search results and 1% of the results are to novel peptides, how can we be sure the novel results are not the majority of the false positives? To address this challenge, we adopted the technique of mass spectrum sequential subtraction (MSSS)[24] as our method for searching multiple proteomic and genomic databases. The process involves searching proteomic and genomic databases in sequence with protein database searches preceding the genomic. For each database searched, the 1% FDR was estimated and all spectra identified at that cutoff were removed from the body of spectra used in the next database search. In cases where estimating a 1% FDR was not possible (e.g., there were fewer than 99 true positives), the ~0% FDR was used. Removing positively identified spectra with each database search not only makes for faster searches, but also neatly addresses the false positive conundrum for the novel, genome-only matches.

RESULTS

From the full body of spectra, 11.4% (117 991) were found to match G12. Subject-specific sequences included 34 unique peptides (from 82 spectra) matched with G12-M, 40 unique peptides (from 94 spectra) with G12-P. Of the peptides containing subject-specific acid variations, 24 were shared between the two sets. Two separate hg19 searches were required due to the fact that different sets of spectra were identified—and therefore removed—following the maternal and paternal protein searches. Three unique peptides were identified in the hg19 genome searches that were not found in the protein databases. No results were

found that would indicate novel forms of expression tied specifically to the maternal or paternal strands. The novel results are as follows:

- **APAGSAAGEGLLPHR** (chr7:48691991-48692036, hg19, reverse strand): The peptide's coding location maps to what was thought to be an intergenic region of chr7. Sequence homologies for the peptide were investigated using the protein BLAST tool;[25] the sequence shares some similarity (11/15 acids) with a subsequence of the bacterial protein leucyl-tRNA synthetase (YP_001505193.1). The highest confidence peptide/spectrum alignment is displayed in Figure 4 and shows significant mass ladder coverage to support this peptide sequence.
- **AVAAAAAAAAAAAAAGGR** (chr18:31158414-31158468, hg19, reverse strand): Located ~70bp upstream of the annotated starting point for the ASXL3 transcript. The peptide implies an extended variation of that gene.
- **NDDIPEQDSLGLSNLQK** (chr20:10401214-10401265, hg19, reverse strand): is located in the annotated 3′ UTR of MKKS. This sequence appears in an unreviewed isoform of the gene (Uniprot Q9HB66), lending credence to that isoform.

The novel matches found here demonstrate how proteogenomic searches can suggest instances of incorrect or incomplete gene annotations. However, given the difficulty of obtaining complete peptide coverage of a protein, these results are best taken as starting points for further investigation. For example, the hypothetical novel protein on chr7 should be confirmed via an orthogonal method such as transcript analysis in order to validate and derive the full coding sequence.

DISCUSSION

Proteogenomic searching is quickly gaining traction as a common component of many proteomic studies.[10b,a,26] Until recently, computing resources for such tasks were significant and proved to be a formidable barrier to entry. Improvements in computer technology have made this kind of search a viable option, however the lack of a tool that unifies all necessary tasks of such a search rendered the work unnecessarily complicated. We have presented Peppy, an integrated proteogenomic search tool that automatically performs the most common proteogenomic search tasks. Peppy's charter is to be fast, accurate and to require as little researcher intervention as possible.

A primary benefit of Peppy is its ability to easily search entire genomes, including the human genome. Peppy is designed to take genome files in FASTA format, negating the need for a predigest and six-frame translation using a separate tool. As genomic locus for each peptide is tracked, a separate mapping tool is not needed. The automation of steps is likewise a significant benefit as it increases usability while reducing the possibility for human error. This is especially true of the automated segmentation of the genome into portions small enough to be handled by the computer's memory. Additionally, the important step in determining the confidence of Peppy's identifications—the FDR analysis—is built into the system itself. The time and effort saved by an integrated proteogenomic system

promotes the possibility of including proteogenomic searches as a standard component of proteomic investigations.

Some Limitations of Six-Frame Searches

Though the possibility for discovering novel peptides is a major benefit of proteogenomic searches, the method is not without its limitations. The high number of false positive peptides present in a six-frame database creates the possibility for deceptively good spectrum matches with incorrect peptides. Proper FDR analysis can control for the saturation of false positives in the final result set, but the greater the number of high-quality false positives, the higher the confidence threshold must be to ensure the desired FDR. This higher confidence threshold then creates a higher degree of false negatives. Second, ~25% of tryptic peptides span exon boundaries[27] and these peptides will elude detection using peptides generated from a straight six-frame translation. Therefore, we recommend using proteogenomic searches in conjunction with standard proteomic searches in a manner akin to the MSSS technique described above. False negatives, especially exon-spanning peptides, will be reduced by including protein database searches; additionally the novel, non-exonic matches found from the genome search will have higher confidence as their FDR estimates will not include known proteins.

Future Work

We are presenting version 1.0 of the Peppy software: refinements are currently in development that will improve the accuracy and usefulness of search results. The inclusion of modifications such as post-translational modifications (PTMs), labels and selenocysteine could greatly improve the number of true positives returned. However, their inclusion will also increase the search space and therefore the chance of false positives, so we are exploring some novel techniques to address this issue, while striving to maintain Peppy's speed capabilities.

Since the genome and spectral data sets are typically large, a cross-analysis of the two will produce copious data, necessitating a way to navigate that data easily. Thus, efforts are underway to facilitate integration with the UCSC Genome Browser as well as user-friendly reports that combine the results from several proteogenomic experiments and provide graphical representations of PSMs.

Given the heterogeneity of mass spectrometers, sample preparation techniques and analyses, there is a need for the automatic discovery of optimal settings for mass tolerances, PTMs, missed cleavages, etc. In the above analysis, we took measures to find an optimal precursor tolerance using FDR analysis and efforts are in progress to automate this process. Further, we aim to implement the use of the MSSS technique directly into the software as the process of sequentially identifying spectra then removing them from subsequent searches is a rote task that is ripe for automation. MSSS confers the benefits of both protein and six-frame database searches while creating result sets that can be unified to a homogeneous FDR–characteristics we feel many researchers will appreciate.

Availability

The Peppy application, source code and documentation are available at http://geneffects.com/peppy.

Acknowledgments

This research was made possible by grant R01 HG003700 to M.C.G.; 1RC2 HG005591-01 to M.C.G.; and U24 CA160035 to M.C.G.

References

1. (a) Brinkman BM. Splice variants as cancer biomarkers. Clin Biochem. 2004; 37(7):584–94. [PubMed: 15234240] (b) Xu Q, Lee C. Discovery of novel splice forms and functional analysis of cancer-specific altern5ative splicing in human expressed sequences. Nucleic Acids Res. 2003; 31(19):5635–5643. [PubMed: 14500827] (c) Wulfkuhle JD, Liotta LA, Petricoin EF. Proteomic applications for the early detection of cancer. Nat Rev Cancer. 2003; 3(4):267–275. [PubMed: 12671665]
2. Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP. Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci USA. 2008; 105(52):21034–21038. [PubMed: 19098097]
3. Payne SH, Huang ST, Pieper R. A proteogenomic update to Yersinia: enhancing genome annotation. BMC Genomics. 2010; 11:460. [PubMed: 20687929]
4. Hanash SM, Beretta LM. Operomics: integrated genomic and proteomic profiling of cells and tissues. Briefings Funct Genomics Proteomics. 2002; 1(1):10–22.
5. (a) Eng J, McCormack A, Yates J. 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. [PubMed: 24226387] (b) Bafna V, Edwards N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics. 2001; 17(Suppl 1):S13–S21. [PubMed: 11472988] (c) Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom. 2003; 17(20):2310–2316. [PubMed: 14558131] (d) Wan Y, Yang A, Chen T. PepHMM: a hidden Markov model based scoring function for mass spectrometry database search. Anal Chem. 2006; 78(2):432–437. [PubMed: 16408924] (e) Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res. 2007; 6(2):654–661. [PubMed: 17269722] (f) Narasimhan C, Tabb DL, Verberkmoes NC, Thompson MR, Hettich RL, Uberbacher EC. MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. Anal Chem. 2005; 77(23): 7581–7593. [PubMed: 16316165]
6. Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem. 2003; 75(4):768–774. [PubMed: 12622365]
7. Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A. The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. Mol Cell Proteomics. 2004; 3(6):531–533. [PubMed: 15075378]
8. Tabb DL. What's driving false discovery rates? J Proteome Res. 2008; 7(1):45–46. [PubMed: 18081243]
9. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007; 4(3):207–214. [PubMed: 17327847]
10. (a) Castellana NE, Pham V, Arnott D, Lill JR, Bafna V. Template proteogenomics: sequencing whole proteins using an imperfect database. Mol Cell Proteomics. 2010; 9(6):1260–1270. [PubMed: 20164058] (b) Renuse S, Chaerkady R, Pandey A. Proteogenomics. Proteomics. 2011; 11(4):620–630. [PubMed: 21246734]

11. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57–74. [PubMed: 22955616]
12. Smith RF, Wiese BA, Wojzynski MK, Davison DB, Worley KC. BCM Search Launcher—an integrated interface to molecular biology data base search and analysis services available on the World Wide Web. Genome Res. 1996; 6(5):454–462. [PubMed: 8743995]
13. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem. 2005; 77:4626–4639. [PubMed: 16013882]
14. (a) Gertz EM, Yu YK, Agarwala R, Schäffer AA, Altschul SF. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 2006; 4:41. [PubMed: 17156431] (b) Sanders WS, Wang N, Bridges SM, Malone BM, Dandass YS, McCarthy FM, Nanduri B, Lawrence ML, Burgess SC. The proteogenomic mapping tool. BMC Bioinf. 2011; 12:115.(c) Specht M, Stanke M, Terashima M, Naumann-Busch B, Janflen I, Hohner R, Hom E, Liang C, Hippler M. Concerted action of the new Genomic Peptide Finder and AUGUSTUS allows for automated proteogenomic annotation of the Chlamydomonas reinhardtii genome. Proteomics. 2011; 11(9):1814–1823. [PubMed: 21432999]
15. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999; 20(18):3567.
16. Mujezinovic N, Schneider G, Wildpaner M, Mechtler K, Eisenhaber F. Reducing the haystack to find the needle: improved protein identification after fast elimination of non-interpretable peptide MS/MS spectra and noise reduction. BMC Genomics. 2010; 11(Suppl 1):S13. [PubMed: 20158870]
17. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol. 2006; 24(10): 1285–1292. [PubMed: 16964243]
18. Risk B, Edwards NJ, Giddings MC. A peptide-spectrum scoring system based on ion alignment, intensity and pair probabilities. J Proteome Res. 2013 submitted for publication.
19. Wenger CD, Coon JJ. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. J Proteome Res. 2013; 12(3):1377–1386. [PubMed: 23323968]
20. Pagano, M.; Gauvreau, K. Principles of Biostatistics. Duxbury/Thomson Learning; Belmont, CA: 2000.
21. (a) Khatun J, Yu Y, Wrobel JA, Risk BA, Gunawardena HP, Secrest A, Spitzer WJ, Xie L, Wang L, Chen X, Giddings MC. Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. BMC Genomics. 2013; 14(1):141. [PubMed: 23448259] (b) Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, Lee BT, Barber GP, Harte RA, Diekhans M, Long JC, Wilder SP, Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. 2013; 41(Database issue):D56–D63. [PubMed: 23193274]
22. Rozowsky J. Coordination between Allele Specific Expression and Binding in a Network Framework. 2013 submitted for publication.
23. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders AG, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22(9):1760–1774. [PubMed: 22955987]
24. Helmy M, Sugiyama N, Tomita M, Ishihama Y. Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics. Genes Cells. 2012; 17(8):633–644. [PubMed: 22686349]
25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25(17):3389–3402. [PubMed: 9254694]

26. Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Briefings Funct Genomics Proteomics. 2008; 7(1):50–62.

27. Tanner S, Shen Z, Ng J, Florea L, Guigó R, Briggs SP, Bafna V. Improving gene annotation using peptide mass spectrometry. Genome Res. 2007; 17(2):231–239. [PubMed: 17189379]

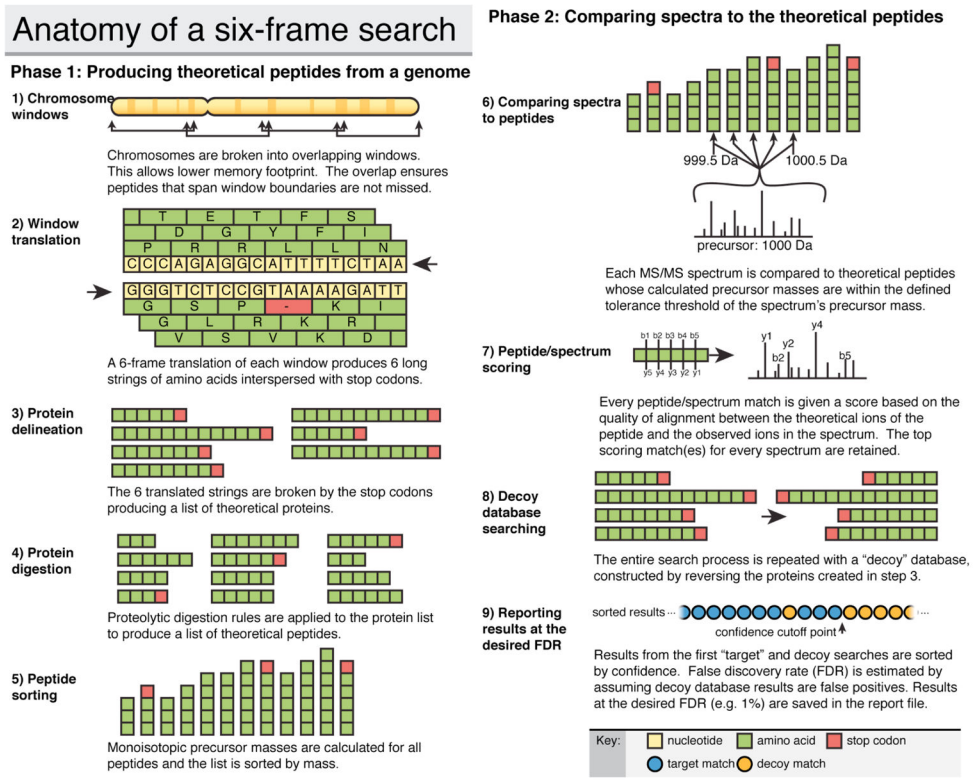


Figure 1.

The essential steps required in a six-frame search. Variations on the approach entail how to segment the genome; what delineates a theorized protein; how to create a decoy database and the method for estimating FDR.

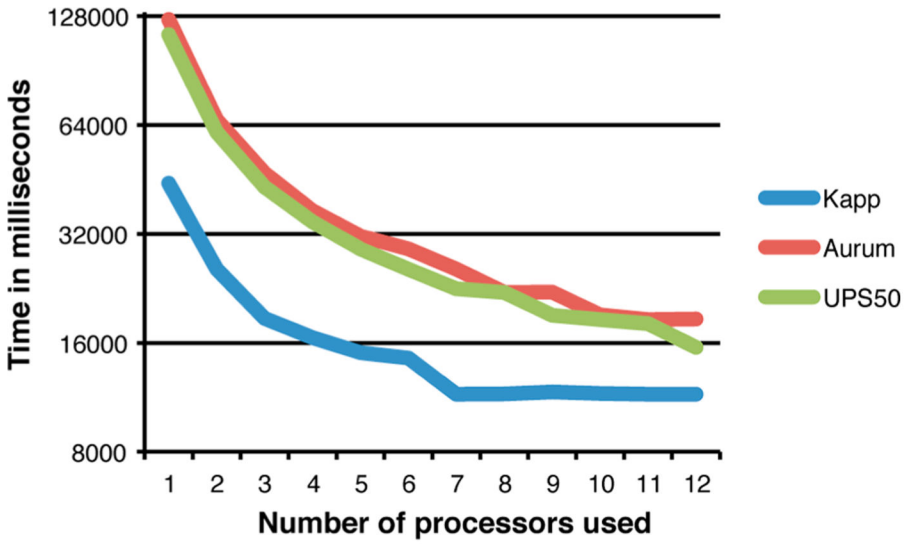


Figure 2.

Time needed to complete a proteomic search vs the number of processors available to the software. Tests were done on a 12-core machine, varying the number of cores used from 1 to 12. “Kapp”, “Aurum”, and “UPS50” are data sets with 517, 1349, and 1473 spectra, respectively; the database searched was UniProt human.

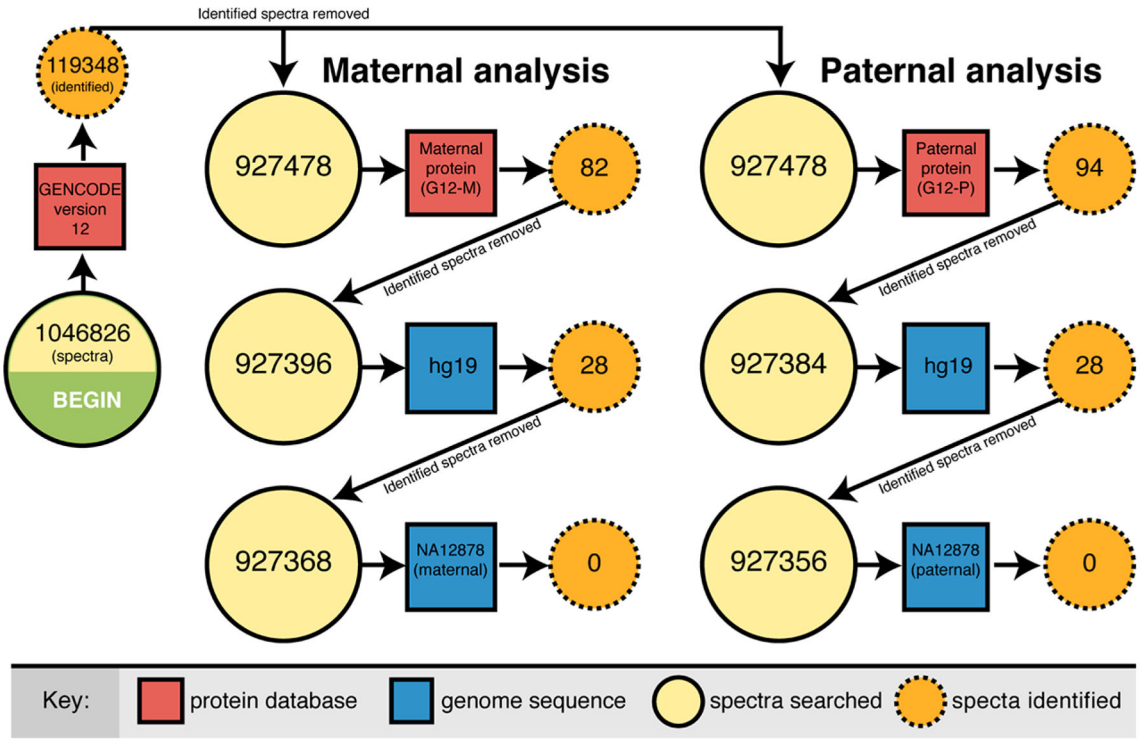


Figure 3.

Proteogenomic analysis workflow of the GM12878 cell line spectra using the NA12878 diploid genome.

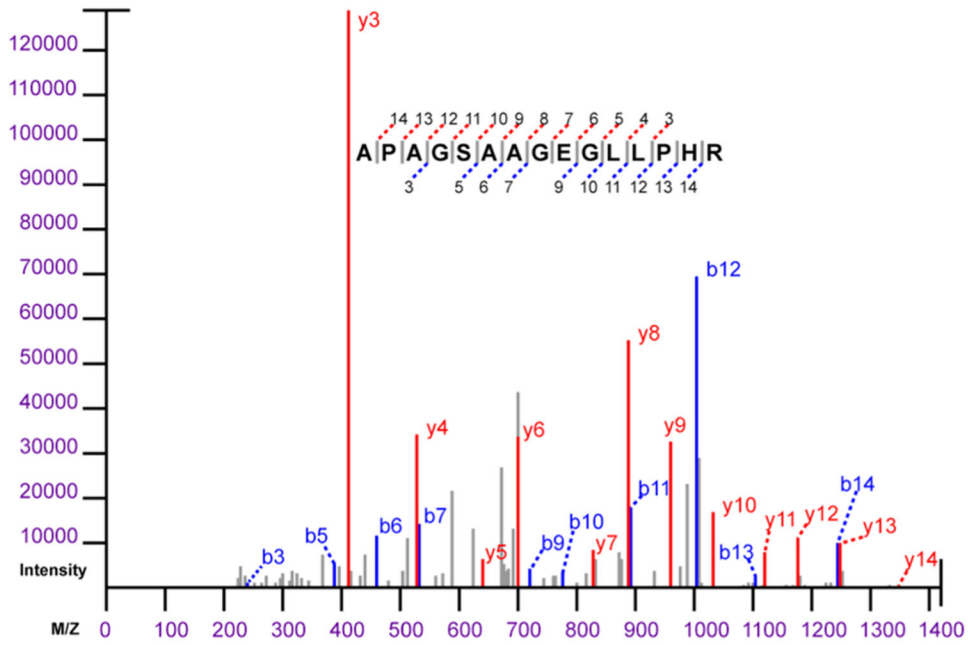


Figure 4.

Peptide/spectrum alignment for the sequence APAGSAAGEGLLPHR. Every amino acid is delineated by at least one fragment ion.