HHS Public Access

Author manuscript

J Proteome Res. Author manuscript; available in PMC 2016 September 22.

Published in final edited form as:

J Proteome Res. 2016 July 1; 15(7): 2309–2320. doi:10.1021/acs.jproteome.6b00344.

JUMPg: an Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells

Yuxin Li[*], **Xusheng Wang**[%], **Ji-Hoon Cho**[%], **Tim Shaw**[#], **Zhiping Wu**[*], **Bing Bai**[*], **Hong Wang**[*], **Suiping Zhou**[%], **Thomas G. Beach**[¶], **Gang Wu**[#], **Jinghui Zhang**[#], and **Junmin Peng**[*,%,&]

[*]Departments of Structural Biology and Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, United States

[%]St. Jude Proteomics Facility, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, United States

[#]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, United States

[¶]Banner Sun Health Research Institute, Sun City, Arizona 85351, United States

Abstract

Proteogenomics is an emerging approach to improve gene annotation and interpretation of proteomics data. Here we present JUMPg, an integrative proteogenomics pipeline including customized database construction, tag-based database search, peptide-spectrum match filtering, and data visualization. JUMPg creates multiple databases of DNA polymorphisms, mutations, splice junctions, partially trypticity, as well as protein fragments translated from the whole transcriptome in all six frames upon RNA-seq *de novo* assembly. We use a multistage strategy to search these databases sequentially, in which the performance is optimized by researching only unmatched high quality spectra, and re-using amino acid tags generated by the JUMP search engine. The identified peptides/proteins are displayed with gene loci using the UCSC genome browser. Then the JUMPg program is applied to process a label-free mass spectrometry dataset of Alzheimer's disease postmortem brain, uncovering 496 new peptides of amino acid substitutions, alternative splicing, frame shift, and "non-coding gene" translation. The novel protein *PNMA6BL* specifically expressed in the brain is highlighted. We also tested JUMPg to analyze a stable-isotope labeled dataset of multiple myeloma cells, revealing 991 sample-specific peptides that include protein sequences in the immunoglobulin light chain variable region. Thus, the JUMPg program is an effective proteogenomics tool for multi-omics data integration.

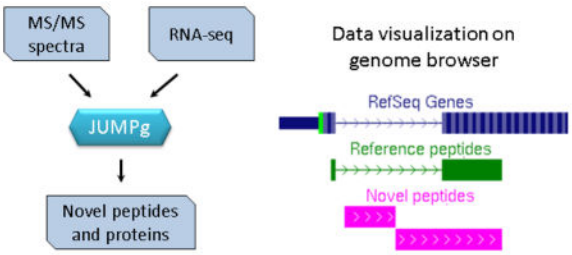
Graphical abstract

[&]Corresponding Author: Junmin Peng, tel: 901-336-1083; junmin.peng@stjude.org.

The authors declare no competing financial interest.

Supporting Information Notes

Supporting Data: MS/MS spectra assigned to RNA 6FT peptides of one-hit wonders.



Keywords

Genomics; proteomics; mass spectrometry; proteogenomics; RNA-seq; database search; multistage analysis; spectrum quality control

INTRODUCTION

The flow of genetic information from DNA to RNA to proteins is the "central dogma" of molecular biology. Global analysis and annotation of these biomolecules in cells is one of the ultimate goals in analytical biochemistry. Although whole genome and transcriptome can now be analyzed by advanced next-generation sequencing technology,[1, 2] it is still not clear whether some DNA sequences truly encode proteins, such as short open reading frames.[3] As to the individual human being, the proteome is shaped by personalized genome containing genetic DNA polymorphism and sporadic mutations. In cancer patients, up to thousands of DNA mutations can be detected in tumor cells.[4] A recurring question is whether these mutations are transcribed and translated into proteins. Furthermore, instead of one gene-one protein hypothesis, the proteome is greatly enlarged by alternative RNA splicing[5] and alternative translation.[6] For example, in repeat-associated non-ATG translation, homopolymeric proteins in three reading frames can be unexpectedly expressed from numerous expanded repeat mutations (e.g. CAG, CGG and GGGGCC) that are linked to neurodegenerative diseases.[7] Thus, many protein forms in cells may be missing in commonly used protein reference databases.

Proteogenomics is an emerging approach to enhance the annotation of genome and proteome, often relying on next-generation sequencing and deep mass spectrometry (MS) data.[8–10] Indeed, the approach is initially designed by proteomics pioneers for interpreting reference genomes.[11–14] With the development of genome-wide sequencing techniques, genomics variant sequence information (e.g., DNA polymorphisms and alternative splicing) has been rapidly accumulated and incorporated into customized database to identify protein variants.[15, 16] But genome-wide proteomics studies had been difficult until recent improvement in the depth of MS-based proteomics.[17–19]

Proteogenomics pipeline usually includes multiple steps, such as customized database construction, database search, peptide-spectrum match (PSM) filtering, novel peptide identification and visualization.[10, 20–22] The whole process can be laborious and error-prone. A number of computational methods have been proposed for generating customized database in recent years,[23–33] which can be classified into three categories: (i) DNA

polymorphism/mutation (hereafter using "mutation" for simplicity) database, (ii) splice junction database, and (iii) genome/transcriptome six-frame translation (6FT) database for detecting non-canonical translation events in unknown coding DNA sequence regions. However, most of these studies do not address peptides in all categories simultaneously. Hence, we attempt to include all possible novel peptides derived from genomic and transcriptomic data in the customized database in an effective and efficient way.

Another challenge in proteogenomics study is how to overcome the effect of database size (i.e. search space).[34] Whereas enlarging customized database potentially increases the chance of true PSMs, it also raises an associated probability that MS/MS spectra are wrongly matched to peptides, especially when the majority of sequences in the search space are false (e.g. in whole genome 6FT database).[35] One effective solution is to use multistage database search strategy,[36–38] in which MS/MS spectra are first searched against a well-annotated database (e.g. UniProt), and then the unmatched spectra are explored by searching large but less confident search space. In this way, most true PSMs are obtained at the first stage of analysis, and further increased by subsequent stage(s). To be more efficient, this multistage process can be further accelerated by removing low quality spectra before the second stage of analysis.[38] Low quality spectra could be readily discriminated by amino acid tags extracted from the spectra.[39] However, these advanced techniques have not been applied in the majority of proteogenomics studies. In addition, most existing proteogenomics tools are not well integrated for high usability.

Here we present JUMPg, a computational pipeline that automates complex proteogenomics analysis. The program provides customized database methods from all three categories; supports multistage analysis to maximize peptide identification, and exhibits detected peptides/proteins in the UCSC genome browser. For high sensitivity and accuracy, the JUMP algorithm[40] is implemented to produce high quality tag information for assessing MS/MS spectrum quality and performing database search. Application of JUMPg in two deep proteomics datasets revealed a variety of sample-specific peptides/proteins in Alzheimer's disease (AD) brain and multiple myeloma (MM) cells. The JUMPg software, source code and documentation are freely available (https://github.com/gatechatl/JUMPg).

EXPERIMENTAL PROCEDURES

The JUMPg algorithm takes as input MS raw data, commonly used protein database (e.g. UniProt), and corresponding genomics data (e.g. mutations detected by next-generation sequencing, RNA-seq raw data and derived splice junctions and assembled transcripts), and outputs identified peptides/proteins with annotations. There are five modules in the algorithm (Figure 1 **and** Figure S1 in the Supporting Information).

Module 1: Customized Database Construction

JUMPg provides three options of constructing customized databases (Figure 2), including mutations, splice junctions, and non-canonical translation from all six frames.

Mutation peptides are directly input by genomic variant files (e.g. VCF files) or indirectly derived from RNA-seq raw data (FASTAQ file). RNA reads are aligned to human reference

genome (hg19) by STAR[41] with UCSC gene annotations.[42] Mutations are detected by GATK following the "GATK Best Practices" protocol[43, 44], stored in VCF format, and classified by AnnoVar (version 23, August 2013)[45] into different functional categories (e.g. non-silent, intronic, intergenic, etc.). Amino acid (AA) sequences that flank (±30 AAs) non-silent mutations are extracted according to UCSC gene annotations into FASTA format.

Splice junction database is generated with a similar strategy described by Sheynkman et al. 2013.[26] Splice junction peptides are input by junction files from the STAR program,[41] and filtered with following criteria: i) at least 2 uniquely mapped reads, and ii) containing canonical splice sites (GT/AG, GC/AG, or AT/AC). The two genomic sequence fragments that flank a pair of splice donor and acceptor sites are retrieved from reference genome, and translated in three open reading frames (ORFs). If the ORF contains a stop codon, the sequence is trimmed. If the stop codon locates upstream of the splice site, the sequence is discarded because no novel amino acid sequence is produced.

Non-canonical translated peptides are made from RNA-seq reads through all 6 frames. RNAseq reads are first assembled into transcripts using trinity (version 2.1.1) [46] with default parameters (Supporting Information Notes). Each transcript is then translated in 6 frames, split by stop codons and only ORFs with at least 66 AAs (i.e., ~200 bp) are retained.

Module 2: Database Search

Our tag-based database search engine JUMP [40] was modified to adapt multistage strategy and maintain regular search function (version 13.1.0). In the regular search, JUMP performs preprocessing, tag generation, MS/MS pattern matching and scoring as previously reported,[40] and outputs two temporary files: a DTAS file containing mass-corrected and intensity-normalized MS/MS peak information, and a TAGS file containing *de novo* tags for each MS/MS scan. We evaluated MS/MS quality by the two files and kept only unmatched high quality spectra for multistage analysis (**see Module 5**). In addition, we also used these files to bypass two time-consuming steps (preprocessing and tag generation) during the multistage analysis, which is termed tag recycling. Common parameters include ±6 ppm for precursor ion mass tolerance, dynamic mass shift for oxidized Met (+15.9949 Da), and the consideration of *a*, *b*, and *y* ions. Fixed modifications on the N termini and Lys (+229.1629 Da) are also included for the TMT dataset.

Module 3: Peptide-spectrum Match Filtering

The peptide-spectrum matches (PSMs) are filtered as previously reported.[47] The target-decoy strategy is used to evaluate false discovery rate.[48, 49] PSMs are first filtered by user-specified parameters (e.g. minimal peptide length and minimum search score), then by precursor ion mass accuracy. The resulting PSMs are further grouped by precursor ion charge state, tryptic ends, and then filtered by matching scores (Jscore and ΔJscore) to achieve the user-specified level of false discovery rate.

Module 4: Peptide Annotation and Visualization

Accepted peptides from mutation or junction search space are examined respectively for the specific amino acid(s) mutated or that span the splicing junction. The survived peptides are

condensed into events (e.g., by collapsing different modification forms), with genomic positions stored in BED format for visualization in the UCSC genome browser.[50]

For peptides identified from RNA 6FT database, peptides from known search space are first removed and the corresponding RNA transcripts are aligned to reference genome (hg19) by BLAT (v36, default parameters).[51] Top-scored alignment result for each transcript is retained from which the genomic position of each peptide is derived. Peptides are further classified according to UCSC gene annotations. In case one peptide can be assigned to multiple isoforms, the following priority will be applied: CDS > UTR > intron > non-coding genes > intergenic regions. Peptides that cannot be mapped to human genome are currently not considered.

Module 5: Spectrum Quality Scoring

The quality of tandem MS spectra is scored by the method of linear discriminant analysis (LDA)[38] with modification, using the 1st stage database search result as a training set. The method integrates features reflecting tag information (i.e. the top tag length) and spectrum data (i.e. square root of MS2 peak number and logarithm transformed MS2 maximum peak intensity) by the equation:

$$S_{quality}=\sum_{i=1}^{n} c_i f_i$$

Where f_i and c_i denote value and coefficient of each feature, respectively. The coefficients are obtained by training the LDA model. The threshold of spectrum quality score is adjusted to allow discarding less than 1% of accepted PSMs.

To evaluate whether the LDA method introduced overfitting during the analysis, we performed 10-fold cross validation analysis in the following steps: i) the whole dataset was randomly partitioned into 10 equal sized subsamples; ii) of these 10 subsamples, 9 subsamples were used for training the LDA model and deriving a training receiver operating characteristic (ROC) curve, while the remaining subsample was used as the testing data for a testing ROC curve; iii) we repeated the 2nd step for 10 times to evaluate the variation of the analysis. If overfitting occurred, the area under ROC curve (AUC) of the training dataset would be larger than that of the independent testing dataset. Our results showed that the training and testing AUCs were nearly identical (Table S4 in Supporting Information), indicating no overfitting of LDA modeling in our analysis.

Data Analysis of Two Large Dataset (AD and MM cells)

The first dataset was collected from human Alzheimer's disease postmortem brain, including label-free high resolution MS raw data (1.7 million MS/MS scans) and RNA-seq data (50 million reads; Table S1 in Supporting Information).[52] The second dataset was obtained from a multiple myeloma cell line (ANBL6),[53] including tandem mass tag (TMT, Thermo Fisher Scientific) labeled raw data (6.2 million MS/MS scans) and RNA-seq data (242 million reads).

Proteogenomics analysis was performed by JUMPg in three stages. i) MS data were searched against a database combining mutations and splice junction peptides as well as UniProt human proteins (downloaded in February 2015) with fully tryptic restriction. The resulting PSMs in the 1st stage were filtered to 1% unique protein FDR. ii) Partially tryptic search was carried out with the unmatched high quality spectra against proteins accepted during the 1st stage. iii) The remaining spectra were matched to RNA-seq 6FT database, and filtered to 0% FDR. However, we recognized that the FDR estimation may not be accurate and further calculated the standard deviation of the FDR by previously reported method.[49] The actual FDR falls between 0% and 0.2% within a 99% confidence interval. Moreover, we compared peptides identified from novel search space with those from reference database, and found that all peptides showed highly similar distributions of matching score, precursor mass error and tag length (Figure S2 in Supporting Information), indicating that our stringent filtering yielded high quality PSMs in novel search space.

RESULTS AND DISCUSSION

Improved Peptide Identification by Multistage Customized Databases and Partially Tryptic Search

The JUMPg program implements three independent methods for constructing customized peptide databases including mutations, splice junctions and RNA-seq 6FT translated peptides (Figure 1). While mutations and splice junctions focus on variant peptides in annotated CDS regions, there is emerging evidence for functional proteins encoded by unknown open reading frames.[3] Therefore JUMPg constructs another database by RNA-seq *de novo* assembly and six frame translation[25] (Figure 2). This method shows the best balance between database comprehensiveness (the whole transcriptome) and specificity (compact search space), compared to whole genome translation (Supporting Information Note, Table S2 and Figure S3). As to search space, the mutation, splice junction and 6FT databases are equivalent to 1.1%, 3.0% and 71.4% of fully tryptic UniProt protein database (default reference database), respectively (Table 1). In addition, whole proteome data often contain a significant amount of partially tryptic peptides possibly due to endogenous protease activities.[48] The partially tryptic database is ~10 times larger than the fully tryptic counterpart.

To increase the search efficiency against these databases (Figure S4 in Supporting Information), a multistage strategy is developed to perform the analysis sequentially (Figure 1A). First, the MS/MS spectra are matched to a fully tryptic peptide database pooled from UniProt, mutations and splice junctions. The matching results are filtered to ~1% false discovery rate at protein level. Then only the unmatched *high quality spectra* are subjected to the 2nd stage of analysis against the large partially tryptic database. Moreover, this partially tryptic database is shrunk by considering proteins only identified in the 1st stage (Table S3 in Supporting Information). Finally, the remaining high quality spectra are searched against the 6FT database, as the majority of the peptides inside are false owing to six frame translation.

JUMPg exports the identified peptides in a text table, showing the peptide sequences, PSM counts, tags and scores of the best PSM, and database entries (Figure 1B). JUMPg also provides a visualization function by converting peptides to genomic locations via UCSC

known gene annotations, which can be co-displayed with other genomic information on the UCSC genome browser (Figure 1C).[50] For example, two novel peptides (in magenta) were identified to indicate an event of intron retention, whereas another peptide (in green) was also identified to support the canonical protein isoform, annotated with known transcript (in navy) of the *CASKIN1* gene in chromosome 16. Taken together, JUMPg maximizes novel peptide identification by three different customized database methods in an effective manner, and displays results for integrative visualization.

Acceleration of Multistage Database Search by Spectrum Quality Control and Amino Acid Tag Recycling

To reduce multistage database search time, we discard low quality spectra before the 2nd stage analysis, based on MS/MS spectrum quality. To calculate a quality score for each spectrum, we use three features, spectrum intensity, total peak number, and the best tag length. All three features are highly correlated with PSM identification successful rate defined by accepted PSMs in the 1st stage analysis (Figure 3A), consistent with previous reports.[39, 54] Of these, the tag length (in black) shows the best performance by receiver operating characteristic (ROC) curves (Figure 3B). To maximize the discriminant capacity, we use the LDA method[38] to combine these features into one single quality score (in red), which outperforms any individual feature (Figure 3B). Next, a threshold of the quality score is selected to balance the removal of MS/MS spectra and the recovery of high quality spectra, depending on the score distribution and the accumulative curve of accepted PSMs in the 1st stage analysis (Figure 3C). When preserving 99% of accepted PSMs, the threshold is able to eliminate ~35% of the MS/MS spectra. As database search time is linearly correlated with the spectrum number,[55] this filtering step can save ~35% of computational time (Figure S4C in Supporting Information).

To further improve database search speed, JUMPg implements a novel function by recycling amino acid tags of each MS/MS spectrum. The tag information enables high sensitivity and accuracy of peptide identification,[40, 56, 57] but the process to derive tags is time-consuming. The tags generated at the 1st stage analysis are recorded and re-used in the 2nd and later stage analysis, which further accelerate database search by ~25%. Taken together, the spectrum quality control and tag recycling reduce search time by more than 50% for multistage analysis.

We also considered one potential caveat of multistage analysis, in which PSMs accepted in the 1st stage could have higher scores in the 2nd or 3rd stage. To address this, more than 156,000 MS/MS spectra were searched against fully tryptic database, resulting in 52,114 accepted PSMs (1% FDR). These spectra were re-searched against partially tryptic database, leading to higher scores of only 97 (0.2% out of 52,114) PSMs. This small percentage is lower than the specified FDR level (1%). Thus, the multistage strategy saves database search time by omitting PSMs accepted in the 1st stage, which does not introduce extra false positives of PSMs.

Application to Label Free Data: Discovery of a Novel Protein Coding Gene in Human Brain

We previously analyzed Alzheimer's disease brain proteome by long gradient LC/LC-MS/MS and the transcriptome by RNA sequencing, identifying ~10,000 reference proteins.[52] However, the sample-specific events such as DNA polymorphisms and brain specific alternative splicing events have not been addressed. Here we used the JUMPg pipeline to analyze the same dataset (dataset 1 with ~1.7 million label free MS/MS spectra), discovering a total of 496 novel peptides in three stage analyses (Table 1). These 496 novel peptides are classified into 14 types of events (Table 2), including amino acid substitution, deletion, frame shift, exon skipping and other alternative splicing events, as well as translation from antisense strand, non-coding genes, untranslated regions (UTR) and intergenic regions. Four examples are illustrated in detail (Figure 4): (i) a single amino acid substitution in *ACO2* supported by three peptides, (ii) an in-frame deletion of *PRUNE2*, (iii) an exon skipping in *SLC12A5*, and (iv) 5′ UTR translation / alternative start codon demonstrated by the original and Met oxidized forms of a fully tryptic peptide.

As a high percentage of mutated peptides identified in proteogenomics studies are also found in public single nucleotide polymorphism database (dbSNP),[24, 31, 37] we analyzed in this sample the 255 novel single-amino acid substitution events, and found that 239 (94%) are collected in the dbSNP (version 138). This consistent result provides additional support to the confidence of JUMPg-identified novel peptides.

Interestingly, we uncovered a novel protein coding gene specifically expressed in the human brain. During the 3rd stage analysis with the 6FT database, we found nine novel peptides that locate within a 1.5 kb region of chromosome X (Figure 5A, 5B), a genomic region with no protein coding gene models recorded in databases (including refSeq, UCSC, GENCODE v19, Ensembl v75 and AceView). Closer examination of these nine peptides (in magenta) confirmed that they share the same open reading frame, suggesting that these peptides are translated from one novel protein coding gene. To obtain the full protein sequence of this new gene, we aligned the assembled RNA transcripts to the reference genome, translated this genomic region according to the reading frame defined by JUMPg-identified peptides, and derived 578 amino acid protein sequence (in green, Figure S4 in Supporting Information).

To functionally annotate this novel protein, the sequence was aligned to NCBI non-redundant protein database.[58] The top hit was a computationally predicted protein termed "putative paraneoplastic antigen-like protein 6B-like protein (PNMA6BL)" in a primate (*Cercocebus atys*) with 86% identity. Moreover, one "PNMA" domain was identified by Pfam[59] within the novel protein sequence, suggesting that this protein belongs to the paraneoplastic antigens Ma (PNMA) gene family. In human, this family includes 7 known members (PNMA1, 2, 3, 5, 6A and PNMAL1, 2). Phylogenetic analysis indicated that this novel protein was clustered within the PNMA6 sub-branch (Figure 5C), consistent with its homologue in the primate. Thus, we named this novel protein as "*PNMA6BL*". In addition, pan-tissue RNA-seq analysis indicates that the *PNMA6BL* gene is almost exclusively expressed in human brain (Figure 5D), which is typical for the PNMA gene family.[60] This example demonstrates that our JUMPg pipeline allows the identification of previously unannotated genes by combining deep proteomics and transcriptomics data.

Application to Isotopically labeled Data: Identification of Immunoglobulin Light Chain Variable Region Peptide in Multiple Myeloma Cells

Isobaric labeling techniques (e.g. iTRAQ/TMT) have been widely used in quantitative proteomics,[61, 62] but peptide identification is more challenging due to noise peaks induced by labeling reagents.[63] To further test the performance of JUMPg, we applied the same procedure (Figure 1A) to a large TMT dataset (dataset 2 with 6.2 million MS/MS spectra) from multiple myeloma (MM) cell lines. Intriguingly, we observed a high percentage of partially tryptic peptides in this dataset (Table 1), with 67,368 partially tryptic peptides (33% of total peptides) from 182,988 PSMs (20% of total PSMs). As MM cells are known to produce a high level of immunoglobulin, cell survival relies on excessive proteasome activity,[64] which might explain the presence of additional partially tryptic peptides. Similarly, high frequency of partial trypticity was also detected in human serum,[65] suggesting that *in vivo* proteases may be linked to this elevated partial cleavage.

In the three stage analyses, we accepted 198,103 peptides from 909,223 PSMs (Table 1). As expected, the PSM identification successful rate of the TMT dataset (18%) is lower than that of the label free AD dataset (36%). Nonetheless, we determined a total of 991 novel peptides, which were summarized into 531 mutation, 158 novel splicing and 48 non-canonical translation events (Table 2). Notably, the most abundant peptide matched by 211 PSMs displays three substitutions in the immunoglobulin light chain variable region (Figure 6). The result exemplifies the advantage of the 6FT method, in which *bona fide* RNA information is retained in the 6FT database. Because the immunoglobulin light chain variable region is unique to individual B cell clones, the capability of identifying these unique peptides highlights JUMPg as a potential tool towards personalized proteome.[66]

CONCLUSIONS

We have demonstrated that JUMPg is an automated proteogenomics pipeline to identify and visualize reference and novel peptides from proteomics and functional genomics data. Several key features distinguish JUMPg from previous proteogenomics software: i) JUMPg streamlines the analysis by integrating five consecutive modules (**see Experimental procedures**); ii) JUMPg maximizes peptide identifications by three customized database methods and partially tryptic search; iii) JUMPg supports multistage database search; and iv) JUMPg uses the tag-based hybrid search engine JUMP for high sensitivity and accuracy. In addition, multistage database search speed is optimized by removing low quality spectra and recycling tags. Application of JUMPg to two different datasets identified thousands of novel peptides, including a novel protein coding gene in human brain and immunoglobulin light chain variable region in multiple myeloma cells. Thus, JUMPg is a comprehensive and integrative tool for proteogenomics analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank for all other members in the Peng and Zhang labs and in St. Jude Proteomics Facility for insightful discussion. This work was partially supported by National Institutes of Health grant R01AG047928, R01GM114260, U24NS072026, P30AG19610, Arizona Department of Health Services (contract 211002), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05-901 and 1001), the Michael J. Fox Foundation, and ALSAC (American Lebanese Syrian Associated Charities). The MS analysis was performed in the St. Jude Children's Research Hospital Proteomics Facility, partially supported by NIH Cancer Center Support Grant (P30CA021765). The authors have declared no conflicts of interest.

ABBREVIATIONS

MS/MS	tandem mass spectrometry
FDR	false discovery rate
PSM	peptide-spectrum match
6FT	six-frame translation
ROC	receiver operating characteristic
TMT	Tandem Mass Tag

References

1. Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem. 2013; 6:287–303.
2. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014; 30(9):418–26. [PubMed: 25108476]
3. Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet. 2014; 15(3):193–204. [PubMed: 24514441]
4. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013; 339(6127):1546–58. [PubMed: 23539594]
5. Lee Y, Rio DC. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu Rev Biochem. 2015; 84:291–323. [PubMed: 25784052]
6. Jackson RJ, Hellen CU, Pestova TV. The mechanism of eukaryotic translation initiation and principles of its regulation. Nat Rev Mol Cell Biol. 2010; 11(2):113–27. [PubMed: 20094052]
7. Cleary JD, Ranum LPW. Repeat associated non-ATG (RAN) translation: new starts in microsatellite expansion disorders. Curr Opin Gen Dev. 2014; 26:6–15.
8. Low TY, Heck AJ. Reconciling proteomics with next generation sequencing. Curr Opin Chem Biol. 2015; 30:14–20. [PubMed: 26590485]
9. Renuse S, Chaerkady R, Pandey A. Proteogenomics. Proteomics. 2011; 11(4):620–30. [PubMed: 21246734]
10. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods. 2014; 11(11):1114–1125. [PubMed: 25357241]
11. Yates JR, Eng JK, Mccormack AL. Mining Genomes - Correlating Tandem Mass-Spectra of Modified and Unmodified Peptides to Sequences in Nucleotide Databases. Anal Chem. 1995; 67(18):3202–3210. [PubMed: 8686885]
12. Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, Mortensen P, Shevchenko A, Boucherie H, Mann M. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. Proc Natl Acad Sci U S A. 1996; 93(25):14440–5. [PubMed: 8962070]
13. Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics. 2004; 4(1):59–77. [PubMed: 14730672]

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

15. Schandorff S, Olsen JV, Bunkenborg J, Blagoev B, Zhang Y, Andersen JS, Mann M. A mass spectrometry-friendly database for cSNP identification. Nat Methods. 2007; 4(6):465–466. [PubMed: 17538625]

16. Edwards NJ. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol. 2007; 3:102. [PubMed: 17437027]

17. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, Faerber F, Kuster B. Mass-spectrometry-based draft of the human proteome. Nature. 2014; 509(7502):582–7. [PubMed: 24870543]

18. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TS, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, Pandey A. A draft map of the human proteome. Nature. 2014; 509(7502):575–81. [PubMed: 24870542]

19. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, Davies SR, Wang S, Wang P, Kinsinger CR, Rivers RC, Rodriguez H, Townsend RR, Ellis MJC, Carr SA, Tabb DL, Coffey RJ, Slebos RJC, Liebler DC, Cptac N. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014; 513(7518):382–7. [PubMed: 25043054]

20. Pang CN, Tay AP, Aya C, Twine NA, Harkness L, Hart-Smith G, Chia SZ, Chen Z, Deshpande NP, Kaakoush NO, Mitchell HM, Kassem M, Wilkins MR. 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. [PubMed: 24152167]

21. Shi Z, Wang J, Zhang B. NetGestalt: integrating multidimensional omics data over biological networks. Nat Methods. 2013; 10(7):597–8. [PubMed: 23807191]

22. Jagtap PD, Johnson JE, Onsongo G, Sadler FW, Murray K, Wang YB, Shenykrnan GM, Bandhakavi S, Smith LM, Griffin TJ. Flexible and Accessible Workflows for Improved Proteogenomic Analysis Using the Galaxy Framework. J Proteome Res. 2014; 13(12):5898–5908. [PubMed: 25301683]

23. Li J, Su ZL, Ma ZQ, Slebos RJC, Halvey P, Tabb DL, Liebler DC, Pao W, Zhang B. A Bioinformatics Workflow for Variant Peptide Detection in Shotgun Proteomics. Mol Cell Proteomics. 2011; 10(5):M110.006536.

24. Wang XJ, Slebos RJC, Wang D, Halvey PJ, Tabb DL, Liebler DC, Zhang B. Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data. J Proteome Res. 2012; 11(2):1009–1017. [PubMed: 22103967]

25. Evans VC, Barker G, Heesom KJ, Fan J, Bessant C, Matthews DA. De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat Methods. 2012; 9(12): 1207–U111. [PubMed: 23142869]

26. Sheynkman GM, Shortreed MR, Frey BL, Smith LM. Discovery and Mass Spectrometric Analysis of Novel Splice-junction Peptides Using RNA-Seq. Mol Cell Proteomics. 2013; 12(8):2341–2353. [PubMed: 23629695]

27. Risk BA, Spitzer WJ, Giddings MC. Peppy: Proteogenomic Search Software. J Proteome Res. 2013; 12(6):3019–3025. [PubMed: 23614390]

28. Prabakaran S, Hemberg M, Chauhan R, Winter D, Tweedie-Cullen RY, Dittrich C, Hong E, Gunawardena J, Steen H, Kreiman G, Steen JA. Quantitative profiling of peptides from RNAs classified as noncoding. Nat Commun. 2014; 5:5429. [PubMed: 25403355]

29. Woo S, Cha SW, Merrihew G, He YP, 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. [PubMed: 23802565]
30. Zickmann F, Renard BY. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. Bioinformatics. 2015; 31(12):106–115.
31. Ruggles KV, Tang ZJ, Wang XY, Grover H, Askenazi M, Teubl J, Cao S, McLellan MD, Clauser KR, Tabb DL, Mertins P, Slebos R, Erdmann-Gilmore P, Li SQ, Gunawardena HP, Xie L, Liu T, Zhou JY, Sun SS, Hoadley KA, Perou CM, Chen X, Davies SR, Maher CA, Kinsinger CR, Rodland KD, Zhang H, Zhang Z, Ding L, Townsend RR, Rodriguez H, Chan D, Smith RD, Liebler DC, Carr SA, Payne S, Ellis MJ, Fenyo D. An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. Mol Cell Proteomics. 2016; 15(3):1060–1071. [PubMed: 26631509]
32. Wang XJ, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics. 2013; 29(24):3235–3237. [PubMed: 24058055]
33. Sheynkman GM, Johnson JE, Jagtap PD, Shortreed MR, Onsongo G, Frey BL, Griffin TJ, Smith LM. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. Bmc Genomics. 2014:15. [PubMed: 24405808]
34. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics. 2010; 73(11):2092–2123. [PubMed: 20816881]
35. Blakeley P, Overton IM, Hubbard SJ. Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. J Proteome Res. 2012; 11(11):5221–5234. [PubMed: 23025403]
36. Tharakan R, Edwards N, Graham DR. Data maximization by multipass analysis of protein mass spectra. Proteomics. 2010; 10(6):1160–71. [PubMed: 20082346]
37. Woo S, Cha SW, Bonissone S, Na SJ, Tabb DL, Pevzner PA, Bafna V. Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer. J Proteome Res. 2015; 14(9):3555–3567. [PubMed: 26139413]
38. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data - Toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteomics. 2006; 5(4):652–670. [PubMed: 16352522]
39. Ma ZQ, Chambers MC, Ham AJ, Cheek KL, Whitwell CW, Aerni HR, Schilling B, Miller AW, Caprioli RM, Tabb DL. ScanRanker: Quality assessment of tandem mass spectra via sequence tagging. J Proteome Res. 2011; 10(7):2896–904. [PubMed: 21520941]
40. Wang X, Li Y, Wu Z, Wang H, Tan H, Peng J. JUMP: a tag-based database search tool for peptide identification with high sensitivity and accuracy. Mol Cell Proteomics. 2014; 13(12):3663–73. [PubMed: 25202125]
41. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29(1):15–21. [PubMed: 23104886]
42. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hickey G, Hinrichs AS, Hubley R, Karolchik D, Learned K, Lee BT, Li CH, Miga KH, Nguyen N, Paten B, Raney BJ, Smit AF, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. 2015; 43(Database issue):D670–81. [PubMed: 25428374]
43. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303. [PubMed: 20644199]
44. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA.

From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; 11(1110):1–11.

45. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164. [PubMed: 20601685]
46. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD, Chen ZH, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29(7):644–U130. [PubMed: 21572440]
47. Xu P, Duong DM, Peng JM. Systematical Optimization of Reverse-Phase Chromatography for Shotgun Proteomics. J Proteome Res. 2009; 8(8):3944–3950. [PubMed: 19566079]
48. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res. 2003; 2(1):43–50. [PubMed: 12643542]
49. 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]
50. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002; 12(6):996–1006. [PubMed: 12045153]
51. Kent WJ. BLAT - The BLAST-like alignment tool. Genome Res. 2002; 12(4):656–664. [PubMed: 11932250]
52. Wang H, Yang Y, Li Y, Bai B, Wang X, Tan H, Liu T, Beach TG, Peng J, Wu Z. Systematic optimization of long gradient chromatography mass spectrometry for deep analysis of brain proteome. J Proteome Res. 2015; 14(2):829–38. [PubMed: 25455107]
53. Jones DR, Wu ZP, Chauhan D, Anderson KC, Peng JM. A Nano Ultra-Performance Liquid Chromatography-High Resolution Mass Spectrometry Approach for Global Metabolomic Profiling and Case Study on Drug-Resistant Multiple Myeloma. Anal Chem. 2014; 86(7):3667–3675. [PubMed: 24611431]
54. Na SJ, Paek E. Quality assessment of tandem mass spectra based on cumulative intensity normalization. J Proteome Res. 2006; 5(12):3241–3248. [PubMed: 17137325]
55. 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]
56. Tabb DL, Ma ZQ, Martin DB, Ham AJL, Chambers MC. DirecTag: Accurate sequence tags from peptide MS/MS through statistical scoring. J Proteome Res. 2008; 7(9):3838–3846. [PubMed: 18630943]
57. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJL, Tabb DL. TagRecon: High-Throughput Mutation Identification through Sequence Tagging. J Proteome Res. 2010; 9(4):1716–1726. [PubMed: 20131910]
58. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005; 33(Database issue):D501–4. [PubMed: 15608248]
59. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. 1998; 26(1):320–2. [PubMed: 9399864]
60. Schuller M, Jenne D, Voltz R. The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease. J Neuroimmunol. 2005; 169(1–2):172–6. [PubMed: 16214224]
61. Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Hamon C. Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem. 2003; 75(8):1895–1904. [PubMed: 12713048]
62. Mertz J, Tan HY, Pagala V, Bai B, Chen PC, Li YX, Cho JH, Shaw T, Wang XS, Peng JM. Sequential Elution Interactome Analysis of the Mind Bomb 1 Ubiquitin Ligase Reveals a Novel Role in Dendritic Spine Outgrowth. Mol Cell Proteomics. 2015; 14(7):1898–1910. [PubMed: 25931508]

63. Sheng QH, Li RX, Dai J, Li QR, Su ZD, Guo Y, Li C, Shyr Y, Zeng R. Preprocessing Significantly Improves the Peptide/Protein Identification Sensitivity of High-resolution Isobarically Labeled Tandem Mass Spectrometry Data. Mol Cell Proteomics. 2015; 14(2):405–417. [PubMed: 25435543]

64. Morgan GJ, Walker BA, Davies FE. The genetic architecture of multiple myeloma. Nat Rev Cancer. 2012; 12(5):335–48. [PubMed: 22495321]

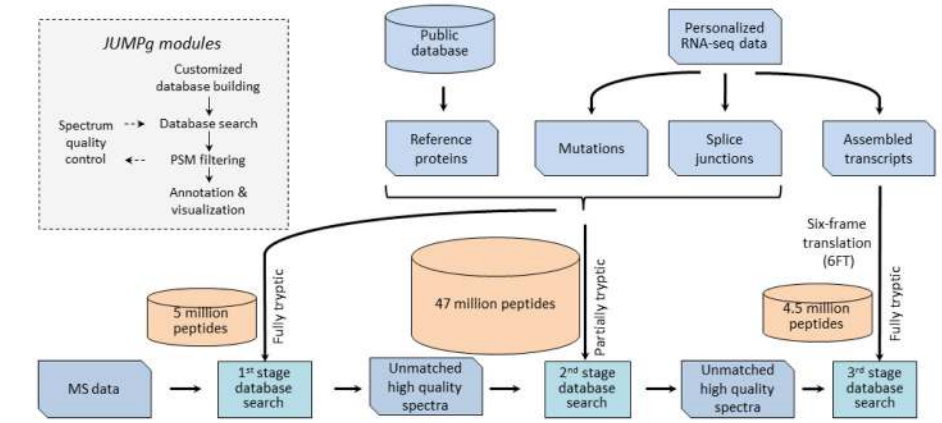
65. Wang H, Tang HY, Tan GC, Speicher DW. Data Analysis Strategy for Maximizing High-Confidence Protein Identifications in Complex Proteomes Such as Human Tumor Secretomes and Human Serum. J Proteome Res. 2011; 10(11):4993–5005. [PubMed: 21955121]

66. Koomen JM. Immunoglobulins: Expanding the Role for Mass Spectrometry in Protein Biomarker Quantification. Clin Chem. 2014; 60(8):1034–1035. [PubMed: 24938750]

67. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3. 0. Syst Biol. 2010; 59(3):307–21. [PubMed: 20525638]

68. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, Johnson R, Segre AV, Djebali S, Niarchou A, Wright FA, Lappalainen T, Calvo M, Getz G, Dermitzakis ET, Ardlie KG, Guigo R. Consortium G. The human transcriptome across tissues and individuals. Science. 2015; 348(6235):660–665. [PubMed: 25954002]

A



B

Peptide	# PSMs	Best PSM	Tag	Jscore	dJn	Database entry	Database
DMGTVVLGK	4	t025.80202.1.2	DMGTVVLGK	69.67	0.53	sp|P15170|ERF3A_HUMAN	Reference
FPASVPTGAQDLISK	3	t034.184014.1.2	FPASVPTGAQDII	91.95	1	AURKB|p.M298T|chr17:8108331.A_G	Mutation
QGVMVGMGQK	2	t030.110766.1.2	QGVMVGMGQK	79.82	1	ACTB|chr7:5569031-5570155	Junction
TEPLYQEK	2	t001.28552.1.2	TEPLYQEK	66.72	0.67	ST8SIA4|chr5:100238701-100238725	6FT

C

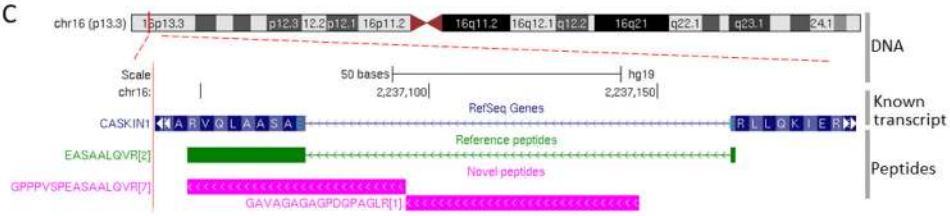


Figure 1. The JUMPg proteogenomic pipeline

(A) Schematic flowchart of JUMPg for peptide identification, with its simplified structure shown in a box. (B) An example of peptide identification table in the JUMPg output. The best scored PSMs are indicated by concatenated LC-MS/MS run name, scan number, the rank of precursor ion intensity in the isolation window, and charge state. (C) Data visualization to display related genes, transcripts and peptides/proteins. JUMPg generates peptide files containing the information of genomic locations (in BED format), which are uploaded into UCSC genome browser.

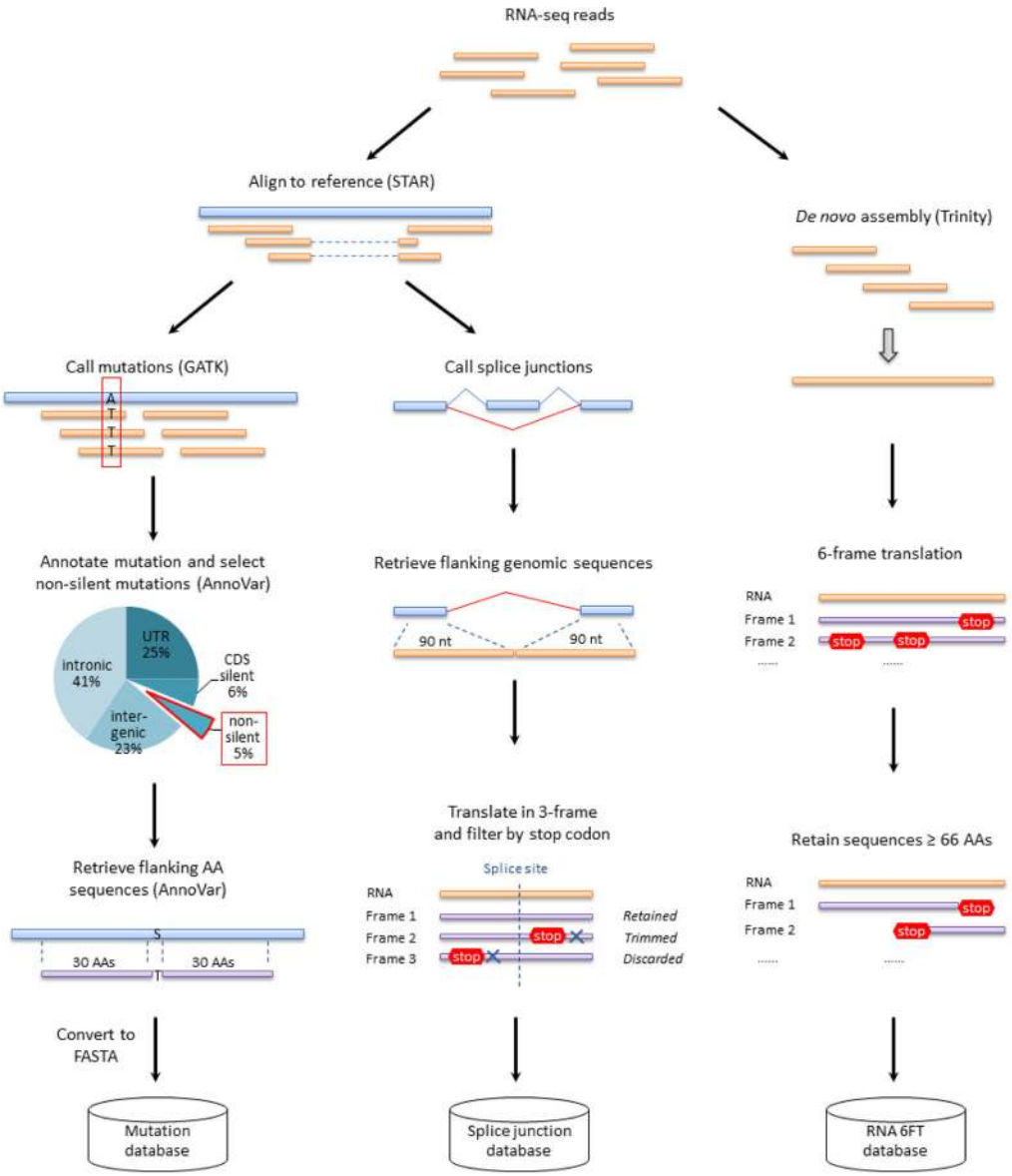


Figure 2. Three customized databases derived from RNA-seq results

RNA-seq reads are aligned to genome reference. Mutations are called, annotated and translated to generate mutation peptide database, while splice junctions are called with flanking genomic sequences translated to build splice junction peptide database. In addition, RNA-seq reads are *de novo* assembled into transcripts and subjected to 6-frame translation to produce another database.

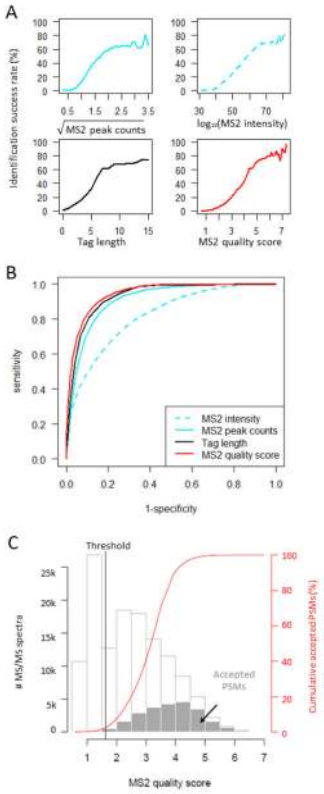


Figure 3. Multistage database search accelerated by MS/MS quality control

(A) PSM identification success rate is positively correlated with MS/MS intensity, MS/MS peak counts, the derived tag length, and MS/MS quality score that integrates the previous three parameters. PSM identification success rate is equal to the accepted PSMs divided by the whole PSMs. (B) ROC curves indicate that MS/MS quality score (red line) outperforms any single feature. (C) The distribution of MS/MS quality score of all scans acquired in one LC-MS/MS run (dataset 1, fraction 7). MS/MS quality score threshold (vertical black line) is selected to retain 99% true positives (grey bars) whereas removing 35% of MS/MS spectra due to poor quality. The red line indicates the accumulative curve of accepted PSMs in the 1st stage analysis.

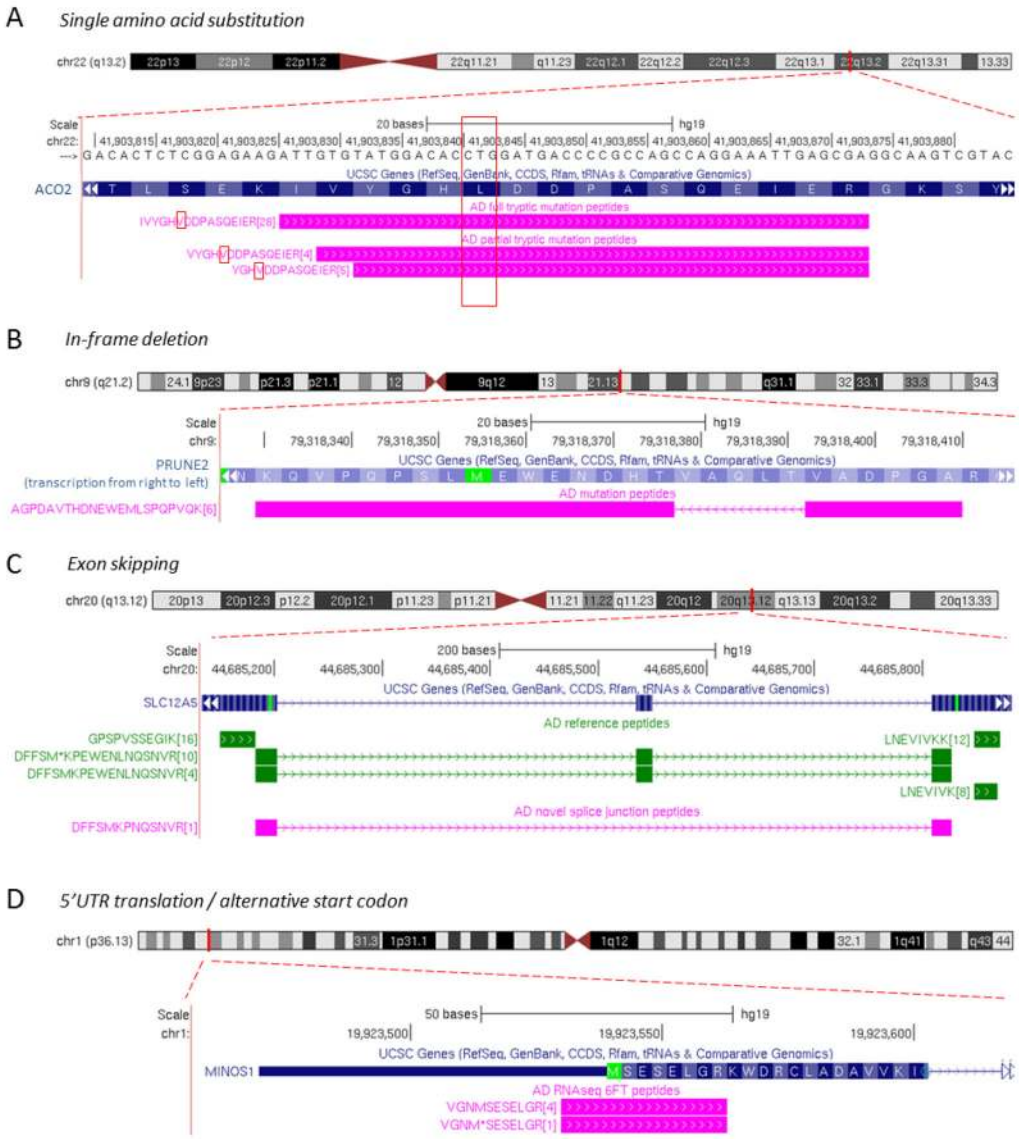


Figure 4. The JUMPg identified novel peptides in the AD brain sample

(A) A single amino acid substitution event (Leucine to Valine) in gene *ACO2* is supported by one fully tryptic and two partially tryptic peptides. The number of PSMs is shown in brackets. (B) A 15-bp in-frame deletion event in gene *PRUNE2* is revealed by a novel peptide (in magenta). (C) An exon-skipping event in gene *SLC12A5*. Reference peptides (in green) are also shown for comparison. (D) Identified peptides (in magenta, Met oxidation indicated by asterisk) support a 5′UTR translation event, indicating alternative start codon usage in the *MINOS1* gene.

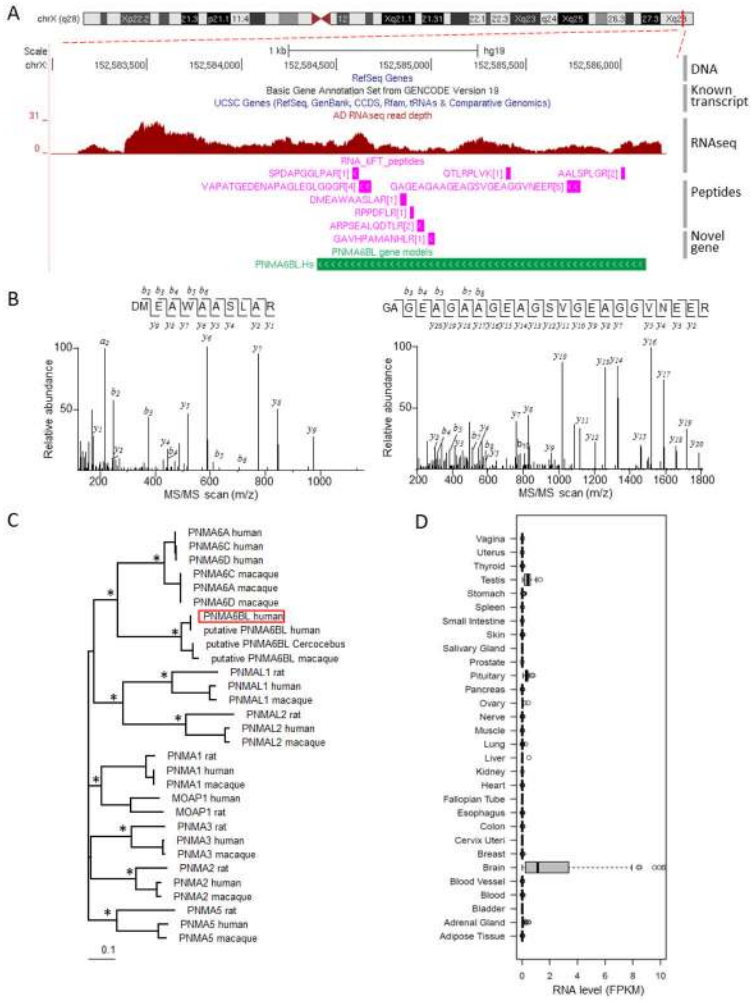


Figure 5. Identification of novel protein coding gene *PNMA6BL* in human brain

(A) A novel gene named *PNMA6BL* is defined by nine peptides in the AD sample. RNA-seq read depth, identified novel peptides, and the newly defined gene model *PNMA6BL* are shown. (B) Examples of MS/MS spectra assigned to *PNMA6BL* peptides. (C) *PNMA6BL* protein sequence is clustered within the PNMA6 branch of the Ma gene family. Asterisks indicate branches with PhyML[67] supported bootstrap values greater than 90%. (D) *PNMA6* is highly expressed in brain, supported by FPKM values in the GTEx dataset. [68]

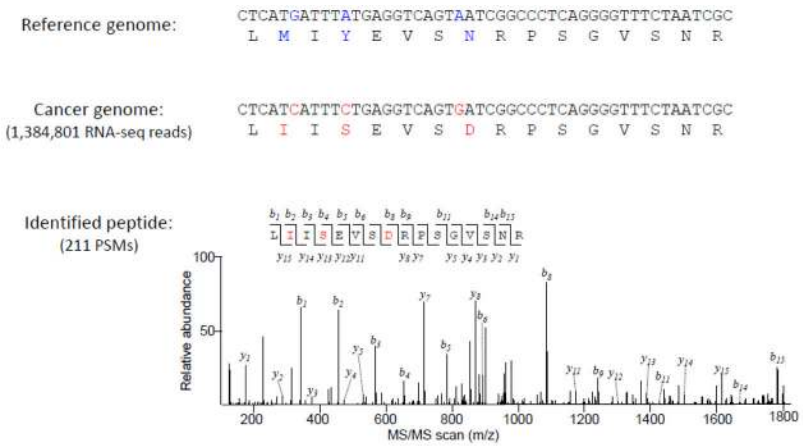


Figure 6. Peptide identification in the immunoglobulin light chain variable region in the MM sample

The peptide carries three amino acid substitutions (highlighted in red) compared to reference genome (hg19), which is supported by 211 PSMs and 1.3 million RNA-seq reads.

Table 1

Summary of peptides identified in Alzheimer's disease and cancer samples by the JUMPg program

Database	Dataset 1: Alzheimer's disease brain			Dataset 2: Multiple myeloma		
	Search space*	PSMs	Peptides	Search space*	PSMs	Peptides
1st stage (fully tryptic)						
UniProt	5,086,244	556,551	87,806	5,086,244	723,521	129,861
Mutation	56,085	1,401	288	63,708	1,931	584
Splice junction	152,662	155	51	742,585	316	163
2nd stage (partially tryptic)#						
UniProt	43,998,844	26,030	6,621	47,273,429	182,737	67,251
mutation	42,761	33	10	98,348	237	114
junction	14,317	0	0	51,899	14	3
3rd stage (fully tryptic)						
RNAseq 6FT	3,633,953	569	147	4,523,368	467	127
Total		584,739	94,923		909,223	198,103

*The search space is indicated by the number of unique theoretical peptides in a particular database.

#Only proteins/events identified in the 1st stage are considered in partially tryptic database search.

Table 2

JUMPg identified novel peptide events

Different Events	AD	MM
Mutation		
Single-AA substitution	255	518
Multi-AA substitution	1	6
Insertion	0	1
Deletion	5	6
Splicing		
Skipped exons	11	58
Within intron / new CDS exon	11	24
Alternative donor	3	17
Alternative acceptor	16	52
Within intergenic regions	4	7
Non-canonical translation		
Frame-shift	5	9
UTR translation	14	19
Intron translation	14	14
Antisense translation	1	1
Non-coding gene translation	1	3
Within intergenic regions	2	2
USCS reference proteins *	17	23
Total	360	760

* Matched to the UCSC database but not in UniProt.