Identification of Novel Peptides from Human Brain Samples: A Pilot Study

Abstract

Identification of novel peptides is very important to reveal the cause of different diseases, and the field of Proteogenomics plays an important role in this regard. Here, novel peptides are identified by searching MS/MS spectra against customized protein sequence databases. These databases contain both known and predicted novel protein sequences, as well as sequence variants that are generated based on genomic and transcriptomic sequence information. Previously our lab has performed a strand-specific RNA sequencing on transcripts from 59 human orbitofrontal cortex samples, and discovered a large number of transcribed and exonized Repetitive Elements (RE). However, it is not known if these novel RE-containing exons are translated into humans, which might have neurological disease relevance. The overall goal of this capstone project was to determine if any of these putative RE-exons are translated to produce proteins in human cells using the proteogenomics approach. Here, we used PGA, an R/Bioconductor package, that enables an automatic process for constructing customized proteomic databases based upon RNA-Seq data, and subsequently search peptides using MS/MS data from publicly available proteomic databases. In this study, we used PRIDE, the most widely used proteomic database which has a very rich deposit of proteomics samples in MS/MS form. After searching against 88 human brain cortex samples, we have discovered 33 different peptides in 26 samples., among them 16 are isomers. Based on these data, we predict that transcripts of RE-exons are potentially translated in the human orbitofrontal cortex. Our study suggests that the PGA based Proteogenomic approach could be a useful tool to identify novel peptides in RNA seq data.

Background

Peptide molecules are building blocks of hormones, toxins, proteins, enzymes etc. NGS-based genomic studies continuously identify new genomic abnormalities such as SNVs, INDELs, RNA edits, novel junctions, and novel transcription regions. In addition, some abnormal DNA and RNA sequences encode novel, disease-relevant proteins, which are promising candidates of disease biomarkers and drug targets. Therefore, today's researchers are very interested to identify novel peptides (1).

Proteomic data is generally obtained using LC-MS/MS, which is called shotgun proteomics. Database-dependent searching is a popular approach for peptide identification by using tandem mass spectrometry (MS/MS) data. Peptides are identified by matching MS/MS spectra against theoretical spectra of all candidate peptides represented in a reference protein sequence database such as Ensembl, RefSeq, or UniProtKB. This type of searching relies on the completeness and quality of the reference database of the proteome (2). However, if a corresponding peptide sequence is not listed in the reference database, an MS/MS spectrum, even at high quality, would fail to identify a peptide. Therefore, generating a comprehensive reference database is a challenging task in bioinformatics analysis of MS/MS signals. Some common databases, such as Ensembl (3), RefSeq (4), and UniProt (5), cannot satisfactorily meet this urgent requirement.

The field of Proteogenomics is developing an alternative comprehensive approach to identify novel peptides. This field combines proteomics, genomics and transcriptomics (study of transcript/RNA) to aid in the discovery and identification of novel peptides. In this way novel peptides are identified by searching MS/MS spectra against customized protein sequence databases which contain predicted novel protein sequences and sequence variants. These databases are generated using genomic and transcriptomic sequence information (6). RNA-Seq technology provides qualitative or quantitative gene expression information on a whole-genome scale at a single-base resolution. As transcriptomic and proteomic analyses could be done on the same cells or tissues, a sample-specific database based upon RNA-Seq data would significantly enhance sensitivity for peptide identification and improve accuracy for finding novel peptides. It is also very important for non-model species whose genome sequences are absent. Therefore, the transcript sequences derived from RNA-Seq data by de novo transcriptome assembly (7) or other methods would be beneficial to construct the proteomic database for MS/MS search .

In this project the RNA seq data came from Darby et al' s published work on repetitive elements (8). The genomes of eukaryotes contain millions of copies of transposable elements and other repetitive sequences. More recent Bioinformatics analyses indicate that repetitive elements (RE) in the human genome might be as high as two-thirds of the whole genome. Under normal conditions endogenous retrotransposons are repressed in human cells, mainly via silencing by promoter DNA methylation (9). However, RE can positively shape the host genome by accelerating novel coding sequences, alternate gene promoters, conserve non-coding elements, and gene networks. They are responsible for the combinations of new DNA that brings an evolutionary advantage to their host. Despite the high abundance of RE in human genome, they are mostly excluded from different genomic, epigenetic and RNA expression studies. However, previous studies demonstrate abundant expression of RE in the human brain cortex and about 10% of the RNA sequencing reads were originated from a RE (8). However, the extent to which each locus was transcribed was still unknown until Darby et al. investigated expression from individual RE in the cortex by performing strand-specific RNA sequencing in 59 human orbitofrontal cortex samples (OFC). In this study, they identified more than 30,000 repetitive elements that are consistently transcribed in the human brain. Most RE expression they detected was likely read through from gene and other promoters. Additionally, there are sequencing reads overlapping splice junctions between the repetitive elements (RE) and annotated exons of known genes, indicating that at least some of the expressed RE are novel exons in previously unannotated mRNA isoforms. They identified transcripts from 10238 different genes in the human OFC that had at least one RE splice junction (8). RE splice junctions occurred predominantly in coding regions of gene transcripts and give rise to novel splice variants with altered coding potential. Similar novel exons were also observed in data from a published mouse study that specifically sequenced RNAs bound to polyribosomes, indicating that they were being translated to make proteins. They found a total of 11041 different transcripts containing RE splice junctions that were associated with translating ribosomes in three different mouse cerebellum cell types. Their results indicate that RE expression is more complex than previously envisioned and raise the possibility that RE splicing may generate novel protein isoforms by extending the open reading frame of endogenous gene transcripts. However, there is no evidence yet that proteins containing the novel RE-containing exons are actually produced in humans. The overall goal of this capstone project was to determine

if any of these putative RE-exons are translated to produce protein in human cells using information from various protein databases.

In order to determine whether any of the novel transcript isoforms discovered by Darby et. al. (8) are translated or not, we used PGA(11), an R/Bioconductor package that enables an automatic process for constructing customized proteomic databases based upon RNA-Seq data with or without guidance from a reference genome, searching peptides using MS/MS data, post-processing and generating an HTML-based report with a visualized interface. In this study publicly available proteomic database PRIDE was used. For this study, we used a total of 88 human brain cortex samples. Among them, 33 different peptides were found in 26 samples. We found 16 isomers. After that, we used UCSC genome browser to determine their genomic location.

Materials and Methods

Data Collection:

Identification of Novel Exons:

The source of the expressed repetitive elements RNA-seq data was described previously (7). The each of the putative RE-exons listed in supplementary table 20 was collected. The RE-exons on the + strands were translated in +1,+2,+3 frames whereas RE exons on the minus strands were translated on the -1, -2 and -3 frames. The Stops Frames list the number of stops in the translated frame.

Download Brain Samples Datasets from PRIDE Database in the Linux Server:

Bioinformatics Analysis:

An R Bioconductor Package PGA was used for Novel Peptide Identification. This package provides functions for construction of customized protein databases based on RNA-Seq data, database searching, post-processing and report generation; therefore, the following steps were performed by us.

- **Customized Protein Database creation from de novo assembly of RNA-seq data without a reference genome:**

Since the novel exons in the dataset were already mapped in the genome (8) the customized protein database was constructed without the annotation information from Ensembl or UCSC genomic database.

- **FASTA file generation from the CSV file containing novel exons:**

The DNA sequences of novel exons from the CSV file were converted in to a FASTA format by using dataframe2fas function from R library SeqRFLP.

- **Creating denovo_txFinder.fasta by PGA**

The FASTA file was input into PGA and it created a denovo_txFinder.fasta file by the function createProDB4DenovoRNASeq.In this fasta file the transcript sequences were translated to protein sequences by three-frame or six-frame translation or based on the longest ORF in all reading frames.

- **MS/MS data searching:**

X!Tandem (11) is a well-accepted and open-source search engine, and was taken as the default database searching method in PGA. In the workflow of PGA, the R package rTANDEM (12) , an R encapsulation of X!Tandem, was automatically used to search the customized proteomic database against MS/MS spectra.

- **Post-processing**

The function parserGear was used to parse the search result after the MS/MS data searching completed. It calculated the q-value for each peptide spectrum matches (PSMs) and then utilized the Occam's razor approach (13). This approach deals with degenerated wild peptides by finding a minimum subset of proteins that covered all of the identified wild peptides. It exported some tab-delimited files containing the peptide identification result and protein identification result.

Visualization of novel peptides

Using the UCSC genome browser we tried to determine the genomic location of novel peptides. In this step at first, we collected the novel exon sequence of predicted novel peptides corresponding gene and submitted that to the blat search in the USCS genome browser for confirming the genomic location.

Results

Enumeration of Novel Peptides:

In this mini project our goal was to see the evidence of translated novel peptides in the human brain samples. Therefore, at first, we selected 22 human brain samples shotgun proteomics datasets from the PRIDE database and downloaded those in the BIFX server.

Since it was a pilot study one dataset was selected for further analysis. Its PRIDE ID was PXD006537.Samples of this dataset were from the human brain cortex region. This dataset contained a total of 754 samples which were divided into test sets and validation sets. We combined the control samples from both the validation and test set to make one large discovery set (330 samples) as our initial dataset.

Over the course of our pilot study, we searched the spectra generated by shotgun proteomics sequencing of 88 samples. Among them 26 samples had at least one novel peptide while 62 samples didn't have any of our target peptides (table1). Total 32 novel peptides were detected. The peptide AQPDRCLGR in FBLN7 gene was found most frequently.

Total how many samples	88
How many samples had peptides	26
How many samples didn't have peptides	62
How many isoforms	16
Which peptide found mostly in which gene	AQPDRCLGR found in FBLN7 gene
How many different peptides found	33

Figure Table 1: Summarize the whole findings

Differentiate Peptides based on E value:

The identified peptides were differentiated based on the E value (<0.05). In this study we got 22 peptides which e value is less than 0.05. These peptides were considered most significant among 32 identified peptides. In addition, some novel exons had isomers. We got 16 novel peptide

isomers. Table 4 represents all the information about novel peptides such as sample name, gene name, position in the chromosome, sequences of peptide and their corresponding E value.

Datasets Name	Gene name	Chromosome location	Novel Peptide Sequences	E value
Alz_P11_C11_872_13Jul12_Roc_12-04-09.mgf_xtandem.xml	RPN2	chr20:35868015-35868128	RVHFSADKLQLH	0.03
Alz_P10_H01_373_14May12_Roc_12-03-29.mgf_xtandem.xml	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.012
Alz_P09_E02_242_7May12_Roc_12-03-30.mgf_xtandem.xml	MFSD6	chr2:191291584-191291871	ESGLSPLASSK	0.025
Alz_P08_C09_609_22Jun12_Roc_12-03-29.mgf_xtandem.xml	SMARCA5	chr4:144455321-144455430	QVTFPSASK	0.023
Alz_P07_F06_066_16Apr12_Roc_12-03-26.mgf_xtandem.xml	::::|+|F1|1	chr16:2032660-2032752	QSPCLGLPK	0.025
Alz_P06_F01_947_7Sep12_Roc_12-03-29.mgf_xtandem.xml	::::|+|F1|1	chr16:2032660-2032752	QSPCLGLPK	0.024
	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.025
Alz_P06_D07_929_7Sep12_Roc_12-04-08.mgf_xtandem.xml	FBLN7		AQPDRCLGR	0.027

		chr2:11292 1446-112921572			
Alz_P06_D02_924_7Sep12_Roc_12-03-30.mgf_xtandem.xml	FBLN7	chr2:11292 1446-112921572	AQPDRCLGR	0.012	
Alz_P05_H10_574_12Jun12_Roc_12-03-30.mgf_xtandem.xml	PLA2G61	chr22:3852 7996-38528118	DLGSPQPPPPR	0.027	
Alz_P05_G07_559_12Jun12_Roc_12-04-08.mgf_xtandem.xml	FBLN7	chr2:11292 1446-112921572	AQPDRCLGR	0.0045	
Alz_P05_F04_544_7Jun12_Roc_12-04-09.mgf_xtandem.xml	LINC01140	chr1:87627 704-87627922	VFLLPLCK	0.038	
Alz_P05_B05_497_7Jun12_Roc_12-03-29.mgf_xtandem.xml	LINC01140	chr1:87627 704-87627922	VFLLPLCK	0.0098	
Alz_P05_B05_497_7Jun12_Roc_12-03-29.mgf_xtandem.xml	FBLN7	chr2:11292 1446-112921572	AQPDRCLGR	0.021	
Alz_P05_A01_481_6Jun12_Roc_12-03-29.mgf_xtandem.xml	FBLN7	chr2:11292 1446-112921572	AQPDRCLGR	0.012	

	LINC01140	chr1:87627704-87627922	VFLLPLCK	0.012	
Alz_P04_G08_848_2Jul12_Roc_12-04-09.mgf_xtandem.xml	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.011	
	SYT338	chr19:51141332-51141590	DDMEPATGGGQWR	0.0058	
Alz_P02_H09_765b_11Sep12_Roc_12-03-29.mgf_xtandem.xml	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.0140	
Alz_P02_G04_748b_11Sep12_Roc_12-04-09.mgf_xtandem.xml	RPN2	chr20:35868015-35868128	RVHFSADKLQLH	0.032	
	RARS2 FBLN7	chr6:88274261-88274380	IVPLHSSLGDK AQPDRCLGR	0.015	
Alz_P02_F02_734b_11Sep12_Roc_12-03-30.mgf_xtandem.xml	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.018	

Visualization of the Genomic Location of Novel peptide:

At first the novel exon sequence of predicted novel peptides corresponding gene were collected. After that the sequence were submitted to the BLAT search in the UCSC genome browser for confirming the genomic location. However, the peptides were too short to be confirmed by BLAT. Therefore, future direction will be to find method to visualize protein results.

Summary

In summary we can say that PGA, a Bioconductor package based on Proteogenomic is a way to identify novel peptides in RNA seq data. We have discovered that protein isoforms containing repetitive elements are potentially translated in the human orbitofrontal cortex. Our applied workflow has opened up a new potential to discover many previously unidentified novel peptides expressed in the human brain samples.

Future Direction:

The importance of this study to identify novel peptide is vast because these will be suitable for further investigation in future studies. In future, we will develop method to confirm and visualize peptide results. In addition, we will identify expressed RE in a large number of brains samples. We will perform a literature search to identify proteins with the most critical functions and prioritize those. Furthermore, we will determine their presence in the normal vs diseases patients and determine their relationship to disease by literature search and wet lab experiment.

Supplementary Table:

Datasets Name	Number of peptides found	Gene name	Chromosome location	Novel Peptide Sequences	E value
Alz_P11_C11_872_13Jul12_Roc_12-04-09.mgf_xtandem.xml	1	RPN2	chr20:35868015-35868128	RVHFSADKLQLH	0.03
Alz_P10_H01_373_14May12_Roc_12-03-29.mgf_xtandem.xml	1	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.012
Alz_P09_F10_262_8May12_Roc_12-03-30.mgf_xtandem.xml	1	BAIAP2	chr17:79036601-79036699	LPAPTAAVFR	0.063
Alz_P09_E02_242_7May12_Roc_12-03-30.mgf_xtandem.xml	1	MFSD6	chr2:191291584-191291871	ESGLSPLASSK	0.025
Alz_P08_C09_609_22Jun12_Roc_12-03-29.mgf_xtandem.xml	3	SMARCA5	chr4:144445321-144455430	QVTFPSASK	0.023
		CCSER1	chr4:91325074-91325170	PLKELDHR	0.06
		FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.099
Alz_P07_F06_066_16Apr12_Roc_12-03-26.mgf_xtandem.xml	1	::::|+|F1|1		QSPCLGLPK	0.025

			chr16:2032660-2032752		
Alz_P06_F01_947_7Sep12_Roc_12-03-29.mgf_xtandem.xml	2	::::|+|F1|1	chr16:2032660-2032752	QSPCLGLPK	0.024
		FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.025
Alz_P06_D07_929_7Sep12_Roc_12-04-08.mgf_xtandem.xml	1	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.027
Alz_P06_D02_924_7Sep12_Roc_12-03-30.mgf_xtandem.xml	1	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.012
Alz_P05_H10_574_12Jun12_Roc_12-03-30.mgf_xtandem.xml	2	PLA2G6	chr22:38527996-38528118	DLGSPQPPPPR	0.027
		LINC01140	chr1:87627704-87627922	VFLLPLCK	0.056
Alz_P05_G07_559_12Jun12_Roc_12-04-08.mgf_xtandem.xml	1	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.0045

Alz_P05_F04_544_7Jun12_Roc_12-04-09.mgf_xtandem.xml	1	LINC01140	chr1:87627704-87627922	VFLLPLCK	0.038
Alz_P05_D10_526_7Jun12_Roc_12-03-30.mgf_xtandem.xml	1	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.05
Alz_P05_B05_497_7Jun12_Roc_12-03-29.mgf_xtandem.xml	2	LINC01140	chr1:87627704-87627922	VFLLPLCK	0.0098
		FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.021
Alz_P05_A01_481_6Jun12_Roc_12-03-29.mgf_xtandem.xml	2	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.012
		LINC01140	chr1:87627704-87627922	VFLLPLCK	0.012
Alz_P04_G11_851_2Jul12_Roc_12-04-08.mgf_xtandem.xml	1	TNFRSF11A	chr18:60020588-60020849	IDFGVQINFIEQ	0.099
Alz_P04_G08_848_2Jul12_Roc_12-04-09.mgf_xtandem.xml	1	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.011

Alz_P04_G08_848_2Jul12_Roc_12-04-09.mgf_xtandem.xml	2	SYT3 38:50	chr19:51141332-51141590	DDMEPATGGGQWR	0.0058
Alz_P04_G08_848_2Jul12_Roc_12-04-09.mgf_xtandem.xml		FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.09
Alz_P04_A11_779_2Jul12_Roc_12-04-08.mgf_xtandem.xml	1	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.09
Alz_P03_G07_463_21May12_Roc_12-04-08.mgf_xtandem.xml	1	RPN2	chr20:35868015-35868128	RVHFSADKLQLH	0.091
Alz_P02_H09_765b_11Sep12_Roc_12-03-29.mgf_xtandem.xml	1	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.0140
Alz_P02_G04_748b_11Sep12_Roc_12-04-09.mgf_xtandem.xml	3	RPN2	chr20:35868015-35868128	RVHFSADKLQLH	0.032
		RARS2		IVPLHSSLGDK	0.015

		FBLN7	chr6:88274261-88274380	AQPDRCLGR	
Alz_P02_F02_734b_11Sep12_Roc_12-03-30.mgf_xtandem.xml	1	FBLN7	chr2:112921446-112921572	AQPDRCLGR	0.018
Alz_P02_F02_734b_11Sep12_Roc_12-03-30.mgf_xtandem.xml	1	RPN2	chr20:35868015-35868128	RVHFSADKLQLH	0.082

References

1. Boycott K.M., Vanstone M.R., Bulman D.E., MacKenzie A.E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat. Rev.Genet. 2013;14(10):681–691.
2. Mann M, Kulak NA, Nagaraj N, Cox J. The coming age of complete, accurate, and ubiquitous proteomes. Mol Cell. 2013; 49:583–590. [PubMed: 23438854]
3. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. Ensembl 2013. Nucleic Acids Res. 2013;41(Database issue):D48–55.
4. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 2012;40(Database issue):D130–135.
5. UniProt C. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 2013;41(Database issue):D43–47.
6. Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics. 2004; 4:59–77. [PubMed: 14730672]

7. 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;15:703. doi: 10.1186/1471-2164-15.
8. Darby MM[1], Leek JT[2], Langmead B[3], Yolken RH[1], Sabunciyan S[1]. Widespread splicing of repetitive element loci into coding regions of gene transcripts. Hum Mol Genet. 2016 Nov 15;25(22):4962-4982. doi: 10.1093/hmg/ddw321.
9. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 7:e1002384.

10. Bo Wen, Shaohang Xu,[#] Ruo Zhou, Bing Zhang, Xiaojing Wang, Xin Liu, Xun Xu, and Siqi Liu. PGA: an R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq. BMC Bioinformatics. 2016; 17: 244.doi: 10.1186/s12859-016-1133-3

11. R Craig and R C Beavis. Tandem: matching proteins with tandem mass spectra. Bioinformatics, 20(9):1466–7, Jun 2004. doi:10.1093/bioinformatics/bth092.
12. Frederic Fournier, Charles Joly Beauparlant, Rene Paradis, and Arnaud Droit. rTANDEM: Encapsulates X!Tandem in R., 2013. R package version 1.2.0.
13. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003;75(17):4646–58.