Published in final edited form as:

J Proteome Res. 2017 September 01; 16(9): 3336–3347. doi:10.1021/acs.jproteome.7b00324.

Integrating Next-Generation Genomic Sequencing and Mass Spectrometry to Estimate Allele-Specific Protein Abundance in Human Brain

Thomas S. Wingo^{1,2,3,*}, Duc M. Duong⁴, Maotian Zhou^{2,4}, Eric B. Dammer⁴, Hao Wu⁵, David J. Cutler³, James J. Lah², Allan I. Levey², and Nicholas T. Seyfried^{2,4,*}

¹Division of Neurology, Department of Veterans Affairs Medical Center, Decatur, GA 30033

²Department of Neurology, Emory University, Atlanta, Georgia 30322

³Department of Human Genetics, Emory University, Atlanta, Georgia 30322

⁴Department of Biochemistry, Emory University, Atlanta, Georgia 30322

⁵Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia 30322

Abstract

Gene expression contributes to phenotypic traits and human disease. To date, comparatively less is known about regulators of protein abundance, which is also under genetic control and likely influences clinical phenotypes. However, identifying and quantifying allele-specific protein abundance by bottom-up proteomics is challenging since single nucleotide variants (SNVs) that alter protein sequence are not considered in standard human protein databases. To address this, we developed the GenPro software and used it to create personalized protein databases (PPDs) to identify single amino acid variants (SAAVs) at the protein level from whole exome sequencing. *In silico* assessment of PPDs generated by GenPro revealed only a 1% increase in tryptic search space compared to a direct translation of all human transcripts and an equivalent search space compared to the UniProtKB reference database. To identify a large unbiased number of SAAV peptides, we performed high-resolution mass spectrometry-based proteomics for two human post-mortem brain samples and searched the collected MS/MS spectra against their respective PPD. We found an average of ~117,000 unique peptides mapping to ~9,300 protein groups for each sample and of these 977 were unique variant peptides. We found that over 400 reference and SAAV peptide pairs were, on average, equally abundant in human brain by label-free ion intensity measurements and confirmed the absolute levels of three reference and SAAV peptide pairs using

Corresponding Authors: Thomas S. Wingo, 505K Whitehead Building, 615 Michael Street NE, Atlanta, GA 30322-1047, thomas.wingo@emory.edu; Nicholas T. Seyfried, 4133 Rollins Research Center, Atlanta, GA 30322-1047, nseyfri@emory.edu.

Supporting Information

Software available at <https://github.com/wingolab-org/GenPro>

The following files are available free of charge at ACS website <http://pubs.acs.org>:

Author Contributions

T.S.W. and N.T.S. designed the study. T.S.W. designed and wrote GenPro and performed experiments. D.M.D., M.Z., and E.B.D. performed experiments. All authors contributed to the analysis and interpretation of the data. T.S.W. and N.T.S. wrote the manuscript with contributions from all authors. All authors have given approval to the final version of the manuscript.

heavy labeled peptides standards coupled with parallel reaction monitoring (PRM). Our results highlight the utility of integrating genomic and proteomic sequencing data to identify sample-specific SAAV peptides and support the hypothesis that most alleles are equally expressed in human brain.

TOC image

Keywords

proteogenomics; proteomics; neurodegeneration; precision medicine; allele-specific protein abundance

Introduction

Gene expression is an important contributor to phenotypic traits and disease. Much of the variation in gene expression is attributable to genetic variants, mostly single nucleotide variants (SNVs)¹⁻². Although protein abundance has become increasingly recognized as an important trait under genetic control³, the majority of the focus of gene expression is on mRNA abundance, typically measured by gene microarrays or RNA sequencing (RNA-seq). SNVs adjacent to genes of interest are often correlated with differences in either RNA or protein abundance, known as expression quantitative trait loci (eQTL) or protein quantitative trait loci (pQTL)³, respectively. Recent evidence suggests eQTLs and pQTLs are heritable, common throughout the genome, and enriched for SNVs that associate with human disease⁴⁻⁸. Both eQTLs/pQTLs are statistical associations between transcript RNA or protein abundance and nearby SNVs are influenced by cis- and trans- regulatory mechanisms⁹. By contrast, allele-specific expression reflects the extent each allele is expressed in a diploid organism and directly assesses the effect of cis-regulatory variants. The natural expectation is that each allele will be equally expressed, yet up to 30% of

expressed genes show allele-specific differences in RNA expression⁹. Allele-specific expression studies have generally focused on mRNA expression and inquiries have yielded insights into traits and disease⁹. An important, yet technically challenging question that remains largely unaddressed is whether allele-specific differences are observed at the protein-level. Allele-specific protein expression could be used to understanding the influence of regulatory or missense SNVs on protein expression and is particularly relevant for neurodegenerative diseases since pathological aggregation is considered a hallmark of most of those diseases¹⁰.

Advances in liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) now allow detection of thousands of expressed proteins at unprecedented depth from disease-relevant tissues¹¹. However, unlike DNA and RNA, which are directly sequenced, the sequence of peptides in bottom-up proteomics is inferred by matching MS/MS spectra against the theoretical spectra of all potential candidate peptides represented in a protein database^{12–13}. All individuals carry missense SNVs that will generate novel variant proteins or proteoforms¹⁴ that are not present in standard protein databases. Without incorporating knowledge of missense SNVs into protein databases prior to LC-MS/MS data analysis, the identification of single amino acid variant (SAAV) peptides is difficult. To overcome this limitation, customized protein sequence databases that integrate genomic or transcriptomic information are being used to identify SAAV peptides¹².

One proteogenomic approach for detecting novel coding regions and variant peptides includes using databases generated from 6-frame translation of the reference genome from known or predicted transcripts^{15–17}. Although 6-frame translation has the advantage of being independent of any *a priori* annotation, the approach has major drawbacks because it is computationally expensive and introduces an artificial six-fold increase in the protein database size¹⁷, which can result in a bias in peptide identification and a substantially inflated false-discovery rate (FDR) since most peptides generated *in silico* are likely not present *in vivo*¹⁸. Other approaches include creation of protein databases that incorporate knowledge of known SNVs and other genomic variants from public repositories (*e.g.*, dbSNP)¹⁹ or even cancer-specific SNPs, such as CanVarPro (human cancer proteome variation)²⁰ or The Cancer Genome Atlas (TCGA)²¹, to construct databases. The main drawback to these approaches is the reduced sensitivity of peptide identification since most of the individual samples being analyzed are not expected to contain most coding variants present in these databases, particularly SAAV peptides that result from rare SNVs²². Conversely, a SAAV peptide that results from an SNV not in these public databases would be missed. More recently, the advent of next-generation sequencing has enabled creation of sample-specific databases¹². One common approach uses RNA-seq data to generate customized cell- or tissue-specific protein databases derived from translation of annotated transcripts^{12, 23–24}. The advantages of this approach are the inclusion of SNVs²², alternative transcripts (*e.g.*, splice junction peptides)²⁵, and alternative start sites²⁶ that are more likely to be present in the sample. However, the main drawback is that it requires abundant and high-quality RNA to make SNV calling from RNA-seq even moderately reliable²⁷. Furthermore, tissue-specific protein databases generated from RNA-seq will not incorporate proteins that are produced in one organ (*e.g.*, liver) and transported to another (*e.g.*, brain). For example, in the brain we have shown that peripheral proteins often found in plasma are

commonly deposited in the brain and associated with neurodegenerative disease²⁸. Genomic sequencing, *e.g.* whole-exome sequencing (WES) or whole genome sequencing (WGS), can be used instead of RNA-seq to construct protein databases^{29–30} and may be the only option for human postmortem tissues with poor quality RNA but intact protein.

The goal of this study was to determine the feasibility of measuring allele-specific protein abundance in post-mortem human tissue, and to determine the relative abundance of SAAV and reference peptide pairs. To do this, we first developed an open-source software, GenPro, to create personalized protein databases (PPDs) using variant calls from next-generation genomic sequencing data (*e.g.*, WES, WGS, or RNA-seq). As proof of principle, we then used GenPro to generate PPDs from WES from two individuals. We show these databases do not appreciably differ from standard reference databases (*i.e.*, UniProtKB), aside from variant peptides they introduce. Next, we tested whether we could identify predicted SAAV peptides in postmortem human brain from the same two individuals. To achieve this goal, we deeply profiled the brain proteomes to enhance our probability of detecting sufficient number of SAAV peptides. We employed off-line fractionation with electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) and liquid-chromatography coupled to tandem mass spectrometry (LC-MS/MS) on an Orbitrap Fusion. We found an average of approximately 117,000 unique peptides mapping to approximately 9,300 protein groups for each sample. Of these, 977 unique variant peptides were identified across the two samples. Label-free quantification was used to estimate allelic balance of reference and SAAV peptide pairs, *i.e.* pairs that differ by a single amino acid. On average, reference:SAAV peptides were equally abundant in human brain. We confirmed these findings with measures of the absolute levels of select reference:SAAV peptide pairs using parallel reaction monitoring (PRM). This highlights the utility of our discovery-validation pipeline to estimate allele-specific protein balance, which supports a hypothesis that most alleles are equally expressed at in human brain.

Experimental Procedures

Tissue Homogenization and Digestion

We obtained fresh-frozen post-mortem human brain (dorsolateral prefrontal cortex; broadman area 9) from two cases from the Emory Brain Bank. Each tissue sample was processed as described²⁸ with slight modification. Briefly, each tissue was individually weighed (approximately 0.1 g) and homogenized (Dounce homogenizer) in 500 μ L of urea lysis buffer (8M urea, 50 mM Tris-HCl, pH 7.8), including both protease inhibitors (Roche) and the HALT (Pierce) phosphatase inhibitor cocktail, 0.6% (v/v). Samples were sonicated (Sonic Dismembrator, Fisher Scientific) 5 times for 5 sec with 15 sec intervals of rest at 30% amplitude to disrupt nucleic acids, and then centrifuged at 22,800 rcf at 4°C for 5 min. Protein concentration was determined by the bicinchoninic acid (BCA) method, and then samples were frozen at -80°C. Protein samples (1 mg) were treated with 1 mM dithiothreitol (DTT) at 37°C for 30 min, followed by 5 mM iodoacetamide (IAA) at 37°C for 30 min. Samples were first digested with 1:200 (w/w) lysyl endopeptidase (LysC; Wako) at 37°C for 4 h, and then diluted with 50 mM NH_4HCO_3 to a final concentration of 1.6 M urea and digested overnight with 1:50 (w/w) trypsin (Promega) at 37°C. For LysC digestion alone,

samples were diluted with 50 mM NH_4HCO_3 to a final concentration of 0.6 M urea and digested overnight with 1:50 (w/w) lysyl endopeptidase (LysC). All peptides were desalted on C_{18} Sep-Pak (Waters Corporation) and lyophilized to dryness.

Electrostatic Repulsion-Hydrophilic Interaction Chromatography (ERLIC)

ERLIC fractionation was performed as previously described³¹ with slight modifications. We used the concatenated ERLIC pooled strategy as it was shown that the number of unique peptides identified in concatenated ERLIC is significantly higher than pooling adjacent fractions using the same LC-MS/MS gradient³². Briefly, trypsin or LysC-derived peptides generated from 1 mg of protein were dissolved in 100 μL of 80% (v/v) loading buffer (10 mM NH_4Ac , 85% ACN/1% acetic acid), injected completely with an auto-sampler, and fractionated using a PolyWAX LP anion-exchange column (200×3.2 mm, 5 μm , 300 $\text{\AA}$; PolyLC, Columbia, MD) on an Agilent 1100 HPLC system monitored at 280 nm. Forty fractions were collected with a 66-min gradient of 100% mobile phase A (90% ACN/0.1% acetic acid) for 3 min, 0%–20% mobile phase B (30% ACN/0.1% FA) for 50 min, 20%–100% B for 5 min, followed by 8 min at 100% B at a flow rate of 0.3 ml/min. The 40 fractions were pooled into 20 fractions by combining in the following manner, (1, 40); (2, 39); (3, 38), and so on.

LC-MS/MS Analysis

An equal volume of each of the 20 ERLIC peptide fractions (LysC or trypsin) was re-suspended in loading buffer (0.1% formic acid, 0.03% trifluoroacetic acid, 1% acetonitrile), and peptide eluents were separated on a self-packed C18 (1.9 μm Dr. Maisch, Germany) fused silica column (25 cm $\times$ 75 μm internal diameter (ID); New Objective, Woburn, MA) by a Dionex UltiMate 3000 RSLCnano system (ThermoFisher Scientific) and monitored on an Thermo Orbitrap Fusion mass spectrometer (ThermoFisher Scientific). Elution was performed over a 120-min gradient at a rate of 300 nl/min with buffer B ranging from 1% to 90% (buffer A: 0.1% formic acid in water, buffer B: 0.1 % formic in acetonitrile). The mass spectrometer cycle was programmed to sequence at top speed with a cycle time of 5 sec. For samples digested with both LysC and trypsin, the MS scans were collected at a resolution of 60,000 (300–1800 m/z range, 200,000 AGC, 50 ms maximum ion time). For samples digested with only LysC, the MS scans were collected with the same resolution and ion time but with a m/z range of 350–1500 and an AGC of 500,000. All HCD MS/MS spectra were acquired at a resolution of 17,500 (0.7 m/z isolation width, 30% collision energy, 20,000 AGC target, 35 ms maximum ion time) in the Orbitrap. Dynamic exclusion was set to exclude previously sequenced peaks for 20 sec within a 10-ppm isolation window. Only those precursors with charge state 2–6 were sampled for MS/MS. All raw mass spectrometry files and matched peptides will be made public on Synapse, <https://www.synapse.org/#!/Synapse:syn6126101>.

Whole-Exome Sequencing

We extracted genomic DNA from fresh-frozen brain tissue using a Gentra PureGene kit (Qiagen, Germany) per the manufacturer's protocol. Whole-exome sequencing was performed using the SeqCap EZ Exome v2 Human Exome Capture Kit (Roche, Switzerland) according to the manufacturer's protocol, with 100-bp paired-end sequencing on an Illumina

HiSeq 2000 instrument to an average read depth of 100X. Raw sequencing reads were mapped to the whole human genome (GRCh38) using the PEMapper and variant sites were identified using PECO with a theta of 0.001 and probability to call a variant site of 95% within the approximately 32 million bases that constitute the consensus coding domain³³.

Personalized Protein Database Construction

The GenPro software is an open-source application developed to generate PPDs. A complete description of the software is available at <https://github.com/wingolab-org/GenPro> where it may be downloaded. Briefly, prior to creating PPDs a binary database for the organism of interest must be created. This database contains knowledge of transcripts (*i.e.*, translation start, translation stop, and exon boundaries) and genomic sequence. GenPro can create a binary database for the following organisms: *Homo sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, or *S. cerevisiae*. The software generates the reference proteome through *in silico* in-frame translation of each transcript in the binary database (Supplementary Figure 1). To generate a PPD for a given sample, the software reads variants from either a VCF- or SNP-formatted file^{33–34}, and creates all novel proteins for each missense SNV on a per sample basis. Each variant protein is then digested *in silico* with a user-specified protease and only if the variant protein results in one or more novel peptides (*i.e.*, peptides not present in the reference proteome) is the variant protein added to the PPD. The user may impose a maximum or minimum length requirement on the peptides to consider (*e.g.*, default minimum and maximum peptide length of 6 and 40). A final PPD consists of all full-length translated proteins (reference and variant) are written to a single FASTA file that may be used as input for Proteome Discoverer 2.0 software (Thermo Scientific, Bremen, Germany). The FASTA header encodes information about whether the transcript is reference or variant, and for variant proteins, it encodes the location(s) and amino acid substitution of the variant(s).

For this study, we used the human genome (GRCh38) and UCSC knownGene transcripts (<http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg38&g=knownGene>, accessed: 4/15/2014)³⁵, which aimed to be a complete set of well-described genes and incorporates data from RefSeq, GenBank, the Consensus Coding Domain Project, and comparative genomics. We use trypsin and LysC protease-specific settings to construct the PPDs with a minimum and maximum peptide length of 6 and 40, respectively. The upper length requirement was chosen to limit database size and because under the same experimental conditions we observed that only approximately 1% of all peptides sequenced are longer than 40 amino acids (data not shown).

Database Searching

MS/MS data obtained were searched against personal protein databases, which are described in results 3.1, and the UniProt Knowledgebase (UniProtKB) containing both Swiss-Prot and TrEMBL human reference protein sequences (90,411 target sequences downloaded April 21, 2015) using the SEQUEST algorithm through the Proteome Discoverer 2.0 platform (Thermo Scientific, Bremen, Germany). Searching parameters against a target-decoy database included mass tolerance of precursor ions (± 10 ppm), 0.05 Da mass tolerance for product ions, fully tryptic or LysC restriction, dynamic modifications for oxidized Met

(+15.9949 Da), deaminated Asn and Gln (+0.98480 Da), protein N-terminal acetylation (+42.03670 Da), static modification for carbamidomethyl Cys (+57.0215 Da), 4 maximal modification sites, and a maximum of 2 missed cleavages. Only b and y ions were considered for scoring, and the embedded Percolator algorithm³⁶ was used to filter the peptide spectral matches (PSMs) to achieve a false-discovery rate (FDR) of <1%. Peptide and protein grouping were followed per Proteome Discoverer's algorithm while applying a strict maximum parsimony principle. For variant-containing peptide identification, we applied the following filters on identified PSMs: amino acid changes that could be due to either spontaneous N/Q-deamination (+0.98 Da) or leucine (L)-to-isoleucine (I) coding changes were not considered.

Validation of Variant Peptides with Synthetic Standards and Label-free Quantification of Allele-Specific Abundance

Custom Peptide Synthesis (ThermoFisher Scientific) was used to validate variant peptides harboring single amino acid substitutions in PLEC, HSPA12A, GAA, TJP1, APOA4, and MINK1. All peptides were mixed together at equal weight (1 µg total) and analyzed by LC-MS/MS using a targeted (data-independent acquisition) method on a high-resolution Q-Exactive Orbitrap. Each cycle consisted of one MS1 scan for each precursor collected at a resolution of 70,000 followed by 10 MS/MS (2 m/z isolation width, 30% collision energy, 50,000 AGC target, 50-ms maximum ion time) at 17,500 resolution targeting predefined precursor ion masses (m/z) derived from the matched variants peptides from human brain. MS/MS HCD spectra for targeted synthetic variant peptides were searched by SEQUEST against the personalized protein database (target-decoy) to confirm identification using the same search and peptide filtering parameters as described above. Estimates of the relative abundance of SAAV and reference peptides in human brain were determined from the chromatographic peak area calculation in Proteome Discover (Thermo Scientific, Bremen, Germany) of the extracted ion chromatograms (XICs) corresponding to the variant and reference peptides with the same charge state (m/z). The stoichiometry or allele-specific abundance was quantified as a ratio of the chromatographic peak area of the variant peptide to the reference peptides. A peptide precursor mass tolerance of ±4 ppm for extracted ion chromatograms and a minimum signal-to-noise threshold of 1.0 was used as default parameters in Proteome Discoverer. The log₂ ratios of the peak area for the reference and variant were plotted versus the average of the log₂ intensities (*i.e.*, as an MA plot). For the plot, we calculated a sliding window standard deviation (SD) and plot ± the 2-fold change lines in red using R (v3.2.3, 2015-12-10)³⁷. All statistical testing was performed using built-in functions implemented in the same version of R.

Parallel Reaction Monitoring (PRM) for Absolute Quantification of Allelic Abundance

Since the above XIC-based approach assumes similar ionization efficiency of SAAVs and reference peptides, we quantified allele-specific expression by PRM analysis³⁸. This was performed using a Q-Exactive-Plus hybrid quadrupole-orbitrap mass spectrometer (ThermoFisher Scientific), operated under Xcalibur software 3.0.63, equipped with a nanoelectrospray ion source and coupled to a nano-Acquity system (Waters). Tryptic peptides were loaded and elute off a self-packed 1.9 µm C₁₈ analytical column (New Objective, 35 cm × 75 µm inner diameter; 360 µm outer diameter). Elution was performed

over a 250-min gradient at a rate of 210 nl/min with buffer B ranging from 1% to 80% (buffer A: 0.1% formic acid with 5% DMSO in water, buffer B: 0.1 % formic with 5% DMSO in 80% acetonitrile). The Q-Exactive-Plus MS was operated with the PRM mode using an inclusion list, consisting of m/z of targeted peptides shown in Supplementary Figure 2, a normalized collision energy of 30 was employed for fragmentation. Typical MS conditions included: spray voltage, 2.6kV; heated capillary temperature, 350°C; isolation windows 1.6 m/z with an offset 0.5 m/z; the resolution set to 17,500 at m/z 200; automatic gain control (AGC) target at 2×10^5 . The heavy isotope labeled peptides were purchased from ThermoFisher Scientific (AQUA-Basic). The dynamic range was determined from lower end and upper end of calibration curve of synthesized peptides, which were spiked into tryptic digests of human brain. Heavy stable isotope labelled peptides were spiked into digested human brain samples (1µg/µl) in 100 fmol concentration. The linear range quantitation of heavy isotope labelled peptides is shown in Supplementary Figure 3.

Before importing Thermo raw files into Skyline software (Version 3.6)³⁹, a spectral library of heavy stable isotope labelled peptides was built. Heavy peptides were mixed in 100 fmol concentration and then analyzed by Q-Exactive-Plus mass spectrometer with the same normalized collision energy as PRM analysis described above. The raw file was processed by Maxquant (Version 1.5.7.4), using a human whole proteome sequence FASTA file was downloaded from UniProt (UP000005640, date: 2016-02-16), with variant protein sequences concatenated to the file. The MS/MS spectrum of each heavy peptide is shown in Supplementary Figure 2. For each peptide, the six strongest product ions were selected in Skyline. The Skyline software settings included: Precursor mass analyzer, Centroided; MS1 mass accuracy of 20 ppm; Product mass analyzer, Centroided; MS/MS Mass accuracy of 20 ppm; Retention time filtering, include all matching scans. All product ion peak areas were calculated by Skyline, and endogenous peptides quantification was based on the normalized peak areas of at least 4 product ions.

Results

In Silico Characterization of Personal Protein Database Generated from DNA Sequencing

Whole-exome sequencing was performed on genomic DNA extracted from fresh-frozen brain from two female donors who died at 50 and 76 years old, respectively. Each sample has >31.5 million bases with >8x coverage, which is more than sufficient to allow base calling, and the mean depths of coverage for the cases were 228X and 135X, respectively. On average, the cases had 28,351 single nucleotide variants (SNVs), including 6,700 intronic, 10,731 replacements (*i.e.*, sites that change the primary coding sequence), and 10,409 silent SNVs. Both cases had ~1:1 silent-to-replacement sites and a transition-to-transversion ratio of ~2.3 for the human whole exome, indicating good quality sequencing³³. A PPD was constructed using GenPro for each sample (Figure 1) using the variant base calls. Each PPD is a complete list of all full-length reference and variant proteins for the sample.

To understand the influence variant proteins in the PPDs had on the proteotypic peptide search space, we performed an *in silico* analysis of PPDs. Each PPDs had approximately 10,000 and 6,000 novel tryptic and LysC peptides, respectively (Figure 2A), which is

approximately 1% more theoretical fully tryptic peptides than the reference (*i.e.*, KnownGene track). Next, we compared the PPDs to UniProtKB, which contains both Swiss-Prot and TrEMBL protein sequences (Supplementary Table S1). We chose UniProtKB because it is one of the most popular reference protein databases used for bottom-up proteomics^{40–41} and its size and tryptic search space has been described⁴². Although UniProtKB is slightly larger than custom PPDs, with approximately 2% more fully tryptic peptides, they share about 91–95% peptides. Approximately, 5–9% of peptides were found to be unique to each database. Differences in the databases are not surprising since Swiss-Prot contains a greater number of alternative predicted proteoforms and because TrEMBL includes protein that are generated from translations of EMBL (European Molecular Biology Laboratory) nucleotide sequence entries. Together this increases the total number of protein entries of the UniProtKB^{40–41}. Although our findings for tryptic search space and average peptide lengths for reference peptides in the PPDs are comparable to UniProtKB⁴² (Supplementary Table S2), the theoretical variant peptides (6–40 residues in length) in the PPDs are significantly longer than reference peptides (independent sample t-test; p-value <0.01), which is a consequence of how GenPro databases are constructed. Namely, for a protein with a SAAV peptide to be included in the PPD, the SAAV peptide must be unique among all reference and variant peptides. Unique peptides also tend to be longer since the probability of a peptide being unique is proportional to length.

Comparing MS/MS Search Results from PPDs to the Human UniProtKB Reference Database

We generated a comprehensive number of human brain tryptic peptides following proteolytic digestion using two separate proteases (*i.e.*, Trypsin and LysC) followed by fractionation utilizing ERLIC (Supplementary Figure 4). In total, we analyzed the resulting 40 fractions per case by nanoLC-MS/MS on an Orbitrap Fusion mass spectrometer collecting precursor peptide ions (MS scans) at high resolution (120,000 FWHM) and fragment ions measurement generated by HCD at high resolution (17,500 FWHM). Table 1 shows the overall protein groups, proteins, total peptides, and peptide spectral matches (PSMs) identified for the UniProtKB and case-specific PPD searches, respectively. We had a similar number of target and decoy matches for the PPD and UniProtKB (Supplementary Table S3), indicating the same FDR for all searches of approximately 0.7%. Overall, for both trypsin and LysC digests, PPDs and UniProtKB performed similarly for the number of peptides and PSMs that were matched. We found slightly fewer protein groups, and dramatically fewer total proteins were identified in the PPDs, which reflects redundant proteoforms in UniProtKB compared to KnownGene. For the Emory brain bank case E05-90, we matched 1,324 and 859 variant PSMs for trypsin and LysC, respectively, which map to 553 and 426 unique peptides and 389 and 299 unique genes (Figure 2A). As expected, in a single individual, there was significant overlap between peptides identified by trypsin and LysC with 43–53% of peptides identified in both experiments (Figure 2B). A total of 977 unique variant peptides were identified across the two individual cases, of which 40–52% were identified in both samples (Figure 2C). Table 2 shows the overall protein groups, proteins, total peptides, and PSMs for the unique peptide matches. We found 4% of all gene products had evidence for a variant peptide per case. These variant peptides reflect approximately 0.5% of all peptides identified for either protease (LysC or trypsin) and

represent approximately 5–7% of the theoretical variant peptides in the PPD. By comparison we found evidence for approximately 17% of all possible theoretical reference peptides.

Next, we asked whether SAAV peptides alter the biochemical properties of peptides compared to reference peptides. This was done by assessing whether SAAV peptides are present in different ERLIC fractions than their reference counterparts. Variants that introduce/replace an ionizable or hydrophobic amino acid would be hypothesized to influence ERLIC elution times, which resolves peptides by hydrophilic interaction and electrostatic repulsion⁴³. Our findings show that reference and variant peptide variant that differed by just a single amino acid showed a strong correlation in the average ERLIC elution position ($R = 68\%$) (Supplementary Figure 5) indicating that most SAAV peptides do not dramatically alter the biochemical properties of the peptide.

We found that SAAV peptides may come from SNVs that are common or rare in the general population. We find that 35% of SAAV peptides come from SNVs with minor allele frequencies (MAF) from dbSNP below 5%, and 65% of SAAVs having MAF above 5% (Supplementary Figure 6). The distribution of XCorr values for reference and variant PSMs identified in the 2 cases (Figure 3) revealed that the average XCorr values for SAAV PSMs were not significantly different from those observed for reference PSMs (independent sample t-test; p-value 0.41–0.87) (Table S3). Of note, searches using the UniProtKB identified an average of 20 variant peptides per case (compared to the GRCh38 reference genome), demonstrating that some entries within the UniProtKB contain variants, which is not apparent to the end user.

Confirmation of Variant Peptides in Human Brain and Estimation of Allele-Specific Abundance by Label-Free Quantification

We confirmed single amino acid coding changes using synthetic peptides. The MS/MS spectra of six synthetic peptides were identical or very similar to those of SAAV peptides detected in human brain. Notably, rare and novel SAAV peptide sequences were also matched with high confidence. Figure 4 shows a representative example of a rare (MAF approximately 1%) missense variant that was identified at the peptide level in *HSPA12A*. This missense variant encodes an E365G (rs61753067) coding change. We identified the reference and variant peptides (IFGE₃₆₅DFIEQFK, and IFGG₃₆₅DFIEQFK, respectively) by LC-MS/MS. An E-to-G substitution results in a loss of 72 Da on the corresponding y-fragment ions, which was observed by MS/MS. The chemically synthesized IFGG₃₆₅DFIEQFK peptide was used confirm unambiguous MS/MS spectrum assignment of the variant peptide identified in brain. MS/MS spectra of additional selected synthetic SAAV peptides were also identical or very similar to those of SAAV peptides detected in human brain (Supplementary Figures 7 and 8). These data demonstrate we can confidently identify replacement SAAVs in human brain.

For diploid organisms, genes are generally expressed from both chromosomes in roughly equal proportion; however, this is not invariably the case⁹. Allele-specific expression influences both traits and disease, but systematic investigations of it are generally restricted to mRNA⁴⁴. To extend this to proteins, we identified 477 SAAV and reference peptide pairs. Among those we estimated allele-specific abundance for the 429 SAAV and reference

peptide pairs with identical charge states graphed those as an MA plot⁴⁵ (Figure 5A). We found that allele-specific abundance for reference to variant peptides did not significantly differ from the expected ratio of 1:1 for each either sample or when considering all pairs from both samples together (t-test, p-value 0.26–0.45). Collectively, the median ratio was 0.92 with 215 and 214 observations above and below the median, respectively. To control for biochemical differences that might have influenced these findings, we removed peptide pairs that changed a Met, Asn, or Gln to avoid the influence of common post-translational modifications on estimates of abundance and found comparable allelic balance (Supplementary Figure 9) which showed the same results. Interestingly, we found that approximately 22% (95/429) of the SAAV and reference pairs showed where one allele was expressed greater than 4-fold higher compared to the other. Of those, 40% (57/95) were showed higher SAAV peptide and 60% (38/95) showed higher reference peptide.

Parallel Reaction Monitoring (PRM) for Absolute Quantification of Allele-Specific Protein Abundance in Human Brain

We found SAAV and reference peptide pairs were generally in similar ERLIC fractions, suggesting that the biochemical properties did not dramatically differ between these peptide pairs. However, one drawback of shotgun approaches is variable peptide ionization and potential under-sampling of the peptide chromatographic peak in the MS1 scan, which together may increase the variance of our label-free quantitative measurements. Thus, we sought to confirm a subset of these SAAV to reference peptide ratios by parallel reaction monitoring (PRM). This targeted mass spectrometry approach permits the parallel detection of all target product ions in one high resolution mass analysis in the Orbitrap³⁸. Like selective reaction monitoring, PRM has the sensitivity that the data dependent proteomics approach lacks, and also offers the ability to determine absolute abundance when employing known quantities of isotopically labeled peptide standards³⁸. We chose three SAAV and reference peptide pairs (six peptides total) for PRM utilizing heavy isotope internal standards, (Figure 5B).

For more accurate absolute quantification of peptide levels we analyzed the total brain proteome using a single-shot LC PRM analysis without prior off-line fractionation. The MS/MS spectra and linear range quantitation of heavy isotope labelled peptides are shown in Supplementary Figures 2 and 3, respectively. Product ion pattern and chromatogram overlay for both the targeted endogenous brain derived (light) and heavy isotope labeled standards for Lysosomal alpha-glucosidase (GAA), Heat shock 70 kDa protein 12A (HSPA12A), and protein AHNK nucleoprotein 2 (AHNAK2) are shown in Figure 6. The absolute levels in femtomole (fmol) of reference and variant peptide pairs were determined in both samples from the standard curve (Figure 7 and Supplementary Figure 3). Variant peptides from GAA and HSPA12A were specific for E05-90 and, as expected, were only identified and quantified in this sample, but the reference peptide was identified in both samples (Figure 6 and 7). For GAA, 4.39 ± 0.66 fmol of reference and 4.36 ± 0.41 fmol variant peptide were measured in E05-90, which disagrees with the label-free approach in the ERLIC fractions that suggested a ratio above 1:2 (Figure 5B). By contrast, HSPA12A had nearly identical levels of the reference (2.24 ± 0.20 fmol) and variant (2.28 ± 0.14 fmol) consistent with the label-free peptide intensities (Figure 7, middle panel). Interestingly, the inter-sample level of

the HSPA12A reference peptide (IFGEDFIEQFK) varied between each case, which may be due to biological or tissue sampling differences. Finally, the SAAV peptide for AHNAK2 was shared by both cases (Figure 7). In E05-90, a higher level of reference (2.23 ± 0.10 fmol) compared to variant (1.03 ± 0.05 fmol) was consistent with >2 -fold less SAAV:reference peptide, which agrees with the label-free approach (Figure 5B). In case OS02-252, the reference peptide (0.62 ± 0.05 fmol) was slightly less abundant than the variant (0.99 ± 0.07 fmol) giving a ratio of 1.4 variant to reference that is consistent with the label-free finding suggesting a ratio of below a 2-fold difference (Figure 5B). We note that case E05-90, which showed imbalance of the variant:reference peptide ratio also had two SNVs in and around AHNAK2 that were not shared with OS02-252. Both were coding SNVs, rs12890949 (L2145V) and rs78014542 (S5185A), 3' to the shared SNV, rs61996045 (P1562L). These SNVs might explain differences in allele-specific protein abundance across the two individuals. In sum, the label-free and targeted PRM approaches support the hypothesis that most alleles are equally expressed in human brain.

Discussion

In this study, we developed GenPro to generate PPDs from next-generation sequencing and applied it to the study of allelic-specific protein abundance in human brain. *In silico* comparison of PPDs generated by GenPro showed a minimal increase ($\sim 1\%$) in the total number of theoretical tryptic peptides compared other databases, which demonstrated suitability to using our PPDs in standard search algorithms. To illustrate this point, we show that PPDs and UniProtKB searches perform similarly on whole-brain proteomes generated from two human subjects. Searching the sample-specific PPDs led us to identify 977 unique variant peptides (*i.e.*, SAAV peptides and peptides that alter proteolytic sites), and we identified 477 unique SAAV and corresponding reference peptide pairs that allowed us to estimation allele-specific expression at steady state in brain using peptide extracted ion intensities. Notably, SAAV and reference peptide pairs were found to have equal abundance in the brain, on average, which is expected for diploid genomes, and, similar to studies examining RNA expression⁹, we found 22% of the SAAV and reference peptides showed evidence for differential expression. Finally, we confirmed the allelic-specific abundance by targeted absolute quantification and found it was generally concordant with the label-free approach. These data support the hypothesis that most alleles are equally expressed in human brain.

This study lays the ground work for future investigation into allele-specific abundance at the protein level in disease-relevant tissue. Previously, investigations into allele-specific abundance generally focus on RNA, which is not always feasibly obtained from post-mortem human tissue and not highly correlated with protein abundance based²⁸. Mass spectrometry-based proteomics has been used to test this in a limited way for common SNVs of blood proteins⁴⁶, but more comprehensive approaches are needed, and, here, we present a pipeline for future investigations. Recent genetic studies have shown that loss-of-function (LoF) SNVs are remarkably common (~ 100 per genome), yet for SNVs that were predicted to cause nonsense mediated decay (a form of LoF) no discernable effect on the RNA level was detected⁴⁷. Allele-specific abundance illuminate our understanding of LoF SNVs on human disease. This has obvious implications for complex neurodegenerative

diseases, such as Alzheimer's disease (AD). For example, intronic SNVs in ABCA7 were found to associate with AD by genome-wide association studies, and, more recently, rare SNVs in ABCA7 that are predicted to cause LoF are enriched among AD cases versus controls⁴⁸.

Our approach to PPDs construction assumes each base called have the same probability of being correct by imposing probability threshold for every base called. Thus, *in silico* translation of an individual genome will yield peptides that equally probable of existing in the sample. By contrast, approaches that incorporate variants from publicly available repositories into a proteomic database without specific genetic knowledge of the sample being sequenced will include peptides with different levels of probability of being in the sample. In that scenario, a two-step approach to variant identification with a more stringent FDR for variant peptides seems reasonable¹². In fact, the FDR may need to incorporate the prior probability of the variant peptide existing in the sample. For dbSNP, this would simply be the minor allele frequency, which would dramatically lower the FDR threshold for variant peptides of very low frequency.

Our approach to finding sample-specific variant peptides is similar to those using RNA-seq to generate sample-specific proteome from immortalized mammalian cell lines^{22, 49}, or whole-exome sequencing coupled to RNA-seq in cancerous tissue²⁴. Indeed, our software would accept variant calls from generated from either DNA or RNA sequencing. However, for complex somatic tissue, there are several advantages to using whole-exome sequencing as the source of SNVs. Base calling is inherently more challenging with RNA-seq because the abundance of RNA molecules adds a complication to most base-calling procedures, and rare mRNA may not meet the minimum abundance threshold to call an SNV. Furthermore, peripheral sources of proteins are often found in complex tissue, *e.g.*, albumin and other serum-associated proteins are found in human brain⁵⁰. These proteins are not normally expressed in brain and would not be identified by RNA-seq from brain tissue. Exclusion of such proteins from a protein database could increase the mis-identification and/or quantification of the remaining proteins. For neurodegenerative diseases, proteins that accumulate extracellularly, which are highly disease relevant, may no longer have appreciable mRNA expression at the time the brain tissue is being studied. Finally, biofluids (*e.g.*, cerebrospinal fluid or blood) are generally composed of proteins from multiple organs, which makes whole-exome sequencing a more logical choice.

Our analysis of the 477 variant and corresponding reference peptide pairs shows the feasibility of using PPDs to estimate allele-specific protein abundance in complex human tissue, and using label-free quantification we showed that most SAAV and reference peptides are expressed in roughly equal amounts. This finding was then confirmed for a subset of SAAV and reference peptide pairs by absolute quantification using PRM. These results were consistent with previous reports that measured SAAV and reference peptide pairs in Jurkat cell lines and found them to be generally in balance²². Interestingly, 22% of SAAV and reference peptides pairs fell outside of the expected 1:1 ratio. One explanation for this is that SAAV peptides could be influencing protein structure and ultimately allele-specific protein levels. Alternatively, genetic variants in the 5' - or 3' -UTR could also influence RNA stability or translation and cause an allelic imbalance at the protein/peptide

level. Future work aimed at studying regions implicated in the genetics of AD and related neurodegenerative disease should measure allelic balance using targeted quantification to uncover SNVs implicated in altering allele-specific protein abundance, which may underlie some of the genetic risk for these diseases.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We gratefully acknowledge the generosity of the individual research volunteers and their families who made this work possible.

Funding Sources

This work was supported by the Veterans Health Administration (BX001820 to T.S.W.), National Institutes of Health (AG025688 and AG046161 to A.I.L.), Alzheimer's Association (NIRG12-242297 to N.T.S.), and the Emory Integrated Genomics Core (EIGC), which is subsidized by the Emory University School of Medicine and is one of the Emory Integrated Core Facilities. Mass spectrometry instrument time was subsidized by an Emory Neuroscience NINDS Core Facilities P30 grant (NS055077 to A.I.L.). N.T.S. is supported in part by an Alzheimer's Association, Alzheimer's Research UK, The Michael J. Fox Foundation for Parkinson's Research, and the Weston Brain Institute Biomarkers Across Neurodegenerative Diseases grant (11060). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Veterans Health Administration or the National Institutes of Health.

Abbreviations

SNV	single nucleotide variant
SAAV	single amino acid variant
UniProtKB	Universal Protein Knowledgebase
PPD	personalized protein database
VCF	variant call format

References

1. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. 2010; 464(7289):768–72. [PubMed: 20220758]
2. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M, Price A, Raj T, Nisbett J, Nica AC, Beazley C, Durbin R, Deloukas P, Dermitzakis ET. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet*. 2012; 8(4):e1002639. [PubMed: 22532805]
3. Wu L, Snyder M. Impact of allele-specific peptides in proteome quantification. *Proteomics Clin Appl*. 2015; 9(3–4):432–6. [PubMed: 25676416]
4. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M, Hubner N, Aitman TJ. Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet*. 2006; 2(10):e172. [PubMed: 17054398]
5. Knight JC. Approaches for establishing the function of regulatory genetic variants involved in disease. *Genome Med*. 2014; 6(10):92. [PubMed: 25473428]

6. Albert FW, Treusch S, Shockley AH, Bloom JS, Kruglyak L. Genetics of single-cell protein abundance variation in large yeast populations. *Nature*. 2014; 506(7489):494–7. [PubMed: 24402228]
7. Battle A, Khan Z, Wang SH, Mitran A, Ford MJ, Pritchard JK, Gilad Y. Genomic variation Impact of regulatory variation from RNA to protein. *Science*. 2015; 347(6222):664–7. [PubMed: 25657249]
8. Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, Raghupathy N, Svenson KL, Churchill GA, Gygi SP. Defining the consequences of genetic variation on a proteome-wide scale. *Nature*. 2016; 534(7608):500–5. [PubMed: 27309819]
9. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet*. 2010; 11(8):533–8. [PubMed: 20567245]
10. Ross CA, Poirier MA. Protein aggregation and neurodegenerative disease. *Nature medicine*. 2004; 10(Suppl):S10–7.
11. Mann M, Kulak Nils A, Nagaraj N, Cox J. The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. *Molecular Cell*. 49(4):583–590. [PubMed: 23438854]
12. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Methods*. 2014; 11(11):1114–25. [PubMed: 25357241]
13. Sadygov RG, Cociorva D, Yates JR. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nat Meth*. 2004; 1(3):195–202.
14. Smith LM, Kelleher NL, The Consortium for Top Down, P. Proteoform: a single term describing protein complexity. *Nature methods*. 2013; 10(3):186–187. [PubMed: 23443629]
15. Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ. Novel gene gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biology*. 2006; 7(4):R35–R35. [PubMed: 16646984]
16. Risk BA, Spitzer WJ, Giddings MC. Peppy: proteogenomic search software. *J Proteome Res*. 2013; 12(6):3019–25. [PubMed: 23614390]
17. Zickmann F, Renard BY. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. *Bioinformatics*. 2015; 31(12):i106–15. [PubMed: 26072472]
18. Blakeley P, Overton IM, Hubbard SJ. Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. *Journal of Proteome Research*. 2012; 11(11): 5221–5234. [PubMed: 23025403]
19. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*. 2001; 29(1):308–311. [PubMed: 11125122]
20. Li J, Duncan DT, Zhang B. CanProVar: A Human Cancer Proteome Variation Database. *Human mutation*. 2010; 31(3):219–228. [PubMed: 20052754]
21. The Cancer Genome Atlas Research, N. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013; 45(10):1113–1120. [PubMed: 24071849]
22. Sheynkman GM, Shortreed MR, Frey BL, Scalf M, Smith LM. Large-scale mass spectrometric detection of variant peptides resulting from non-synonymous nucleotide differences. *Journal of proteome research*. 2014; 13(1):228–240. [PubMed: 24175627]
23. Wang X, Slebos RJC, Wang D, Halvey PJ, Tabb DL, Liebler DC, Zhang B. Protein identification using customized protein sequence databases derived from RNA-Seq data. *Journal of proteome research*. 2012; 11(2):1009–1017. [PubMed: 22103967]
24. Park H, Bae J, Kim H, Kim S, Kim H, Mun DG, Joh Y, Lee W, Chae S, Lee S, Kim HK, Hwang D, Lee SW, Paek E. Compact variant-rich customized sequence database and a fast and sensitive database search for efficient proteogenomic analyses. *Proteomics*. 2014; 14(23–24):2742–9. [PubMed: 25316439]
25. 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):2341–53. [PubMed: 23629695]
26. Koch A, Gawron D, Steyaert S, Nda E, Crappé J, De Keulenaer S, De Meester E, Ma M, Shen B, Gevaert K, Van Crielinge W, Van Damme P, Menschaert G. A proteogenomics approach

integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics*. 2014; 14(0):2688–2698. [PubMed: 25156699]

27. Oszolák F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nature reviews Genetics*. 2011; 12(2):87–98.
28. Seyfried NT, Dammer EB, Swarup V, Nandakumar D, Duong DM, Yin L, Deng Q, Nguyen T, Hales CM, Wingo T, Glass J, Gearing M, Thambisetty M, Troncoso JC, Geschwind DH, Lah JJ, Levey AI. A Multi-network Approach Identifies Protein-Specific Co-expression in Asymptomatic and Symptomatic Alzheimer's Disease. *Cell Systems*. 4(1):60–72.e4.
29. Chisanga D, Keerthikumar S, Pathan M, Ariyaratne D, Kalra H, Boukouris S, Mathew NA, Al Saffar H, Gangoda L, Ang CS, Sieber OM, Mariadason JM, Dasgupta R, Chilamkurti N, Mathivanan S. Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues. *Nucleic Acids Res*. 2016; 44(D1):D969–74. [PubMed: 26496946]
30. Keerthikumar S, Gangoda L, Liem M, Fonseka P, Atukorala I, Ozcitti C, Mechler A, Adda CG, Ang CS, Mathivanan S. Proteogenomic analysis reveals exosomes are more oncogenic than ectosomes. *Oncotarget*. 2015; 6(17):15375–96. [PubMed: 25944692]
31. Hao P, Guo T, Li X, Adav SS, Yang J, Wei M, Sze SK. Novel Application of Electrostatic Repulsion-Hydrophilic Interaction Chromatography (ERLIC) in Shotgun Proteomics: Comprehensive Profiling of Rat Kidney Proteome. *Journal of Proteome Research*. 2010; 9(7):3520–3526. [PubMed: 20450224]
32. Hao P, Ren Y, Dutta B, Sze SK. Comparative evaluation of electrostatic repulsion–hydrophilic interaction chromatography (ERLIC) and high-pH reversed phase (Hp-RP) chromatography in profiling of rat kidney proteome. *Journal of Proteomics*. 2013; 82:254–262. [PubMed: 23486160]
33. Johnston HR, Chopra P, Wingo TS, Patel V, International Consortium on, B; Behavior in 22q11.2 Deletion, S. Epstein MP, Mülle JG, Warren ST, Zwick ME, Cutler DJ. PEMapper and PECaller provide a simplified approach to whole-genome sequencing. *Proc Natl Acad Sci U S A*. 2017; 114(10):E1923–E1932. [PubMed: 28223510]
34. The Variant Call Format Specification. <http://samtools.github.io/hts-specs/VCFv4.3.pdf> (accessed 2017-01-27)
35. 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:D670–81. Database issue. [PubMed: 25428374]
36. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Meth*. 2007; 4(11):923–925.
37. R: A Language and Environment for Statistical Computing, v3.2.3. R Foundation for Statistical Computing; Vienna, Austria: 2016.
38. Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ. Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Molecular & Cellular Proteomics: MCP*. 2012; 11(11):1475–1488. [PubMed: 22865924]
39. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*. 2010; 26(7):966–968. [PubMed: 20147306]
40. UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res*. 2015; 43:D204–12. Database issue. [PubMed: 25348405]
41. The UniProt, C. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research*. 2011; 39(Database issue):D214–D219. [PubMed: 21051339]
42. Alpi E, Griss J, da Silva AW, Bely B, Antunes R, Zellner H, Rios D, O'Donovan C, Vizcaino JA, Martin MJ. Analysis of the tryptic search space in UniProt databases. *Proteomics*. 2015; 15(1):48–57. [PubMed: 25307260]

43. Alpert AJ. Electrostatic Repulsion Hydrophilic Interaction Chromatography for Isocratic Separation of Charged Solutes and Selective Isolation of Phosphopeptides. *Analytical Chemistry*. 2008; 80(1):62–76. [PubMed: 18027909]
44. Buckland PR. Allele-specific gene expression differences in humans. *Hum Mol Genet*. 2004; 13:R255–60. Spec No 2. [PubMed: 15358732]
45. Peng S, Alekseyenko AA, Larschan E, Kuroda MI, Park PJ. Normalization and experimental design for ChIP-chip data. *BMC Bioinformatics*. 2007; 8:219. [PubMed: 17592629]
46. Su ZD, Sun L, Yu DX, Li RX, Li HX, Yu ZJ, Sheng QH, Lin X, Zeng R, Wu JR. Quantitative detection of single amino acid polymorphisms by targeted proteomics. *J Mol Cell Biol*. 2011; 3(5): 309–15. [PubMed: 22028381]
47. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IH, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Genomes Project, C. Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. 2012; 335(6070):823–8. [PubMed: 22344438]
48. Steinberg S, Stefansson H, Jonsson T, Johannsdottir H, Ingason A, Helgason H, Sulem P, Magnusson OT, Gudjonsson SA, Unnsteinsdottir U, Kong A, Helisalmi S, Soininen H, Lah JJ, DemGene. Aarsland D, Fladby T, Ulstein ID, Djurovic S, Sando SB, White LR, Knudsen GP, Westlye LT, Selbaek G, Giegling I, Hampel H, Hiltunen M, Levey AI, Andreassen OA, Rujescu D, Jonsson PV, Bjornsson S, Snaedal J, Stefansson K. Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. *Nat Genet*. 2015; 47(5):445–7. [PubMed: 25807283]
49. Krug K, Popic S, Carpy A, Taumer C, Macek B. Construction and assessment of individualized proteogenomic databases for large-scale analysis of nonsynonymous single nucleotide variants. *Proteomics*. 2014; 14(23–24):2699–708. [PubMed: 25251379]
50. Liu HM, Atack JR, Rapoport SI. Immunohistochemical localization of intracellular plasma proteins in the human central nervous system. *Acta Neuropathologica*. 1989; 78(1):16–21. [PubMed: 2735186]

Figure 1. Personal protein database creation using genomic sequencing

This flow diagram provides the general scheme for generating a personal protein database. For each missense variant, GenPro creates all variant proteins by in-frame *in silico* translation from genomic DNA. The resultant variant protein is digested with the specified protease. Peptides that pass a user-specified length-exclusion threshold are compared to the existing collection of peptides, which are initially populated by a complete *in silico* digestion of all reference proteins. If the variant protein results in a new SAAV peptide then then the variant protein entry is saved to the database.

A. Overview of Variant Peptide Identification in 1 Sample

B. Variant Peptides Identified in 1 Sample

C. Variant Peptides Identified In 2 Samples

Figure 2. Identification of variant peptides from personal proteomic databases

A. General overview of exome sequencing. About 31.5 million bases per individual were sequenced, and an average of 28,351 SNVs per sample were found; of those, an average of 10,731 were replacement sites (*i.e.*, missense). Considering two commonly used enzymes (trypsin and LysC), we show the number of theoretical peptides generated, total number of PSMs, unique peptides, and proteins/genes identified for a single case. B. Venn diagram of unique peptides identified from 1 sample using trypsin and LysC. C. Venn diagram of unique peptides identified in both samples.

XCorr Distribution for Reference and Variant PSMs

Figure 3. XCorr distribution for reference and variant peptide spectral matches

Panels A and B give the distribution of XCorr values of peptide spectral matches (PSMs) identified for each sample, respectively. Trypsin is shown in black, LysC in blue. The solid line represents PSMs that correspond to reference peptides, and the broken line represents PSMs that correspond to variant peptides. There is no significant difference between XCorr values for PSMs that correspond to either reference or variant peptides by t-test.

Representative Mass Spectra for Variant Containing Peptide

Figure 4. Confirmation of a HSPA12A missense variant (rs61753067) in human brain

A MS/MS spectrum corresponding to the doubly protonated reference peptide IFGE₃₆₅DFIEQFK (top) and the doubly protonated HSPA12A variant-specific peptide, IFGG₃₆₅DFIEQFK, sequenced from brain (middle). The single amino acid (E₃₆₅G) coding change is highlighted as green to red. A synthetic standard peptide was used to confirm the coding change (bottom panel).

Figure 5. Allele-specific abundance of reference and variant pairs

(A) We identified 477 SAAV and reference peptide pairs in both samples from two separate digests (Trypsin and LysC). The sum of the log₂ of the signal intensities plotted against the log₂ of the ratio of those intensities for each sample, and the red broken line shows the 2-fold threshold. The average log intensity ratio does not significantly differ from zero, which reflects most peptides are in an approximately 1:1 ratio (*i.e.*, in allelic balance). (B) The reference and SAAV peptide pairs are labeled for each sample that underwent further validation and quantification using PRM analysis.

Figure 6. Absolute quantification of reference and variant peptides in human brain by parallel reaction monitoring

(A) Product ion pattern and chromatograph of endogenous and heavy isotope labeled reference peptide (APSPLYSEFSEEPFGVIVR) and the variant peptide (APSPLYSEFSEEPFGLIVR) for the Lysosomal alpha-glucosidase (GAA) protein from human brain samples. (B) Product ion pattern and chromatograph of endogenous and heavy isotope labeled peptide (IFGDFIEQFK) and the variant peptide (IFGDFIEQFK) for the Heat shock 70 kDa protein 12A (HSPA12A) from human brain samples. (C) Product ion pattern and chromatograph of endogenous and heavy isotope labeled reference peptide (LPEGPVPEGAGLK) and variant peptide (LSEGPVPEGAGLK) for the protein AHNK2 (AHNK2). For all reference and variant peptides, the dot-product values were shown to indicate the similarity of product ion pattern with spectra library. The underlined amino acid is the position of the coding change.

Figure 7. Absolute quantification of reference and variant peptides in human brain samples
The absolute levels (fmol) of reference and variant peptide pairs for GAA, HSPA12A and AHNAK2 were determined in both samples (E05-90 and OS02-252) from the standard curve (Supplementary Figure 3). The variant for GAA and HSPA12A were specific to sample 1 (E05-90), whereas the AHNAK2 variant peptides was common across both samples. The upper and lower limits of the whisker plots represent the standard deviation of product ion measurements. Quantification by PRM was based on the peak area from least four product ions.

Table 1
Average summary data for human brain proteome searched with UniProtKB and PPD.

Sample	Enzyme	Protein Group	Proteins		Total Peptides		PSMs	
			UniProtKB	PPD	UniProtKB	PPD	UniProtKB	PPD
E05-90	Trypsin		9,582	9,555	129,489	129,926	451,947	453,732
	LysC		8,239	8,203	83,747	83,976	242,650	243,003
OS02-252	Trypsin		9,038	9,009	104,343	104,458	342,861	342,923
	LysC		7,832	7,797	76,543	76,676	215,126	214,823

Table 2

Summary data for non-reference peptides.

Sample	Enzyme	Protein Group	Proteins	Total Peptides	PSMs
E05-90	Trypsin	389	404	553	1,324
	LysC	299	311	426	859
OS02-252	Trypsin	313	326	434	1,031
	LysC	257	266	373	774