De novo identification and quantification of single amino-acid variants in human brain

Zhi-Duan Su\textsuperscript{1,\ast}, Quan-Hu Sheng\textsuperscript{1,\dagger}, Qing-Run Li\textsuperscript{1}, Hao Chi\textsuperscript{2}, Xi Jiang\textsuperscript{3}, Zheng Yan\textsuperscript{3}, Ning Fu\textsuperscript{1}, Si-Min He\textsuperscript{2}, Philipp Khaitovich\textsuperscript{3}, Jia-Rui Wu\textsuperscript{1,\ast}, and Rong Zeng\textsuperscript{1,\ast}

\textsuperscript{1}Key Laboratory of Systems Biology, Chinese Academy of Sciences, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Shanghai 200031, China
\textsuperscript{2}Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
\textsuperscript{3}Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai 200031, China
\dagger These authors contributed equally to this work.
\ast Correspondence to: Jia-Rui Wu, E-mail: wujr@sibs.ac.cn; Rong Zeng, E-mail: zr@sibs.ac.cn

The detection of single amino-acid variants (SAVs) usually depends on single-nucleotide polymorphisms (SNPs) database. Here, we describe a novel method that discovers SAVs at proteome level independent of SNPs data. Using mass spectrometry-based \textit{de novo} sequencing algorithm, peptide-candidates are identified and compared with theoretical protein database to generate SAVs under pairing strategy, which is followed by database re-searching to control false discovery rate. In human brain tissues, we can confidently identify known and novel protein variants with diverse origins. Combined with DNA/RNA sequencing, we verify SAVs derived from DNA mutations, RNA alternative splicing, and unknown post-transcriptional mechanisms. Furthermore, quantitative analysis in human brain tissues reveals several tissue-specific differential expressions of SAVs. This approach provides a novel access to high-throughput detection of protein variants, which may offer the potential for clinical biomarker discovery and mechanistic research.

Keywords: single amino-acid variants (SAVs), \textit{de novo}, proteomics, human brain tissues

Introduction

Genome-wide association studies (GWAS) and other genomic approaches are widely used to identify single-nucleotide polymorphisms (SNPs) that are associated with particular physiological or pathological traits in individuals (Yoshiura et al., 2006; Boyle et al., 2010; Boyle et al., 2012; Zhang et al., 2012b). Nowadays, identification of single amino-acid variants (SAVs) is generally based on databases including SAVs of certain proteins derived from SNPs information (Yip et al., 2008; Li et al., 2011; Su et al., 2011). However, those databases are restricted when protein sequence variations are independent of SNPs. Emerging evidences have indicated that there are SAVs arising from RNA editing at transcription level (Wulff et al., 2011; Bahn et al., 2012; Jepson et al., 2012; Park et al., 2012). In addition, aminoacyl-tRNA synthetase-induced translational infidelity (Jordanova et al., 2006; Schepet et al., 2007) and some other undiscovered mechanisms at post-transcriptional level can also lead to protein variants inconsistent with either genome or transcriptome. Therefore, detecting SAVs directly at proteome level is important to reveal protein variations that not only confirm the SNPs and RNA variants, but also increase new information on diversity of protein sequences and functions (Wu and Zeng, 2012).

Mass spectrometry-based proteomic methods are able to identify thousands of proteins in a single experiment (Domon and Aebersold, 2006; Kiyonami et al., 2011; Thakur et al., 2011; Geiger et al., 2012). Nevertheless, mainstream proteomic pipelines rely on database-searching and proteins cannot be identified if they are absent from database, such as protein variations with SAVs. As the development of mass spectrometry technology, especially the appearance of high-precision scan modes (Olsen et al., 2007), \textit{de novo} identification of database-excluded peptides is increasingly feasible. But evaluation and confirmation of new peptides are still challenging in this unrestricted strategy. Therefore, the application of \textit{de novo} sequencing has been limited.

Here, we developed a novel method for \textit{de novo} identification of SAVs at proteome level. We first made use of \textit{de novo} sequencing, which was unrestricted and sensitive, to identify peptide-candidates. After that, we compared these \textit{de novo} results with theoretical protein database to determine the identification only as SAVs and their corresponding database-included peptides, termed SAV-pairs. Eventually, we searched all MS/MS spectra against a SAV-database combining \textit{de novo}-identified SAVs with original protein sequences, and took advantage of target-decoy strategy (Elias and Gygi, 2007) to control false discovery rate (FDR).
Results

Pipeline description of de novo-based SAVs identification

This workflow consisted of the following steps (Figure 1A): (i) All tandem mass spectra were acquired via higher-energy C-trap dissociation (HCD) mode on Orbitrap mass spectrometry. High-accuracy measurements of both peptide-precursor ions and their fragment ions as well as nearly complete ion series owing to large detected mass range facilitated the precision of de novo sequencing. Next, we used pNovo (Chi et al., 2010), a de novo peptide sequencing tool, to sequence all HCD MS/MS spectra, and only peptide-candidates with C-score resulting from pNovo higher than 0.65 were kept. Furthermore, the remaining candidates were strictly filtered via...
whether they were in accordance with experimental proteolytic forms. In our assays, we used trypsin to digest proteins and therefore peptides whose C-terminals were neither K nor R were discarded. (ii) A human UniProt protein database (complete proteome, 2012-01-25) was theoretically digested in silico and compared with the peptide-candidates from de novo results by pNovoValidator (software in ProteomicsTools in house). The peptides found in the database were defined as database-included peptides. According to the pairing strategy, only database-excluded de novo peptide-candidates with single amino-acid substitution for database-included peptides were considered as SAVs and reserved for further analysis. (iii) By incorporating all the above SAVs into this database, we established a SAV-database, in which each genuine sequence was followed by a reversed sequence again. Here, each SAV-sequence was treated as an independent protein. Then, all the HCD MS/MS spectra were searched against this SAV-database by Mascot. The search results were filtered by 1% peptide/protein FDR following target-decoy strategy, and SAVs and their corresponding database-included peptide were reserved as SAV-pairs only if both peptides were identified again by Mascot. (iv) Optionally, three more strict criteria might be used to decrease false discovery: (a) whether eliminating N-terminal substitutions; (b) whether eliminating deamidated substitutions, such as N→D or Q→E; and (c) whether keeping SNPs-derived SAVs only.

Identification of SAVs in synthetic peptide-mixtures

Without assistance of an available database, de novo peptide sequencing extracts a peptide sequence directly from a spectrum (Ma and Johnson, 2012). Therefore, one potential advantage is to discovery both known and novel variations at proteome level, which is also independent of SNPs information at genome level. With the advancement of mass spectrometry technology and appearance of novel computational methods, de novo algorithm has been greatly improved. However, it is still not comparable with database-searching for common protein identification and needs manual checks by proteomics experts, which is time consuming and of low-throughput (Pevtsov et al., 2006; Kim et al., 2009a). In this study, we exploited the predominance of de novo peptide sequencing to identify SAV-peptides, and used mature database-searching strategy to monitor false discovery.

To evaluate the sensitivity and precision of this method, we synthesized 39 pairs of SAV-peptides with stable isotope labeling (Supplementary Table S1). Pooling them into a mixture, we analyzed it on a Q-Exactive mass spectrometry (Thermo Scientific) via HCD mode. First of all, we constructed a SAV-database, which combined these 39 SAVs with a human UniProt database and contained corresponding reversed sequences of each entry. By searching against this database, we identified 32 out of 39 SAV-pairs. We used this procedure to simulate database-searching. On the other hand, from the same 37094 MS/MS spectra we collected 791 SAV candidates by de novo identification and pairing strategy screening. Finally, we could also detect 30 SAV-pairs (Figure 1B, Supplementary Tables S2 and S3). This result indicated that the sensitivity of our method was comparable with database-searching.

Besides these targeted detections, we also identified 124 non-targeted SAVs, all of which could be paired with one of our synthetic database-included peptides (Figure 1C, Supplementary Table S3). To further evaluate these identifications, we used PEAKS 6, another de novo sequencing assisted database-searching software, and Andromeda (integrated into MaxQuant v1.3.0.5) to deal with these raw data. We made further efforts to detect each synthetic SAV-peptide by MALDI-TOF mass spectrometry. As a result, 120 out of 124 non-targeted SAV-peptides were also identified (Figure 1D, Supplementary Table S3). Taken together, only four non-targeted SAV-peptides were unconfirmed, and we considered them as false positive discoveries. The other 120 non-targeted SAV-peptides very likely resulted from impurities of synthesized peptides, and all of them together with 30 targeted SAVs were regarded as true positive discoveries. Finally, the FDR of SAVs detection was 2.67%.

SAVs identification in human brain tissues

Next, we extended the application of this method to real complex samples. Proteins extracted from pairs of normal prefrontal cortex (PFC) and cerebellums (CB) of three human individuals were digested. Both label-free assay and TMT-labeling assay were performed. After separation, all fractions were identified on an Orbitrap-Velos mass spectrometry (Thermo Scientific) via HCD mode. From 1024097 MS/MS spectra, we successfully received 48436 SAV-candidate entries. And by database-searching validation, we totally identified 248 SAV-pairs, of which 169 and 101 SAV-pairs were from label-free assay and TMT-assay, respectively (Figure 2, Supplementary Figure S1, Tables S2, S4, and S5). They would be annotated as known SAVs if they were marked as ‘Natural Variant’ or ‘Sequence Conflict’ in UniProtDB with identical amino-acid substitutions. As a result, 55 and 7 known SAVs were separately marked.

GO-annotation results suggested that SAVs were significantly enriched in structural molecule activity of molecular function item (Supplementary Figure S2A). In biological process item, there were three categories enriched, i.e. cellular component organization, developmental process as well as cell adhesion (Supplementary Figure S2B). Moreover, we found proteins located in plasma membrane varied with high frequency (Supplementary Figure S2C), which might be related to various functions and signaling pathways of brain tissues and heterogeneity of individuals.

Identification of SAVs derived from DNA mutations

SNPs were recognized as the most common type of genetic variants in the human genome, and non-synonymous SNPs could give rise to amino-acid substitutions at protein level. At present, a number of SNP-databases are available. However, none of them could comprise whole SNPs information that was updated every day. Our method was independent of SNPs; thus we could identify both known and novel SAVs.

Succinate-semialdehyde dehydrogenase (ALDH5A1) catalyzed one step in the degradation of the inhibitory neurotransmitter gamma-aminobutyric acid, whose defects might cause succinate semialdehyde dehydrogenase deficiency (Kim et al., 2009b). It was reported that several variations on this protein were tightly associated with its activity (Akaboshi et al., 2003), including
ALDH5A1_H180Y that could result in loss of 17% activity, which we also identified. In our results, ALDH5A1_180H and ALDH5A1_180Y were identified in both CB1 and PFC1, which was considered as a heterozygous subject, while only ALDH5A1_180H was identified in both CB2 and PFC2, which was considered as a homozygous individual. Curiously, both ALDH5A1_180H and ALDH5A1_180Y were in PFC3, but only ALDH5A1_180H was in CB3. Our DNA sequencing results were completely consistent with SAV identifications in these three pairs of tissues (Figure 3, Supplementary Table S6).

Hydroxyacyl-CoA dehydrogenase (HADH) played an essential role in the mitochondrial β-oxidation of short chain fatty acids, and some mutations could lead to HADH deficiency. Here we identified that Pro on 215th site was changed to Thr, which was not annotated in UniProtDB. It was recorded as rs140413151 in another SNP database, NCBI, although little previous attention was paid to it and there was no validation information. Compared with DNA sequencing results, there was no false positive discovery but several false negative discoveries (Figure 3, Supplementary Table S6), probably due to bias of mass spectrometry toward high abundance. Next, we used selected reaction monitoring (SRM) (Picotti et al., 2010; Jovanovic et al., 2012), a more sensitive targeted proteomics method than shotgun proteomics, to detect both HADH_215P and HADH_215T in all these six samples, and our finding was excellently complementary to false negative discoveries of shotgun SAVs identification (Supplementary Figure S3).

Retinol-binding protein 1 (RBP1) was a universal protein that was detected nearly in all the tissues and functioned as intracellular transport of retinol. In this study, we identified RBP1_50M in CB1, PFC1, CB3, and PFC3 as homozygous ones, and both RBP1_50M and RBP1_50V in CB2 as heterozygote, all of which were validated by DNA sequencing (Figure 3, Supplementary Table S6). Notably, one missing detection of RBP1_50M in PFC2 was made up by SRM results (Supplementary Figure S4). Moreover, these three SAV-pairs were synthesized by stable isotope-labeling and their MS/MS patterns were very similar with endogenous ones (Supplementary Figures S5—S7). In conclusion, our de novo-based method could provide a robust access to high-throughput accurate identification of known and novel SAVs in real complex biological samples.
Identification of SAVs derived from RNA variations

Except the SNPs-derived protein variations, a proportion of SAVs were irrelevant to SNPs. As an important origin, variations derived from RNA level, such as RNA editing (Ekdahl et al., 2012) and alternative splicing (Sommer et al., 1990), might also result in amino-acid variations after translation.

Glutamate receptor 1 (GRIA1) was widely expressed in brain and acted as Ionotropic receptor to bind L-glutamate, which was an excitatory neurotransmitter at many synapses in the central nervous system. Here, we found that Ser of its 768th site was substituted for Asn, which was annotated in UniProt as alternative sequence. Next, we designed pairs of alternative exon-specific primers and used PCR assay to confirm this variation origin. We found that both Flop and Flip exons, existed at DNA level in all tissues. At RNA level, Flop was expressed in two PFC and two CB tissues, while Flip was tissue specifically expressed in three CB tissues, which was exactly consistent with RNA sequencing results (Figure 4, Supplementary Table S6). Although mass spectrometric discovery was difficult for plasma membrane proteins, we successfully identified GRIA1_768S in two CB tissues. This SAV-pair was validated via MS/MS-spectrum comparison between endogenous one and stable isotope-labeling synthetic one (Supplementary Figure S8). Hence, this workflow could efficiently discover, although not much in this assay, protein variants that might result from RNA level. Due to their low abundance, stronger separation methods would be needed to increase detection sensitivity.

Identification of novel SAVs originating from unknown post-transcriptional mechanisms

Apart from the above protein variants originating from either DNA mutations or RNA variations, several SAVs might happen during translational or some other unknown post-transcriptional processes. For example, infidelity of aminoacyl-tRNA synthetase would lead to protein variants that especially dramatically affected
phenotypes in nervous system (Scheper et al., 2007). Together with DNA/RNA sequencing results, we also identified some SAVs independent of DNA/RNA variations.

Cell division control protein 42 homolog (CDC42) was a plasma membrane-associated small GTPase, which played multiple roles in a variety of cellular processes (Serebriiskii et al., 2002; Oceguera-Yanez et al., 2005; Melendez et al., 2011). Combining shotgun SAVs identification and SRM SAVs validation, we found both CDC42_163K and CDC42_163R in all six brain tissues, which were also annotated in UniProt as alternative sequence (Figure 5, Supplementary Figure S9). Interestingly, we only detected one isoform, CDC42_163R (codon AGA), at corresponding DNA or RNA sites (Figure 5, Supplementary Table S6). This SAV-pair was also validated via MS/MS-spectrum comparison (Supplementary Figure S10). Similarly, a few other SAVs were also verified, even though no DNA/RNA mutations were detected before (Supplementary Figures S11–S13). In short, this unusual finding demonstrated that our de novo-based method was powerful to identify protein variants arising from unknown post-transcriptional mechanisms with high confidence (Table 1).

Measurement of SAVs in homozygote/heterozygote
For human and other diploid-genome organisms, SNPs/SAVs implied two possible genetic structures in a single allele: either homozygous state or heterozygous state (Tewhey et al., 2011).
Previously, some work had reported that different genetic structures of coding SNPs were tightly associated with various traits of individuals (Kubo et al., 2007). To address this issue, we selected a few SAV-peptides from above identification results to synthesize corresponding heavy ones with stable isotope-labeling, and measured their relative abundance in all six brain tissues using SRM strategy. HADH_P215T was only detected in the first human subject, but not in the other two (Figure 3). SRM results revealed that both HADH_P215 and HADH_P215T were expressed much higher in CB than in PFC in a heterozygous individual, while in the other two homozygous individuals, only HADH_P was detected at a lower expression level in CB tissues (Figure 6A and B). For RBP1_M50V, the relative expression pattern of RBP1_M50V was similar in all three individuals, which was higher in CB than in PFC, although it was less significant in heterozygous individual 2 (Figure 6C). Analogously, RBP1_M50V was also expressed more in heterozygous CB2 than in PFC2 (Figure 6D).

Table 1 Summary of validated SAVs.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Site</th>
<th>SAVs</th>
<th>DNA-mutation</th>
<th>RNA-variant</th>
<th>SAVs origins</th>
<th>Probable mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALDH5A1</td>
<td>180</td>
<td>H180Y</td>
<td>Y</td>
<td>N.A.</td>
<td>Genome-level</td>
<td>SNPs</td>
</tr>
<tr>
<td>HADH</td>
<td>215</td>
<td>P215T</td>
<td>Y</td>
<td>N.A.</td>
<td>Genome-level</td>
<td>SNPs</td>
</tr>
<tr>
<td>RBP1</td>
<td>50</td>
<td>M50V</td>
<td>Y</td>
<td>N.A.</td>
<td>Genome-level</td>
<td>SNPs</td>
</tr>
<tr>
<td>GRIA1</td>
<td>768</td>
<td>N768S</td>
<td>Y</td>
<td>N.D.</td>
<td>Transcriptome-level</td>
<td>Alternative splicing</td>
</tr>
<tr>
<td>CDC42</td>
<td>163</td>
<td>K163R</td>
<td>N.D.</td>
<td>N.D.</td>
<td>Proteome-level</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

Y: corresponding sites were detected; N.D: corresponding sites were not detected; N.A.: corresponding sites were not sequenced.
From these cases, we concluded that protein variants could bring about differential expressions between homozygote and heterozygote, which might be related to their underlining effects.

**Tissue-specific differential expressions of SAVs**

Based on SNPs frequency in a given population, GWAS and some other traditional SNPs work were mainly focused on qualitative analysis of associations between genotype and particular physiological or pathological traits (Jin et al., 2012; Deloukas et al., 2013). Equally important, SAVs studies could offer more quantitative information. Quantitative analysis of SAVs at proteome level was also critical to understand the associations of molecular variants with diverse characteristics, particularly in heterozygous state (Su et al., 2011). From the above consideration, we combined TMT-labeling approach and applied this method to the same human PFC and paired CB samples to profile several tissue-specific differential expressions of SAVs (Supplementary Figure S14). Totally, we quantified 74 SAV-pairs in all six tissues (Supplementary Table S4).

As we all know, microtubule-associated protein 1B (MAP1B) could facilitate tyrosination of α-tubulin in neuronal microtubules and have effect on cytoskeletal changes. In this work, we identified a novel SAV-site, MAP1B_A2335E, and measured its relative abundance. In all three individuals, MAP1B_2335A did not present any significant difference (Figure 7). In contrast, MAP1B_2335E was dramatically higher in PFC of two human subjects and 1.55-fold more in the remaining one. We also calculated this protein expression ratios between PFC and paired CB, all of which were less than 2-fold changed. As mentioned above, it was not surprising that our workflow was able to uncover differential expressions of SAVs among distinct samples after integrating in vitro labeling or spike-in quantitative strategies.

**Discussion**

Due to the large proportion of noncoding region in genome and degeneracy of genetic codes, section of SNPs information at genome level was redundant. Also, not all amino-acid substitutions resulted from SNPs, because of RNA alternative splicing and some other unknown post-transcriptional mechanisms. Therefore, it made sense to detect SAVs at proteome level directly that facilitated revealing their biological significance. In this study, we developed a novel method for protein variants discovery, according to the characteristics of SAV-peptides.

**Evaluation of de novo-based workflow**

Without limits of database, de novo peptide sequencing derived peptide sequences directly from MS/MS spectra and was able to discover peptides with amino-acid variations. Conventional de novo peptide sequencing could discover peptides under this unrestricted condition with high sensitivity but low accuracy. After preliminary filtered via some experimental information such
as proteolytic forms, screening of de novo-derived peptide candidates excluded from the database (usually the rare one) is more feasible and accurate based on our pairing strategy, since SAVs contained only one amino-acid substitution for the peptides included in the database (usually the dominant one). However, these crude results still contained substantial false identifications. Consequently, as a validation step, database-searching was necessary to evaluate the data precision. Here, only if SAV candidates were identified again from the same spectra by Mascot, they could be reserved with their identified corresponding database-included peptide as SAV-pairs. Furthermore, we took advantage of target-decoy strategy to control the FDR of
identification (Figure 1A). This semi-restricted method provided a high throughput, sensitive, robust and confident approach to discover SAV-pairs, which was a so-called ‘bottom-up’ strategy to reveal both SNP-dependent and independent protein variations at proteome-level, complementary to current ‘top-down’ strategy relying on DNA sequencing.

In synthesized SAV-peptide mixtures assay, we successfully identified 30 SAV-pairs from 39 ones we mixed, which was comparable with classical database-searching results (32 SAV-pairs identified) (Figure 1B). What was more important, our result was of high accuracy, and FDR was acceptable 2.67% (Figure 1C). For this analysis, we used two strict criteria to filter results, eliminating both N-terminal substitutions and deamidated ones. If not used, it was worth noting that 13 non-targeted deamidated substitutions and 22 non-targeted N-terminal substitutions would be reserved as well. Deamidation universally spontaneously happened during process of sample preparation, especially in alkaline environment such as NH₄HCO₃ or TEAB buffer for TMT labeling (Clarke, 1987). We could not distinguish whether the origins of deamidated SAVs were endogenous or not. Therefore, we abandoned all of them conservatively. For the same purpose, we ignored all N-terminal SAVs, because it was difficult to identify plentiful N-terminal substituted particular product ions such as yₙ₋₁ (n = peptide length) and series of b ions in HCD mode. As argued above, we strongly recommend using these two criteria to strictly filter data in order to reduce FDR. For real biological samples, it was well known that SNPs were dominant where protein variations originated from. For this reason and particular objective, we could use an additional strict criterion, keeping SNPs-derived SAVs only, to further improve results.

**Development prospect of SAVs identification**

Attributed to the 1000 Genomes Project (Abecasis et al., 2012), recent efforts to map human genetic variation by sequencing exons and whole genomes had characterized the vast majority of common SNPs across the genome (Le and Durbin, 2011; Mills et al., 2011; Tennessen et al., 2012). Overall, the 1000 Genomes Project Consortium genotyped 38 million SNPs from the genomes of 1092 individuals, which were sampled from 14 populations drawn from Europe, East Asia, sub-Saharan Africa, and the Americas. On average, each person was found to carry ~8000 non-synonymous variants and 250–300 loss-of-function variants in annotated genes.

In our results, we totally detected 248 protein variants in paired PFC and CB tissues from three individuals. As pointed out above, it was estimated that these SAVs covered ~1% non-synonymous variants of each person. It should be underlined here that expressions of SAVs were also tissue-specific, as their corresponding proteins. On the other hand, a large proportion of novel SAVs independent of DNA/RNA sequence detected by this work further indicate extensive protein variants in tissues. Further efforts could be made to increase identification sensitivity by more effective fractionation or targeted protein enrichment. In particular some models of human diseases, such as cancer, protein variations might play more important roles in disease progression and deterioration. For this consideration, it was more significant to apply SAVs discovery to clinical disease-related samples, so as to uncover disease mechanisms and assist in disease diagnosis and treatment.

Recently, several groups integrated RNA-Seq data into classical protein database to help identify SAVs using MS-based workflows (Low et al., 2013; Sheynkman et al., 2014), which could effectively detect SAVs in certain samples. However, there is an obvious limit that RNA-Seq data are vital for this method and it is untransferrable without the assist of RNA-Seq, because it might reduce accuracy as applying RNA-Seq data from other samples or populations. What is more, some SAVs derived from post-transcriptional mechanisms will not be included. According to the consideration of the above factors, our de novo-based method is complementary to their RNA-Seq-assisted SAVs discoveries.

**Essentiality of quantitative SAVs**

Usually, biomarker discovery mainly was focused on differentially expressed proteins among various conditions of cells or clinical tissues. Obviously, our method provided a new dimension for these researches. In TMT assay, we quantified 94 SAV-pairs in all six tissues. In particular cases, we found heterozygous alleles were expressed differentially in paired tissues (Figure 7). Besides, some different SAV expressions were more significant than corresponding proteins (Figure 7). One reason for this was that if only one of paired SAVs presented dramatic difference but the other one not or presented contrary variance, the latter would conceal the dramatic one. This inconsistency of SAV-pairs might weaken the significance of protein, when we measured it using traditional protein-quantitative assays, such as western blot or ELISA. On the other hand, we also detected that the relative expression of SAV-pairs might show difference between homozygote and heterozygote (Figure 6). This result demonstrated that different genetic structures of coding SNPs were diversely associated with physiological traits of individuals. Our study quantitatively revealed tissue-specific expressions of SAVs for the first time. No doubt, applying this approach in disease models for quantitative analysis of known and novel SAVs would create a new direction to probe the relationship between protein function and pathogenesis. Using more precise and sensitive technology, such as SRM, these clues could be validated in extensive samples for further exploring their profound biological significance.

**Preliminary functional analysis of SAVs in human brain tissues**

Individual regions of the brain were likely to have distinct protein compositions, particularly during development and in pathological disorders. Many brain proteins of primary interest were trans-membrane and membrane-bound proteins, including neurotransmitter receptors, ion channels, and G-proteins, which were hardly identified owing to their low abundance and lack of effective methods. In our results, the identified plasma membrane proteins were not enriched. Nevertheless, it was interesting that the percentage of plasma membrane protein variants were significantly high (Supplementary Figure S2C). Here, we totally detected 57 proteins containing SAVs were located in plasma membrane.
From the predicted protein–protein interaction network (Supplementary Figure S15), we found Ras-related C3 botulinum toxin substrate 1 (RAC1) was a key node, which was concordant with its multiplied functions involving in a variety of regulation pathways. Accordingly, we discovered two SAVs sites on RAC1, A146S, and L155V, and the latter one was also quantified by TMT, although there was no difference between two protein variants. Akin to RAC1 was CDC2, which was plasma membrane-associated small GTPase as well. As mentioned above, we identified and quantified its K163R SAVs site from label-free, TMT and SRM assays. According to DNA/RNA sequencing results we knew that it was relative to neither DNA nor RNA mutations. On account of their critical roles, it could be expected that these high-frequency protein variants of plasma membranes might serve as potential clues to reveal numerous functions of distinct sections of human brain and heterogeneity of individuals.

Materials and methods

Synthesized SAV-peptides preparation

Pairs of SAV-peptides were synthesized with stable-isotope labeling by GL Biochem (Shanghai) Ltd, China and JPT Peptide Technologies GmbH, Berlin, Germany. Approximate 10 pmol of each SAV-peptide was mixed together for further analysis.

Tissue collection

All human brain tissues were obtained from the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD (Hu et al., 2011). All subjects were defined as normal controls by forensic pathologists at the NICHD Brain and Tissue Bank. No subjects who suffered a prolonged agonal state were used. For the PFC, samples were taken from the frontal part of the superior frontal gyrus: a cortical region approximately corresponding to Brodmann Area 9.

Tissue lysis

Each human brain tissue, CB and paired PFC tissue were added into SDS-lysis buffer (2% SDS, 0.1 M DTT, 0.1 M Tris–HCl, pH = 7.6), and the ratio of buffer to tissue was 10:1. After homogenization for 3 min, the mixtures were incubated in boiling water for another 3 min. Then, the crude extract was sonicated for 1 min and centrifuged at 16000 g for 10 min at 20°C. The protein concentration was determined by measurements of tryptophan fluorescence as described (Nielsen et al., 2005). Briefly, 1 μl of sample was added into 3 ml of 8 M urea buffer (8 M urea, 20 mM Tris–HCl, pH = 7.6). Fluorescence was excited at 295 nm and measured at 350 nm. A series of distinct concentration of Tryptophan was used as a standard curve, and protein concentration was calculated by assuming the mean tryptophan content in proteins of 1.3%.

Sample digestion

Two hundred micrograms of samples were digested by the FASP procedure as described (Wisniewski et al., 2009). Each sample was transferred to a 10 k filter (Pall Corporation) and centrifuged at 10000 g for 20 min at 20°C. Two hundred microliters of UA buffer (8 M urea, 0.1 M Tris–HCl, pH = 8.5) was added and centrifuged at 10000 g for another 20 min. This step was repeated. Then, the concentrate was mixed with 100 μl of 50 mM IAA in UA buffer and incubated for an additional 40 min at room temperature in darkness. After that, IAA was removed by centrifugation at 10000 g for 20 min. Following washing filter with 200 μl of UA buffer and centrifugation twice, 200 μl of 50 mM NH₄HCO₃ or 0.5 M TEAB buffer (for TMT labeling) was added and centrifuged at 10000 g for 20 min. This step was repeated. Finally, 100 μl of 50 mM NH₄HCO₃ or 0.5 M TEAB buffer and Trypsin (1:25, enzyme to protein) was added and incubated with the mixtures at 37°C for 16 h. The tryptic peptide mixtures were lyophilized and collected.

All three pairs of CB and PFC digestion were also used for TMT labeling following protocol which TMT Mass Tagging Kit (Thermo Scientific) suggested using. For forward assay, 50 μg of CB1, PFC1, CB2, PFC2, CB3, PFC3 were labeled by from 126-tag to 131-tag, respectively. In addition, 50 μg of PFC1, CB1, PFC2, CB2, PFC3, CB3 were separately labeled by from 126-tag to 131-tag again for reverse assay as an experimental duplicate.

Peptide fractionation and desalination

Using a 3100 OFFGEL Fractionator (Agilent Technologies), each peptide mixtures were separated into six fractions or eight fractions (for TMT assays). OFFGEL low resolution kit (pH 3–10, 13 cm, GE Healthcare) was used. The limit of volt-hour, voltage, current, and power were set at 20 kV-h, 8000 V, 50 μA, and 200 mW, respectively. Besides, 20°C of platform temperature were set. A half peptides of each fraction was desalted by self-packed stage-tip (Rappsilber et al., 2007).

nanoHPLC-Q-Exactive system

Synthesized SAV-peptide mixtures were analyzed on nanoHPLC-Q-Exactive system. Liquid chromatography separation was performed on nanoHPLC, Easy-nLC-1000 system (Thermo Scientific). The home-made tip column (15 cm in length, 75 μm inner diameter) was packed with ReproSil-Pur C18-AQ 3 μm resin (Dr Maisch GmbH) in 100% methanol. For high mass accuracy data acquisition, a Q-Exactive mass spectrometry (Thermo Scientific) was used and equipped with a nano-electrospray ion source. Peptide mixtures were loaded onto the column and separated with a linear gradient of 1%–28% buffer B (ACN with 0.1% formic acid) at a flow rate of 200 nL/min over 104 min. The total time for reversed phase analysis was 120 min. Data were acquired in the data-dependent 'top10' mode, in which ten most abundant precursor ions were selected with high resolution (70000 at m/z 200) from the full scan (300–1800 m/z) for HCD fragmentation. Precursor ions with singly charged and charge information unassigned were excluded. Resolution for MS/MS spectra was set to 17500 at m/z 200, target value was 5E5 (AGC control enabled), and isolation window was set to 1.6 m/z. Normalized collision energy was 27 eV.

nanoHPLC-LTQ-Orbitrap-Velos system

Label-free and TMT-labeled CB and paired PFC samples were performed on nanoHPLC-LTQ-Orbitrap-Velos system. Liquid chromatography separation was performed on a nanoflow-HPLC
Exclusion settings: repeat counts, Scientific was used to acquire SRM data. For ionization, triple quadrupole mass spectrometer (TSQ Vantage, Thermo Scientific) was operated with a lock mass option enabled for the most intense ions for MS/MS with charge ≥ +2 with the following Dynamic Exclusion™ settings: repeat counts, 1; repeat duration, 30 sec; exclusion duration, 180 sec. For the HCD strategy, both precursor and product ions were acquired in Orbitrap analyzer. Full scans and MS/MS spectra were acquired at resolution of 30000, profile-mode and 7500, centroid-mode, respectively, with a lock mass option enabled for the 445.120025 ion (Olsen et al., 2005).

SRM assay

RP-HPLC fractionation was performed on a self-packed tip column using a gradient B (0.1% formic acid in acetonitrile) from 2% to 35% in 35 min at a flow rate of ~300 nl/min after split. A triple quadrupole mass spectrometer (TSQ Vantage, Thermo Scientific) was used to acquire SRM data. For ionization, 1800 V of spray voltage and 200 °C of capillary temperature were set. The selectivity of Q1 and Q3 was set at 0.2 and 0.7 Da (FWHM), respectively, and the collision gas pressure of Q1 was set at 1.2 mTorr argon. Using the formula of CE = 0.034 × m/z (precursor ion) + 3.314 for charge +2 (precursor ion) and CE = 0.044 × m/z (precursor ion) + 3.314 for charge +3 (precursor ion), the collision energy (CE) was calculated (Su et al., 2011). For each peptide, >10 SRM transitions were optimized and 5 of them were monitored, and a scan time of 10 msec was used. The median of all detected transitions was used to quantify the corresponding peptide. It was considered as NA if the signal-to-noise ratio (S/N) was <3 (Keshishian et al., 2007).

De novo peptide sequencing

Using pNovo algorithm, all HCD MS/MS spectra were sequenced via de novo sequencing strategy. Of the 20 standard amino-acids and their combinations, only leucine and isoleucine were considered the same. For synthesized SAV-peptide mixtures assay, heavy-Ala (+3.010064), heavy-Val (+6.013809), heavy-Leu (+7.017164), heavy-Lys (+8.014199), heavy-Arg (+10.008269), and carbamidomethylation on cysteine (+57.021464) were set as variable modifications. For human brain tissues label-free assay, carbamidomethylation on cysteine (+57.021464) and oxidation on methionine (+15.994915) were set as variable modifications, and carbamidomethylation on cysteine (+57.021464), oxidation on methionine (+15.994915), TMT-labeling on N-terminal and lysine (+229.162932) were set as variable modifications for human brain tissues TMT-labeling assay. The tolerances of both precursor and fragment ions were set at 20 ppm. All the top 10 C-scored candidates for one reliable MS/MS spectrum were reserved for further analysis.

Database searching

All SAV candidates were incorporated into human Uniprot database (20120202, 58720 entries) where each SAV was treated as a protein. Using Mascot algorithm (Version 2.2.2), all MS/MS spectra were searched against the combined database, in which each genuine protein sequence was followed by a reversed amino-acid sequence. The modifications were set the same as corresponding pNovo setting mentioned above, except that carbamidomethylation on cysteine (+57.021464) was set as fixed modification for human brain tissue assays. Only one missing cleavage site was allowed. The tolerances of peptides and fragment ions were set at 10 ppm and 0.02 Da, respectively. Using software BuildSummary (ProteomicsTools 3.2.2) (Sheng et al., 2012), the output results were combined together and filtered. Both the peptide and protein FDRs were calculated and fixed at no more than 0.01.

Besides, we also used PEAKS 6 (Zhang et al., 2012a) and Andromeda (integrated into MaxQuant v1.3.0.5) (Cox et al., 2011) to validate our SAV-identification in synthesized SAV-peptide mixtures assay. PEAKS was a de novo sequencing assisted database-searching software, and the parameter of its de novo sequencing part was set the same as pNovo. Andromeda database-searching parameter was set the same as MASCOT.

GO analysis

The molecular function item and biological process item of GO analysis was performed by PANTHER online (http://www.pantherdb.org/), and the cellular component item of GO analysis was performed by software Perseus (version1.2.0.17). All the P-values were calculated by the hypergeometric distribution test comparing all identified proteins (ID) with SAV-proteins (SAV) and adjusted by Bonferroni correction.

Protein–protein interaction analysis

Protein–protein interaction analysis was performed by STRING online (http://string-db.org/). High confidence was required and score was set at >0.700. All seven active prediction methods were used and 10 additional key nodes were added, which were white but not identified in our data.

Supplementary material

Supplementary material is available at Journal of Molecular Cell Biology online.

Funding

This work is supported by the grants from Ministry of Science and Technology (2011CB910200, 2011CB910601), a grant from the National Natural Science Foundation of China (31130034), and a grant from the Science and Technology Commission of Shanghai Municipality (13DZ1940100).

Conflict of interest: none declared.

References


