Large-Scale Quantification of Single Amino-Acid Variations by a Variation-Associated Database Search Strategy

Chunxia Song,‡‡†† Fangjun Wang,†† Kai Cheng,‡‡ Xiaoluan Wei,‡‡ Yangyang Bian,†† Keyun Wang,†† Yexiong Tan,§†§ Hongyang Wang,§§ Mingliang Ye,*,†† and Hanfa Zou,*,††

†CAS Key Laboratory of Separation Sciences for Analytical Chemistry, National Chromatographic R&A Center, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, 457 Zhongshan Road, Dalian 116023, China
‡‡SINOPEC Research Institute of Petroleum Processing, 18 Xueyuan Road, Beijing 100083, China
§§The International Cooperation Laboratory on Signal Transduction of Eastern Hepatobiliary Surgery Institute, Second Military Medical University, 225 Changhai Road, Shanghai 200438, China

ABSTRACT: Global quantification of the single amino-acid variations (SAAVs) is essential to investigate the roles of SAAVs in disease progression. However, few efforts have been made on this issue due to the lack of high-throughput approach. Here we presented a strategy by integration of the stable isotope dimethyl labeling with variation-associated database search to globally quantify the SAAVs at the first time. A protein database containing 87,745 amino acid variant sequences and 73,910 UniProtKB/Swiss-Prot canonical protein entries was constructed for database search, and higher energy collisional dissociation combined with collision-induced dissociation fragmentation modes were applied to improve the quantification coverage of SAAVs. Compared with target proteomics in which only a few sites could be quantified, as many as 282 unique SAAVs sites were quantified between hepatocellular carcinoma (HCC) and normal human liver tissues by our strategy. The variation rates in different samples were evaluated, and some interesting SAAVs with significant increase normalized quantification ratios, such as T1406N in CPS1 and S197R in HTATIP2, were observed to highly associate with HCC progression. Therefore, the newly developed strategy enables the large-scale comparative analysis of variations at the protein level and holds a promising future in the research related to variations.

KEYWORDS: single amino-acid variations, quantification, proteomics, liver cancer

INTRODUCTION

Single amino-acid variations (SAAVs) in functional domains of important proteins can alter the progression of many human diseases.1−3 Direct assessment of SAAVs under different physiological conditions provides unparalleled opportunities for SAAV-associated biomarker discovery. Although hundreds of thousands of single nucleotide polymorphisms (SNPs) can be characterized at the genome level,4,5 the large-scale analysis of SAAVs at the protein level is still limited.6,7 Schandorff et al. created a modified protein database by elongating original protein entries with additional variant sequences.7 This enhanced database was then used to identify variant peptides in lysosome isolated from MCF7 breast cancer cell line, and 15 variant peptides were identified. In contrast, Bünget al. identified 36 variant peptides using a separate variant database.8 Recently, an integrated bioinformatics workflow for the detection of variant peptides was developed by Zhang et al., and a total of 204 variant peptides were detected from three colorectal tumor specimens.9 Although these studies paid attention to the variant peptide identification, they did not address the expression level of mutated proteins. This is particularly important in cancer, in which allele copy number variations (CNVs) can vary greatly and epigenetic regulations of alleles can be different.10,11 Thus, the quantitative analysis of SAAVs under different physiological conditions is absolutely needed.12 Theoretically, the quantification information of SAAVs can be obtained at the transcriptome and proteome levels.2 However, some studies have revealed that protein and mRNA copy numbers for any given gene are low-correlated.13,14 Therefore, quantitative analysis of SAAVs at the protein level rather than at mRNA level is vital to better understand their roles.

Multiple reaction monitoring (MRM) approaches have been applied to the absolute quantification of variant peptides;2,12 however, only a few candidates can be monitored and the quantification throughput is limited. Moreover, it is important not only to measure changes in SAAVs but also to distinguish them from protein abundance changes. Alternatively, the quantitative proteomics with shotgun approach can be used for the large-scale quantitative analysis of variant peptide and
protein expression levels. However, few efforts have been made on this important issue.

In this study, a variation-associated database with low redundancy was constructed and applied to the large-scale quantification of SAAV peptides by integrating with stable isotope dimethyl-labeling approach. Two different mass spectrometry (MS) fragmentation modes were applied to improve the identification coverage and reliability of variant peptides. A total of 282 SAAVs peptides were quantified between HCC and normal human liver samples, representing the largest data set of variants quantified by MS. This simple strategy provides in-depth quantitative profiling of SAAVs from different biological samples and might be useful to uncover the associations between variation at the protein levels and disease development.

### EXPERIMENTAL PROCEDURES

#### Chemicals and Materials

Acetonitrile (ACN, HPLC grade) was purchased from Merck (Darmstadt, Germany). All the other chemicals and reagents were purchased from Sigma (St. Louis, MO). The water used in this experiment was prepared using a Mill-Q system (Millipore, Bedford, MA).

#### Sample Preparation and Digestion

The hepatocellular carcinoma (HCC) and normal human liver tissues were provided by Eastern Hepatobiliary Surgery Hospital (Shanghai, China), and the study was approved by the Institutional Review Board of this hospital. Informed consent was obtained from patients enrolled in this study. The normal human liver tissues were the nontumor liver tissues obtained from the advanced stage of the HCC patients removed by the surgical operation. The liver tissues were provided by Eastern Hepatobiliary Surgery Hospital (Shanghai, China), and the study was approved by the Institutional Review Board of this hospital. Informed consent was obtained from patients enrolled in this study. The normal human liver tissues were the nontumor liver tissues obtained from the advanced stage of the HCC patients removed by the surgical operation.

The protein samples were prepared as we previously described. In brief, the liver tissues were lysed in an ice-cold homogenization buffer and precipitated by acetone/ethanol precipitation. After resuspension in the denaturing buffer containing 8 M urea and 100 mM TEAB (triethyl ammonium bicarbonate, pH 8.0), the proteins were reduced by DTT and alkylated by iodoacetamide. Then, the solutions were diluted with 100 mM TEAB and trypsin was added with a weight ratio of trypsin to protein at 1/25 and incubated at 37 °C overnight. The resulting peptide solutions were stored under −80 °C.

#### Dimethyl Labeling Reaction

For the light and heavy labeling, 500 μL of CH₃O (4%, v/v) and CD₃(O) (4%, v/v) was added to the tryptic digests (1 mg in 1 mL 100 mM TEAB solution), respectively; then, 500 μL of freshly prepared NaBH₄CN (0.6 M) was subsequently added. The labeled peptides were subjected to a 1:1 ratio on the basis of the total protein amount. Then, the mixture was desalted on a solid-phase extraction column (SPE, Waters), and lyophilized in a vacuum concentrator.

#### Peptide Fractionation

For the large-scale quantification of HCC-related SAAVs, the off-line strong cation exchange-reversed phase (SCX-RP) multidimensional separation was adopted, and the polySUL-FOETHYL A column (100 mm × 4.6 mm i.d., PolyLC) was used for the 1-D fractionation. Seven mM KH₂PO₄, pH 2.8, 20% ACN (v/v) was used as mobile phase A, and the gradient was performed by 0–15% B (7 mM KH₂PO₄, 500 mM KCl, pH 2.8, 20% ACN (v/v)) in 60 min and 15–100% B in 10 min, then 100% B lasted 10 min. A total of 20 fractions (4 min/fraction) were collected from the first dimension and then were desalted by the C18 SPE column. After resuspension in 200 μL of 0.1% FA, 20 μL of the sample was loaded onto the 2-D RP separation, followed by LC–MS/MS analysis.

#### Nano-LC–MS/MS Analysis

For the test samples, the online SCX-RP multidimensional separation system was constructed as previously described. In brief, 20 μg tryptic digests were loaded onto the SCX monolithic column and then were eluted by 5 salt stepwise elutions to perform the 2-D RPLC–MS/MS analysis.

For the 2-D RPLC separation, a 15 cm × 75 μm i.d. capillary column packed with C18 AQ particles (3 μm, 120 Å, DAISO chemical) was used as the separation column, and 0.1% FA in water and in ACN was used as mobile phases A and B, respectively. For the 13 h online SCX-RP multidimensional separation, the gradient elution was performed with a gradient of 0–3% B in 2 min, 3–35% B in 90 min, 35–80% B in 8 min, 80% B in 10 min, 80–0% B in 5 min, and 100% A for 10 min. For the off-line SCX-RP system, gradient elution was performed by 3–35% B in 150 min.

The LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific) was operated in data-dependent MS/MS acquisition mode. For the higher energy collisional dissociation (HCD) method, full mass scans were performed in the Orbitrap analyzer at a resolution of 30 000 at m/z 400. The 10 most intense ions with charge states ≥2 from the full scan were selected to fragmentation by HCD. The resolution for MS/MS in the Orbitrap analyzer was set to 7500. The normalized collisional energy for HCD was set to 35% and activation time was 0.1 ms. For the collision-induced dissociation (CID) method, the resolution for the precursor ions in the Orbitrap was set to 60 000 at m/z 400. Up to 20 most intense ions were selected for fragmentation by CID in the linear ion trap. The normalized collisional energy for CID was set to 35%, and activation time was 10 ms with a q value of 0.25. The dynamic exclusion function was set as follows: repeat count 1, repeat duration 30 s, and an exclusion duration of 60 s.

#### Database Sources

A conventional UniProtKB/Swiss-Prot protein database (human) including 73 910 protein entries that was released on 11/25/2011 was downloaded from http://www.uniprot.org. Data set “humsavar.txt” containing human polymorphisms and disease mutations was downloaded from www.uniprot.org/docs/humsavar on 12/14/2011. MS-CanProVar database (EnsemblV53, 97 637 variation sequences) was downloaded from http://bioinfo.vanderbilt.edu/canprovar, and Ensembl protein IDs were mapped to UniProtKB/Swiss-Prot protein IDs based on the ID mapping tools in http://www.uniprot.org. The unique variant sequences from humsavar database and MS-CanProVar database were integrated into the UniProtKB/Swiss-Prot canonical protein database. Then, a more comprehensive variation-associated database, named SwissCanSAAVs database, was constructed. For each single amino-acid variation, an independent tryptic peptide with two missed cleavage sites surrounding the centric variation site was extracted from the protein sequences, and an identifier prefixed with “SAAV” was adopted to distinguish with canonical protein
sequence. At last, a total of 87,745 amino acid variant sequences and 73,910 UniProtKB/Swiss-Prot canonical protein entries are included in this new Swiss-CanSAAVs database. It can be accessed from http://bioanalysis.dicp.ac.cn/proteomics/Publications/SAAV/SAAV-Database.htm and will be updated annually for 2 years.

**Protein Identification and Relative Quantification**

Protein identification and quantification are based on the MaxQuant software version 1.1.1.36 according to the standard workflow.\(^{19}\) The database search was performed on the Andromeda search engine against the constructed Swiss-CanSAAVs database concatenated with reversed sequences for evaluating the false discovery rate (FDR).\(^{20}\) Carbamidomethylation on cysteine was set as a fixed modification, whereas oxidation on methionine was set as a variable modification. Trypsin was set as the specific proteolytic enzyme with up to two missed cleavages allowed. For the CID spectra, the mass tolerances for the precursor ions and fragment ions were set to 7 ppm and 0.5 Da, respectively, and these parameters for HCD were set to 7 and 20 ppm, respectively. Doublets were selected as the quantification mode with the dimethyl Lys 0 and N-term 0 as light labels and dimethyl Lys 4 and N-term 4 as heavy labels. All other parameters are the default setting in MaxQuant. After filtering by the FDR criteria, the matched decoy hits and known contaminants were excluded from the list of identified peptides, and only the unique peptides with the highest score were retained. For the final identification results, the variations for isoleucine (I) to leucine (L), which had no mass shift, were excluded from the final list of SAAV peptides.

**RESULTS AND DISCUSSION**

**Construction of a Variation-Associated Protein Database**

Integrating variant sequences into the protein database is an effective way to achieve comprehensive characterization of SAAVs by shotgun proteomics.\(^ {7,9}\) Recently, Zhang et al. constructed a MS-CanProVar database to identify the cancer-related variant peptides.\(^ {9,21}\) However, redundant sequences were observed in this variation-containing database, and the identification sensitivity of variant peptide was decreased (Supporting Information). To further improve the identification coverage of SAAVs, we combined the unique variant sequences from MS-CanProVar database and humsavar database (http://www.uniprot.org/docs/humsavar) to construct a more comprehensive variation-associated database, which was named Swiss-CanSAAVs database. A total of 87,45 amino-acid variant sequences and 73,910 UniProtKB/Swiss-Prot canonical protein entries are included in this database.

In general, unambiguously identification of SAAVs peptide requires a refined FDR calculation due to their lower score distribution.\(^ {8,9}\) In our data set, we did not observe lower identification score distributions for the variant peptides compared with the canonical peptide (Figure 1A). However, SAAVs and canonical peptides exhibited different occurrences.\(^ {9}\) Therefore, only variant peptides and corresponding decoys were considered for the FDR\(_\text{SAAV}\) estimation. Consistent with the results of Zhang et al.,\(^ {9}\) a higher FDR\(_\text{SAAV}\) (6.1%) was observed for the variant peptides. Thus, a more stringent score criterion for variant peptide was set to control the FDR\(_\text{SAAV}\) < 1%. Finally, a slight shift to higher score distribution was observed and most of the SAAVs peptides with lower identification scores were excluded (Figure 1B). To guarantee the confidence of variant peptide identifications, only the variant peptides that passed manual validation were kept in our final data set.

**Figure 1.** Search score distribution of single amino acid variation (SAAV) and canonical peptides before (A) and after (B) filtering based on the separate FDR estimation <1%.

**Figure 2.** Overlap of canonical peptides (A) and variant peptides (B) quantified by HCD and CID.
SAAVs Peptides Identification by Different MS Fragmentation Modes

HCD has been reported to provide better proteome identification due to its high efficiency in peptide fragmentation.\textsuperscript{22−24} Moreover, by performing the MS/MS in an orbitrap, higher mass accuracy is achieved and more informative low mass reporter ions can be detected. These characteristics are helpful for unambiguous identification of SAAVs peptides, especially for excluding the interference of various peptide modifications with similar mass shifts.\textsuperscript{25} However, the performance of HCD and CID fragmentation modes used for the identification of SAAV peptides has not been investigated.

We first tested and compared the performances of HCD and CID for variant peptides identification. In brief, 20 μg tryptic digests of protein extracted from normal and cancer human liver tissues were isotopically labeled and analyzed by the online SCX-RP 2D LC−MS/MS system. HCD and CID fragmentation modes were employed in two independent 2D analyses, respectively. After controlling the FDRs <1% for both canonical and SAAVs peptides, 5792 and 5905 unique peptides, including 69 and 53 variant peptides were quantified by HCD and CID, respectively (Figure 2). Similar identification capability was observed for these two different fragmentation modes. However, HCD provides better identification results for longer variant peptides with higher charge states, whereas CID outperforms HCD for the identification of 2+ and 3+ peptides (Supporting Information). In addition, some low mass reporter ions (y\textsubscript{1}+ and y\textsubscript{2}+ ions) can be detected only in HCD spectra (Figure S3 in the Supporting Information), whereas much more informative b-type ions are observed in CID spectra (Figure S4 in the Supporting Information). Overall, these two fragmentation modes are complementary for the identification of SAAVs peptides, and combination of them emerges as a good choice to improve the identification coverage and accuracy of variant peptides.

Large-Scale Quantification of HCC-Related SAAVs Peptide

HCC is one of the most common cancers, and over 1 million people die of this cancer per year.\textsuperscript{26} Many risk factors, such as chronic hepatitis and liver cirrhosis, contribute to the development of HCC.\textsuperscript{27,28} However, the molecular mechanism of HCC, especially the changes at the gene and protein expression levels, remains poorly understood. Numerous genomic studies revealed that many variations were responsible for the development and progression of HCC.\textsuperscript{29−31} Nevertheless, there are few studies of the associations between variations at the protein level and HCC development. As an important part of the Human Proteome Project (HPP), the Chromosome-centric HPP (C-HPP) made great efforts to understand the linkages between genetic perturbations and the dysregulation of proteins.\textsuperscript{32,33} Recently, Chinese scientists initiated proteomic exploration on human chromosomes 1, 8, and 20, and comprehensive proteome data sets on these chromosomes were obtained by using normal and carcinomatous liver tissues.\textsuperscript{34−36} However, the SAAVs were not considered in their studies.

As one of the most popular quantitative proteomics strategies, stable isotope dimethyl-labeling approach has been
widely used in the large scale quantification of proteins. It is based on the reductive amination for primary amines of peptides and has been proved to be a robust and cost-effective quantitative method, especially for the tissue samples. Therefore, we applied the stable isotope dimethyl-labeling approach to the comparative analysis of SAAVs in HCC and normal human liver tissues (Figure 3). In brief, 1 mg tryptic digests of protein extracted from normal and cancer human liver tissues were isotopically labeled and fractionated by SCX. One tenth of each fraction was analyzed by the 2-D RPLC-MS/MS analysis, and CID and HCD modes were adopted as fragmentation methods, respectively. After controlling the separate FDRs <1% for both canonical and SAAVs peptide, 27,499 and 22,187 unique peptides, including 251 and 206 variant peptides, were quantified by CID and HCD, respectively. After integrating the quantification results from the same variation site in two fragmentation modes, a total of 282 unique variation sites were obtained (Table S1 in the Supporting Information). To the best of our knowledge, this is the first quantification data set of HCC-associated variations at the protein level, and the quantitative throughput is much higher than that in other studies by MRM method.

To better understand the significance of SAAVs at the protein level, the variation rates (the fraction of a given variation in total protein expression) in different samples should be evaluated. For example, a 2-fold increase in relative quantification may correspond to changes from 0.1 to 0.2% or from 50 to 100% in variation rate. Obviously, these two different conditions represent distinct effects on the cellular states and may have different impact on disease development. However, how to determine the variation rates on the large scale is still in its infancy. Here we calculated the variation rates in different samples based on three different ratios of variant peptides, canonical peptides and proteins. (For details, see the Supporting Information.) In particular we assumed that (i) the sum of the copy number of variants and canonical peptides equals the copy number of the corresponding proteins; (ii) all of the quantification results are accurate; and (iii) the ratios between different states change. Finally, the variation rates of 91 SAAVs sites were calculated (Table S1 in the Supporting Information). For most of the variant peptides, it is difficult to detect their canonical peptides simultaneously because of the complexity of samples and the data-dependent acquisition mode in MS/MS analysis. Therefore, for 88 variation sites, which only have the corresponding protein ratios, we introduce a notion of “normalized quantification ratio” to remove the contribution of protein abundance changes to the abundance changes of variant peptides (Table S1 in the Supporting Information). Here the normalized quantification ratio for each variant peptide was calculated by normalizing against the ratio of corresponding protein.

The associations between HCC development and SAAVs with increased normalized quantification ratios were emphasized in our quantitative results. For example, the T1406N variation of carbamoyl phosphate synthase (CPS1), the rate-determining enzyme in the urea cycle, showed a 10-fold
increase in HCC samples. The T1406N in CPS1 is located at the critical N-acetylglutamate-binding domain and is associated with neonatal pulmonary hypertension and hepatic veno-occlusive disease.\textsuperscript{39,40} A series of $y_{19}$, $y_{21}$, $y_{22}$, and $y_{26}$ ions indicating a +13 Da mass shift corresponding to the substitution of asparagine (Asn) for threonine (Thr) at position 1406 were observed in our study (Figure 4). The variant peptide was up-regulated by 11-fold, whereas the canonical peptide and protein expression did not change (Figure S4A,B). The variation rates in HCC and normal liver tissues were 19.3 and 1.9%, respectively, which emphasized the role of T1406N in HCC development.

HIV-1 TAT-interactive protein 2 (HTATIP2) is an important tumor suppressor that acts by promoting apoptosis and inhibiting angiogenesis.\textsuperscript{41} Five mutations in this protein have been observed in HCC samples, and some of them have been reported to dramatically accelerate tumorigenesis in HepG2-cell-injected mice.\textsuperscript{28,42} A S197R variation in HTATIP2 with a 3-fold increase in normalized quantification ratio was observed by HCD and CID. The consecutive $y$ ions with a high score demonstrated the high confidence identification of this SAAVs peptide (Figure S4 in the Supporting Information). A significant increase in this mutation in tumor samples emphasized its important role in HCC progression.

**CONCLUSIONS**

In this study, we developed a strategy for the comparative analysis of SAAVs by integration of stable isotope dimethyl labeling with a variation-associated database search approach. As many as 282 unique variation sites were quantified in the human liver tissues, and the associations between variations with increased normalized quantification ratio and HCC development were discussed. Although only the known genomic sequence variations can be identified by this database-search approach, the reduced search space improves the identification of variants at the protein level. With the improvement of the variation-related database and MS-related techniques, the variation-associated quantitative proteomics will play a more important role in the study of SAAVs and will be highly applicable to the ongoing C-HPP.\textsuperscript{32,33}

**ASSOCIATED CONTENT**

**Supporting Information**

Construction of the Swiss-SAAV database and comparison with the MS-CanProVar database; comparison of the HCD and CID fragmentation modes in variant peptide identification; calculation of variation rates for variant peptides; a figure comparing the search score distribution of peptides with different charge states identified by HCD and CID; a figure comparing average score distribution of peptides with different charge states identified by HCD and CID; MS/MS spectra of the variant peptide HVGDVGNVTADKDGADVSIEDSVISLSDHCIGR in CID and HCD mode; MS/MS spectra of the variant peptide KFFGSFLDSWAR in CID and HCD mode; MS/MS spectra for the SAAV peptides with significant changes which identified/quantified based on a single identification; a table listing 282 HCC-related variation sites and their quantification results; Swiss-CanSAAVs database containing 87 745 amino acid variant sequences and UniProtKB/Swiss-Prot canonical protein entries. This material is available free of charge via the Internet at http://pubs.acs.org.
Feasibility of large-scale phosphoproteomics with higher energy sequencing speed.
M.; Wouters, E. R.; Senko, M.; Makarov, A.; Mann, M.; Horning, S. A.


