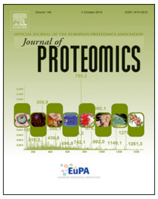
Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot



Quality control of single amino acid variations detected by tandem mass spectrometry

Xinpei Yi[a,c,1], Bo Wang[b,1], Zhiwu An[a,c], Fuzhou Gong[a,c,*], Jing Li[b,**], Yan Fu[a,c,*]

[a] NCMIS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[b] Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
[c] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:
Single amino acid variations
Mass spectrometry
Peptide identification
False discovery rate
Unrestrictive mass shift relocalization

ABSTRACT

Study of single amino acid variations (SAVs) of proteins, resulting from single nucleotide polymorphisms, is of great importance for understanding the relationships between genotype and phenotype. In mass spectrometry based shotgun proteomics, identification of peptides with SAVs often suffers from high error rates on the variant sites detected. These site errors are due to multiple reasons and can be confirmed by manual inspection or genomic sequencing. Here, we present a software tool, named SAVControl, for site-level quality control of variant peptide identifications. It mainly includes strict false discovery rate control of variant peptide identifications and variant site verification by unrestrictive mass shift relocalization. SAVControl was validated on three colorectal adenocarcinoma cell line datasets with genomic sequencing evidences and tested on a colorectal cancer dataset from The Cancer Genome Atlas. The results show that SAVControl can effectively remove false detections of SAVs.

Significance: Protein sequence variations caused by single nucleotide polymorphisms (SNPs) are single amino acid variations (SAVs). The investigation of SAVs may provide a chance for understanding the relationships between genotype and phenotype. Mass spectrometry (MS) based proteomics provides a large-scale way to detect SAVs. However, using the current analysis strategy to detect SAVs may lead to high rate of false positives. The SAVControl we present here is a computational workflow and software tool for site-level quality control of SAVs detected by MS. It accesses the confidence of detected variant sites by relocating the mass shift responsible for an SAV to search for alternative interpretations. In addition, it uses a strict false discovery rate control method for variant peptide identifications. The advantages of SAVControl were demonstrated on three colorectal adenocarcinoma cell line datasets and a colorectal cancer dataset. We believe that SAVControl will be a powerful tool for computational proteomics and proteogenomics.

1. Introduction

Single nucleotide variations (SNVs) or single nucleotide polymorphisms (SNPs) resulting from single base mutations, are recognized as the most frequent type of genetic variations in the human genome [1]. These variations are often associated with particular physiological or pathological traits in individuals [2]. Protein sequence variations caused by SNVs are single amino acid variations (SAVs). The investigation of SAVs provides a chance for understanding the relationships between genotype and phenotype [3–8].

Over the past decade, tandem mass spectrometry (MS/MS) based shotgun proteomics has developed rapidly as a high throughput method to investigate proteins in biological and clinical samples [9, 10]. To identify the variant peptides with SAVs, the widely used approach is to search the MS/MS spectra against a protein sequence database containing variation information [11–13]. In recent years, we have witnessed the advent of such databases in a variety of studies [14–18]. One typical example is the human Cancer Proteome Variation (CanProVar) database [19] that integrated variation information from databases of dbSNP [15], Catalogue of Somatic Mutations in Cancer [16] and Online Mendelian Inheritance in Man [20].

There are many reasons that can lead to false positive matches in database searching. Therefore, rigorous quality control of peptide identifications is necessary [21–24]. In shotgun proteomics, false

[*] Corresponding authors at: NCMIS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.
[**] Corresponding author.
E-mail addresses: fzgong@amt.ac.cn (F. Gong), jing.li@sjtu.edu.cn (J. Li), yfu@amss.ac.cn (Y. Fu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jprot.2018.07.004
Received 7 March 2018; Received in revised form 26 June 2018; Accepted 2 July 2018

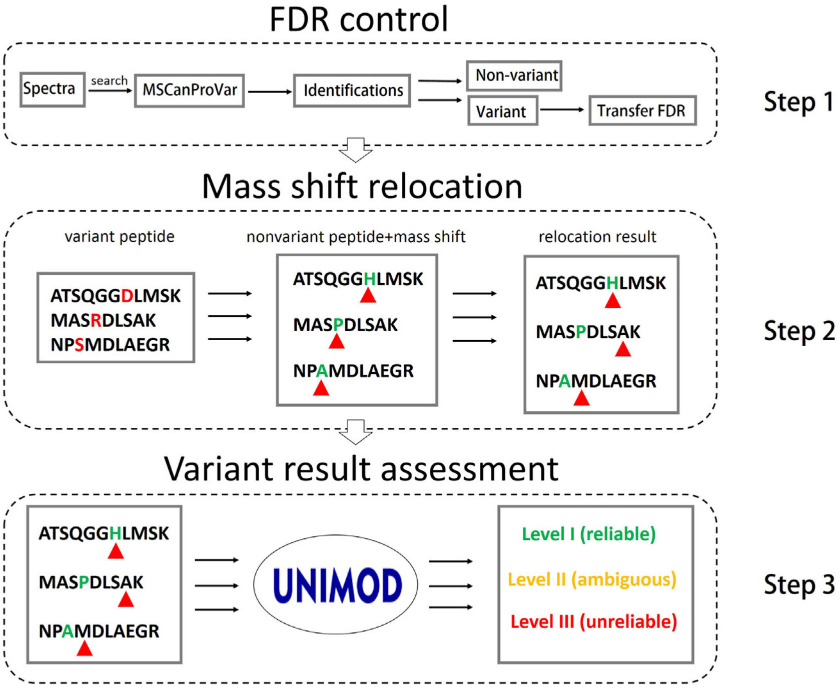


Fig. 1. SAVControl workflow for site-level quality control of variant peptide identifications. After searching the MS/MS spectra against MSCanProVar, a protein sequence database containing variation information, SAVControl processes the search results with three steps: 1) transfer FDR control of the variant peptide identifications, 2) mass shift relocalization with PTMiner, and 3) search of the Unimod database and classification of variant sites into three levels.

discovery rate (FDR) [25] is a commonly used statistical confidence measure for this purpose. The target-decoy database search strategy [26, 27] is widely used to estimate the FDR of peptide identifications, where the database is composed of forward protein sequences and their reversed (or randomized, shuffled) ones. Since an incorrect identification has an equal chance of being a match to the target sequences or the decoy sequences, the number of decoy matches above a score threshold can be used to estimate the number of random target matches. Consequently, the FDR is estimated by the number of decoy identifications divided by the number of target identifications above the score threshold.

However, for identification of variant peptides, the above FDR control method may be unreliable. One reason is the FDR heterogeneity of different groups of peptide identifications [12, 28], e.g., the variant and non-variant ones here. At the same score threshold, the FDR of variant peptide identifications that are of interest may be significantly different from the global FDR of all peptide identifications that are usually obtained and processed together. Several approaches have been proposed to calculate the FDR of variant identifications dedicatedly [17, 29, 30]. The simplest one is to apply the target-decoy FDR estimation to the group of variant peptide identifications separately [17, 30]. However, this separate FDR strategy suffers from large variances when the number of variant peptide identifications is small. In order to overcome this drawback, Li et al. [17] proposed a refined separate FDR for variant peptide identifications, in which the number of false positives in variant identifications above a score threshold is estimated by the number of all decoy matches above this threshold multiplied by the proportion of variant sequences among all decoy matches below the threshold. Fu et al. [31] proposed a more accurate method, called transfer FDR, for quality control of small groups of peptide identifications. Transfer FDR estimates the proportion of decoy matches belonging to the group as a function of peptide score. Although proposed for modified peptide identification, transfer FDR is in principle applicable to variant peptides.

More importantly, even if the FDR of peptide identifications can be properly controlled, there is often a high error rate on the variant sites. For example, Li et al. [17] verified some of the identified variant sites using genomic sequencing and found that their real error rate was much larger than the FDR control level. In this paper, we also confirmed this phenomenon by manual analysis and genomic sequencing. For most of the incorrectly identified variant sites, their responsible fragment ions are often missing or have very low intensities. The main reason that these false positives passed the FDR control is the sequence homologies between the identified variant peptides and the true peptides [12]. A common scenario is that the observed mass shift of peptide precursor is caused by some amino acid modification or precursor mass error instead of mutation [12]. Unconsidered modifications have been recognized as an important reason for un-identified or mis-identified spectra [12, 32–37]. However, there still lacks a powerful tool to evaluate the confidence of variant peptide identifications by taking into account this issue.

In this paper, we present a computational workflow and software tool, named SAVControl, for site-level quality control of variant peptide identifications. It first filters variant peptide identifications by transfer FDR control, and then evaluates the reliability of the variant sites by unrestrictive mass shift relocalization and introducing alternative interpretations, e.g. modifications. Finally, all identified variant sites are classified into three levels: Level I (reliable), Level II (ambiguous) and Level III (unreliable).

To validate SAVControl, we analyzed three datasets of colorectal cancer cell lines SW480, RKO and HCT-116. Genomic sequencing evidences showed that SAVControl successfully detected all the falsely identified variations. We further applied the workflow to a colorectal cancer dataset from The Cancer Genome Atlas (TCGA). For this dataset, ~29% of the variant peptide identifications that passed the FDR control were recognized as unreliable or ambiguous by SAVControl. To the best of our knowledge, SAVControl is the first attempt at automated site-level quality control of variant peptide identifications.

2. Methods

Fig. 1 depicts the SAVControl workflow for the quality control of variant peptide identifications in shotgun proteomics. It processes the results of database searching by three steps: FDR control, mass shift relocalization, and variant site assessment.