Contents lists available at ScienceDirect

Journal of Proteomics

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

mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry

Guoshou Teo [a,b], Sinae Kim [c], Chih-Chiang Tsou [d], Ben Collins [e], Anne-Claude Gingras [f,g], Alexey I. Nesvizhskii [d,h], Hyungwon Choi [b,*]

[a] *Department of Applied Probability and Statistics, National University of Singapore, Singapore*
[b] *Saw Swee Hock School of Public Health, National University of Singapore, Singapore*
[c] *Department of Biostatistics, School of Public Health, Rutgers University, Piscataway, NJ, USA*
[d] *Department of Pathology, University of Michigan, Ann Arbor, MI, USA*
[e] *Department of Biology, Institute of Molecular Systems Biology, ETH Zürich, Zürich, Switzerland*
[f] *Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada*
[g] *Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada*
[h] *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA*

ARTICLE INFO

Article history:
Received 13 August 2015
Received in revised form 5 September 2015
Accepted 8 September 2015
Available online xxxx

Keywords:
Data independent acquisition
Data preprocessing
Normalization
Differential expression

ABSTRACT

Data independent acquisition (DIA) mass spectrometry is an emerging technique that offers more complete detection and quantification of peptides and proteins across multiple samples. DIA allows fragment-level quantification, which can be considered as repeated measurements of the abundance of the corresponding peptides and proteins in the downstream statistical analysis. However, few statistical approaches are available for aggregating these complex fragment-level data into peptide- or protein-level statistical summaries. In this work, we describe a software package, mapDIA, for statistical analysis of differential protein expression using DIA fragment-level intensities. The workflow consists of three major steps: intensity normalization, peptide/fragment selection, and statistical analysis. First, mapDIA offers normalization of fragment-level intensities by total intensity sums as well as a novel alternative normalization by local intensity sums in retention time space. Second, mapDIA removes outlier observations and selects peptides/fragments that preserve the major quantitative patterns across all samples for each protein. Last, using the selected fragments and peptides, mapDIA performs model-based statistical significance analysis of protein-level differential expression between specified groups of samples. Using a comprehensive set of simulation datasets, we show that mapDIA detects differentially expressed proteins with accurate control of the false discovery rates. We also describe the analysis procedure in detail using two recently published DIA datasets generated for 14–3-3β dynamic interaction network and prostate cancer glycoproteome.

Availability: The software was written in C++ language and the source code is available for free through SourceForge website http://sourceforge.net/projects/mapdia/.

1. Introduction

The data dependent acquisition (DDA) mode of analysis has long been the prevailing platform in mass spectrometry (MS)-based shotgun proteomics. In the DDA mode, more abundant precursor peptide ions are preferentially isolated and fragmented to generate tandem mass (MS/MS) spectra. These MS/MS spectra are then computationally analyzed to identify the peptides and to infer the corresponding proteins. In this strategy, peptides are quantified using the intensity of the precursor peptide signal detected in the first stage of MS analysis (MS1 quantification). A well-known limitation of the DDA strategy is that precursor selection is systematically biased in favor of more abundant peptides, which leads to inconsistent detection and quantification of lower abundance peptides across multiple samples. This is particularly a problem in complex samples where the number of co-eluting species to be sequenced exceeds the duty cycle of the mass spectrometer [1,2].

An alternative mode of analysis, called data independent acquisition (DIA), has the potential to provide more consistent peptide quantification [3,4]. In the currently favored DIA set-ups, the entire mass range relevant to the experimentalist is covered using a set of wide windows,

Abbreviation: AP-MS, Affinity purification – mass spectrometry; DDA, Data dependent acquisition; DIA, Data independent acquisition; FDR, False discovery rate; pROC, Pseudo receiver operating characteristic; RT, Retention time; TIC, Total ion chromatogram; TIS, Total intensity sum.

* Corresponding author.
E-mail address: hyung_won_choi@nuhs.edu.sg (H. Choi).

http://dx.doi.org/10.1016/j.jprot.2015.09.013
1874-3919/© 2015 Elsevier B.V. All rights reserved.

which allows segmented acquisition of MS/MS spectra for an unbiased set of precursors. All precursor peptide ions within each window are co-isolated and subjected to fragmentation to produce multiplex MS/MS spectra. Although DIA had been initially proposed years ago [3,5], it was not until recently that advances in the instrumentation enabled faster scans with improved resolution or resolving power, allowing practical implementations of this strategy. One commonly used DIA strategy, SWATH-MS, was first implemented on a Qq-TOF AB SCIEX instrument using a sequence of 25 m/z-wide precursor isolation windows [2], and related methods are now available on MS instruments from other manufacturers, including on the Thermo Fisher Q Exactive system. For example, a variant of this strategy, called MSX, uses a stochastic selection of smaller (e.g. 4 m/z wide) precursor isolation windows and has been shown to reduce the fragment ion interference and increased precursor selectivity [6].

Because virtually every peptide ion is selected for fragmentation, DIA theoretically allows more consistent peptide detection and quantification across multiple samples, resulting in more complete quantitative coverage (i.e., less missing data) [7]. In addition, DIA data changes the way quantitative data are analyzed compared to the traditional quantitative DDA proteomics analysis. The volume of quantitative information in the DIA data is considerably larger than that of the DDA data, since the intensity data can be extracted not only at the peptide level from MS1 data but also at the MS/MS fragment level from MS2 data. The current approaches for DIA data analysis, however, do not take full advantage of this extended (fragment-level) data and instead use peptide/protein intensities summed over the fragments [8,9,10].

The fragment intensity data can be viewed as repeated measures of the intensity of their parent peptides (this information is lost once the intensity data are aggregated). From a statistical point of view, these data create the opportunity to improve the reliability of statistical analysis, since the fragment intensity data allow us to estimate the reproducibility of relative quantification provided that they are correlated with the (unknown) quantitative level of their parent peptides across the samples. In other words: there are much more data to work with to draw inferences for protein expression changes per protein basis in the DIA data in comparison to the DDA data analyzed at the level of MS1 only.

Nevertheless, the complexity of the DIA data poses numerous challenges to its extraction and analysis. At present, the default data analysis strategy for DIA data is targeted quantification using tools such as OpenSWATH [8], Skyline [11] or PeakView (AB Sciex) that all use spectral assay libraries generated by DDA for matching peaks and extracting their areas. This requirement for external spectral libraries is however not absolute, and can be alleviated using, for example, the new computational workflow DIA-Umpire that enables untargeted identification and quantitative extraction [9]. In either case, the MS2 DIA data may contain fragments that are shared across multiple co-eluting precursor ions within the same isolation window, creating a difficult problem for quantification. Furthermore, after data extraction for each sample, the fragment maps will not necessarily be reproducible across multiple runs if the chromatographic elution patterns are distorted by factors such as pressure and temperature changes in the column or fragment ion interference. Therefore a reliable set of fragments has to be selected carefully before the statistical analysis is performed.

Several different types of challenging cases (non-reproducible peptides; too little data) are simultaneously present in any DIA-MS dataset, and these challenges have direct ramifications for statistical analysis of large DIA datasets. Supplementary Figs. 1 and 2 demonstrate real examples of fragment intensity data in the 14–3-3β dynamic interactome dataset we will analyze later. In these figures, the intensity data from a time course affinity purification experiment with three biological replicates were transformed into log scale (base 2), and the data for each fragment were centered by median within each biological replicate. Supplementary Fig. 1 shows example proteins in which most fragments from these peptides are well correlated with one another and faithfully represent their parent protein abundance. By contrast, Supplementary Fig. 2 shows the other side of the reality. Here, MYCBP2 and YWHAB (14–3-3β/α) contain many peptides with several associated fragments, yet they both suffer from poor reproducibility across peptides within each protein. On the other hand, while the reproducibility within and between time points is fair for CYB5R3, there are only two peptides and, in contrast to the two proteins above, they provide relatively limited evidence to draw precise statistical inference for this protein. Thus, careful post-extraction processing of fragment-level intensity data is necessary to preclude spurious findings (i.e. inaccurately quantified fragments) that percolate through the final stage of statistical significance analysis.

The data analysis challenges from DIA are not entirely addressed by the currently existing statistical software tools. For instance, the majority of statistical analysis software packages are designed for protein or peptide intensity data, but not fragment intensity data. For example, the DANTE software package offers regression model-based analysis of peptide intensity data [12]. The MaxQuant-Perseus packages enable protein quantification via the LFQ (label-free quantification) or iBAQ (intensity-based, absolute quantification) values and perform subsequent statistical analysis of these data [13]. MSstats (version 2.3.4) is currently the only statistical software capable of differential expression analysis using fragment intensity data, since it was originally written for S/MRM (selected/multiple reaction monitoring) data [14]. However, whether the regression-based framework currently implemented in MSstats is adaptive to far more complex DIA data has not been rigorously examined. In particular, as illustrated in Supplementary Figs. 1 and 2, the fragment intensities in DIA data can vary significantly between different peptide precursors from the same protein. This type of data may expose any statistical model to erroneous quantification and false discoveries more easily than the S/MRM data that uses specifically isolated transitions that have been carefully selected by the experimentalists.

In light of these issues, and with the number and scope of DIA studies rapidly expanding, it is therefore of great importance to evaluate the existing options and develop new tools, if necessary, which will render the statistical significance analysis of fragment-level intensity data as robust as possible. In this work, we present mapDIA, the first comprehensive software package specifically designed for the fragment-level intensity data generated in the DIA mode. mapDIA tackles the challenges associated with these data in three major steps: normalization, fragment/peptide selection, and statistical modeling.

1.1. Experimental Procedures

Here we describe the detailed methods for data preprocessing and statistical analysis implemented in the mapDIA workflow. We also present additional details regarding experimental designs and simulation setup in this section. The input data to mapDIA can be acquired from the targeted data extraction tools such as OpenSWATH [8] and Skyline [11] with a prebuilt spectral assay library, or DIA-Umpire [9] that does not required a spectral assay library.

1.2. Data preprocessing and statistical model in mapDIA

1.2.1. Step 1: Intensity normalization

Using the extracted fragment intensity (or peak area) data, the first data preprocessing step in mapDIA begins with the normalization of intensity data (Fig. 1a). Here the goal is to remove systematic variations in the chromatography across different samples, specifically the variations in the total intensity sum in short periods of chromatographic time or retention time (RT). A commonly used data normalization strategy is to divide fragment intensities by the total intensity sum (TIS), i.e. the sum of intensities of all detected fragments in each sample. Denoting the entire dataset by $\mathbf{Y} = \{y_{fs}\}$, a $F \times S$ matrix of intensity values