Bioinformatics, 35(5), 2019, 898–900
doi: 10.1093/bioinformatics/bty727
Advance Access Publication Date: 23 August 2018
Applications Note

Systems biology

PANDA: A comprehensive and flexible tool for quantitative proteomics data analysis

Cheng Chang [1,*,†], Mansheng Li [1,†], Chaoping Guo [2], Yuqing Ding [2], Kaikun Xu [1], Mingfei Han [1], Fuchu He [1] and Yunping Zhu [1,*]

[1] State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Lifeomics, National Center for Protein Sciences (Beijing), Beijing 102206, Peoples Republic of China and [2] Beijing Key Laboratory of Human Computer Interactions, Institute of Software Chinese, Academy of Sciences, Beijing 100190, P.R. China

*To whom correspondence should be addressed.
[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Oliver Stegle

Received on December 10, 2017; revised on June 29, 2018; editorial decision on August 19, 2018; accepted on August 21, 2018

Abstract

Summary: As the experiment techniques and strategies in quantitative proteomics are improving rapidly, the corresponding algorithms and tools for protein quantification with high accuracy and precision are continuously required to be proposed. Here, we present a comprehensive and flexible tool named PANDA for proteomics data quantification. PANDA, which supports both label-free and labeled quantifications, is compatible with existing peptide identification tools and pipelines with considerable flexibility. Compared with MaxQuant on several complex datasets, PANDA was proved to be more accurate and precise with less computation time. Additionally, PANDA is an easy-to-use desktop application tool with user-friendly interfaces.

Availability and implementation: PANDA is freely available for download at https://sourceforge.net/projects/panda-tools/.

Contact: 1987ccpacer@163.com or zhuyunping@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online

1 Introduction

Quantitative proteomics is gaining its popularity by providing a global and systematic view on biological processes and cellular functions (Schubert *et al.*, 2017). There are two kinds of approaches to protein quantification according to whether the sample is isotope labeled, i.e. the label-free and labeled quantifications. Nowadays, numbers of algorithms and software tools have been proposed and developed to facilitate label-free or labeled quantification of proteomics data.

Due to the variety of experiment designs and strategies in quantitative proteomics, current quantification software tools are usually only suitable for a few specific quantitative experiment strategies, such as PyQuant (Mitchell *et al.*, 2016) and SILVER (Chang *et al.*, 2014) for stable isotope labeling quantification, RIPPER (Van Riper *et al.*, 2016) and LFQuant (Zhang *et al.*, 2012) for label-free quantification. Even the famous tool MaxQuant (Cox and Mann, 2008), which contains many methods for label-free and labeled quantifications, cannot support ^{15}N labeling method. Moreover, MaxQuant consists of its own mass spectrometry (MS) data analysis algorithms, which are not compatible with other tools or pipelines. In brief, there is a lack of comprehensive and flexible quantification tools for the rapidly developing quantitative proteomics.

Here, we present a new tool named PANDA for accurate and precise analysis of quantitative proteomics with high comprehensiveness and flexibility. PANDA can process MS data from different instrument manufacturers by reading the standard formats mzXML and mzML. It is also able to be compatible with existing peptide identification tools (e.g. Mascot) by supporting the standard format mzIdentML. PANDA contains multiple methods to deal with MS data produced in various kinds of quantitative strategies. Further, by integrating the advanced algorithms of our previous quantification tools LFQuant and SILVER, PANDA has

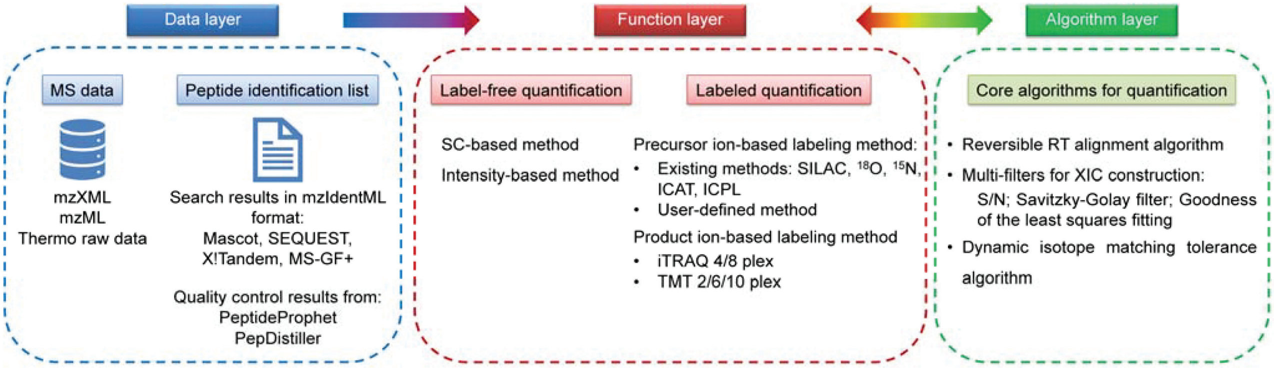
Data layer
Function layer
Algorithm layer
MS data
Peptide identification list
mzXML
mzML
Thermo raw data
Search results in mzIdentML format:
Mascot, SEQUEST, X!Tandem, MS-GF+
Quality control results from:
PeptideProphet
PepDistiller
Label-free quantification
Labeled quantification
SC-based method
Intensity-based method
Precursor ion-based labeling method:
• Existing methods: SILAC, 18O, 15N, ICAT, ICPL
• User-defined method
Product ion-based labeling method
• iTRAQ 4/8 plex
• TMT 2/6/10 plex
Core algorithms for quantification
• Reversible RT alignment algorithm
• Multi-filters for XIC construction:
S/N; Savitzky-Golay filter; Goodness of the least squares fitting
• Dynamic isotope matching tolerance algorithm

Fig. 1. The schema of PANDA workflow. PANDA consists of three core components, i.e., the data layer, the function layer and the algorithm layer

been demonstrated to be accurate and precise for protein quantification.

2 Materials and methods

2.1 Benchmark datasets

For label-free quantification, the yeast samples with a serial dilution of UPS2 (Proteomics Dynamic Range Standard, Sigma-Aldrich) standard proteins (1 µg, 0.2µg, 0.04µg, 0.008µg) spiked in from (Chang *et al.*, 2016) were analyzed in this study. For labeled quantification, a large-scale complex dataset obtained from HeLa cells (Cox and Mann, 2008) with stable isotope labeling by amino acids in cell culture (SILAC) was used. Moreover, several phosphoproteomic datasets (Hogrebe, *et al.*, 2018) and a ^{15}N labeling dataset (Arsova, *et al.*, 2012) were also used for evaluation in this study. See Supplementary Methods for details.

2.2 PANDA workflow

PANDA is designed for comprehensive and flexible analysis of both label-free and labeled quantitative proteomics data. As shown in Figure 1, PANDA consists of three core layers, i.e. the data layer, the function layer and the algorithm layer. (i) The data layer includes two kinds of input data in PANDA: MS data and peptide identification. For MS data, PANDA can directly process Thermo raw files through MSFileReader. Besides, it can also take the MS data standard formats mzXML and mzML as input. For peptide identification, being able to access the mzIdentML format proposed by the Human Proteome Organization Proteomics Standards Initiative makes it possible for PANDA to quantify the results of the commonly-used peptide identification tools, such as Mascot, SEQUEST, X! Tandem and MS-GF+. Meanwhile, PANDA can read the quality control results of PeptideProphet (Keller *et al.*, 2002) and PepDistiller (Li *et al.*, 2012), which further broadens its usage and flexibility. (ii) The function layer contains the current mainstream quantification methods. For label-free quantification, spectral count (SC) method and extracted-ion chromatography (XIC) (also named as intensity-based) method were implemented in PANDA. As to labeled quantification, PANDA supports the prevalent precursor ion labeling methods, i.e., SILAC, ^{18}O, ^{15}N, isotope-coded affinity tags (ICAT) and isotope-coded protein labels (ICPL), as well as product ion labeling methods, i.e. isobaric tag for relative and absolute quantitation (iTRAQ) and tandem mass tag (TMT). Furthermore, users can define their own labeling methods in PANDA. (iii) The algorithm layer includes the basic algorithms for MS data processing and peptide/protein quantification (Supplementary Note 1). Part of them are adapted from LFQuant and SILVER, such as the reversible retention time (RT) alignment algorithm in LFQuant, the multi-filters for XIC construction and the dynamic isotopic matching tolerance algorithm in SILVER.

3 Results

In this study, PANDA was compared with MaxQuant (v1.6.0.13, released on Aug 2017) on a yeast dataset with four concentration levels of UPS2 standard proteins spiked in (A–D groups) for label-free quantification and a large-scale HeLa dataset with SILAC labeling as well as several SILAC and TMT labeling phosphoproteomic datasets for labeled quantifications, respectively.

3.1 Accuracy evaluation

In the yeast dataset, the theoretical ratios of the spiked-in UPS2 proteins for A/B, A/C and A/D should be 5, 25 and 125. As shown in Supplementary Figure S1, the quantification results of PANDA were closer to the theoretical ratios than those of MaxQuant. In the HeLa dataset, the SILAC ratios of the 3471 proteins commonly quantified by PANDA and MaxQuant were shown in Supplementary Figure S2. The ratio distribution of PANDA was also closer to the theoretical ratio (1: 1) than that of MaxQuant. In the phosphoproteomic datasets, PANDA owns a similar accuracy compared with MaxQuant (Supplementary Figs S3 and S4). These results demonstrated PANDA has a high accuracy for both label-free and labeled quantifications in a wide dynamic range. Specially, another advantage of PANDA is that it can handle ^{15}N labeling data with high accuracy (Supplementary Fig. S5).

3.2 Precision evaluation

In the yeast dataset, PANDA showed a lower coefficient of variation (CV) distribution of the yeast proteins for the technical replicates within each group (A–D) than MaxQuant, indicating the high precision of PANDA for label-free quantification (Supplementary Fig. S6). In the HeLa dataset, the protein intensity CVs of the three technical replicates for both SILAC labeled and unlabeled samples were calculated and PANDA also displayed a lower CV distribution than MaxQuant, which proved that PANDA is precise for labeled quantification (Supplementary Fig. S7). More details are provided in Supplementary Notes 2-3.

Finally, PANDA is efficient due to the refinement of its source codes and the inclusion of popular third-party libraries, such as GNU scientific library. It spent less computation time than MaxQuant on all the datasets (Supplementary Table S1).

4 Conclusion

In summary, PANDA contains a comprehensive algorithm collection for label-free and labeled quantifications and supports all the main methods in quantitative proteomics. Being able to read proteomics data in public format, PANDA is very flexible and compatible with existing peptide identification tools or MS data analysis pipelines. Most importantly, PANDA is proved to be accurate and precise for label-free and labeled quantifications. Although PANDA can only run in Windows at present, other operating systems will be supported in the future. At last, the quantification results of PANDA can be further analyzed in its affiliated tool PANDA-view (Chang *et al.*, 2018) for statistical analysis and data visualization.

Funding

This work was supported by the National Key Research and Development Program of China [2017YFA0505002 and 2017YFC0906602] and the National Natural Science Foundation of China [21605159 and 21475150].

Conflict of Interest: none declared.

References

Arsova,B. *et al.* (2012) Precision, proteome coverage, and dynamic range of Arabidopsis proteome profiling using (15)N metabolic labeling and label-free approaches. *Mol. Cell Proteomics*, **11**, 619–628.

Chang,C. *et al.* (2018) PANDA-view: An easy-to-use tool for statistical analysis and visualization of quantitative proteomics data, *Bioinformatics*.

Chang,C. *et al.* (2014) SILVER: an efficient tool for stable isotope labeling LC-MS data quantitative analysis with quality control methods. *Bioinformatics*, **30**, 586–587.

Chang,C. *et al.* (2016) Quantitative and in-depth survey of the isotopic abundance distribution errors in shotgun proteomics. *Anal. Chem.*, **88**, 6844–6851.

Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.

Hogrebe,A. *et al.* (2018) Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nat. Commun.*, **9**, 1045.

Keller,A. *et al.* (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, **74**, 5383–5392.

Li,N. *et al.* (2012) PepDistiller: a quality control tool to improve the sensitivity and accuracy of peptide identifications in shotgun proteomics. *Proteomics*, **12**, 1720–1725.

Mitchell,C.J. *et al.* (2016) PyQuant: a versatile framework for analysis of quantitative mass spectrometry data. *Mol. Cell Proteomics*, **15**, 2829–2838.

Schubert,O.T. *et al.* (2017) Quantitative proteomics: challenges and opportunities in basic and applied research. *Nat. Protoc.*, **12**, 1289–1294.

Van Riper,S.K. *et al.* (2016) RIPPER: a framework for MS1 only metabolomics and proteomics label-free relative quantification. *Bioinformatics*, **32**, 2035–2037.

Zhang,W. *et al.* (2012) LFQuant: a label-free fast quantitative analysis tool for high-resolution LC-MS/MS proteomics data. *Proteomics*, **12**, 3475–3484.