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

Cheng Chang[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, P.R. China.*

[2]*Beijing Key Laboratory of Human Computer Interactions, Institute of Software, Chinese Academy of Sciences, Beijing 100190, P.R. China.*

Correspondence should be addressed to:

Cheng Chang (1987ccpacer@163.com)

Yunping Zhu (zhuyunping@gmail.com)

Supplementary Methods

Experimental datasets

The yeast dataset. To evaluate the accuracy and precision of label-free quantification, the yeast mass spectrometry (MS) data obtained from (Chang, et al., 2016) was used in this study. In this dataset, a series of UPS2 standard proteins (Proteomics Dynamic Range Standard, Sigma-Aldrich) with four levels of amounts (1µg, 0.2µg, 0.04µg, 0.008µg) were spiked into the yeast samples, named as A-D groups. The loading amount of yeast samples were equal in all the four groups. Each group contains three technique replicates. The UPS2 standard proteins consist of 48 synthesized proteins with six levels of concentrations, ranging from 5000 fmol to 0.05 fmol.

The HeLa dataset. This dataset was obtained by analyzing the MS data from (Cox and Mann, 2007) and used for evaluation of labeled quantification. The unlabeled HeLa cells were mixed with the same amount of SILAC labeled samples. The SILAC labels consist of lysine with +8.014199Da and arginine with +10.008269Da. The mixed samples were separated by isoelectric focusing into 24 fractions and analyzed by MS in triplicate.

Peptide identification

For PANDA, all MS data in the two datasets were re-analyzed at first. MS raw files were processed by msconvert (ProteoWizard suite version 3.0.11516) using the default parameters. The acquired MS/MS peak list files (MGF files) were searched by Mascot (version 2.3.2) search engine against the Swiss-Prot yeast database (release 2013_11) with 48 UPS2 standard proteins sequences for the yeast dataset and the Swiss-Prot human database (release 2013_06) for the HeLa dataset. The detailed search parameters were kept the same as the descriptions in their original papers.

For MaxQuant, all MS raw files were loaded into MaxQuant (v1.6.0.13) for peptide identification and quantification. MS data were searched by Andromeda (Cox, et al., 2011) against the same protein sequence databases mentioned above. The search parameters were also kept unchanged for a fair comparison of PANDA and MaxQuant.

Finally, for quality control of the peptide identification results, both peptide and protein false discovery rates (FDRs) were kept below 0.01 in this study.

Supplementary Notes

Quantification accuracy evaluation

To evaluate the quantification accuracy of PANDA and MaxQuant, the commonly quantified proteins by the two software tools were selected for further comparison. Before the selection, the technical replicates are merged at first in either dataset as follows:

$$\text{protein merged intensity} = \frac{\sum_{i=1}^{N} protein\ intensity_i}{N}$$

where N is the number of replicates with intensity larger than zero, $\text{protein intensity}_\text{i}$ is the protein intensity in the *i-th* replicate.

In the yeast dataset, there were 20 UPS2 proteins quantified in all the four groups

(A-D) by PANDA and MaxQuant. The theoretical ratios of these proteins for A/B, A/C and A/D are 5, 25, 125. As shown in Supplementary Figure 1a, PANDA showed a closer ratio distribution to the theoretical value than MaxQuant in all the situations. Moreover, when we split these proteins with their actual amounts, the results showed the similar trends (Supplementary Figure 1b-d). Note that the 20 UPS2 proteins consist of all the 16 proteins in 5000 fmol level and 500 fmol level, as well as four proteins in 50 fmol level. Due to the very few number of quantified proteins in 50 fmol level, the results from 50 fmol level are easily influenced by outliers and thus not so reliable. Therefore, only the proteins in 5000 fmol and 500 fmol levels were shown in boxplot in Supplementary Figure 1b-d.

In the HeLa dataset, after merging the three technical replicates, there were 3471 proteins both quantified by PANDA and MaxQuant. As shown in Supplementary Figure 2, the protein SILAC ratios of PANDA were significantly closer to the theoretical value (1:1) than those of MaxQuant with Kruskal-Wallis test *p-value*<0.001. The median ratio of these proteins quantified by PANDA is 0.85, while the median ratio of MaxQuant is 0.78.

Thus, we can conclude that PANDA owns a high accuracy for label-free and labeled quantifications in a wide dynamic range.

Quantification precision evaluation

In the yeast dataset, since the loading amount of the yeast samples remained the same in A-D groups as background, the coefficient of variations (CVs) of the three technical replicates within each group were calculated respectively to evaluate the precision of the technical replicates for label-free quantification. As shown in Supplementary Figure 3, there was an obvious difference between the protein CV distributions within each group using PANDA and MaxQuant indicating that PANDA is of high precision for label-free quantification.

In the HeLa dataset, there were 1905 and 1958 proteins quantified in all the three technical replicates with two or more spectral counts using PANDA and MaxQuant, respectively. Among them, a total of 1573 proteins were both quantified by the two software tools. As shown in Supplementary Figure 4, the CV distribution of PANDA is significantly closer to zero than that of MaxQuant for both heavy and light labeling samples (Kruskal-Wallis test *p-value*<0.001), proving that PANDA is also precise for labeled quantification.

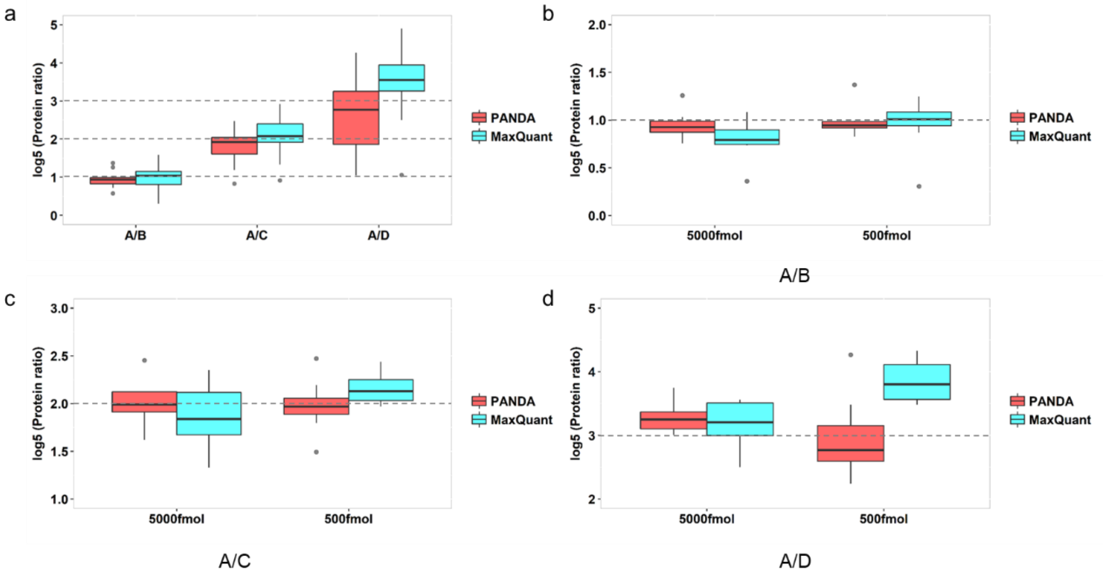
Supplementary References

Chang, C.*, et al.* (2016) Quantitative and In-Depth Survey of the Isotopic Abundance Distribution Errors in Shotgun Proteomics, *Anal Chem*, **88**, 6844-6851.

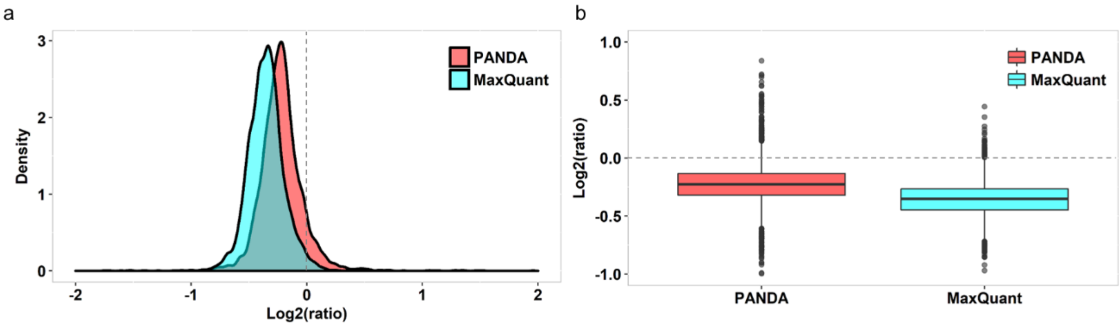
Cox, J. and Mann, M. (2007) Is proteomics the new genomics?, *Cell*, **130**, 395-398.

Cox, J.*, et al.* (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment, *J Proteome Res*, **10**, 1794-1805.

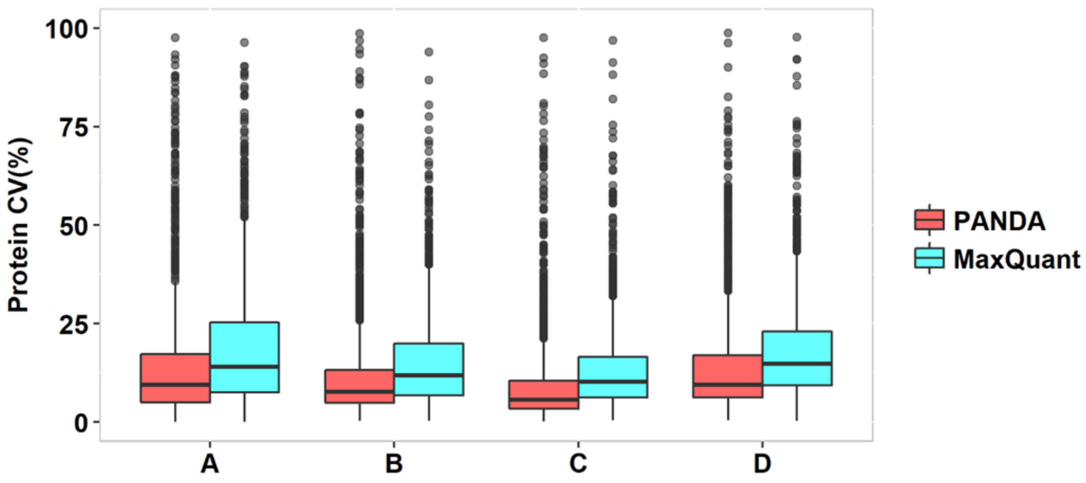
Supplementary Figures and Tables



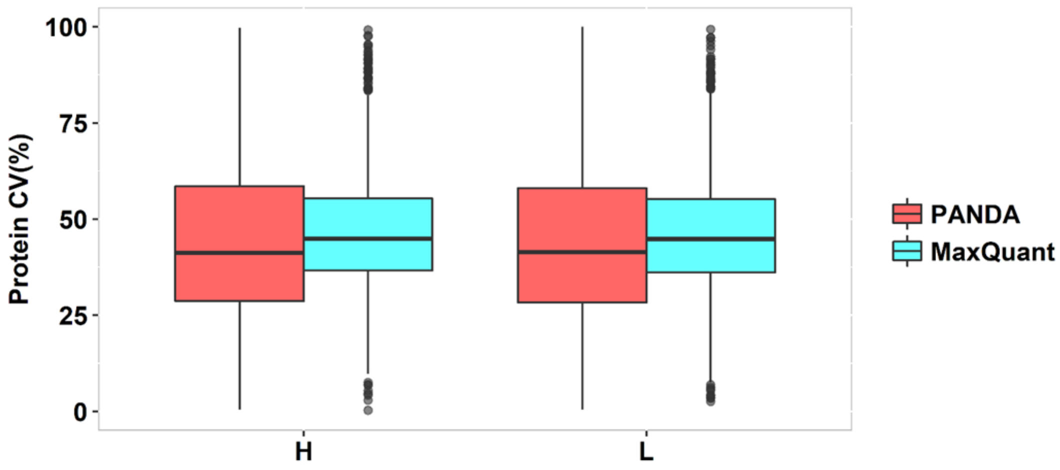
Supplementary Figure 1. Accuracy evaluation for label-free quantification. Boxplots of the spike-in UPS2 protein ratios between A-D groups in the yeast dataset using PANDA (red) and MaxQuant (cyan) separately. A-D indicate the four dilution concentrations of the UPS2 proteins spiked in the yeast samples. (a) Protein ratio boxplots of all the UPS2 proteins for A/B, A/C and A/D. (b-d) Protein ratio boxplots of UPS2 proteins with 5000 fmol and 500 fmol, respectively. The protein ratios are shown in base-5 logarithm scale. The gray dashed lines represent the theoretical ratios.



Supplementary Figure 2. Accuracy evaluation for labeled quantification. Distributions of the protein ratios between the SILAC labeled and unlabeled samples in the HeLa dataset using PANDA (red) and MaxQuant (cyan). (a) Density plots of the protein ratios commonly quantified by PANDA and MaxQuant. (b) Boxplots of the protein ratios commonly quantified by PANDA and MaxQuant. The protein ratios are shown in base-2 logarithm scale. The gray dashed lines represent the theoretical ratios.



Supplementary Figure 3. Precision evaluation for label-free quantification on the yeast dataset. Boxplots of yeast protein intensity CVs of the three technical replicates within each group (A-D). The red box indicates PANDA and the cyan one indicates MaxQuant.



Supplementary Figure 4. Precision evaluation for labeled quantification on the HeLa dataset. Boxplots of the human protein intensity CVs of the three replicates in the HeLa dataset. H and L indicate SILAC labeled (heavy) and unlabeled samples (light), respectively.

Supplementary Table 1. List of the quantification time for the yeast and HeLa datasets using PANDA and MaxQuant.

Software	Yeast dataset (12 raw files, ~8.9 GB) (minute)	HeLa dataset (72 raw files，~15.6 GB) (minute)
PANDA	16	27
MaxQuant (v1.6.0.13)	35	133

Note:

1) For MaxQuant, only the quantification time (i.e. starting from the step "Re-quantification" to the last step) was considered, not including the time for data searching and quality control.
2) PANDA and MaxQuant were tested using one thread on the same computer: Windows7 64-bit operating system, Intel Core E3–1230 v3 CPU 3.30-GHz processors, 2 TB SATA3 hard disk with 7200 rpm, and 8 GB RAM.