Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files

To the Editor: Discovery proteomics has limited quantification capabilities because of stochastic precursor-ion selection. Several data-independent acquisition (DIA) methods have been proposed to overcome this limitation[1–4], including the sequential-window acquisition of all theoretical mass spectra (SWATH-MS)[4].

We developed an untargeted analysis method named Group-DIA, which can analyze multiple DIA data files simultaneously. In contrast to DIA-Umpire[5], another untargeted analysis method recently published in *Nature Methods*, Group-DIA combines the elution profiles of precursor ions and fragment ions from all data files to determine precursor-fragment pairs. Those pairs make up pseudo-spectra that can be searched using conventional sequence database–searching software (**Fig. 1a**). The Group-DIA method includes the following main steps (**Supplementary Fig. 1** and **Supplementary Note 1**).

Retention-time alignment. The retention times of different data files are aligned first on the basis of the chromatographic signals extracted from MS1 spectra and then by the correlation coefficients of extracted ion chromatograms (XICs) of the product ions (**Supplementary Fig. 2**).

Similarity comparison and spectra generation. For each feature, all the possible product ions are determined, and the XICs of precursor ions and fragment ions are extracted. The target spectrum is generated through the selection of fragment ions with high similarity to the precursor ion. For the estimation of errors in the generation of the pseudo-spectrum, a decoy spectrum is generated by random selection of the same number of remaining fragment ions (**Supplementary Note 2**).

Peak rediscovery and interference removal. Group-DIA handles various data files in an experiment as a group to rediscover the peak and then performs quantification. It also removes interfering ions by comparing their transition intensities across all data files.

We compared the performance of Group-DIA and DIA-Umpire for analyzing the SWATH-MS Gold Standard (SGS) data set used by Röst *et al.*[6]. We diluted 422 stable isotope–labeled standard (SIS) peptides twofold in yeast or HeLa cell lysate in ten steps and then analyzed them in triplicate with SWATH-MS. We analyzed the resulting 60 DIA data files (termed the yeast and human SGS data sets) in parallel with DIA-Umpire and Group-DIA. The more data files that were analyzed, the better Group-DIA performed in identifying peptides compared with DIA-Umpire (**Fig. 1b** and **Supplementary Fig. 3**). We also used different search engines and various cutoff values in the analyses, and Group-DIA performed better than DIA-Umpire in each setting (**Supplementary Figs. 4** and **5**). About 90% of the peptides identified by DIA-Umpire were also identified by Group-DIA, but less than 60% of the peptides revealed by Group-DIA

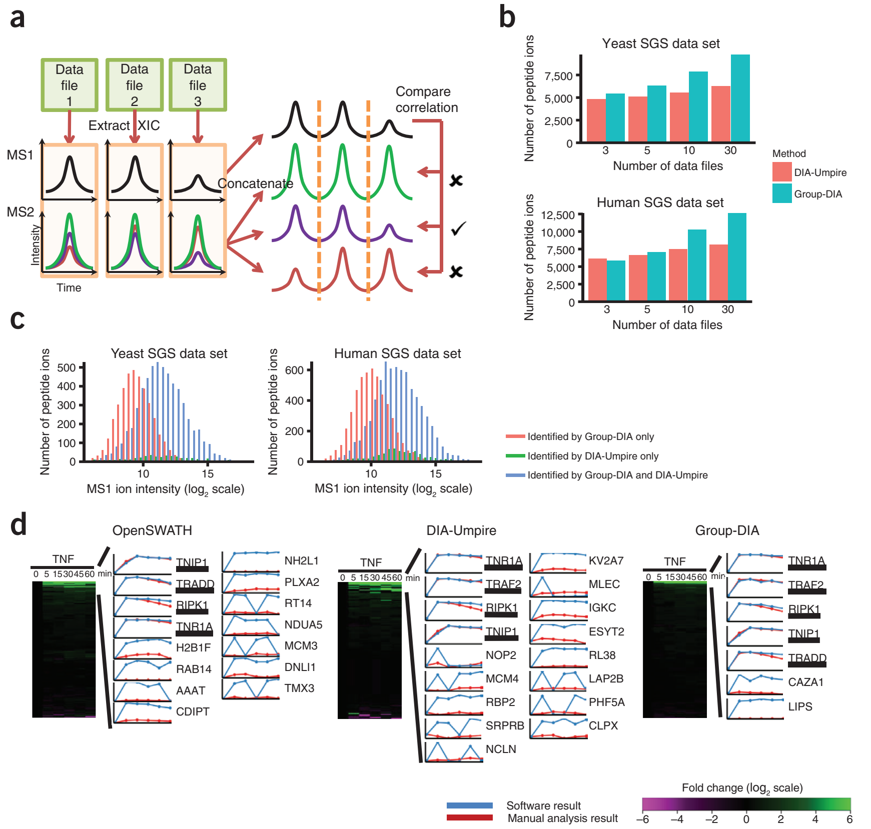

Figure 1 | The principle of Group-DIA and evaluation of its performance in analyzing DIA-MS data files. (**a**) Group-DIA pseudo-spectra–generating algorithms. (**b**) Comparison of the numbers of peptides identified by DIA-Umpire and Group-DIA in analyses of 3, 5, 10 and 30 data files from the yeast SGS data set and from the human SGS data set. The pseudo-spectra were searched by Mascot and validated by PeptideProphet, and then they were combined and rescored using iProphet. The identified peptides were filtered at an iProphet probability cutoff of 0.9. (**c**) Distribution of MS1 ion intensities of peptides identified by DIA-Umpire and Group-DIA from the yeast SGS data set and the human SGS data set. (**d**) Heat maps of the protein intensities quantified by Group-DIA, DIA-Umpire and OpenSWATH in analyses of the TNFR1 complex data set. Temporal profiles of the upregulated proteins identified by OpenSWATH, DIA-Umpire and Group-DIA are shown as blue lines in the plots, and manually checked results are shown as red lines. The intensities were normalized by an untreated control. The names of the proteins that we confirmed by manual check are underlined.

were identified by DIA-Umpire (**Supplementary Fig. 6** and **Supplementary Tables 1** and **2**). Nearly 70% of the additional peptides identified by Group-DIA were multiple hits (**Supplementary Fig. 7**). A comparison of peptide intensities suggested that Group-DIA was more efficient in identifying low-abundance peptides (**Fig. 1c**). We manually checked the XICs of all SIS peptides identified by Group-DIA (but not DIA-Umpire) and confirmed that they were true positives (**Supplementary Figs. 8–10** and **Supplementary Data**).

To prove the validity of the decoy spectra, we investigated their properties. Product-ion intensities of target and decoy spectra had similar distributions (**Supplementary Fig. 11**). Decoy spectra were mapped to target and decoy databases with similarly low confidence (**Supplementary Fig. 12**). Additionally, receiver operating characteristic plots suggested that target spectra could be distinguished from decoy spectra when they were mapped to the target database (**Supplementary Fig. 13**). These results suggested that decoy spectra could be used for error estimation in the generation of pseudo-spectra.

We also compared the quantification accuracy of Group-DIA with that of OpenSWATH[6], a targeted analysis strategy. SIS peptide intensities suggested that the two tools had similar quantification accuracy (**Supplementary Fig. 14**). However, Group-DIA obtained more consistent quantification data in replicates than OpenSWATH did (**Supplementary Fig. 15**).

We then evaluated the performance of Group-DIA in analyzing immunoprecipitation (IP) samples. We immunoprecipitated TNFR1 (tumor necrosis factor receptor 1) complex from L929 cells treated with TNF for six different time periods and analyzed these IP samples using shotgun MS to build a spectral library for OpenSWATH analysis and SWATH-MS for generating DIA files. Group-DIA identified more peptides than DIA-Umpire did (**Supplementary Fig. 16** and **Supplementary Table 3**). The majority of the peptides identified by these two workflows can be found in the spectral library (**Supplementary Fig. 17**, **Supplementary Table 4** and **Supplementary Data**). Temporal profiles of the proteins revealed by Group-DIA, DIA-Umpire and OpenSWATH are shown in **Figure 1d** and **Supplementary Table 5**. Comparison of the quantifications of replicates showed that Group-DIA was more consistent than OpenSWATH (**Supplementary Fig. 18**).

Group-DIA, DIA-Umpire and OpenSWATH revealed 7, 17 and 15 proteins, respectively, whose levels increased time-dependently in TNF IP (**Fig. 1d**). We performed a manual check, which showed that Group-DIA could identify more truly regulated proteins with less background noise than the other approaches could (**Fig. 1d** and **Supplementary Table 5**).

Finally, we compared the performance of Group-DIA with that of OpenSWATH in an analysis of SWATH-MS data from whole-cell lysates (**Supplementary Note 3**, **Supplementary Methods, Supplementary Figs. 19–21**, **Supplementary Tables 6–9** and **Supplementary Data**). We concluded that the two methods are essentially equivalent for analyzing DIA data from highly complex samples. About half of the hits obtained with both methods were false positives and needed to be removed via a manual check.

Group-DIA source code and documentation are available as **Supplementary Software** and at http://yuanyueli.github.io/group-dia/.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.3593).

ACKNOWLEDGMENTS

We thank Z. Xu and Y. Yu for help with the high-performance computer. This work was supported by the National Science Foundation (NSF) of China (grants 91429301 and 31221065), 973 Program 2015CB553800, National Major Project 2013ZX10002-002, 111 Project B12001, funding from Xiamen City (grant 3502Z20130027) and the NSF of China for Fostering Talents in Basic Research (grant J1310027).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Yuanyue Li[1,3], Chuan-Qi Zhong[1,3], Xiaozheng Xu[1], Shaowei Cai[1], Xiurong Wu[1], Yingying Zhang[1], Jinan Chen[1], Jianghong Shi[2], Shengcai Lin[1] & Jiahuai Han[1]

[1]State Key Laboratory of Cellular Stress Biology, Innovation Center for Cell Signaling Network, School of Life Sciences, Xiamen University, Xiamen, Fujian, China. [2]School of Information Science and Engineering, Xiamen University, Xiamen, Fujian, China. [3]These authors contributed equally to this work.
e-mail: andyzcq@gmail.com or jhan@xmu.edu.cn

PUBLISHED ONLINE 5 OCTOBER 2015; DOI:10.1038/NMETH.3593

1. Venable, J.D., Dong, M.Q., Wohlschlegel, J., Dillin, A. & Yates, J.R. *Nat. Methods* **1**, 39–45 (2004).
2. Plumb, R.S. *et al. Rapid Commun. Mass Spectrom.* **20**, 1989–1994 (2006).
3. Panchaud, A. *et al. Anal. Chem.* **81**, 6481–6488 (2009).
4. Gillet, L.C. *et al. Mol. Cell. Proteomics* **11**, 0111.016717 (2012).
5. Tsou, C.C. *et al. Nat. Methods* **12**, 258–264, (2015).
6. Röst, H.L. *et al. Nat. Biotechnol.* **32**, 219–223 (2014).