Published in final edited form as:
J Proteome Res. 2020 January 03; 19(1): 537–542. doi:10.1021/acs.jproteome.9b00328.

ThermoRawFileParser: modular, scalable and cross-platform RAW file conversion

Niels Hulstaert[1,2], **Jim Shofstahl**[3], **Timo Sachsenberg**[4], **Mathias Walzer**[5], **Harald Barsnes**[6,7], **Lennart Martens**[1,2], **Yasset Perez-Riverol**[5]

[1]VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium

[2]Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

[3]Thermo Fisher Scientific, 355 River Oaks Parkway San Jose, California 95134, United States

[4]Applied Bioinformatics, Department for Computer Science, University of Tuebingen, Sand 14, 72076 Tuebingen, Germany

[5]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

[6]Computational Biology Unit (CBU), Department of Informatics, University of Bergen, Norway

[7]Proteomics Unit (PROBE), Department of Biomedicine, University of Bergen, Norway

Abstract

The field of computational proteomics is approaching the big data age, driven both by a continuous growth in the number of samples analysed *per* experiment, as well as by the growing amount of data obtained in each analytical run. In order to process these large amounts of data, it is increasingly necessary to use elastic compute resources such as Linux-based cluster environments and cloud infrastructures. Unfortunately, the vast majority of cross-platform proteomics tools are not able to operate directly on the proprietary formats generated by the diverse mass spectrometers. Here, we present ThermoRawFileParser, an open-source, cross-platform tool that converts Thermo RAW files into open file formats such as MGF and the HUPO-PSI standard file format mzML. To ensure the broadest possible availability, and to increase integration capabilities with popular workflow systems such as Galaxy or Nextflow, we have also built Conda package and BioContainers container around ThermoRawFileParser. In addition, we implemented a user-friendly interface (ThermoRawFileParserGUI) for those users not familiar with command-line tools. Finally, we performed a benchmark of ThermoRawFileParser and msconvert to verify that the converted mzML files contain reliable quantitative results.

Introduction

The field of computational proteomics is approaching the big data age (1), driven both by a continuous growth in the number of samples analysed *per* experiment, as well as by the growing amount of data obtained in each analytical run. At the same time, more data is now publicly available in proteomics repositories, which in turn means that there is increasing benefit to be had from the reanalysis of millions of mass spectra (2–5) to find new biological insights (e.g. novel variants and post-translational modifications (6)). However, in order to process these large amounts of (public) data, it is increasingly necessary to use elastic compute resources such as Linux-based cluster environments and cloud infrastructures (7).

The development of computational proteomics tools has historically been favoured the Microsoft Windows operating systems with tools such as ProteomeDiscover, MaxQuant (8), PeaksDB and Mascot Distiller (9). An important driver for this bias has been the lack of cross-platform libraries to access instrument output data files (RAW files) from major instrument providers (10). Several approaches have been devised to overcome this challenge, including the use of dedicated Windows machines in workflows (11) for conversion to RAW data to standard file formats such as mzML (12), the encapsulation of Windows tools such as ReAdW (13) and msconvert (14) into WineHQ (http://tools.proteomecenter.org/wiki/index.php?title=Msconvert_Wine) to make these tools Linux-compatible, and even the creation of reverse-engineered RAW file readers (15).

An important breakthrough was achieved in 2016, when Thermo Scientific released the first cross-platform application programming interface (API) that enables access to Thermo RAW files from all their instruments on all commonly used operating systems (e.g. Linux/Unix, Mac OX or Microsoft Windows). Importantly, this provides the enticing possibility to move proteomics into Linux/UNIX environments, including scalable clusters and cloud environments. This library has already led to a new version of the popular MaxQuant framework that is compatible with Linux/UNIX environments (16), and it has also been incorporated into the cross-platform, cluster-oriented quantification tool moFF (17).

While the Thermo cross-platform library thus enables specially-developed software to access Thermo Raw files on diverse operating systems, most open-source computational proteomics workflows (e.g. OpenMS (18), Galaxy (19), and the Trans-Proteomics pipeline (TPP) (20)) are based on generic, open data formats such as Mascot Generic File (MGF) or mzML. In order to allow these tools to benefit maximally from the cross-platform access to Thermo Raw files, we here present ThermoRawFileParser, an open-source, cross-platform tool that converts Thermo RAW files into open file formats such as MGF and mzML similar to other tools such as msconvert (14) and RawTools (21). To ensure the broadest possible availability, and to increase integration capabilities with popular workflow systems such as Galaxy (22) or Nextflow (23), we have also built a Conda package (24) and a BioContainers (25) container around ThermoRawFileParser. Finally, we performed a benchmark of ThermoRawFileParser and msconvert to verify that the converted mzML files contain reliable quantitative results.

Materials

Tool Design and Integration

ThermoRawFileParser (https://github.com/compomics/ThermoRawFileParser) has been implemented following a modular design (Figure 1). Every file specific exporter is implemented as an independent module, which enables easy extension to include more exporters in the future. Currently, the tool can export to MGF (**MGFSpectrumWriter**), mzML (**MzMLSpectrumWriter**), and JSON (for the metadata only) (**MetadataWriter**). This modular design has already enabled the community to extend the library for other novel file formats such as Parquet (**ParquetSpectrumWriter**), which is designed for distributed big data processing clusters of Hadoop or Spark. The JSON export of ThermoRawFileParser can optionally be used to only extract various metadata elements (including instrument settings and scan settings; see https://github.com/PRIDE-Archive/pride-metadata-standard) (Supplementary Note 1). This specific feature is currently used by the PRIDE Database to re-annotate thousands of RAW files with the correct instrument metadata. For peak picking, data centroiding, and noise removal, ThermoRawFileParser relies on the native methods provided by the Thermo API.

A key feature of any open-source tool is its ability to integrate with other frameworks (26). We have therefore created a BioConda recipe (24) for ThermoRawFileParser (https://github.com/bioconda/bioconda-recipes/tree/master/recipes/thermorawfileparser), which can be used to automatically build a Docker Container. This Docker is pushed to the BioContainer project (25), which in turn enables easy reuse of the tool by both the Galaxy (22) and the Nextflow (23) environments. As an illustration of such integration, we have developed a Nextflow workflow for the proteomics community, which converts an entire ProteomeXchange project using the ThermoRawFileParser container (https://github.com/bigbio/nf-workflows).

In addition to the command-line tool, we have implemented a graphical user interface that makes the use of ThermoRawFileParser easier and highly intuitive, enabling the user to perform conversions of RAW files. The GUI includes all main options of ThermoRawFileParser, and a report system to report errors during the conversion. ThermoRawFileParserGUI is an open source Java program, available in a cross-platform package that incorporates ThermoRawFileParser executables for the main operating systems. It can be downloaded from https://github.com/compomics/ThermoRawFileParserGUI.

Benchmark datasets

Four different datasets and three different Thermo models were used to compare the conversion from RAW files into mzML with the ProteoWizard msconvert tool and the ThermoRawFileParser: PXD014195 and PXD006336 (Orbitrap Q-Exactive), PXD014346 (Orbitrap Fusion Lumos), PXD014772 (Orbitrap Velos).

IPRG-2015 dataset

We used the IPRG2015 dataset (https://www.ebi.ac.uk/pride/archive/projects/PXD010981) (27) to benchmark the quality of the mzML files produced by ThermoRawFileParser. This dataset is based on four artificially constructed samples of known composition, each containing a constant background of 200ng of tryptic digests of *S. cerevisiae* (ATCC strain 204508/S288c). Each sample was separately spiked in with different quantities of six individual protein digests. Samples were analysed in three LC-MS/MS using a Thermo Scientific Q-Exactive mass spectrometer (12 runs). Both MS and MS/MS data were acquired in profile mode in the Orbitrap, with resolution 70 000 for MS and 17 500 for MS/MS. The MS1 scan range was 300-1650 m/z, the normalized collision energy was set to 27%, and singly charged ions were excluded (27).

Benchmark analysis workflows

Identification-free workflow—We built a Nextflow (23) identification-free workflow using OpenMS (28) to benchmark different metrics such as: number of spectra MS1/MS2, number of peaks by spectrum at MS1 and MS2 levels and the charge state distribution (https://github.com/bigbio/nf-workflows/tree/master/qc-idfree_from_raw).

Identification workflow—To perform the benchmarking, we built a workflow using OpenMS (18, 28) in which raw files were converted from Thermo Scientific RAW files to mzML using the msconvert tool from ProteoWizard (14) on the one hand, and with ThermoRawFileParser on the other hand. The resulting spectra were centroided and searched using MS-GF+ (v2018.01.30) (29), executed via the OpenMS search engine wrapper MSGFPlusAdapter, allowing 10 ppm precursor mass tolerance, and setting carbamidomethylation of cysteine as fixed, and methionine oxidation as variable modification. All other settings were kept at their default values. PSMs were filtered (q-value < 5%). The workflow for comparison was developed using Nextflow (23) and BioContainers (25) to ensure the reproducibility of the present results (https://github.com/bigbio/nf-workflows/tree/master/benchmark-converter-nf).

Results and Discussion

We compare msconvert and ThermoRawFileParser conversion to mzML using four different metrics: Number of spectra MS1 and MS2, number of peaks by MS1 and MS2, and the precursor charge distribution (Figure 1).

The results show major differences between both tools with regards to the number of peaks reported, and this on each MS level (MS1 and MS2). On average, the number of peaks *per* spectrum is ten-fold higher for msconvert mzML files as compared to ThermoRawFileParser mzML files. This occurs because the new peak picking method implemented in the Thermo API used by ThermoRawFileParser improved drastically with regards to the removal of noise peaks that do not contribute to identification. As a result, despite the substantial difference in the number of peaks retained, there is no major difference in the identification map and precursor charge distribution (Supplementary Figure 1) between the tools.

Table 1 shows the number of MS1 and MS2 spectra, and the number of identified peptides and proteins for both workflows. The number of MS1 and MS2 in the mzML files were the same for all RAW files converted with msconvert and ThermoRawFileParser. Across all samples and replicates, the number of identified peptides and proteins is higher for the ThermoRawFileParser workflow when compared to the msconvert workflow, despite the abovementioned higher number of peaks retained in the msconvert workflow. This identification advantage for ThermoRawFileParser derived mzML files amounts to 10% on average at the peptide level, and to 4% on average at the protein level (Table 1). Benchmarking protein quantification between both approaches shows no major differences between the two approaches (Figure 2).

As a final benchmark, we analysed the IPRG2015 dataset to verify whether the mzML files obtained by the ThermoRawFileParser pipeline could replicate the quantification of the spike-in proteins in the sample using the approach described in the original publication (27). The results show that there is no appreciable difference between the IPRG 2015 analysis and the results from the ThermoRawFileParser workflow (Figure 2).

In addition to msconvert, the recently published RawTools (21) allows to convert RAW files into MGF files. In addition, it provides multiple options to perform QC metrics. However, RawTools is not design as a conversion tool and does not provides support for standard HUPO-PSI file formats such as mzML.

Conclusions

ThermoRawFileParser is an open-source software tool for the conversion of Thermo Raw files into open formats. Because of the growing need for more scalable and distributed computational proteomics approaches, ThermoRawFileParser has been designed to easily plug into large-scale workflow systems such as Galaxy, Nextflow, or OpenMS. The current implementation also provides support for native writing into Amazon web service object stores (S3), making the tool highly portable to cloud architectures. Finally, the modular design of the library, along with its open source nature, allows other researchers to contribute to and extend ThermoRawFileParser for new file formats in the future. Benchmarking tests on gold standard datasets against the ProteoWizard exporter show major improvements in peak detection, and noticeable increases in peptide and protein identifications while maintaining quantitative accuracy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by one ELIXIR Implementation Study. ELIXIR-EXCELERATE is funded by the European Commission within the Research Infrastructures programme of Horizon 2020, grant agreement numbers 676559. YP-R acknowledge the Wellcome Trust (grant number 208391/Z/17/Z) and the EPIC-XS project (grant number 823839), funded by the Horizon 2020 programme of the European Union. HB is supported by the Bergen Research Foundation and the Research Council of Norway.

References

1. Lam MP, Lau E, Ng DC, Wang D, Ping P. Cardiovascular proteomics in the era of big data: experimental and computational advances. Clin Proteomics. 2016; 13:23. [PubMed: 27980500]

2. Martens L, Vizcaino JA. A Golden Age for Working with Public Proteomics Data. Trends Biochem Sci. 2017; 42(5):333–341. [PubMed: 28118949]

3. Vaudel M, Verheggen K, Csordas A, Raeder H, Berven FS, Martens L, Vizcaino JA, Barsnes H. Exploring the potential of public proteomics data. Proteomics. 2016; 16(2):214–25. [PubMed: 26449181]

4. Griss J, Perez-Riverol Y, Lewis S, Tabb DL, Dianes JA, del-Toro N, Rurik M, Walzer M, Kohlbacher O, Hermjakob H. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. Nature Methods %@. 2016:1548–7091.

5. Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaino JA. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics. 2015; 15(5-6):930–49. [PubMed: 25158685]

6. Griss J, Perez-Riverol Y, Lewis S, Tabb DL, Dianes JA, Del-Toro N, Rurik M, Walzer MW, Kohlbacher O, Hermjakob H, Wang R, et al. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. Nat Methods. 2016; 13(8):651–656. [PubMed: 27493588]
7. Verheggen K, Barsnes H, Martens L. Distributed computing and data storage in proteomics: many hands make light work, and a stronger memory. Proteomics. 2014; 14(4-5):367–77. [PubMed: 24285552]
8. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008; 26(12):1367–72. [PubMed: 19029910]
9. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. ELECTROPHORESIS: An International Journal. 1999; 20(18):3551–3567.
10. Martens L, Nesvizhskii AI, Hermjakob H, Adamski M, Omenn GS, Vandekerckhove J, Gevaert K. Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. Proteomics. 2005; 5(13):3501–5. [PubMed: 16041670]
11. Verheggen K, Maddelein D, Hulstaert N, Martens L, Barsnes H, Vaudel M. Pladipus Enables Universal Distributed Computing in Proteomics Bioinformatics. J Proteome Res. 2016; 15(3):707–12. [PubMed: 26510693]
12. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Rompp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, et al. mzML--a community standard for mass spectrometry data. Mol Cell Proteomics. 2011; 10(1):R110.
13. Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, et al. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol. 2004; 22(11):1459–66. [PubMed: 15529173]
14. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol. 2012; 30(10):918–20. [PubMed: 23051804]
15. Kelchtermans P, Silva AS, Argentini A, Staes A, Vandenbussche J, Laukens K, Valkenborg D, Martens L. Open-Source, Platform-Independent Library and Online Scripting Environment for Accessing Thermo Scientific RAW Files. J Proteome Res. 2015; 14(11):4940–3. [PubMed: 26477298]
16. Sinitcyn P, Tiwary S, Rudolph J, Gutenbrunner P, Wichmann C, Yilmaz S, Hamzeiy H, Salinas F, Cox J. MaxQuant goes Linux. Nat Methods. 2018; 15(6):401. [PubMed: 29855570]
17. Argentini A, Staes A, Gruning B, Mehta S, Easterly C, Griffin TJ, Jagtap P, Impens F, Martens L. Update on the moFF Algorithm for Label-Free Quantitative Proteomics. J Proteome Res. 2019; 18(2):728–731. [PubMed: 30511867]
18. Pfeuffer J, Sachsenberg T, Alka O, Walzer M, Fillbrunn A, Nilse L, Schilling O, Reinert K, Kohlbacher O. OpenMS - A platform for reproducible analysis of mass spectrometry data. J Biotechnol. 2017; 261:142–148. [PubMed: 28559010]
19. Chambers MC, Jagtap PD, Johnson JE, McGowan T, Kumar P, Onsongo G, Guerrero CR, Barsnes H, Vaudel M, Martens L, Gruning B, et al. An Accessible Proteogenomics Informatics Resource for Cancer Researchers. Cancer Res. 2017; 77(21):e43–e46. [PubMed: 29092937]
20. Deutsch EW, Mendoza L, Shteynberg D, Slagel J, Sun Z, Moritz RL. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. Proteomics Clin Appl. 2015; 9(7-8):745–54. [PubMed: 25631240]
21. Kovalchik KA, Colborne S, Spencer SE, Sorensen PH, Chen DDY, Morin GB, Hughes CS. RawTools: Rapid and Dynamic Interrogation of Orbitrap Data Files for Mass Spectrometer System Management. J Proteome Res. 2019; 18(2):700–708. [PubMed: 30462513]
22. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Gruning BA, Guerler A, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018; 46(W1):W537–W544. [PubMed: 29790989]

23. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017; 35(4):316–319. [PubMed: 28398311]

24. Gruning B, Dale R, Sjodin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Koster J, Bioconda T. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018; 15(7):475–476. [PubMed: 29967506]

25. da Veiga Leprevost F, Gruning BA, Alves Aflitos S, Rost HL, Uszkoreit J, Barsnes H, Vaudel M, Moreno P, Gatto L, Weber J, Bai M, et al. BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics. 2017; 33(16):2580–2582. [PubMed: 28379341]

26. Wang R, Perez-Riverol Y, Hermjakob H, Vizcaino JA. Open source libraries and frameworks for biological data visualisation: a guide for developers. Proteomics. 2015; 15(8):1356–74. [PubMed: 25475079]

27. Choi M, Eren-Dogu ZF, Colangelo C, Cottrell J, Hoopmann MR, Kapp EA, Kim S, Lam H, Neubert TA, Palmblad M, Phinney BS, et al. ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC-MS/MS Experiments. J Proteome Res. 2017; 16(2):945–957. [PubMed: 27990823]

28. Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Andreotti S, Ehrlich HC, Gutenbrunner P, Kenar E, Liang X, et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods. 2016; 13(9):741–8. [PubMed: 27575624]

29. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun. 2014; 5

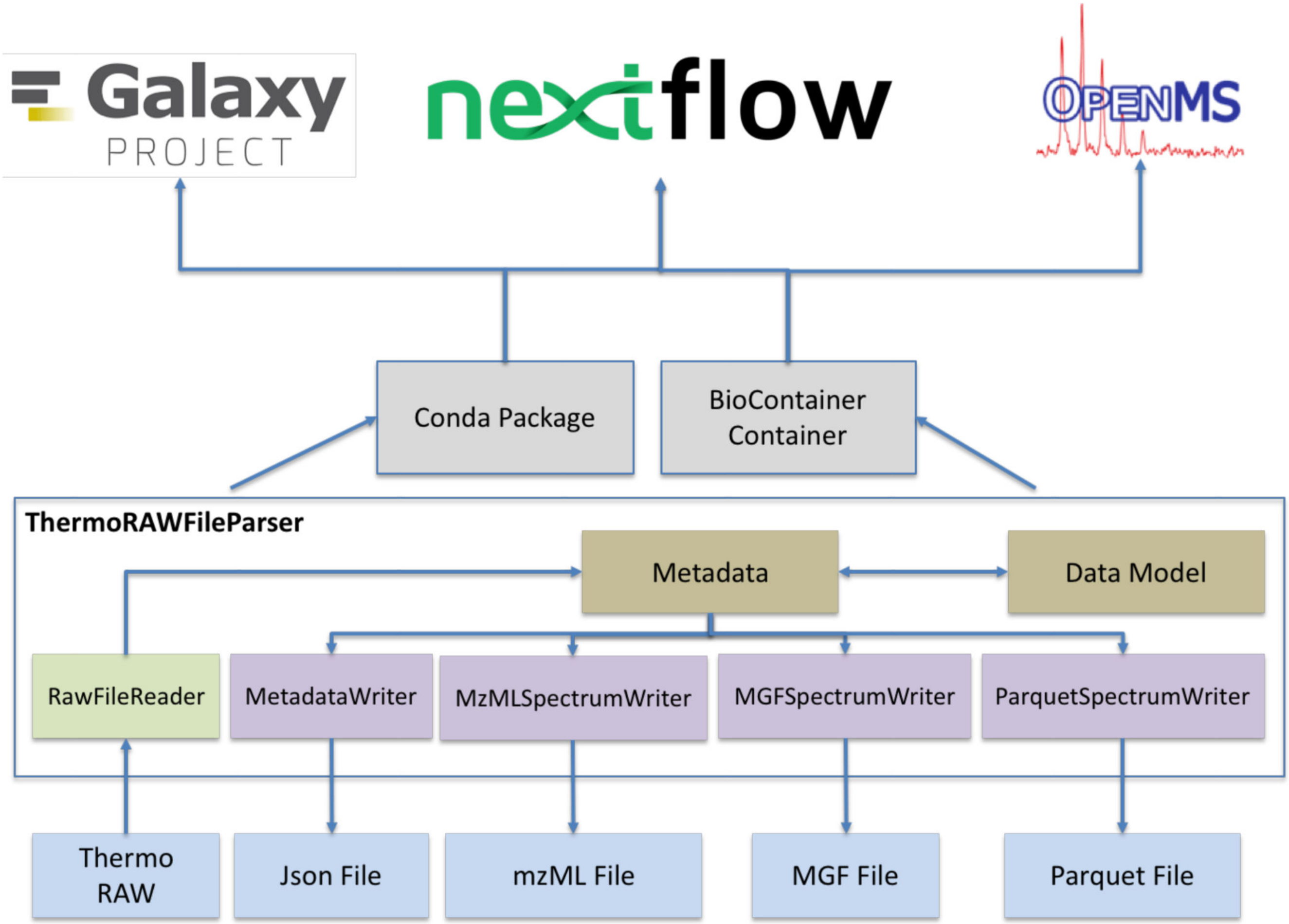

Figure 1. Modular design of ThermoRawFileParser includes exporters to MGF, mzML, Parquet, and Json Metadata. A Conda package and corresponding BioContainer is available for reuse in workflow engines such as Nextflow, Galaxy and OpenMS.

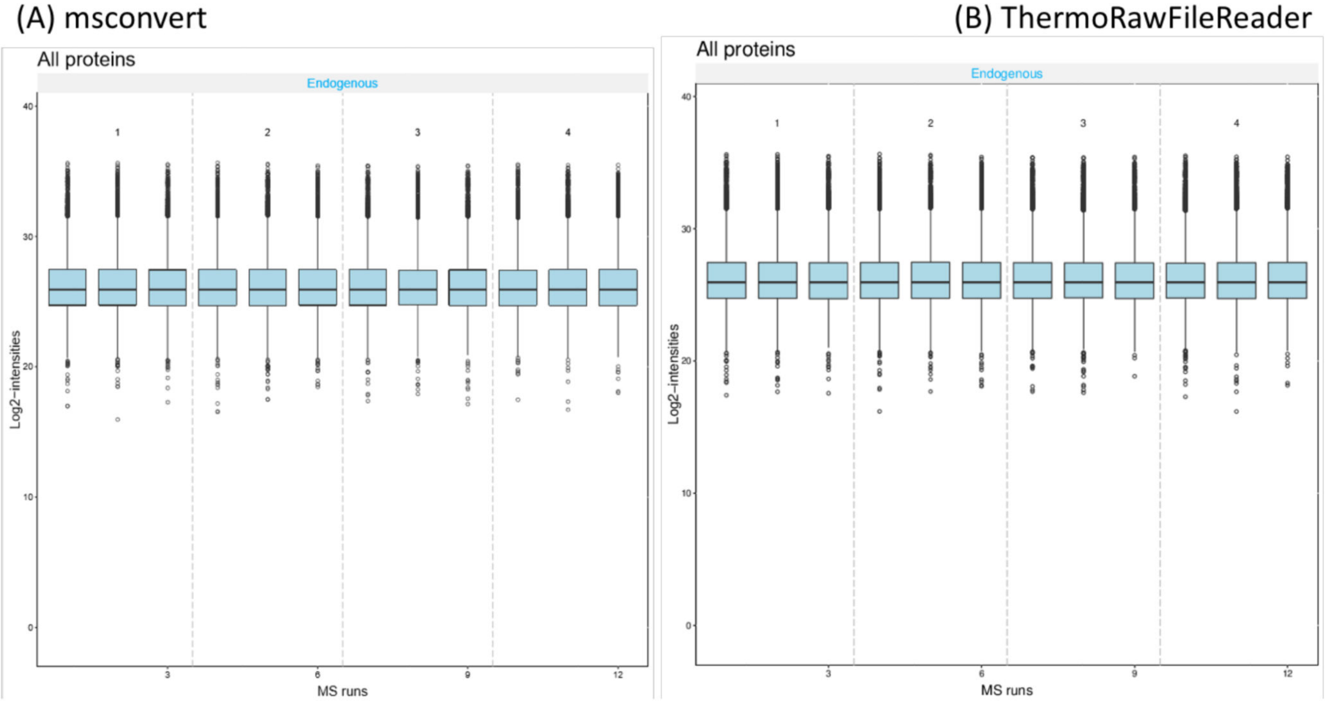

Figure 2. Log2-transformed intensity boxplot for four samples and twelve technical replicates for (A) msconvert-derived mzML files, and (B) ThermoRawFileParser-derived mzML files.

Table 1

Statistics on number of spectra at MS1 and MS2 level, identified peptides, and proteins for each MS run and software workflow (ThermoRawFileParser and msconvert).

		ThermoRawFileParser	**msconvert**
MS1 spectra count	**Sample 1A**	7787	7787
MS2 spectra count		49514	49514
Total number of peptides		24377	21727
Total number of proteins		3686	3607
MS1 spectra count	**Sample 1B**	7764	7764
MS2 spectra count		49633	49633
Total number of peptides		25435	22174
Total number of proteins		3815	3591
MS1 spectra count	**Sample 1C**	7802	7802
MS2 spectra count		49334	49334
Total number of peptides		23075	20034
Total number of proteins		3573	3442
MS1 spectra count	**Sample 2A**	7812	7812
MS2 spectra count		49293	49293
Total number of peptides		22736	20422
Total number of proteins		3555	3452
MS1 spectra count	**Sample 2B**	7740	7740
MS2 spectra count		49766	49766
Total number of peptides		24576	21323
Total number of proteins		3685	3542
MS1 spectra count	**Sample 2C**	7796	7796
MS2 spectra count		49455	49455
Total number of peptides		23174	20188
Total number of proteins		3597	3452
MS1 spectra count	**Sample 3A**	7702	7702
MS2 spectra count		49905	49905
Total number of peptides		22966	20564
Total number of proteins		3573	3489
MS1 spectra count	**Sample 3B**	7636	7636
MS2 spectra count		50417	50417
Total number of peptides		24445	21552
Total number of proteins		3684	3588
MS1 spectra count	**Sample 3C**	7806	7806
MS2 spectra count		49657	49657
Total number of peptides		23906	20311
Total number of proteins		3645	3436

		ThermoRawFileParser	**msconvert**
MS1 spectra count	**Sample 4A**	7757	7757
MS2 spectra count		49592	49592
Total number of peptides		21998	19691
Total number of proteins		3452	3366
MS1 spectra count	**Sample 4B**	7713	7713
MS2 spectra count		49930	49930
Total number of peptides		23902	21104
Total number of proteins		3597	3461
MS1 spectra count	**Sample 4C**	7791	7791
MS2 spectra count		49589	49589
Total number of peptides		22683	19348
Total number of proteins		3532	3364