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,*], and 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-P (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) (**Figure 2**). 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/tree/master/thermo-convert-nf).

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 (**Figure 3**). The GUI includes all main options of ThermoRawFileParser, and a report system to report errors during the conversion. ThermoRawFileParserGUI (**Figure 3**) 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

Three different public Thermo datasets were used to compare the conversion from RAW files into mzML with the ProteoWizard msconvert tool and the ThermoRawFileParser: PXD006336 (Orbitrap Q-Exactive), PXD014346 (Orbitrap Fusion Lumos), PXD001502 (Orbitrap Velos). We used a Nextflow workflow and the Identification-free OpenMS quality control (27) tools to benchmark different metrics such as: Number of spectra MS1/MS2, number of peaks by spectrum (https://github.com/bigbio/nf-workflows/tree/master/qc-idfree_from_raw).

We used the IPRG2015 dataset (https://www.ebi.ac.uk/pride/archive/projects/PXD006336) (28) 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 (28).

Quantification workflow

To perform the quantification benchmark using the PXD006336 dataset, we built a workflow using OpenMS (18, 27) in which raw files were converted from Thermo Scientific RAW files to mzML using ThermoRawFileParser tool. The resulting spectra were 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. PSMs were filtered (q-value < 5%) and used for feature detection using the semi-targeted approach implemented in the OpenMS tool FeatureFinderIdentification (30). Prior to identification, nonlinear retention time alignment was performed using the MapAlignerIdentification and the identified proteins were then quantified using unique peptides only (31). 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 MS1, number of MS2, MS1 peak count distribution, MS2 peak count distribution, identification map, and the precursor charge distribution. We observed no major differences between both tools (msconvert and ThermoRawFileParser) for the number of MS1/MS2 and the peak count distributions

(PXD006336 – **Supplementary Information S1**, PXD014346 – **Supplementary Information S2**, PXD001502 – **Supplementary Information S3**).

We analysed the IPRG2015 (28) dataset (PXD006336) using OpenMS framework and MSstats (32). We conducted a high-level analysis of 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 (28). In Figure 4, high values correspond to statistically significant changes. The x-axis is the log2-fold change between two samples, which in statistical language is sometimes called the “practical significance” of a change. Similar to best workflows reported in Choi *et. al*. (28), the estimates of log2 fold changes among the spiked proteins (A-F) were close to the true values, while most background proteins did not show significant differential expression. We computed the number of false positive (FP=3), true positive (TP=27) and positive predictive value (PPV=0.9) as defined by Choi et. al. (32). The results are within the 10 tops analysis performed in Choi et. al. for intensity-base methods. We performed the same analysis using msconvert to transform RAW data to mzML (**Supplementary Information S4, Figure 1 and Figure 2**). The number of peptides and proteins identified with both workflows (msconvert and ThermoRawFileParser) were similar.

In addition to msconvert, the recently published RawTools (21) allows to convert RAW files into MGF files. 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 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.

Acknowledgements

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. TS was supported by a grant from the German Federal Ministry of Education and Research (BMBF) under grant no. 031A535A (German Network for Bioinformatics, de.NBI/CIBI). We thank Meena Choi for the review of our data.

References

1. Lam, M. P.; Lau, E.; Ng, D. C.; Wang, D.; Ping, P., Cardiovascular proteomics in the era of big data: experimental and computational advances. *Clin Proteomics* **2016,** 13, 23.
2. Martens, L.; Vizcaino, J. A., A Golden Age for Working with Public Proteomics Data. *Trends Biochem Sci* **2017,** 42, (5), 333-341.
3. Vaudel, M.; Verheggen, K.; Csordas, A.; Raeder, H.; Berven, F. S.; Martens, L.; Vizcaino, J. A.; Barsnes, H., Exploring the potential of public proteomics data. *Proteomics* **2016,** 16, (2), 214-25.
4. Griss, J.; Perez-Riverol, Y.; Lewis, S.; Tabb, D. L.; Dianes, J. A.; 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 %@ 1548-7091* **2016**.
5. Perez-Riverol, Y.; Alpi, E.; Wang, R.; Hermjakob, H.; Vizcaino, J. A., Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* **2015,** 15, (5-6), 930-49.
6. Griss, J.; Perez-Riverol, Y.; Lewis, S.; Tabb, D. L.; Dianes, J. A.; Del-Toro, N.; Rurik, M.; Walzer, M. W.; Kohlbacher, O.; Hermjakob, H.; Wang, R.; Vizcaino, J. A., Recognizing

millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat Methods* **2016,** 13, (8), 651-656.
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.
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.
9. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S., 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, A. I.; Hermjakob, H.; Adamski, M.; Omenn, G. S.; 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.
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.
12. Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Rompp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P. A.; Deutsch, E. W., mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics* **2011,** 10, (1), R110 000133.
13. Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* **2004,** 22, (11), 1459-66.
14. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **2012,** 30, (10), 918-20.
15. Kelchtermans, P.; Silva, A. S.; 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.
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.
17. Argentini, A.; Staes, A.; Gruning, B.; Mehta, S.; Easterly, C.; Griffin, T. J.; Jagtap, P.; Impens, F.; Martens, L., Update on the moFF Algorithm for Label-Free Quantitative Proteomics. *J Proteome Res* **2019,** 18, (2), 728-731.

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.
19. Chambers, M. C.; Jagtap, P. D.; Johnson, J. E.; McGowan, T.; Kumar, P.; Onsongo, G.; Guerrero, C. R.; Barsnes, H.; Vaudel, M.; Martens, L.; Gruning, B.; Cooke, I. R.; Heydarian, M.; Reddy, K. L.; Griffin, T. J., An Accessible Proteogenomics Informatics Resource for Cancer Researchers. *Cancer Res* **2017,** 77, (21), e43-e46.
20. Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L., Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin Appl* **2015,** 9, (7-8), 745-54.
21. Kovalchik, K. A.; Colborne, S.; Spencer, S. E.; Sorensen, P. H.; Chen, D. D. Y.; Morin, G. B.; Hughes, C. S., RawTools: Rapid and Dynamic Interrogation of Orbitrap Data Files for Mass Spectrometer System Management. *J Proteome Res* **2019,** 18, (2), 700-708.
22. Afgan, E.; Baker, D.; Batut, B.; van den Beek, M.; Bouvier, D.; Cech, M.; Chilton, J.; Clements, D.; Coraor, N.; Gruning, B. A.; Guerler, A.; Hillman-Jackson, J.; Hiltemann, S.; Jalili, V.; Rasche, H.; Soranzo, N.; Goecks, J.; Taylor, J.; Nekrutenko, A.; Blankenberg, D., The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* **2018,** 46, (W1), W537-W544.
23. Di Tommaso, P.; Chatzou, M.; Floden, E. W.; Barja, P. P.; Palumbo, E.; Notredame, C., Nextflow enables reproducible computational workflows. *Nat Biotechnol* **2017,** 35, (4), 316-319.
24. Gruning, B.; Dale, R.; Sjodin, A.; Chapman, B. A.; Rowe, J.; Tomkins-Tinch, C. H.; Valieris, R.; Koster, J.; Bioconda, T., Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* **2018,** 15, (7), 475-476.
25. da Veiga Leprevost, F.; Gruning, B. A.; Alves Aflitos, S.; Rost, H. L.; Uszkoreit, J.; Barsnes, H.; Vaudel, M.; Moreno, P.; Gatto, L.; Weber, J.; Bai, M.; Jimenez, R. C.; Sachsenberg, T.; Pfeuffer, J.; Vera Alvarez, R.; Griss, J.; Nesvizhskii, A. I.; Perez-Riverol, Y., BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* **2017,** 33, (16), 2580-2582.
26. Wang, R.; Perez-Riverol, Y.; Hermjakob, H.; Vizcaino, J. A., Open source libraries and frameworks for biological data visualisation: a guide for developers. *Proteomics* **2015,** 15, (8), 1356-74.
27. Rost, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H. C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmstrom, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O., OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods* **2016,** 13, (9), 741-8.
28. Choi, M.; Eren-Dogu, Z. F.; Colangelo, C.; Cottrell, J.; Hoopmann, M. R.; Kapp, E. A.; Kim, S.; Lam, H.; Neubert, T. A.; Palmblad, M.; Phinney, B. S.; Weintraub, S. T.; MacLean, B.; Vitek, O., 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.
29. Kim, S.; Pevzner, P. A., MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* **2014,** 5, 5277.
30. Weisser, H.; Choudhary, J. S., Targeted Feature Detection for Data-Dependent Shotgun Proteomics. *J Proteome Res* **2017,** 16, (8), 2964-2974.

31. Griss, J.; Jones, A. R.; Sachsenberg, T.; Walzer, M.; Gatto, L.; Hartler, J.; Thallinger, G. G.; Salek, R. M.; Steinbeck, C.; Neuhauser, N.; Cox, J.; Neumann, S.; Fan, J.; Reisinger, F.; Xu, Q. W.; Del Toro, N.; Perez-Riverol, Y.; Ghali, F.; Bandeira, N.; Xenarios, I.; Kohlbacher, O.; Vizcaino, J. A.; Hermjakob, H., The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol Cell Proteomics* **2014,** 13, (10), 2765-75.
32. Choi, M.; Chang, C. Y.; Clough, T.; Broudy, D.; Killeen, T.; MacLean, B.; Vitek, O., MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **2014,** 30, (17), 2524-6.

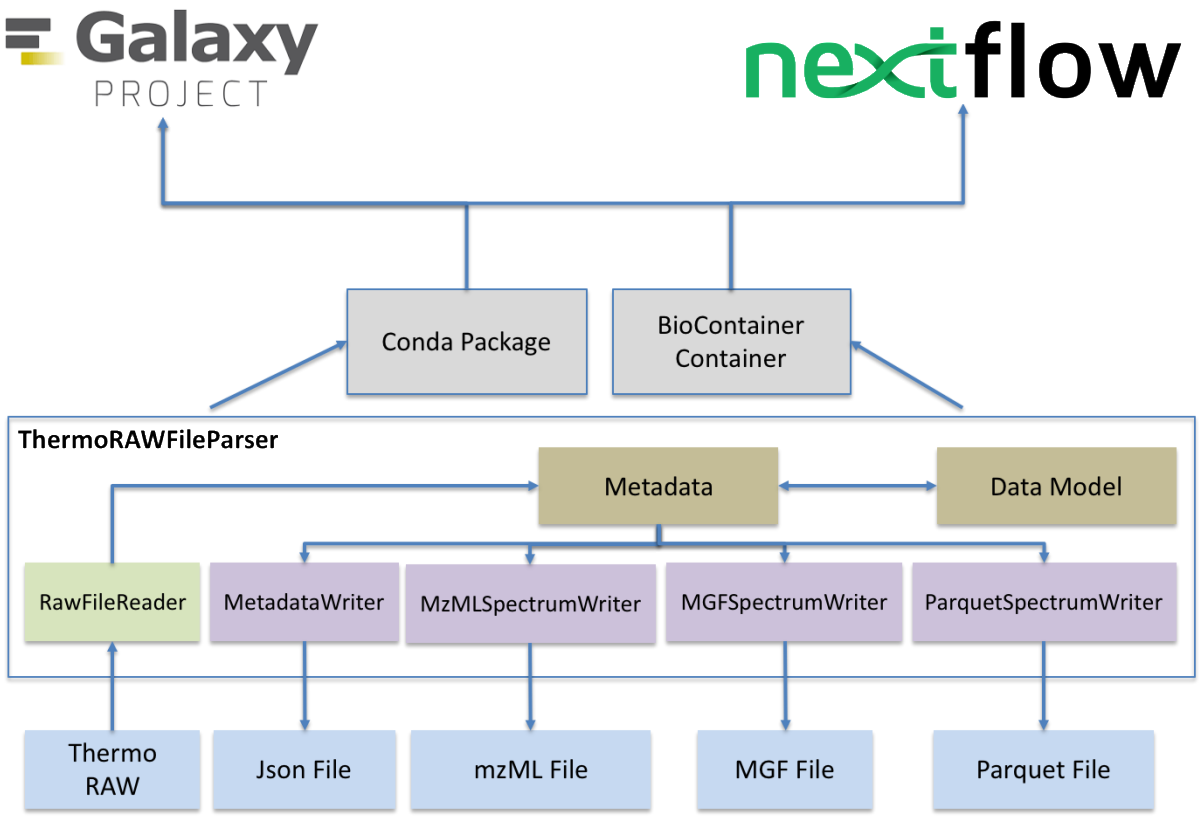

Figure 1: Modular design of ThermoRawFileParser includes exporters to MGF, mzML, Parquet, and Json Metadata. A Conda package and corresponding BioContainer is available for reuse in workflow engines such as Nextflow and Galaxy.

Figure 2: JSON representation for one File run in PXD006336 dataset. The JSON file contains metadata information about the file (FileProperties), the instrument (InstrumentProperties), mass spectrometry information (MSData), sample (SampleData) and scan settings (ScanSettings).

Figure 3: ThermoRawFileParserGUI provides a user-friendly user interface to convert Thermo Raw files to mzML, mgf and metadata formats.

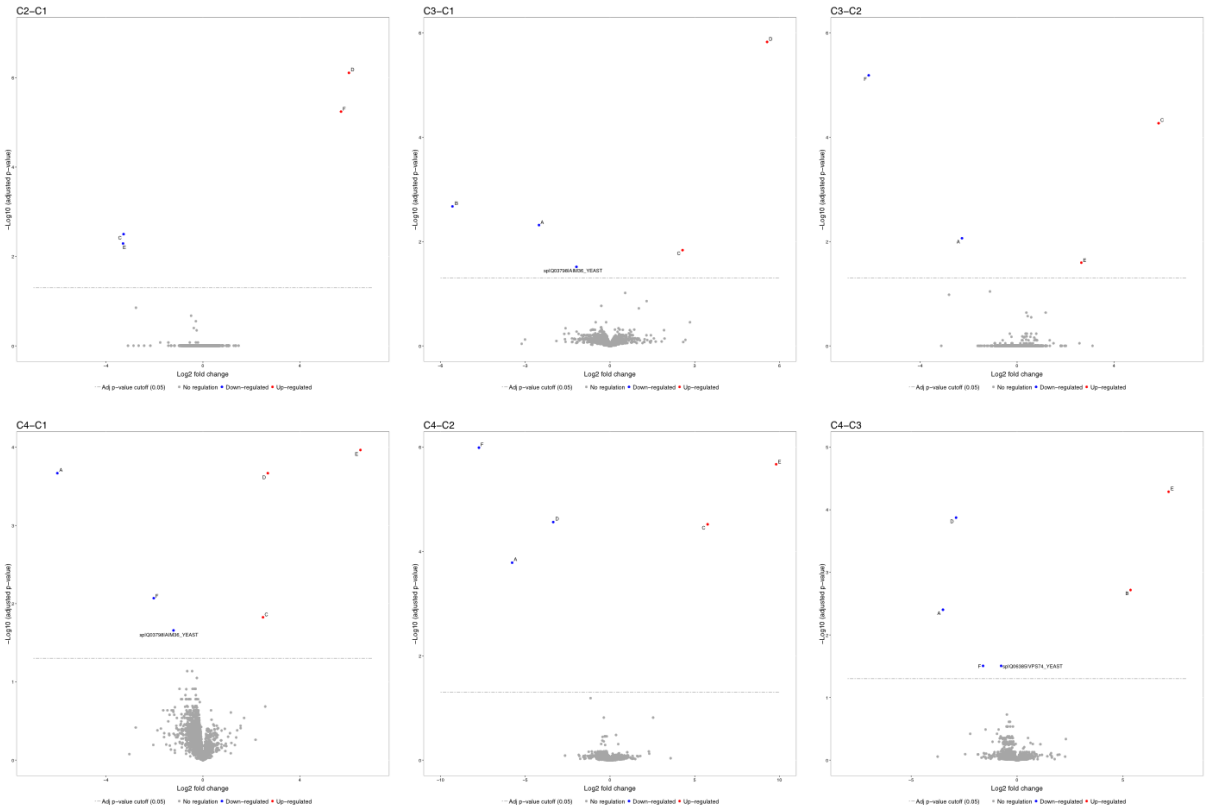

Figure 4: Volcano plot display of the results of the statistical analysis with OpenMS and MSstats of LC–MS converted using ThermoRawFileParser. Y axis: minus log10p-value of a pairwise comparison between two samples, adjusted to control the False Discovery Rate in the list of differentially abundant proteins in this comparison. X axis: log2-fold change between two samples.