Supplementary Material

SWATH-MS Gold Standard data experiment

Sample preparation:

422 stable isotope-labeled standard (SIS) peptides (AAA- quantified, Sigma or Thermo) were pooled in a master sample of equal concentration. 10 two-fold dilution steps of the master sample were conducted, resulting in a 512-fold concentration range from 50fmol/µL to 0.097fmol/µL. 15µL of each sample were spiked into 7.5 µL of HeLa cell lysate (human background), Saccharomyces cerevisiae BY4741 (yeast background) or water (no background). Three correspondingly diluted back- ground samples were prepared without SIS peptides as negative controls. All samples were finally supplemented with 2.5µL of reference peptides (iRT-Kit, Biognosys AG), yielding a final sample volume of 25 µL. For each replicate, 2 µL of the samples were injected, yielding a final protein amount of 1 µg of HeLa and yeast cell lysate, respectively, loaded on the HPLC column and SIS peptide amounts ranging from 60 fmol to 0.117 fmol.

Data acquisition:

An Eksigent nanoLC (AS-2 / 2Dplus) system was used to deliver the sample to the SWATH MS-enabled AB SCIEX TripleTOF 5600 System. The reversed phase LC-MS/MS analysis was conducted as follows. The HPLC solvent system consisted of buffer A (2% acetonitrile and 0.1% formic acid in water) and buffer B (2% water with 0.1% formic acid in acetonitrile). The samples were separated in a 75 µm- diameter PicoTip emitter (New Objective) packed with 20 cm of Magic 3 µm, 200A C18 AQ material (Bischoff Chromatography). The loaded material was eluted from the column at a flow rate of 300 nl/min with the following gradient: linear 2-35% B over 120 min, linear 35-90% B for 1 min, isocratic 90% B for 4 min, linear 90-2% B for 1 min and isocratic 2% solvent B for 9 min.

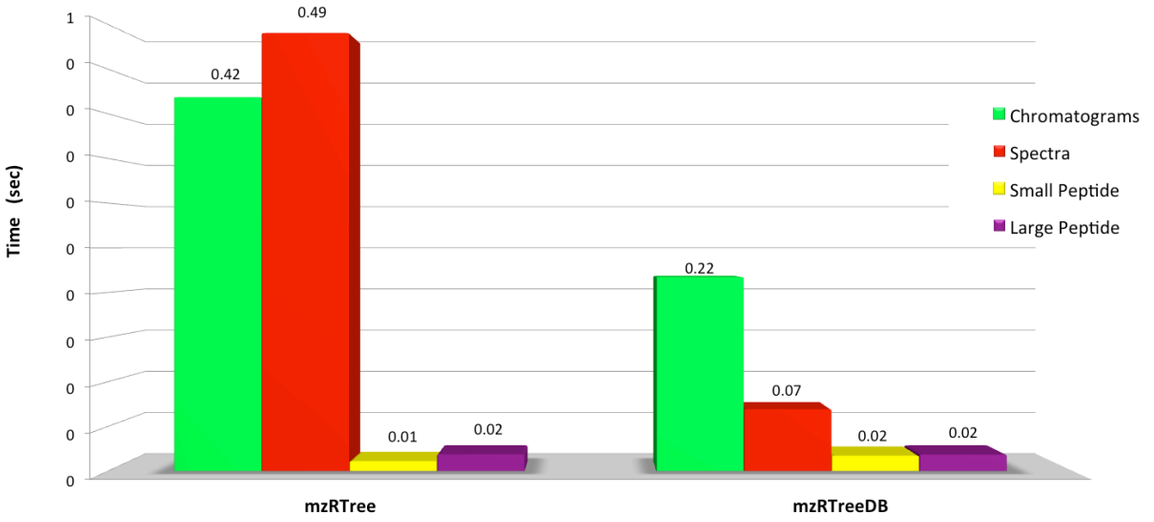
The mass spectrometer was operated essentially as described in (11),using 32 windows of 25 Da effective isolation width (with an additional 1 Da overlap on the left side of the window) and with a dwell time of 100 ms to cover the

mass range of 400 - 1200 m/z in 3.3 s. Before each cycle, an MS1 scan was acquired; then the MS2 scan cycle started (399-425 m/z precursor isolation window for the first scan, 424-450 m/z for the second ...1174-1200 m/z for the last scan). The collision energy for each window was set using the collision energy of a 2+ ion centered in the middle of the window with a spread of 15 eV.

To minimize sample carry over, for each dilution series, the samples were injected from the least concentrated to the most concentrated. The 10 dilution steps in 3 different backgrounds were acquired in triplicates in DIA mode with SWATH-MS settings on a TripleTOF 5600 System, giving rise to 90 LC-MS/MS data files.

Feasibility study outcome

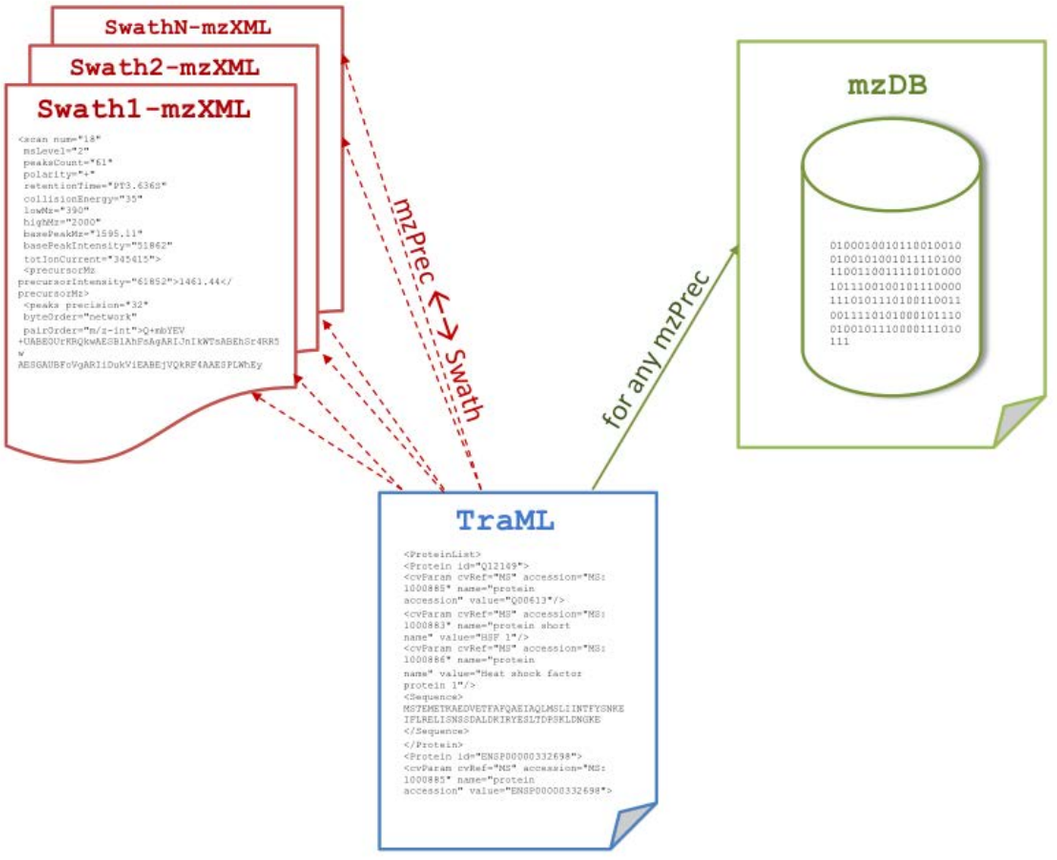
mzDB has been designed to implement a multidimensional data indexing model that allows for efficient queries along both time and m/z dimensions and it is based on mzRTree (42), an efficiency oriented data format. Here, the mzRTree structure was implemented as an SQLite file format, taking advantage of the SQLite standardization and its built-in R*Tree index. As a preliminary step to the mzDB project, we evaluated how the SQLite adaptation of mzRTree would have affected access times performance. Our feasibility study proved that the SQLite implementation could even improve upon published results: we attained, indeed, either comparable performance or speed gain of 2 to 10 fold (see Figure 1).



Supplementary figure 1: Performance comparison from the feasibility study. The times on 4 different accesses are reported for both the mzRTree data structure and its prototype SQLite implementation. The 4 tested accesses are the same as from the original mzRTree paper: Chromatograms (5Da x whole LC gradient), Spectra (20 retention times x whole m/z), Small Peptide (5Da x 60 retention times), Large Peptide (5Da x 200 retention times).

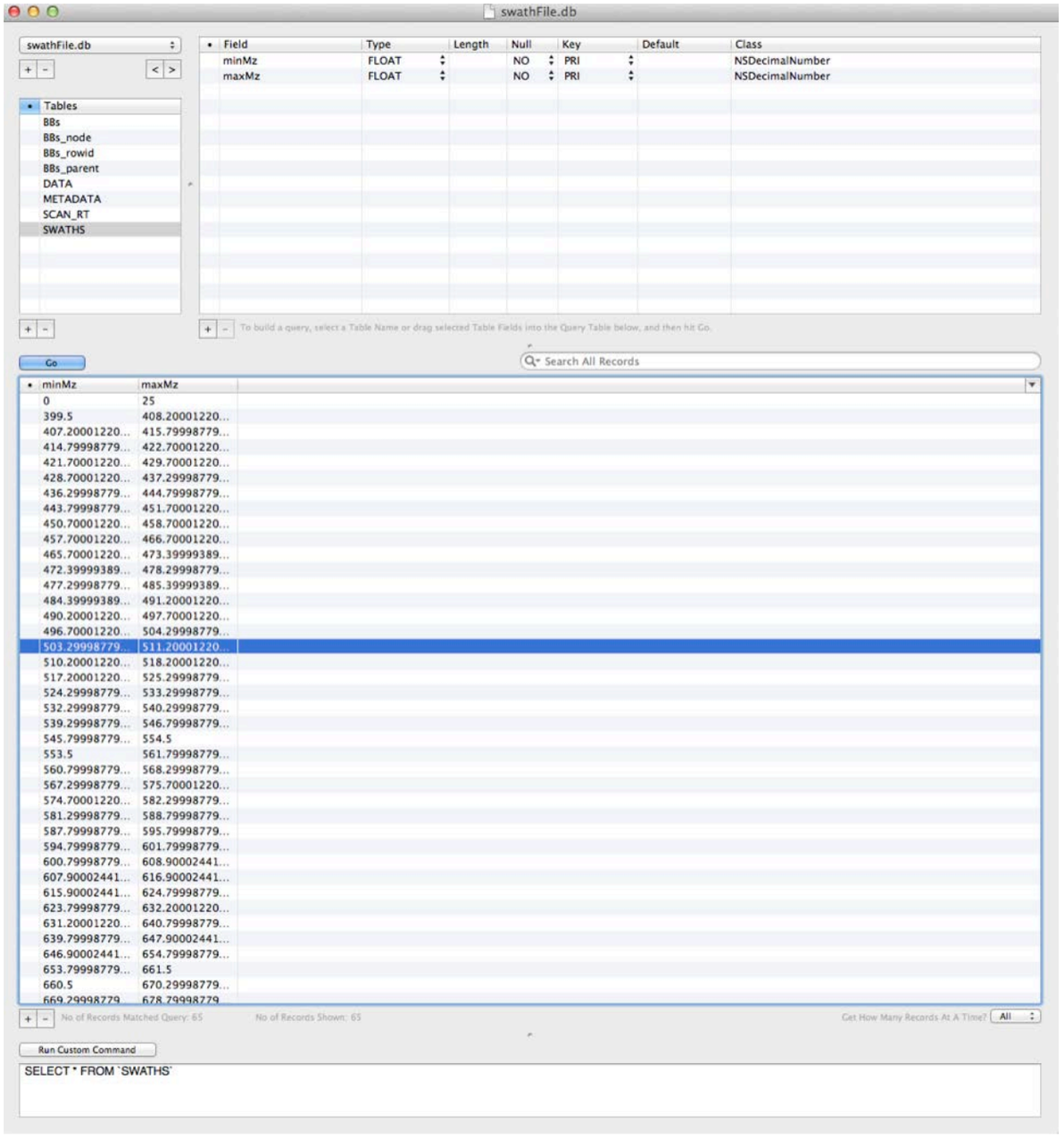
Further mzDB advantages

State-of-the-art software for SWATH-MS data analysis that use open formats (i.e., OpenSwath) is using a splitted version of each mzXML file. Essentially, the single mzXML is splitted in many mzXML files, one for each swath: as a result, depending on the precursor ion a different data file will be accessed to retrieve the desired data. Clearly, the adoption of mzDB, besides the performance gain, would dramatically simplify the process. A visual comparison of the 2 processes is represented in Figure 2.



Supplementary figure 2: Visual comparison of the data access procedure using the splitted mzXML or mzDB. Starting from the same library (or transition list), the splitted mzXML require the software to access different data files depending on the queried precursor ion, resulting in a resources overhead that can be avoided using mzDB.

Since it is based on the standard SQLite technology, an additional advantage of the mzDB is that can be intuitively browsed by non-specialized users by means of existing dedicated SQLite GUIs (see Figure 3). Hence, it is much easier to access the stored metadata information with mzDB than with XML based formats.



Supplementary figure 3: Browser screenshot of the mzDB for a SWATH-MS file. mzDB can be browsed with existing dedicated SQLite GUIs. Hence, it is much easier to access the stored metadata information with mzDB than with XML based formats.

Supplementary figure 4: File size comparison of several raw data formats. A Thermo RAW file of 1.57 GB has been converted into mzML, mz5 and mzDB file formats using ProteoWizard tools (see Experimental Procedures). mzML and mz5 conversions were performed while enabling (comp) or not (no comp) a zlib compression in msconvert. mzDB was created in profile or fitted mode using the "raw2mzDB.exe" tool. mzML file need 2 to 3 times more space compared to raw file, while mz5 and mzDB are close to the original file size respectively in compressed and fitted modes.