HUPO 2023 – PP05-610

TIMSquant™: Precise and scalable MS1 quantification for DDA and DIA using transfer learning, targeted analysis and semi-supervised machine learning

Tharan Srikumar[1], Ignacio Jauregui[2], Michael Krause[2], Sven Brehmer[2], Sebastian Wehner[2], Dennis Trede[2], Jonathan Krieger[1], George Rosenberger[3]
[1]Bruker Ltd., Milton, ON; [2]Bruker Daltonics GmbH & Co. KG, Bremen, Germany; [3]Bruker Switzerland AG, Faellanden, Switzerland

Introduction

Bruker ProteoScape™ (BPS; formerly PaSER) has been transforming into a comprehensive proteomic data analysis platform that integrates all common data processing steps within a single framework.

In label-free analysis of data-dependent acquisition (DDA) datasets, different approaches have been established either based on MS1 feature finding (MS1-FF) or the extraction of ion chromatograms (MS1-XIC). Although both solutions can typically quantify the identified peptide precursors accurately within single runs, MS1-XIC-based approaches frequently have a higher recovery rate when the targeted signals are guided by peptide-precursor-defined properties such as expected retention time (RT), ion mobility (IM), and isotopic pattern. However, defining these properties based on peptide-spectrum-matches (PSMs) can be difficult, because in DDA, due to the stochastic and heuristic selection of precursors for fragmentation, the proportion of missing values dramatically increases when multiple runs are quantitatively compared. To alleviate this issue, "match-between-run" (MBR) algorithms are typically employed, aligning LC gradients and transferring peptide identifications to runs with missing values. While this represents a suitable solution for small- to medium-sized sample cohorts, the approach struggles to scale to the alignment of hundreds or thousands of samples.

To address these challenges, we have developed TIMSquant, a novel MS1-XIC-based algorithm that replaces MBR by run-wise transfer learning of global RT and IM prediction models. Using the additional IM separation dimensions of timsTOF instruments provides increased specificity, while advanced isotope ion chromatogram scoring in combination with semi-supervised machine learning and statistical validation provides consistent quantification consistency with controlled error rates of quantitative features of identified peptides and aligned missing values.

Methods

TIMSquant uses confident PSMs in run-wise and global contexts from upstream database search engines in addition to MS1 spectra as input. For each run separately, global machine learning models based on AlphaPeptDeep[PMID:36433986] for the prediction of RT and IM are adapted to local sample and instrument conditions using transfer learning and a randomly sampled subset of several hundreds to thousands of confidently identified peptides. Using the full set of peptides confidently identified in global context, the missing values not identified in run-specific context are selected and the local models are used to predict the run-specific RT and IM values within each run. In addition, RT and IM-dependent window widths based on the deviation of measured and predicted values of identified peptides are estimated.

The full set of peptide precursors, their measured or predicted RT and IM coordinates and windows, are then used to extract precursor ion chromatograms from the MS1 scans within predefined boundaries. Chromatograms are extracted for the first three isotopes. Chromatographic peak picking based on OpenSWATH[PMID:24727770] is then used to define peak borders of the candidates using the first isotope, whereas chromatographic scoring assesses cross-correlation and mutual information between isotopes. Deviations of expected values in m/z, isotope pattern, RT and IM dimensions complement the set of scores that is derived for each candidate signal. Using an XGBoost-based semi-supervised learning approach provided by PyProphet[PMID:33333029], a classifier is trained to separate the identified peptides from a null model derived by predicting RT and IM coordinates for mutated peptide sequences. This classifier and the null model then allow to score and statistically validate quantitative features based on the predicted coordinates for missing values[PMID: 28673088].

TIMSquant exports quantitative values on peptide-precursor and protein levels using the MaxLFQ algorithm[PMID:24942700, PMID:31909781]. The analysis run-time scales linearly with the number of samples and all steps can be run independently, allowing for full parallelization of the workflow.

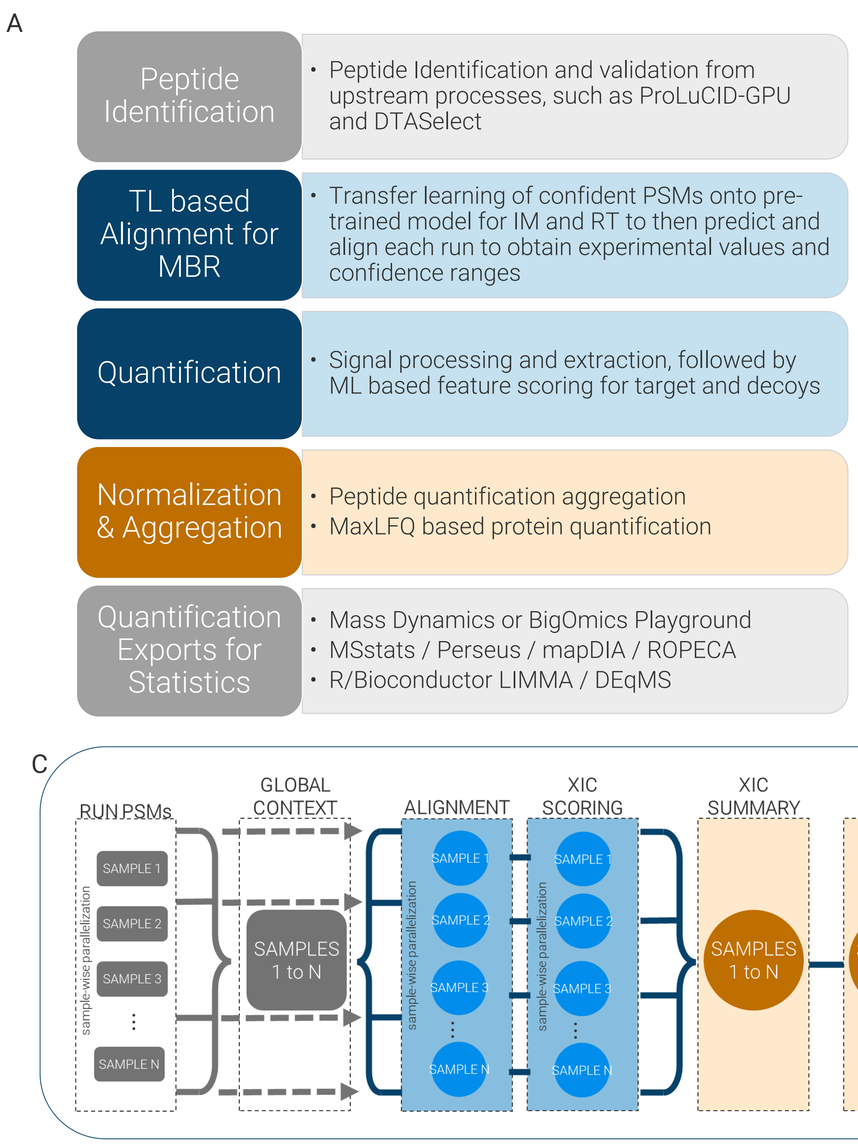

Fig. 1: The TIMSquant workflow. A) TIMSquant requires identified peptides in run-specific and global contexts as input, as well as the MS1 raw data. B) Transfer-learning-based alignment is used as replacement for match-between-runs. Quantification is conducted by an ion chromatogram extraction (XIC)-based approach. C) All steps of TIMSquant can be parallelized. Global context confidence estimates based on PSMs of all individual runs need to be provided to define the total set of peptides that will be quantified. D) Classification of candidate signals based on XIC and mass & ion mobility scores. E) Cross-correlation shape and coelution scores across retention time (RT) typically are among the most discriminative scores. F) Mass and Ion Mobility (IM)-based scores provide orthogonal evidence for the correctness of candidate peptide features. G) The combined XGBoost classifier, combining several different partial scores, provides superior performance to the individual components.

Results

To assess the performance of our method, we used the established LFQbench strategy[PMID:27701404] in combination with dda-PASEF measurements of differentially mixed human, yeast, and E.coli samples, identically as described previously. In total 5 replicates were measured for both Sample A and B. As reference, we used MaxQuant (2.4.2.0) with default parameters, with MBR enabled and disabled. Our comparison shows that MaxQuant (without MBR) and TIMSquant quantify similar numbers of peptides (Fig. 2). However, the number of complete quantification events (5 quantifications in both Sample A & B) is substantially higher for TIMSquant than MaxQuant. Whereas the number of partial quantifications (5 quantifications either in Sample A or B) is similar, MaxQuant (with MBR) has many more incomplete quantification events (at least 1 quantification in Sample A or B).

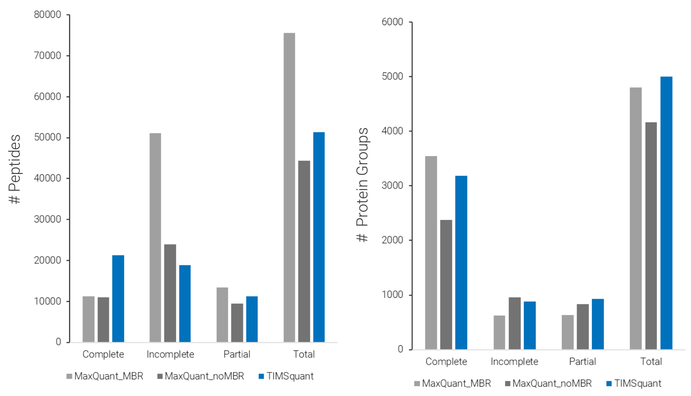

Figure 2: LFQbenchmark quantification performance. Peptide and protein-level coverage is depicted for TIMSquant, MaxQuant (with MBR), and MaxQuant (without MBR).

Assessment of quantification accuracy (Fig. 3) suggests that TIMSquant is similarly accurate as MaxQuant (without MBR), while providing much more complete quantification across the replicates. In contrast, MaxQuant (with MBR) provides lower quantitative accuracy.

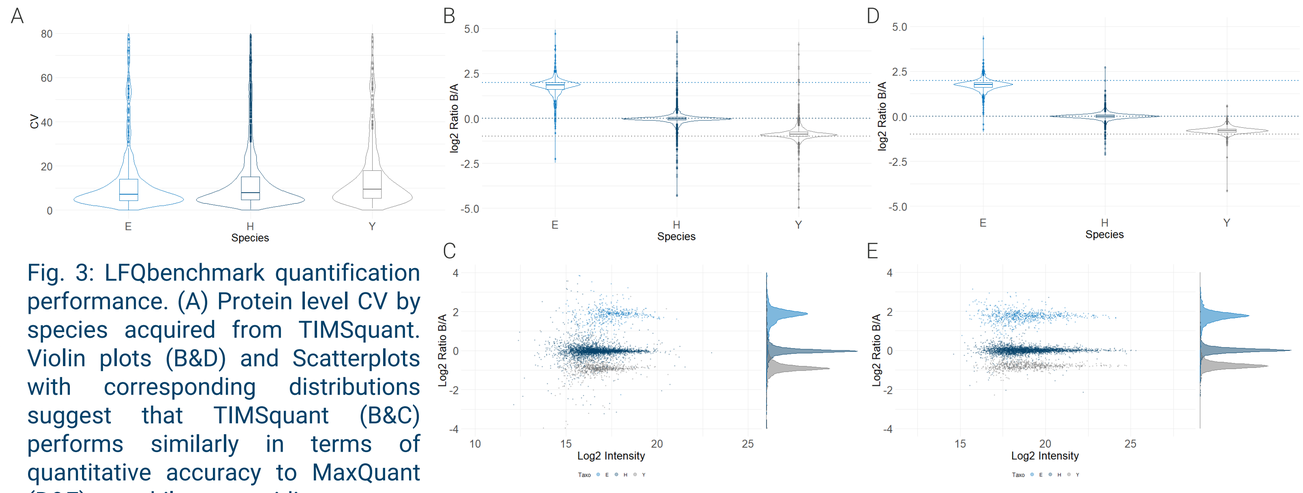

Fig. 3: LFQbenchmark quantification performance. (A) Protein level CV by species acquired from TIMSquant. Violin plots (B&D) and Scatterplots with corresponding distributions suggest that TIMSquant (B&C) performs similarly in terms of quantitative accuracy to MaxQuant (D&E), while providing more consistent quantification performance.

- TIMSquant represents an accurate, scalable MS1-XIC-based quantification approach, replacing MBR algorithms by state-of-the-art machine learning techniques.
- Together with mass accuracy and isotope assessment, ion mobility represents the most important criteria for selection of true quantitative signals.
- TIMSquant is natively integrated within Bruker ProteoScape, allowing the extension of Run & Done to quantitative applications.

Technology