BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 4 2015, pages 555–562
doi:10.1093/bioinformatics/btu686

Gene expression

Advance Access publication October 27, 2014

DIANA—algorithmic improvements for analysis of data-independent acquisition MS data

Johan Teleman[1,2], Hannes L. Röst[3], George Rosenberger[3], Uwe Schmitt[4], Lars Malmström[5], Johan Malmström[1,*] and Fredrik Levander[2,*]

[1]Department of Clinical Sciences, Lund University, BMC B14 221 84 Lund, [2]Department of Immunotechnology, Lund University, Medicon Village (Building 406) 223 81 Lund, Sweden, [3]Department of Biology, Institute of Molecular Systems Biology, [4]ITS Scientific IT Services, ETH Zurich and [5]SIT, University of Zurich, Winterthurerstrasse 190, 8057, Zurich, Switzerland

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Motivation: Data independent acquisition mass spectrometry has emerged as a reproducible and sensitive alternative in quantitative proteomics, where parsing the highly complex tandem mass spectra requires dedicated algorithms. Recently, targeted data extraction was proposed as a novel analysis strategy for this type of data, but it is important to further develop these concepts to provide quality-controlled, interference-adjusted and sensitive peptide quantification.

Results: We here present the algorithm DIANA and the classifier PyProphet, which are based on new probabilistic sub-scores to classify the chromatographic peaks in targeted data-independent acquisition data analysis. The algorithm is capable of providing accurate quantitative values and increased recall at a controlled false discovery rate, in a complex gold standard dataset. Importantly, we further demonstrate increased confidence gained by the use of two complementary data-independent acquisition targeted analysis algorithms, as well as increased numbers of quantified peptide precursors in complex biological samples.

Availability and implementation: DIANA is implemented in scala and python and available as open source (Apache 2.0 license) or precompiled binaries from http://quantitativeproteomics.org/diana. PyProphet can be installed from PyPi (https://pypi.python.org/pypi/pyprophet).

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 27, 2014; revised on September 26, 2014; accepted on October 15, 2014

1 INTRODUCTION

Accurate and precise quantification of proteins is a critical component of life science and systems biology applications. The prevailing method for quantification of complete proteomes was until recently data-dependent acquisition (DDA), also referred to as shotgun mass spectrometry (MS). Shotgun MS can provide extensive maps of the measurable and expressed proteomes of a large numbers of organisms, tissues, organs and organelles. However, <50% of identified peptides are typically shared between two replicate shotgun MS injections (Tabb *et al.*, 2010), requiring multiple injections of the same sample to reproducibly measure peptides in all samples (Liu *et al.*, 2004; Vincent *et al.*, 2013; Bailey *et al.*, 2014). The limited analytical reproducibility observed in shotgun MS has fuelled the development of targeted MS strategies such as selected reaction monitoring (SRM), to increase reproducibility and specificity compared with shotgun MS (Wolf-Yadlin *et al.*, 2007).

To perform targeted MS strategies requires a priori determined information on how to target a given peptide sequence. Such information typically consists of the peptide sequence, the preferred charge state, the empirical or predicted high performance liquid chromatography (HPLC) elution time, as well as the relative intensities and masses of the *n* most prominent fragments. The construction of these MS assays requires a substantial effort, which has resulted in the assembly of public repositories of peptides and MS assays, to simplify further studies (Desiere *et al.*, 2006; Picotti *et al.*, 2008; Farrah *et al.*, 2012; Karlsson *et al.*, 2012). Although the targeted MS strategies such as SRM provides reproducible and accurate protein quantification, the throughput is normally limited to up to a few hundred peptides per injection (Picotti *et al.*, 2009; Waldemarson *et al.*, 2012), limiting the technique for whole-proteome studies.

Data-independent acquisition MS (DIA-MS) was originally used to improve peptide identification rates (Purvine *et al.*, 2003; Plumb *et al.*, 2006; Panchaud *et al.*, 2009), but lately workflows using DIA-MS combined with targeted data extraction have been described in attempts to combine the reproducibility of SRM with the throughput of shotgun MS (Gillet *et al.*, 2012; Weisbrod *et al.*, 2012; Egertson *et al.*, 2013). Data acquisition in DIA-MS relies on deterministic splitting of the survey scan peptide-ion mass range into one or more subsets, followed by co-fragmentation of all precursor masses in one entire subset, while leaving the de-convolution of the peptide ions in these complex MS^2 spectra to the post-acquisition analysis. The acquisition method yields complete MS^2-retention time maps, compared with the discontinuous maps of shotgun MS, and can be seen as a complete digitization of the sample as seen by the mass spectrometer.

Targeted extraction DIA-MS has the sensitivity, precision, reproducibility and dynamic range to allow deep large-scale measurement of the proteomes of biological systems (Collins *et al.*,

*To whom correspondence should be addressed.

2013). However, the strategy's data analysis needs further improvement, and currently only a few tools exist that are capable of large-scale robust targeted extraction DIA-MS analysis (Bernhardt *et al*., 2012; Egertson *et al*., 2013; Röst *et al*., 2014). The poor availability of data analysis tools limits robustness, as users are blindly exposed to any potential error in the particular algorithm used, and lowers sensitivity, as parts of the proteome might be unreachable due to the potential preferences of a given algorithm. For example, the combination of multiple algorithms in shotgun MS has been shown to increase the amount of peptide-spectrum matches with up to 50% compared with a single algorithm (Häkkinen *et al*., 2009; Jones *et al*., 2009; Nahnsen *et al*., 2011; Shteynberg *et al*., 2013).

We have previously described algorithms for detection of the correct signals in SRM chromatograms based on fragmentation patterns (Teleman *et al*., 2012), and we hypothesized that these concepts can further improve the targeted data analysis in DIA data, and also provide complementarity towards existing tools. Here, we investigate this by combining the DIA-MS targeted analysis strategy with our previous efforts in SRM data analysis, and present a new algorithm and software for automated analysis of DIA-MS data. The algorithm, called DIANA, introduces a new function for computing chromatographic peak sub-scores based on expected ratios between fragments, as well as a new interference-corrected measure of quantity. These factors, together with the new semi-supervised classification tool PyProphet, increase the amounts of peptide quantifications and enable more accurate quantifications in complex samples. Finally, we also demonstrate that DIANA is complementary to the previously published OpenSWATH software (Röst *et al*., 2014), and that the combination of results from the two engines can further improve on the confidence in and number of peptide quantifications.

2 METHODS

The DIANA analysis workflow has similarities to classical shotgun MS data analysis workflows. For each target peptide ion, chromatograms are extracted, followed by chromatogram peak detection and scoring by several sub-scores. The same procedure is applied to a large number of decoy peptides to allow for significance estimations, analogous to shotgun MS/MS database searching (Elias and Gygi, 2007). In addition, a number of retention time peptides are targeted using the same method, which are used to normalize retention times for the peptide-ion assays in the current injection. Target and decoy peptide peaks are then subjected to a semi-supervised learner to merge the sub-scores into a final score, to select the best peak in each chromatogram, and to estimate false discovery rates (FDRs).

Input MS data to DIANA should be in mzML format (Martens *et al*., 2011) with optional MS-numpress (Teleman *et al*., 2014) and gzip compression. Apart from the raw MS data, three assay lists in TraML format (Deutsch *et al*., 2012) are required; one with target assays, one with decoy assays and one with retention time normalization assays.

2.1 Targeted data extraction

DIANA is based on a targeted data analysis approach, which in turn requires targeted data extraction. During acquisition, the mass spectrometer systematically collects one MS^1 spectrum followed by MS^2 spectra of preselected subsets of the MS^1 range (Fig. 1A). For the targeted extraction of a peptide ion, DIANA relies on an MS assay consisting of a number of *channels*, describing the most prevalent fragments and the most prominent natural isotopes of the peptide (Fig. 1B). Chromatograms for each channel are extracted from the MS^1 and relevant MS^2 spectra, using a given window size and deconvolution function, to give a multi-channeled measurement of the targeted peptide ion (Fig 1C). The extracted chromatogram for a channel is here referred to as a *trace*, and all the traces for a peptide-ion assay will be collectively called a peptide-ion *assay trace* (Fig. 2).

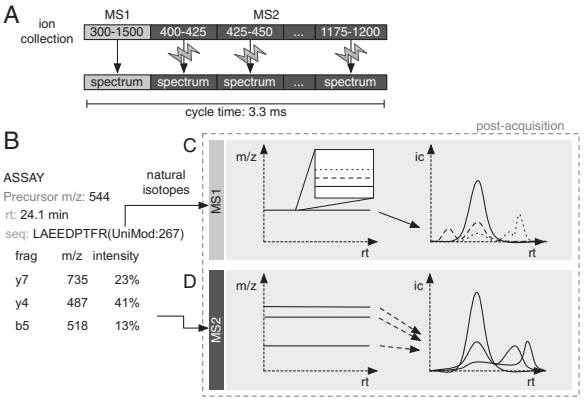

Fig. 1. Overview of DIA-MS and targeted extraction. (**A**) The instrument typically performs a single full-range MS1 scan, followed by a number of MS2 scans on subsets of the precursor range. (**B**) For the targeted extraction and analysis, a peptide ion assay is used, with information on the prevalent isotopes and fragments for the peptide. Assay isotopes (**C**) and fragments (**D**) are extracted from the MS1 and relevant MS2 spectra to get chromatograms related to the target peptide ion

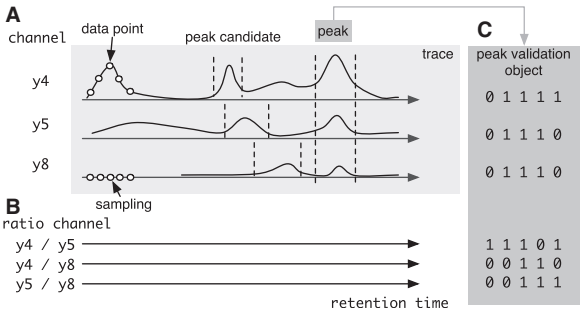

Fig. 2. Nomenclature for chromatogram extraction and peak picking. Target fragments and isotopes can be thought of as data *channels* (**A**), a ratio between two channels a *ratio channel* (**B**), and the measurements for a channel are called a *trace*. Local maxima in channels are called *peak candidates* (A); of which several aligned is a multi-channeled *peak*. (**C**) For all peaks a Boolean peak-validation object is computed. The peak-validation object displays the data points that are close to the target ratios (within the target tolerance window), which is an indication of a correct peak

2.2 Peak detection and initial scoring

Each trace under analysis is smoothed by taking the second level Laplace eight-point wavelet decomposition, and a baseline is also calculated as the maximum of 1.0 and the median of a 20-point sliding window, resulting in a minimum value of 1. The smoothed trace is partitioned by its local minima, resulting in a number of *peak candidates* (Fig. 2A). These are considered further if the smoothed curve intensity at the apex (local maximum for the candidate) is larger than twice the baseline at the same time.

Peak candidates from the different channels in the peptide-ion assay trace are then grouped if peak candidate apices are maximally 1 data

point (Fig. 2A) off, and filtered to only leave *peaks* (Fig. 2A) with at least two fragment candidate peaks, or at least one fragment candidate peak that has a sufficiently large area (default cutoff 25.0).

Peaks are initially scored by four different scores—the fragment Markov ratio probability (FMRP), the fragment correlation score (FCS), the isotope Markov ratio probability (IMRP) and the isotope correlation score (ICS). As indicated by the names, the scores represent two types of calculations (Markov ratio probability (MRP) and correlation score) for two types of inputs (fragments and precursor isotopes).

2.3 Markov ratio probability

MRP is here introduced as a type of *P*-value that can be calculated for sections in an *n*-channeled input, with the goal to find sections where the ratio between each pair of channels maintains a target value. In our case, the channels and ratios are either peptide fragments and an empirical fragmentation pattern, or peptide isotopes and a natural isotope distribution. To calculate the MRP for a peak of width *w* data points, first all the pairwise ratio channels (Fig. 2B) between the channels are calculated as previously described (Teleman *et al*., 2012), followed by the computation of a Boolean peak-validation object (Fig. 2C). This object consists of Boolean vectors of length *w*: the vectors v_i, $0 \leq i < n$, correspond to the input channels c_i, and the vectors $w_{i,j}$, $i<j \leq n$ correspond to the ratio channels $r_{i,j}$, giving a total of *m* vectors, $m = n + (n * (n-1))/2$. The purpose of the peak-validation object is to specify in detail which data points in each channel and ratio channel that provide evidence of the target relationship. For a description of the population of the peak-validation object, see the Supplementary Methods.

With the peak-validation object, a *P*-value is calculated for each ratio channel using a two-state Markov model (see Supplementary Methods for motivation). The states represent agreement or non-agreement with the target ratio, and the likelihoods for the four state-transitions in the model are chosen by frequency counting of the measured state-transitions using all data points in the ratio channel that are inside any peak. If the peak has *t* of *w* data points in agreement with the target ratio in a ratio channel, the *P*-value is calculated as the likelihood of getting *t* or more data points in agreement given the above Markov model. These *P*-values are combined to one according to Kost and McDermott (Kost and McDermott, 2002), as they are calculated on pair-wise ratios and therefore dependent. This final *P*-value is the MRP.

2.4 Interference correction/signal estimation

The peak-validation object is a detailed map of the data points in the channels believed to support the target ratios, but inversely also a map of possibly noisy data points. These noisy data points will heavily influence the reported quantity if some high-abundant alternative ion is causing the deviation. Therefore, for any data point that is not validated according to the peak-validation object, we calculate an estimated intensity as the average of all validated channels at that time, multiplied by the expected ratios. If this estimated intensity is less than half the measured one, the estimated intensity is used in place of the measured.

2.5 Correlation sub-score

As a complement to the MRP, a correlation sub-score is calculated using the corrected assay trace over the peak. The Pearson correlation between each pair of corrected traces is computed, and the correlation score is calculated as the mean of these correlations. The correlation score is calculated separately for the precursor isotope traces (ICS) and the fragment traces (FCS).

2.6 Retention time normalization

Once the first four sub-scores are calculated, the FMRP is used to select the best peak (lowest FMRP) in each decoy and retention time assay trace, and *q*-values are calculated using a simple non-parametric method (Käll *et al*., 2008). Retention time peptide peaks at *q*-value <15% are selected. For these a linear regression is made for measured versus expected retention time. To correct for possible false positive identifications, peaks with residuals outside 3 standard deviations are discarded, and the linear regression is performed again on the remaining peaks. The linear transformation specified by the regression is used to map the expected (assay) retention times for all target- and decoy-assays to the specific chromatographic profile of this injection, and a retention time score is calculated as the absolute deviation of a peak apex from the adjusted expected retention time. This supports assay libraries with iRT retention times (Escher *et al*., 2012), although any linear retention time scale can be used.

2.7 Semi-supervised classification with PyProphet

Using the above five sub-scores, decoy and target peaks are used to perform semi-supervised learning, using the new tool PyProphet. PyProphet is a Python reimplementation of mProphet (Reiter *et al*., 2011), using optimized C code and NumPy (http://numpy.org) calculations to decrease computation times and memory usage by several orders of magnitude for large input sets. Apart for optimizations, PyProphet also extends the original mProphet functionality with (i) multiple inner learners from SkLearn (Pedregosa *et al*., 2011) like stochastic gradient descent, logit and support vector machine (SVM), apart from the original linear discriminant analysis, (ii) non-parametric and log-normal null-distribution models, (iii) *q*-value calculation according to Storey (Storey, 2002) and (iv) traditional cross-validation. To account for the non-gaussian sub-score null distributions, DIANA uses a SVM with an rbf-kernel, and the non-parametric null model. DIANA uses a traditional cross-validation where all samples are used for learning, instead of the random sampling cross-validation used by mProphet and OpenSWATH.

2.8 Implementation

The software representing the DIANA algorithm is packaged into a set of stand-alone Java Virtual Machine 1.6 command line applications. These can be run individually, or combined and scheduled using a small toolbox of python programs that is also provided. All software exists as self-contained binary packages, which can be downloaded from http://quantitativeproteomics.org/diana, and should be compatible with any operating system. PyProphet is a stand-alone Python tool, which can be installed from PyPi (https://pypi.python.org/pypi/pyprophet/0.9.1) or downloaded and compiled from source from https://github.com/fickludd/pyprophet. For installation of PyProphet, we currently recommend a Linux environment.

2.9 Analysis of gold standard dataset and *Streptococcus pyogenes* dataset

Vendor data files and assay lists were obtained from (Röst *et al*., 2014). Decoys for the gold standard dataset were generated by shuffling each target peptide trice, giving 1026 decoy assays, whereas a random subsample of 3000 *Streptococcus pyogenes* peptide ions were shuffled once to give 3000 *S.pyogenes* decoy assays. Decoy generation was performed using an in-house tool called DecoyGenerator, which shuffles the peptide amino acid sequence randomly, but preserves the c-terminal amino acid. Data files were processed through the DIANA workflow, using an in-house MS-Numpress enabled Msconvert build (essentially performing equally to msconvert in current ProteoWizard builds (Chambers *et al*., 2012)). Chromatograms were extracted with a ±20 ppm uniform extraction window, using DianaExtractor, unless other window sizes are indicated. Apart from additionally using the three most abundant precursor isotopes, the extracted fragments and retention time peptides were as previously described (Röst *et al*., 2014). In PyProphet we used weighted

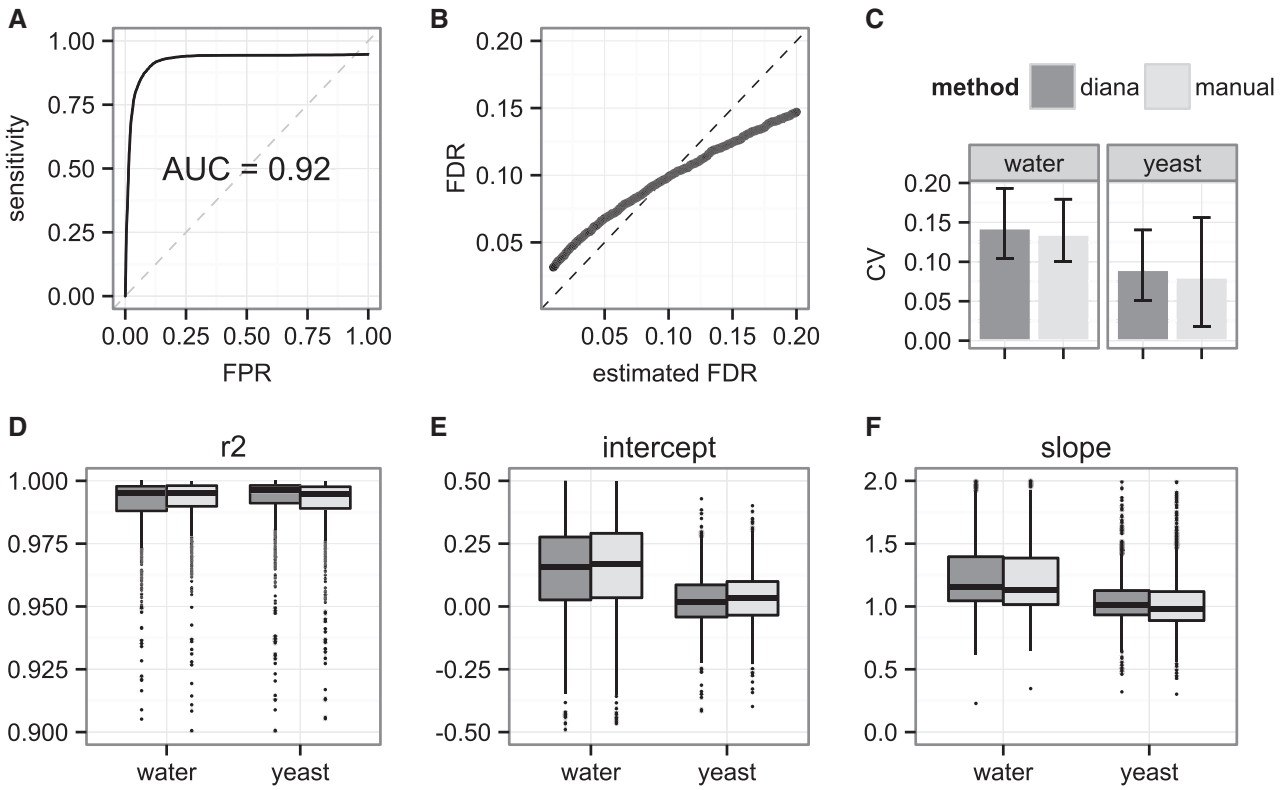

Fig. 3. Validation of DIANA of gold standard data set compromising 342 manually analyzed peptides in 60 injections. (**A**) ROC-curve of DIANA and PyProphet semi-supervised classification. (**B**) Evaluation of true FDR according to the manual analysis as a function of estimated FDR by DIANA. (**C**) Coefficients of variation calculated on the three technical replicates for each peptide and dilution level. Precision is similar to manual analysis. (**D-E**) Linearity of peptides. Log-log scale linear regression on each peptide and dilution series reveals confidently high coefficients of determination (r^2) and intercepts and slopes close to theoretical values (0.0 and 1.0), with performance identical to manual analysis

classes, 10 iteration traditional cross-validation (xeval.type = split), all peptides were used in the cross-validation, and an rbfSVM inner learner with 1 GB cache size was used. The non-parametric null distribution was used with Storey FDR calculation (Storey, 2002), and mProphet (Reiter *et al.*, 2011) statistics calculation and sampling. Applied data analysis was done using custom R scripts, mainly using reshape and ggplot2 packages.

3 RESULTS

Targeted data extraction for the analysis of complex DIA-MS data was recently demonstrated as a promising data analysis strategy. With the goal to further explore and improve this strategy, we have designed a new peak detection algorithm, as well as four new peak sub-scores—FMRP, FCS, IMRP and ICS. Analogous to previous published work on SRM data analysis (Teleman *et al.*, 2012), the sub-scores consider the data channels in a pair-wise manner to provide robustness toward interfering signals from non-targeted compounds. All sub-scores are used to calculate a *q*-value, which allows the user to determine the strength of the found evidence for a certain peptide, and to filter the results at a target FDR.

To evaluate DIANA performance, we used the gold standard water and yeast background datasets from the OpenSWATH publication (Röst *et al.*, 2014). The gold standard dataset consists of 342 detectable stable isotope labeled peptides, diluted in 10 concentrations from 1:1 to 1:512 in water and yeast lysate backgrounds, yielding 20 separate samples, analyzed in triplicates, resulting in 60 DIA-MS maps. Targeted extraction of the spiked-in peptide traces from the 60 DIA-MS maps generated 20 520 extracted chromatograms, which were analyzed manually in the OpenSWATH publication (Röst *et al.*, 2014).

3.1 Parameter optimization and quantity calculation

Before analyzing the 60 DIA-MS maps from the gold standard dataset using DIANA, we selected a subset of six DIA-MS maps (yeast background, dilutions 1:2 to 1:64) to optimize the chromatogram extraction parameters. We tested extraction window sizes of ±5, ±10, ±20, ±40 and ±80 ppm from the theoretical mass, using either uniform or top-hat extraction profiles, resulting in 10 extracted chromatogram sets per map. Each chromatogram set was scored and classified using DIANA and the gold standard target and decoy assays, and evaluation was based on the amount of significant peptides at 1% FDR. The ±20 ppm uniform window resulted in the highest number of significant peptides in the more diluted samples (Supplementary Fig. S1) and was used for the rest of this study. Note that the selected extraction window shape and size should be close to optimal for any measurement using the same method on the same instrument. Other DIANA parameters (Supplementary Table S1 for

full list) are treated as constants for MS methods similar to the used method, and do not need to be changed. For example, DIANA is robust with respect to the retention time mapping parameters (Supplementary Figs S2 and S3).

Larger extraction windows could decrease the accuracy of quantification because of a higher risk of co-extraction of interfering compounds, but linear regression of measured versus theoretical quantities showed excellent linearity, with r^2 values for the interference corrected extracted ion current (XIC) of >0.92 for 75% of the peptides, as well as median slope of 0.89 ± 0.32 (median $\pm$ SD) and median intercept of -0.13 ± 0.18 (Supplementary Fig. S4). DIANA reports three measures that can be used to represent the quantity of a peptide: the XIC of the measured fragments, the XIC of the measured precursor isotopes and the interference-corrected XIC of the measured fragments. The interference-corrected XIC values gave higher r^2 values compared with the raw XICs of the fragments or the isotopes, and they also resulted in slopes and intercepts closer to their expected theoretical values (Supplementary Fig. S4). Therefore the interference corrected XIC is used as the measure of quantity throughout this article.

3.2 DIANA compared with gold standard

To evaluate the performance of DIANA, we analyzed the complete 60 DIA-MS gold standard data using the optimized parameters and original assays. The results were comparable to those from manual analysis performed previously (Röst *et al.*, 2014). At a 1% FDR, DIANA detected 7004 out of 7689 manually detected peptides in water, and 4786 out of 5716 peptides in yeast, representing sensitivities of 91.1% and 83.7%, respectively. A pseudo-roc curve of sensitivity versus false positive rate results in an area under the curve (AUC) of 0.92 (Fig. 3A). As the presumed correct peak in a few cases was not the highest scoring of the peaks for that assay, the sensitivity did not reach 1.0 even at the maximal score cutoff. The gold standard dataset also enabled the evaluation of the quality of FDR estimations. DIANA estimated the true FDR according to the gold standard reasonably well, with exact estimation at 10% FDR, underestimation for lower FDRs and overestimation for higher FDRs (Fig. 3B).

For the purpose of hypothesis-driven quantitative experiments, the precision and accuracy of a method is equally important to classification power. Precision calculations for DIANA and manual analysis yielded similar coefficients of variation (CV) across the technical replicates, with median CVs of 14.3% and 13.5% in water and 9.0% and 8.0% in yeast, respectively (Fig. 3C). Orthogonally, reported quantities from DIANA are also as accurate as manual analysis, and closely follow the theoretical dilutions. In log-log scale, 95% of peptide dilutions have $r^2 > 0.96$ in both water and yeast (0.964 and 0.968 with DIANA, 0.972 and 0.960 with manual) (Fig. 3D). Apart from r^2, the slope and intercept of a linear regression can be used to evaluate quantification. We normalized peptide quantities by division of the most concentrated sample followed by $\log_2$ transform. As theoretical $\log_2$ concentrations were set to [–9, –8, . . . , 0], slopes should theoretically be 1 and intercepts 0. Log-log scale intercepts had a median value of 0.18 ± 0.25 and 0.02 ± 0.10 (DIANA) compared with 0.19 ± 0.25 and 0.03 ± 0.11 (manual) for water and yeast backgrounds (Fig. 3E), whereas slopes were

Fig. 4. True FDR and number of identifications when combining search engines for DIA-MS analysis. Peptide ions were either detected as the same peak by both engines (green), as different peaks by the two engines (red), exclusively by DIANA (blue) or OpenSWATH (yellow), or not at all (gray). (**A**) True FDRs depending on detection status. Peptide ions selected at 1% FDR by both DIANA and OpenSWATH (green) had a true FDR of <0.5%, whereas the peptide ions exclusively quantified by only one engine (blue/yellow) had true FDRs of 20–27%. (**B**) Number of peptides ions quantified, stratified by detection status. The two engines agreed on a majority of the peptide peaks, but there are still exclusive contributions from both engines in both backgrounds

1.20 ± 0.46 and 1.01 ± 0.22 (DIANA) compared with 1.26 ± 0.45 and 0.98 ± 0.24 (manual) (Fig. 3F). We conclude that DIANA is well fit for targeted analysis of DIA-MS data, with high sensitivity and very accurate quantification.

3.3 DIANA compared with OpenSWATH

The novelty and utility of new algorithms should not only be evaluated compared with time-consuming gold standard manual analysis, but also by comparison with existing software. At a controlled FDR-level of 1%, DIANA reported similar peptide quantification results compared with the main existing software OpenSWATH (12174 versus 11932) in the gold dataset. This global trend was preserved across the entire dilution series in water, whereas the algorithms diverge in the yeast dilution, with DIANA performing better in the concentrated half, and OpenSWATH better in the diluted half (Fig. 4B and Supplementary Fig. S5).

The classification parameters precision and recall were used to compare the semi-supervised learning strategies of DIANA and OpenSWATH. Overall, the behavior over dilution in both parameters is similar, with a high precision (as forced by the target FDR of 1%), and a high recall that is declining with spiked peptide concentration (Supplementary Fig. S6). As indicated by the number of quantifications, the only difference lies in the recall of the yeast dilution series, where DIANA has a higher recall in the concentrated samples, while OpenSWATH has a higher recall in the diluted samples.

The corrected quantity measure of DIANA consistently yields minor increases in accuracy on peptides significantly and

correctly detected by both engines, compared with the OpenSWATH quantity measure (Supplementary Fig. S7). Both DIANA (DI) and OpenSWATH (OS) had high performance, with 95% of peptide dilutions having an $r^2 > 0.966$ (DI) and $r^2 > 0.963$ (OS) in water, and $r^2 > 0.972$ (DI) and $r^2 > 0.969$ (OS) in yeast. Intercepts were 0.17 ± 0.29 (DI) and 0.20 ± 0.25 (OS) in water, and 0.023 ± 0.10 (DI) and 0.028 ± 0.10 (OS) in yeast. Finally, slopes of the regressions were 1.18 ± 0.43 (DI) and 1.19 ± 0.45 (OS) in water and 1.01 ± 0.20 (DI) and 0.97 ± 0.22 (OS) in yeast.

3.4 Combination of DIANA and OpenSWATH

Previous studies have reported that the successful combination of multiple search engines improves both the number and quality of reported peptides detected in DDA data (Häkkinen *et al*., 2009; Jones *et al*., 2009; Nahnsen *et al*., 2011; Shteynberg *et al*., 2011). To investigate the possibility of similar performance gains in DIA data analysis, we studied the extent of overlap in reported peptides between DIANA and OpenSWATH. We observe that the confidence in the peptides identified by both engines is considerably increased. Using the gold standard manual analysis, the actual FDRs for the identification status groups could be calculated (Fig. 4A). Across all samples the agreeing identifications have actual FDRs of 0.5%, well below the target 1%. In contrast, the few conflicting identifications have actual FDRs ranging between 13% and 86%, with OpenSWATH being correct in a majority of cases, whereas single algorithm identifications have actual FDRs of 20–27% in both backgrounds. The vast majority of the identified peptides were detected in consensus by both engines and therefore in the high-confidence group (green), emphasizing the robustness of the targeted analysis strategy (Fig. 4B). Nonetheless, the total number of detectable peptides does increase when considering exclusive quantifications, and these could prove to be suitable targets for further study, for example by SRM.

3.5 Analysis of a bacterial lysate proteome using DIANA

To complement the strictly controlled setting of spiked-in synthetic peptides, we reanalyzed 4 MS injections of *S.pyogenes* grown with or without 10% human plasma from a previous study (Malmström *et al*., 2012), which was also used in the OpenSWATH manuscript (Röst *et al*., 2014). *Streptococcus pyogenes* is a major microbial pathogen, responsible for millions of cases of pharyngitis and 500 000 deaths annually (Carapetis *et al*., 2005). Apart from this very relevant reason for study, the bacterium's proteome of 1905 open reading frames makes it suitably complex for whole-proteome measurements. The data was evaluated using the pre-existing assay library (Röst *et al*., 2014), generated from 10 shotgun MS measurements of fractions of the *S. pyogenes* proteome, consisting of 1322 proteins represented by 20 027 proteotypic peptide precursors.

The reanalysis of the streptococcal lysates with DIANA yielded similar numbers of measured peptide ions at 1% FDR compared with OpenSWATH without inter-sample alignment, resulting in 38 776 versus 38 272 peptide ion identifications (Supplementary Fig S8). Together, 47 467 identifications were reported as significant by at least one algorithm. Of these, 29 397 (61%) peptide-ion identifications could be considered high confidence because of consensus identification by the two algorithms (Fig. 5). In addition, DIANA identified 8713 (18%) peptide ions exclusively, whereas OpenSWATH added another 8353 (18%). The algorithms gave conflicting results for only 824 (1.7%) peptide ions. According to the gold standard analysis we would expect consensus identifications to have a true FDR of 0.5% and single identifications to have a true FDR of about 20-27%. However, as the number of single identifications is larger compared with consensus identifications in this dataset, true FDRs are likely to be closer to 1% if the FDR estimates of algorithms are correct. We conclude that the combination of two search engines improves the total number of detected peptide ions at 1% FDR but also importantly increases the confidence for the majority of the detected peptides.

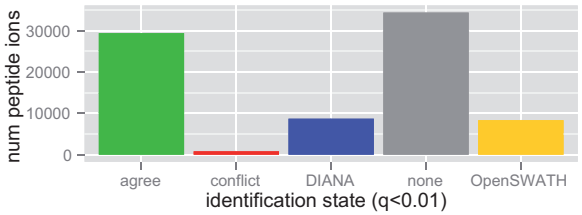

Fig. 5. DIA-MS analysis of four *Streptococcus pyogenes* lysates grown with 0% or 10% plasma supplement. Combined analysis using OpenSWATH and DIANA confirms close to 30–000 peptide ion quantifications, but each engine also quantifies over 8000 peptides exclusively. Less than 1000 peptide ions have conflicting quantifications from the two algorithms

4 DISCUSSION

The presented work demonstrates three advancements in targeted extraction DIA-MS analysis. First, the invention of a probabilistic score for fragmentation patterns is shown to give high analytical power in the complex bacterial and yeast backgrounds, which should be considered the minimal expected sample complexity for in vivo or cell-line studies. Second, the adopted interference-corrected measure of quantity from our previous SRM work is shown to provide increased accuracy in quantification in the noisy DIA-MS data. Third, we demonstrate the advance of combining two analysis tools for DIA-MS data processing. The combined output from DIANA and OpenSWATH generated both an increased number of identifications and considerably increased confidence in the peptides identified by both engines.

The DIANA algorithm is very reliant on the peptide fragmentation pattern, both for scoring and interference correction. This is both a strength and a weakness. The advanced probabilistic score is very powerful as the probabilities are individually calculated based on the noise in each specific ratio channel, and this allowed us to rely completely on extracted chromatograms for the analysis. On the other hand, the algorithm depends on conserved fragmentation, and changes in instrument collision energy tuning or mass-dependent ion transmission could hinder detection of true peptides. Nevertheless, we have demonstrated powerful classification (AUC < 0.92) and accurate quantification (95% of peptides have $r^2 > 0.96$) of the new scoring software DIANA and classifier PyProphet in a gold standard dataset. Further,

even if performance is largely similar, DIANA is shown to improve performance in samples from bacterial whole cell lysates with sufficient amounts of true positives, compared with OpenSWATH.

The structure of DIANA and OpenSWATH are conceptually similar. The observed differences in performance between the engines can likely be explained by the detailed differences such as the different sub-scores or the exact chromatogram extraction or peak detection. We believe that further improvement of DIANA could be achieved by including something similar to OpenSWATH's intensity and signal-to-noise sub-scores, as well as a preliminary score to initiate the semi-supervised learning better. The lack of such sub-scores could well explain DIANA's lower sensitivity in samples with very few true positives.

In the performed gold standard and streptococcal lysate analysis, we demonstrate the usefulness of utilizing multiple analysis tools, to increase the confidence and amounts of detected peptides. Increasing peptide identification rates and confidence using a combination of search engines is an attractive option, as it only requires computer hardware investments that are minor compared with instrument investment and maintenance costs. Being standard procedure in shotgun MS data analysis, we believe this study to validate the approach also in DIA-MS.

Although shotgun MS data analysis is a mature field with tens of different tools available, research on the analysis of targeted analysis DIA data has only begun. It can be anticipated that several powerful concepts for DIA analysis remain to be discovered. We believe DIANA demonstrates some such new analysis concepts, and their successful application to the complex task of detecting and quantifying peptide ions.

ACKNOWLEDGEMENTS

The authors thank Ufuk Kirik for the helpful discussions on the algorithms.

Funding: J.T. and J.M. were supported by the Swedish Research Council (projects 2008:3356 and 621-2012-3559), the Swedish Foundation for Strategic Research (grant FFL4), the Crafoord Foundation (grant 20100892), Stiftelsen Olle Engkvist Byggmästare, the Wallenberg Academy Fellow KAW (2012.0178) and European research council starting grant (ERC-2012-StG-309831). H.L.R. was funded by ETH (ETH-30 11-2). G.R. was funded by the Swiss Federal Commission for Technology and Innovation CTI (13539.1 PFFLI-LS). L.M. was support by ETH Zurich, Department of Biology, within the frame of an IT-strategy initiative. F.L. was funded by the Swedish Foundation for Strategic Research (RBb08-0006) and Mistra Biotech.

Conflict of interest: none declared.

REFERENCES

Bailey,D.J. *et al.* (2014) Intelligent data acquisition blends targeted and discovery methods. *J. Proteome Res.*, **13**, 2152–2161.

Bernhardt,O.M. *et al.* (2012) Spectronaut: a fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data. In: *Proceedings of 60th American Society for Mass Spectometry Conference* American Society for Mass Spectrometry, Vancouver 2012.

Carapetis,J.R. *et al.* (2005) The global burden of group A streptococcal diseases. *Lancet Infect. Dis.*, **5**, 685–694.

Chambers,M.C. *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, **30**, 918–920.

Collins,B.C. *et al.* (2013) Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nat. Methods*, **10**, 1246–1253.

Desiere,F. *et al.* (2006) The PeptideAtlas project. *Nucleic Acids Res.*, **34**, D655–658.

Deutsch,E.W. *et al.* (2012) TraML–a standard format for exchange of selected reaction monitoring transition lists. *Mol. Cell. Proteomics*, **11**, R111.015040.

Egertson,J.D. *et al.* (2013) Multiplexed MS/MS for improved data-independent acquisition. *Nat. Methods*, **10**, 744–746.

Elias,J.E. and Gygi,S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, **4**, 207–214.

Escher,C. *et al.* (2012) Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics*, **12**, 1111–1121.

Farrah,T. *et al.* (2012) PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics*, **12**, 1170–1175.

Gillet,L.C. *et al.* (2012) Targeted data extraction of the MS/MS spectra generated by data independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics*, **11**, O111.016717.

Häkkinen,J. *et al.* (2009) The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J. Proteome Res.*, **8**, 3037–3043.

Jones,A.R. *et al.* (2009) Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics*, **9**, 1220–1229.

Käll,L. *et al.* (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.*, **7**, 29–34.

Karlsson,C. *et al.* (2012) Proteome-wide selected reaction monitoring assays for the human pathogen *Streptococcus pyogenes*. *Nat. Commun.*, **3**, 1301.

Kost,J.T. and McDermott,M.P. (2002) Combining dependent P-values. *Stat. Probab. Lett.*, **60**, 183–190.

Liu,H. *et al.* (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.*, **76**, 4193–4201.

Malmström,J. *et al.* (2012) *Streptococcus pyogenes* in human plasma: adaptive mechanisms analyzed by mass spectrometry-based proteomics. *J. Biol. Chem.*, **287**, 1415–1425.

Martens,L. *et al.* (2011) mzML–a community standard for mass spectrometry data. *Mol. Cell. Proteomics*, **10,** R110.000133.

Nahnsen,S. *et al.* (2011) Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. *J. Proteome Res.*, **10**, 3332–3343.

Panchaud,A. *et al.* (2009) Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal. Chem.*, **81**, 6481–6488.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Picotti,P. *et al.* (2008) A database of mass spectrometric assays for the yeast proteome. *Nat. Methods*, **5**, 913–914.

Picotti,P. *et al.* (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell*, **138**, 795–806.

Plumb,R.S. *et al.* (2006) UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun. Mass Spectrom.*, **20**, 1989–1994.

Purvine,S. *et al.* (2003) Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics*, **3**, 847–850.

Reiter,L. *et al.* (2011) mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods*, **8**, 430–435.

Röst,H.L. *et al.* (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.*, **32**, 219–223.

Shteynberg,D. *et al.* (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics*, **10,** M111.007690.

Shteynberg,D. *et al.* (2013) Combining results of multiple search engines in proteomics. *Mol. Cell. Proteomics*, **12**, 2383–2393.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**, 479–498.

Tabb,D.L. *et al.* (2010) Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.*, **9**, 761–776.

Teleman,J. *et al.* (2012) Automated selected reaction monitoring software for accurate label-free protein quantification. *J. Proteome Res.*, **11**, 3766–3773.
Teleman,J. *et al.* (2014) Numerical compression schemes for proteomics mass spectrometry data. *Mol. Cell. Proteomics*, **13**, 1537–1542.
Vincent,C.E. *et al.* (2013) Segmentation of precursor mass range using 'tiling' approach increases peptide identifications for MS1-based label-free quantification. *Anal. Chem.*, **85**, 2825–2832.
Waldemarson,S. *et al.* (2012) Protein expression changes in ovarian cancer during the transition from benign to malignant. *J. Proteome Res.*, **11**, 2876–2889.
Weisbrod,C.R. *et al.* (2012) Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. *J. Proteome Res.*, **11**, 1621–1632.
Wolf-Yadlin,A. *et al.* (2007) Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc. Natl Acad. Sci. USA*, **104**, 5860–5865.