Combining Precursor and Fragment Information for Improved Detection of Differential Abundance in Data Independent Acquisition

Authors

Ting Huang, Roland Bruderer, Jan Muntel, Yue Xuan, Olga Vitek, and Lukas Reiter

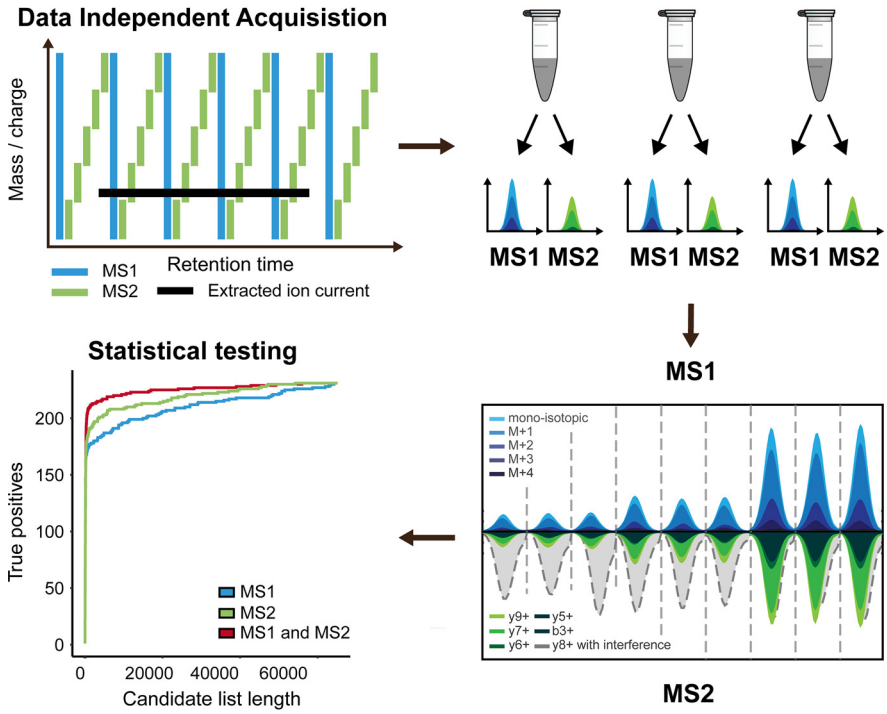
Correspondence

lukas.reiter@biognosys.com;
o.vitek@northeastern.edu

In Brief

DIA profiles of complex biological matrices such as tissues can contain quantitative interferences, and the interferences at the MS1 and the MS2 signals are often independent. We developed a statistical procedure incorporating both MS1 and MS2 quantitative information of DIA. We benchmarked the performance of the MS1-MS2-combined method to the individual use of MS1 or MS2 in DIA. In the majority of the comparisons, the combined method outperformed the individual use of MS1 or MS2.

Graphical Abstract



Highlights

- Modern DIA methods contain high quality MS1 and MS2.
- We developed a statistical procedure incorporating MS1 and MS2.
- Benchmarking, the combined method outperformed the individual use of MS1 or MS2.

Huang et al., 2020, Molecular & Cellular Proteomics *19*, 421–430
February 2020
https://doi.org/10.1074/mcp.RA119.001705

⌘ *Author's Choice*

Combining Precursor and Fragment Information for Improved Detection of Differential Abundance in Data Independent Acquisition*[S]

Ting Huang‡**, Roland Bruderer§**, Jan Muntel§, Yue Xuan¶, Olga Vitek‡‡‡, and Lukas Reiter§‖

In bottom-up, label-free discovery proteomics, biological samples are acquired in a data-dependent (DDA) or data-independent (DIA) manner, with peptide signals recorded in an intact (MS1) and fragmented (MS2) form. While DDA has only the MS1 space for quantification, DIA contains both MS1 and MS2 at high quantitative quality. DIA profiles of complex biological matrices such as tissues or cells can contain quantitative interferences, and the interferences at the MS1 and the MS2 signals are often independent. When comparing biological conditions, the interferences can compromise the detection of differential peptide or protein abundance and lead to false positive or false negative conclusions.

We hypothesized that the combined use of MS1 and MS2 quantitative signals could improve our ability to detect differentially abundant proteins. Therefore, we developed a statistical procedure incorporating both MS1 and MS2 quantitative information of DIA. We benchmarked the performance of the MS1-MS2-combined method to the individual use of MS1 or MS2 in DIA using four previously published controlled mixtures, as well as in two previously unpublished controlled mixtures. In the majority of the comparisons, the combined method outperformed the individual use of MS1 or MS2. This was particularly true for comparisons with low fold changes, few replicates, and situations where MS1 and MS2 were of similar quality. When applied to a previously unpublished investigation of lung cancer, the MS1-MS2-combined method increased the coverage of known activated pathways.

Since recent technological developments continue to increase the quality of MS1 signals (*e.g.* using the BoxCar scan mode for Orbitrap instruments), the combination of the MS1 and MS2 information has a high potential for future statistical analysis of DIA data. ***Molecular & Cellular Proteomics 19: 421–430, 2020. DOI: 10.1074/mcp.RA119.001705.***

Liquid chromatography-mass spectrometry (LC-MS)[1] has proven to be a powerful and versatile tool to quantify changes in protein abundance (1). In bottom-up proteomics, proteins are digested into peptides, which are then subjected to mass analysis. Two types of spectra are independently recorded: 1) the intact peptide mass (more precise *m/z*) or MS1 isotope envelope and 2) after fragmenting the peptide, the fragment ion spectrum (MS2). The accurate masses in the MS1 and MS2 spectra are used to identify and/or quantify the peptide. Unfortunately, MS1 and MS2 spectra can contain interferences that distort the quantitative signals. Interferences in MS1 spectra are typically caused by other coeluting peptide precursor isotope envelopes. Interferences in MS2 spectra are typically caused by fragments from coeluting peptides, which occur independently from the interferences in MS1 (2).

The two main applications of LC-MS in discovery-oriented investigations are data-dependent acquisition (DDA) and data-independent acquisition (DIA). In DDA, the MS1 precursor isotope envelope is used to generate extracted ion currents or three-dimensional-peak reconstructions for identification and quantification (3). Additionally, dependent on the MS1 scan, a limited number of peptide precursor peaks are isolated, fragmented, and subjected to secondary mass analysis. These MS2 scans are used for identification but do not contain direct quantitative information. They are not triggered at a defined point in the peptide precursor elution (Fig. 1*A*). In DDA with isobaric labeling, only MS2 information can be used for quantification, *e.g.* reporter ions with ITRAQ or tandem mass tag labels (4, 5) fragment with label remnants for tandem mass tag or easily abstractable sulfoxide-based isobaric-tag (6–8).

In contrast, in sequential window acquisition of all theoretical fragment ions (SWATH-type) DIA, MS1 and MS2 data are generated and recorded at a high enough frequency and quality to robustly sample the chromatographic peak (9, 10). A peak

From the ‡Northeastern University, Boston, MA 02115; §Biognosys, Wagistrasse 21, 8952 Schlieren, Switzerland; ¶Thermo Fisher Scientific, 28199 Bremen, Germany

Received August 23, 2019, and in revised form, December 16, 2019
Published, MCP Papers in Press, December 30, 2019, DOI 10.1074/mcp.RA119.001705

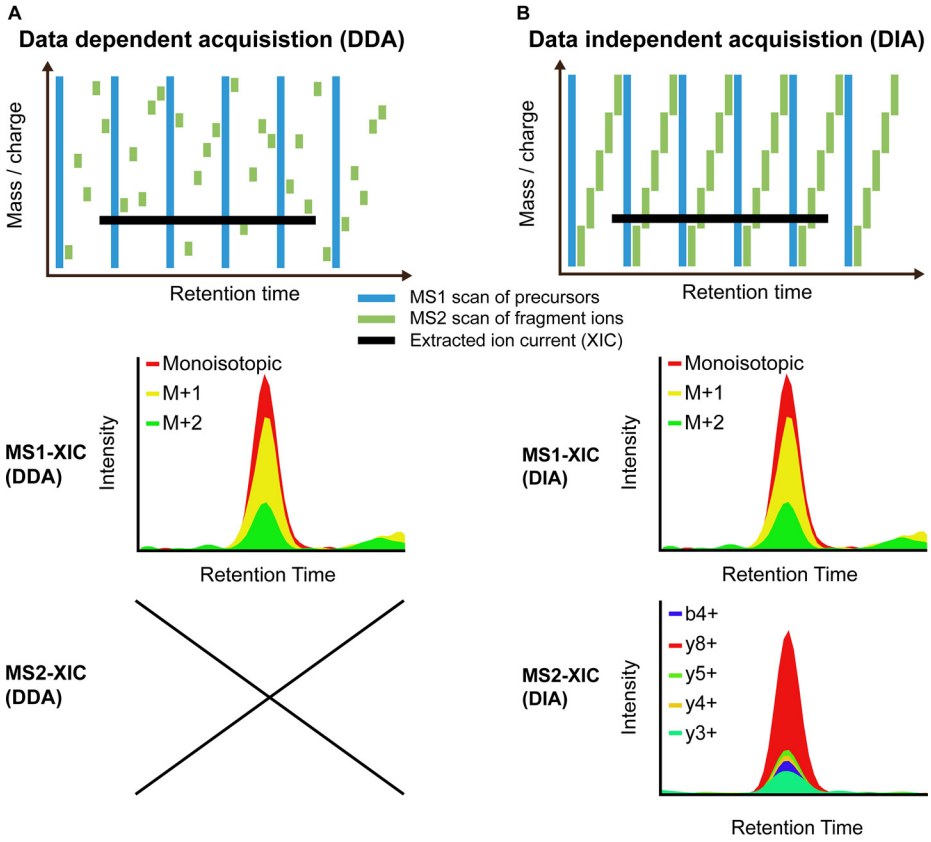


FIG. 1. **Quantitative data structure label-free discovery proteomics.** (*A*) Schematic representation of the acquisition layout of data-dependent acquisition methods with regular MS1 scans. The lower panels show the extracted ion current in MS1, which can be used for quantification. (*B*) Schematic representation of the acquisition layout of data-independent acquisition experiment with a regular MS1 and MS2 pattern. The lower panels show the two extracted ion currents, which can be used for quantification.

group of an extracted ion currents (precursor isotopes or fragments) exists for every peptide precursor in both the MS1 and MS2 space (Fig. 1*B*). In MS1, isotopic variants of an intact peptide precursor are present. In MS2, a precursor is fragmented into multiple fragment ions with isotopic variants. Hence, both quantitative spaces are fundamentally different, and the coeluting interferences do not correlate. Interferences on both levels can be corrected independently in order to increase signal to noise ratios and improve the detection and the quantification of peptide fragments. Several approaches currently perform such MS2 refinement for DIA. These include DIA-Umpire (11), Spectronaut (10), SWATHProphet (12), NOFI (13), TargetedMSQC (14), Encyclopedia (15), and Avant-garde (16). The improved quality of quantitative information after the interference correction can be relevant for projects such as those focusing on posttranslational modifications or peptidomics (2).

Traditionally in SWATH-type DIA, quantification relies on the MS2 information. An (optionally recorded) survey scan is followed by consecutive DIA segments covering the entire *m/z* range (9, 17). These settings are inherited from the initial development of SWATH-type DIA on Triple TOF instruments with moderate resolution, combined with the targeted data analysis strategy borrowed from selected reaction monitoring (*e.g.* in Spectronaut (10), OPENSWATH (18), Skyline (18)) and based on mProphet (19). Some researchers have also explored the value of MS1-level signals. Schilling *et al.* introduced MS1 data extraction from DIA with filtering and quantification but did not implement a procedure for characterizing the associated FDR, thereby limiting it to experiments where the numbers of peptides and runs are small enough to facilitate manual inspection (20). Rardin *et al.* (2) performed an exploratory analysis of MS1 and MS2 extracted ion currents of SWATH DIA data and found a strong quantitative correlation between the two. Specifically, they found that MS1 information can be especially relevant for studies of posttranslational modifications. DIA-Umpire (11) extracts the MS1 and MS2 information during a direct analysis of DIA data from a fasta database and correlates this information for identification using a search engine. In the Spectronaut software suite, MS1 and MS2 are fully implemented for identification and quantification (10).

Since recent progress has produced new MS instrumentation with higher resolution at faster speed, both the identification and the quantification of peptides in DIA can benefit from the increased quality of MS1 (21, 22). In particular, high quality and quantitative MS1 and MS2 data now enable statistical inference of differential peptide and protein abun-

[1] The abbreviations used are: LC-MS, liquid chromatography-mass spectrometry; DDA, data-dependent acquisition; DIA, data-independent acquisition; SWATH, sequential window acquisition of all theoretical fragment ions; MP, mixed proteomes; SFC and LFC, small/large fold change; OT, orbitrap; TTOF, triple quadrupole time of flight; CV, coefficient of variation; BH, Benjamini-Hochberg; FDR, false discovery rate.

dance. While this can be done separately for MS1 or MS2 (as *e.g.* is currently done in Spectronaut), it may be advantageous to simultaneously model MS1 and MS2 quantitative signals by viewing them as technical replicates from the same biological samples (Fig. 2*A*). To the date, the value of a direct join statistical modeling of MS1 and MS2 information for detecting differentially abundant proteins has not been systematically attempted and evaluated. This is due in part to the lack of controlled DIA datasets that make available both MS1 and MS2 quantitative information.

This manuscript contributes a statistical approach for the detection of differential abundant proteins that systematically leverages both MS1 and MS2 information. We applied this procedure to six sets of controlled mixtures. This was including mixtures with realistic biological background variation, and recorded on various instruments. A comparison of the performance of the MS1-MS2-combined method to the individual MS1- or MS2-based methods found consistent improvement in our ability to detect differentially abundant proteins, as judged by the quality of the candidate lists. The influence of the quality of the MS1 data was apparent, enabling the generation of optimal DIA methods for the combined use of MS1 and MS2. Finally, we applied the MS1-MS2-combined method to a clinical investigation and demonstrated that the MS1-MS2-combined method increased the coverage of known activated pathways.

EXPERIMENTAL PROCEDURES

Overview—We evaluated the impact of the use of MS1 and MS2 quantitative information in DIA using multiple datasets. To ensure the generality of the evaluation, we relied on six diverse sets of controlled mixtures with known ground truth and on a clinical investigation of lung cancer. Specifically, one set of controlled mixtures had defined spike-in of few proteins in a constant background (Spike-in-HEK293-OT dataset). Additionally, a set of controlled mixtures had defined spike-in of few proteins in a background with realistic biological variation (Spike-in-biol-var-OT) (23). The third type consisted of sets of controlled mixtures at defined ratios (MP-LFC-OT, MP-SFC-OT, MP-LFC-TTOF, and MP-LFC-MS1var-OT). Finally, we evaluated the MS1-MS2-combined method in a clinical investigation, comparing healthy tissues to tissues with lung cancer (BioIDS-OT). Data from these experiments were acquired on different instruments of two main classes (time of flight and Orbitrap). The quality of the candidate lists resulting from the statistical testing was used as a criterion for performance evaluation.

Sample Preparation for the Controlled Dataset with Biological Background Variation/Spike-in-biol-var-OT—To generate samples for the controlled mixtures with biological background variation (Spike-in-biol-var-OT), 25 mouse cerebellum samples were ordered from AMS Biotechnology (Abingdon, UK). For tissue lysis, half of a cerebellum was lysed in reducing lysis buffer (8 M urea, 0.1 M ammonium bicarbonate, 10 mM TCEP) in a bead mill (3 × 30 beats per second for 30 s, TissueLyser II, Qiagen, Hilden, Germany). To shear the DNA, lysates were sonicated in a Bioruptor (Diagenode, Seraing, Belgium) for five cycles at a high intensity (30 s on, 30 s off). The lysates were cleared by centrifugation (20 min, 16,000 × *g*). 60 μl of the lysate were used for digestion. For the alkylation of the samples, 60 μl of alkylation buffer (8 M urea, 0.1 M ammonium bicarbonate, 40 mM CAA) were added, and the samples were incubated for 1 h at 37 °C. Subsequently, samples were diluted with 600 μl of 0.1 M ammonium bicarbonate buffer including 5 μg of trypsin. Digestion was carried out overnight at 37 °C and constant shaking at 500 rpm. To stop the digestion, samples were acidified with 20% TFA. Peptide mixtures were purified using 96-well plate clean-up plates (NEST group, Southborough, MA) following the manufacturer's protocol. The samples were completely dried by vacuum centrifugation and resuspended in solvent A (1% acetonitrile, 0.1% formic acid in water) including iRT peptides (Biognosys, Schlieren, Switzerland). The UPS2 standard (Sigma, St Louis, MO) was digested separately using the same protocol, but MicroSpin columns were used for the cleanup (NEST group).

The concentrations of the cerebellum samples were adjusted to 1 μg/μl. Next, the samples were spiked five different concentrations of UPS2 standard, each in five different cerebellum samples (in total 25). Based on the lowest abundant UPS2 proteins, the spike-in concentrations were (assuming no losses during UPS2 sample preparation): S1: 0.75 amol/μl, S2: 0.83 amol/μl, S3: 1.07 amol/μl, S4: 2.04 amol/μl, and S5: 7.54 amol/μl.

For library generation, aliquots from all spiked samples were pooled and basified using ammonium hydroxide. The peptide pool was separated by high pH reverse-phase chromatography on a 2.1*150 mm Acquity CSH 1.7-μm column (Waters, MA) using a Dionex Ultimate 3000 LC (Thermo Scientific, Sunnyvale, CA) by a 30-min nonlinear gradient from 1% buffer B (100% acetonitrile)/99% buffer A (20 mM ammonium formate, pH 10) to 40% buffer B. A fraction was taken every 45 s and pooled into 10 final fractions.

Preparation of the Samples for the Controlled Dataset with Varying MS1 Resolution/MP-LFC-MS1var-OT—To generate the controlled mixtures with varying MS1 resolution (MP-LFC-MS1var-OT), HeLa, *Caenorhabditis elegans* and *Saccharomyces cerevisiae* digests were prepared as described before (21). Subsequently, two samples were generated with HeLa constant, *C. elegans* 30% change and *S. cerevisiae* 100% change to generate the samples for the controlled dataset with varying MS1 resolution (MP-LFC-OT).

Preparation of the Lung Cancer Samples/BioIDS-OT—For the clinical cancer dataset (BioIDS-OT), 12 nonsmall cell lung cancer and 12 matching normal adjacent tissue were purchased from Proteogenex (Culver City, CA). Around 30 mg per tissue were cut and lysed in lysis buffer (8 M urea, 0.1 M ammonium bicarbonate) in a bead mill (3 × 30 beats per second for 30 s, TissueLyser II, Qiagen, Hilden, Germany). DNA was digested by benzonase (Sigma-Aldrich, St. Louis, MO) treatment according to manufacturer's instructions. Lysates were cleared by centrifugation (20 min, 16,000 × *g*). 80 μl of the lysate were used for digestion. For reduction and alkylation of the samples, 80 μl of reduction/alkylation buffer (8 M urea, 0.1 M ammonium bicarbonate, 10 mM TCEP, 40 mM CAA) were added, and the samples were incubated for 1 h at 37 °C. Subsequently, the samples were diluted with 500 μl of 0.1 M ammonium bicarbonate buffer including 10 μg of trypsin. Digestion was carried out overnight at 37 °C and constant shaking at 500 rpm. To stop the digestion, samples were acidified with 20% TFA. Peptide mixtures were purified using 96-well plate cleanup plates (NEST group) following the manufacturer's protocol. The samples were completely dried by vacuum centrifugation and resuspended in solvent A (1% acetonitrile, 0.1% formic acid in water) including iRT peptides (Biognosys). Prior LC-MS analysis peptide concentrations were adjusted to 1 μg/μl. For library generation, aliquots from all samples were pooled (in total 200 μg) and basified using ammonium hydroxide. The peptide pool was fractionated by high pH reverse phase chromatography as described above. Pooled fractions were completely dried by vacuum centrifugation and resuspended in solvent A (1% acetonitrile, 0.1% formic acid in water) including iRT peptides (Biognosys).

LC/MS Acquisition of Spike-in-biol-var-OT, MP-LFC-MS1var-OT and BioIDS-OT—For DDA and DIA, 2 μg of each sample were separated using a self-packed analytical PicoFrit column (75 μm × 50 cm

length) (New Objective, Woburn, MA) packed with ReproSil-Pur 120A C18-AQ 1.9 μm (Dr. Maisch GmbH, Ammerbuch, Germany) with a 2-h segmented gradient using an EASY-nLC 1200 (Thermo Scientific). The datasets were acquired in a block randomized manner. The Spike-in-biol-var-OT and the MP-LFC-MS1var-OT were acquired on a Q Exactive HF mass spectrometer (Thermo Scientific) with methods modified from (21). The BiolDS-OT dataset was acquired on a Q Exactive HF-X mass spectrometer. The DIA method contained 43 DIA segments of 30,000 resolution with injection time set to auto and automatic gain control of $3*10^6$ and a survey scan of 120,000 resolution with 60ms max injection time and automatic gain control of $3*10^6$. The mass range was set to 350–1650 *m/z*. The default charge state was set to 3. Loop count 1 and normalized collision energy stepped at 25.5, 27, and 30. For the dataset with varying MS1 resolutions, the number of DIA segments was adapted to maintain a constant method cycle time. For the acquisition of the fractionated sample for the library, a DDA method was applied. The TOP15 method was modified from (24) (MS-Methods.xlsx).

Mass Spectrometric Data Analysis of Spike-in-biol-var-OT, MP-LFC-MS1var-OT and BiolDS-OT—DIA data were analyzed with Spectronaut Pulsar X 12.0.20491.6, (Biognosys (10)). The default settings were used for the targeted analysis of DIA data in Spectronaut. The initial mass tolerance for MS1 and MS2 was 15 ppm. High precision iRT calibration was used (25). The analysis was performed with and without the built-in interference correction (10). The FDR was calculated according to (19). The DDA spectra were analyzed with the MaxQuant (Version 1.5.6.5) analysis software (26, 27) using default settings (Trypsin/P, two missed cleavages). The search criteria included carbamidomethylation of cysteine as a fixed modification and oxidation of methionine and acetyl (protein N terminus) as variable modifications. The initial mass tolerance for the precursor was 4.5 ppm and for the fragment ions was 20 ppm. The Spike-in-biol-var-OT DDA were searched against the mouse isoform UniProt fasta database (state 11.12.2014, 24,723 entries) and the Biognosys iRT peptides fasta database (uploaded to the public repository). The DDA from the BiolDS-OT dataset was searched against the UniProt fasta database (state 11.12.2014, 20,215 entries) and iRT fasta. The library was generated in Spectronaut by importing the MaxQuant search results using the default settings. Supplemental Table S1 shows the number of entries in the libraries.

Analysis of the Spike-in-biol-var-OT Dataset—The dataset was normalized in Spectronaut Pulsar X by the default normalization option. The blood proteins ALBU_MOUSE, ALBU_HUMAN, TRFE_MOUSE, and TRFE_HUMAN, were removed due to the interference between the spiked-in proteins and the background proteins. Precursor and protein FDR were set to 1%. Since the detection limit of MS1 signal was around 100, normalized MS1 intensities below 100 were considered as missing values. Precursors with any missing MS1 intensities or MS2 intensities over all the MS runs were filtered out.

Analysis of the Spike-in-HEK293-OT Dataset—The published controlled dataset from Bruderer *et al.* (10) was analyzed with Spectronaut Pulsar X using default settings using the published library. The dataset was normalized in Spectronaut Pulsar X by the default normalization option. Shared peptides of the spike-in proteins with the human background were removed. Four proteins, P07724, Q921I1, P02768, and P02787, were removed due to the interference between the spiked-in proteins and the background of human proteins. The three replicates from group S8 were not used for statistical testing (28). Next, the data were filtered by FDR and intensity thresholds like the Spike-in-biol-var-OT dataset.

Analysis of the MP-LFC-OT, MP-SFC-OT, MP-LFC-TTOF, MP-LFC-MS1var-OT, and BiolDS-OT Datasets—The published mixed proteome controlled datasets MP-LFC-OT and MP-SFC-OT from Bruderer *et al.* (21) and MP-LFC-TTOF from Navarro *et al.* (29) were analyzed with Spectronaut Pulsar X using default settings. The MP-LFC-MS1var-OT dataset was analyzed with Spectronaut Pulsar X using default settings using the spectral libraries from Bruderer *et al.* (21). The BiolDS-OT dataset was analyzed with Spectronaut Pulsar X using default settings using the project spectral library described above. All the datasets (MP-LFC-OT, MP-SFC-OT, MP-LFC-TTOF, MP-LFC-MS1var-OT, and BiolDS-OT) were normalized based on housekeeping proteins using a global median approach for MS1 and MS2 separately (30) (supplemental Table S2). Shared peptides between the proteomes were removed. Next, the data were filtered by FDR and intensity thresholds, as in the Spike-in-biol-var-OT dataset.

Statistical Modeling—For a given protein, let X_{iprg} be the log_2 intensity of peptide precursor ion p in replicate r from group g, where $i\epsilon\{1, 2\}$, $p\epsilon\{1, \ldots, P\}$, $r\epsilon\{1, \ldots, R\}$, and $g\epsilon\{1, \ldots, G\}$. The index $i = 1$ indicates that X_{1prg} is estimated from MS1 signal, and $i = 2$ indicates that X_{2prg} is estimated from MS2 signal.

The methods of statistical analysis are summarized in supplemental Fig. S1. When separately analyzing MS1-based quantification, we first summarized the protein abundances in an LC-MS run, by calculating the median MS1 intensities across all its matching peptide ions. In other words, the summarized MS1 intensity of one protein is $Z_{1rg} = \text{median}_p\{X_{1prg}\}$, the median of all X_{1prg} across all peptide ions p, in replicate r from group g. Next, since all the experiments in this manuscript had a multiple group design, we fit a one-way analysis of variance (ANOVA) model:

$$Z_{1rg} = \mu + Group_{1g} + \varepsilon_{1rg} \quad \text{(Eq. 1)}$$

$$\sum_{g=1}^{G} Group_{1g} = 0$$

$$\varepsilon_{1rg} \sim N(0, \sigma_1^2)$$

In this notation, $Group_{1g}$ describes the deviation of the MS1-based expected protein abundance in group g from the average expected protein abundance in all the groups. This term is of the main interest in MS1-based quantification. The term ε_{1rg} is the random experimental error with mean 0 and a constant variance. Hypothesis testing for differential abundance was performed based on this model. *P*-values of all the proteins were adjusted for multiple testing by the method of Benjamini and Hochberg (31).

We analyzed MS2-based quantification using the same method (supplemental Fig. S1).

$$Z_{2rg} = \mu + Group_{2g} + \varepsilon_{2rg} \quad \text{(Eq. 2)}$$

$$\sum_{g=1}^{G} Group_{2g} = 0$$

$$\varepsilon_{2rg} \sim N(0, \sigma_2^2)$$

Here Z_{2rg} is the summarized MS2 intensity of the same protein, *i.e.* $Z_{2rg} = \text{median}_p\{X_{2prg}\}$, the median of all X_{2prg} across all peptide ions p, in replicate r from group g. $Group_{2g}$ describes the deviation of the MS2-based expected protein abundance in group g from the average expected protein abundance in all the groups. This term is of the main interest in MS2-based quantification. The term ε_{2rg} is the random experimental error with mean 0 and a constant variance. *p*-values of all the proteins were also adjusted for multiple testing by the method of Benjamini and Hochberg.

In order to jointly analyze the MS1 and MS2 precursor intensities in a statistical model, we first separately normalized MS1 and MS2 intensities of each peptide precursor ion to have zero median across

all the replicates. That is, $X'_{iprg} = X_{iprg} - \text{median}_{rg}\{X_{iprg}\}$ is the normalized log_2 intensity of peptide precursor ion p in replicate r from group g. Then normalized protein intensity $Z'_{irg} = \text{median}_p\{X'_{iprg}\}$, the median of all X'_{iprg} across all peptide ions p, in replicate r from group g. Next, we extended the ANOVA models in Equations (1) and (2) above, to express both MS1 and MS2 signals.

$$Z'_{irg} = \mu + MS_i + Group_g + Replicate_{r(g)} + \varepsilon_{irg} \quad \text{(Eq. 3)}$$

$$\sum_{i=1}^{2} MS_i = 0$$

$$\sum_{g=1}^{G} Group_g = 0$$

$$Replicate_{r(g)} \sim N(0,\sigma_R^2)$$

$$\varepsilon_{irg} \sim N(0,\sigma^2)$$

In this notation MS_i is the contribution of MS1 signal and MS2 signal to the estimate of protein abundance Z'_{irg}. $Group_g$ describes the deviation of the expected protein abundance in group g from the average expected protein abundance in all the groups, on average over MS1 and MS2 signals. This term is of the main interest in joint MS1- and MS2-based quantification. $Replicate_{r(g)}$ expresses the variability of protein abundance in replicate r within group k. ε_{irg} is the random experimental error with mean 0 and a constant variance. As above, hypothesis testing for differential abundance was performed based on this model. Importantly, the combination of MS1 and MS2 signals allows us to distinguish the sources of biological and technological variation (supplemental Fig. S1). p-values of all the proteins were adjusted for multiple testing by the method of Benjamini and Hochberg.

We evaluated the performance of the statistical methods in the six controlled mixtures by sorting the candidate lists by p-value and examining the true positives as a function of candidate list length. Additionally, for simpler visualization, cuts were taken at fixed candidate list lengths (Top200 for spike-in and Top2000 for proteome mixtures). Finally, the candidate lists were analyzed in terms of the number of true positives and false positives determined based on the ground truth. The BioIDS-OT was evaluated by comparing the candidate lists against an independent study of the same cancers in lung (32). The R script for the analysis was uploaded to the proteomeXchange repository.

RESULTS

Separately Characterizing the MS1 and MS2 Quantification in DIA—Before integrating the MS1 and MS2 in the statistical modeling, we sought to separately characterize the strengths and weaknesses of the quantitative information of MS1 and MS2. In each dataset, we visually inspected the peptide precursor and fragment signals of the truly differentially abundant proteins. We compared the precision, accuracy, and correlation of precursor quantification by MS1 and MS2 globally across all the proteins.

Visual inspection of the DIA data using Spectronaut showed that the MS1 and MS2 interferences occur independently. In the controlled mixtures, interferences of peptide precursors with differential abundance can be clearly spotted because they do not follow the expected patterns of abundance between conditions. In an example with interference on MS1, a coeluting peptide was of the same *m/z* (Fig. 2*B*, *upper panel*) In an example with interference on MS2, a fragment of a coeluting and cofragmented peptide precursor was of the same mass (Fig. 2*B*, *lower panel*).

To globally assess the precision of MS1 and MS2 quantification, we calculated the CVs on the precursor level for the controlled mixtures within conditions. For all the tested datasets, the precision of MS2 quantification was higher than that of MS1, when counting precursors with CVs <20%. The number of precursors with CVs below 20% (on MS2) was between 11 and 61% higher than MS1 (Fig. 2*B*). Consistently, the medians of the CVs of MS2 quantification were consistently lower than those of MS1 (supplemental Fig. S3*A*). The CVs of precursors on MS1 had a median between 9% and 25%, and CVs on MS2 it was between 7% and 15%.

In the controlled mixtures based on mixed proteomes acquired on the Orbitrap instruments, the accuracy of fold change estimation was similar between MS1 and MS2 (supplemental Fig. S3*B*). In the MP-LFC-TTOF dataset recorded on the time of flight instrument, the fold change was more compressed in the MS1 space than in MS2, while the fold change estimate of the combined MS1 and MS2 information was between the estimates of the individual (supplemental Fig. S3*B*). The Pearson correlation between MS1 and M2 quantification had a median value of 0.5 (Fig. 2*C*). These analyses demonstrated that, although MS1 and MS2 are of generally similar quality, they can be of variable quality for specific analytes.

Combining the Quantitative Information in MS1 and MS2—Our next step was to evaluate the ability of the proposed MS1-MS2-combined method to statistically detect differentially abundant proteins in the controlled mixtures. First, we compared the true positive and false positive differentially abundant proteins (as defined by the ground truth), detected by MS1 alone, by MS2 alone, and by the MS1-MS2-combined method. The MS1-MS2-combined method always outperformed the individual tests (Fig. 3*A*, supplemental Fig. S4*A*, supplemental Table S4, and statistical-inference-results.zip). In the controlled mixtures with spike-in proteins, the top differentially abundant proteins consisted mostly of the true positives (the number of true positives among the top 200 differentially abundant proteins were as follows. Spike-in-HEK293-OT, MS1: 129, MS2: 160, and MS1-MS2: 164; Spike-in-biol-var-OT, MS1: 72, MS2: 111, and MS1-MS2: 113). In the proteome mixtures, the performance of the MS1-MS2-combined method was similar or better than the individual (Fig. 3*B*, supplemental Fig. S4*B*, supplemental Table S4, and statistical-inference-results.zip).

In order to distinguish the role of increasing the number of replicates from that of reducing the undesirable effects of interferences in MS1 or MS2 signals, we reanalyzed the datasets with and without interference correction implemented in Spectronaut. The interference correction improved the statistical power in all the datasets (supplemental Fig. S5*A*). The combination of MS1 and MS2 information could mitigate the negative impact of interferences. The performance of the combined

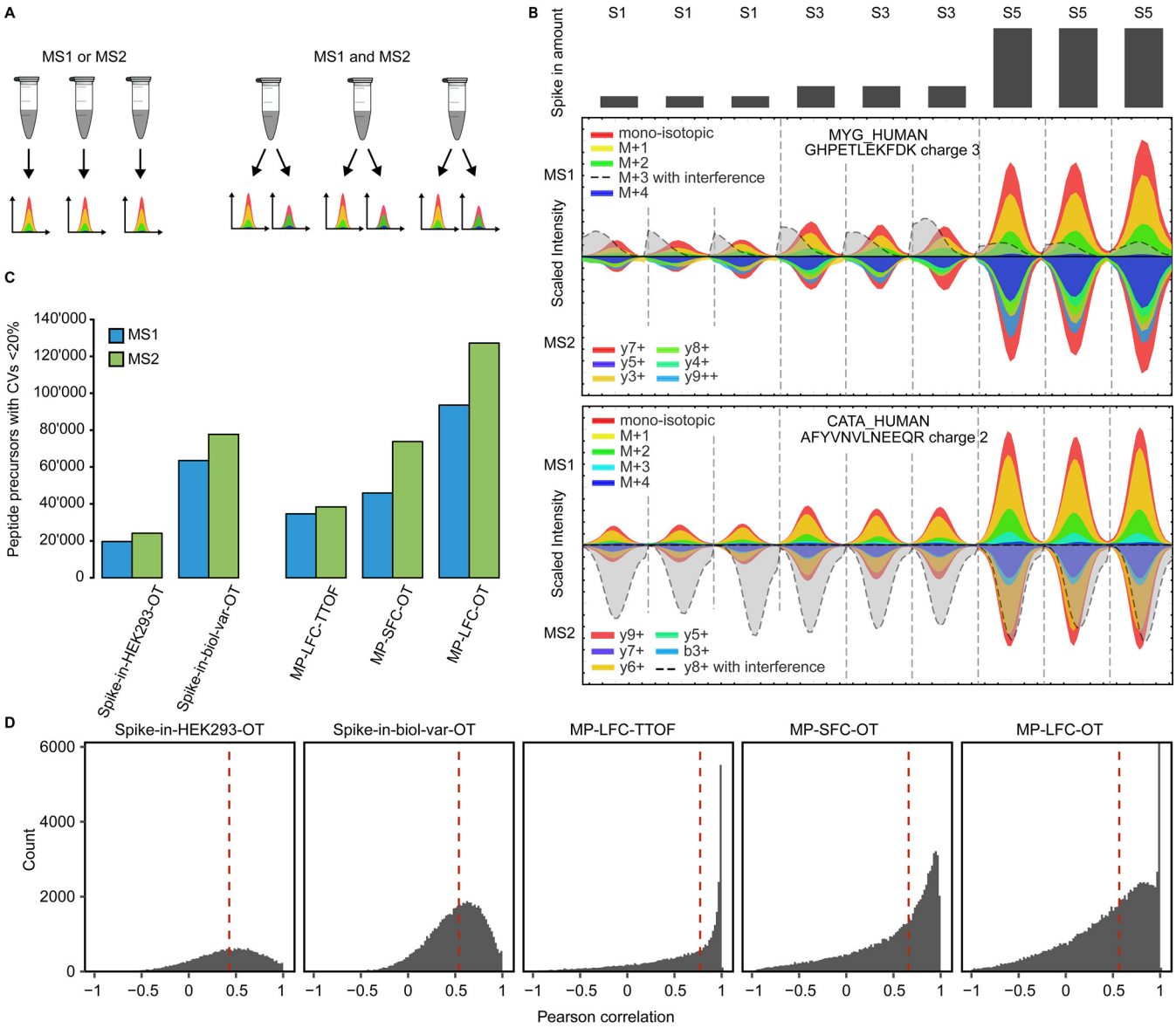


FIG. 2. **MS1 and MS2 quantification characteristics in DIA** (*A*) The MS1 and MS2 quantitative signals can be viewed as technical replicates from the same biological samples. (*B*) Extracted ion currents of two peptides derived from spike-in proteins from the Spike-in-biol-var-OT dataset of sample 1, 3, and 5. The interferences were manually identified as not following the predefined pattern of differential abundance. (*C*) The CVs for all precursors on condition level were calculated for the controlled datasets separately for MS1 and MS2 and separately for each condition. The graph displays the counts of precursors with CVs below 20%. (*D*) Pearson correlation between the precursor abundances in MS1 and MS2 space in the controlled datasets. The median is indicated by the red dotted line.

method without interference correction was comparable to the performance of MS2 alone with interference correction in all but one set of controlled mixtures (supplemental Fig. S5*B*).

Next, we analyzed the effect of the magnitude of the fold changes on the performance of the statistical models. We split the pairwise comparisons of the Spike-in-biol-var-OT into subsets with decreasing maximal fold changes from 900% to 10% (Fig. 3*C*). Additionally, for the controlled mixtures Spike-in-HEK293-OT and MP-SFC-OT (supplemental Fig. S6), the MS1-MS2-combined method performed better with smaller fold changes than the individual MS1- or MS2-based quantification. Detecting small fold changes proved challenging because the observed maximum of true positives was reached more slowly in all the three methods. This challenge also manifested itself by the earlier deviation from the perfect candidate list.

We further evaluated the effect of replication on the outcome of the statistical analyses. We used the three statistical approaches to analyze the Spike-in-biol-var-OT dataset with a decreasing number of replicates. The MS1-MS2-combined method maintained higher statistical power than the individual tests (Fig. 3*D*). At the lowest number of replicates (two), the

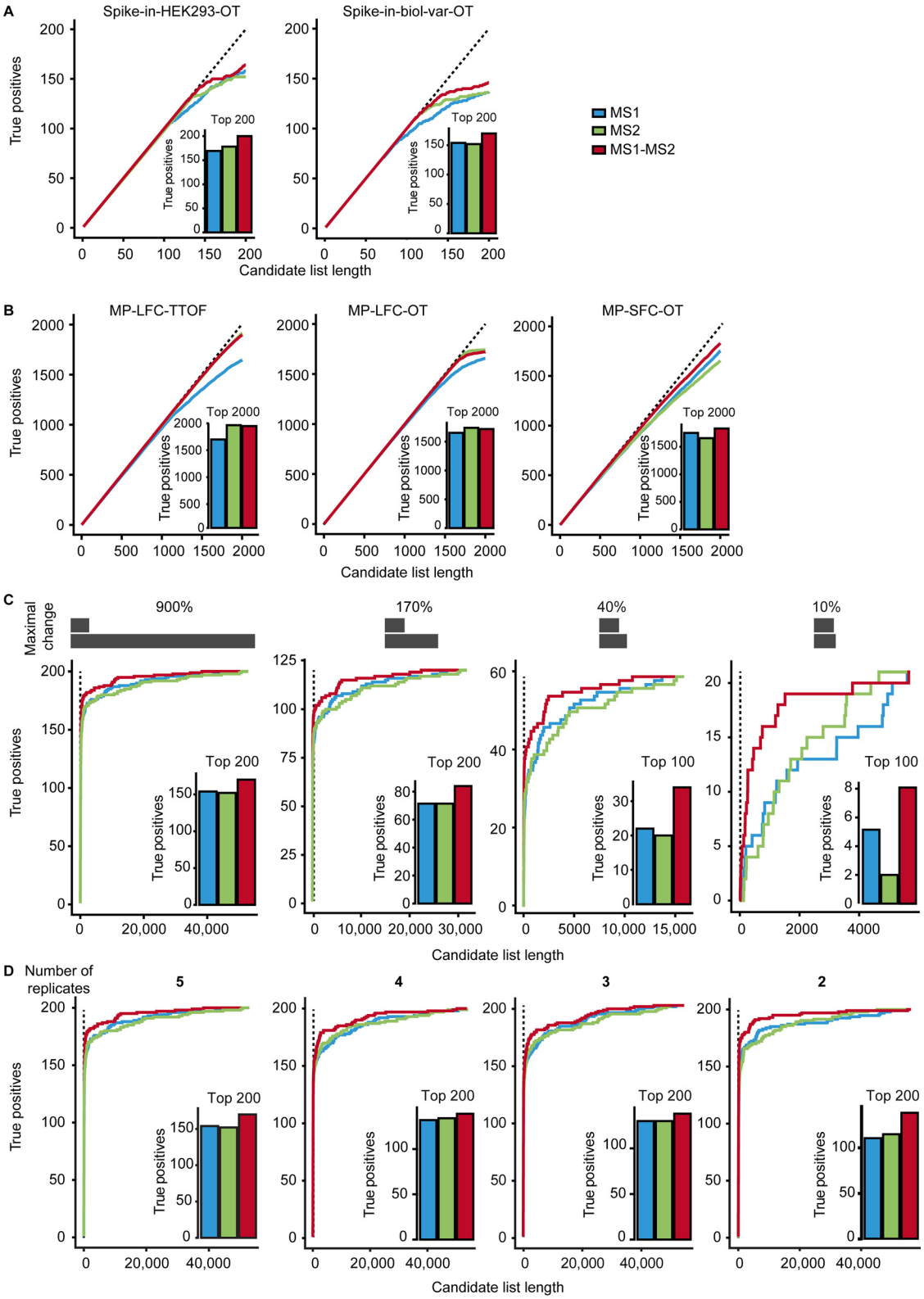


FIG. 3. **Benchmarking of the MS1-MS2 combined method** (*A*) Statistical inference of differential abundance was performed for spike-in datasets. The 200 proteins with the smallest adjusted *p* values were sorted by their *p* value. Next, the number of true positive differentially abundant proteins was displayed as a function of the candidate list containing true and false positives. The dotted line indicates a perfect

MS1-MS2-combined method had a 13% better sensitivity as compared with MS1 and 17% as compared with MS2.

An important parameter of DIA is the time allocated to MS1 and MS2, respectively (thus affecting the resolution for Orbitrap instruments). Because current DIA methods have been mostly optimized for MS2-based quantification, we evaluated the influence of the experimental MS1 resolution on an Orbitrap mass spectrometer on the MS1-MS2-combined statistical procedure. The controlled mixture MP-LFC-MS1var-OT was profiled with DIA while varying MS1 resolutions (30,000 to 240,000) and balancing MS2 time (at a constant resolution of 30,000) to keep the cycle time of the methods constant. We then characterized the quantitative precision, as well as the detection of differentially abundant proteins in these settings (supplementary File MP-LFC-MS1var-OT.zip). The DIA method with 120,000 MS1 resolution consistently produced most precursors with CVs below 20%, both for MS1 and MS2 (supplemental Fig. S7*A*). This can be explained by the fact that peak picking and integration in Spectronaut is dependent on MS1 and MS2. For all three statistical models, the best candidate lists were obtained at the 120,000 MS1 resolution. The MS1-MS2-combined approach showed the best overall performance, likely due to a sufficient MS1 resolving power and a balance in the number of MS2 segments (supplemental Figs. S7*B* and S7*C*). This is practical because it corresponds to the widely used DIA methods.

Testing Proteins for Differential Abundance in the Set of Clinical Samples with Lung Cancer—Finally, we evaluated the performance of the MS1-MS2-combined method in a clinical investigation of 12 healthy lung and 12 cancer samples (six adenocarcinomas and six squamous cell carcinomas). We performed an exploratory analysis, statistical analysis of differential abundance using the MS1- or the MS2-based method or the MS1-MS2-combined method (supplementary File BioIDS-OT.xlsx), and biological pathway analysis using ingenuity pathway analysis (IPA) (33). Principal component analysis-based clustering revealed a clear separation of healthy and tumor samples, indicating the biological separation between the two sample sets (Fig. 4*A*). The MS1-MS2-combined method produced the largest list of differentially abundant proteins (multiple testing correction BH, FDR $<1\%$). The three methods shared 65% of the union of differentially abundant proteins (Fig. 4*B*). The MS2-based and the MS1-MS2-combined method showed a large portion of unique candidates, respectively (7.5% and 5.8% of the union). The results of the MS1-MS2-combined method had a larger overlap with the MS1 approach, missing 79 candidates. The MS2-based method missed 238 differentially abundant proteins that were reported by MS1.

Because the true differentially abundant proteins are unknown, we compared our results to the results of an independent study based on the same cancer type by Tenzer *et al.* (32). The results of the MS1-MS2-combined method had the largest overlap with Tenzer *et al.* The overlap had 70 proteins more than the overlap based on the MS2 method (and 79 more than the MS1-based) (Fig. 4*C*). Upon investigation of pathway enrichments in the candidate lists, we found that among the 64 most enriched pathways there was a high degree of overlap between all three different methods. The pathways were consistently either activated or deactivated (Fig. 4*D*). The proteins uniquely identified by the MS1-MS2-combined method belonged to the same pathways, such as the pathways for the shared proteins, indicating a more comprehensive description (supplemental Fig. S8).

DISCUSSION

In complex mixtures, peptides of the same mass can coelute, and the coeluting peptides can share fragments of the same mass. Modern DIA experiments allow us to characterize these two independent quantitative spaces. The combined use of MS1- and MS2-level information increases the number of technical replicates and, therefore, the precision of the measurement. The improved precision, combined with the ability to separate the sources of biological and technological variation by the statistical modeling, in turn improve the power of detecting differentially abundant proteins.

Moreover, because interferences in MS1 and MS2 usually do not correlate, another strength of the combined use of MS1 and MS2 is in reducing the negative impact of the interferences in one of the quantitative spaces on the downstream statistical analysis (Supplemental Fig. S5).

To take advantage of both layers of quantification, we developed a statistical model that combines the use of MS1 and MS2. We demonstrated the advantages of this approach on multiple sets of controlled mixtures from multiple instrument platforms. We noticed the largest improvement in the detection of differential abundance for small fold changes, a finding that is particularly relevant for plasma studies and in cases of a low number of replicates (34–36). On a clinical investigation of lung cancer, the proteins uniquely identified by the MS1-MS2-combined method increased the coverage of the pathway enriched by MS1- and MS2-based quantification. Despite this improvement, the MS1-MS2-combined method is not a substitute for adequate biological replication.

candidate list containing only true positives (slope = 1). *Inset*: the number of true positives in the list of 200 proteins with the smallest adjusted *p* values. (*B*) As in (*A*), but for the mixed proteome datasets as in (*A*). (*C*) Statistical detection of differentially abundant proteins was performed as above for subsets of the Spike-in-biol-var-OT dataset with decreasing maximal true fold change, by selecting subsets of the dataset. The first plot to the left is based on the samples 1 to 5, the second to the left on 1 to 4, the third from the left on 1 to 3, and the right plot shows 1 to 2. The resulting candidate lists were analyzed as above. (*D*) Statistical detection of differentially abundant proteins was performed as above for subsets of the Spike-in-biol-var-OT dataset with decreasing numbers of replicates. The resulting candidate lists were analyzed as above.

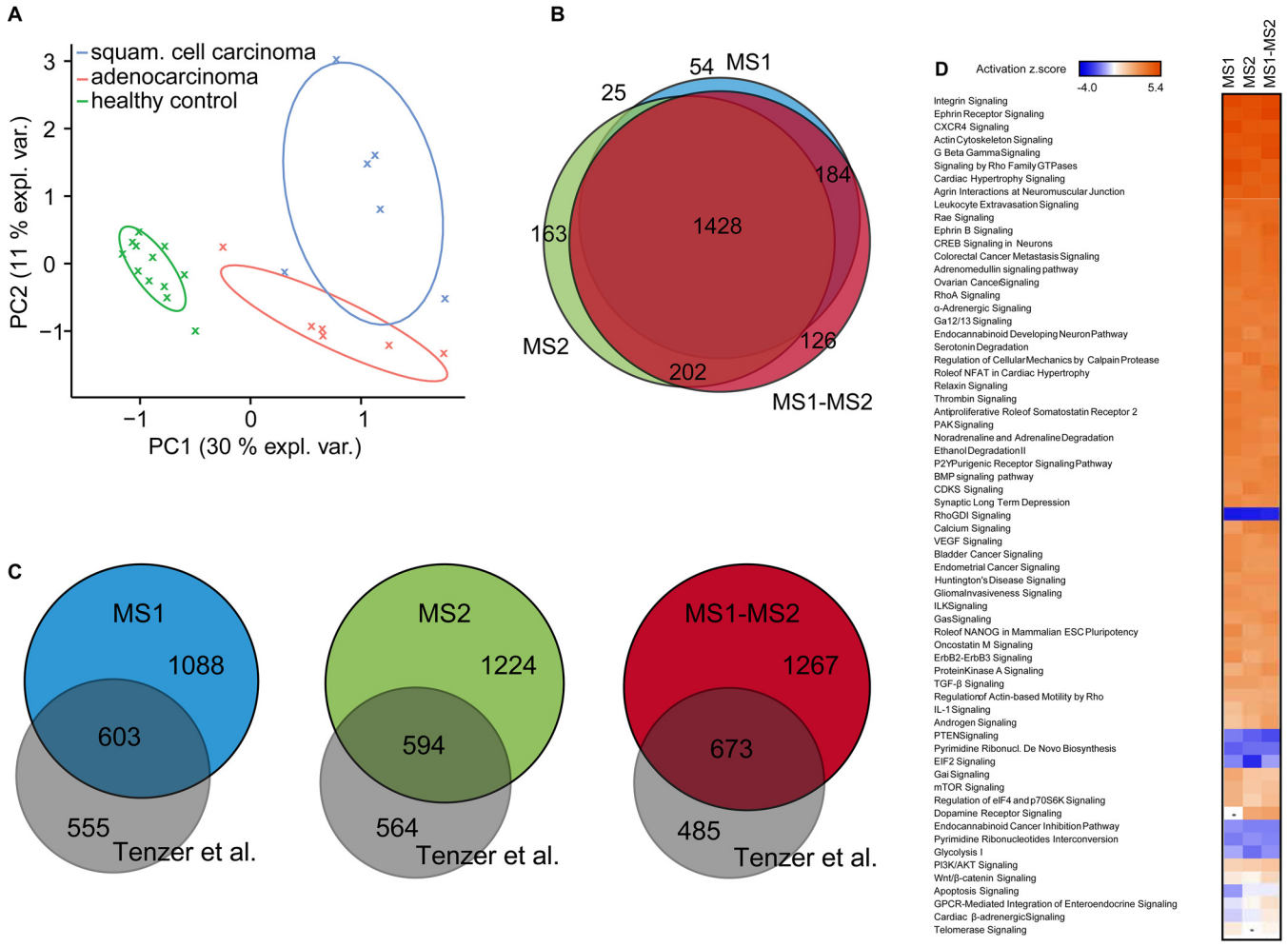


FIG. 4. **MS1-MS2-combined method based differential abundance testing in clinical samples** (*A*) 12 healthy lung and 12 cancer (six adenocarcinomas and six squamous cell carcinomas) were analyzed by mass spectrometry. The resulting data were subjected to principal component analysis. (*B*) Statistical detection of differentially abundant proteins was performed with MS1-, MS2-based and the MS1-MS2-combined method. The overlap of differentially abundant proteins (FDR <0.05) was calculated on the protein level. (*C*) The candidate lists from the testing approaches were compared with the candidate list of an independent lung cancer study by Tenzer *et al.* (32). (*D*) The functional analyses were generated through the use of IPA (33). The figure plots the activation states of the pathways according to IPA.

We believe that the proposed approach may have a high potential in the future as the technology evolves. For example, the recent development of BoxCar MS1 acquisition (37) improves the intrascan dynamic range, which leads to better MS1 quantification. Therefore, it is conceivable that the approach presented here will become even more powerful and will lead to sensitive and robust statistics when combining BoxCar MS1 and DIA.

Acknowledgment—We want to thank Nigel Beaton for proofreading.

DATA AVAILABILITY

The raw mass spectrometric data, the spectral libraries, and the quantitative data tables have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (38) with the dataset identifier PXD016647. The saved projects from Spectronaut can be reviewed with the Spectronaut Viewer (www.biognosys.com/spectronaut-viewer).

* This project has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 686282. Jan Muntel was supported by an innovation project grant from Innosuisse (Project Number 18365.2 PFLS-LS). Competing financial interests: The authors R.B., J.M., and L.R. are employees of Biognosys AG (Zurich, Switzerland).

S This article contains supplemental material files BioIDS-OT.xlsx, MS-Methods.xlsx, and statistical-inference-results.zip; Tables S1, S2, and S4; and Figs. S1 and S3-S7.

** Authors contributed equally.

‖ To whom correspondence may be addressed. Tel.: +41 (0) 44 738 20 40; Email: lukas.reiter@biognosys.com.

‡‡ To whom correspondence may be addressed. Tel.: +41 (0) 44 738 20 40; Email: o.vitek@northeastern.edu.

Author contributions: T.H., R.B., Y.X., O.V., and L.R. designed research; T.H. and R.B. performed research; T.H. and R.B. analyzed data; T.H. and R.B. wrote the paper; and R.B. and J.M. contributed new reagents/analytic tools.

REFERENCES

1. Aebersold, R., and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355
2. Rardin, M. J., Schilling, B., Cheng, L. Y., MacLean, B. X., Sorensen, D. J., Sahu, A. K., MacCoss, M. J., Vitek, O., and Gibson, B. W. (2015) MS1 Peptide ion intensity chromatograms in MS2 (SWATH) data independent acquisitions. Improving post acquisition analysis of proteomic experiments. *Mol. Cell Proteomics* **14,** 2405–2419
3. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
4. Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., and Hamon, C. (2003) Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75,** 1895–1904
5. Unwin, R. D., Pierce, A., Watson, R. B., Sternberg, D. W., and Whetton, A. D. (2005) Quantitative proteomic analysis using isobaric protein tags enables rapid comparison of changes in transcript and protein levels in transformed cells. *Mol. Cell Proteomics* **4,** 924–935
6. Wühr, M., Haas, W., McAlister, G. C., Peshkin, L., Rad, R., Kirschner, M. W., and Gygi, S. P. (2012) Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. *Anal. Chem.* **84,** 9214–9221
7. Sonnett, M., Yeung, E., and Wühr, M. (2018) Accurate, sensitive, and precise multiplexed proteomics using the complement reporter ion cluster. *Anal. Chem.* **90,** 5032–5039
8. Virreira Winter, S., Meier, F., Wichmann, C., Cox, J., Mann, M., and Meissner, F. (2018) EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nat. Methods* **15,** 527–530
9. Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (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
10. Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinović, S. M., Cheng, L. Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C., Vitek, O., Rinner, O., and Reiter, L. (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell Proteomics* **14,** 1400–1410
11. Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A. C., and Nesvizhskii, A. I. (2015) DIA-Umpire: Comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12,** 1–14
12. Zirlik, A., Htun, N., Iphöfer, A., Jänsch, L., and Mischak, H. (2015) Automated validation of results and removal of fragment ion interferences in targeted analysis of data independent acquisition MS using SWATHProphet. *Mol. Cell Proteomics* **14,** 1411–1418
13. Bilbao, A., Zhang, Y., Varesio, E., Luban, J., Strambio-De-Castillia, C., Lisacek, F., and Hopfgartner, G. (2015) Ranking fragment ions based on outlier detection for improved label-free quantification in data-independent acquisition LC-MS/MS. *J. Proteome Res.* **14,** 4581–4593
14. Toghi Eshghi, S., Auger, P., and Mathews, W. R. (2018) Quality assessment and interference detection in targeted mass spectrometry data using machine learning. *Clin. Proteomics* **15,** 33
15. Searle, B. C., Pino, L. K., Egertson, J. D., Ting, Y. S., Lawrence, R. T., MacLean, B. X., Villén, J., and MacCoss, M. J. (2018) Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9,** 5128
16. Jacome, A. S. V., Peckner, R., Shulman, N., Krug, K., DeRuff, K. C., Officer, A., MacLean, B., MacCoss, M. J., Carr, S. A., and Jaffe, J. D. (2019) Avant-garde: An automated data-driven DIA data curation tool. *bioRxiv* 565523
17. Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., and Aebersold, R. (2018) Data-independent acquisition-based SWATH-MS for quantitative proteomics: A tutorial. *Mol. Syst. Biol.* **14,** e8126
18. MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., and MacCoss, M. J. (2010) Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26,** 966–968
19. Reiter, L., Rinner, O., Picotti, P., Hüttenhain, R., Beck, M., Hengartner, M. O., and Aebersold, R. (2011) mProphet : Automated data processing and statistical validation for large scale SRM experiments. *Nat. Methods* **8,** 430–435
20. Schilling, B., Rardin, M. J., MacLean, B. X., Zawadzka, A. M., Frewen, B. E., Cusack, M. P., Sorensen, D. J., Bereman, M. S., Jing, E., Wu, C. C., Verdin, E., Kahn, C. R., Maccoss, M. J., and Gibson, B. W. (2012) Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: Application to protein acetylation and phosphorylation. *Mol. Cell Proteomics* **11,** 202–214
21. Bruderer, R., Bernhardt, O. M., Gandhi, T., Xuan, Y., Sondermann, J., and Schmidt, M. (2017) Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell Proteomics* **16,** 2296–2309
22. Muntel, J., Gandhi, T., Verbeke, L., Bernhardt, O. M., Treiber, T., Bruderer, R., and Reiter, L. (2019) Surpassing 10,000 identified and quantified proteins in a single run by optimizing current LC-MS instrumentation and data analysis strategy. *Molecular Omics* **15,** 348–360
23. Muntel, J., Kirkpatrick, J., Bruderer, R., Huang, T., Vitek, O., Ori, A., and Reiter, L. (2019) Comparison of protein quantification in a complex background by DIA and TMT workflows with fixed instrument time. *J. Proteome Res.* **18,** 1340–1351
24. Kelstrup, C. D., Young, C., Lavallee, R., Nielsen, M. L., and Olsen, J. V. (2012) Optimized fast and sensitive acquisition methods for shotgun proteomics on a quadrupole Orbitrap mass spectrometer. *J. Proteome Res.* **11,** 3487–3497
25. Bruderer, R., Bernhardt, O. O. M., Gandhi, T., and Reiter, L. (2016) High precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. *Proteomics* **16,** 1–20
26. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
27. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805
28. Suomi, T., and Elo, L. L. (2017) Enhanced differential expression statistics for data-independent acquisition proteomics. *Sci. Rep.* 1–8
29. Navarro, P., Kuharev, J., Gillet, L. C., Bernhardt, O. M., MacLean, B., Rost, H. L., Tate, S. A., Tsou, C.-C., Reiter, L., Distler, U., Rosenberger, G., Perez-Riverol, Y., Nesvizhskii, A. I., Aebersold, R., and Tenzer, S. (2016) A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34,** 1130–1136
30. She, X., Rohl, C. A., Castle, J. C., Kulkarni, A. V., Johnson, J. M., and Chen, R. (2009) Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics* **10,** 269
31. Hochberg, Y., and Benjamini, Y. (1990) More powerful procedures for multiple significance testing. *Stat. Med.* **9,** 811–818
32. Tenzer, S., Leidinger, P., Backes, C., Huwer, H., Hildebrandt, A., Lenhof, H.-P., Wesse, T., Franke, A., Meese, E., and Keller, A. (2016) Integrated quantitative proteomic and transcriptomic analysis of lung tumor and control tissue: A lung cancer showcase. *Oncotarget* **7,** 14857–14870
33. Krämer, A., Green, J., Pollard, J., Jr, and Tugendreich, S. (2014) Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30,** 523–530
34. Moreno, S. O., Cominetti, O., Galindo, A. N., Irincheeva, I., Corthésy, J., Astrup, A., Saris, W. H. M., Hager, J., Kussmann, M., and Dayon, L. (2017) The differential plasma proteome of obese and overweight individuals undergoing a nutritional weight loss and maintenance intervention. *Proteomics Clin. Appl.* **12,** 160015
35. Geyer, P. E., Wewer Albrechtsen, N. J., Tyanova, S., Grassl, N., Iepsen, E. W., Lundgren, J., Madsbad, S., Holst, J. J., Torekov, S. S., and Mann, M. (2016) Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Mol. Syst. Biol.* **12,** 901
36. Bruderer, R., Muntel, J., Müller, S., Bernhardt, O. M., Gandhi, T., Cominetti, O., Macron, C., Carayol, J., Rinner, O., Astrup, A., Saris, W. H. M., Hager, J., Valsesia, A., Dayon, L., and Reiter, L. (2019) Analysis of 1508 plasma samples by capillary-flow data-independent acquisition profiles proteomics of weight loss and maintenance. *Mol. Cell Proteomics* **18,** 1242–1254
37. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J., and Mann, M. (2018) BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* **15,** 440–448
38. Vizcaino, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q. W., Wang, R., and Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** 447–456