A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics

Bart van Puyvelde, Simon Daled, Sander Willems, Ralf Gabriels, Anne Gonzalez de Peredo, Karima Chaoui, Emmanuelle Mouton-Barbosa, David Bouyssié, Kurt Boonen, Christopher J Hughes, et al.

scientific data

OPEN

DATA DESCRIPTOR

A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics

Bart Van Puyvelde[1], Simon Daled[1], Sander Willems[2], Ralf Gabriels[3,4], Anne Gonzalez de Peredo[5], Karima Chaoui[5], Emmanuelle Mouton-Barbosa[5], David Bouyssié[5], Kurt Boonen[6,7], Christopher J. Hughes[8], Lee A. Gethings[8], Yasset Perez-Riverol[9], Nic Bloomfield[10], Stephen Tate[10], Odile Schiltz[5], Lennart Martens[3,4], Dieter Deforce[1] & Maarten Dhaenens[1] ✉

In the last decade, a revolution in liquid chromatography-mass spectrometry (LC-MS) based proteomics was unfolded with the introduction of dozens of novel instruments that incorporate additional data dimensions through innovative acquisition methodologies, in turn inspiring specialized data analysis pipelines. Simultaneously, a growing number of proteomics datasets have been made publicly available through data repositories such as ProteomeXchange, Zenodo and Skyline Panorama. However, developing algorithms to mine this data and assessing the performance on different platforms is currently hampered by the lack of a single benchmark experimental design. Therefore, we acquired a hybrid proteome mixture on different instrument platforms and in all currently available families of data acquisition. Here, we present a comprehensive Data-Dependent and Data-Independent Acquisition (DDA/DIA) dataset acquired using several of the most commonly used current day instrumental platforms. The dataset consists of over 700 LC-MS runs, including adequate replicates allowing robust statistics and covering over nearly 10 different data formats, including scanning quadrupole and ion mobility enabled acquisitions. Datasets are available via ProteomeXchange (PXD028735).

Background & Summary

Hypothesis-driven biochemical assays have been the foundation of molecular biology for well over a century, with great success. However, the lack of a more holistic view on the biomolecular complexity requires trial-and-error experimentation. Therefore, the past few decades were characterized by a shift towards an experimental design wherein a broader biomolecular perspective of the system is first generated in order to contextualize the hypothesis and the targeted biochemical assays beforehand. These "omics" approaches were enabled by two technical revolutions, i.e. the sequencing of nucleotides and the accurate mass measurement of biomolecules by mass spectrometry (MS).

In its barest form, the output from an MS instrument is merely a list of m/z's with intensities measured at very precise moments in time. However, MS is quickly evolving towards capturing the full complexity of a biological sample. To this end, not only the accuracy of instruments has improved greatly, they now also incorporate analytical techniques that select or separate analytes based on other physico-chemical properties. In proteomics nowadays, a mass spectrometer thus rarely only measures the m/z coordinate and intensity of (fragment) ions. The ion coordinates are mostly supplemented with precursor m/z, retention time (t_R) and/or ion mobility coordinates, depending on acquisition strategy. This creates a multidimensional data matrix that captures the complexity of the sample to an unprecedented depth[1].

[1]ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium. [2]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [3]VIB-UGent Center for Medical Biotechnology, VIB, 9000, Ghent, Belgium. [4]Department of Biomolecular Medicine, Ghent University, 9000, Ghent, Belgium. [5]Institut de Pharmacologie et de Biologie Structural (IPBS), Université de Toulouse, CNRS, UPS, Toulouse, France. [6]VITO Health, Mol, Belgium. [7]Centre for Proteomics, University of Antwerpen, Antwerp, Belgium. [8]Waters Corporation, Wilmslow, UK. [9]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. [10]SCIEX, Concord, Ontario, Canada. ✉e-mail: maarten.dhaenens@ugent.be

The field of mass spectrometry has diversified greatly, driven by a fast sequence of innovations from many vendors. Modern MS instruments allow the manipulation of ions in countless of different ways, including different ionization methodologies, fragmentation techniques, multipoles, time-of-flight tubes, ion mobility separation devices and trap designs, including the now very dominant orbitrap. The way in which these different ion manipulations are combined has ballooned the number of different acquisition strategies available to the end user today but all these instrumental and strategic innovations are futile if no data analysis pipeline is available to translate the data back into biology. For bottom-up proteomics this implies reconstructing the peptide backbone sequences from their fragment ions because the latter encompasses the specificity for identifying the hundreds of millions of different protein sequences that make up the biotic world.

Conventionally, MS instruments have been operated using data dependent acquisition (DDA) wherein the data from a precursor scan at low energy is used to pinpoint potentially interesting analytes which are then sequentially selected for fragmentation at high energy. These fragmentation spectra can then be identified by a plethora of different algorithms[2,3]. Data-independent acquisition (DIA) however, is the more intuitive way of analyzing a sample, because it captures all (fragment) ions without any instrumental bias. However, interpreting such complex data matrices has proven difficult and an additional separation dimension, such as ion mobility, was added to increase the discriminating power[4–8]. Alternatively, configuring a quadrupole to sequentially scan the entire mass range - while still operating "data independent" - alleviates the complexity of the resulting fragmentation spectra even more[9,10]. This has opened up the way for the many different spectrum-centric and peptide-centric data analysis strategies available today[11–16]. The latest reduction in complexity or "chimericy" of DIA spectra encompasses continuously scanning the quadrupole as is done with SONAR[17] and Scanning SWATH[18] and combining quadrupole selection and ion mobility separation, as is done with diaPASEF[19]. Unsurprisingly, machine learning is taking center stage in mining the various resulting data architectures[20–27].

Here, we created a comprehensive dataset on a single benchmark experimental design adapted from Navarro *et al.*[28]. It contains a ground truth that serves as a quality control for bioinformatics algorithm development and evaluation. This sample was acquired with adequate replicates on many of the current day instrumental platforms – partially in nano flow LC and partially in capillary flow LC - by most of the available acquisition strategy families, covering all commonly measured ion coordinates (Fig. 1). Far from being complete, it still is the most comprehensive repository of its kind for algorithmic development and validation, both at the level of identification and quantification. Instead of being yet another way of attaining the highest number of identified peptides, we hope it becomes a resource for compatibility assessment and data analysis quality control. Above all, it is a snapshot of current day completeness of our digital image of the protein world.

Methods

Sample preparation. Mass spectrometry-compatible *Human* K562 (P/N: V6951) and *Yeast* (P/N: V7461) protein digest extracts were purchased from Promega (Madison, Wisconsin, United States). Lyophilised MassPrep *Escherichia.coli* digest standard (P/N:186003196) was purchased from Waters Corporation (Milford, Massachusetts, United States). The extracts were reduced with dithiothreitol (DTT), alkylated with iodoacetamide (IAA) and digested with sequencing grade Trypsin(-Lys C) by the respective manufacturers. The digested protein extracts were reconstituted in a mixture of 0.1% Formic acid (FA) in water (Biosolve B.V, Valkenswaard, The Netherlands) and spiked with iRT peptides (Biognosys, Schlieren, Switzerland) at a ratio of 1:20 v/v. Two master samples A and B were created similar to Navarro *et al.*, each in triplicate, as shown in Fig. 1. Sample A was prepared by mixing *Human*, *Yeast* and *E.coli* at 65%, 30% and 5% weight for weight (w/w), respectively. Sample B was prepared by mixing *Human*, *Yeast* and *E.coli* protein digests at 65%, 15%, 20% w/w, respectively. The resulting samples have logarithmic fold changes (log2FCs) of 0, −1 and 2 for respectively *Human*, *Yeast* and *E.coli*. One sixth of each of the triplicate master batches of A and B were mixed to create a QC sample, containing 65% w/w *Human*, 22.5% w/w *Yeast* and 12.5% w/w *E.coli*.

LC-MS/MS. In this section, a detailed description of the different LC-MS/MS parameters is given for each LC-MS/MS instrumental setup applied to generate this comprehensive dataset. All instruments were operated according to the lab's best practice, i.e. not necessarily the best attainable, but rather most realistic data quality. Sample load was chosen based on LC setup (nano flow = 1 µg on column vs capillary flow = 5 µg on column) and instrument sensitivity. Thus, differences in absolute number of identified peptides and proteins can be attributed to sample load, LC flow rate, MS instrumentation, operator's choices and search algorithmic compatibility; direct conclusions on MS instrument performance can therefore not be drawn from this dataset.

As a rule of thumb, data-dependent acquisition (DDA) methods use high energy fragmentation spectra (MS2) of narrow mass selections for identification and use the area under the curve of the precursor (MS1) for quantification. Therefore, a cycle time needs to be attained wherein enough datapoints across the precursor elution peak are sampled for accurate quantification. In most data-independent acquisition (DIA) strategies, a broader precursor selection window is used and both identification and quantification can be done at the fragment level, taken that the cycle time for both MS1 and MS2 is adapted to the LC gradient. Finally, Precursor Acquisition Independent From Ion Count (PAcIFIC) is a method that is acquired solely to extend the size of the peptide library for detecting peptides in DIA data and is therefore not strictly dependent on the cycle time. Of note, by scanning the quadrupole instead of acquiring different mass windows separately, acquisition strategies like SONAR and Scanning SWATH create an additional dimension in the data matrix, akin to how ion mobility separation is perceived. Since these are similar to PAcIFIC acquisition, we also acquired gas phase (GP) fractions in SONAR and Scanning SWATH for library building, i.e. with no emphasis on cycle time.

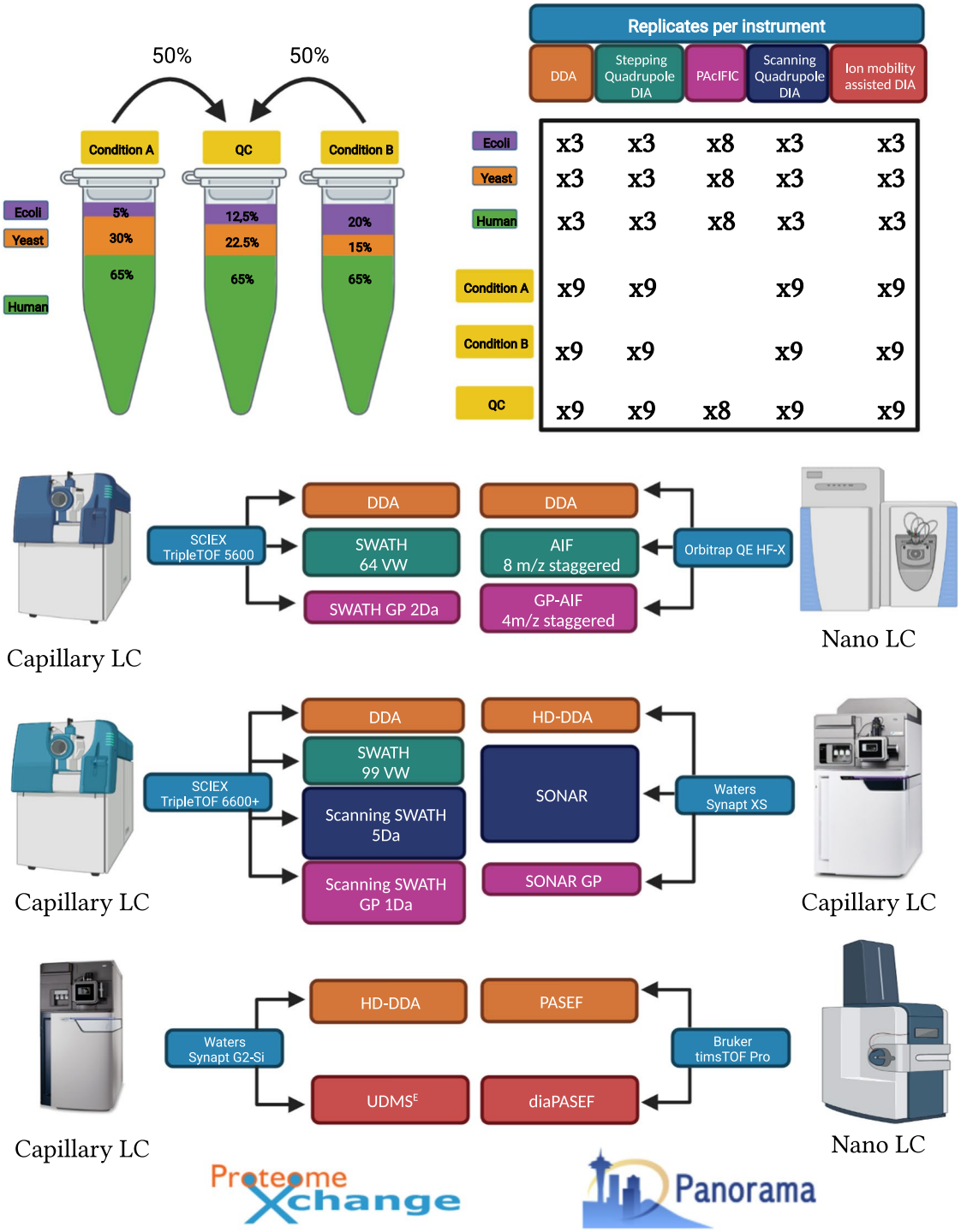

Fig. 1 Schematic overview of the different acquisition strategies/instruments applied in the study. A comprehensive LC-MS/MS dataset was generated using samples composed of commercial *Human* K562, *Yeast* and *Escherichia coli (E.coli)* full proteome digests. Two hybrid proteome samples A and B containing known quantities of *Human*, *Yeast* and *E.coli* tryptic peptides, as described by Navarro *et al*. were prepared in three consecutive times to include handling variability. Additionally, a QC sample was created by mixing one sixth of each of the six master batches (65% w/w *Human*, 22.5% w/w *Yeast* and 12.5% w/w *E.coli*). These commercial lysates were measured individually and as triple hybrid proteome mixtures each in triplicate using DDA and DIA acquisition methodologies available on six LC-MS/MS platforms, i.e. SCIEX TripleTOF 5600 and TripleTOF 6600+, Thermo Orbitrap QE HF-X, Waters Synapt G2-Si and Synapt XS and Bruker timsTOF Pro. The complete dataset was made publicly available to the proteomics community through ProteomeXchange with dataset identifier: PXD028735. In addition, a system suitability workflow (AutoQC) was incorporated on each instrument using commercial *E.coli* lysate digest which were acquired at multiple timepoints throughout each sample batch. The AutoQC data was automatically imported in Skyline and uploaded to the Panorama AutoQC server using AutoQC loader, enabling system suitability assessment of each LC-MS/MS system used in the dataset.

SCIEX TripleTOF 5600 (Capilary flow LC). A TripleTOF 5600 mass spectrometer (Sciex, Concord, Ontario, Canada) fitted with a Duospray ion source operating in positive ion mode, was coupled to an Eksigent NanoLC 400 HPLC system (Eksigent, Dublin, CA). 5 µL of each sample was loaded at 5 µL/min with 0.1% Trifluoroacetic acid (TFA) in water and trapped on a YMC TriArt C18 guard column (id 500 µm, length 5 mm, particle size 3 µm) for 5 minutes. Peptides were separated on a microLC YMC TriArt C18 column (id 300 µm, length 15 cm, particle size 3 µm) maintained at 55 °C at a flow rate of 5 µL/min by means of trap-elute injection. Mobile phase A consisted of UPLC-grade water with 0.1% (v/v) FA and 3% (v/v) DMSO, and mobile phase B consisted of UPLC-grade ACN with 0.1% (v/v) FA. Peptide elution was performed at 5 µL/min using the following gradient: i) 2% to 30% mobile phase B in 120 min, ii) ramp to 90% mobile phase B in 1 min. The washing step at 90% mobile phase B lasted 4 min and was followed by an equilibration step at 2% mobile phase B (starting conditions) for 10 min. Ion source parameters were set to 5.5 kV for the ion spray voltage, 30 psi for the curtain gas, 13 psi for the nebulizer gas and 80 °C as source temperature.

Data-dependent acquisition. For DDA (a cycle time of 3.5 s), MS1 spectra were collected between 399–1200 m/z for 500 ms. The 20 most intense precursors ions with charge states 2–5 that exceeded 250 counts per second were selected for fragmentation, and the corresponding fragmentation MS2 spectra were collected between 50–2000 m/z for 151 ms. After the fragmentation event, the precursor ions were dynamically excluded from reselection for 20 s.

SWATH 64 variable windows. For SWATH (a cycle time of 3.4 s), a 64 variable window acquisition scheme as described by Navarro *et al.* was used for all samples (Supplementary Table 1)[28]. Briefly, SWATH MS2 spectra were collected in high-sensitivity mode from 50–2000 m/z, for 50 ms. Before each SWATH MS cycle an additional MS1 survey scan in high sensitivity mode from 400–1200 m/z was recorded for 150 ms.

PAcIFIC. For PAcIFIC (a cycle time of 4 s), the TripleTOF5600 was configured to acquire eight gas phase fractionated acquisitions with isolation windows of 4 m/z using an overlapping window pattern from narrow mass ranges, as described by Searle *et al.* (i.e., 396.43–502.48; 496.48–602.52;596.52–702.57;696.57–802.61; 796.61–902.66; 896.6–1002.70; 996.70–1102.75; 1096.75–1202.80)[29]. See Supplementary Table 2 for the actual windowing scheme. MS2 spectra were collected in high-sensitivity mode from 360–1460 m/z, for 75 ms. An MS1 survey scan was recorded per cycle from 360–1460 m/z for 50 ms.

SCIEX TripleTOF 6600+ (Capilary flow LC). A TripleTOF 6600+ mass spectrometer (Sciex, Concord, Ontario, Canada) fitted with an Optiflow ion source operating in positive ion mode, was coupled to an Eksigent NanoLC 425 HPLC system (Eksigent, Dublin, CA). 5 µL of each sample was loaded at 5 µL/min with 0.1% FA in water by means of direct injection. Peptides were separated on a Phenomenex Luna Omega Polar C18 column (150 × 0.3 mm, particle size 3 µm) at a column temperature of 30 °C. Mobile phase A consisted of UPLC-grade water with 0.1% (v/v) FA, and mobile phase B consisted of UPLC-grade ACN with 0.1% (v/v) FA. Peptide elution was performed at 5 µL/min using the following gradient: i) 2% to 30% mobile phase B in 120 min, ii) ramp to 90% mobile phase B in 1 min. The washing step at 90% mobile phase B lasted 4 min and was followed by an equilibration step at 2% mobile phase B (starting conditions) for 10 min. Ion source parameters were set to 4.5 kV for the ion spray voltage, 25 psi for the curtain gas, 10 psi for nebulizer gas (ion source gas 1), 20 psi for heater gas (ion source gas 2) and 100 °C as source temperature.

Data-dependent acquisition. For DDA acquisition (a cycle time of 3.3 s), MS1 spectra were collected between 400–1200 m/z for 250 ms. The 30 most intense precursor ions with charge states 2–4 that exceeded 300 counts per second were selected for fragmentation, and the corresponding fragmentation MS2 spectra were collected between 100–1500 m/z for 100 ms. After the fragmentation event, the precursor ions were dynamically excluded from reselection for 10 s.

SWATH 99 Variable windows. For SWATH (a cycle time of 4 s), a 99 variable window acquisition scheme was used (see Supplementary Table 3). Briefly, SWATH MS2 spectra were collected in high sensitivity mode from 100–1500 m/z, for 37.5 ms. Before each SWATH MS cycle an additional MS1 survey scan in high sensitivity mode was recorded for 250 ms.

Scanning SWATH GP 1Da. A Scanning SWATH beta version was installed on the Analyst TF control software in collaboration with Sciex Research. Scanning SWATH Q1 calibration was confirmed by directly infusing a tuning solution (ESI Positive Calibration Solution for the SCIEX X500B System - P/N: 5049910) and by acquiring a pre-built calibration batch (SSCalibration.dab). Afterwards, the calibration was verified by i) running a verification calibration batch and inspect the data in PeakView. Calibration of the Q1 after initial calibration was controlled automatically by internal recalibration of each data file at run time removing the need for subsequent calibration events.

The gas-phase fractionation approach usually acquired in PAcIFIC, were also acquired by Scanning SWATH because this uniquely allows to apply DIA annotation algorithms for library building of subsequent full mass range DIA acquisition. Precursor isolation window was set to 1 m/z and a mass range of 100 m/z was covered in 6 s (average accumulation time per precursor: 59.57 ms). An MS1 scan was included and data was acquired in high resolution mode. Raw data was converted to standard SCIEX data files with an effective precursor isolation of 0.2 m/z bins and Q1 calibration was obtained by running rawSSProcessor.exe.

Scanning SWATH Fixed 5 Da WindowThe precursor mass range of 400–900 m/z was covered in 4 s with a 5 m/z isolation window (average accumulation time per precursor: 37.5 ms). A TOF MS scan was included and data was acquired in high sensitivity mode. Raw data was converted to standard SCIEX data files with an effective precursor isolation of 1 m/z bins and Q1 calibration was obtained by running the rawSSProcessor.exe.

Thermo Orbitrap QE HF-X (Nano flow LC). A Thermo Orbitrap QE HF-X (Thermo Fisher Scientific, Waltham, Massachusetts, United States) was coupled to an UltiMate 3000 LC-system (NCS-3500RS Nano/Cap System, Thermo Fisher Scientific). Peptides were separated on an Acclaim PepMap C18 column (id 75 µm, length 50 cm, particle size 2 µm, Thermo Fisher Scientific ref 164942) at a flow rate of 350 nL/min by means of trap-elute injection (Acclaim PepMap C18, id. 300 µm × 5 mm) after 3 min desalting on a nano-trap cartridge (id.300 µm, length 5 mm, Thermo Fisher Scientific ref 160454).

Mobile phase A consisted of UPLC-grade water with 0.1% (v/v) FA, and mobile phase B consisted of UPLC-grade ACN with 0.1% (v/v) FA. Peptide elution was performed at 350 nL/min using the following gradient: i) 2% to 30% mobile phase B in 120 min, ii) ramp to 90% mobile phase B in 1 min. The washing step at 90% mobile phase B lasted 4 min and was followed by an equilibration step at 2% mobile phase B (starting conditions) for 21 min.

Data-dependent acquisition. The data-dependent acquisition runs on the Q Exactive HF-X were acquired with MS survey scans (350–1400 m/z) at a resolution of 60,000, and an AGC target of 3e6. The 12 most intense precursor ions, were selected for fragmentation by high-energy collision–induced dissociation, and the resulting fragments were analyzed at a resolution of 15,000 using an AGC target of 1e5 and a maximum fill time of 22 ms. Dynamic exclusion was used within 30 s to prevent repetitive selection of the same peptide.

Narrow window gas-phase fractionation (GP) DIA. Narrow-window GP-DIA data was acquired as described by Searle *et al.*[29]. Briefly, 6 GP runs (400–500, …, 900–1000 m/z) using staggered 4 m/z DIA spectra (4 m/z precursor isolation windows at 30,000 resolution, AGC target 1e6, maximum inject time 60 ms, NCE 27, +3H assumed charge state) were acquired using an overlapping window pattern, described by Pino *et al.*[30]. In each run, full MS scans matching each part of the fractionated mass range (i.e., either 395–505, 495–605, 595–705, 695–805, 795–905, or 895–1005 m/z), acquired at a resolution of 60,000 using an AGC target of 1e6 and a maximum inject time of 60 ms, were interspersed every 25 MS/MS spectra.

All ion fragmentation (AIF). The AIF DIA data was acquired using a staggered pattern of 75 × 8 m/z isolation windows over the mass range 400–1000 m/z as described by Pino *et al.* (Supplementary Table 4)[30]. DIA MS/MS scans were acquired at 15,000 resolution, with an AGC target of 1e6, a maximum inject time 20 ms, and a NCE of 27. Full MS scans over the range 390–1010 m/z at 60,000 resolution, AGC target 1e6, maximum inject time 60 ms were interspersed every 75 MS/MS spectra.

Waters Synapt G2-Si (Capillary flow LC). An M-class LC system (Waters Corporation, Milford, MA) was equipped with a 1.7 µm CSH 130 C18 300 µm × 100 mm column, operating at 5 µL/min with a column temperature of 55 °C. Mobile phase A was UPLC-grade water containing 0.1% (v/v) FA and 3% DMSO, mobile phase B was ACN containing 0.1% (v/v) FA. Peptides were separated using a linear gradient of 3−30% mobile phase B over 120 minutes. All experiments were conducted on a Synapt G2-Si mass spectrometer (Waters Corporation, Wilmslow, UK). The ESI Low Flow probe capillary voltage was 3 kV, sampling cone 60 V, source offset 60 V, source temperature 80 °C, desolvation temperature 350 °C, cone gas 80 L/hr, desolvation gas 350 L/hr, and nebulizer pressure 2.5 bar. A lock mass reference signal of GluFibrinopeptide B (m/z 785.8426) was sampled every 30 s.

HD-DDA. Data was acquired according to Helm *et al.* with minor adaptations[7]. Briefly, in data-dependent mode, the MS automatically switches between MS survey and MS/MS scans based upon a set of switching criteria, including ion intensity and charge state. Full scan MS and MS/MS spectra (m/z 50–5000) were acquired in sensitivity mode. MS survey spectra were acquired using a fixed acquisition time of 250 ms and the ions present in each scan were monitored for the following criteria: more than 3000 intensity/sec and only 2,3,4,5+ charge states. Once criteria were satisfied, the precursor ion isolation width of the quadrupole was set to 1.0 Th around each precursor sequentially. Tandem mass spectra of up to 12 precursors were generated in the trapping region of the ion mobility cell by using a collisional energy ramp from 6/9 V (low mass 50 Da, start/end) to up to 147/183 V (high mass 5000 Da, start/end), with actual values applied dependent upon the precursor m/z. The MS2 scan time was set to 100 ms and the "TIC stop" parameter was set to 100,000 intensity/s allowing a maximum accumulation time of 300 ms (i.e. up to three tandem MS spectra of the same precursor). IMS wave velocity was ramped from 2400 m/s to 450 m/s (start to end) and the pusher/ion mobility synchronized for singly charged fragment ions in MS/MS spectra, with up to 85% duty cycle efficiency.

UDMS^{E}. Data was acquired according to Distler *et al.* with minor adaptations[31]. Briefly, two data functions were acquired over a mass range of m/z 50 to 2000 in alternating mode, differing only in the collision energy applied to the gas cell. In low-energy MS1 mode, data was collected at a constant gas cell collision energy of 4 eV. In elevated energy MS2 mode, the gas cell collision energy was ramped from 10 to 60 eV according to a collision energy look up table in function of drift time. The spectral acquisition time in each mode was 0.6 s with a 0.015 s interscan delay.

Waters Synapt XS (Capilary flow LC). An M-class LC system (Waters Corporation, Milford, MA) equipped with a 1.7 µm CSH 130 C18 300 µm × 100 mm column, operating at 7 µL/min with a column temperature of 55 °C was coupled to a Synapt XS quadrupole oa-ToF mass spectrometer (Waters Corporation, Wilmslow, UK) operating at a mass resolution of 30000, FWHM. The ESI Low Flow probe capillary voltage was 1.8 kV, sampling cone 30 V, source offset 4 V, source temperature 100 °C, desolvation temperature 300 °C, cone gas disabled, desolvation gas 600 L/hr, and nebulizer pressure 3.5 bar was used. The time-of-flight (TOF) mass analyzer of the mass spectrometer was externally calibrated with a NaCsI mixture from m/z 50 to 1990. A lock mass reference signal of GluFibrinopeptide B (m/z 785.8426) was sampled every two minutes. Mobile phase A was water containing 0.1% (v/v) FA, while mobile phase B was ACN containing 0.1% (v/v) FA. The peptides were eluted and separated with a gradient of 5−40% mobile phase B over 120 minutes.

HD-DDA. In data-dependent mode, the MS instrument automatically switches between MS survey and MS/MS scans based upon a set of switching criteria, including ion intensity and charge state. Full scan MS and MS/MS spectra (m/z 50–5000) were acquired in resolution mode. MS survey spectra were acquired using a fixed acquisition time of 200 ms and the ions present in each scan were monitored for criteria intensity more than 5000 intensity/sec and 2,3,4+ charge states. Once criteria were satisfied, the precursor ion isolation width of the quadrupole was set to 1.0 Th around each precursor sequentially. Tandem mass spectra of up to 15 precursors were generated in the trapping region of the ion mobility cell by using a collisional energy ramp from 6/9 V (low mass 50 Da, start/end) to up to 147/183 V (high mass 5000 Da, start/end), with actual values applied dependent upon the precursor m/z. The MS2 scan time was set to 70 ms and the "TIC stop" parameter was set to 100,000 intensity/s allowing a maximum accumulation time of 100 ms (i.e. up to two tandem MS spectra of the same precursor). IMS wave velocity was ramped from 2450 m/s to 550 m/s (start to end) and the pusher/ion mobility synchronized for singly charged fragment ions in MS/MS spectra, with up to 85% duty cycle efficiency.

SONAR GP. As for Scanning SWATH, the GP fractionation approach which is usually acquired in PAcIFIC was also analysed by SONAR purely to extend the peptide library. Therefore, the mass scale from m/z 400 to 1200 was divided into 100 Da sections, thus requiring 8 injections for each sample. The quadrupole was continuously scanned from the start mass to end mass of each section and a transmission window of 4 Da was used. In low-energy MS1 mode, data were collected at constant gas cell collision energy of 6 eV. In elevated energy MS2 mode, the gas cell collision energy was ramped with values calculated from the start and end mass of the 100 Da mass range being scanned by the quadrupole, and are shown in Supplementary Table 5. The spectral acquisition time in each mode was 0.5 s with a 0.02 s interscan delay.

SONAR. The quadrupole was continuously scanned between m/z 400 to 900, with a quadrupole transmission width of ~20 Da. Two data functions are acquired in an alternating mode, differing only in the collision energy applied to the gas cell. In low-energy MS1 mode, data were collected at constant gas cell collision energy of 6 eV. In elevated energy MS2 mode, the gas cell collision energy was ramped from 16 to 36 eV (per unit charge). The spectral acquisition time in each mode was 0.5 s with a 0.02 s interscan delay.

Bruker timsTOF pro (Nano flow LC). An Acquity UPLC M-Class System (Waters Corporation) was fitted with a nanoEase™ M/Z Symmetry C18 trap column (100 Å, 5 µm, 180 µm × 20 mm) and a nanoEase™ M/Z HSS C18 T3 Column (100 Å, 1.8 µm, 75 µm × 250 mm, both from Waters Corporation). The sample was loaded onto the trap column in 2 min at 5 µl/min in 94% mobile phase A and 6% mobile phase B. Mobile phase A is UPLC-grade water with 0.1% FA, while mobile phase B is 80% ACN with 0.1% FA. The Acquity UPLC M-Class system was coupled online to a TimsTOF Pro via a CaptiveSpray nano-electrospray ion source (Bruker Daltonics, Bremen, Germany), with an ion transfer capillary temperature at 180 °C. Liquid chromatography was performed at 40 °C and with a constant flow of 400 nL/min. Peptides were separated using a linear gradient of 2−30% mobile phase B over 120 minutes. The timsTOF Pro elution voltages were calibrated linearly to obtain reduced ion mobility coefficients (1/K0) using three selected ions of the Agilent ESI-L Tuning Mix (m/z, $1/K_0$: 622.0289 Th, 0.9848 Vs cm^{-2}; 922.0097 Th, 1.1895 Vs cm^{-2}; 1222.9906 Th, 1.3820 Vs cm^{-2}).

PASEF. Parallel Accumulation–Serial Fragmentation DDA (PASEF) was used to select precursor ions for fragmentation with 1 TIMS-MS scan and 10 PASEF MS/MS scans, as described by Meier *et al.* in 2018[32]. The TIMS-MS survey scan was acquired between 0.70–1.45 V.s/cm^2 and 100–1700 *m/z* with a ramp time of 100 ms. The 10 PASEF scans contained on average 12 MS/MS scans per PASEF scan with a collision energy of 10 eV. Precursors with 1–5 charges were selected with the target value set to 20 000 a.u and intensity threshold to 2500 a.u. Precursors were dynamically excluded for 0.4 min.

diaPASEF. The diaPASEF method was implemented as described by Meier *et al.* in 2019[19]. The DIA parameters that define the windows can be found in Supplementary Table 6. The DIA range was set to 400–1200 m/z with 16 diaPASEF scans of 25 m/z isolation windows, including an overlap of 1 Da. Each diaPASEF scan consisted of two steps (measuring two 25 Da intervals), with each step spanning an IMS range of 0.3 V.s/cm^2. The lower IMS value increased linear from 0.6 to 0.834375 for the diaPASEF scans. The TIMS-MS scan was identical to the PASEF method.

Data Records

Data record 1. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[33] with the dataset identifier, PXD028735[34]. For every instrument, a separate Sample and Data Relationship File (SDRF) and an Investigation

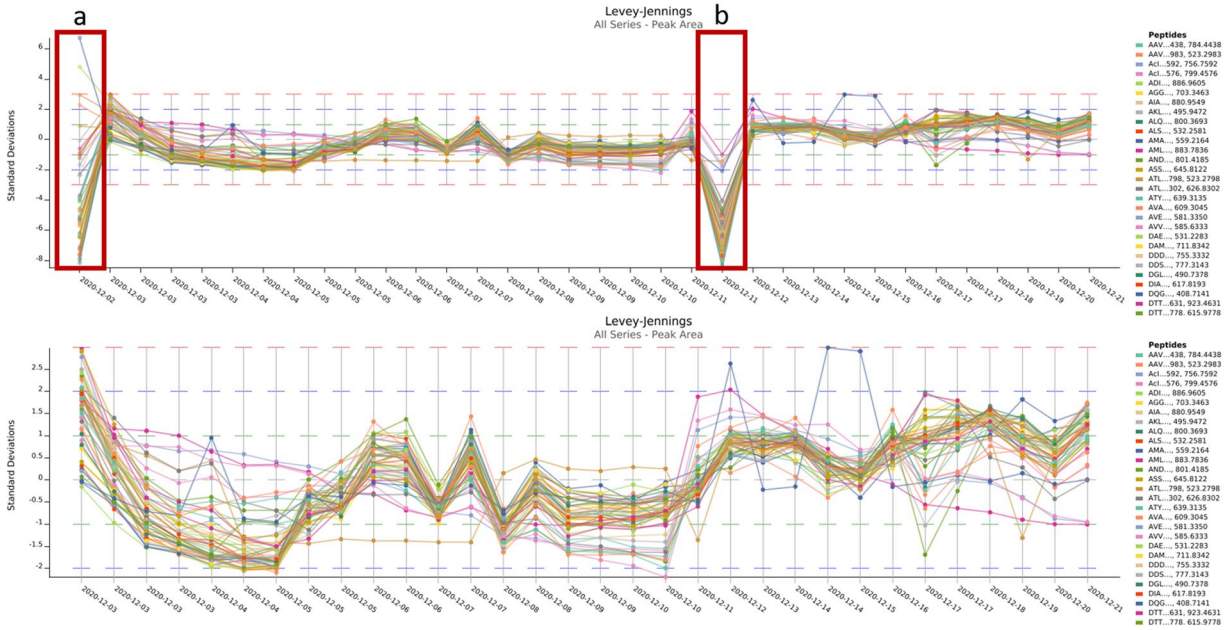

Fig. 2 Levey-Jennings plot of the standard deviation in peak area for 50 selected precursors acquired in DDA with the TripleTOF 6600+. The upper chart shows two distinct outliers, acquired respectively on the 2nd and 12th of December (red boxes). Manual inspection of the data shows these were caused by (**a**) a wrong vial in the sample tray and (**b**) an empty vial. When these two samples are excluded from the Levey-Jennings plot (lower chart), a significant drop in standard deviation over the time period of data acquisition is seen.

Description File (IDF) have been uploaded to ProteomeXchange. Both the SDRF and IDF file formats are relatively new in Proteomics and were developed in a collaboration between EuBIC-MS and the Proteomics Standards Initiative (PSI)[35,36]. These files are used to annotate the sample metadata and link the metadata to the corresponding data file(s) and thus will improve the reproducibility and reanalysis of this comprehensive benchmark dataset.

Data record 2. The AutoQC data analysed in Skyline is available from Panorama Public with the link https://panoramaweb.org/LFQBenchmark.url[37].

Technical Validation

We continuously performed system suitability procedures to monitor LC-MS/MS performance in a longitudinal fashion. Therefore, we ran an AutoQC complex lysate, i.e. a commercial E.coli protein digest extract (Supplementary Table 7), every 3 to 4 samples over all acquired runs on all instruments. All the AutoQC samples were acquired in DDA on each LC-MS/MS instrument, except for the Synapt XS, Orbitrap QE-HF and the timsTOF Pro, where incidentally DDA and DIA acquisitions were alternated. The same mobile phase A and B composition as for the benchmark samples was used as for the benchmark samples, but the gradient applied was modified to reduce the time required to acquire the complete sample batch: linear 3–40% B in 60 minutes, up to 85% B in 2 minutes, isocratic at 85% B for 7 minutes, down to 3% B in 1 minute and isocratic at 3% B for 10 minutes. Note that the timsTOF Pro AutoQC samples were acquired using the 120 min gradient similar to the actual hybrid proteome samples.

System suitability assessment was performed by monitoring peptide-identification free metrics (i.e. retention time, peak area, mass accuracy, etc.) extracted with the vendor neutral Panorama AutoQC framework[38,39]. To isolate a set of peptides that can be used for this, triplicate AutoQC samples acquired on each instrument were peak picked using MSConvert (version 3.0.20070) and the corresponding.MGF files were searched against an *E.coli* FASTA database using Mascot Daemon (v2.7). The searches were performed with following settings: (i) 20 ppm peptide mass tolerance, (ii) 50 ppm fragment mass tolerance and (iii) two allowed missed cleavages. The peptide and protein identification results were exported as Mascot .DAT file and imported into Skyline Daily (version 21.1.1.160). The five highest ranked proteins were retained in the target list and after importing one of the AutoQC .raw files, we manually verified and removed each precursor with co-eluting peptides and low MS1 signal intensity before a Skyline file was saved as template file. Finally, a configuration file for each setup was created with the AutoQC Loader software (version 21.1.0.158) which leads to the automatic import of every sample, with the pattern "AutoQC" in the .raw file or folder structure, in the Skyline template.The data and skyline reports were published to the PanoramaWeb folder "U of Ghent Pharma Biotech Lab - LFQBenchmark across Instrument Platforms" containing six subfolders for each instrumental platform.

For each instrument, peak area, retention time and mass accuracy were manually checked by plotting these metrics in Levey-Jennings plots. For almost every instrument a few outliers were detected, as can be expected on a dataset of over 600 LC-MS runs. Fortunately, most of these can be explained by inspecting the raw data and by personal communication with the technicians acquiring the respective datasets. Figure 2 illustrates one such

case. More specifically, two AutoQC samples display a near-complete loss in peak area in the TripleTOF 6600+ DDA dataset. Indeed, these were caused by (a) a wrong vial in the sample tray and (b) an empty vial. When these two samples are removed from the QC plot, a more coherent perspective on the variation in standard deviation is seen in the Levey-Jennings plot. Other instances that we have already found include (i) a significant shift in standard deviation in peak area reported for the Orbitrap QE HF-X, timsTOF Pro and Synapt XS dataset because AutoQC samples were incidentally acquired in two different acquisition methodologies, i.e. in DDA and DIA; (ii) For the timsTOF Pro, a drift in retention time was seen, indicating LC related technical variation which could have been caused by e.g. too short column equilibration times; (iii) In the Orbitrap QE HF-X AutoQC data one peptide (EEAIIK) was undetectable in all the AutoQC samples acquired in AIF. Manual inspection of the acquisition in Skyline (easily accessible through the Panorama QC pipeline) surfaced that it fell out of the precursor m/z range (351.7053 *m/z*) that was acquired.

As expected in such a massive MS proteomics experiments, it seems that for every instrument some outliers were recorded, most of which have explanations common to the field. Above all, this demonstrates the necessity of a performant system suitability workflow to increase the reproducibility and quality of LC-MS/MS proteomics datasets[40].

Usage Notes

A comprehensive dataset such as the one presented here is inspired by the bioinformatics need to cope with the recent expansion of novel acquisition strategies and data dimensions. Apart from being a repository for validating both performance and compatibility of (new) bioinformatic pipelines, it can serve as a reference for general proteomics courses (e.g. Skyline tutorials, SWATH/DIA course) and be applied for training and validating machine/deep learning algorithms. As such, it is intended to facilitate our understanding of the impact of instrumentation on the perspective that is generated on protein biology and to a larger extent to help unify the field of proteomics.

To demonstrate the applicability of this data repository, we assessed the impact of instrumentation on the most conventional data format acquired by all instruments i.e. DDA. Therefore, for every instrumental platform, triplicate *Human*, *Yeast* and *E.coli* DDA runs were peak picked with MSConvert (version 3.0.21285) and exported as a Mascot .MGF file (Fig. 3a). MS Convert software was chosen as it is vendor-independent and contains each vendor's implementation for peak picking, with the exception of UNIFI i.e. Waters. Therefore, Progenesis QI for Proteomics (version 4.2.7207) was used for the Waters DDA data. By doing so, the MS1 precursor space was aligned in the retention time dimension before peak picking. Subsequently, all MS/MS spectra were exported as .MGF file for peptide identification. A standard search with carbamidomethylation of cysteine as fixed modification was performed using a database containing the *Human*, *Yeast* and *E.coli* protein sequences (downloaded from Uniprot on 19/01/2021) and with following parameters: i) mass error tolerances for the precursor ions and the fragment ions were set at 20 ppm and 50 ppm, respectively; ii) enzyme specificity was set to trypsin, allowing up to one missed cleavage. Next, the results were exported as .DAT file and imported into Skyline to create a non-redundant spectral library with BiblioSpec. Afterwards, the .BLIB file was converted to a .dlib and .msp file format respectively using EncyclopeDIA[29]. The resulting .msp file was converted using a Python conversion tool (speclib_to_mgf.py) built-in MS²PIP, to create a peptide record file (.PEPREC) and .MGF file. Next the proportion (amount) of peptide identifications overlapping between the different instruments was assessed using a custom Python script (Fig. 3b).

Since each instrument analysed the same commercial protein digests, a large overlap in peptide identifications would be expected. However, while the timsTOF Pro, TripleTOF 6600+ and Orbitrap QE HF-X roughly identify a similar number of peptide sequences (approximately 40,000 over nine LC-MS runs), the overlap in identified sequences is in the order of 50%, with that between the Orbitrap QE HF-X and the TripleTOF 5600 and 6600+ being overall 10% higher than the overlap with the timsTOF Pro. While at the protein level a lot more coherency is found, it is still striking how the Orbitrap QE HF-X and especially the timsTOF Pro have significantly more proteins from a similar number of annotated peptides (Fig. 3c).

A benchmark like the one presented here however, allows to investigate potential assertions attributed to different stages of the workflow.

Acquisition. Even with DDA, the acquisition parameters can be modified substantially. A first, crucial difference is the used flow rate of the LC system: the TripleTOF 6600+ was coupled to a 5 µl/min flow, while the Orbitrap QE HF-X and timsTOF Pro used nanoflow (350 and 400 nL/min, respectively). While nano flow is more sensitive, it is also more fragile, which is reflected in some of the quality control samples uploaded to Panorama QC. A second important parameter is dynamic exclusion, i.e. how long a certain precursor is excluded from fragmentation to avoid redundancy. As found in the methods section, this was set to 10 s, 24 s and 30 s for the TripleTOF 6600+, the timsTOF Pro and the Orbitrap QE HF-X, respectively. This parameter usually is a function of LC gradient length and of consistent MSMS spectral quality, which can be facilitated through e.g. automatic gain control (AUC) in orbitrap instruments, but not in QTOF designs. As a result, the orbitrap sampled the peptides most efficiently, with 20% of the peptides only sampled once in a run, i.e. three times over triplicate runs of the same proteome (Fig. 3d). A third aspect of acquisition is the addition of ion mobility separation (IMS) capabilities in Waters and Bruker instruments. Importantly, they both use IMS in a very different way, forcing a deeper understanding of instrumental architecture and how ion mobility is applied by both vendors in DDA relative to the quadrupole (Q) selection and the collision-induced dissociation (CID). Briefly, the order of ion manipulation is Q-CID-IMS for the so-called High Definition DDA (HD-DDA) in the Waters series and it is IMS-Q-CID in the Bruker instruments. Therefore, Waters separates the fragment ions in IMS and leverages the efficient charge state separation to synchronize the pusher of the TOF tube with singly charged fragment ions in order to only and

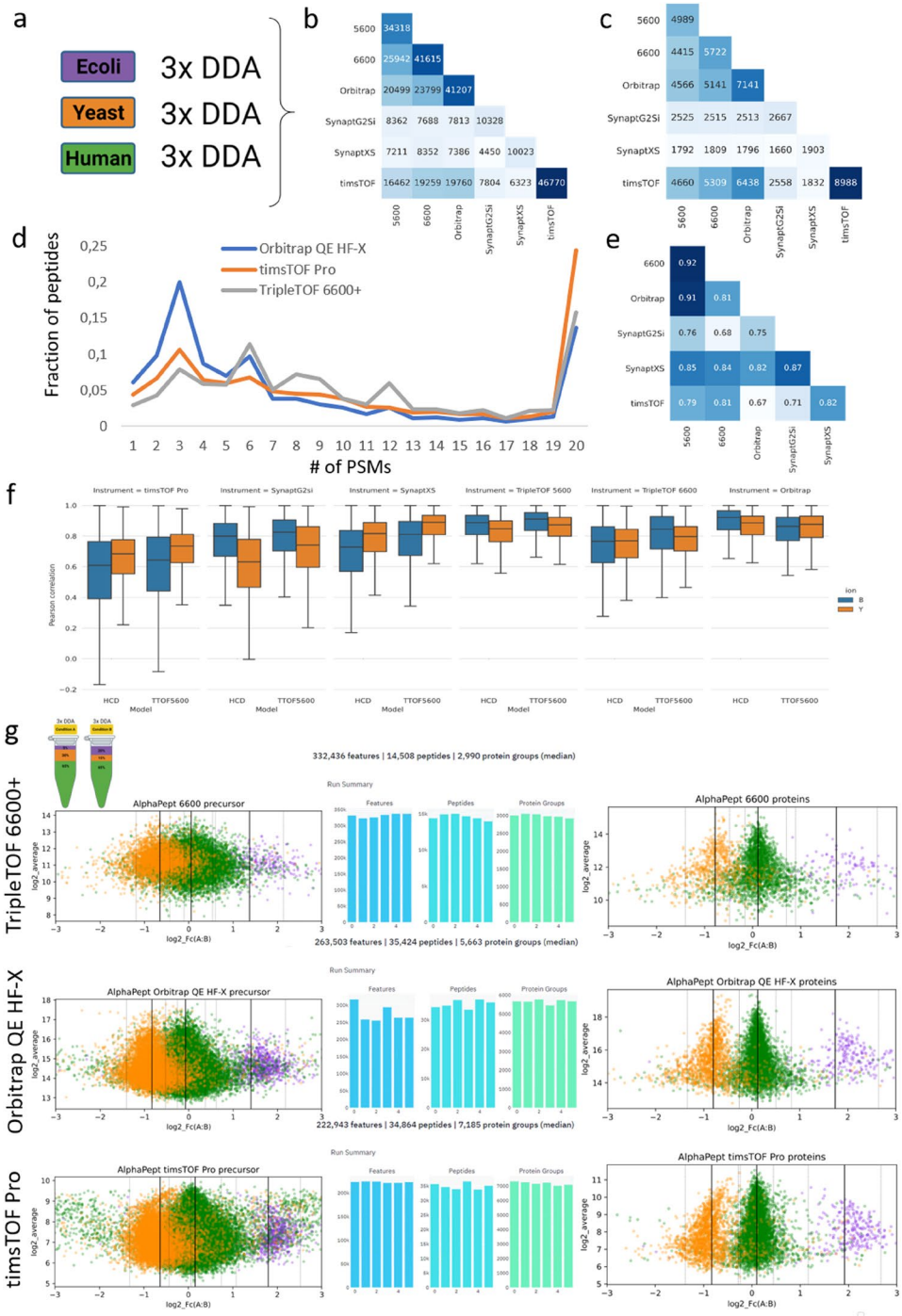

Fig. 3 Comparing the DDA data of six different instruments. Experimental design (**a**) Triplicate measurements of three individual proteomes. (**b**) The overlap in uniquely identified peptide sequences and (**c**) proteins between the six instruments. (**d**) Number of PSMs per peptide identification throughout nine DDA runs on three different proteomes for three instruments. (**e**) Pearson Correlation Coefficient (PCC) of the fragment intensities were calculated between the shared identified peptides from the DDA replicates between each instrument. The numbers in each box correspond to the median spectrum PCC between the instrument on the x-axis and the instrument on the y-axis. Dark blue color indicates a higher degree of overlap or higher median PCC. (**f**) Boxplots of the Pearson correlation coefficients (PCC) between the MS²PIP predicted (HCD and TTOF5600 model) and experimental fragment ion intensities across the six different LC-MS instruments. (**g**) The benchmark design of mixed proteomes for three instruments as annotated and quantified using AlphaPept. Here, triplicate runs of Condition A and Condition B were used, resulting in the six bars depicted in the middle, respectively representing the number of MS1 features, the number of identified peptides and the number of identified proteins for each instrument. The log-fold plots to the left depict the distribution of the peptide ratios in the x-axis as a function of their intensity in the y-axis; protein log fold changes are depicted to the right.

nearly one hundred percent efficiently sample the single charged fragments[7]. This can be expected to impact the fragment intensities, especially when detector saturation in MSMS occurs. In the timsTOF Pro on the other hand, the IMS is in fact resolving the precursor ion space before fragmentation, leading to a different selection of the peptide precursor space compared to other devices that do not use IMS (in this way).

Raw data. Next, the differences in the fragmentation process between the instruments (all of which are beam-type CID) can be assessed by mutually mapping the MS2 intensities. Figure 3e shows their median Pearson correlation coefficients (PCC). By definition, this only plots commonly found peptides. The largest differences in fragment intensities are found between timsTOF Pro and Orbitrap QE HF-X, potentially underlying part of the differences seen in peptide and protein identification. Importantly, the recent introduction of machine learning algorithms has illustrated the potential of fragment intensity prediction for proteomics[24,25,41–43]. Therefore, to attain an even deeper understanding of these fragmentation differences, we calculated the PCC of both b- and y-ions compared to two prediction models (i.e. HCD and TTOF 5600) from MS²PIP[44] (Fig. 3f). This confirms that the Orbitrap and TripleTOF designs have a very similar fragmentation pattern. Note that the normalized collision energy was not taken into account as input feature[25,45,46]. Additionally, the Synapt and timsTOF instruments are sampling the ion beam more efficiently, through ion mobility (Waters) or trapping and pusher design (Bruker), potentially causing more frequent detector saturation in the fragment space, leading to skewed intensity patterns compared to the predicted spectra. Indeed, augmented gain controlled in Orbitrap devices can efficiently avoid this.

Sample complexity. We also acquired the proteomes as mixtures in known ratios, and selected a triplicate series of Condition A and B for a subsequent analysis (Fig. 3g). This higher sample complexity is considerably more challenging for DDA acquisition, because the instrument needs to do a faster sampling of the precursor ion space. To create a more comprehensive picture of the resulting data, we turned to AlphaPept, an alternative annotation algorithm that also quantifies the number of features in MS1, i.e. potential peptide species in the sample[47]. Strikingly, while the TripleTOF 6600+ generated the most features (330k compared to 260k and 220k for the Orbitrap QE HF-X and the timsTOF Pro), it annotated considerably less peptides as well as proteins compared to the two other instruments (Fig. 3g). Still, the peptide/protein bias persists for the Orbitrap QE HF-X and timsTOF Pro comparison. Importantly, AlphaPept support for SCIEX was only preliminarily implemented through mzML preprocessing, potentially underlying the lower annotation rate and making it a perfect example of how the benchmark can support the development of compatibility in the future.

Data analysis: Quantification. Next, we used the benchmark design to verify to what extent the annotated peptides and proteins were found in their respective ratios (Fig. 1). Apart from being a benchmark target for quantification algorithms, this ground truth can also be considered a very rough empirical estimation of the false discovery rate (FDR)[1]. At the peptide level the spread of especially *Human* peptides into other ratios is increasingly visible in the order TripleTOF 6600+ < Orbitrap QE HF-X < timsTOF Pro, which alludes to increasing misannotation (Fig. 3g). While ratio estimation is known to become compromised with decreasing intensities of the MS1 features selected, the relative position of these peptides on the y-axis argues against this intensity explanation. Unfortunately, the exported ion counts of the timsTOF Pro differ at least 1–2 orders of magnitude compared to both the Orbitrap and the TripleTOF 6600+ and more advanced (conversion) algorithms are needed to compare these. However, at the protein level there was no overall skewing in the AlphaPept results that would point towards a significant increase in FDR in any of the instruments.

In conclusion, considerable bioinformatics challenges remain to be tackled before the intricacies of the data from contempory instrumentation and their acquisition parameters can truly be defined.

In conclusion, it is clear from this preliminary data usage case on the simple and most commonly applied DDA strategies that a lot of insights on data structure and bias were still hidden. Especially in light of the more recent DIA strategies, of which only a few were analysed in this Scientific Data report (Supplementary Fig. 1), a lot of work needs to be done before the field truly understands how data acquisition and data analysis effectively perform. Importantly, each step in the data processing can greatly impact the final outcome and we especially anticipate a renewed interest in (multidimensional) peak picking algorithms as a currently underappreciated preprocessing step of the complex DIA data[1]. Additionally, recent annotation tools have doubled the number of peptides that can be extracted from the same DIA data, an asset not expected for novel DDA annotation algorithms[21,28]. We would therefore like to invite the developers of all current bioinformatics tools to make use of this comprehensive dataset to benchmark their performance for each instrument individually and adapt their algorithms to increase the performance on all. This dataset was compiled from different labs and therefore also captures differences in instrument usage peculiar to the field, yet it all can be assessed in the Panorama QC metrics that are also publically available. Especially a better understanding of the detectable and annotatable ion space will move the field forward and help researchers make informed decisions on the best acquisition strategies for their application or biological question under investigation.

Code availability

MS²PIP is open source, licensed under the Apache-2.0 License, and hosted on https://github.com/compomics/ms2pip_c. The Jupyter notebooks used to generate Fig. 3(b,c,e,f) are available through Zenodo, under https://doi.org/10.5281/zenodo.5714380[48].

Received: 23 December 2021; Accepted: 23 February 2022;
Published online: 30 March 2022

References

1. Willems, S. *et al.* Ion-networks: A sparse data format capturing full data integrity of data independent acquisition mass spectrometry. *bioRxiv* (2019).
2. Vaudel, M. *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology* **33**, 22–24 (2015).
3. Verheggen, K., Martens, L., Berven, F. S., Barsnes, H. & Vaudel, M. Database Search Engines: Paradigms, Challenges and Solutions. in *Advances in Experimental Medicine and Biology* 147–156 (2016).
4. Geromanos, S. J., Hughes, C., Ciavarini, S., Vissers, J. P. C. & Langridge, J. I. Using ion purity scores for enhancing quantitative accuracy and precision in complex proteomics samples. *Analytical and bioanalytical chemistry* **404**, 1127–1139 (2012).
5. Richardson, K. *et al.* A Probabilistic Framework for Peptide and Protein Quantification from Data-Dependent and Data-Independent LC-MS Proteomics Experiments. *OMICS: A Journal of Integrative Biology* **16**, 468–482 (2012).
6. Li, G.-Z. *et al.* Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *PROTEOMICS* **9**, 1696–1719 (2009).
7. Helm, D. *et al.* Ion Mobility Tandem Mass Spectrometry Enhances Performance of Bottom-up Proteomics. *Molecular & Cellular Proteomics* **13**, 3709–3715 (2014).
8. Shliaha, P. V., Bond, N. J., Gatto, L. & Lilley, K. S. Effects of Traveling Wave Ion Mobility Separation on Data Independent Acquisition in Proteomics Studies. *Journal of Proteome Research* **12**, 2323–2339 (2013).
9. Gillet, L. C. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics* **11**, O111.016717 (2012).
10. Ludwig, C. *et al.* Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular Systems Biology* **14** (2018).
11. Ting, Y. S. *et al.* Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Molecular & cellular proteomics: MCP* **14**, 2301–7 (2015).
12. Li, Y. *et al.* Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. *Nature Methods* **12**, 1105–1106 (2015).
13. Kuharev, J., Navarro, P., Distler, U., Jahn, O. & Tenzer, S. In-depth evaluation of software tools for data-independent acquisition based label-free quantification. *PROTEOMICS* **15**, 3140–3151 (2015).
14. Teleman, J. *et al.* DIANA–algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics (Oxford, England)* **31**, 555–562 (2015).
15. Peckner, R. *et al.* Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. *Nature methods* **15**, 371–378 (2018).
16. Wang, J. *et al.* MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nature methods* **12**, 1106–1108 (2015).
17. Moseley, M. A. *et al.* Scanning Quadrupole Data-Independent Acquisition, Part A: Qualitative and Quantitative Characterization. *Journal of Proteome Research* **17**, 770–779 (2018).
18. Messner, C. B. *et al.* Ultra-fast proteomics with Scanning SWATH. *Nature Biotechnology* **39**, 846–854 (2021).
19. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nature Methods* **17**, 1229–1236 (2020).
20. Van Puyvelde, B. *et al.* Removing the Hidden Data Dependency of DIA with Predicted Spectral Libraries. *PROTEOMICS* **20**, 1900306 (2020).
21. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods* **17**, 41–44 (2019).
22. Bouwmeester, R., Gabriels, R., Van Den Bossche, T., Martens, L. & Degroeve, S. The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows. *Proteomics* **20**, 1900351 (2020).
23. Silva, A. S. C., Bouwmeester, R., Martens, L. & Degroeve, S. Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics* **35**, 5243–5248 (2019).
24. Zhou, X.-X. *et al.* pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Analytical Chemistry* **89**, 12690–12697 (2017).
25. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods* **16**, 509–518 (2019).
26. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods* **18**, 1363–1369 (2021).
27. Mann, M., Kumar, C., Zeng, W.-F. & Strauss, M. T. Artificial intelligence for proteomics and biomarker discovery. *Cell Systems* **12**, 759–770 (2021).
28. Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nature Biotechnology* **34**, 1130–1136 (2016).
29. Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications* **9**, 5128 (2018).
30. Pino, L. K., Just, S. C., MacCoss, M. J. & Searle, B. C. Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Molecular & Cellular Proteomics* **19**, 1088–1103 (2020).
31. Distler, U. *et al.* Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nature Methods* **11**, 167–170 (2014).
32. Meier, F. *et al.* Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Molecular & Cellular Proteomics* **17**, 2534–2545 (2018).
33. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research* **47**, D442–D450 (2019).
34. Dhaenens, M. & Perez-Riverol, Y. A comprehensive LFQ benchmark dataset to validate data analysis pipelines on modern day acquisition strategies in proteomics. *PRIDE Archive* https://identifiers.org/pride.project:PXD028735 (2021).
35. Dai, C. *et al.* A proteomics sample metadata representation for multiomics integration and big data analysis. *Nature Communications* **12**, 5854 (2021).
36. Bittremieux, W. *et al.* The European Bioinformatics Community for Mass Spectrometry (EuBIC-MS): an open community for bioinformatics training and research. *Rapid Communications in Mass Spectrometry* e9087 (2021).
37. Van Puyvelde, B. A comprehensive LFQ benchmark dataset to validate data analysis pipelines on modern day acquisition strategies in proteomics. *Panorama Public* https://doi.org/10.6069/ffcw-g217 (2021).
38. Bereman, M. S. *et al.* An Automated Pipeline to Monitor System Performance in Liquid Chromatography–Tandem Mass Spectrometry Proteomic Experiments. *Journal of Proteome Research* **15**, 4763–4769 (2016).
39. Sharma, V. *et al.* Panorama: A Targeted Proteomics Knowledge Base. *Journal of Proteome Research* **13**, 4205–4210 (2014).
40. Bereman, M. S. Tools for monitoring system suitability in LC MS/MS centric proteomic experiments. *Proteomics* **15**, 891–902 (2015).
41. Degroeve, S. & Martens, L. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics (Oxford, England)* **29**, 3199–3203 (2013).
42. Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods* **16**, 519–525 (2019).

43. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature Communications* **11**, 1–11 (2020).
44. Gabriels, R., Martens, L. & Degroeve, S. Updated MS^{2}PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Research* **47**, 295–299 (2019).
45. Diedrich, J. K., Pinto, A. F. M. & Yates, J. R. Energy Dependence of HCD on Peptide Fragmentation: Stepped Collisional Energy Finds the Sweet Spot. *Journal of the American Society for Mass Spectrometry* **24**, 1690–1699 (2013).
46. Tarn, C. & Zeng, W.-F. pDeep3: Toward More Accurate Spectrum Prediction with Fast Few-Shot Learning. *Analytical Chemistry* **93**, 5815–5822 (2021).
47. Strauss, M. T. *et al.* AlphaPept, a modern and open framework for MS-based proteomics. *bioRxiv* (2021).
48. Gabriels, R. & Van Puyvelde, B. Code to generate Fig. 3 and 4. A comprehensive LFQ benchmark dataset to validate data analysis pipelines on modern day acquisition strategies in proteomics. *zenodo* https://doi.org/10.5281/zenodo.5714380 (2021).

Acknowledgements

This research was funded by grants from the Research Foundation Flanders (FWO) awarded to BVP (grant number: 11B4518N), RG (1S50918N), and MD (12E9716N). Hans Vissers is acknowledged for his assistance with data conversion and formatting. This work was supported in part by the French Ministry of Research with the Investissement d'Avenir Infrastructures Nationales en Biologie et Santé program (ProFi, Proteomics French Infrastructure project; ANR-10-INBS-08).

Author contributions

B.V.P., S.W., S.D. and M.D. conceived the study, B.V.P. performed the TripleTOF 5600 and 6600+ data acquisition, S.D. performed the Synapt G2-Si data acquisition, A.G.P., D.B., E.M. and K.C. performed the Orbitrap data acquisition, K.B. performed the timsTOF Pro data acquisition. C.H. and L.G. performed the Synapt XS data acquisition. B.V.P. and S.D. prepared the samples and R.G. and S.W. wrote the scripts to generate Fig. 3. Y.P.R. organised the ProteomeXchange submission. N.B. and S.T. helped us to set up the acquisition methodologies on the TripleTOF 6600+ and provided valuable insights on the differences seen for each instrument. B.V.P. and M.D. wrote the draft manuscript with contributions from all authors. M.D. supervised and D.D. and L.M. co-supervised the experiment.

Competing interests

Chris Hughes and Lee Gethings are employed by Waters Corporation. Nic Bloomfield and Stephen Tate are employed by SCIEX.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01216-6.

Correspondence and requests for materials should be addressed to M.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.