







Generating high quality libraries for DIA MS with empirically corrected peptide predictions

Brian C. Searle ^{1,2}✉, Kristian E. Swearingen ¹, Christopher A. Barnes³, Tobias Schmidt ⁴,
Siegfried Gessulat ^{4,5}, Bernhard Küster ^{4,6} & Mathias Wilhelm ⁴

Data-independent acquisition approaches typically rely on experiment-specific spectrum libraries, requiring offline fractionation and tens to hundreds of injections. We demonstrate a library generation workflow that leverages fragmentation and retention time prediction to build libraries containing every peptide in a proteome, and then refines those libraries with empirical data. Our method specifically enables rapid, experiment-specific library generation for non-model organisms, which we demonstrate using the malaria parasite *Plasmodium falciparum*, and non-canonical databases, which we show by detecting missense variants in HeLa.

¹Institute for Systems Biology, Seattle, WA, USA. ²Proteome Software, Inc., Portland, OR, USA. ³Novo Nordisk Research Center Seattle, Inc., Seattle, WA, USA. ⁴Technical University of Munich, Freising, Germany. ⁵SAP SE, Potsdam, Germany. ⁶Bavarian Center for Biomolecular Mass Spectrometry, Freising, Germany. ✉email: bsearle@systemsbiology.org

Data-independent acquisition (DIA) mass spectrometry (MS) is a powerful label-free technique for deep, proteome-wide profiling^{1,2}. With DIA, mass spectrometers are tuned to systematically acquire tandem mass spectra at regular retention time and m/z intervals, freeing the method of the intensity-triggering biases introduced by data-dependent acquisition (DDA). To accomplish this, precursor isolation windows are widened such that multiple peptides are usually co-fragmented in the same MS/MS scan. DIA methods generally identify peptides with library search engines^{3–5} using experiment-specific spectrum libraries⁶ from DDA experiments. In peptide-centric searching⁷, library entries are scored according to retention time, such that the best-scoring time point for each peptide is reported. Only peptides present in the libraries can be detected, and the peptide detection reports must be corrected to limit the number of potential false discoveries⁸. Most importantly, these libraries are built at the expense of time, sample, and considerable effort with offline fractionation, especially considering that they are typically not reusable across laboratories or instrument platforms⁹.

When experiment-specific spectrum library generation is either impossible or impractical, as is frequently the case with non-model organisms, sequence variants, splice isoforms, or scarce sample quantities, software tools such as Pecan¹⁰ and DIA-Umpire¹¹ can detect peptides from DIA experiments without a spectrum library by directly searching every peptide in FASTA databases. Gas-phase fractionation¹² (GPF) improves detection rates with these tools¹⁰ by injecting the same sample multiple times with tiled precursor isolation windows, allowing each injection to have narrower windows (and thus fewer co-fragmented peptides) with the same instrument duty cycle. While offline fractionation requires an additional liquid chromatography (LC) step using orthogonal separation modes to online LC-MS, GPF occurs completely within the mass spectrometer, making it both more reproducible and easier to perform. While this method is often prohibitively expensive because it requires enough instrument time and protein content for multiple injections for each sample, the use of multiple GPF injections can be applied just to pooled samples to generate DIA-only chromatogram libraries that make it easier to detect peptides in single-injection DIA experiments¹³. However, even when using GPF, these tools still generally detect fewer peptides than library search engines, which can leverage previously acquired instrument-specific fragmentation and measured retention times.

Recently it has become possible to accurately predict spectra from peptide sequences^{14,15}, but direct searching of single-injection DIA data has remained problematic, in part due to the false discovery rate (FDR) correction required when considering all possible tryptic peptides in a FASTA database. Proteins show 3–4 orders of magnitude difference in intensity between the best- and worst-responding tryptic peptides¹⁶, and only considering the best-responding peptides in libraries can improve detection rates by lessening the required FDR correction. This approach has been applied by generating independent assay libraries for each DIA injection, either by searching the DIA data directly^{5,11}, or using paired DIA/DDA experiments^{6,17}.

Here we demonstrate an approach to generate DIA-only chromatogram libraries from GPF-DIA injections using peptide fragmentation and retention time predictions. This method creates empirically corrected libraries that sidestep the issues of directly searching predicted libraries, because the GPF-DIA injections use the same acquisition parameters, chromatographic conditions, and sample matrix as quantitative single-injection DIA experiments. We observe improved peptide detection rates when searching these empirically corrected libraries over searching sample-specific DDA libraries.

Empirically corrected libraries are built directly from protein sequence databases, allowing our workflow to enable experiments that identify protein-level genetic variants and quantify peptides from non-model organisms.

Results

Empirically corrected libraries from peptide predictions. Here we report on a DIA-only workflow that produces higher-quality libraries than those generated by DDA while simultaneously supplanting the need for any offline fractionation. Our workflow (Fig. 1, Methods) uses a recently developed deep neural network, Prosit¹⁴, to generate a predicted spectrum library of fragmentation patterns and retention times for every +2H and +3H tryptic peptide in a FASTA database, with up to one missed cleavage. Fragmentation prediction in Prosit adjusts based on normalized collision energy (NCE), and we tune the NCE parameter for each peptide charge state to account for DIA-specific fragmentation.

Building on the chromatogram library method¹³, we make a pool of sub-aliquots from a representative subset of biological samples in our experiment. In addition to analyzing each biological sample using single-injection DIA (typically 4- to 12 m/z -wide precursor isolation windows after staggered-window demultiplexing¹⁸, depending on the instrumentation) with a 90-min gradient, we collect six additional GPF-DIA acquisitions (typically 2 m/z -wide precursor isolation windows after demultiplexing, regardless of instrumentation) of the sample pool using the same gradient. Considering column washes, these GPF-DIA acquisitions take approximately 12 total additional hours of MS acquisition. We find that single-injection DIA can benefit from tuning the precursor isolation window to suit the mass spectrometer acquisition rate. However, GPF-DIA acquisitions tend to be ion population-limited, and narrowing precursor isolation windows below 2 m/z (after demultiplexing) does not improve detection rates (data not shown). In part, this may be related to smaller windows breaking up isotopic envelopes, resulting in lower overall sensitivity.

We search the GPF-DIA acquisitions of the pool against the predicted spectrum library using EncyclopeDIA. Searching with the predicted spectrum library has multiple disadvantages over experiment-specific libraries. First, correcting for false discoveries requires more stringent thresholds when considering every possible peptide in a proteome, rather than just those previously detected in a pooled sample. Secondly, while Prosit typically produces higher-quality spectrum libraries with deep learning than other, more conventional approaches¹⁹, the predictions are less accurate than experiment-specific libraries generated on the same instrumentation. However, these two disadvantages are mitigated by the use of GPF-DIA with precursor isolation windows as narrow as those used in targeted parallel reaction monitoring²⁰ (PRM) or DDA.

Finally, we use the detections made with GPF-DIA to construct a new, empirically corrected library removing the disadvantages of the predicted spectrum library. Assuming that virtually every consistently quantifiable peptide in an experiment is detectable from the pool using GPF-DIA, we filter the predicted library to remove peptides that cannot be found in the pool. In addition, we select the highest-scoring (and therefore easiest to detect) charge state for each peptide and remove other, lower-scoring charge states from the library. Then, for each identified peptide, we calculate the aggregate peak shape across all of the identified fragment ions and extract fragment peak area intensities for all possible B- or Y-type ions that correlate with this shape. Since the GPF-DIA injections are performed using the same instrumentation setup as the single-injection DIA injections, we use these intensities as the fragmentation patterns in the empirically

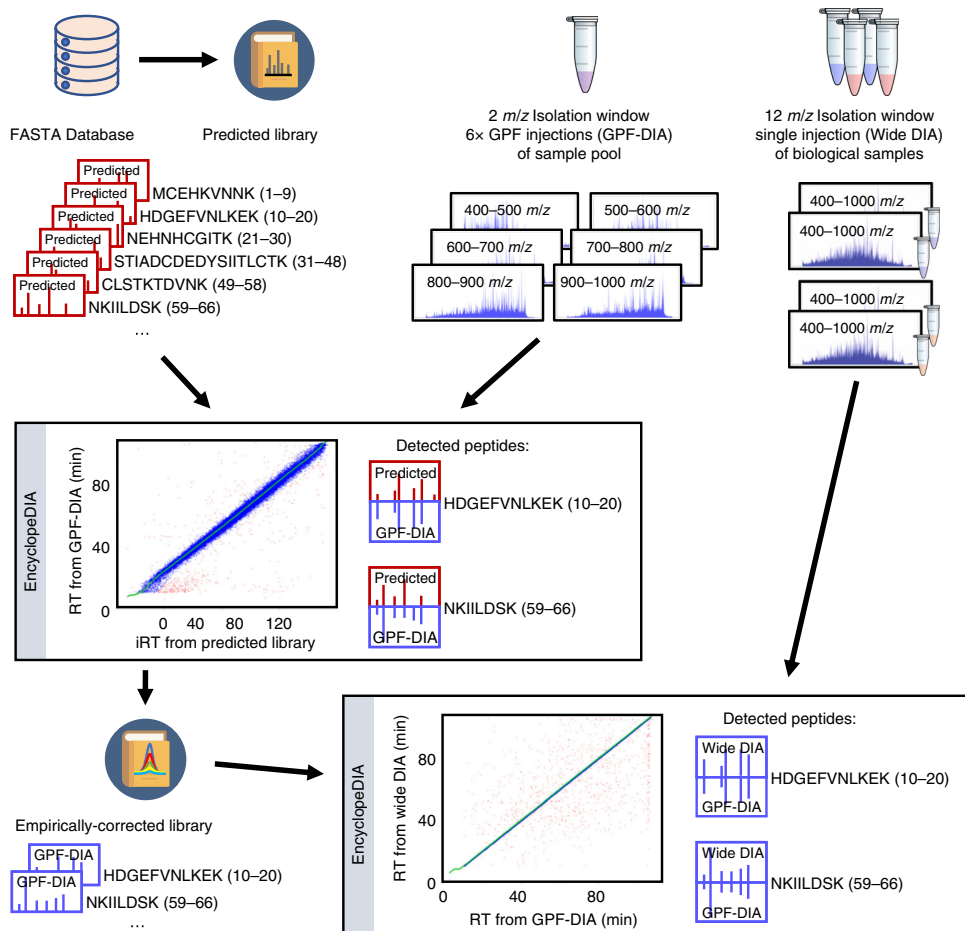


Fig. 1 Workflow for generating empirically corrected libraries. Fragmentation patterns and indexed retention times (iRTs) are generated with Prosit for all possible tryptic peptides in a FASTA database, and these predictions are compiled into a predicted spectrum library. In this example, peptides from CDPK2 are shown with start/stop indices within the protein indicated in parentheses (red predicted spectra). We use EncyclopeDIA to search GPF-DIA acquisitions of a sample pool with that library, and peptides detection results are compiled into a experiment-specific, empirically corrected library. This new library contains fragmentation patterns and retention times extracted from the GPF-DIA data for only the detected peptides (blue empirical spectra). Since GPF-DIA and single-injection DIA have the same instrumentation and on-column matrix, retention times and fragmentation patterns in the empirically corrected library are more closely aligned than the original predictions.

corrected library. Similarly, we use the time point of the apex intensity of the aggregate peak shape as the retention time in the new library. We find that while peptide ordering on the same HPLC platform with the same column and method is typically very high, we still benefit from retention time alignment to account for fluctuations in run-to-run chromatography stability. In addition, we perform the six GPF-DIA injections of the pool near the middle of an experiment after at least one full set of biological replicates, to limit variability caused by column (re)conditioning to a new proteome composition. We recommend reacquiring the six GPF-DIA injections if the column or gradient change while an experiment is being conducted.

Validating the empirically corrected library methodology. We first applied our method to analyze a yeast tryptic digest on a Thermo Fusion Lumos MS. After column conditioning, we acquired four replicate injections of single injection 400–1000 m/z DIA using 4 m/z-wide windows (after demultiplexing) at 20 ms ion injection time. We followed this by six GPF-DIA injections from 400 to 500 m/z, 500 to 600 m/z, 600 to 700 m/z, 700 to 800 m/z, 800 to 900 m/z, and 900 to 1000 m/z with 2 m/z-wide windows (after demultiplexing) at 60 ms ion injection time. Using

a Uniprot *Saccharomyces cerevisiae* protein database (6729 entries), we produced a predicted spectrum library containing 456,511 total entries with 320,150 unique peptide sequences, assuming an NCE of 33. After empirical correction, the new library contained 64,597 unique peptide sequences from 4464 protein groups at a 1% peptide and protein FDR.

Although searching single-injection DIA acquisitions directly with predicted spectrum libraries is highly dependent on prediction accuracy, we found that our workflow produced high-quality libraries even if the predictions were not precisely tuned for the instrument, which suggests broad cross-platform applicability. We demonstrated this by modulating the NCE setting in Prosit (Fig. 2a) and comparing with two libraries: (1) an experiment-specific offline high-pH reversed-phase (HpH-RP) fractionated DDA spectral library containing 39,612 unique peptide sequences acquired at the same time as the DIA experiment and (2) a sample-specific offline SCX fractionated DDA spectral library containing 45,987 unique peptide sequences from another study. Even across a wide range of NCE settings, searching GPF-DIA spectra produced between 33% and 60% larger empirically corrected libraries than we could obtain from the experiment-specific 10-fraction HpH-RP fractionated DDA library (Fig. 2b). Interestingly, the optimal NCE setting for Prosit

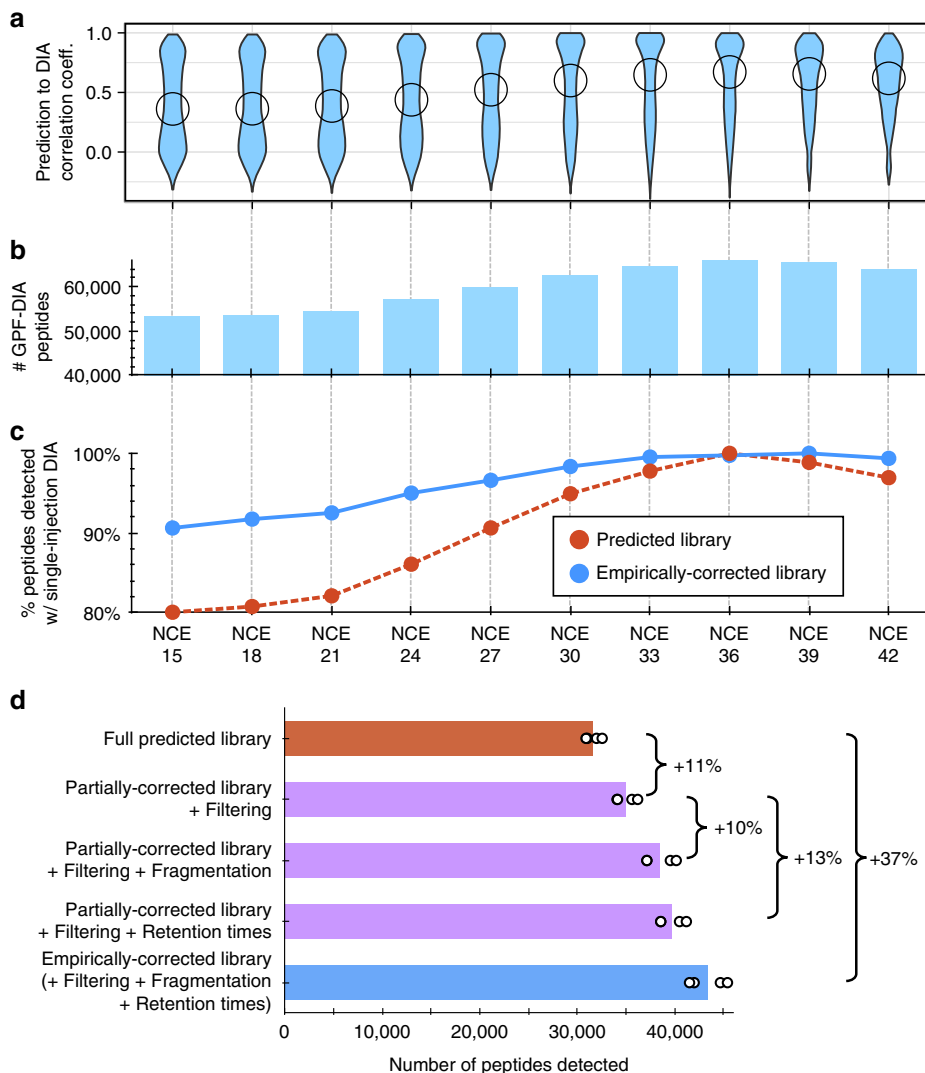


Fig. 2 Evaluating empirically corrected libraries made with peptide predictions. **a** Violin plots showing spectral correlation between predicted library for yeast peptides and single-injection DIA ($N = 1$) at various NCE settings (circles indicate medians) when the instrument was configured for NCE = 33. **b** The total numbers of empirically corrected library entries detected from GPF-DIA at various NCE settings ($N = 1$). **c** The fraction of peptides detected in single-injection DIA ($N = 1$) relative to the optimal NCE for empirically corrected libraries (blue line) are more consistent than predicted libraries (red dashed line) across a wide range of NCE settings. **d** The number of yeast peptides detected at 1% peptide FDR in single-injection DIA acquisitions ($N = 4$) using the NCE = 33 chromatogram library, where either the retention times or fragmentation patterns have been switched with the predicted Prosit values (purple bars). Compared to the predicted spectrum library search (red bar), an 11% increase comes from simply using a narrowed peptide selection. DIA-based retention times and fragmentation patterns provide a 13% and 10% increase over this, respectively. Comparing the empirically corrected library and the predicted library detections (blue bar, 37% increase), these percentage gains appear to be nearly multiplicative (i.e., $111\% \times 110\% \times 113\% = 138\%$), indicating that all three factors are independent and of roughly equal importance. Source data are provided as a Source Data file.

was 36 (not the instrument method-specified 33), which likely reflects calibration drift and variation across Orbitrap instruments²¹. Fewer peptides will be detectable in both single-injection DIA and GPF-DIA data at incorrect NCE settings. However, since there is less interference in GPF-DIA, these detection rates do not drop as quickly. After empirical correction, the library will contain fragmentation patterns observed in the GPF-DIA data rather than the original library tuning parameters (Supplementary Fig. 1), and any peptide that can be detected in the GPF-DIA data will be easier to detect in single-injection DIA. In this way, the GPF-DIA functions as a calibration step that corrects the Prosit NCE setting, making searches of single-injection DIA less sensitive to NCE drift or other sources of prediction inaccuracies after empirical correction (Fig. 2c).

Retention time is affected by chromatographic conditions, but also by matrix effects. Single-injection DIA and DDA both measure peptides within the full matrix. Offline fractionation, such as SCX or high-pH reversed-phase, change the matrix by fractionating the peptide mixture into multiple samples. Consequently, each peptide sees a different matrix as it elutes relative to the single-injection injections, causing errors in retention time estimates. Unlike offline fractionation, GPF-DIA uses the quadrupole for fractionation, maintaining the same full matrix complexity of single-injection DIA and improving retention time consistency. This process produced libraries with better retention time accuracy (80% of peptides within 35 s) than both the predicted (80% within 5.4 min) and fractionated DDA libraries (80% within 4.6 min), even when the DDA libraries were acquired

on the same instrument (Supplementary Fig. 2). Coupled with smaller library size and improved fragmentation patterns (Supplementary Fig. 3), these three factors had roughly equal and orthogonal improvements over directly searching single-injection DIA with predicted libraries (Fig. 2d).

We were interested to determine if empirically corrected libraries could be reused for different experiments. To test this, we reanalyzed yeast datasets¹³ from a Thermo QE-HF MS at a different location using the empirically corrected library generated in this study on a Thermo Fusion Lumos MS. We found we were able to detect more peptides using the empirically corrected library than could be detected by analyzing the same data with a Prosit-predicted library or FASTA-only approach using Pecan. However, even better results were produced if the data were

analyzed with a library built from GPF-DIA injections collected on the same instrument (Supplementary Fig. 5). In this case, collecting additional GPF-DIA injections and building an empirically corrected library for each experiment improved peptide detection rates by 30%.

While HpH-RP and SCX fractionated DDA produced similar-sized libraries, these libraries draw from different pools of peptides (Supplementary Fig. 4), demonstrating that combining both fractionation methods is necessary for building comprehensive DDA libraries. We observed that searching a combined HpH-RP and SCX library produced more detections in single-injection DIA datasets than either DDA library individually, but that overall, searching single-injection DIA acquisitions with an empirically corrected library detected 31% more yeast peptides (Fig. 3a). Both

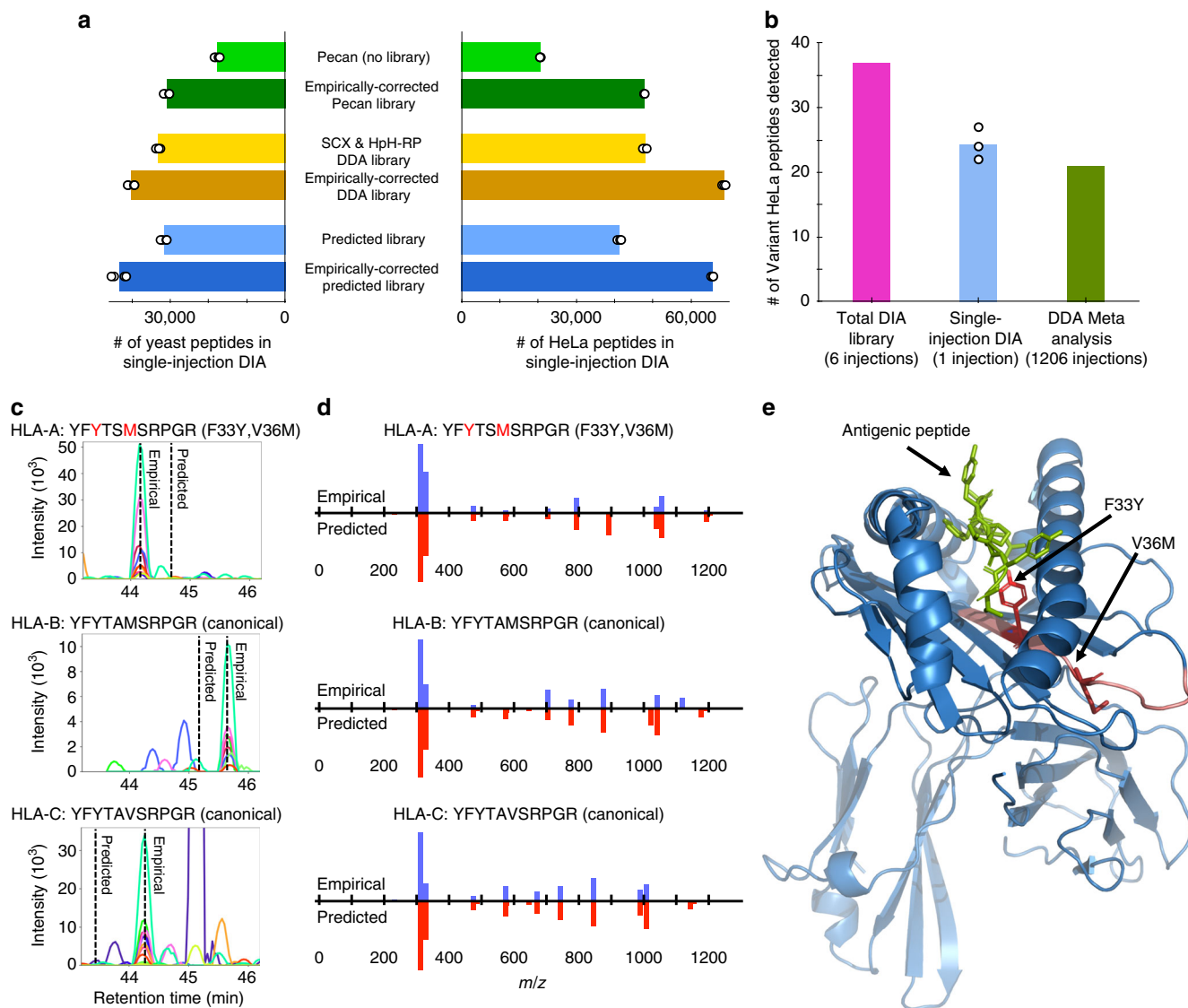


Fig. 3 Empirically corrected libraries improve peptide and missense variant detection rates. **a** The average number of yeast ($N = 4$) and HeLa ($N = 3$) peptides in single-injection DIA detected using library-free searching with Pecan, searching with a combined SCX fractionated and high-pH reverse-phase fractionated DDA spectrum library, or searching with a predicted spectrum library, before or after empirical correction with the chromatogram library method. Match-between-runs was not enabled for any search strategy so that replicates are independent measurements demonstrating the technical variability for each approach. **b** The number of HeLa missense variants detected in the total DIA library, single-injection DIA ($N = 3$), or a meta-analysis of 40 published DDA experiments²⁷, each filtered to a 1% peptide/protein FDR. **c** Retention time predictions differ somewhat from empirical data from GPF-DIA for homologous peptides: YFYTSMRPGR from HLA-A (F33Y,V36M), YFYTAMSRPGR from HLA-B, and YFYTAVSRPGR from HLA-C. **d** Relative +1H y-type and b-type fragmentation patterns above $200m/z$ for the same peptides are shown as butterfly plots with empirical intensities (blue) above predicted intensities (red). **e** All three peptides are in the HLA peptide binding/presentation region, as indicated by YFYTSMRPGR (red) in HLA-A (blue) relative to an antigenic peptide (green) in PDB structure 1AO7. Source data are provided as a Source Data file.

protein kinase CDPK2 (PF3D7_0610600). *P. falciparum* parasites lacking CDPK2 develop normally through asexual stages, but male gametocytes are incapable of undergoing exflagellation to become gametes, thereby preventing transmission to the mosquito vector³⁵. Our work validates that CDPK2 is indeed present at measurable levels in gametocytes, paving the way to monitor dynamic expression of this kinase over the course of parasite maturation.

Discussion

In conclusion, empirical correction of predicted spectrum libraries enables rapid experiment-specific library generation for non-canonical proteomes or non-model organisms without offline fractionation. DDA-based spectral libraries can become stale over time as columns are changed or NCE tuning drifts within an instrument. While the method we propose to create empirically corrected libraries requires an extra 6 GPF-DIA injections for each new experiment, the procedure has the advantage of ensuring that the library is always up-to-date, and even accounts for variation across different instrument platforms.

In addition to DIA applications, this method is applicable for building accurate mass and time tag^{36,37} libraries for MS1-only data acquisition strategies, such as BoxCar³⁸, an approach that forgoes collecting MS/MS and relies on highly accurate mass and retention time indices to identify peptides using match-between-runs. Error rates for match-between-runs peptide detection without MS/MS spectra are often higher than 1% FDR and are hard to estimate without controlled experiments³⁹. Errors caused by approaches such as BoxCar are likely exacerbated when the on-column matrix changes, such as between fractionated and unfractionated proteomes. Since our approach builds libraries using the same on-column matrix, retention time tags built with GPF-DIA will likely remove this source of variability.

We also developed a graphical user interface in EncyclopeDIA (Supplementary Note 1) to facilitate making empirically corrected libraries for new proteomes from any FASTA database, which can be converted for external use in both Skyline⁴⁰ and OpenSwath⁴. While our interface currently does not support analyzing peptides with PTMs, we believe that when prediction software improves for PTMs, our approach to library building will work for those peptides as well. To encourage the reuse of our method, we have released a growing repository of pre-generated predicted libraries compatible with EncyclopeDIA, Skyline, and Scaffold DIA, which are available for download at ProteomicsDB [<https://www.proteomicsdb.org/prosit/libraries>].

Methods

***S. cerevisiae* culture and sample preparation.** As described in Searle et al.,¹³ *S. cerevisiae* strain BY4741 (Dharmacon) was cultured at 30 °C in YEPD and harvested at the mid-log phase. Cells were pelleted and lysed in a buffer of 8 M urea, 50 mM Tris (pH 8), 75 mM NaCl, 1 mM EDTA (pH 8) followed by seven cycles of 4 min bead beating with glass beads. After a 1 min rest on ice, lysate was collected by piercing the tube and centrifuging for 1 min at 3000 × g and 4 °C into an empty eppendorf. After further centrifugation at 21,000 × g and 4 °C for 15 min, the protein content of the supernatant was removed and estimated using BCA. Proteins were then reduced with 5 mM dithiothreitol for 30 min at 55 °C, alkylated with 10 mM iodoacetamide in the dark for 30 min at room temperature, and diluted to 1.8 M urea, before digestion with sequencing-grade trypsin (Pierce) at a 1:50 enzyme-to-substrate ratio for 16 h at 37 °C. In all, 5 N HCl was added to approximately pH 2 to quench the digestion, and the resulting peptides were desalted with 30 mg MCX cartridges (Waters). Peptides were dried with vacuum centrifugation and brought to 1 µg/3 µl in 0.1% formic acid (buffer A) prior to MS acquisition. All measurements of yeast were performed on the same biological replicate to assess technical variability in the method.

***P. falciparum* culture and red blood cell sample preparation.** Human O + erythrocytes (RBCs) were obtained from Valley Biomedical (Winchester, VA; catalog number HP10020). Three biologically replicate flasks of stage IV/V *P.*

falciparum NF54 gametocytes were prepared. Asexual cultures were synchronized with sorbitol and set up at 5% hematocrit and 1% parasitemia. Gametocytogenesis was induced by withholding fresh blood and allowing parasitemia to increase. N-acetyl glucosamine was added to media for 4 days beginning 7 days after setup in order to remove asexual parasites. Gametocyte-infected erythrocytes (giRBC) were enriched from uninfected erythrocytes (uiRBC) by magnetic-activated cell sorting at stage III. Stage IV/V gametocytes were collected on day 15 post-setup. Additional uiRBC were also prepared by washing multiple times in RPMI and stored at 50% hematocrit. giRBC and uiRBC cells were lysed in a buffer of 10% sodium dodecyl sulfate (SDS), 100 mM ammonium bicarbonate (ABC), cComplete EDTA-free Protease Inhibitor Cocktail (Sigma), and Halt Phosphatase Inhibitor Cocktail (Thermo Scientific). Proteins were then reduced with 20–40 mM tris(2-carboxyethyl)phosphine (TCEP) for 10 min at 95 °C and alkylated with 40–80 mM iodoacetamide in the dark for 20 min at room temperature. After centrifugation at 16,000 × g to pellet insoluble material, proteins were purified with methanol: chloroform extraction⁴¹ and dried and resuspended in 8 M urea buffer before the content was estimated using BCA. After dilution to 1.8 M urea, proteins were digested with sequencing-grade trypsin (Promega) at a 1:40 enzyme-to-substrate ratio for 15 h at 37 °C. The resulting peptides were desalted with Sep-Pak cartridges (Waters), dried with vacuum centrifugation, and brought to 1 µg/3 µl in 0.1% formic acid (buffer A) prior to MS acquisition. In addition, several digested peptide mixtures were made by diluting peptides from one flask of giRBC cells with peptides from uiRBC cells at ratios of 1:0, 2:1, 7:8, 4:15, 1:9, 2:41, 2:91, and 1:99 giRBC:uiRBC.

LC MS (*S. cerevisiae*). Tryptic *S. cerevisiae* peptides were separated with a Thermo Easy nLC 1200 on self-packed 30 cm columns packed with 1.8 µm ReproSil-Pur C18 silica beads (Dr. Maisch) inside of a 75 µm inner diameter fused silica capillary (#PF360 Self-Pack PicoFrit, New Objective). The 30 cm column was coiled inside of a Sonation PRSO-V1 column oven set to 35 °C prior to ionization into the MS. The HPLC was performed using 200 nl/min flow with solvent A as 0.1% formic acid in water and solvent B as 0.1% formic acid in 80% acetonitrile. For each injection, 3 µl (approximately 1 µg) was loaded and eluted with a linear gradient from 7% to 38% buffer B over 90 min. Following the linear separation, the system was ramped up to 75% buffer B over 5 min and finally set to 100% buffer B for 15 min, which was followed by re-equilibration to 2% buffer B prior to the subsequent injection. Data were acquired using DIA.

The Thermo Fusion Lumos was set to acquire six GPF-DIA acquisitions of a biological sample pool using 120,000 precursor resolution and 30,000 fragment resolution. The automatic gain control (AGC) target was set to 4e5, the maximum ion inject time (IIT) was set to 60 ms, the NCE was set to 33, and +2H was assumed as the default charge state. The GPF-DIA acquisitions used 4m/z precursor isolation windows in a staggered-window pattern with optimized window placements (i.e., 398.4 to 502.5m/z, 498.5 to 602.5m/z, 598.5 to 702.6m/z, 698.6 to 802.6m/z, 798.6 to 902.7m/z, and 898.7 to 1002.7m/z). Individual samples for proteome profiling acquisitions used single-injection DIA acquisitions (120,000 precursor resolution, 15,000 fragment resolution, AGC target of 4e5, max IIT of 20 ms) using 8m/z precursor isolation windows in a staggered-window pattern with optimized window placements from 396.4 to 1004.7m/z.

For generation of an *S. cerevisiae* spectral library, 80 µg of the same tryptic digests described above were separated into 10 total fractions using the Pierce high-pH reversed-phase peptide fractionation kit (Thermo, #84868). Briefly, peptides were loaded onto hydrophobic resin spin column and eluted using the following 8 acetonitrile steps: 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20.0%, and 50%, keeping both the wash and flow-through. The resulting peptide fractions were injected into the same Thermo Fusion Lumos using the same chromatography setup and column described above, but configured for DDA. After adjusting each fraction to an estimated 0.5–1.0 µg on column, the fractions were measured in a top-20 configuration with 30 s dynamic exclusion. Precursor spectra were collected from 300–1650m/z at 120,000 resolution (AGC target of 4e5, max IIT of 50 ms). MS/MS were collected on +2H to +5H precursors achieving a minimum AGC of 2e3. MS/MS scans were collected at 30,000 resolution (AGC target of 1e5, max IIT of 50 ms) with an isolation width of 1.4m/z with a NCE of 33.

LC MS (*P. falciparum*). Tryptic *P. falciparum* and RBC peptides were separated with a Thermo Easy nLC 1000 and emitted into a Thermo Q-Exactive HF. In-house laser-pulled tip columns were created from 75 µm inner diameter fused silica capillary and packed with 3 µm ReproSil-Pur C18 beads (Dr. Maisch) to 30 cm. Trap columns were created from Kasil fritted 150 µm inner diameter fused silica capillary and packed with the same C18 beads to 2 cm. The HPLC was performed using 250 nl/min flow with solvent A as 0.1% formic acid in water and solvent B as 0.1% formic acid in 80% acetonitrile. For each injection, 3 µl (approximately 1 µg) was loaded and eluted using a 84-min gradient from 6% to 40% buffer B, followed by steep 5-min gradient from 40% to 75% buffer B and finally set to 100% buffer B for 15 min, which was followed by re-equilibration to 0% buffer B prior to the subsequent injection. Data were acquired using either DDA or DIA.

The Thermo Q-Exactive HF was set to acquire DDA in a top-20 configuration with auto dynamic exclusion. Precursor spectra were collected from 400 to 1600m/z at 60,000 resolution (AGC target of 3e6, max IIT of 50 ms). MS/MS were

collected on +2H to +5H precursors achieving a minimum AGC of 1e4. MS/MS scans were collected at 15,000 resolution (AGC target of 1e5, max IIT of 25 ms) with an isolation width of 1.4m/z with a NCE of 27. Additionally, six GPF-DIA acquisitions were acquired of a biological sample pool (60,000 precursor resolution, 30,000 fragment resolution, AGC target of 1e6, max IIT of 60 ms, NCE of 27, +3H assumed charge state) using 4m/z precursor isolation windows in a staggered-window pattern with optimized window placements (i.e., 398.4–502.5m/z, 498.5–602.5m/z, 598.5–702.6m/z, 698.6–802.6m/z, 798.6–902.7m/z, and 898.7–1002.7m/z). Individual samples used single-injection DIA acquisitions (60,000 precursor resolution, 30,000 fragment resolution, AGC target of 1e6, max IIT of 60 ms) using 16m/z precursor isolation windows in a staggered-window pattern with optimized window placements from 392.4 to 1008.7m/z.

FASTA databases and predicted spectrum libraries. Species-specific reviewed FASTA databases for *Homo sapiens* (25 April 2019, 20415 entries) and *Saccharomyces cerevisiae* (25 January 2019, 6729 entries) were downloaded from Uniprot. The *Plasmodium falciparum* FASTA database⁴² was downloaded from PlasmoDB³⁴ version 43 (24 April 2019, 5548 entries). The Ensembl-based HeLa-specific FASTA database²⁷ was downloaded from the ACS Publications website and modified to be compatible with EncyclopeDIA (47,305 entries, including both canonical and variant protein sequences). Each database was digested in silico to create all possible +2H and +3H peptides with precursor m/z within 396.43 and 1002.70, assuming up to one missed cleavage. Peptides were further limited to be between 7 and 30 amino acids to match the restrictions of the ProSight tool¹⁴. In general, NCE were assumed to be 33 (yeast was processed using NCE from 15 to 42 in 3 NCE increments) but modified to account for charge state. Since DIA assumes all peptides are of a fixed charge, we adjusted the NCE setting as if peptides were fragmented at the wrong charge state using the formula:

$$\text{Adjusted NCE} = \text{NCE} \times \frac{\text{factor}(\text{default charge})}{\text{factor}(\text{peptide charge})}, \quad (1)$$

where the factors were 1.0 for +1H, 0.9 for +2H, 0.85 for +3H, 0.8 for +4H, and 0.75 for +5H and above⁴³. After submitting to ProSight, predicted MS/MS and retention times were converted to the EncyclopeDIA DLIB format for further processing. Scripts to produce ProSight input from FASTAs and build EncyclopeDIA-compatible spectrum libraries from ProSight output are available as functions in EncyclopeDIA 1.0.

DDA data processing. All Thermo RAW files were converted to.mzML format using the ProteoWizard package⁴⁴ (version 3.0.18299) using vendor peak picking. DDA data were searched with Comet⁴⁵ (version 2017.01 rev. 1), allowing for fixed cysteine carbamidomethylation, variable peptide n-terminal pyro-glu, and variable protein n-terminal acetylation. Fully tryptic searches were performed with a 50 ppm precursor tolerance and a 0.02 Da fragment tolerance permitting up to two missed cleavages. High-pH reversed-phase fractions were combined and search results were filtered to a 1% peptide-level FDR using PeptideProphet⁴⁶ from the Trans-Proteomic Pipeline⁴⁷ (TPP version 5.1.0). A yeast-specific Biblispec⁴⁸ DDA spectrum library was created from Thermo Q-Exactive DDA data using Skyline^{40,49} (Daily version 19.0.9.149).

P. falciparum and RBC DDA data were additionally processed with MaxQuant³³ (version 1.6.5.0) to perform label-free quantitation with precursor ion integration. MaxQuant was configured to use default parameters, briefly fixed cysteine carbamidomethylation, variable methionine oxidation, and variable protein n-terminal acetylation. Fully tryptic searches were performed with a 20 ppm fragment tolerance using both the human and *P. falciparum* FASTA databases, as well as common contaminants and filtered to a 1% peptide-level FDR. Quantification was performed using unique and razor peptides with the match-between-runs setting turned on.

DIA data processing. DIA data were overlap demultiplexed¹⁸ with 10 ppm accuracy after peak picking in ProteoWizard (version 3.0.18299). Searches were performed using EncyclopeDIA (version 0.8.3), which was configured to use default settings: 10 ppm precursor, fragment, and library tolerances. EncyclopeDIA was allowed to consider both B and Y ions and trypsin digestion was assumed.

FDR estimation. All searches are performed using the target/decoy strategy⁵⁰. As previously described¹³, EncyclopeDIA generates decoy peptide sequences by keeping the first and last amino acids in place, but reversing the remaining inbetween sequence. Decoy spectra are generated by moving all fragment ions corresponding to amino acids to the mass appropriate for the new decoy sequence. Each decoy peptide retains the same retention time as the corresponding target peptide. Retention time is only used as a feature (not a filter), so every peptide (decoy or target) can be assigned at any retention time. This approach is designed to give decoys a chance to produce higher scores and better model truly incorrect peptides. EncyclopeDIA search results were filtered to a 1% peptide-level using Percolator 3.1 (refs. ^{51,52}). Proteins are then parsimoniously allocated to protein groups and filtered to a 1% protein-level FDR.

Empirically corrected library generation. Predicted libraries were corrected with EncyclopeDIA using the chromatogram library method described previously¹³, and a tutorial for this process is outlined in Supplementary Note 1. Briefly, GPF-DIA injections for a given study were loaded into EncyclopeDIA using the above parameters, where the search library was set to the appropriate predicted spectrum library. Peptides detected by EncyclopeDIA were exported as a chromatogram library.

Percolator is rerun on peptides detected from the GPF-DIA injections to globally filter peptide detections to a 1% FDR. Only peptides detected at a 1% peptide FDR in both the individual GPF-DIA injection and the global analysis are retained for the empirically corrected library. For each detected peptide, fragment ion chromatograms are Savitzky-Golay smoothed⁵³, normalized to the same peak area, and a peptide peak shape is calculated using median smoothing between these chromatograms. A Pearson's correlation score is calculated for every fragment ion indicating the agreement between the overall peptide peak shape and the fragment peak shape.

A peptide entry in a chromatogram library is similar to a peptide entry in a spectrum library in that it contains a precursor mass, retention time, and a fragmentation spectrum. In addition, a chromatogram library entry also contains the peptide peak shape and a correlation score for each fragment ion. This score provides an indication of the likelihood the fragment ion was interfered with in the GPF-DIA injection, with the expectation that it will also likely be interfered with in single-injection DIA as well. This process created a new empirically corrected library containing only peptides found in the GPF-DIA samples, and also retained empirical fragment ion intensities and retention times observed from the DIA data. These libraries were made to be compatible with both EncyclopeDIA and Skyline and were used for downstream analysis of single-injection DIA.

After library generation, FDR estimation for single-injection DIA experiments was performed twice: once at the individual injection level, and again globally across all quantitative samples. For peptide detection experiments, the match-between-runs approach was not used. For the quantitative *P. falciparum* experiments, match-between-runs was applied for peptides not detected in every injection, but only if the peptide was detected at a 1% FDR in the global analysis and at a 1% FDR in at least one individual injection.

Further validation was used for constructing the HeLa-specific library. Here, peptides with similar sequences that fall in the same precursor isolation window can be incorrectly identified by shared fragment ions alone. This class of peptides falls outside of target/decoy-based false discovery estimation and require additional FDR control. Missense variants in the HeLa empirically corrected library were manually validated by checking for variant-specific ions that follow the peak shape. Peptide detections made with no variant-specific ions were considered likely false discoveries and removed from the library.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw data from the yeast and *P. falciparum* studies are available at MassIVE (MSV000084000 [<https://doi.org/10.25345/C5BD2H>]), ProteomeXchange PXD: PXD017705) and file descriptions are listed in Supplementary Data 3. The raw data from the HeLa reanalysis are available as originally published at MassIVE (MSV000082805 [ftp://massive.ucsd.edu/MSV000082805/]). The source data underlying Figs. 2a–d, 3a, and 4a–c and Supplementary Figs. 2, 3 and 8 are provided as a Source Data file. All other data are available from the corresponding author on reasonable request.

Code availability

ProSight (<https://www.proteomicsdb.org/prosight>) and EncyclopeDIA 1.0 (<https://bitbucket.org/searleb/encyclopedia>) are both available under the Apache 2 open-source license.

Received: 21 August 2019; Accepted: 28 February 2020;

Published online: 25 March 2020

References

- 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. *Mol. Cell. Proteomics* **11**, O111.016717 (2012).
- Doerr, A. DIA mass spectrometry. *Nat. Methods* **12**, 35 (2014).
- Weisbrod, C. R., Eng, J. K., Hoopmann, M. R., Baker, T. & Bruce, J. E. Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. *J. Proteome Res.* **11**, 1621–1632 (2012).
- Röst, H. L. et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219–223 (2014).
- Wang, J. et al. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nat. Methods* **12**, 1106–1108 (2015).

6. Schubert, O. T. et al. Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.* **10**, 426–441 (2015).
7. Ting, Y. S. et al. Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data. *Mol. Cell. Proteomics* **14**, 2301–2307 (2015).
8. Rosenberger, G. et al. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* **14**, 921–927 (2017).
9. Bruderer, R. et al. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteomics* **16**, 2296–2309 (2017).
10. Ting, Y. S. et al. PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat. Methods* **14**, 903–908 (2017).
11. Tsou, C.-C. et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258 (2015).
12. Panchoaud, A. et al. Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal. Chem.* **81**, 6481–6488 (2009).
13. Searle, B. C. et al. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9**, 5128 (2018).
14. Gessulat, S. et al. ProSight: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
15. Tiwary, S. et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525 (2019).
16. Searle, B. C., Egertson, J. D., Bollinger, J. G., Stergachis, A. B. & MacCoss, M. J. Using Data Independent Acquisition (DIA) to model high-responding peptides for targeted proteomics experiments. *Mol. Cell. Proteomics* **14**, 2331–2340 (2015).
17. Egertson, J. D., MacLean, B., Johnson, R., Xuan, Y. & MacCoss, M. J. Multiplexed peptide analysis using data-independent acquisition and Skyline. *Nat. Protoc.* **10**, 887–903 (2015).
18. Amodei, D. et al. Improving precursor selectivity in data-independent acquisition using overlapping windows. *J. Am. Soc. Mass Spectrom.* **30**, 669–684 (2019).
19. Röst, H. L. Deep learning adds an extra dimension to peptide fragmentation. *Nat. Methods* **16**, 469–470 (2019).
20. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol. Cell. Proteomics* **11**, 1475–1488 (2012).
21. Zolg, D. P. et al. PROCAL: a set of 40 peptide standards for retention time indexing, column performance monitoring, and collision energy calibration. *Proteomics* **17**, <https://doi.org/10.1002/pmic.201700263> (2017).
22. Rosenberger, G. et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1**, 140031 (2014).
23. Zolg, D. P. et al. Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **14**, 259–262 (2017).
24. Wang, M. et al. Assembling the community-scale discoverable human proteome. *Cell Syst.* **7**, 412–421.e5 (2018).
25. Liu, Y. et al. Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat. Biotechnol.* **37**, 314–322 (2019).
26. Tate, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
27. Robin, T., Bairoch, A., Müller, M., Lisacek, F. & Lane, L. Large-scale reanalysis of publicly available HeLa cell proteomics data in the context of the Human Proteome Project. *J. Proteome Res.* **17**, 4160–4170 (2018).
28. Garboczi, D. N. et al. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* **384**, 134–141 (1996).
29. Rosenberger, G. et al. Inference and quantification of peptidofoms in large sample cohorts by SWATH-MS. *Nat. Biotechnol.* **35**, 781–788 (2017).
30. Searle, B. C., Lawrence, R. T., MacCoss, M. J. & Villén, J. Thesaurus: quantifying phosphopeptide positional isomers. *Nat. Methods* **16**, 703–706 (2019).
31. Lasonder, E. et al. Integrated transcriptomic and proteomic analyses of *P. falciparum* gametocytes: molecular insight into sex-specific processes and translational repression. *Nucleic Acids Res.* **44**, 6087–6101 (2016).
32. Pino, L. K. et al. Matrix-matched calibration curves for assessing analytical figures of merit in quantitative proteomics. *J. Proteome Res.* **19**, 1147–1153 (2020).
33. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
34. Aurrecochea, C. et al. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* **37**, D539–D543 (2009).
35. Bansal, A., Molina-Cruz, A., Brzostowski, J., Mu, J. & Miller, L. H. Plasmodium falciparum calcium-dependent protein kinase 2 is critical for male gametocyte exflagellation but not essential for asexual proliferation. *MBio* **8**, e01656-17 (2017).
36. Nepomuceno, A. I. et al. Dual electrospray ionization source for confident generation of accurate mass tags using liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **75**, 3411–3418 (2003).
37. Pasa-Tolić, L., Masselon, C., Barry, R. C., Shen, Y. & Smith, R. D. Proteomic analyses using an accurate mass and time tag strategy. *Biotechniques* **37**, 621–624, 626–633, 636 passim (2004).
38. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* **15**, 440–448 (2018).
39. Lim, M. Y., Paulo, J. A. & Gygi, S. P. Evaluating false transfer rates from the match-between-runs algorithm with a two-proteome model. *J. Proteome Res.* **18**, 4020–4026 (2019).
40. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
41. Wessel, D. & Flügge, U. I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **138**, 141–143 (1984).
42. Gardner, M. J. et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
43. Orsburn, B. Normalized collision energy calculation for Q Exactive. <http://proteomicsnews.blogspot.com/2014/06/normalized-collision-energy-calculation.html> (2014).
44. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
45. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
46. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
47. Deutsch, E. W. et al. Trans-proteomic pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteom. Clin. Appl.* **9**, 745–754 (2015).
48. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S. & MacCoss, M. J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **78**, 5678–5684 (2006).
49. Pino, L. K. et al. The Skyline ecosystem: informatics for quantitative mass spectrometry proteomics. *Mass Spectrom. Rev.* <https://doi.org/10.1002/mas.21540> (2017).
50. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
51. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
52. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. Fast and Accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **27**, 1719–1727 (2016).
53. Savitzky, A. & Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627–1639 (1964).

Acknowledgements

We thank S. Kappe for insightful discussions, N. Carnago for providing the gametocyte samples, L. Pino for providing yeast lysates, B. Kim for technical assistance, and J. Haskin for editorial assistance. We also thank M. MacCoss, K. Grove, and M. Guldbbrandt for providing instrument time. B.C.S. is supported by the Translational Research Fellows Program (TRFP) from the Institute for Systems Biology. K.E.S. is supported by K25AI119229. This work was supported by NIH grant R01GM133981, the German Federal Ministry of Education and Research (BMBF, grant no. 031L0008A and no. 031L0168) and the EU Horizon 2020 grant EPIC-XS (grant no. 823839).

Author contributions

B.C.S. conceived the study. B.C.S., K.E.S., and C.A.B. performed the experiments. B.C.S., T.S., and S.G. developed the software. B.C.S., B.K., and M.W. supervised the work. All authors wrote and approved the manuscript.

Competing interests

The authors declare the following competing interests: B.C.S. is a founder and shareholder in Proteome Software, which operates in the field of proteomics. M.W. and B.K. are founders and shareholders of OmicScouts GmbH and msAId GmbH. T.S. and S.G. are founders and shareholders of msAId GmbH. OmicScouts and msAId operate in the field of proteomics. The other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-15346-1>.

Correspondence and requests for materials should be addressed to B.C.S.

Peer review information *Nature Communications* thanks Birgit Schilling for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020