SCIENTIFIC DATA

Check for updates

OPEN

DATA DESCRIPTOR

A comprehensive CHO SWATH-MS spectral library for robust quantitative profiling of 10,000 proteins

Kae Hwan Sim[1], Lillian Chia-Yi Liu[1], Hwee Tong Tan[1], Kelly Tan[1], Daniel Ng[1], Wei Zhang[1], Yuansheng Yang[1], Stephen Tate[2] & Xuezhi Bi[1,3]✉

Sequential window acquisition of all theoretical fragment-ion spectra (SWATH) is a data-independent acquisition (DIA) strategy that requires a specific spectral library to generate unbiased and consistent quantitative data matrices of all peptides. SWATH-MS is a promising approach for in-depth proteomic profiling of Chinese hamster Ovary (CHO) cell lines, improving mechanistic understanding of process optimization, and real-time monitoring of process parameters in biologics R&D and manufacturing. However, no spectral library for CHO cells is publicly available. Here we present a comprehensive CHO global spectral library to measure the abundance of more than 10,000 proteins consisting of 199,102 identified peptides from a CHO-K1 cell proteome. The robustness, accuracy and consistency of the spectral library were validated for high confidence in protein identification and reproducible quantification in different CHO-derived cell lines, instrumental setups and downstream processing samples. The availability of a comprehensive SWATH CHO global spectral library will facilitate detailed characterization of upstream and downstream processes, as well as quality by design (QbD) in biomanufacturing. The data have been deposited to ProteomeXchange (PXD016047).

Background & Summary

Chinese hamster ovary (CHO) cells are widely studied in biomedical research and have been used for the production of nearly 70% recombinant therapeutic proteins, including the blockbuster monoclonal antibodies (mAbs) such as adalimumab (Humira), bevacizumab (Avastin), and trastuzumab (Herceptin)[1–4]. As with other biopharmaceuticals, the production of recombinant mAbs using CHO cells is strictly regulated by the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA)[5]. The quality by design (QbD) approach is emphasized in order to ensure all aspects of CHO mAb product development and manufacturing are evaluated and consistent[6–9]. Hence, in order to enhance the mAb product quality while minimizing potential adverse effects, a thorough understanding of the recombinant mAbs, the early-stage product development process as well as the production and purification process is critical.

Mass spectrometry (MS)-based proteomics techniques have been applied to facilitate the QbD approach for optimal bioprocess design, and enhanced quality and yield of biotherapeutic products[10–13]. For example, the data-dependent acquisition (DDA) shotgun proteomics has been demonstrated to complement the traditional immunoassays (western blotting and ELISA) to monitor and analyze known and/or previously unseen host cell proteins (HCPs) in the drug manufacturing processes[14–16]. Nevertheless, due to the stochastic nature of peptide sampling by DDA-MS technique, multi-dimensional separation or fractionation steps and lengthy chromatographic gradients are often required to increase the proteome coverage over a large dynamic range of protein concentrations. Additionally, the advanced MS methodologies usually require well-trained professionals to operate the instrument and analyze the MS data in a consistent manner. These factors limit the throughput and militate the viability of MS-based techniques as a routine method for bioprocess development and optimization.

[1]Bioprocessing Technology Institute (BTI), Agency for Science, Technology and Research (A*STAR), Singapore, 138668, Singapore. [2]SCIEX, Concord, Ontario, Canada. [3]Duke-NUS Medical School, Singapore, 169857, Singapore. ✉e-mail: bi_xuezhi@bti.a-star.edu.sg

Sequential window acquisition of all theoretical fragment-ion spectra (SWATH), a specific variant of data-independent acquisition (DIA)-based MS technique introduced by SCIEX, provides the possibility to overcome the limitations imposed by traditional DDA-based MS techniques[17–19]. In SWATH-MS, all ionized compounds of a given sample within a specified mass range (window) are fragmented in a systematic and unbiased fashion through a sequential series of predefined precursor isolation windows. The complete information of precursors and product ions will be permanently recorded in a single scan, offering highly specific and multiplexed MS data to quantitate the analytes in a manner equivalent to selected reaction monitoring (SRM). Furthermore, SWATH-MS coupled with peptide-centric data extraction strategy enables quantification of multiple analytes, and this has recently been demonstrated in bioprocessing research[20–23]. Importantly, a specific spectral library containing the empirically accurate SWATH assay coordinates is preferred over the computational prediction methods (library-free analysis) to specifically extract high-quality quantitative measurements from the SWATH-MS data[23–27]. However, no CHO spectral library has been publicly available thus far; the potential commercial value and competing interest may have prevented the release of the in-house spectral libraries generated by the biopharmaceutical companies and contract research organizations. With the aim to facilitate and improve the consistency in bioprocess development as well as to benefit the academic research, we believe that the construction of a publicly accessible CHO-specific SWATH-MS spectral library is significant for future SWATH-MS applications in biopharma R&D and biomanufacturing.

Here, we present a comprehensive SWATH-MS CHO global spectral library generated through a series of systematic analyses of mAb-producing CHO-K1 intracellular proteome using advanced LC-MS technique coupled with multi-dimensional fractionation strategies. More than 10,000 proteins were identified in the library, and stringent filtering criteria were implemented in the subsequent fine-tuning steps to include only the assays which are reliably detected. The robustness, accuracy and consistency of the CHO global spectral library were demonstrated by SWATH-MS analyses of different samples obtained from CHO whole cell lysates (WCL), harvested cell culture fluids (HCCF) and downstream processing (DSP) mAb samples, and across multiple instrumental setups. These results have also shown the feasibility of SWATH-MS with the in-house CHO global spectral library as a potential process analytical technology (PAT) to resolve the bioprocessing issues within the QbD paradigm for future biopharmaceuticals manufacturing system. The original MS data and the CHO spectral library demonstrated in current study are deposited to the ProteomeXchange (PXD016047)[28].

Methods

Reagents. Reagents used in the experiments were purchased from Sigma-Aldrich (St Louis, MO, US) unless stated otherwise.

CHO cell cultures. The mAb-producing CHO-K1 (CCL-61, ATCC) cells were cultured in fully chemically-defined protein-free culture medium as described previously[29]. The cells were subcultured every 3 days in 125 mL disposable Erlenmeyer shake flasks (Corning, Acton, MA), and incubated on a shaker platform at 110 rpm in a humidified 37 °C / 8% CO_2 incubator. Viable cell density and viability were measured by Vi-CELL™ XR 2.04 cell viability analyzer (Beckman Coulter, Brea, California) according to the manufacturer's instructions. Harvested cell culture fluid (HCCF, collected on day 14 of CHO-K1 fed-batch cell cultures), CHO-DG44 and CHO-S cell cultures were obtained from the Animal Cell Technology group in Bioprocessing Technology Institute (BTI), A*STAR, Singapore.

Protein sample preparation and enzymatic digestion. CHO cells were pelleted at 1,500 *g*, washed thrice with ice-cold phosphate buffered saline (PBS) and re-suspended in a SDS cell lysis buffer containing 50 mM triethylammonium bicarbonate buffer (TEAB) pH 8.5, 5% SDS, and 1x Halt™ protease inhibitor cocktail (Thermo Fisher Scientific, Waltham, MA). The cell lysates were further disrupted using an UP50H ultrasonic processor (Hielscher Inc. Teltow, Germany) at 30% amplitude, with 0.5 s pulse on and 0.5 s pulse off, for 20 times on ice, and clarified by centrifugation at 20,000 *g* for 10 min. Protein concentration was determined using the BCA Protein Assay Kit (Pierce™, Thermo Fisher Scientific) according to the manufacturer's instructions. The DSP mAb samples were processed using a typical DSP purification procedures (Fig. 1): starting with clarified HCCF original material (OM), mAb was captured with protein A affinity chromatography using MabSelect SuRe LX resin (GE Healthcare, Uppsala, Sweden), followed by cation exchange chromatography for intermediate purification using POROS XS resin (Thermo Fisher Scientific, Waltham, MA) with salt gradient elution, and anion exchange chromatography for polishing using POROS HQ resin (Thermo Fisher Scientific, Waltham, MA) in a flow-through mode.

mAb concentrations were measured by analytical SEC with a TSKgel G3000SWXL column (Tosoh Bioscience, South San Francisco, CA) on a Dionex UltiMate™ 3000 HPLC system (Thermo Fisher Scientific, Waltham, MA) operated at a flow rate of 0.6 mL/min, using a buffer with the formulation of 50 mM MES, 20 mM EDTA, 200 mM arginine, pH 6.5. The sample injection volume was 100 µL. mAb IgG concentrations were calculated by comparing the experimental results with a calibration curve prepared from the known concentrations of purified mAb, determined by SoloVPE (C Technologies, Inc. Bridgewater Township, NJ). HCP content in DSP samples was determined by ELISA using a Generation III CHO HCP kit (Cygnus Technologies, Southport, NC) according to the manufacturer's instructions.

HCCF and DSP mAb samples obtained from CHO-K1 were concentrated using 10,000 MWCO Vivaspin® 20 centrifugal concentrators (#VS2002), (Sartorius, Göttingen, Germany), and the proteins were precipitated using methanol-chloroform precipitation method as described previously[30].

Enzymatic digestion: 200 µg protein of each sample was reduced and alkylated, followed by digestion on S-Trap™ mini spin columns (ProtiFi™, Farmingdale, NY) using Trypsin Gold, MS-grade (Promega, Madison,

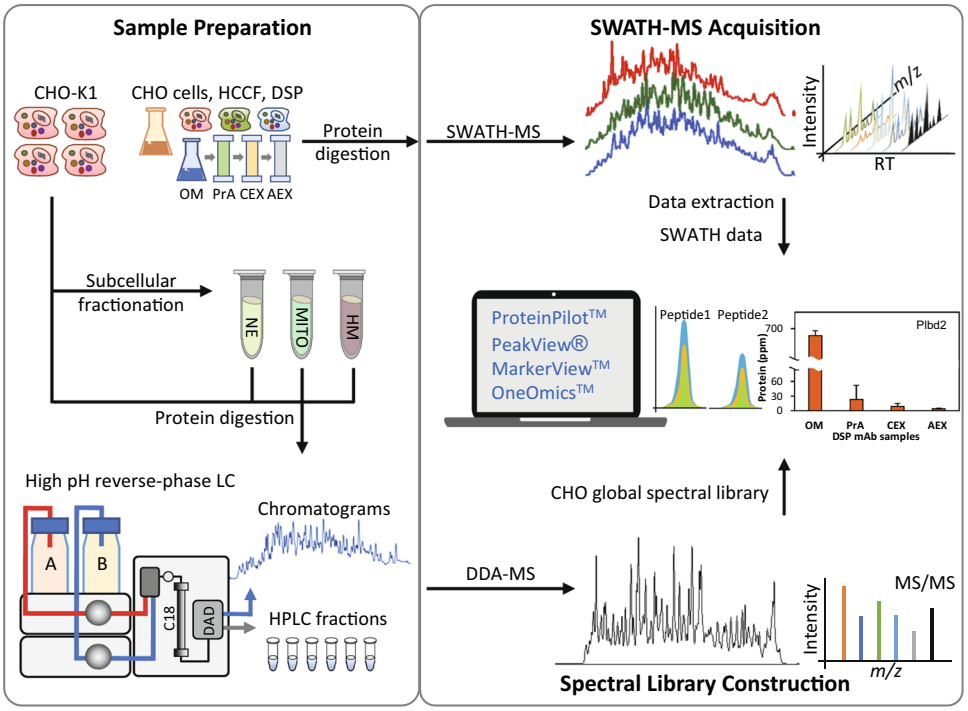

Fig. 1 Workflow for creating and using the SWATH CHO global spectral library. The CHO-derived samples were processed using in-house multi-dimensional separation protocol. Briefly, the CHO-K1 cells were lysed and fractionated using differential ultracentrifugation to isolate nuclear (NE), mitochondrial (MITO), and heavy-membrane (HM) compartments. The protein lysates from whole cell (WCL) and subcellular-organelle compartments were tryptic digested, subsequently fractionated using basic reverse-phase liquid chromatography separation, and subjected to DDA-MS analysis. Protein digest from harvested cell culture fluid (HCCF) and downstream processing (DSP) mAb samples were directly subject to SWATH-MS in TripleTOF 6600. The raw DDA data was searched locally in ProteinPilot™ software and the results were uploaded to OneOmics™ for spectral library construction. The SWATH-MS data sets were processed locally using PeakView® and MarkerView™ or using OneOmics™. The applicability and robustness of the CHO global spectral library were evaluated with SWATH-MS data sets of different CHO-derived samples, including WCL of different cell lines, HCCF and DSP mAb samples, and using various LC-MS instrumental setups.

Wisconsin) according to the manufacturer's protocol[31]. The eluted peptide mixtures were dried in a SpeedVac vacuum concentrator at room temperature and stored at −80 °C for future use.

Subcellular organelle fractionation of CHO-K1. CHO-K1 cell pellets were resuspended in cell lysis buffer containing 250 mM sucrose, 20 mM HEPES-NaOH (pH 7.9), 10 mM KCl, 1.5 mM $MgCl_2$, 1 mM EDTA, 1 mM EGTA, and 1x Halt™ protease inhibitor cocktail. The cells were incubated on ice for 5 min with occasional vortex followed by passing through 27-gauge needle for lysate homogenization. The cell lysate was centrifuged at 800 *g* for 10 min at 4 °C to pellet down the nuclear fraction, cell debris and unbroken cells. The supernatant was collected and centrifuged at 10,000 *g* for 20 min at 4 °C to pellet the mitochondrial fraction. The supernatant was further centrifuged at 100,000 *g* for 60 min at 4 °C using a micro-ultracentrifuge (#CS150FNX) (Hitachi, Tokyo, Japan). The pellet obtained was heavy-membrane fraction (mainly rough endoplasmic reticulum) while the supernatant was cytosolic fraction. The nuclear, mitochondrial and heavy-membrane fractions were tryptic digested using the S-Trap aided protein digestion protocol.

High pH reverse phase liquid chromatography (bRPLC) peptide fractionation. A total of 200 µg of peptide samples from CHO-K1 WCL and subcellular-organelle fractions were each subjected to a 60-min multi-step gradient performed on a Kinetex core shell C18 column (2.6 µm, EVO C18, 150 mm × 3.0 mm, 100 Å, Phenomenex, Brechbuhler AG, Schlieren, Switzerland) using an UltiMate 3000 UHPLC system (Thermo Fisher Scientific). The mobile phase was 10 mM ammonium formate (buffer A, adjusted to pH 10 using ammonium hydroxide) and 10 mM ammonium formate in 80% acetonitrile (ACN) (buffer B, adjusted to pH 10 using ammonium hydroxide). The peptide samples were separated with a gradient from 0–35% buffer B for 55 min at a flow rate of 0.8 mL/min. The fractions were collected and pooled in concatenation into 30 fractions for CHO-K1 WCL; and 10 fractions for each subcellular-organelle sample. The fractions were dried in a SpeedVac vacuum concentrator at room temperature and stored at −80 °C for future use.

Data-dependent acquisition (DDA) – MS. For spectral library generation, estimated 1 µg of peptide mixtures from each fraction were analyzed on a TripleTOF 6600 mass spectrometer (SCIEX) coupled to a

NanoSpray III Ion Source (SCIEX) and interfaced with a Waters nanoAcquity UPLC system (Waters, Milford, MA) or Eksigent Ekspert nanoLC 425 (NanoLC Ultra 2D, Eksigent, Toronto, Canada). The peptide samples were separated at 50 °C on an ACQUITY UPLC M-class peptide BEH C18 column (75 µm x 200 mm, particle size of 1.7 µm, and pore size of 130 Å). The iRT reference peptides (Biognosys AG, Schlieren, Switzerland) were added to all samples at 1:10 ratio prior to MS injection for retention time calibration. The LC system was operated with 0.1% formic acid (FA) in water (buffer A) and 0.1% FA in ACN (buffer B) at a flow rate of 300 nL/min. The separation gradient was from 5–35% of buffer B over the period of 110 min. The MS was operated in DDA top 20 mode with the following parameters: MS1 spectra were collected at 400–1,500 *m/z* for 500 ms, 20 most intense precursors with charge states 2–5 that exceeded 125 counts/s were selected for fragmentation, and the corresponding fragmentation MS2 spectra were collected at 50–2,000 *m/z* for 100 ms. Rolling collision energy (equation: (0).0625 × *m/z* – 10.5) (derived from SCIEX) with a collision energy spread (CES) of 15 eV was set as the fragmentation patterns used in SWATH-MS analysis.

Data-independent acquisition (DIA) SWATH-MS. SWATH-MS acquisition was operated using the same LC-MS instrumental setup as described above with some modifications. Briefly, a 100-variable-window setup was generated using the SWATH® Variable Window Calculator 1.1 (SCIEX) with a 1 *m/z* window overlap on the lower side of the window. The MS1 survey scan was acquired from 400–1,250 *m/z* for 250 ms and MS2 spectra were acquired in high-sensitivity mode from 100–1,500 *m/z* for 30 ms. The total cycle time was ~3.3 s. The collision energy used in SWATH-MS acquisition was that applied to a doubly charged precursor centered in the middle of the isolation window calculated with the same collision energy equation for DDA, and with a CES of 5 eV. For the analyses conducted in capillary flow and microflow rate, the SWATH-MS data were recorded on a TripleTOF 6600 mass spectrometer coupled to a DuoSpray Analytical Ion Source (SCIEX). For capillary flow LC setup, the Eksigent Ekspert nanoLC 425 system was connected to a Waters CSH C18 column (300 µm × 150 mm, 1.7 µm, 130 Å) or a Eksigent ChromXP C18 column (300 µm × 150 mm, 3 µm, 120 Å) and operated in trap-elute mode for 1-h gradient at a flow rate of 5 µL/min. For the microflow LC setup, a Waters nanoAcquity LC system connected with a Waters CSH C18 column (1 mm × 150 mm, 1.7 µm, 130 Å) was utilized and the system was operated in direct-inject mode at a flow rate of 50 µL/min. The 100-variable-windows setup was applied and optimized in both the capillary flow and microflow rate instrumental system.

Generation of CHO spectral library. The CHO spectral library was constructed using the workflow established by SCIEX[32]. Briefly, the raw DDA data files were processed using the ProteinPilot™ software (version 5.0.1) against a Chinese hamster (CH) protein sequence database. The database was the latest release of CH RefSeq assembly (downloaded on August 2019 from ncbi.nlm.nih.gov; GCF_003668045.1_CriGri-PICR_protein.faa), appended with the Biognosys iRT fusion protein sequence and in-house mAb heavy- and light-chain protein sequences. The alkylation reagent was iodoacetamide. The search effort setting was Rapid ID with carbamidomethyl (C) as a fixed modification, and oxidation (M), deamidation (NQ) and pyroglutamic acid conversion (EQ) as variable modifications. Each of the DDA raw data files was searched in ProteinPilot™ and the result files were used as the input libraries for spectral library generation. The input libraries and SWATH-MS acquisition data were uploaded to the Illumina BaseSpace Cloud (www.basespace.illumina.com) using the CloudConnect software and processed in SCIEX Cloud OneOmics™ (Fig. 1). For the construction of spectral library, each input library was filtered by 1% FDR and 99% confidence threshold to remove low-confidence peptide identifications. The largest filtered input library, which contained the highest number of high-confidence peptides, were selected as base, and peptides identified from smaller filtered input libraries were merged to the base library using a non-linear calibration strategy in the OneOmics™ [32]. Any newly identified peptides are added to the existing proteins and any newly detected proteins are added if not present in the base library. The merged spectral library was further processed with an in-house script to identify and remove any shared and/or modified peptides.

The combined spectral library was constructed by searching all the 63 DDA raw data files together in the ProteinPilot™ software. The 1D spectral library was generated using unfractionated peptide mixtures from a CHO-K1 WCL sample.

SWATH-MS data analysis. The SWATH-MS data analysis was performed in the OneOmics™ or the SWATH™ Processing software in local PC. In the OneOmics™, the CHO endogenous peptides were extracted according to the precursor *m/z*, intensity and confidence of identification across the entire time range, and the best scoring peak groups were used for RT calibration. The data was filtered by 1% FDR and the comparisons of proteins and/or peptides were further filtered by 20% coefficient of variation (CV) cutoff between replicates. In local SWATH-MS processing workflow, the spectral library and SWATH-MS acquisition data were loaded into SWATH™ Processing microApp in PeakView® software. The iRT, mAb and shortlisted endogenous CHO peptides were manually selected as reference peptides for RT calibration. The peak groups were extracted with a 99% peptide confidence threshold and a 1% peptide FDR cutoff. The RT extraction window and the fragment ion mass tolerance were set to 5 min and 75 ppm, respectively. After data extraction, the results were imported into MarkerView™ (version 1.3.1) for further data processing and analysis. Microsoft® Excel, Python programming and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis[33] were also applied in the downstream data browsing and analysis.

Data Records

The raw DDA data files for library generation, the search result (group files), the CHO spectral library, and the SWATH-MS acquisition data applied in the current study have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org)[34] *via* the PRIDE partner repository[35,36] with the data set identifier PXD016047[28].

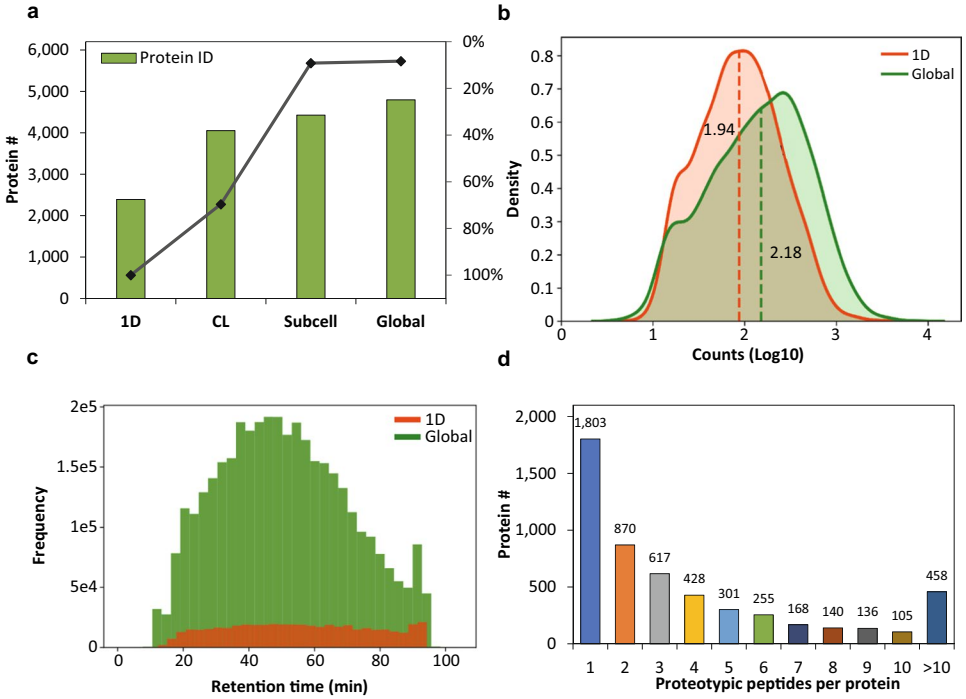

Fig. 2 Characteristics of the SWATH CHO global spectral library. (**a**) The stacked bar chart showed the numbers of confidently quantified protein ID (at 20% CV cutoff and 1% peptide FDR) when using 1D library (1D, 3 fractions), cell-lysate library (CL, 33 fractions), subcellular-organelle library (Subcell, 33 fractions), and CHO global spectral library (Global, 63 fractions). The percentage of increase was calculated by dividing the number of increased protein ID by the total number of protein ID. (**b**) The distributions of the assays per protein in the CHO global spectral library and 1D library. The dash lines indicated the median values of the assays per protein. (**c**) The histogram distributions of the coverage and the number of peptides eluting across the LC separation gradient for the CHO global spectral library and 1D library. (**d**) The bar chart showed the confidently quantified proteotypic peptides per protein (at 20% CV cutoff and 1% peptide FDR) in the CHO-K1 WCL SWATH-MS data set using the CHO global spectral library.

Technical Validation

Error rate control during spectral library generation. The number of proteins and peptides identified from the raw DDA data files are significantly affected by the quality and redundancy of the protein sequence database used. UniProtKB/Swiss-Prot[37], a freely accessible database of protein sequence and functional information, is manually annotated and reviewed to contain high-quality and non-redundant protein information and hence highly recommended for the purpose of public spectral library generation. However, the UniProtKB/Swiss-Prot contains only 14 and 237 reviewed entries for Chinese hamster/CHO-K1 and *Cricetulus griseus*, respectively. We therefore used the most updated NCBI reference sequence (RefSeq) proteome database of Chinese hamster in this study. NCBI RefSeq[38] is also a curated non-redundant collection of sequences which records genomes, transcripts and proteins information from multiple sources. We searched the mass spectra from the DDA raw data against the Chinese hamster proteome database containing 46,750 protein entries. The error rate in the analysis of these large sample cohorts was carefully handled to prevent the accumulation of false positive identifications during the construction of the spectral library. We utilized the library merging algorithm established in the OneOmics™ to filter the lists of peptides in ProteinPilot™ group files to a 1% global protein FDR[32]. All the peptides that passed the 99% peptide confidence threshold were selected and merged to the base library using a non-linear retention time calibration strategy[32]. Any shared or modified peptides (except those with carbamidomethylation on cysteine residue) were identified and removed from the spectral library using an in-house Python script to avoid any possible redundancy and inaccuracy in SWATH-MS data extraction and quantification using the spectral library. A total of 10,974 proteins, comprised of 199,102 peptides, 247,135 precursors and 741,405 transitions, were represented in the finalized library, which was known as CHO global spectral library hereafter.

Coverage of the CHO global spectral library. During the construction of CHO global spectral library, it was found that the number of confidently quantified proteins (at 20% CV cutoff and 1% peptide FDR) reached plateau despite introducing more DDA files into the spectral library (Fig. 2a). As shown in Fig. 2a, the Subcell library (contained 30 fractions of bRPLC-separated subcellular compartments and 1D library) quantitated 9.18% more proteins than the CL library (contained 30 fractions of bRPLC-separated cell lysate and 1D library). By using the CHO global spectral library (consisted of the fractions from Subcell library, CL library and 1D library), we quantitated slightly more proteins (8.34%) than the previous work (Fig. 2a). It is noteworthy that a large number of library inputs will critically increase the time required for loading libraries into OneOmics™. Therefore,

	DDA files #	Protein #[a]	Protein #[a] (≥2 peptides)	Peptide #[a,b]	Confidence cutoff	Data processing pipeline
CHO global spectral library	63	10,974	9,549	199,102[b]	99% peptides confidence & 1% protein FDR	ProteinPilot & OneOmics
CHO-K1 1D library	3	3,134	2,810	20,249[b]	99% peptides confidence & 1% protein FDR	ProteinPilot & OneOmics
CHO-DG44 exponential[13]	48	N.A	5,950	58,128	1% FDR	Proteome Discoverer
CHO-DG44 stationary[13]	48	N.A	5,593	50,194	1% FDR	Proteome Discoverer
CHO-S exponential[13]	48	N.A	6,089	53,958	1% FDR	Proteome Discoverer
CHO-S stationary[13]	48	N.A	5,538	48,205	1% FDR	Proteome Discoverer
Combined CHO-DG44 and CHO-S[13]	192	N.A	9,359	N.A	1% FDR	Proteome Discoverer

Table 1. Comparison of the proteome coverage between CHO global spectral library, 1D library and the latest proteomic studies on CHO-DG44 and CHO-S cells. [a]The iRT, reverse sequences, and the proteins containing only non-CAM modified peptides were excluded. [b]The modified peptides except for carbamidomethylation on cysteine residue were excluded.

we did not further fractionate the CHO-K1 protein lysate as these DDA inputs have provided sufficient CHO proteome coverage while allowing reasonable downstream SWATH-MS data processing time. In summary, the CHO global spectral library was constructed by using the MS/MS spectral data obtained from the unfractionated (1D) CHO-K1 cell lysates as the base, followed by adding the MS/MS spectral data sets derived from the three subcellular organelles (10 fractions of nuclear compartment, 10 fractions of mitochondrial compartment and 10 fractions of heavy-membrane compartments) and an extensively separated CHO-K1 whole-cell protein lysate (30 fractions).

The CHO global spectral library has reached a 23.48% proteome coverage of the 46,750 protein entries represented in the CHO RefSeq proteome database. This is 2.5 times more proteins and 8.8 times more peptides than a non-fractionated 1D spectral library (comprising of 3,134 proteins and 20,249 peptides identification) (Table 1). We also compared the CHO global spectral library (which was built based on the CHO-K1 cell proteome) to the latest proteomics study of CHO-DG44 and CHO-S cell lines (Table 1)[13]. When filtered with at least two unique peptides, the CHO global spectral library identified significantly more proteins (+56.82% to +72.42%) than individual CHO-DG44 and CHO-S data set, respectively, and a slightly higher number of proteins (+2.03%) than the combination of these two data sets. Remarkably, the CHO global spectral library represented a 242.52% to 313.03% improvement over the previous work in peptide identification (Table 1). The CHO global spectral library is thus able to provide a significantly higher number of empirically detected peptide identities for better protein identification and quantification. As displayed in Fig. 2b, the CHO global spectral library showed high similarity in the distribution of the number of assays per protein as that of 1D library, but the former covered a wider effective separation gradient with its large number of peptide identities (Fig. 2c).

A CHO-K1 WCL SWATH-MS data set was processed with the CHO global spectra library. Up to 3,478 proteins (65.86%) could be confidently quantified (at 20% CV cutoff and 1% peptide FDR) with at least two unique peptides (Fig. 2d). Additionally, we perceived the paramount significance to quantitatively identify the critical components in CHO biological pathways relating to the stability, quality and productivity of recombinant protein production as well as the cell viability. By using KEGG pathway mapping analysis, we demonstrated that proteins associated with Protein Processing in ER (83.81%, Fig. 3a), Cell Cycle (82.98%, Fig. 3b) and N-glycan Biosynthesis (82.05%, data not shown) were well represented in the CHO global spectral library (proteins boxed in yellow color in Fig. 3). In addition, more than half of these proteins could be confidently identified (at 20% CV cutoff and 1% peptide FDR) in the SWATH-MS analysis (proteins highlighted in red color in Fig. 3).

Calibrated retention time in CHO global spectral library. One of the advantages of adopting nanoflow instrumental setup in the construction of SWATH spectral library was that nanoflow LC-MS provides the highest sensitivity in DDA-MS without requiring huge amounts of starting materials as compared to higher flow LC-MS. However, retention time drift[39] is an inherent issue in the nanoflow instrumental setup which would lead to retention time misalignment in combined database search (the first step in constructing spectral library). Therefore, we utilized the library merging algorithm in OneOmics™ to improve the retention time correlation and alignment between input libraries during the library construction. In subsequent SWATH-MS data extraction procedure in OneOmics™, a non-linear autoRT calibration algorithm was applied[32]. Briefly, one hundred high-abundance, confidently detected peptides, including those of iRT peptides, mAb protein and endogenous CHO proteins, were selected as reference peptides across the entire scan time range. The retention time frame of spectral library was matched to that of SWATH-MS data based on the reference peptides to fine-tune the retention time from run to run. Here we compared the CHO global spectral library and the uncalibrated combined-searched spectral library in the SWATH-MS analysis of CHO-K1 WCL data set. The CHO global spectral library identified a total of 6,445 proteins, which was 49.54% more than that of the combined-searched library (1% peptide FDR) (Table 2). At 20% CV cutoff, the CHO global spectral library quantitated higher numbers of protein (4,717) and peptide (16,919) identities as compared to those of the combined-searched spectral library (2,537 proteins and 8,252 peptides) (Table 2). On top of that, the former has reported lower median CV values in both protein- and peptide-level quantitation (Table 2). These results highlighted the effectiveness of retention time calibration during the library

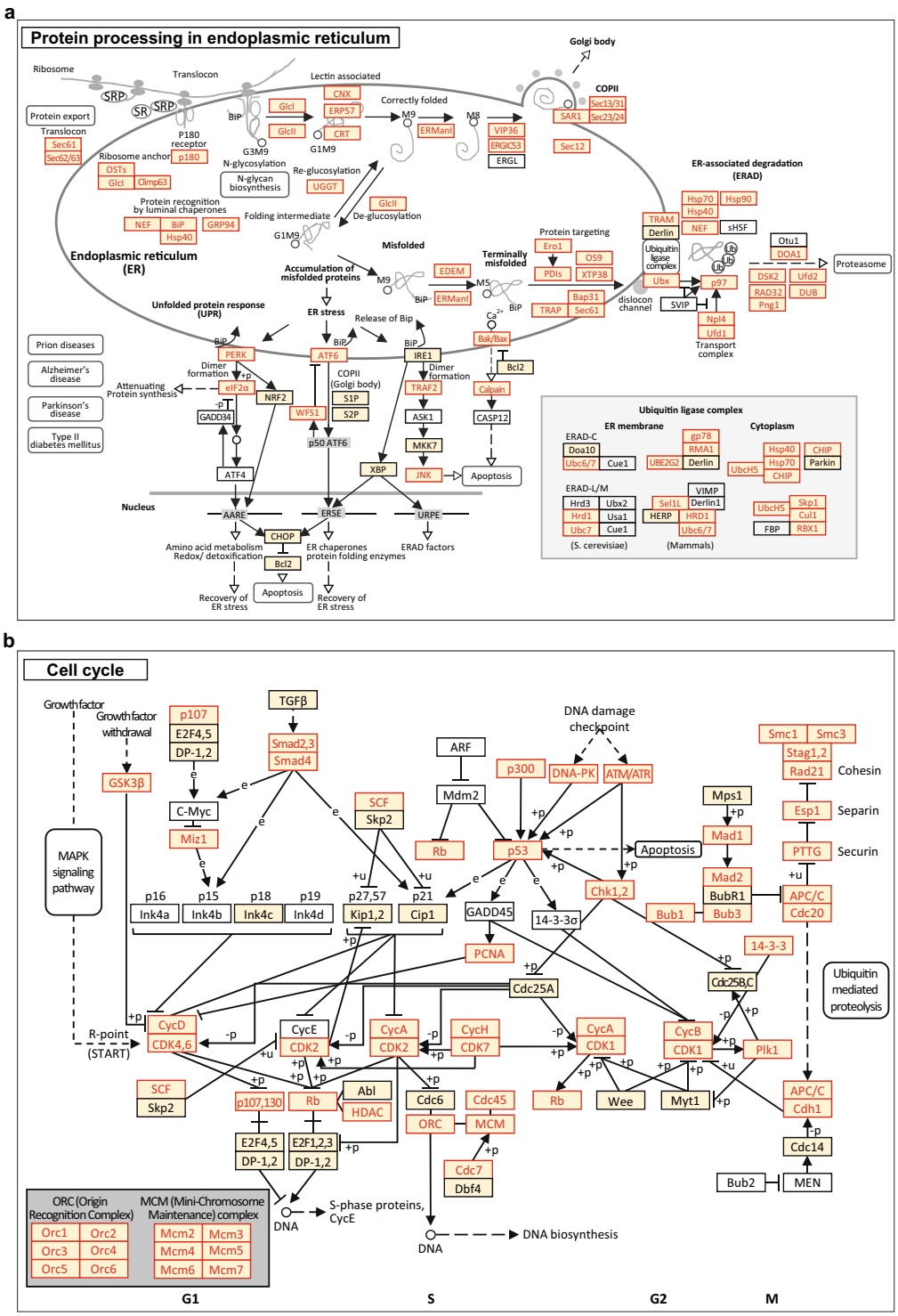

Fig. 3 KEGG pathway mapping of CHO global spectral library. KEGG pathway analysis was applied to study the protein coverage in the (**a**) Protein processing in endoplasmic reticulum and (**b**) Cell cycle pathways. The proteins represented in the CHO global spectral library were boxed in yellow color while those confidently identified in the CHO-K1 SWATH-MS data set were highlighted in red color.

merging and data extraction procedure. By using the CHO global spectral library which contained highly accurate predicted retention times, we could execute correct and consistent SWATH-MS data analysis.

Portability of the CHO global spectral library. Different laboratories may acquire their SWATH-MS data sets using various instrumental setups. For example, in biopharmaceutical industries, higher flow-rate LC-MS analytical instruments are often preferred in order to achieve high reproducibility and throughput with

Spectral library	CHO global spectral library	Uncalibrated combined-search spectral library
Total identified proteins	6,445	4,310
20% CV cutoff	4,717	2,537
10% CV cutoff	3,183	1,414
Median CV	10.16%	15.47%
Total identified peptides	23,971	14,061
20% CV cutoff	16,919	8,252
10% CV cutoff	10,325	4,585
Median CV	11.73%	15.42%

Table 2. Comparison of CHO global spectral library and uncalibrated combined-search spectral library in CHO-K1 WCL SWATH-MS data.

LC-MS Setup	NF6600 (TripleTOF 6600 coupled to nanoflow LC)	CF5600 (TripleTOF 5600 + coupled to capillary flow LC)	MF6600 (TripleTOF 6600 coupled to microflow LC)
MS	SCIEX TripleTOF 6600 (A)	SCIEX TripleTOF 5600+	SCIEX TripleTOF 6600 (B)
Acquisition mode	SWATH-MS	SWATH-MS	SWATH-MS
LC model	Eksigent Ekspert nanoLC 425 (nanoflow module)	Eksigent Ekspert nanoLC 425 (microflow module)	Water nanoAcquity UPLC (microflow)
Injection mode	Trap-elute	Trap-elute	Direct-inject
Column	Waters BEH C18 75 µm × 200 mm	Eksigent ChromXP C18 300 µm × 150 mm	Waters CSH C18 1 mm × 150 mm
Flow rate	300 nL/min	5 µL/min	50 µL/min
LC gradient	100 min	95 min	90 min
Sample amount	1 µg	5 µg	20 µg

Table 3. Details of three LC-MS instrumental setups for the SWATH-MS acquisition of CHO-K1 WCL samples.

a shorter duration of data acquisition and processing. Since the CHO global spectral library was established in nanoflow LC-MS setup (TripleTOF 6600 coupled to Waters nanoAcquity LC), we validated its performance with three CHO-K1 WCL data sets acquired in nanoflow LC (Eksigent) coupled to TripleTOF 6600 (NF6600 data), capillary flow LC (Eksigent) coupled to TripleTOF 5600+ (CF5600 data), and microflow LC (Waters) coupled to TripleTOF 6600 (MF6600 data) (Table 3). Noted that in CF5600 and MF6600 we analyzed higher loading amounts of sample to compensate for the lower sensitivity at higher flow rate LC-MS. The relationship between the delta retention times (ΔRT = observed retention times – predicted retention times) and the predicted retention times were studied. As shown in Fig. 4a, the ΔRT maintained within ±2.5 minutes across the entire predicted retention time range (red colored "x" indicated the eleven iRT reference peptides). Graphical demonstration using violin plot showed the distribution of observed and predicted retention times were highly consistent in the three LC-MS setups (Fig. 4b). Total 4,222 proteins were confidently quantified from three data sets (at 20% CV cutoff and 1% peptide FDR): In NF6600 there were 3,631 proteins (15,311 peptides) quantified when 1 µg of WCL was analyzed; In CF5600 3,519 proteins (12,825 peptides) were quantifiable from 5 µg of WCL; In MF6600 up to 4,087 proteins (17,071 peptides) were quantified from 20 µg of WCL (Fig. 4c). Hundreds to thousands of proteins and peptides could be consistently quantified in three data sets when more stringent CV cutoffs (5 and 10%) were applied (Fig. 4c). Right-skewed distribution of CV values with low median CVs ranging from 9.5% to 12.3% in both the protein- and peptide-level quantification demonstrated the portability as well as the robustness of the CHO global spectral library for the analyses of SWATH-MS data sets acquired across three LC-MS setups (Fig. 4d). These results have further elaborated the effectiveness of retention time calibration in our pipeline, and supported the potential transfer of the CHO global spectral library to industrial scale application which often requires high throughput with short sample and data handling duration.

In addition to CHO-K1 cells, there are other CHO-derived cell lines that are commonly studied and widely utilized in the biomanufacturing process. We evaluated the performance of the CHO global spectral library, which was built using the CHO-K1 cell proteome searched against the Chinese hamster RefSeq database, in the SWATH-MS analysis of another two CHO cell lines, namely CHO-DG44 and CHO-S. In this analysis, more than 4,700 proteins were confidently quantified (at 20% CV cutoff and 1% peptide FDR). A total of 1,664 proteins (35.24%) were commonly identified in the three cell lines and 2,979 proteins (63.09%) were quantified in at least two cell lines (Fig. 4e). The median CV values for the protein abundances in each cell line ranged from 8.72% to 12.28% while the median CV values for all the quantified proteins across the three cell lines was 9.97% (Fig. 4f). The SWATH-MS results of CHO-DG44 and CHO-S cells suggested that the CHO global spectral library can be effectively used to quantify thousands of proteins across multiple CHO cell lines.

Robustness of protein identification and quantification in CHO HCCF and HCPs in DSP mAb samples. Being the most widely utilized host cell in biopharma mAb production, the systematic and

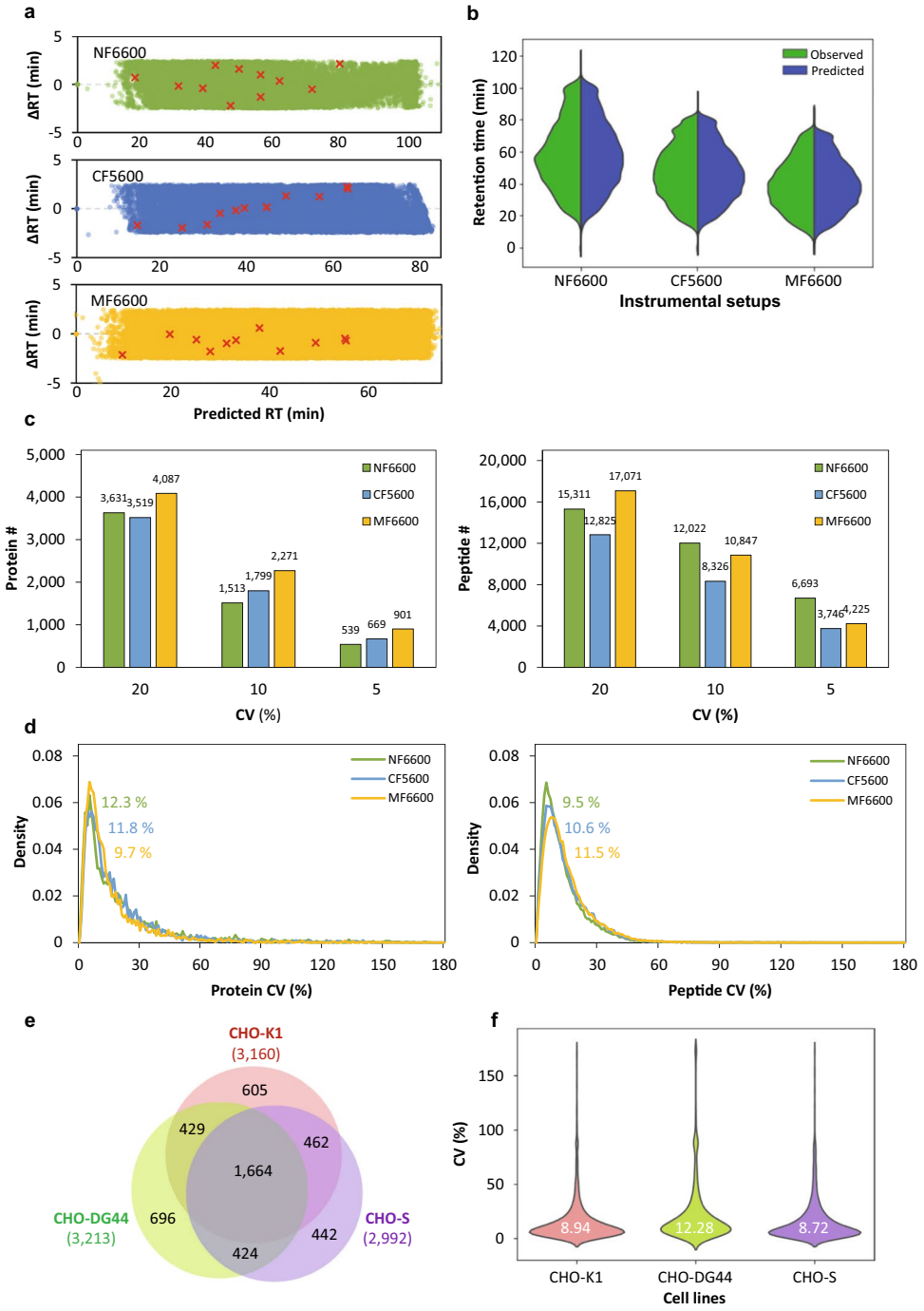

Fig. 4 The robustness of the SWATH CHO global spectral library in quantification of CHO proteome. (**a**) The correlation between the ΔRT and the predicted retention times in three LC-MS instrumental setups. The red colored "×" represented the 11 iRT reference peptides. (**b**) The violin plot showed the distribution of observed and predicted retention times in three instrumental setups. (**c**) The bar charts showed the numbers of confidently quantified proteins (left) and peptides (right) after 20%, 10% and 5% CV cutoff. (**d**) The CV distribution of all the identified proteins (left) and peptides (right) in three LC-MS instrumental setups. The median CV values in each condition were highlighted. (**e**) Venn diagram analysis of confidently quantified protein ID (at 20% CV cutoff and 1% peptide FDR) between three CHO cell lines. The total protein IDs of respective cell lines were indicated in brackets. (**f**) The violin plot showed the distribution of the protein CV values across three CHO cell lines with the median CV values highlighted in white color. NF6600: nanoflow LC coupled to TripleTOF 6600; CF5600: capillary flow LC coupled to TripleTOF 5600 +; MF6600: microflow LC coupled to TripleTOF 6600.

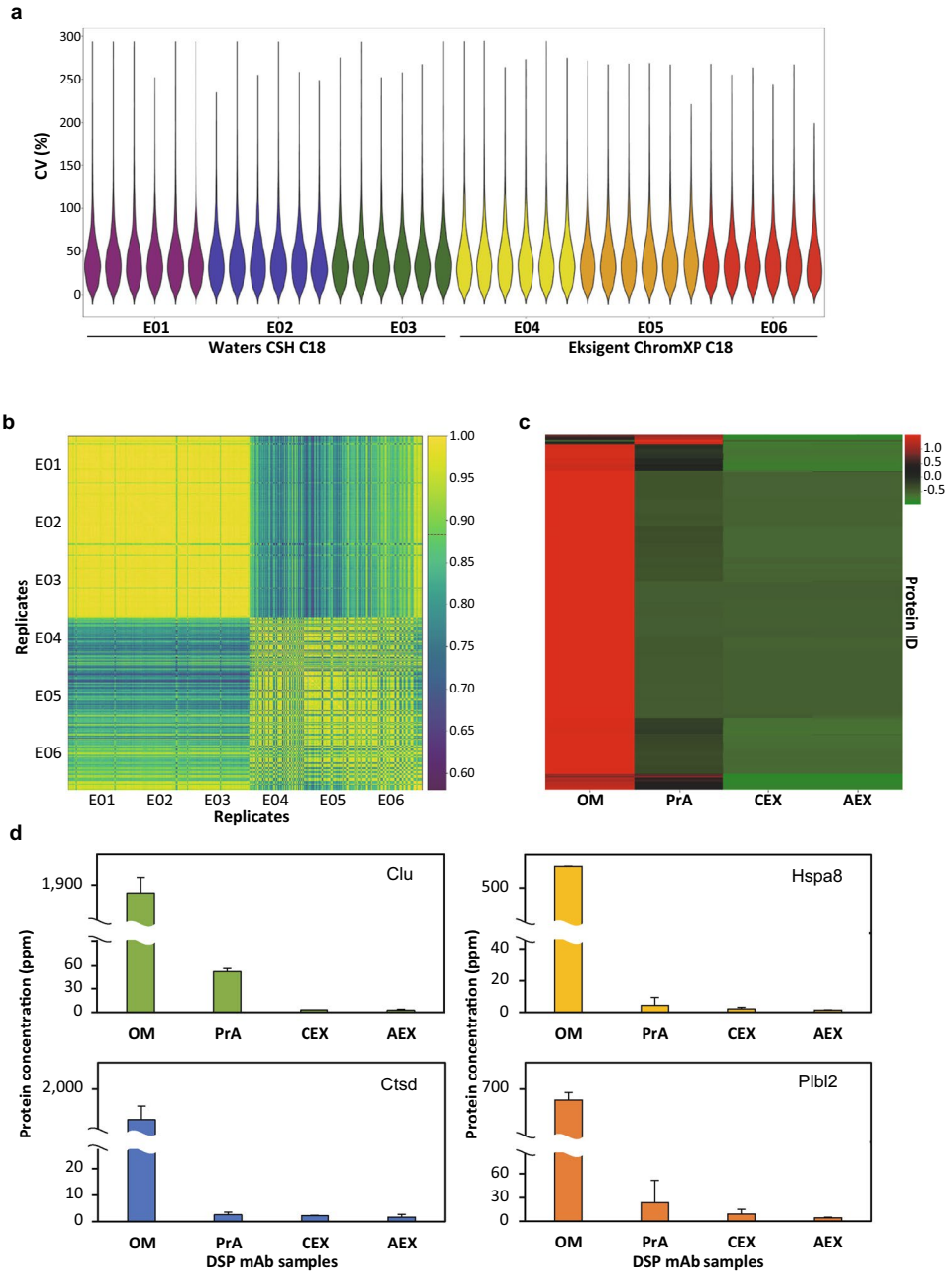

Fig. 5 The application of CHO global spectral library in proteome profiling and quantification of HCCF and DSP mAb samples. (**a**) The violin plot showed the distribution of CV values across the large-scale analysis of HCCF samples. The CV values were calculated using the technical replicates for each of the six biological replicates in E01 – E06. (**b**) The Pearson correlation matrix of the abundance of proteins identified in the HCCF samples from experimental group E01 – E06. The median correlation coefficient r value (0.88) was represented as a red colored dash line in the color bar. (**c**) Heat map analysis of confidently identified proteins (at 20% CV cutoff and 1% peptide FDR in OM) in the typical DSP purification workflow. (**d**) The bar chart showed the average protein abundances of four well-known difficult-to-remove HCPs, including clusterin (Clu), cathepsin D (Ctsd), heat shock cognate 71 kDa protein (Hspa8) and phospholipase B-like 2 (Plbl2), identified in the DSP purification process. The error bars indicated the standard deviations between replicates. OM: original material; PrA: protein A eluate; CEX: cation exchange; AEX: anion exchange.

comprehensive analysis of the intracellular CHO proteome using DIA SWATH-MS will help to better understand this model cell line and contribute towards the rational improvement of CHO cells performance[40]. In addition, the profiling of HCCF protein dynamics will also facilitate the bioprocess design and optimization, and the monitoring of HCP impurity in the biologics product[41].

Subsequently, we validated the robustness of the CHO global spectral library by performing a large-scale SWATH-MS stress test experiment on unclarified HCCF samples from CHO-K1 mAb producing cultures

over a period of four months. A total of 360 SWATH-MS runs were carried out in six batches of experimental groups, E01 – E03 were separated using Waters CSH C18 column; while E04 – E06 were separated using Eksigent ChromXP C18 column. Each experimental group consisted of six biological replicates (B01 – B06), and each biological replicate comprised of ten technical replicates (T01 – T10). Except for nine technical replicate runs (four from E05-B06 and five from E06-B06) which were excluded due to spectrum quality failure, total of 351 SWATH-MS data sets were analyzed in OneOmics™ using the abovementioned workflow, and total 4,985 proteins were identified (at 1% peptide FDR) after SWATH-MS data extraction. The CV values for each of the six biological replicates in E01 – E06 were also determined. The violin plot showed highly similar pattern of CV distributions (median CV values of ~35%) although the MS data was acquired over a long period of four months, and two analytical columns from different vendors were used (Fig. 5a). Higher median CVs observed in this experiment might be due to the inherently higher margin of error for quantifying low abundance proteins[42] and much wider dynamic range of protein abundances in the HCCF samples as compared to that of cell lysate samples. We have also observed that lower median CV values (~22%) is achievable when subsets of the replicates were acquired within a shorter duration (data not shown). A Pearson correlation matrix was constructed using the intensities of all identified proteins. While E01 – E03 (180 runs) and E04 – E06 (171 runs) were visually distinct from each other, corresponding to the use of two different analytical columns, the Pearson correlation matrix showed that the 351 runs were positively correlated with each other with a median correlation coefficient r value of 0.88 (indicated as a red colored dash line on the color bar in Fig. 5b and Supplementary Table 1). These results have demonstrated that extracellular CHO proteins can be identified and quantified when performing the SWATH-MS analyses of HCCF using the CHO global spectral library in OneOmics™.

We next evaluated the performance of CHO global spectral library in monitoring HCP clearance in the DSP workflow. The DSP mAb samples collected from a generic DSP practice, including original material (OM), post protein A eluate (PrA), post cation exchange flow through (CEX) and post anion exchange flow through (AEX), were digested and analyzed accordingly (Fig. 1). ELISA analysis of these samples showed that the total HCP concentrations were greatly reduced from 510,762 ppm in OM, to 1,081 ppm in PrA, 557 ppm in CEX, and finally only 7 ppm in the AEX. In the SWATH-MS analysis, total 1,900 proteins were quantified (at 20% CV cutoff and 1% peptide FDR in OM) from one microgram of each DSP sample, including previously reported difficult-to-remove and problematic HCPs[43–45] (Online-only Table 1). Heat map analysis of all the identified proteins showed that majority of the impurities were removed after protein A affinity chromatography (Fig. 5c). Although some HCPs were found to be enriched together with mAb (at the top and bottom of heat map of PrA), these impurities were properly removed by the following polishing steps to generate a highly purified sample (AEX) (Fig. 5c). The protein concentration (in ppm) of four well-known difficult-to-remove HCPs, including clusterin (Clu), cathepsin D (Ctsd), heat shock cognate 71 kDa protein (Hspa8) and phospholipase B-like 2 (Plbl2), were illustrated as examples: these HCPs were successfully removed after the DSP purification steps (Fig. 5d). The SWATH-MS results were not only in line with the ELISA result, but also provided critical information to monitor individual HCP abundance in each step of DSP purification. Taken all together, the CHO global spectral library can be effectively applied in the identification and quantification of thousands of proteins from CHO-derived samples in our SWATH-MS analysis pipeline.

Usage Notes

Generating alternative SWATH-MS spectral library. The current CHO spectral library has been constructed using SCIEX spectral library generation pipeline and could be directly applied in SCIEX SWATH™ Processing software and OneOmics™. With the library merging strategy implemented in the OneOmics™, users could easily expand the proteome coverage of spectral libraries in future. However, future efforts are necessary to expand the library to include cell line-specific protein or peptide identification using other CHO cell lines.

Application of the CHO spectral library in other LC-MS instruments. The CHO spectral library has been successfully applied in the analyses of SWATH-MS data sets acquired using different LC-MS instruments (TripleTOF 5600+) from the same vendor. To apply the spectral library in the DIA analysis using a LC-MS instrument provided by a different vendor, a user needs to ensure the similarity of the ion fragmentation and the optimization of collision energy settings when setting up a DIA acquisition method. Besides, it is recommended to use an analytical column of similar property and spike iRT peptides into the samples for accurate RT alignment and calibration in subsequent data processing.

Absolute protein quantification of HCPs. It is feasible to perform absolute HCP quantification using the CHO global spectral library in the SWATH-MS analysis pipeline. Internal standard proteins with known concentration relative to the mAb product are spiked into the samples prior to LC-MS injection. During data analysis, the CHO global spectral library, appended with the assays of standard proteins, will be used in OneOmics™. The absolute protein concentration (ppm) of targeted HCPs can then be obtained by directly comparing the protein abundance of HCPs to a calibration curve constructed from the standard proteins.

Received: 14 April 2020; Accepted: 6 July 2020;
Published online: 11 August 2020

References

1. Santos, R. *et al*. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16**, 19–34, https://doi.org/10.1038/nrd.2016.230 (2017).
2. Urquhart, L. Market watch: Top drugs and companies by sales in 2017. *Nat. Rev. Drug Discov.* **17**, 232, https://doi.org/10.1038/nrd.2018.42 (2018).

3. Singh, S. *et al*. Monoclonal Antibodies: A Review. *Curr Clin Pharmacol* **13**, 85–99, https://doi.org/10.2174/1574884712666170809124728 (2018).
4. Urquhart, L. Top drugs and companies by sales in 2018. *Nat. Rev. Drug Discov.*, https://doi.org/10.1038/d41573-019-00049-0 (2019).
5. Howie, L. J., Hirsch, B. R. & Abernethy, A. P. A comparison of FDA and EMA drug approval: implications for drug development and cost of care. *Oncology (Williston Park)* **27**, 1195, 1198–1200, 1202 passim (2013).
6. Patel, B. A. *et al*. On-Line Ion Exchange Liquid Chromatography as a Process Analytical Technology for Monoclonal Antibody Characterization in Continuous Bioprocessing. *Anal. Chem.* **89**, 11357–11365, https://doi.org/10.1021/acs.analchem.7b02228 (2017).
7. Ortiz-Enriquez, C. *et al*. Optimization of a recombinant human growth hormone purification process using quality by design. *Prep. Biochem. Biotechnol.* **46**, 815–821, https://doi.org/10.1080/10826068.2015.1135467 (2016).
8. Bade, P. D., Kotu, S. P. & Rathore, A. S. Optimization of a refolding step for a therapeutic fusion protein in the quality by design (QbD) paradigm. *J. Sep. Sci.* **35**, 3160–3169, https://doi.org/10.1002/jssc.201200476 (2012).
9. Esmonde-White, K. A., Cuellar, M., Uerpmann, C., Lenain, B. & Lewis, I. R. Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing. *Anal. Bioanal. Chem.* **409**, 637–649, https://doi.org/10.1007/s00216-016-9824-1 (2017).
10. Xu, N. *et al*. Comparative Proteomic Analysis of Three Chinese Hamster Ovary (CHO) Host Cells. *Biochem. Eng. J.* **124**, 122–129, https://doi.org/10.1016/j.bej.2017.05.007 (2017).
11. Kumar, A. *et al*. Elucidation of the CHO Super-Ome (CHO-SO) by Proteoinformatics. *J Proteome Res.* **14**, 4687–4703, https://doi.org/10.1021/acs.jproteome.5b00588 (2015).
12. Kildegaard, H. F., Baycin-Hizal, D., Lewis, N. E. & Betenbaugh, M. J. The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. *Curr. Opin. Biotechnol.* **24**, 1102–1107, https://doi.org/10.1016/j.copbio.2013.02.007 (2013).
13. Heffner, K. M. *et al*. Lessons from the Hamster: Cricetulus griseus Tissue and CHO Cell Line Proteome Comparison. *J. Proteome Res.* **16**, 3672–3687, https://doi.org/10.1021/acs.jproteome.7b00382 (2017).
14. Vanderlaan, M. *et al*. Experience with host cell protein impurities in biopharmaceuticals. *Biotechnol. Prog.* **34**, 828–837, https://doi.org/10.1002/btpr.2640 (2018).
15. Levy, N. E., Valente, K. N., Lee, K. H. & Lenhoff, A. M. Host cell protein impurities in chromatographic polishing steps for monoclonal antibody purification. *Biotechnol. Bioeng.* **113**, 1260–1272, https://doi.org/10.1002/bit.25882 (2016).
16. Levy, N. E., Valente, K. N., Choe, L. H., Lee, K. H. & Lenhoff, A. M. Identification and characterization of host cell protein product-associated impurities in monoclonal antibody bioprocessing. *Biotechnol. Bioeng.* **111**, 904–912, https://doi.org/10.1002/bit.25158 (2014).
17. Geromanos, S. J. *et al*. The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics* **9**, 1683–1695, https://doi.org/10.1002/pmic.200800562 (2009).
18. Law, K. P. & Lim, Y. P. Recent advances in mass spectrometry: data independent analysis and hyper reaction monitoring. *Expert Rev. Proteomics* **10**, 551–566, https://doi.org/10.1586/14789450.2013.858022 (2013).
19. Hu, A., Noble, W. S. & Wolf-Yadlin, A. Technical advances in proteomics: new developments in data-independent acquisition. *F1000Res* **5**, https://doi.org/10.12688/f1000research.7042.1 (2016).
20. Ludwig, C. *et al*. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **14**, e8126, https://doi.org/10.15252/msb.20178126 (2018).
21. Husson, G. *et al*. Dual Data-Independent Acquisition Approach Combining Global HCP Profiling and Absolute Quantification of Key Impurities during Bioprocess Development. *Anal. Chem.* **90**, 1241–1247, https://doi.org/10.1021/acs.analchem.7b03965 (2018).
22. Walker, D. E. *et al*. A modular and adaptive mass spectrometry-based platform for support of bioprocess development toward optimal host cell protein clearance. *MAbs* **9**, 654–663, https://doi.org/10.1080/19420862.2017.1303023 (2017).
23. 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, https://doi.org/10.1074/mcp.O111.016717 (2012).
24. Silva, C., Santa, C., Anjo, S. I. & Manadas, B. A reference library of peripheral blood mononuclear cells for SWATH-MS analysis. *Proteomics Clin. Appl.* **10**, 760–764, https://doi.org/10.1002/prca.201600070 (2016).
25. Rosenberger, G. *et al*. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1**, 140031, https://doi.org/10.1038/sdata.2014.31 (2014).
26. Wu, J. X. *et al*. SWATH Mass Spectrometry Performance Using Extended Peptide MS/MS Assay Libraries. *Mol. Cell Proteomics* **15**, 2501–2514, https://doi.org/10.1074/mcp.M115.055558 (2016).
27. Blattmann, P. *et al*. Generation of a zebrafish SWATH-MS spectral library to quantify 10,000 proteins. *Sci Data* **6**, 190011, https://doi.org/10.1038/sdata.2019.11 (2019).
28. Sim, K. H. *et al*. A comprehensive CHO SWATH-MS spectral library for accurate quantitative profiling of 10k proteins. *PRIDE Archive* https://identifiers.org/pride.project:PXD016047 (2020).
29. Lim, U. M., Yap, M. G., Lim, Y. P., Goh, L. T. & Ng, S. K. Identification of autocrine growth factors secreted by CHO cells for applications in single-cell cloning media. *J Proteome Res.* **12**, 3496–3510, https://doi.org/10.1021/pr400352n (2013).
30. Paulo, J. A. *et al*. Quantitative mass spectrometry-based multiplexing compares the abundance of 5000 S. cerevisiae proteins across 10 carbon sources. *J. Proteomics* **148**, 85–93, https://doi.org/10.1016/j.jprot.2016.07.005 (2016).
31. Zougman, A., Selby, P. J. & Banks, R. E. Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. *Proteomics* **14**, 1006–1000, https://doi.org/10.1002/pmic.201300553 (2014).
32. Huebsch, M. *et al*. Simplifying the Use of Ion Libraries During Data Processing of SWATH® Acquisition Proteomics Data. *SCIEX*, https://sciex.com/x60589 (2018).
33. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361, https://doi.org/10.1093/nar/gkw1092 (2017).
34. Jarnuczak, A. F. & Vizcaino, J. A. Using the PRIDE Database and ProteomeXchange for Submitting and Accessing Public Proteomics Datasets. *Curr. Protoc. Bioinformatics* **59**, 13 31 11–13 31 12, https://doi.org/10.1002/cpbi.30 (2017).
35. Vizcaino, J. A. *et al*. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, 11033, https://doi.org/10.1093/nar/gkw880 (2016).
36. Perez-Riverol, Y. *et al*. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450, https://doi.org/10.1093/nar/gky1106 (2019).
37. Boutet, E. *et al*. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods Mol. Biol.* **1374**, 23–54, https://doi.org/10.1007/978-1-4939-3167-5_2 (2016).
38. O'Leary, N. A. *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–745, https://doi.org/10.1093/nar/gkv1189 (2016).
39. Podwojski, K. *et al*. Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics* **25**, 758–764, https://doi.org/10.1093/bioinformatics/btp052 (2009).
40. Lewis, A. M., Abu-Absi, N. R., Borys, M. C. & Li, Z. J. The use of 'Omics technology to rationally improve industrial mammalian cell line performance. *Biotechnol. Bioeng.* **113**, 26–38, https://doi.org/10.1002/bit.25673 (2016).
41. Farrell, A., McLoughlin, N., Milne, J. J., Marison, I. W. & Bones, J. Application of multi-omics techniques for bioprocess design and optimization in chinese hamster ovary cells. *J. Proteome Res.* **13**, 3144–3159, https://doi.org/10.1021/pr500219b (2014).

42. Oberg, A. L. & Mahoney, D. W. Statistical methods for quantitative mass spectrometry proteomic experiments with labeling. *BMC Bioinformatics* **13**(Suppl 16), S7, https://doi.org/10.1186/1471-2105-13-S16-S7 (2012).
43. Valente, K. N., Lenhoff, A. M. & Lee, K. H. Expression of difficult-to-remove host cell protein impurities during extended Chinese hamster ovary cell culture and their impact on continuous bioprocessing. *Biotechnol. Bioeng.* **112**, 1232–1242, https://doi.org/10.1002/bit.25515 (2015).
44. Kol, S. *et al*. Multiplex secretome engineering enhances recombinant protein production and purity. *Nat. Commun.* **11**, 1908, https://doi.org/10.1038/s41467-020-15866-w (2020).
45. Gilgunn, S. *et al*. Identification and tracking of problematic host cell proteins removed by a synthetic, highly functionalized nonwoven media in downstream bioprocessing of monoclonal antibodies. *J. Chromatogr. A.* **1595**, 28–38, https://doi.org/10.1016/j.chroma.2019.02.056 (2019).

Acknowledgements

The study was financially supported by AB SCIEX Pte. Ltd. and Industry Alignment Fund (IAF) from Biomedical Research Council (BMRC) of A*STAR (Agency for Science, Technology and Research), Singapore (IAF111189).

Author contributions

X.B. conceived and supervised the study; K.S. and C.L. drafted the manuscript, designed and executed the experiments and data analysis; K.S. H.T. and X.B. finalized the manuscript. K.S. and K.T. performed MS data acquisition; D.N. performed the library statistical analysis; W.Z. performed mAb product downstream processing; Y.Y. performed upstream cell culture of mAb production; S.T. provided critical input into the SWATH-MS analysis; all authors have read and approved the final manuscript.

Competing interests

S.T. is an employee of SCIEX, which operates in the field covered by this work. The research group of X.B. is supported by SCIEX by providing access to instrumentation. K.S. and C.L. are supported partially by the fund from SCIEX. The remaining authors declare no conflicts of interest.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41597-020-00594-z.

Correspondence and requests for materials should be addressed to X.B.

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.

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 Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files associated with this article.

© The Author(s) 2020