Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2018 March 26.

Published in final edited form as:

Nat Methods. 2017 March ; 14(3): 259–262. doi:10.1038/nmeth.4153.

Building ProteomeTools based on a complete synthetic human proteome

Daniel P. Zolg[1,§], **Mathias Wilhelm**[1,§], **Karsten Schnatbaum**[2], **Johannes Zerweck**[2], **Tobias Knaute**[2], **Bernard Delanghe**[3], **Derek J. Bailey**[4], **Siegfried Gessulat**[1,5], **Hans-Christian Ehrlich**[5], **Maximilian Weininger**[1], **Peng Yu**[1], **Judith Schlegl**[6], **Karl Kramer**[1], **Tobias Schmidt**[1], **Ulrike Kusebauch**[7], **Eric W. Deutsch**[7], **Ruedi Aebersold**[8], **Robert L. Moritz**[7], **Holger Wenschuh**[2], **Thomas Moehring**[3], **Stephan Aiche**[5], **Andreas Huhmer**[4], **Ulf Reimer**[2], and **Bernhard Kuster**[1,9,10,*]

[1]Chair of Proteomics and Bioanalytics, Technical University of Munich, Freising, Germany [2]JPT Peptide Technologies GmbH, Berlin, Germany [3]Thermo Fisher Scientific, Bremen, Germany [4]Thermo Fisher Scientific, San Jose, CA, USA [5]SAP SE, Potsdam, Germany [6]SAP SE, Walldorf, Germany [7]Institute for Systems Biology, Seattle, WA, USA [8]Department of Biology, Institute of Molecular Systems Biology, ETH Zürich, Switzerland and Faculty of Science, University of Zurich, Zurich, Switzerland [9]Center for Integrated Protein Science Munich, Freising, Germany [10]Bavarian Center for Biomolecular Mass Spectrometry, Freising, Germany

Abstract

The ProteomeTools project builds molecular and digital tools from the human proteome to facilitate biomedical and life science research. Here, we report the generation and multimodal LC-MS/MS analysis of >330,000 synthetic tryptic peptides representing essentially all canonical human gene products and exemplify the utility of this data. The resource will be extended to >1 million peptides and all data will be shared with the community via ProteomicsDB and proteomeXchange.

Keywords

Liquid chromatography; tandem mass spectrometry; bioinformatics; proteomics; synthetic peptides; protein identification; protein quantification; peptide fragmentation; spectral library; reference spectra; reference standards; ProteomeTools

Proteomic research greatly relies on the mass spectrometric and bioinformatic analysis of proteolytic digests of complex protein mixtures to infer protein identity and quantity[1]. Albeit

*Corresponding author: Bernhard Kuster, kuster@tum.de, Tel. +49 8161 715696, Fax: +49 8161 715931.
§equal contribution

Author Contributions

RA, HW, TM, AH, UR and BK conceived the study. DPZ, MWi, KS, JZ, TK, BD, KK, UK, RLM, BK designed experiments. DPZ, MWi, KS, JZ, TK, BD, DB, PY, KK, EWD, TS performed the experiments and analyzed data. MWi, SG, HCE, MWe, JS, TS, SA extended the web resource. DPZ, MWi and BK wrote the manuscript.

Competing Financial Interests Statement

The authors declare no competing financial interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

very powerful, there are technical and conceptual limitations in commonly followed approaches that make the measurement of complete proteomes very challenging. In part, this is due to the vast molecular complexity of proteomes that arises from e.g. gene expression, splicing of mRNAs or post-translational modification of proteins. As a result, the precise composition of a proteome is essentially always unknown. In addition, the measurement of protease digested proteomes by mass spectrometry creates large quantities of spectra of varying quality. The computational tools used in the field all make assumptions as to the presumed content of a proteome by matching mass spectra to peptides and to infer proteins. The statistical methods applied invariably represent compromises in terms of the sensitivity and specificity with which proteins are identified from complex mixtures. In analytical chemistry, verifying the identity of a molecule with certainty often makes use of synthetic reference standards. However, in proteome research the generation or use of such standards is only beginning to be implemented systematically[2–4]. To facilitate this, we have embarked on a project termed ProteomeTools (Fig. 1) in which we aim to synthesize ~1.4 million individual peptides to cover all human proteins. Here, we report on the synthesis and multimodal LC-MS/MS analysis of >330,000 synthetic tryptic peptides covering essentially all canonical human proteins as annotated in Swissprot. Peptides were chosen either based on their experimentally determined proteotypicity[5, 6] or by brute force (all peptides within the typical mass range of a mass spectrometer) for hitherto unobserved proteins or those with little prior experimental evidence. We also included a subset of peptides of the Human SRMAtlas[3] (Supplementary Table 1). As more data on the use of alternative proteases become available[7], such peptides (~300,000) will be systematically incorporated into the project to increase spectral and sequence coverage for any given protein. Another 200,000 peptides are earmarked for protein sequence variants such as protein isoforms or important natural or pathological variants. A substantial part (~350,000) of the capacity will be devoted to post-translationally modified peptides such as phosphorylation, acetylation, methylation, ubiquitinylation and mono-glycosylation[8]. While some of these peptides may be synthetically more challenging, their impact will likely be high as they represent the result of enzymatic activity that often modulates the function of proteins. Finally, we are reserving 200,000 peptides to represent other interesting biology such as disease associated mutations, HLA neo-antigens, protease cleavage products, small open reading frames or translated lincRNAs (Fig. 1a).

Tryptic peptides were individually synthesized, combined into pools of ~1,000 peptides and spiked with 66 non-naturally occurring and 15 stable isotope labeled peptides for retention time calibration. Whenever possible, we designed pools such that peptides do not have identical masses to avoid ambiguity in the mass spectrometric data or cover the entire LC gradient (Supplementary Figure 1). Each pool was subjected to an initial LC-MS/MS analysis using HCD and CID fragmentation on an Orbitrap Fusion Lumos mass spectrometer in order to determine which peptides were successfully synthesized, to determine their chromatographic retention times (RT) and compute retention time indices (iRT; Supplementary Figure 2)[9]. For each peptide pool, an inclusion list was generated to target peptides for fragmentation in further LC-MS experiments using five different fragmentation methods (HCD, CID, ETD, ETHCD, ETCID) with ion trap or Orbitrap readout and HCD spectra were recorded at 6 different collision energies (Supplementary

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Figures 3–7). Peptides ranged from 7 to 40 amino acids in length and up to 96 % of these could be detected by LC-MS/MS in individual pools (78 % average recovery, Supplementary Figure 8–9). Using the open modification search option of MaxQuant/Andromeda[10], we were able to assess the side product profile to estimate the approximate yield of each peptide by measuring the percentage of the total MS signal that can be attributed to the target peptide sequence (Fig. 1b; Supplementary Figure 10). As expected, the purity of the synthesized peptides varied and many of the chemical by-products correspond to incompletely removed protection groups or missing amino acids. The presence of by-products turned out to be useful as truncated peptides provided additional evidence for the presence of the correct full length peptide. They may in the future also serve to refine retention time and fragmentation prediction or to identify some of the many good quality spectra that remain unidentified in typical proteomic experiments[11].

One important goal of the project is to generate reference mass spectra for the identification and quantification of human peptides and proteins. At an arbitrarily high Andromeda score cut-off of 100, indicating very high spectral quality, we obtained a total of 11.3 million peptide spectrum matches (PSMs) mapping to 211,895 peptides and covering each gene by a median of 9 peptides (Fig. 1c). The median precursor intensity fraction (PIF; i.e. the fraction of the precursor signal vs. the total signal selected for fragmentation) was 92%, indicating that the spectra of most peptides are largely free of other contaminating peptides. Very high quality tandem mass spectra were obtained for all eleven tandem MS methods used but with varying degrees of proteome coverage. Analysis of the Andromeda search engine score distribution (Fig. 1d) showed that the 211,895 peptides (peptide FDR <0.002%) led to the identification of 19,735 of the 20,036 human genes deposited in Swissprot thus providing very high quality reference spectra for 98.5% of the human proteome (Supplementary Table 2). The remaining genes/proteins often contain proline-rich repeats or retroviral sequences that cannot be covered by tryptic peptides of reasonable length. Some of these may eventually be covered when synthesizing peptides using other digestion enzyme specificities. As an interesting side note, because of considerable protein sequence conservation between the human and mouse proteomes, the peptide library also contains 60,961 (proteotypic) peptides covering 12,599 (77%) unique mouse genes, thus considerably expanding the scope and utility of these peptides (Supplementary Table 3; Supplementary Notes).

One obvious use of synthetic peptide reference spectra is to confirm identifications of rarely (or newly) observed proteins. At the time of writing, there were only two spectra in ProteomicsDB supporting the identification of the same peptide of Aquaporin 12B with identification Q-scores different from the target-decoy distributions using the ‘picked’ target-decoy approach[12] (Fig. 2a). The mirror plot showing the ion trap spectrum of the endogenous peptide and the corresponding spectrum of the synthetic peptide indicates very good agreement thus validating the identification of this protein from a single peptide.

We recorded HCD spectra at six different normalized collision energies with the aim to identify conditions for the measurement of peptides by targeted assays such as SRM, PRM or SWATH[13, 14]. To evaluate if HCD spectra obtained in this study are suitable for this purpose, we compared our data to ~9,000 peptides from a SWATH spectral library built from

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

proteome digests acquired on a QTOF instrument[15]. The analysis shows that there is very high correlation between the two types of data (R>0.9) and that spectra with poor correlation may represent false positives in the SWATH spectral library (Fig. 2b; Supplementary Figure 11).

As an example to illustrate the usefulness of the data for developing software, we built a prototype classifier based on multiple HCD spectra for the same peptide at a particular collision energy. This classifier predicts the fragment ion intensity of tandem mass spectra of any peptide with Pearson correlation coefficients of around 0.9 (Fig. 2c, Supplementary Figures 12–14). Such tools may complement or eventually replace experimental data for the development of SRM/PRM/SWATH assays or facilitate the transfer of data recorded on a discovery proteomics instrument to an assay instrument. We consciously decided not to synthesize stable isotope encoded peptides for this project because their fragmentation spectra can be easily simulated based on spectra of the unlabeled version. It is also economically more efficient to create heavy peptides tailored to the project at hand. However, we are in the process of measuring peptides following chemical derivatization by tandem mass tags (TMT) and di-methyl labeling to cover the most commonly used stable isotope labeling methods.

An important aspect of the project is to enable and engage the proteomics community and this will be done in a number of ways. We encourage the community to propose sets of peptides to include in the project. Our stocks contain 100 clones of the peptide library which can be handed out to research groups willing to generate data on alternative mass spectrometric platforms such as QTOF instruments, ion mobility devices or different chromatographic systems. All of the current data is available in proteomeXchange and ProteomicsDB[6, 16] and we will do the same for all upcoming data to engage the bioinformatics community to enable re-use and re-analysis of what we believe is a valuable resource.

The tryptic peptides reported here are only the beginning of the project and many further uses of the information generated can be envisaged. We plan to release new data every six months (~250,000 peptides per release) so that the community can access this data while the project progresses. As the physical reagents are finite resources, the long term value of the work lies in the data. It should be valuable for the development of software tools that may include intelligent data acquisition routines within the instrument control software[17] or the development of more powerful database or spectral library search engines utilizing e.g. concepts of machine learning[18]. There is also still a need to develop improved statistical tools for the assessment of large-scale proteomic experiments particularly for data independent measurements such as SWATH or SRM/PRM. The spectral libraries generated in this project should provide ample opportunity to facilitate these applications[19]. Reagents and software aside, we also plan to build targeted assays in the next two years for sets of functionally important proteins such as kinases and phosphopeptides representing the activation status of signaling pathways. The results obtained so far demonstrate that the overall project is conceptually and technically feasible and yield scientifically meaningful and interesting results. We are therefore confident that the molecular and digital tools arising

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

from the ProteomeTools project will become a valuable resources for the proteomics community in the future.

Data availability

Reference spectra are available at https://www.proteomicsdb.org and updates to the resource are available at www.proteometools.org. The mass spectrometric data have been deposited with the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD004732.

Online Methods

Synthetic peptide sets

To achieve extensive coverage of human proteins, three different sets of peptides were created or used in this study. First, a "proteotypic" peptide set covering confidently and frequently identified proteins was derived from prior mass-spectrometric evidence available in ProteomicsDB[20, 21]. We selected between 2 and 10 unique (at gene level) tryptic peptides for each human gene to reach a cumulative proteotypicity of ~95 % (i.e. we stopped selecting further peptides when the selected peptides covered at least 95% of all cases a particular protein was identified). Further constraints included a peptide length of 7 to 40 amino acids and no more than two missed tryptic cleavage sites. The resulting list contained 124,875 peptides covering 15,855 human Uniprot/SwissProt annotated genes. Second, a "missing proteome" set was constructed containing tryptic peptides mapping to genes which lacked confident experimental identification evidence in ProteomicsDB. Here, any gene-unique tryptic peptide between 7 to 30 amino acids in length and allowing for a maximum of one missed cleavage site was included in the selection without restricting the number of peptides per gene. The resulting list contained 140,458 peptides covering 4,818 genes. Third, we obtained a subset of the "SRMAtlas" peptides comprising 90,967 peptides mapping to 19,099 genes also covering proteins with empirical evidence and "missing" proteins[22]. Altogether, the three sets of tryptic peptides contained 330,286 non-redundant peptides covering 19,840 human genes as annotated in Uniprot/SwissProt (Version 2016-07-20; 42,164 protein sequences) (see Supplementary Table 1).

Peptide pool design

Peptide pools for synthesis and LC-MS measurement consisted of approximately 1,000 peptides each. The peptide pools representing the "proteotypic" and "missing proteome" sets were designed to have a narrow peptide length distribution to support optimal synthesis. Near isobaric peptides (± 10 ppm) were distributed across different pools of similar length to avoid ambiguous masses in pools wherever possible (Supplementary Table 1). To this end, peptides were first ordered by length and mass. Second, the peptides were sorted by taking every n^{th} peptide within the ordered list of peptides of one length, where n is the number of pools needed to distribute these peptides. The resulting ordering provided a well sampled sub-population of peptides with the same MW distribution. In a third step, peptides with near isobaric (± 10 ppm) mass were identified and, as long as no additional near isobaric

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

conflict was created, distributed across pools with similar peptide length (max 2 AAs difference).

The "SRMAtlas" peptide set was acquired in 96 well plates with each well containing one individual proteotypic peptide of 7–30 amino acids in length (i. e. one peptide per well, PEPotec Grade 1, suspended in 0.1% TFA in 50% (v/v) acetonitrile/water). These were also pooled into sets of approximately 1,000 peptides. To create plate pools, the peptides from every plate were first manually pooled together resulting in mixtures of 95 peptides (one quality control peptide present in each plate was discarded to avoid accumulation of this peptide in the subsequent pooling process). To create measurement pools of ~1,000 peptides, either 10 (for fully-tryptic peptides; i.e. C-terminal K/R) or 14 (non- and semi-tryptic peptides; i.e. non-K/R C-terminal) plate pools were combined. In order to avoid bias in pools of 1000 peptides regarding mass (MW) or hydrophobicity index (HI)[23], the pooling scheme was computed to best mimic the overall MW and HI distribution of the entire set. We used the Kolmogorov-Smirnov test (KS-test) to quantify the distance of the MW and HI distribution between mixtures of plate pools to the total distribution of the total set. Starting with a plate pool or mixture of plate pools, all other (still available) plate pools were tested to generate a combined mixture that is closest to the overall set. The best match (lowest p-value) was chosen and the process was repeated until the desired number of plate pools for combination was reached (Supplementary Figure 1). The resulting 96 measurement pools were desalted on C18 material (Waters, SepPak) before storage at -20 °C. All peptide sequences and their pool membership are listed in Supplementary Table 1.

Peptide synthesis

All peptides were individually synthesized following the Fmoc-based solid phase synthesis strategy. A carboxyamidomethylated cysteine building block was used to eliminate the need for cysteine modification prior to MS analysis. Peptides of the "proteotypic" and "missing gene" sets were synthesized by SPOT-synthesis on cellulose membranes at a scale of approximately 2–5 nmol of peptide per spot as described[24]. Depending on the length of peptides in a given pool, up to 6 peptide pools (containing at most 6,000 peptides; see Supplementary Notes) were synthesized in parallel using a purpose built peptide synthesizer. Five quality control peptides were synthesized along with every peptide pool. Peptides were cleaved from the membrane into pools of 1,000 peptides following the design criteria described above. Following solvent evaporation, peptides were stored at -20 °C until use. Peptides from the "SRMAtlas" set were synthesized in 96 well synthesizers (Thermo-Fisher Scientific, PEPotec Grade 1) at a scale of 0.1 mg per peptide. They were pooled and stored as described above.

Sample preparation for mass spectrometry

Dried peptide pools were initially solubilized in 100% DMSO to a concentration of 10 pmol/µl by vortexing for 30 min at room temperature. The pools were then diluted to 10% DMSO using 1% formic acid in HPLC grade water to a stock solution concentration of 1 pmol/µl and stored at −20 °C until use. 10 µl of the stock solution were transferred to a 96-well plate and spiked with two retention time (RT) standards. The first set of retention time peptides (JPT Peptide Technologies) consisted of 66 peptides with non-naturally occurring peptide

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

sequences (Supplementary Table 1). 200 fmol per peptide was used per injection. The second RT standard (Pierce, Thermo Scientific) comprised 15 ^{13}C-labelled peptides and 100 fmol per peptide was used per injection. Samples in the resulting 96-well plates were vacuum dried and stored at −20 °C until use.

Nanoscale liquid chromatography

For LC-MS analysis, the peptide pools in the 96 well plates were dissolved in 0.1% formic acid in water to a concentration of 100 fmol/μl per peptide (residual DMSO concentration of ~1%). An estimated amount of 200 fmol of every peptide in a pool was subjected to liquid chromatography using a Dionex 3000 HPLC system (Thermo Fisher Scientific) using in-house packed C18 columns. The setup consisted of a 75 μm × 2 cm trap column packed with 5 μm particles of Reprosil Pur ODS-3 (Dr. Maisch GmbH) and a 75 μm × 40 cm analytical column, packed with 3 μm particles of C18 Reprosil Gold 120 (Dr. Maisch GmbH). Peptides were loaded onto the trap column using 0.1% FA in water. Separation of the peptides was performed by using a linear gradient from 4% to 35% ACN with 5% DMSO, 0.1% formic acid in water over 50 min followed by a washing step (60 min total method length) at a flow rate of 300 nL/min and a column temperature of 50 °C.

Mass spectrometry

The HPLC system was coupled on-line to an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific). Each peptide pool was first measured using a “survey method” consisting of an Orbitrap full MS scan (60k resolution, 5e5 AGC-target, 50 ms maximum injection time, 360–1300 *m/z*, profile mode), followed by MS2 events with a duty cycle of 2 s for the most intense precursors and a dynamic exclusion set to 5 s as follows: (i) HCD scan with 28% normalized collision energy and Orbitrap readout (15k resolution, 1e5 AGC-target, 22 ms maximum injection time, inject ions for all available parallelizable time enabled, 1.3 *m/z* isolation width, centroid mode); (ii) CID scan with 35% normalized collision energy and ion trap readout (rapid mode, 3e4 AGC target, 0.25 activation Q, 22 ms maximum injection time, inject ions for all available parallelizable time enabled, 1.3 *m/z* isolation width, centroid mode). From this data, inclusion lists with retention time constraints were generated for each pool and used for three subsequent LC-MS measurements focusing on different acquisition types. Precursors detected in the “survey method” were scheduled for fragmentation within a 5 min RT window. Peptides lacking identification in the survey run were added to the inclusion as 2+ or 3+ precursor ions, but without retention time scheduling. (1) The “HCD” method consisted of an Orbitrap MS1 scan (120k resolution, 5e5 AGC-target, 50 ms maximum injection time, 360–1300 *m/z*, profile mode) followed by 3 seconds of MS2 scans with consecutive HCD scans at 20, 25 and 30 normalized collision energy and Orbitrap readout (15k resolution, 1e5 AGC-target, 20 ms maximum injection time, inject ions for all available parallelizable time enabled, 1.3 *m/z* isolation width, centroid mode). (2) The “IT” method consisted of an Orbitrap MS1 scan (120k resolution, 5e5 AGC-target, 50 ms maximum injection time, 360–1300 *m/z*, profile mode) followed by 3 seconds of MS2 scans with (i) CID scan with 35 normalized collision energy and ion trap readout (rapid mode, 3e4 AGC target, 0.25 activation Q, 20 ms maximum injection time, inject ions for all available parallelizable time enabled, 1.3 *m/z* isolation width, centroid mode); (ii) HCD scan with 28 normalized collision energy and ion

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

trap, (iii) HCD scan with 20 normalized collision energy and Orbitrap readout; (iv) HCD scan with 23 normalized collision energy and Orbitrap readout. (3) The "ETD" method consisted of an Orbitrap MS1 scan (120k resolution, 5e5 AGC-target, 50 ms maximum injection time, 360–1300 *m/z*, profile mode) followed by 3 seconds of MS2 scans with (i) ETD scan using charge dependent ETD parameters and Orbitrap readout[25]; (ii) EThcD scan using charge dependent ETD parameters and supplemental HCD activation with 28% normalized collision energy and Orbitrap readout; (iii) ETciD scan using charge dependent ETD parameters and supplemental CID activation with 35 normalized collision energy and Orbitrap readout with settings described above.

Data processing

The logistics of data processing and MS method generation was governed by an in-house database (Supplementary Figure 8). RAW data were analysed using MaxQuant version 1.5.3.30 searching individual LC-MS runs against pool-specific databases (see Supplementary Table 2)[26]. If not mentioned otherwise, default parameters were used: Carbamidomethylated cysteine was specified as fixed modification, methionine oxidation as variable modification. First search tolerance was set to 20 ppm, main search tolerance to 4.5 ppm and filtered for peptide and protein FDR of 1 %. Retention time windows of ±5 min were corrected for drifts using the internal retention time standards. The pool-specific inclusion lists were generated from confidently identified precursors (from the survey method) which passed an ad-hoc Andromeda score cut-off of 100. For analysis of synthesis side product, the survey MS run was searched with unspecific digestion and "dependent peptides" enabled.

Conserved peptide sequences in human and mouse

A current mouse protein sequence database representing 16,336 mouse genes was obtained from Swissprot (version dated 07/09/2016, 16,818 sequences). The database was *in silico* digested using tryptic cleavage specificity (no proline rule) and a maximum of 2 missed cleavages. The resulting peptide list was filtered for unique entries and mapped against our sequence list of peptides (see Supplementary Notes and Supplementary Table 3).

Comparison QTOF vs Fusion Lumos spectra

We systematically compared spectra generated in this project on an Orbitrap Fusion Lumos (Thermo) to a spectral library generated on a 5600 TripleTOF (QTOF) mass spectrometer (AB Sciex)[27]. For this, intensities of matching annotated fragment ions of the highest scoring (>100) beam-type CID spectrum per acquired normalized collision energy (Lumos) were correlated using Pearson correlation to the corresponding beam-type CID QTOF spectrum (acquired with rolling collision energy).

Fragmentation Prediction

First, MaxQuant result files were parsed and only spectra of unmodified doubly charged peptides with a PIF > 0.8 and a score of higher than 100 were selected for training. For each combination of amino acids N-terminal (left) and C-terminal (right) of the fragmentation position at a given normalized collision energy, a local polynomial regression (LOESS)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

model was fitted using the peptide length normalized fragmentation position and the base peak intensity (BPI) normalized intensity of the y-ions (see Supplementary Figure 11–13). The resulting models were tested on pool 66 of the "proteotypic" set using the same peptide selection criteria. Each possible y-ion for each peptide passing the filters was predicted using the corresponding LOESS-fit. The predicted y-ion intensities were scored against the measured spectra using the Pearson correlation coefficient.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank numerous colleagues including S. Eluik, G. Tan, X. Sun, X. Liu (Thermo Fisher Scientific), A. Hubauer, J. Mergner, J. Zecha, P. Samaras and the entire Kuster team as well as H. Hahne (OmicScouts), R. Weise, D. Riehn, K. Schrödter, F. Schumacher, N. Kolls and R. A. Castro-Alvaro (JPT) for fruitful discussions and technical assistance. We thank the efforts of D. Campbell and Z. Sun and the entire Moritz group at ISB and the Aebersold group at ETH Zurich for their efforts in peptide selection and synthesis. This work was in part funded by the German Federal Ministry of Education and Research (BMBF; grant No 031L0008A). A postdoctoral fellowship from the Alexander von Humboldt Foundation (to PY) is also gratefully acknowledged. This work was also performed in part with federal funds from the American Recovery and Reinvestment Act (ARRA) funds through NIH, from the National Human Genome Research Institute grant RC2HG005805 (to R.L.M.), the National Institute of General Medical Sciences under grant R01GM087221, S10RR027584 and 2P50 GM076547/Center for Systems Biology (to R.L.M.), the European Research Council grant ERC-2008-AdG 233226 and ERC-2014-AdG 670821 and the Swiss National Science Foundation (grant #31003A-130530) (to R.A.), and DAAD (fellowship to U.K.).

Abbreviations

AGC	Automatic gain control
CE	Collision energy
CID	Collision induced dissociation
DIA	Data independent acquisition
DMSO	Dimethyl sulfoxide
ESI	Electrospray ionization
ETciD	Electron-transfer/collision induced dissociation
ETD	Electron transfer dissociation
EThcD	Electron-transfer/higher-energy collision dissociation
FA	Formic acid
HCD	Higher energy CID
HLA	Human leucocyte antigen
IT	IonTrap
iTRAQ	Isobaric tags for relative and absolute quantification

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

LC-MS/MS	Liquid chromatography-tandem mass spectrometry
MS	Mass spectrometry
NCE	Normalized collision energy
nESI	Nano-ESI
ORF	Open reading frame
OT	Orbitrap
PIF	Precursor intensity fraction
QTOF	Quadrupole time-of-flight
RT	Retention time
SRM/PRM	Selected reaction monitoring/parallel reaction monitoring
SWATH	Sequential window acquisition of all theoretical fragment ion spectra
TMT	Tandem mass tags

References

1. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR 3rd. Protein analysis by shotgun/bottom-up proteomics. Chemical reviews. 2013; 113:2343–2394. [PubMed: 23438204]
2. Ahrens CH, Brunner E, Qeli E, Basler K, Aebersold R. Generating and navigating proteome maps using mass spectrometry. Nature reviews Molecular cell biology. 2010; 11:789–801. [PubMed: 20944666]
3. Kusebauch U, et al. Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. Cell. 2016; 166:766–778. [PubMed: 27453469]
4. Picotti P, et al. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. Nature. 2013; 494:266–270. [PubMed: 23334424]
5. Mallick P, et al. Computational prediction of proteotypic peptides for quantitative proteomics. Nature biotechnology. 2007; 25:125–131.
6. Wilhelm M, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014; 509:582–587. [PubMed: 24870543]
7. Giansanti P, et al. An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas. Cell reports. 2015; 11:1834–1843. [PubMed: 26074081]
8. Marx H, et al. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. Nature biotechnology. 2013; 31:557–564.
9. Escher C, et al. Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics. 2012; 12:1111–1121. [PubMed: 22577012]
10. Cox J, et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. Journal of proteome research. 2011; 10:1794–1805. [PubMed: 21254760]
11. Griss J, et al. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. Nat Meth. 2016; 13:651–656.
12. Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. Molecular & cellular proteomics: MCP. 2015; 14:2394–2404. [PubMed: 25987413]
13. Gallien S, et al. Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. Molecular & cellular proteomics: MCP. 2012; 11:1709–1723. [PubMed: 22962056]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

14. Lawrence RT, Searle BC, Llovet A, Villen J. Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. Nat Meth. 2016; 13:431–434.
15. Guo T, et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat Med. 2015; 21:407–413. [PubMed: 25730263]
16. Vizcaino JA, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nature biotechnology. 2014; 32:223–226.
17. Bailey DJ, McDevitt MT, Westphall MS, Pagliarini DJ, Coon JJ. Intelligent data acquisition blends targeted and discovery methods. Journal of proteome research. 2014; 13:2152–2161. [PubMed: 24611583]
18. Kelchtermans P, et al. Machine learning applications in proteomics research: how the past can boost the future. Proteomics. 2014; 14:353–366. [PubMed: 24323524]
19. Wang J, et al. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. Nature methods. 2015; 12:1106–1108. [PubMed: 26550773]
20. Mallick P, et al. Computational prediction of proteotypic peptides for quantitative proteomics. Nature biotechnology. 2007; 25:125–131.
21. Wilhelm M, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014; 509:582–587. [PubMed: 24870543]
22. Kusebauch U, et al. Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. Cell. 2016; 166:766–778. [PubMed: 27453469]
23. Krokhin OV. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. Analytical chemistry. 2006; 78:7785–7795. [PubMed: 17105172]
24. Wenschuh H, et al. Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides. Biopolymers. 2000; 55:188–206. [PubMed: 11074414]
25. Rose CM, et al. A calibration routine for efficient ETD in large-scale proteomics. Journal of the American Society for Mass Spectrometry. 2015; 26:1848–1857. [PubMed: 26111518]
26. Shanmugam AK, Nesvizhskii AI. Effective Leveraging of Targeted Search Spaces for Improving Peptide Identification in Tandem Mass Spectrometry Based Proteomics. Journal of proteome research. 2015; 14:5169–5178. [PubMed: 26569054]
27. Guo T, et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat Med. 2015; 21:407–413. [PubMed: 25730263]

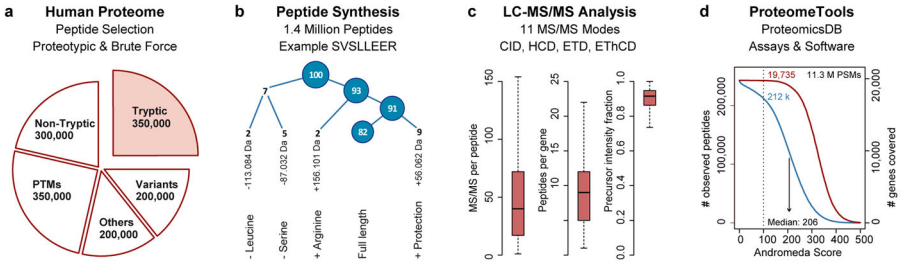

Figure 1. Overview of the ProteomeTools project

(a) Planned segmentation of the 1.4 million peptides that will be selected from the human proteome and synthesized over the course of the project. Here, we report on the analysis of 330,000 individually synthesized tryptic peptides. **(b)** Estimation of synthesis success using peptide precursor intensity information for the peptide SVSLLEER and its by-products. Here, 82% of the total MS signal can be attributed to the full length product. **(c)** Boxplots for the number of tandem mass spectra identifying a given peptide with very high confidence (Andromeda score >100; total of 11.3 million PSMs in 11 types of tandem MS); the number of such peptides (total of 211,895) covering a given protein/gene (total of 19,735) and the average precursor intensity fraction (PIF; see main text) of these peptides. **(d)** Distribution of peptide and protein identifications as a function of the Andromeda score. All data is available in ProteomicsDB and proteomeXchange.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

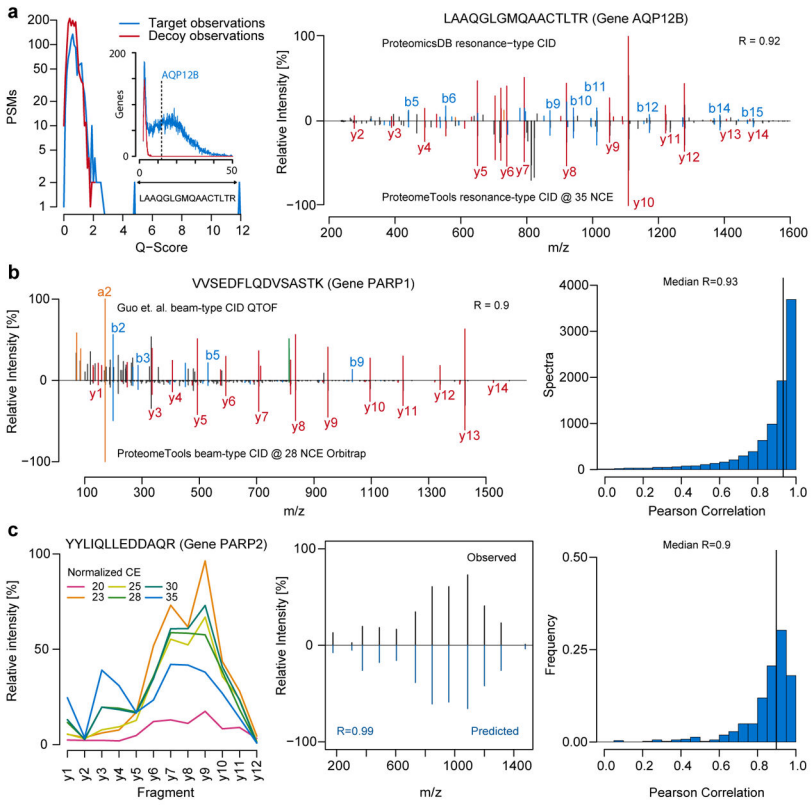

Figure 2. Data analysis and application

(a) Protein identification: target/decoy search results for the peptide LAAQGLGMQAACTLTR of Aquaporin 12B (AQP12B). There are only two spectra in ProteomicsDB with identification Q-Scores distinct from the decoy distribution (left panel). The inset shows the Q-Score distribution of all genes/proteins in ProteomicsDB, placing AQP12B well way from the decoy proteins. The right panel shows the best CID mass spectrum for AQP12B in ProteomicsDB (top) compared to the corresponding CID spectrum of the synthesized reference peptide confirming this identification. **(b)** Transferability between MS instruments: comparison of a spectrum acquired from a complex digest by beam-type CID on a QTOF instrument for the peptide VVSEDFLQDVSASTK compared to the corresponding spectrum of the synthesized reference peptide acquired by beam-type CID on an Orbitrap instrument (left panel). Fragment ion intensities show very high correlation (Pearson correlation of 0.9). Extending this analysis to ~9,000 peptides confirmed the high correlation of these two types of tandem mass spectra (right panel). **(c)** Development of a predictor for tandem mass spectra. HCD data were recorded at six collision energies. The left panel shows the median relative fragment ion intensities of 12 y-fragment ions for the peptide YYLIQLLEDDAQR. Using these characteristics for all spectra of all peptides, a predictive model was trained for each normalized collision energy. The comparison of measured and predicted spectra for YYLIQLLEDDAQR (middle panel) show very good agreement. The histogram on the right shows that the predictor (tested on 529 peptide sequences and 3,248 spectra) of pool 66 of the proteotypic peptide set, is generally able to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

predict the relative y-ion intensity for a given peptide with good quality (see Supplementary Notes for details).