Six alternative proteases for mass spectrometry–based proteomics beyond trypsin

Piero Giansanti[1–3], Liana Tsiatsiani[1–3], Teck Yew Low[1,2] & Albert J R Heck[1,2]

[1]Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, the Netherlands. [2]Netherlands Proteomics Centre, Utrecht, the Netherlands. [3]These authors contributed equally to this work. Correspondence should be addressed to A.J.R.H. (a.j.r.heck@uu.nl).

Published online 28 April 2016; doi:10.1038/nprot.2016.057

Protein digestion using a dedicated protease represents a key element in a typical mass spectrometry (MS)-based shotgun proteomics experiment. Up to now, digestion has been predominantly performed with trypsin, mainly because of its high specificity, widespread availability and ease of use. Lately, it has become apparent that the sole use of trypsin in bottom-up proteomics may impose certain limits in our ability to grasp the full proteome, missing out particular sites of post-translational modifications, protein segments or even subsets of proteins. To overcome this problem, the proteomics community has begun to explore alternative proteases to complement trypsin. However, protocols, as well as expected results generated from these alternative proteases, have not been systematically documented. Therefore, here we provide an optimized protocol for six alternative proteases that have already shown promise in their applicability in proteomics, namely chymotrypsin, LysC, LysN, AspN, GluC and ArgC. This protocol is formulated to promote ease of use and robustness, which enable parallel digestion with each of the six tested proteases. We present data on protease availability and usage including recommendations for reagent preparation. We additionally describe the appropriate MS data analysis methods and the anticipated results in the case of the analysis of a single protein (BSA) and a more complex cellular lysate (*Escherichia coli*). The digestion protocol presented here is convenient and robust and can be completed in ~2 d.

INTRODUCTION

MS–based proteomics is currently the method of choice for the systematic and global analysis of proteins. Besides the identification of protein-protein interaction networks and the quantitative analysis of post-translational modifications (PTMs), MS-based proteomics has enabled the identification of almost the complete proteome of *Saccharomyces cerevisiae*[1,2] and contributed to the first drafts of the human proteome[3,4]. Shotgun proteomics, also termed bottom-up proteomics, focuses on the analysis of protein mixtures after enzymatic digestion of the proteins into peptides. The resulting complex mixture of peptides is analyzed by reverse-phase liquid chromatography (RP-LC) coupled to electrospray ionization tandem mass spectrometry (MS/MS)[5,6], often in combination with other upstream prefractionation methods such as strong cation exchange (SCX) to reduce sample complexity. Finally, identification of peptides and subsequently proteins is performed by matching peptide fragment ion spectra to theoretical spectra generated from protein databases with dedicated computational algorithms, as reviewed in detail numerous times[7–11].

Despite parallel advances of different components within the shotgun proteomics workflow, such as cell lysis, peptide separation, mass spectrometry and database search algorithms, protein digestion is still largely performed using a single enzyme—i.e., trypsin. Other proteases such as chymotrypsin, LysC, LysN, AspN, GluC and ArgC are also used in proteomics but to a lesser extent[12]. Because of their distinctive specificities, these proteases generate different sets of peptides; therefore, by simply subjecting the same proteome to digestion with multiple proteases in parallel, complementary parts of the protein sequence space and thus the proteome can be covered[13–16]. It is evident that this strategy ensures higher sequence coverage and thus enables the discrimination of closely related protein isoforms, as well as robust and precise protein identification and quantification, including that of the various PTM sites[17–19]. In fact, our laboratory has demonstrated the advantages of parallel digestion with multiple proteases for quantification[20] of proteomes obtained from human heart[21], human adenovirus[22] and rat liver[23]. We further showed that analysis of distinct proteolytic peptides improves the identification of novel phosphorylation sites and motifs[24,25].

These results imply that proteolytic digestion by trypsin alone can be a limiting factor in peptide-centric proteomics. It has been demonstrated that simple changes in standard digestion methodologies[26,27] can result in increased number of identified proteins in complex mixtures[28]. Protocols for in-solution protein digestion are available, although they often differ in several aspects, including the amount of enzyme used (enzyme to protein ratio), incubation temperature, duration of digestion, buffering reagents, conditions for reduction and alkylation and the concentration of denaturants used for cell lysis and protein solubilization. Here, we present an optimized protocol for non-tryptic, in-solution protein digestion that can be adapted for the LC-MS/MS analysis of single proteins or mixtures up to whole mammalian cell lysates. As a model system for describing the detailed protocol, we use BSA and the bacterial *E. coli* cell lysate. We selected and compared the six proteases chymotrypsin, LysC, LysN, AspN, GluC and ArgC because of their increasing availability and popularity for use in proteomics. Besides a detailed description of the merits and limitations of each enzyme, enzyme availability and recommendations for buffer composition and optimal digestion conditions are presented. The digestion protocol for six commercially available proteases, as described in this work, has been extensively optimized so as to minimize deviations from the conventional tryptic digestion protocol and to facilitate convenient adoption by the proteomics community at large. We also describe how to perform the MS data analysis and the anticipated results in the case of each of the six presented

proteases. These protocols can be further amended in several aspects, including the use of different components in the lysis buffers, data-dependent decision tree algorithms for peptide fragmentation and alternative search engines so as to improve the overall performance of parallel multiple-enzyme digestions.

Experimental design

Lysate preparation. Depending on the source of the sample, the lysis step can be performed in a number of formats. For most mammalian cell lines, only gentle lysis is necessary. Cells can be collected, washed with PBS and resuspended in lysis buffer, followed by homogenization or sonication. As for bacterial, fungal, plant or tissue samples, more intensive physical disruptions may be necessary because of specific cell wall properties[26,29]. Appropriate protocols for sample disruption have to be optimized before use. Always look out for any cloudiness in the suspension, which indicates insoluble materials. To assist solubilization, protein lysates can be sonicated for a few cycles until they become clear, followed by centrifugation to remove any insoluble material. It is important to perform the above process on ice with the necessary protease inhibitors, so as to prevent undesirable protein degradation by endogenous proteases in the cells whose activity may be enhanced during lysis.

There are several issues associated with the use of protease inhibitors during lysate preparations. These should be used with caution and selected so as not to interfere with the mass spectrometry analyses. For example, 4-(2-aminoethyl) benzenesulfonyl fluoride hydrochloride, a commonly used irreversible serine protease inhibitor, has been shown to modify or derivatize nontarget proteins on multiple residues, thus introducing artifacts in mass spectra interpretation[30]. Importantly, chelating agents such as EDTA should not be used in the lysis buffer when the proteolytic digestion is performed by a metalloprotease. PMSF should be avoided in case of serine proteases. Please refer to **Supplementary Table 1** for additional and specific instructions on the compatibility of detergents and protease inhibitors for each of the different proteases.

If the lysate preparation requires the use of chemicals and/or protease inhibitors that are incompatible with the enzymes selected for the in-solution digestion, we recommend performing removal of the interfering compounds (e.g., protein precipitation, dialysis, ion-exchange chromatography) and resuspension of the protein pellet in the appropriate buffer (**Supplementary Table 1**).

Protein digestion. Before proceeding with digestion, estimate the protein concentration of the lysate using a Bradford or bicinchoninic acid assay. It is always advantageous to have an idea about the amount of protein required before proteomics analysis so as to be able to estimate the appropriate amount of protease to be used for the generation of a sufficient amount of peptides to cater for technical, process or biological replicates. Furthermore, if single-shot LC-MS/MS analyses is to be performed for a proteome digest, we recommend digestion of 50–100 µg of protein sample, as this amount can be comfortably handled, although each injection into state-of-the-art nanoLC-MS/MS system typically should not exceed 1 µg. However, if one intends to cover as much proteome space as possible by applying off-line prefractionation techniques such as SCX or basic RP-LC, depending on the sensitivity of the mass spectrometer used, the LC configuration and type of prefractionation, we recommend using higher initial protein amounts—i.e., 50–200 µg. An even higher amount, i.e., on the milligram scale, is necessary if the aim of the experiment is to enrich for low-abundant proteins/complexes using affinity purification or immunoprecipitation or to isolate particular proteins modified by PTMs—e.g., phosphorylated or ubiquitinated peptides. The above recommendations assume that the amounts of samples are not limiting.

Sample desalting. Sample desalting and cleanup is essential in order to remove undesirable salts, reagents and buffers that may compromise peptide ionization or the LC separation, thus decreasing the resolving power of the mass spectrometer. For our proteomics studies, we find commercially available kits for C18 solid-phase extraction to be very useful. To clean up very small amounts, for example, <10 µg of total peptides, it is more useful to pack C18 STAGE-Tips in-house. For this, a detailed protocol is available[31].

Quality control of the protein digestion. It is prudent to monitor the quality of the digested samples before committing them to LC-MS/MS analysis, as this process can be lengthy and expensive, especially when extensive peptide prefractionation is involved. Estimation of peptide concentration after digestion and/or C18 cleanup can be performed by commercially available colorimetric or fluorometric peptide assays. This is highly recommended when performing quantitative analysis, especially when label-free quantification is performed. If peptide prefractionation is used, the LC-UV trace (when an UV detector is available) can be used to estimate the amount of sample to inject for each collected fraction.

After protein amount estimation and digestion, we recommend injecting ~0.1–0.5 µg of complex sample digest into a LC-MS/MS system to check the sample quality (or 20–50 fmol for single-protein digest). In this process, we always evaluate the elution profile of the peptides, the presence of polymer peaks or any entities that can suppress the ion intensity of peptides. A list of commonly detected contaminants for LC-MS/MS has been published to assist tracing the major contaminants[32]. For data acquired with Thermo Fisher instruments, we use RawMeat, which is a software tool for data quality assessment to quickly monitor the charge states, precursor mass spectra and fragment ion spectra of the peptides based on the proteases used. For proteases that generate longer and more highly charged peptides such as LysC, LysN, ArgC, AspN and GluC[33], if the electron transfer dissociation option is available, we recommend using a data-dependent decision tree setting, as we find this to yield generally more unique peptide identifications. However, in this work, we only present data based on higher-energy collisional dissociation (HCD) because HCD and beam-type collision-induced dissociation (CID) are still the most common methods for peptide fragmentation in the majority of proteomics laboratories operating Orbitrap or quadrupole-time of flight (qTOF) instruments as well as triple-quadrupole systems.

Quality control of LC-MS/MS system. Before analysis of the samples, we evaluate the performance of the LC-MS/MS system by

TABLE 1 | Advantages and limitations of proteases used in typical shotgun proteomics experiments.

Protease	Family	Cleavage site	Advantages	Limitation
ArgC	Cysteine protease	C-terminal of R	ArgC is mostly combined with other proteases to investigate PTMs and to increase the proteome coverage qualitatively	ArgC cleaves at the carboxyl terminus of R residues. It can also cleave at K residues, although with less efficiency ArgC peptides are generally longer than tryptic peptides (**Supplementary Fig. 1**) An improved identification rate could be achieved using complementary and alternative peptide fragmentation strategies
AspN	Metalloprotease	N-terminal of D	AspN can perform hydrolysis of peptide bonds at the amine side of D residues. It also functions within a pH range of 4–9	If detergents are present in the digestion buffer, Asp-N can cleave at the amine side of E residues AspN cleaves more efficiently at the N termini of D than E residues, resulting in many missed cleavages AspN peptides are generally longer than tryptic peptides (**Supplementary Fig. 1**) An improved identification rate could be achieved using complementary and alternative peptide fragmentation strategies
Chymotrypsin	Serine protease	C-terminal of F, Y, L, W and M	Peptides produced from a chymotrypsin digest cover a proteome space that is most orthogonal to that of trypsin, in both a qualitative and quantitative manner. Because of its preference for hydrophobic amino acids, chymotrypsin is particularly useful for covering transmembrane regions of membrane proteins (**Supplementary Fig. 1**)	The efficiency of chymotrypsin toward different hydrophobic amino acid residues varies and results in quite a few missed cleavages
GluC	Serine protease	C-terminal of D	GluC can be combined with other proteases for the study of PTMs and to increase proteome coverage qualitatively	Specificity of GluC depends on the pH and the buffer composition. At pH 4, the enzyme preferentially cleaves at the C terminus of E, whereas at pH 8 it additionally cleaves at D residues GluC peptides are generally longer than tryptic peptides (**Supplementary Fig. 1**) An improved identification rate could be achieved using complementary and alternative peptide fragmentation strategies
LysargiNase[a16]	Metalloprotease	N-terminal of R and K	Its specificity mirrors trypsin by cleaving at the N-terminal side of R and K residues Enables the identification of peptides derived from the protein C termini. These would otherwise not be identifiable after trypsin digestion	

(continued)

TABLE 1 | Advantages and limitations of proteases used in typical shotgun proteomics experiments (continued).

Protease	Family	Cleavage site	Advantages	Limitation
LysC	Serine protease	C-terminal of K	Very efficient and specific	Peptides generated by LysC alone overlap significantly with tryptic peptides, and therefore sequence coverage of proteins may not increase significantly
			Often used to complement trypsin in a serial LysC > trypsin digestion protocol to complement the somewhat lower efficiency of trypsin toward K residues	LysC peptides are generally longer than tryptic peptides (**Supplementary Fig. 1**)
			LysC is resistant to denaturants (such as 8 M urea). This allows proteins to be digested in their optimal denatured state which enhances digestion efficiency	An improved identification rate could be achieved using complementary and alternative peptide fragmentation strategies
LysN	Metalloprotease	N-terminal of K	LysN is more resistant to denaturants than trypsin. It may also be heated to 70 °C	The specificity of LysN toward K residues accounts for 90% of the cleavages in complex protein samples. Occasionally, LysN also cleaves N-terminally to A, S and R
			The combination of LysN with ETD peptide fragmentation provides unique and straightforward sequence interpretation, and it allows facile de novo sequencing	
Pepsin[a16]	Aspartic protease	C-terminal of Y, F and W	Because Pepsin exhibits broad specificity and high activity at a low pH, it is the preferred enzyme for determining disulfide bonds by MS. Digestion of proteins at low pH eliminates disulfide reshuffling	Pepsin has a preference for aromatic residues (Y, F and W) and L, but its specificity is pH-dependent
			Pepsin remains active at low temperature (4 °C) and pH (2.5), which is essential for hydrogen/deuterium exchange experiments using LC-MS/MS, as such conditions inhibit the back exchange of deuterium to hydrogen	The complexity of the peptide mixture impairs spectra interpretation. It is, however, possible to resolve this by the use of dedicated search engines
Trypsin	Serine protease	C-terminal of R and K	Trypsin is very efficient, specific and broadly available at a relatively reasonable cost. It is the gold standard for shotgun proteomics	Tryptic peptides are generally short (6 residues, **Supplementary Fig. 1**). Therefore, trypsin alone covers only a restricted portion of the proteome
			It generates peptides with either R or K at the C termini, making them amenable to peptide fragmentation with CID, generating useful b and y series gaseous ions for peptide sequence annotation	The presence of negatively charged amino acids such as D, E phosphorylated S and T adjacent or in close proximity to R or K residues, prevents tryptic cleavage and thus leads to missed cleavages and longer peptides
				Trypsin exhibits a lower cleavage efficiency toward K than R residues
				Often unable to produce MS-identifiable peptides derived from the C termini of proteins
WaLP and MaLP[a15]	Serine protease	C-terminal of aliphatic amino acids	These α-lytic proteases are specific toward aliphatic residues. This makes them particularly valuable for the study of membrane protein sequences	

[a]Proteases not experimentally used in the present work.

analyzing a 20 fmol trypsin digest of BSA with a short (45 min) gradient. By monitoring the chromatographic peak shapes and retention time for defined BSA peptides, we assess and diagnose the peptide trap and analytical columns. We repeat these 45-min acquisitions after every 1–3 samples, not only to monitor instrument performance but also to condition and clean the columns so as to reduce peptide carry-over from one sample to another. To evaluate the performance of the LC-MS/MS system for the analysis of complex samples, we typically perform LC-MS/MS acquisitions with a 90-min gradient on 50 ng of tryptic *E. coli* digestion. Fragment ion spectra are then searched against an *E. coli* protein database and filtered with 1% false discovery rate (FDR). The final numbers for peptide-to-spectra matches (PSMs) and unique peptides are then compared with reference specification numbers that are accumulated over time for a designated MS instrument. This allows us to decide whether the instrument is optimal for handling complex samples.

Advantages and limitations. The activities of different proteases differ in terms of specificity and digestion efficiency. Given the same protease, the activity may also differ depending on the sources/vendors. Besides, most proteolytic enzymes (including trypsin) are generally unable to withstand the harsh solubilizing conditions often used in proteomics, such as 8 M urea, with the exceptions of LysC and LysN. Therefore, it is beneficial to initially refer to the recommended digestion conditions by the suppliers by carefully considering the pH, amount of unfolding reagents and incubation temperature. Some proteases may produce peptides of atypical lengths or charge states that are thus outside the chromatographic coverage of C18 columns or are beyond the detection range of most LC-MS/MS systems. Database search algorithms that are specialized for bottom-up mass spectrometry may also be less optimal for these unconventional peptides. We note that although LysN and chymotrypsin, just like trypsin, can be easily adapted for in-gel digestion, the other proteases are less efficient in in-gel digestion.

To enable researchers to make the right choice, we have compiled a list of proteases used in large-scale shotgun proteomics experiments and outlined some of their merits and limitations (**Table 1**)[12].

In addition, when performing PTM-based proteomics analysis, one has to consider that a modification may have an effect on a cleavable site, thus hampering, for instance, efficient proteolytic cleavage (acetylation for LysC, LysN, or methylation for LysC and to a lesser extend LysN). This again might result in the generation of peptides unamenable to MS analysis. Reduced proteolytic activity has also been reported when the PTM (e.g., phosphorylation and O-GlcNAc) occurs on residues in the proximity of the cleavable site[34–36], because of steric and/or electrostatic hindrance. To tackle this issue, we generally allow higher numbers of missed cleavages for PTM-modified peptides when performing mass spectra to peptide sequence matching.

A separate discussion can be reserved for ubiquitinated proteins. Ubiquitination sites can be identified by MS through the detection of the ubiquitin remnant–containing peptides. The C terminus of the mature ubiquitin protein contains the amino acid sequence DYNIQKESTLHLVLRLRGG, in which the last G can be conjugated to lysine residues on target proteins. Whereas for ArgC, and for chymotrypsin, the ubiquitin side chain is trimmed down to a small tag (GG and RGG, respectively), with other proteases, the resulting remnant might be too long (13–18 amino acids), which reduces the chance of successful identification of the fragmented peptide.

It is noteworthy that the introduction of multiplexing isotopic labels (i.e., stable isotope labeling by amino acids in cell culture or dimethyl labeling) or isobaric tags (tandem mass tag or isobaric tags for relative and absolute quantification) will have a negative effect on the observed proteome coverage, especially with respect to highly complex proteomes. In the first case, the lower sequence coverage is probably due to the increased complexity of the peptide mixture, whereas for isobaric tag–based labeling, a high-resolution mass spectrometer will be required to resolve and accurately measure the relative intensity of the isobaric reporter ions. Therefore, we advise to carefully evaluate which labeling technique is best to use upon any proteolytic digestion and, when possible, to choose label-free quantification (shotgun and targeted), which is compatible with the peptide products of virtually any protease[13,20,36].

MATERIALS

REAGENTS

- Acetonitrile (ACN; Biosolve, cat. no. 012007)
- Acetic acid (AA; Merck, cat. no. 1.00063)
- Trifluoroacetic acid (TFA; Thermo Scientific, cat. no. TS-28904)
  ! CAUTION TFA solutions and TFA vapors are toxic; prepare solutions in a fume hood.
- Formic acid (FA; Fluka, cat. no. 94318)
- High-purity water obtained from a Milli-Q purification system (Millipore)
- Urea (Merck, cat. no. 66612)
- Ammonium bicarbonate (NH_4HCO_3; Fluka, cat. no. 09830)
- Complete Mini EDTA-free cocktail (Roche, cat. no. 11.836.170.001)
- PhosphoSTOP Phosphatase Inhibitor cocktail (Roche, cat. no. 04.906.845.001)
- DL-dithothreitol (DTT; Sigma-Aldrich, cat. no. 43815)
- Iodoacetamide (IAA; Sigma-Aldrich, cat. no. I6125)
- Trypsin (Promega, cat. no. V528A)
- LysC MS grade (Wako Chemicals, cat. no. 129-02541)
- LysN (Thermo Scientific, cat. no. 90300)
- AspN (Roche, cat. no. 11.054.589.001)
- GluC (Roche, cat. no. 11.047.817.001)
- ArgC (Roche, cat. no. 11.370.529.001)
- Chymotrypsin (Roche, cat. no. 11.428.467.001)
- BSA (Sigma-Aldrich, cat. no. A2153)
- Cells to be analyzed. The procedure is optimized for lysis of *E. coli* strain K12 (Invitrogen, DH5α)
- PBS (PAA Laboratories GmbH)

EQUIPMENT

- Sonicator UP100H (Hielscher Ultrasound Technology)
- Vortex (VWR)
- Eppendorf centrifuge 5417R (Eppendorf)
- Milli-Q purification system (Millipore)
- Sep-Pak C18 cartridges (Waters)
- Sep-Pak vacuum scaffold (Waters)
- SpeedVac (Thermo Scientific)
- Agilent 1290 UPLC system (Agilent)
- Q Exactive Plus quadrupole Orbitrap mass spectrometer (Thermo Fisher Scientific)
- Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific)
- MASCOT (Matrix Science)
- RawMeat (free download from Vast Scientific)
- Proteome discoverer version 2.0 or higher (Thermo Fisher Scientific)

REAGENT SETUP

▲ CRITICAL Please refer to **Table 2** for specific instructions on the use of each of the different proteases for sample digestion.

TABLE 2 | Proteolytic enzymes and digestion conditions that are recommended by the protocol presented here.

Protease	Specificity	Expected missed cleavages	pH	Enzyme/protein (wt/wt)	Temp. (°C)	Hours	Recommendations
C-terminal cleavage							
Chymotrypsin	F, Y, L, W, M	0–4	8	1/75	25	12	Dilute urea concentration to <2 M
LysC	K	0–2	8	1/75	37	12	
GluC	E (D)[a]	0–3 (0–4)[b]	8	1/75	25	12	Add 20 mM methylamine when applying urea. Dilute the urea concentration to <2 M
ArgC	R (K)[c]	0–2 (0–3)[b]	8	1/75	37	12	Add 8.5 mM $CaCl_2$, 5 mM DTT and 0.5 mM EDTA. Add 20 mM methylamine when applying urea. Dilute urea to <2M
Trypsin	R, K	0–2	8	1/75	37	12	Dilute the urea concentration to <2 M
N-terminal cleavage							
AspN	D (E)[d]	0–3 (0–4)[b]	8	1/75	37	12	Add 20 mM methylamine when applying urea. Dilute the urea concentration to <2 M. Do not use metal chelators
LysN	K	0–2	8	1/75	37	12	Dilute the urea concentration to below 6 M. Do not use metal chelators

[a]Asp residues are also cleaved but at a much lower rate than Glu residues. [b]Expected number of missed cleavages when using relaxed specificity settings during database search. [c]Lys residues are also cleaved but at a lower rate than Arg residues. [d]Glu residues are also cleaved but at a much lower rate than Asp residues. However, in both cases the cleavage rate for the secondary amino acid might increase depending on the buffer used, incubation time and amount of protease.

Lysis buffer Lysis buffer is 8 M urea (4.8 g per 10 ml) in 50 mM NH_4HCO_3, pH 8 (40 mg per 10 ml), containing 1 tablet of Complete Mini EDTA-free protease inhibitor cocktail per 10 ml of lysis buffer. ▲ **CRITICAL** First dissolve 4.8 g of urea with a lower volume of NH_4HCO_3 solution, and then bring it to a final volume of 10 ml. ▲ **CRITICAL** Freshly prepare all the reagents, and add the inhibitor tablets just before use. Keep the lysis buffer on ice.

Protein reduction reagent: DTT solution Prepare DTT stock solution by dissolving DTT in water to a final concentration of 0.25 M. ▲ **CRITICAL** DTT is susceptible to oxidation, and it should be freshly prepared.

Protein alkylation reagent: IAA solution Prepare IAA stock solution by dissolving IAA in water to a final concentration of 0.25 M. ▲ **CRITICAL** IAA is sensitive to light, and it should be freshly prepared and kept in the dark. Make sure that the pH solution is above 7.5 to avoid alkylation of lysine and histidine[37].

Chymotrypsin Dissolve lyophilized chymotrypsin sequencing grade in 50 mM NH_4HCO_3, pH 8. Once it is made, the solution is stable at least until the expiration date printed on the label at −80 °C.

LysC Dissolve lyophilized LysC in 50 mM NH_4HCO_3, pH 8. Once it is made, the solution is stable at least until the expiration date printed on the label at −80 °C.

GluC Dissolve lyophilized GluC sequencing grade in 50 mM NH_4HCO_3, pH 8. Once it is made, the solution is stable at least until the expiration date printed on the label at −80 °C.

ArgC Dissolve lyophilized ArgC sequencing grade in 50 mM NH_4HCO_3, pH 8. Once it is made, the solution is stable at least until the expiration date printed on the label at −80 °C.

Trypsin Dissolve lyophilized trypsin sequencing grade in 50 mM NH_4HCO_3, pH 8. Once it is made, the solution is stable at least until the expiration date printed on the label at −80 °C.

AspN Dissolve lyophilized AspN sequencing grade in 50 mM NH_4HCO_3, pH 8. Once it is made, the solution is stable at least until the expiration date printed on the label at −80 °C.

LysN Dissolve lyophilized LysN sequencing grade in 50 mM NH_4HCO_3, pH 8. Once it is made, the solution is stable at least until the expiration date printed on the label at −80 °C. ▲ **CRITICAL** Although it is recommended by the manufacturers to reconstitute proteolytic enzymes in water or acids, we do not experience loss of enzymatic activity when enzyme solutions are prepared in 50 mM NH_4HCO_3, pH 8, and stored at −80 °C long term.

Sep-Pak washing buffer 1 Sep-Pak washing buffer 1 is 100% (vol/vol) ACN. Freshly prepare this buffer on the day of use, and keep it at ambient temperature.

Sep-Pak washing buffer 2 Sep-Pak washing buffer 2 is 0.6% (vol/vol) acetic acid in water. Freshly prepare this buffer on the day of use, and keep it at ambient temperature.

Sep-Pak elution buffer Sep-Pak elution buffer is 80% (vol/vol) ACN and 0.6% (vol/vol) acetic acid in water. Freshly prepare this buffer on the day of use, and keep it at ambient temperature.

Reverse-phase UPLC solvent A Reverse-phase UPLC solvent A is 0.1% (vol/vol) formic acid in water. Mobile phase can be stored at ambient temperature, and it should be replaced every 2 months.

Reverse-phase UPLC solvent B Reverse-phase UPLC solvent B is 80% (vol/vol) ACN and 0.1% (vol/vol) formic acid in water. Mobile phase can be stored at ambient temperature, and it should be replaced every 2 months.

EQUIPMENT SETUP

Mass spectrometry analysis LC-MS/MS is a platform technology for bottom-up proteomics. Typical LC-MS/MS setups in our laboratory feature an Agilent 1290 Infinity UHPLC system connected to a Q-Exactive Plus Orbitrap or Orbitrap Fusion. For rapid sample loading and desalting, the combination of a C18 trap column with an analytical C18 column is preferred. This UHPLC is equipped with a double-frit trapping column (Dr Maisch Reprosil C18, 3 µm, 2 cm × 100 µm) and a single-frit analytical column (Agilent Poroshell 120 EC-C18, 2.7 µm, 50 cm × 75 µm), both packed in-house and configured in a vented column setup[38]. Injected samples are loaded onto the trapping column with a flow rate of 5 µl/min for 10 min with RP solvent A, whereas gradient elution is performed at a column flow rate of ~300 nl/min (split flow from 0.2 ml/min). The column effluent is directly introduced into the NSI source via a coated fused silica emitter (360 µm outer diameter (o.d.), 20 µm inner diameter (i.d.), 10 µm tip i.d.; constructed in-house). Peptide ions are selected on the basis of signal intensity (in a data-dependent mode) for fragmentation using HCD. In this work, peptides are chromatographically separated using 45-min or 90-min LC gradients (**Table 3**).

PROCEDURE

Preparation of digests and quality control

1| To prepare digests for standard BSA protein using the selected proteolytic enzyme, refer to option A. To prepare and digest *E. coli* cell lysate, see option B. In addition, it is also necessary to set up a trypsin digest for both BSA protein and cell lysate by following both options A and B. These digests are used to perform quality control (QC) of the LC-MS/MS system in the next section.

(A) Standard BSA protein ● TIMING ~1 d

(i) Prepare BSA solution at 3.33 mg/ml in 2 M urea and 50 mM NH_4HCO_3, pH 8.0. Add DTT from a 0.25 M stock to obtain a final concentration of 8 mM, and incubate the mixture for 15 min at 50 °C with gentle agitation.

(ii) Bring the protein solution to room temperature (22 ± 3 °C) and add IAA to a 16 mM final concentration. Incubate the mixture at room temperature for 15 min in the dark.

(iii) Add DTT from 0.25 M stock to obtain a final concentration of 8 mM.
▲ **CRITICAL STEP** This step is recommended to prevent overalkylation.

(iv) Dilute the protein solution with 50 mM NH_4HCO_3 to reduce the urea concentration to less than 0.6 M.
▲ **CRITICAL STEP** Please refer to **Table 2** and/or consult the vendor's datasheets for particular enzymatic resistances to urea (**Supplementary Table 1**).

(v) Add the selected proteolytic enzyme at the recommended enzyme-to-protein ratio (**Table 2**) for 12 h at 30–37 °C.
▲ **CRITICAL STEP** A trypsin digest is used as a control sample for evaluating nano-LC-MS/MS systems on the retention time (**Supplementary Table 2**) and fragmentation patterns of frequently observed BSA peptides.

(vi) Quench the digestion by acidification with TFA to 1% (vol/vol).
! **CAUTION** TFA solutions and TFA vapors are toxic; prepare solutions in a fume hood.

(B) *E. coli* cell lysate ● TIMING ~2 d

(i) Add 2 ml of lysis buffer to the collected cells (7×10^{11} cells) after washing them with ice-cold PBS, and lyse them with sonication. Sonicate the lysate three times for 1 min each with at least 1-min rest on ice between each pulse.
▲ **CRITICAL STEP** Protease inhibitors in ice-cold lysis buffer are required in order to minimize undesirable protein degradation by endogenous proteases. For phosphoproteomic studies, besides protease inhibitors, include phosphatase inhibitors in the lysis buffer. It is recommended that these inhibitors be added to buffers just before use.

(ii) Remove cell debris via centrifugation at 20,000*g* for 15 min at 4 °C.

(iii) Perform a protein assay to determine the protein concentration. By using the amount of cells suggested here, a protein concentration of ~4 mg/ml can be expected.
▲ **CRITICAL STEP** For storage, freeze the lysate using liquid nitrogen, and then store it at –80 °C for few months.

TABLE 3 | LC and MS parameters that were used during the 45-min or 90-min methods in Orbitrap Q-Exactive Plus and Orbitrap Fusion.

LC-MS parameters	
Time interval (min)	**LC gradient (% B)**
0–10	0–13
10–30 (10–75)	13–44
30–33 (75–78)	44–100
33–34 (78–79)	100–100
34–35 (79–80)	100–0
35–45 (80–90)	0–0
MS parameters	
Polarity	Positive
MS1	
Microscans	1
Resolution	35,000 (60,000)
Automatic gain control target	3e6 ion counts (4e5)
Maximum ion time	250 ms (50)
Scan range	375–1600 *m/z* (375–1,500)
dd-MS2	
Microscans	1
Automatic gain control target	5e4 ion counts (1e4)
Maximum ion time	120 ms (35)
Loop count	10 (Top3s)
Isolation window	1.5 *m/z* (1.6)
Fixed first mass	180 *m/z* (120)
Normalized collision energy	25 (35)
dd settings	
Underfill ratio	1%
Charge exclusion	Unassigned, 1
Peptide match	Preferred
Exclude isotopes	On
Dynamic exclusion	6s (12)

Although the methods are named 45 min and 90 min, the indicated times refer to total analysis time rather than the actual gradient time, which is 23 and 68 min, respectively. In parentheses are included the LC and MS parameters for the 90-min gradient.

(iv) Reduce 1 ml of lysate (2 mg in total) by adding 25 µl of DTT from 0.25 M stock solution to a final concentration of 4 mM, and then incubate the mixture for 15 min at 50 °C with gentle agitation.

(v) Bring the protein solution to room temperature and add IAA to obtain a final concentration of 8 mM. Incubate the mixture at room temperature for 15 min in the dark.

(vi) Add DTT to a final concentration of 4 mM to quench unreacted IAA.

▲ **CRITICAL STEP** This step is recommended to prevent overalkylation.

(vii) Dilute the sample solution with 50 mM NH_4HCO_3 to reduce urea to a suitable concentration for optimal digestion (**Table 2**).

▲ **CRITICAL STEP** To ensure optimal balance between enzymatic activities and protein solubility, the degree of dilution of the lysis buffer may differ depending on enzymes. For example, LysC and LysN retain proteolytic activity in 6–8 M urea, and therefore dilution is not necessary. As for other proteases—i.e., chymotrypsin, GluC, AspN and ArgC—lowering the concentration of urea to less than 2 M is necessary to ensure optimal proteolytic performance. Furthermore, for these urea-sensitive enzymes, the addition of methylamine is advised to counteract urea. For metalloproteases such as AspN and LysN, it is strongly advised to avoid the use of chelating agents such as EDTA, as these will inhibit enzymatic activity. Other enzyme-specific conditions such as incubation temperatures, pH and buffer components are clearly listed in **Table 2** and **Supplementary Table 1**.

(viii) Add the selected proteolytic enzyme at the recommended enzyme-to-protein (wt/wt) ratio (**Table 2**) for 12 h at the recommended temperature.

▲ **CRITICAL STEP** A tryptic digest of *E. coli* lysate is used for benchmarking nano-LC-MS/MS systems at a proteomics scale using HCD fragmentation, so as to evaluate instrument performance from time to time.

▲ **CRITICAL STEP** Enzyme-to-protein ratio, incubation times and temperatures have been independently optimized for each of the enzymes used here. For a list of these parameters, please refer to **Table 2**. For the availability and source origin of these enzymes from different vendors, please refer to **Supplementary Table 1**.

(ix) Quench the digestion by acidification with TFA to 1% (vol/vol).

! **CAUTION** TFA solutions and TFA vapors are toxic; prepare the solutions in a fume hood.

(x) Centrifuge the mixture at 2,500*g* for 5 min at room temperature, and remove the pellet.

? **TROUBLESHOOTING**

(xi) Condition Sep-Pak C18 cartridge with 2 ml of washing buffer 1, and then equilibrate the mixture with 2 × 1 ml of washing buffer 2.

(xii) Load the peptide digests into the Sep-Pak C18 cartridge.

▲ **CRITICAL STEP** To ensure optimal loading, maintain a slow flow rate, and do not apply too high a pressure to the vacuum scaffold. High pressure may collapse the collection tubing, resulting in blockade. As previously reported[39,40], we choose the cartridge size based on sample input amount and do not allow the cartridges to run dry.

▲ **CRITICAL STEP** It is always useful to collect the flow-through in case it is necessary to repeat the desalting process or for future analysis.

(xiii) Wash the solid-phase extraction columns with 2 × 1 ml of washing buffer 2.

(xiv) Elute the desalted peptides with 2 × 250 µl of elution buffer.

▲ **CRITICAL STEP** To maximally recover the bound peptides, the elution step should take at least 10 min.

▲ **CRITICAL STEP** It is recommended to take a small aliquot of desalted digests and analyze it by LC-MS/MS for sample quality control.

(xv) Lyophilize the desalted peptides with vacuum centrifugation to almost dryness.

▲ **CRITICAL STEP** Care must be taken to avoid complete dryness and thus sample loss.

■ **PAUSE POINT** Sample can be stored at –80 °C for several months until LC-MS/MS analysis.

QC of the LC-MS/MS system using tryptic digests ● **TIMING** ~5 h

2| Evaluate the LC-MS/MS setup using a 20 fmol BSA tryptic digestion (from Step 1A).

▲ **CRITICAL STEP** The analysis of BSA is to test the chromatographic properties of the peptides including separation, peak width and intensity, as well as elution time (**Supplementary Table 2**). Further, MS performance regarding sensitivity and peptide fragmentation is also monitored.

3| Check the retention times and signal intensities of peptides at *m/z* 488.53, 722.32 and 582.31, which should elute in this order from a C18 column with optimal signal intensities, as described (**Supplementary Table 2**). These values depend on the quality of the digest and the chromatographic columns, the ionization conditions and mass analyzers for the LC-MS/MS system in question.

▲ **CRITICAL STEP** The values reported here are specific to the LC-MS/MS specifications in our laboratory. They need to be adjusted for individual setups, although the general principles apply.

? **TROUBLESHOOTING**

4| When the BSA QC run has met the specifications determined by individual laboratories, evaluate the LC-MS/MS systems with a more complex sample, such as an *E. coli* tryptic digest. Inject an appropriate volume of sample of *E. coli* tryptic digest (from Step 1B) corresponding to a total amount of 50 ng using a longer gradient—e.g., 90 min.

5| In our laboratory, raw data are processed using Proteome Discoverer (version 2.0 or higher, Thermo Fisher). All MS/MS spectra are searched with the MASCOT search engine against an *E. coli* SwissProt database. Validate the PSMs using Percolator (through Proteome Discoverer) on the basis of *q* values at a 1% FDR. The numbers of PSMs and unique peptides obtained are then compared with an average reference number accumulated over time. When the *E. coli* QC run has met the specifications determined by individual laboratories, actual samples can be run on selected LC-MS/MS systems.
? TROUBLESHOOTING

LC-MS/MS analysis ● TIMING variable

6| Dilute or resuspend the dried peptides in 10% (vol/vol) FA and inject an appropriate amount sample into the LC-MS/MS system (10–100 fmol for a single protein and 0.5–1 µg for complex samples). For highly complex samples without prefractionation, longer LC gradients and replicate analyses are generally required to increase the coverage of the proteome (e.g., a 90-min run). A shorter gradient is applicable for the enriched sample from less-abundant fractions to increase the sensitivity (e.g., a 45-min run). For a detailed description of the LC and MS parameters used here, please refer to **Table 3**.

Computational proteomics analysis ● TIMING variable

7| In our laboratory, raw data are mostly processed using Proteome Discoverer (version 2.0 or higher), although other software suites are available and equally applicable. All MS/MS spectra are searched with the MASCOT search engine against a *Bovine* (version 2015_04, 5,991 sequences including common contaminants) or *E. coli* strain K12 SwissProt database (version 2015_07, 4,433 sequences). Regardless of the software used, set the enzyme specificity and number of missed cleavages according to the protease (please refer to **Table 2** for the proteases used in this work). Set carbamidomethylation of cysteines as a fixed modification and oxidation of methionines and protein N-terminal acetylation as variable modifications. Search the precursor ion mass tolerances at 50 p.p.m. and the product ion mass tolerance at 0.6 Da for ion trap readout or 0.05 Da for Orbitrap readout. Besides the enzyme specificity rules, the same search settings apply for enzyme-specific or nonspecific searches. The latter type of analysis is particularly informative for enzymes with unknown specificity or if you are interested in evaluating enzymatic performance.
▲ CRITICAL STEP Commercially available proteases from different vendors may possess differential specificities and digestion efficiencies. In addition, some proteases (i.e., AspN, GluC and ArgC) show a relaxed specificity under particular conditions, such as the composition of the buffer, the incubation time and the amount of protease used. On the basis of literature and our own experience, we find that chymotrypsin cleaves C-terminal to Phe, Leu, Tyr, Thr and Met, although the enzymatic rules for chymotrypsin for most search engines omit Met. As for LysN, besides cleaving N-terminally to Lys, we and others have also observed cleavage at Arg and a lower frequency of cleavages N-terminally to Ser and Ala[41,42]. Please refer to **Table 2** for recommendations for protease specificity and expected missed cleavages to be used during enzyme-specific database searches.

8| Validate the PSMs (through Proteome Discoverer) at a 1% FDR using the target-decoy strategy for low-complexity samples (e.g., BSA digest) and Percolator on the basis of *q* values for medium- and high-complexity digests (e.g., *E. coli*). For MASCOT searches, set the peptide score to 20 and peptide confidence to high.
? TROUBLESHOOTING

? TROUBLESHOOTING
Troubleshooting advice can be found in **Table 4**.

TABLE 4 | Troubleshooting table.

Steps	Problem	Possible reason	Solution
1B(x)	Large pellet formed after centrifugation of acidified digests	The lysate is too concentrated	Dilute the lysate to a protein concentration of ~1 µg/µl
		Insufficient digestion	Check the protease activity and ensure that the enzyme has not lost its activity because of storage in suboptimal conditions by analyzing the percentage of missed cleavages after a nonspecific database search. Follow the instructions for optimal enzymatic digestion such as keeping the urea concentration below 2 M and pH between 7.5 and 8.5

(continued)

TABLE 4 | Troubleshooting table (continued).

Steps	Problem	Possible reason	Solution
3	BSA signal intensity is less than acceptable	Electrospray is unstable, signal is weak	Clean the electrospray tip with ethanol, or replace it with a new tip. Check for leaks or blocks in the nanoLC system If necessary, clean the transfer capillary in the ESI source
		Presence of contaminants suppressing peptide signals	Repeat a few BSA runs, flush the column with running buffer B or change the column
		Problems with chromatography	Please refer to the published protocols[26,47,48]
	Missing peak m/z 488.53 (TCVADESHAGCEK peptide precursor)	Oxidized to 493.87	Use a fresh aliquot
		C18 trap column degradation	Replace it with a new column
5, 8	Fewer identifications than expected	Instrument is out of calibration	Search data with a broader MS1 tolerance such as 50 p.p.m. It is also possible to recalibrate the data based on available scripts. Calibrate the instrument
		Fragmentation is poor	Check the level of collision gas if replacement is needed. For ETD, check electron transfer reagent supply
		Desalting did not work	Analyze the flow-through. Ensure that the digest was properly acidified before loading onto the cartridge
		Unexpected peptide modifications that you did not include in the database searches	Avoid keeping your sample in formic acid for extended periods of time, as this can lead to unwanted formylation. Over-alkylation can be minimized by performing alkylation at room temperature for 30 min and by quenching the reaction with DTT
		Suboptimal database search settings (AspN)	Repeat database search with different number of missed cleavages Repeat database search including or excluding the secondary cleavage site at E residues
		Suboptimal database search settings (GluC)	Repeat database search with different number of missed cleavages Repeat database search including or excluding the secondary cleavage site at D residues
		Suboptimal data-base search settings (Chymotrypsin)	Repeat database search with different number of missed cleavages Repeat database search including or excluding M residues as cleavage sites
		Suboptimal database search settings (LysN)	Repeat database search including A and R, the nonspecific cleavage sites, in the specificity settings. However, the efficiency at A and R might be low, and thus higher missed cleavages have to be allowed
		PSMs scores are low	Highly charged precursors may generate fragments that carry charges >2. Some search engines cannot handle these high-charged fragments. Choose an alternative search engine, or apply high-resolution MS2 such as TOF or Orbitrap so that the high-charged fragments can be deconvoluted to counter the limitation imposed by the search engine algorithms

(continued)

TABLE 4 | Troubleshooting table (continued).

Steps	Problem	Possible reason	Solution
		Insufficient digestion	Check the protease activity and ensure that the enzyme has not lost its activity because of storage in suboptimal conditions by determining the percentage of missed cleavages. Follow the instructions for enzymatic digestion such as keeping the urea concentration below 2 M and pH between 7.5 and 8.5 Remove/avoid protease inhibitors that might hamper the activity of the used protease such as EDTA for AspN and LysN, PMSF for chymotrypsin and LysC and E-64 for ArgC
		Less specific digestion	Perform a nonspecific database search and look for increased frequencies of secondary cleavages. If so, then you might consider extending your specificity rules during enzyme-specific database searches
		Too high incubation temperature used	Decrease the incubation temperature
		Too little enzyme used	Increase the amount of enzyme for digestion or incubate for longer time

● TIMING

Step 1A, preparation of digests of standard BSA protein: ~1 d
Step 1B, preparation of digests of cell lysate: ~2 d
Steps 2–5, QC of the LC-MS/MS system using tryptic digests: ~5 h
Step 6, LC-MS/MS analysis: variable
Steps 7 and 8, computational proteomics analysis: variable

ANTICIPATED RESULTS

In this protocol, we describe the recommended conditions to generate peptides by in-solution protein digestion for shotgun proteomics using six different proteases, with no restrictions on the sample origin. The in-solution digestion method has been chosen, as it is the most commonly used procedure in shotgun proteomics[7]; nevertheless, certain proteases can also be used with other popular strategies, such as in-gel[43] or on-filter digestion[44]. Importantly, when applying LysC, AspN, GluC and ArgC for in-gel digestion, we and others have noticed that the protease efficiency tends to decrease, probably because of the lower ability of the higher-molecular-weight enzymes to permeate the polyacrylamide gel matrix[43].

In our laboratory, BSA digestion is used as a first-line control for evaluating the efficiency of a protease chosen for proteomics experiments. **Figure 1** shows the results of BSA digestion based on our digestion protocol. As can be seen from the analysis of a very low amount of protein material (20 fmol injection), many different peptides were identified from each proteolytic digest, thus generating a cumulative sequence coverage of 94%.

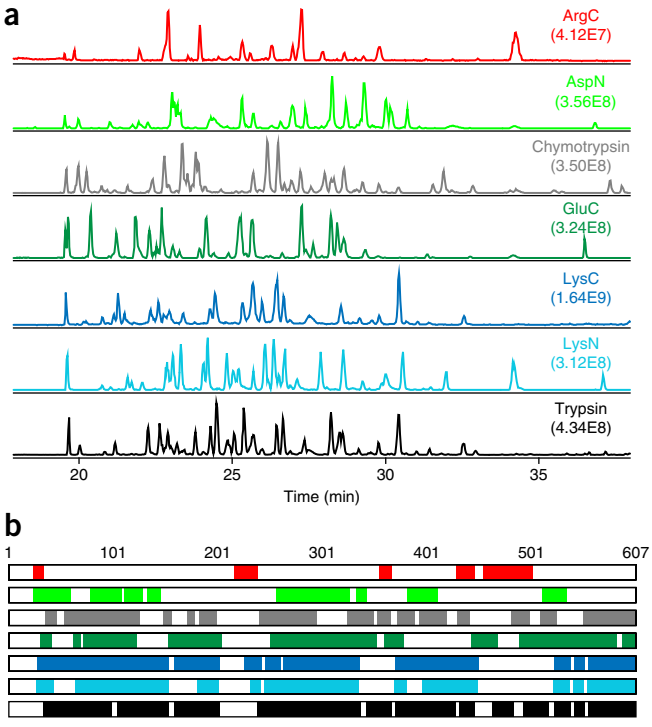

Figure 1 | LC-MS analysis of 20 fmol of BSA digests. (**a**) LC-MS chromatograms acquired in the analysis of 20 fmol of a BSA digest. For each digest, the normalized level of the base peak chromatogram is reported between brackets. (**b**) Graphical representation of the BSA sequence coverage. Filled sections show the relative portion of the entire sequence that was measured and used to identify the protein. The individual sequence coverages are as follows: ArgC: 18.3%, AspN: 38.1%, chymotrypsin: 57.8%, GluC: 61.9%, LysC: 70.8%, LysN: 60.8% and trypsin: 78.4%.

To confirm the efficient digestion of BSA by each protease, we benchmarked the number of identified peptides (**Supplementary Table 3**) against the theoretical number of proteotypic peptide sequences obtained by *in silico* digestion (**Table 5**). On average, 70% of the theoretical peptides overlap with at least one experimentally assigned peptide sequence, including that of the ArgC digestion, where 8 of the 13 theoretical peptides were successfully matched. These results are in line with what has been observed for trypsin digestion—i.e., variable amenabilities of each peptide to MS analysis—mainly as the result of their differential physicochemical properties[45].

Figure 2a shows the results of a more complex mixture, 400 ng of *E. coli* digest, and illustrates the feasibility of high-throughput proteomics with each protease.

To test the reproducibility between LC-MS/MS runs, we performed the analysis in technical triplicates. The number of MS/MS scans, PSMs, unique peptides and protein groups for all data sets are presented in **Supplementary Table 4**.

TABLE 5 | BSA peptide identifications in each of the different protease digestions.

Protease	Identified Peptides[a]	Unique peptides[b]	Proteotypic peptides[c]	Matched peptides[d]
ArgC	14	12	13	8
AspN	24	24	26	20
Chymotrypsin	50	48	37	24
GluC	46	42	43	31
LysC	43	41	43	34
LysN	42	40	43	30
Trypsin	62	59	45	37

[a]Nonredundant PSMs. [b]Unique peptide sequences. [c]Proteotypic peptides obtained by *in silico* digestion allowing zero missed cleavages. [d]Matched proteotypic peptides, including missed cleavages.

On average, more than 65,000 MS/MS scans were acquired by the Orbitrap Fusion per LC-MS/MS analysis, and upon peptide to spectrum matching this translated to an average of 10,544 PSMs and 6,120 unique peptides at 1% FDR.

Notably, although similar numbers of MS/MS events were acquired for each proteolytic digest, an obvious discrepancy lies in the numbers of peptides identified. These differing identification rates among the proteases is in agreement with what has been reported by us and others[14,23,26], and it can probably be attributed to the bias of the search engine toward the different proteases, and/or to the inferior fragmentation of their proteolytic peptides by the chosen fragmentation method (i.e., HCD).

Next, we evaluated the number of identified proteins, as well as sequence coverage for (i) each proteolytic data set and (ii) all data sets combined together. As illustrated in **Figure 2b**, triplicate analysis of any single protease digest results in an average of ~1,500 protein identifications (**Supplementary Table 4**). As additional proteases are included, the mean number of identified proteins increases by an average of 8%. Considering protein identifications from all six data sets, this number increases by 670 proteins to 2,158 (45% increase), which indicates a high complementarity in the multiprotease approach. These results are likely to change substantially depending on the types of sample. Application of this protocol to organisms of higher complexity than *E. coli* (e.g., human cell lines) will lead to an even higher increase in protein identifications.

The in-parallel application of multiple proteases for digestion not only increases protein identification but also the protein sequence coverage. This enables the identification of more post-translational modifications, single-nucleotide substitutions and post-transcriptional editing events[14,23,36]. **Figure 2c** shows the impact of additional proteases on the sequence coverage. Our analysis reveals that the mean number of unique amino acids sequenced in each of the six digest is 97,622. Again, aggregation of all data sets results in an increase of 210,982 additional amino acids for a total of 308,605, reflecting a 216% gain, going from using 1 to 6 proteases.

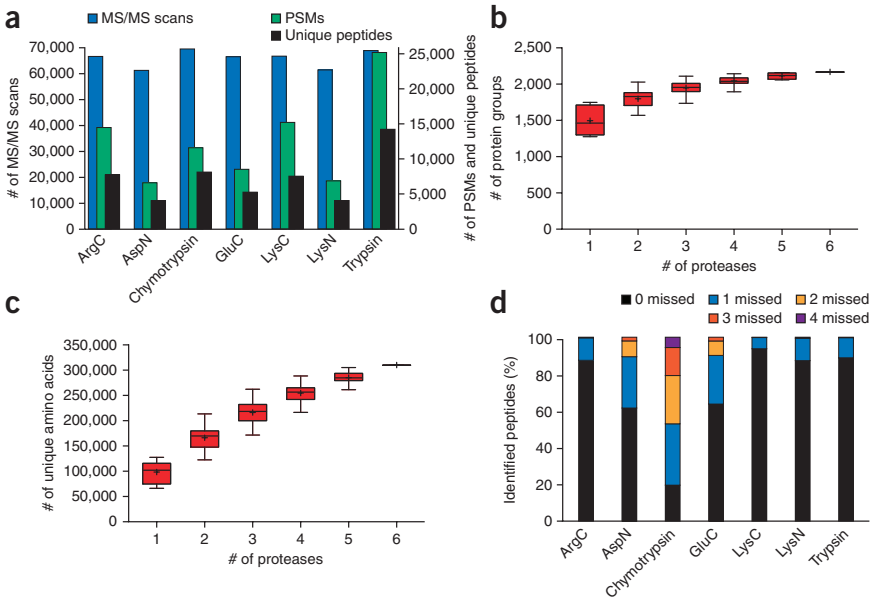

Figure 2 | LC-MS analysis of *E. coli* lysate digests. (**a**) Scan statistics for each of the six protease digests and comparison with trypsin. The number of MS/MS events undertaken, annotated on the left *y* axis, is nearly identical for each protease. The number of PSMs and unique peptides identified (using the right *y* axis) varies substantially using different proteases. (**b**,**c**) The number of (**b**) proteins and (**c**) nonredundant amino acids identified covering the whole proteome when single or multiple enzyme data sets are combined. (**d**) Proportion of peptides identified in each of the different *E. coli* protease digests carrying either 0, 1, 2, 3 or 4 missed cleavages.

Another major factor to consider is the efficiency of proteolytic digestion. This can be investigated by analyzing the specificity and performance of each protease.

Determination of the number of missed cleavages that occurred during each of the proteolytic digestions and evaluation of their sequence context is a valid approach to benchmark enzyme performance. An ideal protease should have high catalytic constants and would not be inhibited by residues at the prime and nonprime sides of the cleavage bond. This would enable the use of stringent search criteria for database searches and thus rule out an increased FDR due to missed cleavages. Unfortunately, such a protease does not exist, and the analysis of the peptides identified in this work reveals notable differences between the six proteases. As shown in **Figure 2d**, for ArgC, LysC and LysN, the vast majority of the identified peptides do not contain any interfering uncleaved residues, whereas AspN, GluC and, especially, chymotrypsin tend to be less efficient in processing all the possible cleavage sites. Further analysis on the composition of the amino acid sequences surrounding the missed cleavage sites[46] provides a closer insight into the effects of specific amino acids on protease specificity (**Fig. 3a**). For ArgC, acidic residues at positions −1, +1 or +2 selectively hamper cleavage. Similar inhibition, but to a lesser extent, is seen in the chymotrypsin data set, in which most of the missed cleavages are observed on leucine, one of the five preferred cleavage residues. In addition, for LysC and LysN, it is well known that Pro or Lys prevent cleavage frequently when located adjacent to the cleavable site[35]. For AspN and GluC, there was no prominent over-representation of particular residues at the neighboring positions of the missed sites, except for their somewhat lower efficiency in cutting D-D and E-E bonds, respectively, which indicates that the efficiency of these enzymes is not biased by the extended sequence context, like for instance LysN and ArgC.

Moreover, we analyzed the degree of specificity that we can expect under these conditions for each of the tested enzymes. For this, we re-analyzed the *E. coli* data using nonspecific search settings (for scan statistics and identifications see **Supplementary Table 5**). As shown in **Figure 3b**, the most frequent cleavage events were in accordance with the strict specificity rules (**Table 2**). This illustrates that using these rules one can map 90% of the peptides within each nontryptic digestion. By including the secondary cleavages that occur with low frequencies, it is possible to increase peptide identifications albeit with the risk of introducing higher FDR because of an increased search space. For this reason, we recommend performing database searches using both the strict and relaxed specificity rules, so that the best settings can finally be selected.

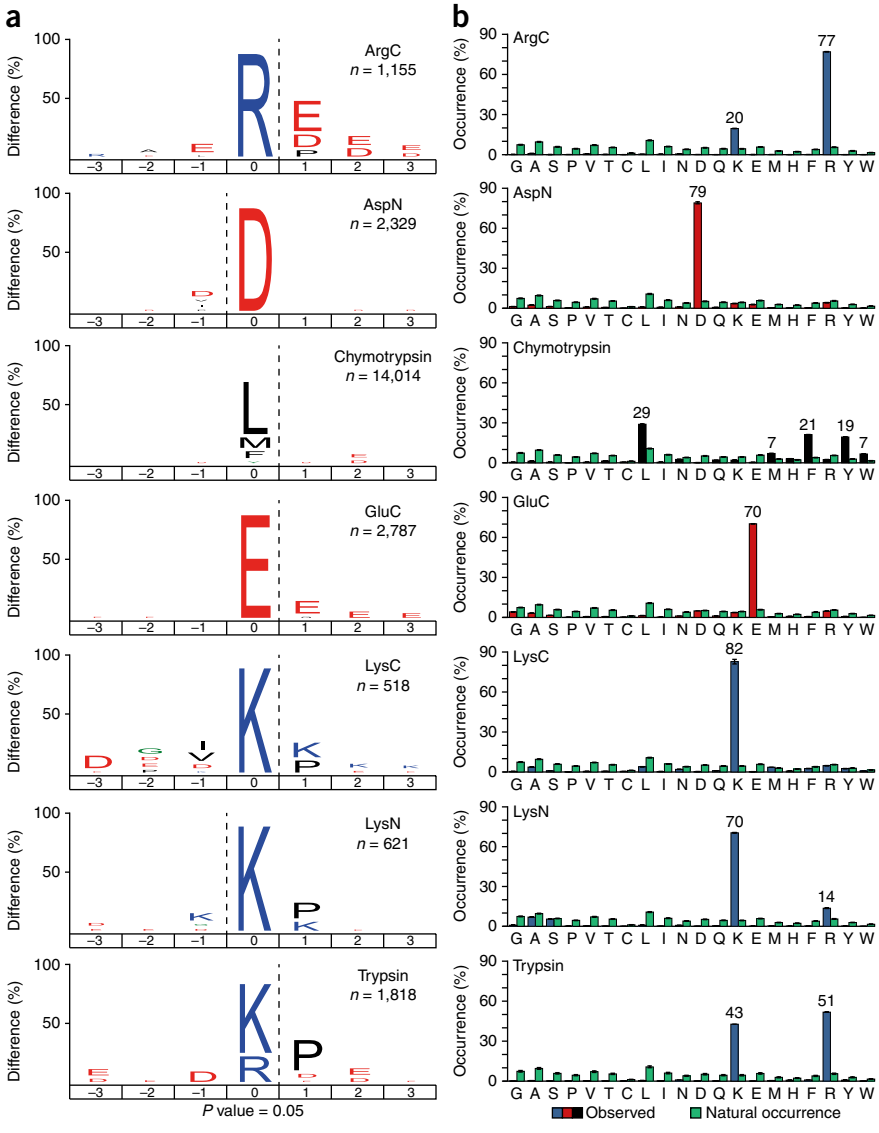

Figure 3 | Specificity and trend for missed cleavages for each of the six enzymes using the protocol presented here for the digestion of an *E. coli* lysate. (**a**) IceLogos of the peptide sequences that contained missed cleavages (position zero) for each of the used enzymes. Dashed line indicates the missed cleavage of peptide bond. Data in **a** were derived from strict enzyme-specific searches using the rules described in **Table 2**. (**b**) Experimental cleavage frequency of residues in *E. coli* proteins as derived from nonspecific MASCOT searches of the SwissProt database. Unique N- and C-terminal flanking regions of the identified peptides were used for the calculation of the residue frequencies at the cleavage site (blue bars), whereas green bars represent the natural occurrence of the amino acids as the frequency mean and s.d. within the reference *E. coli* set. Trypsin data are included for reference.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS This work has been supported by the Netherlands Proteomics Centre, the Netherlands Organization for Scientific Research (NWO) supporting the Roadmap embedded large-scale proteomics facility *Proteins@Work* (project 184.032.201) and by the PRIME-XS project grant agreement number 262067 supported by the European Community's Seventh Framework Programme (FP7/2007-2013) to AJRH. LT was supported by EMBO with a long-term fellowship (ALTF 776-2013).

AUTHOR CONTRIBUTIONS A.J.R.H. conceived the idea for this protocol. P.G. and L.T. designed and performed the experiments and analyzed the data. All authors wrote the manuscript and discussed the experimental results.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

1. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
2. de Godoy, L.M.F. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).
3. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
4. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
5. Link, A.J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682 (1999).
6. Wolters, D.A., Washburn, M.P. & Yates, J.R. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683–5690 (2001).
7. Altelaar, A.F.M., Munoz, J. & Heck, A.J.R. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14**, 35–48 (2013).
8. Yates, J.R., Ruse, C.I. & Nakorchevsky, A. Proteomics by mass spectrometry: approaches, advances, and applications. *Annu. Rev. Biomed. Eng.* **11**, 49–79 (2009).
9. Bensimon, A., Heck, A.J.R. & Aebersold, R. Mass spectrometry-based proteomics and network biology. *Annu. Rev. Biochem.* **81**, 379–405 (2012).
10. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
11. Walther, T.C. & Mann, M. Mass spectrometry-based proteomics in cell biology. *J. Cell Biol.* **190**, 491–500 (2010).
12. Tsiatsiani, L. & Heck, A.J.R. Proteomics beyond trypsin. *FEBS J.* **282**, 2612–2626 (2015).
13. Guo, X., Trudgian, D.C., Lemoff, A., Yadavalli, S. & Mirzaei, H. Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. *Mol. Cell. Proteomics* **13**, 1573–1584 (2014).
14. Swaney, D.L., Wenger, C.D. & Coon, J.J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* **9**, 1323–1329 (2010).
15. Meyer, J.G. *et al.* Expanding proteome coverage with orthogonal-specificity α-lytic proteases. *Mol. Cell. Proteomics* **13**, 823–835 (2014).
16. López-Ferrer, D. *et al.* Pressurized pepsin digestion in proteomics: an automatable alternative to trypsin for integrated top-down bottom-up proteomics. *Mol. Cell. Proteomics* **10**, M110.001479 (2011).
17. Bian, Y. *et al.* Improve the coverage for the analysis of phosphoproteome of HeLa cells by a tandem digestion approach. *J. Proteome Res.* **11**, 2828–2837 (2012).
18. Choudhary, G., Wu, S.-L., Shieh, P. & Hancock, W.S. Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS. *J. Proteome Res.* **2**, 59–67 (2003).
19. Huesgen, P.F. *et al.* LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification. *Nat. Methods* **12**, 55–58 (2015).
20. Peng, M. *et al.* Protease bias in absolute protein quantitation. *Nat. Methods* **9**, 524–525 (2012).
21. Aye, T.T. *et al.* Proteome-wide protein concentrations in the human heart. *Mol. Biosyst.* **6**, 1917–1927 (2010).
22. Benevento, M. *et al.* Adenovirus composition, proteolysis, and disassembly studied by in-depth qualitative and quantitative proteomics. *J. Biol. Chem.* **289**, 11421–11430 (2014).
23. Low, T.Y. *et al.* Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep.* **5**, 1469–1478 (2013).
24. Gauci, S. *et al.* Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach. *Anal. Chem.* **81**, 4493–4501 (2009).
25. Mohammed, S. *et al.* Multiplexed proteomics mapping of yeast RNA polymerase II and III allows near-complete sequence coverage and reveals several novel phosphorylation sites. *Anal. Chem.* **80**, 3584–3592 (2008).
26. Richards, A.L. *et al.* One-hour proteome analysis in yeast. *Nat. Protoc.* **10**, 701–714 (2015).
27. Washburn, M.P., Wolters, D. & Yates, J.R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
28. Klammer, A.A. & MacCoss, M.J. Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J. Proteome Res.* **5**, 695–700 (2006).
29. Wu, X., Xiong, E., Wang, W., Scali, M. & Cresti, M. Universal sample preparation method integrating trichloroacetic acid/acetone precipitation with phenol extraction for crop proteomic analysis. *Nat. Protoc.* **9**, 362–374 (2014).
30. Schuchard, M.D. *et al.* Artifactual isoform profile modification following treatment of human plasma or serum with protease inhibitor, monitored by 2-dimensional electrophoresis and mass spectrometry. *Biotechniques* **39**, 239–247 (2005).
31. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
32. Keller, B.O., Sui, J., Young, A.B. & Whittal, R.M. Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* **627**, 71–81 (2008).
33. Good, D.M., Wirtala, M., McAlister, G.C. & Coon, J.J. Performance characteristics of electron transfer dissociation mass spectrometry. *Mol. Cell. Proteomics* **6**, 1942–1951 (2007).
34. Molina, H., Horn, D.M., Tang, N., Mathivanan, S. & Pandey, A. Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc. Natl. Acad. Sci. USA* **104**, 2199–2204 (2007).
35. Gershon, P.D. Cleaved and missed sites for trypsin, lys-C, and lys-N can be predicted with high confidence on the basis of sequence context. *J. Proteome Res.* **13**, 702–709 (2014).
36. Giansanti, P. *et al.* An augmented multiple-protease-based human phosphopeptide atlas. *Cell Rep.* **11**, 1834–43 (2015).
37. Boja, E.S. & Fales, H.M. Overalkylation of a protein digest with iodoacetamide. *Anal. Chem.* **73**, 3576–3582 (2001).
38. Meiring, H.D., van der Heeft, E., ten Hove, G.J. & de Jong, A.P.J.M. Nanoscale LC-MS(n): technical design and applications to peptide and protein analysis. *J. Sep. Sci.* **25**, 557–568 (2002).
39. Udeshi, N.D., Mertins, P., Svinkina, T. & Carr, S.A. Large-scale identification of ubiquitination sites by mass spectrometry. *Nat. Protoc.* **8**, 1950–1960 (2013).
40. Villén, J., Gygi, S.P. & Villen, J. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat. Protoc.* **3**, 1630–1638 (2008).
41. Hohmann, L. *et al.* Proteomic analyses using Grifola frondosa metalloendoprotease Lys-N. *J. Proteome Res.* **8**, 1415–1422 (2009).
42. Taouatas, N., Heck, A.J.R. & Mohammed, S. Evaluation of metalloendopeptidase Lys-N protease performance under different sample handling conditions. *J. Proteome Res.* **9**, 4282–4288 (2010).
43. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–2860 (2006).
44. Wiśniewski, J.R. *et al.* Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–352 (2009).
45. Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25**, 125–131 (2007).
46. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **6**, 786–787 (2009).
47. Köcher, T., Pichler, P., Swart, R. & Mechtler, K. Quality control in LC-MS/MS. *Proteomics* **11**, 1026–1030 (2011).
48. Köcher, T., Pichler, P., Swart, R. & Mechtler, K. Analysis of protein mixtures from whole-cell extracts by single-run nanoLC-MS/MS using ultralong gradients. *Nat. Protoc.* **7**, 882–890 (2012).