MIT Open Access Articles

Optimization of parameters for coverage of low molecular weight proteins

Massachusetts Institute of Technology

ORIGINAL PAPER

Optimization of parameters for coverage of low molecular weight proteins

Stephan A. Müller · Tibor Kohajda · Sven Findeiß ·
Peter F. Stadler · Stefan Washietl · Manolis Kellis ·
Martin von Bergen · Stefan Kalkhof

Abstract Proteins with molecular weights of <25 kDa are involved in major biological processes such as ribosome formation, stress adaption (e.g., temperature reduction) and cell cycle control. Despite their importance, the coverage of smaller proteins in standard proteome studies is rather sparse. Here we investigated biochemical and mass spectrometric parameters that influence coverage and validity of identification. The underrepresentation of low molecular weight (LMW) proteins may be attributed to the low numbers of proteolytic peptides formed by tryptic digestion as well as their tendency to be lost in protein separation and concentration/desalting procedures. In a systematic investigation of the LMW proteome of *Escherichia coli*, a total of 455 LMW proteins (27% of the 1672 listed in the SwissProt protein database) were identified, corresponding to a coverage of 62% of the known cytosolic LMW proteins. Of these proteins, 93 had not yet been functionally classified, and five had not previously been confirmed at the protein level. In this study, the influences of protein extraction (either urea or TFA), proteolytic digestion

S. A. Müller · M. von Bergen · S. Kalkhof (✉)
Department of Proteomics, UFZ,
Helmholtz-Centre for Environmental Research,
Permoserstraße 15,
04318 Leipzig, Germany
e-mail: Stefan.kalkhof@ufz.de

T. Kohajda · M. von Bergen
Department of Metabolomics, UFZ,
Helmholtz-Centre for Environmental Research,
Permoserstraße 15,
04318 Leipzig, Germany

S. Findeiß · P. F. Stadler
Bioinformatics Group, Department of Computer Science, and
Interdisciplinary Center for Bioinformatics, University of Leipzig,
Härtelstraße 16-18,
04107 Leipzig, Germany

P. F. Stadler
Institute for Theoretical Chemistry, University of Vienna,
Währingerstraße 17,
1090 Wien, Austria

P. F. Stadler
Max Planck Institute for Mathematics in the Sciences,
Inselstrasse 22,
04103 Leipzig, Germany

P. F. Stadler
Nomics Group,
Fraunhofer Institute for Cell Therapy and Immunology,
Deutscher Platz 5e,
04103 Leipzig, Germany

P. F. Stadler
Santa Fe Institute,
1399 Hyde Park Rd,
Santa Fe, NM 87501, USA

S. Washietl · M. Kellis
Computer Science and Artificial Intelligence Laboratory,
Broad Institute, Massachusetts Institute of Technology,
Cambridge, MA 02139, USA

(solely, and the combined usage of trypsin and AspN as endoproteases) and protein separation (gel- or non-gel-based) were investigated. Compared to the standard procedure based solely on the use of urea lysis buffer, in-gel separation and tryptic digestion, the complementary use of TFA for extraction or endoprotease AspN for proteolysis permits the identification of an extra 72 (32%) and 51 proteins (23%), respectively. Regarding mass spectrometry analysis with an LTQ Orbitrap mass spectrometer, collision-induced fragmentation (CID and HCD) and electron transfer dissociation using the linear ion trap (IT) or the Orbitrap as the analyzer were compared. IT-CID was found to yield the best identification rate, whereas IT-ETD provided almost comparable results in terms of LMW proteome coverage. The high overlap between the proteins identified with IT-CID and IT-ETD allowed the validation of 75% of the identified proteins using this orthogonal fragmentation technique. Furthermore, a new approach to evaluating and improving the completeness of protein databases that utilizes the program *RNAcode* was introduced and examined.

Abbreviations

LMW	Low molecular weight (below 25 kDa)
CID	Collision-induced dissociation
ET(ca)D	Electron transfer (collision activation) dissociation
FDR	False discovery rate
FTICR MS	Fourier transform ion cyclotron resonance mass spectrometry
GO	Gene Ontology
HCD	Beam-type collision-activated dissociation
LB medium	Lysogeny broth medium
ORF	Open reading frame

Introduction

Escherichia coli (*E. coli*) is a Gram-negative bacterium of the family *Enterobacteriacae*. It is relatively easy to cultivate, fast growing, and allows for feasible genetic manipulation. Due to these characteristics, *E. coli* is omnipresent in molecular biology, biotechnology and gene technology, and it is one of the most intensively studied and best-characterized prokaryotes. Sequencing and analysis of the 4.6 Mb chromosome of the laboratory strain *E. coli* K12 coding for 4411 protein-coding genes was completed in 1997 [1].

In the last two decades, the *E. coli* proteome has been extensively analyzed by 2D gel electrophoresis (2D-GE) initially and then via LC/MS approaches. Besides investigations of numerous biological questions, the *E. coli* proteome has also been used to validate new technologies and methodologies, including sample prefractionation, protein enrichment and separation by 2D-GE or *n*-dimensional chromatography, and protein identification and quantification by MS [2].

The first proteome study was conducted using 2D-GE and resulted in the identification of 381 proteins [3]. By combining 2D-DIGE with biochemical prefractionation and the analysis of stationary and exponential growth phases, it was possible to detect and quantify 3199 protein species, among which 575 unique proteins could be identified [4]. In several gel-free approaches using *n*-dimensional LC for protein [5] or peptide separation [6–9], the number of proteins was successively increased further (Table 1). Most recently, in 2010, Iwasaki and coworkers used 1D-LC/MS/MS with a 350 cm long monolithic silica–C_{18} capillary column and 41 h of LC gradient time to identify 2602 proteins [10]. However, even with all of these different methods, the identification rate for LMW proteins of <25 kDa listed in the SwissProt protein database is usually below 25%, and is significantly lower than the average identification rate (Table 1).

Proteins that are essential in numerous biological functions, especially ribosome formation (e.g., 18 30S ribosomal protein subunits, 34 50S ribosomal protein subunits), transcription regulation, and stress response (cold shock proteins, universal stress proteins) are of LMW. Coverage of those functional proteins in proteomic studies is of great interest in systems biology in order to gain an in-depth understanding of the reactions of bacteria to external stresses [11], adaption to different substrates, and interdependencies in microbial bacterial communities in the new field of metaproteomics [12]. Furthermore, over 500 LMW proteins of *E. coli* are still classified as "functionally uncharacterized" according to the latest GO annotation database [13]. This number is astonishingly high given the limited genome of *E. coli* and the high feasibility of this organism for culturing and genomic manipulation.

Another challenge is the de novo annotation of open reading frames (ORF) coding for small proteins on a genome-wide scale. In the past, computational gene-finding approaches excluded short ORFs with less than 40 or 50 amino acids. For such short ORFs, typical statistical signals in the sequence (ORF length and codon usage) are very weak, resulting in a high false-discovery rate (FDR). Thus, using standard methods with less stringent filters leads to the prediction of thousands of small ORFs, most of which are not likely to be translated [14]. The methods of choice to verify the existence of these small proteins are LC/MS

Table 1 Summary of total and LMW proteins detected in previous studies based on at least [a] four peptides, [b] two peptides, and [c] one peptide per protein

Study	Method	LMW	Complete proteome	LMW (%)	Complete proteome (%)	Reference
Lopez-Campistrous et al. (2005)	2D-PAGE after prefractionation in periplasm, inner membrane, and outer membrane	164	575	10	13	[4]
Geveart et al. (2002)	Diagonal 2D-LC-MS of methionine-containing peptides	187[c]	872[c]	11	20	[6]
Corbin et al. (2003)	1D-LC-MS with and w/o membrane fractionation (4 h per run)	218[a]– 331[c]	404[a]–1147[b]	13[a]-21[c]	26	[7]
Taoka et al. (2004)	2D-LC-MS (16 h per run)	401	1480	24	34	[8]
Ishihama et al. (2008)	More than 200 2D-LC-MS measurements after 1D-gel protein prefractionation	341	1103	20	25	[9]
Iwasaki et al. (2010)	LC-MS with a 3.5 m non-commercially available monolithic column (41 h per run)	737[b]–820[c]	2404[b]– 2602[c]	44[b]-49[c]	60	[10]

approaches. Since these experimental methods are cost and time intensive, in silico methods are still required for efficient genome annotation. Recently, we developed *RNAcode*, a gene prediction program that uses the principle of comparative genomics [15] to detect protein-coding genes in multiple genome alignments [16]. Since *RNAcode* is based on evolutionary signatures, it can detect statistically significant signals—even in short ORFs—as long as sufficient phylogenetic information from related sequences is available. The fact that *RNAcode* is not based on the detection of complete ORFs also makes it applicable to incomplete data, such as fragments of transcriptome studies [17]. Thus, *RNAcode* fills a specific gap in the current repertoire of protein annotation software. To further investigate the applicability and power of RNAcode, we systematically analyzed the LMW of *E. coli* and compared these results with our proteome data.

The variation in the abundances of cytosolic proteins in *E. coli* ranges from less than 200 to more than 10^8 molecules per cell—in other words, more than six orders of magnitude [9]. The low abundances of some proteins certainly hamper their detection, and not all proteins will be expressed at the same time. Aside from these biological reasons for limited coverage, it has been discussed that losses during protein extraction [18], separation and purification [19], as well as the low number of detectable proteotypic peptides formed by proteolysis [19] are responsible for the low identification rate. Taking into account recent improvements in the coverage of LMW proteins, the best study achieved 49% coverage of LMW in *E. coli* (Table 1). It is obvious that there is plenty of scope for improvement. This can in principle be achieved by separation, fractionation or the complementary usage of multiple proteases, or on the LC/MS side. In order to get information on which strategy to start with in this study, key parameters associated with both prefractionation and LC/MS were tested. With respect to prefractionation and biochemical preprocessing, the following parameters were assessed for their influence on coverage: (i) protein extraction buffers, (ii) enrichment and separation, and (iii) enzymatic proteolysis. In terms of LC/MS, the crucial steps of (iv) the fragmentation procedure and (v) MS/MS data analysis were varied and evaluated with respect to identification rate, average sequence coverage, and validation of identifications.

Materials and methods

Cell culture

Cell lysates of *E. coli* strain K12 were analyzed to assess critical parameters for LMW proteome analysis. Analyses were performed in two (gel-based approach) and three (non-gel-based approach) independent biological replicates. Cells were grown in LB medium to stationary phase. Therefore, 1 l of fresh medium was inoculated with 100 ml of a preparatory culture grown under the same conditions. Cells were collected by centrifugation (10 min, 8,000×*g*, 4 °C).

Protein extraction and small protein enrichment

Cell pellets were resuspended in either urea lysis buffer (40 ml, 8 M urea, 10 mM DTT, 1 M NaCl, 10 mM Tris/HCl, pH 8.0) [20] or acidic lysis buffer (40 ml, 0.1% TFA) [21]. Cell disruption was performed by ultrasonification (5 min, 50% duty cycle, Branson Sonifier 250, Emerson, St. Louis, MO, USA). Undissolved material was removed by centrifugation (15 min, 10,000×*g*, 4 °C). High molecular weight proteins were depleted by centrifugation through a filter membrane (molecular weight cut-off: 50 kDa, Pall

Macrosep 50 K, Pall Life Science, Ann Arbor, MI, USA) [22]. The permeate was split into aliquots of 1.2 ml. TFA lysates were equilibrated to neutral pH with NH_4CO_3 (final concentration: 250 mM) and protein disulfide bonds were reduced by adding DTT (final concentration: 10 mM). Cysteines were alkylated by the addition of 2-iodoacetamide (final concentration: 51.5 mM) to both lysates and incubation for 45 min at room temperature in the dark. Proteins were desalted and concentrated by TCA precipitation (final concentration: 20% (w/v), incubation at 4 °C for 16 h, centrifugation at $20,000 \times g$ for 20 min).

Protein separation and protein digestion

For the non-gel approach, one protein pellet of every biological replicate was dissolved in 500 mM NH_4HCO_3 and the protein concentration was measured with a Bradford assay (Bradford Quick Start, Bio-Rad, Hercules, CA, USA) using bovine serum albumin for calibration. Pellets were redissolved in 100 μl 1.6 M urea in NH_4HCO_3 (100 mM). Trypsin (modified porcine trypsin, Sigma–Aldrich, Steinheim, Germany) was dissolved in 50 mM NH_4HCO_3 containing 10% acetonitrile to a concentration of 125 ng/μl. Trypsin solution was added to the dissolved protein pellets with a molecular weight ratio of 1:50 (trypsin:protein). Digestions were performed overnight at 37 °C and stopped by adding formic acid (final concentration: 4%). Digestion solutions were concentrated to 20 μL using vacuum centrifugation and reconstituted by adding 40 μL 1% formic acid.

For the gel separation, protein pellets were redissolved with SDS loading buffer (2% (w/v) SDS, 12% (w/v) glycerol, 120 mM DTT, 0.0024% (w/v) bromophenol blue, 70 mM Tris/HCl) and adjusted to neutral pH by adding 10× cathode buffer solution (1 M Tris, 1 M tricine, 1% (w/v) SDS, pH 8.25). GE was performed according to a modified protocol of Schaegger [23]. In brief, a 20% T, 6% C separation gel was used in combination with a 4% T, 3% C stacking gel. A prestained LMW protein standard (molecular weight range 1.7–42 kDa, multicolor low-range protein ladder, Fermentas, St. Leon-Rot, Germany) was applied as a molecular weight marker. For each experiment, three lanes were loaded with the LMW protein extract, among which one was stained with colloidal Coomassie. Nine gel slices from each of the two unstained lanes were excised in the molecular weight range 1–25 kDa and used for in-gel digestion.

The gel slices were washed twice with water for 10 min and once with NH_4HCO_3 (10 mM). In-gel digestion was performed by adding modified porcine trypsin (100 ng, Sigma–Aldrich) or endoproteinase AspN (100 ng, Sigma–Aldrich) in NH_4HCO_3 (10 mM, 30 μl volume) to the slices.

The digestions were performed overnight at 37 °C and stopped afterwards by adding formic acid (final concentration: 4%). The supernatant and the two gel elution solutions (first elution step: 40% (v/v) acetonitrile; second elution step: 80% (v/v)) were collected and mixed. The combined mixtures were dried using vacuum centrifugation. Peptides were reconstituted in 0.1% formic acid.

Analysis with nano-HPLC/nano-ESI-LTQ Orbitrap MS

LC/MS/MS analysis was performed on a nano-HPLC system (nanoAcquity, Waters, Milford, MA, USA) coupled to an LTQ Orbitrap mass spectrometer. Chromatography was conducted with 0.1% formic acid in solvents A (100% water) and B (100% acetonitrile).

In-solution digestion samples were injected by the autosampler and concentrated on a trapping column (nano-Acquity UPLC column, C18, 180 μm×2 cm, 5 μm, Waters) with water containing 0.1% formic acid at flow rates of 15 μL/min. After 10 min, peptides were eluted onto a separation column (nanoAcquity UPLC column, C18, 75 μm×150 mm, 1.7 μm, Waters). Peptides were eluted over 150 min with a 2–40% solvent B gradient (0 min, 2%; 3 min 2%;10 min, 6%;100 min, 20%; 150 min, 40%).

Scanning of eluted peptide ions was carried out in positive ion mode between m/z 300 and 1500, automatically switching to MS/MS mode for ions exceeding an intensity of 3,000. Precursor ions were dynamically excluded for MS/MS measurements for 3 min. Six runs with different MS/MS measurements were performed per biological sample. CID and ETD fragmentations were carried out with ion detection in the ion trap or the Orbitrap in separate runs. HCD fragmentations were detected in the Orbitrap. Additionally, a method with a decision tree between CID and ETD in the ion trap was performed.

In-gel digestion samples were injected and concentrated on a trapping column in an identical manner to the analysis of in-solution digestions. Peptides were eluted onto a separation column (nanoAcquity UPLC column, C18, 75 μm×250 mm, 1.7 μm, Waters) and separation was done over 30 min with a 2–40% solvent B gradient (0 min, 2%; 2 min 8%; 20 min, 20%; 30 min, 40%). Scanning of eluted peptide ions was carried out in positive ion mode in the range m/z 350–2000, automatically switching to CID-MS/MS mode for ions exceeding an intensity of 2,000. For CID-MS/MS measurements, a dynamic precursor exclusion of 3 min was applied.

Data analysis

Database searching was performed with *Proteome Discoverer* (version 1.0; Thermo Fisher Scientific, San Jose, CA, USA) using the MASCOT (version 2.2; Matrix Science,

London, UK) and SEQUEST (version 1.0.43.0; Thermo Fisher Scientific) algorithms that search through a target and decoy database containing all proteins of *E. coli* strain K12 in the SwissProt protein database. In-gel digestions with trypsin were searched with maximum of one missed cleavage, while two missed cleavages were allowed for in-gel digestion with AspN and in-solution digestions. For trypsin C-terminal cleavage to arginine and lysine, and for endoprotease AspN N-terminal cleavage to aspartic and glutamic acid were considered. MS/MS spectra were grouped with a precursor mass tolerance of 4.0 ppm and a retention time tolerance of 5 min. MASCOT and SEQUEST searched with a parent ion tolerance of 5.0 ppm. Fragment ion mass tolerances were specified as 0.5 Da when fragment ions were detected in the ion trap and 0.05 Da when detection was performed in the Orbitrap. Carbamidomethylation of cysteines was specified in MAS-COT and SEQUEST as a fixed modification, and the oxidation of methionine as a variable modification. Additionally, deamidations of asparagine and glutamine were considered variable modifications for in-solution digestion samples.

SCAFFOLD (version SCAFFOLD_2_06_01_pre3; Proteome Software Inc., Portland, OR, USA) was used to validate MS/MS-based peptide and protein identifications. Peptide and protein identification parameters were adjusted to a false-positive rate of lower than 5% using the target and decoy database. False-positive rates were calculated as described by Elias et al. [24]. Peptide identifications were accepted if they could be established at a probability of greater than 70.0% as specified by the *Peptide Prophet* algorithm [25]. Peptide identifications were accepted by exceeding specific database search engine thresholds. MASCOT identifications required ion scores of greater than 10.0. SEQUEST identifications required deltaCn scores of greater than 0.10 and XCorr scores of greater than 1.7, 2.0, and 2.3 for doubly, triply and quadruply charged peptides. Protein identifications were accepted if they could be established at greater than 95.0% probability and contained at least two identified peptides. Protein probabilities were assigned by the *Protein Prophet* algorithm [26]. Proteins that contained similar peptides and which could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. GO annotations were obtained with STRAP [27] from the EBI GO database (http://www.ebi.ac.uk/GOA/, version 05/07/2010).

ProtStat: protein statistics and peptide predictions

The software *ProtStat* is an in-house tool programmed with C# which calculates protein as well as proteolotytic peptide properties. The program has three different modes:

protein pre-statistics, protein post-statistics and peptide statistics.

For the protein statistics, various data can be obtained for every protein, including molecular weight, protein sequence, GRAVY score, protein database ID, protein description, and a calculation of the p*I* value. p*I* values are calculated using the advanced algorithm suggested by Kozlowski (http://isoelectric.ovh.org/) with a selectable set of amino acid p*K* increments according to EMBOSS, DTASelect, Solomon, Sillero or Rodwell.

The protein pre-statistic allows an in silico simulation of a proteolytic digestion by calculating the number and sequences of proteolytic peptides, the expected possible sequence coverage, and performing a comparison in terms of unique peptides and sequence coverage to other proteolytic digestions (e.g., those using other proteases). In terms of digestion parameters, several specific proteases as well as their combinations and fixed modifications are allowed.

In the protein post-processing mode, the same analysis is possible for a list of identified proteins, and this enables the comparison of experimental and theoretical LC/MS measurements.

The peptide statistics mode allows the calculation of inclusion or exclusion lists based on the results of a theoretical or experimental proteolytic digestion. Therefore, exact *m/z* values in a given *m/z* range were calculated for the charge states 1+ to 4+. Again, fixed protein modifications are taken into account. Additionally, p*I* values of all potential proteolytic peptides for every protein inside a protein FASTA database are calculated.

Prediction of protein coding regions in genome-wide alignments of nucleotide sequences by *RNAcode*

We used the Multiz pipeline [28] to align 54 fully sequenced enterobacteria species from GenBank (Electronic supplementary material Table S1). The alignments were screened using the default parameters of *RNAcode* (software available at http://wash.github.com/rnacode) and a *p*-value cutoff of 0.05. This resulted in 20,528 high-scoring coding segments. Multiple sequence alignments of such a high number of species tend to be fragmented into relatively small blocks. Therefore, high-scoring coding segments in the same reading frame and less than 15 nucleotides apart were combined. This reduced the number of high-scoring coding segments to 6,542.

The SwissProt protein database was downloaded (http://pir.uniprot.org/downloads, May 2010 release). For each registered *E. coli* protein, the ID, the type of evidence, and the amino acid sequence was extracted. In order to compare the *RNAcode* predictions, which are based on nucleotide alignments, with the protein sequences from SwissProt and

our peptide data, we blasted all peptide sequences (TBLASTN, *E*-value 10^{-3} and 98% identity) against the *E. coli* genome. Using this conservative method, 1574 proteins were mapped to 1605 distinct genomic loci.

Results and discussion

General experimental strategy

In this paper, our experiences relating to the large-scale identification of LMW proteins (molecular weights <25 kDa) using gel-based and gel-free approaches are summarized. By combining different methods, a total of 455 LMW proteins of *E. coli* were identified with high certainty (Electronic supplementary material Tables S2 and S3).

As a starting point for optimization, the procedure published in 2007 by Klein et al. [20] was used, as this study reported an identification rate of 35% of the LMW subproteome of *Halobacterium salinarum*. The outline of this study consisted of high molecular weight protein depletion, separation by 1D-GE using a modified protocol according to Schaegger [23], and ESI-LC/MS³ analysis with FTICR MS.

Here we vary this strategy stepwise in order to estimate the influence of the critical parameters in (i) protein extraction, (ii) enrichment and separation, (iii) proteolysis, (iv) MS and MS/MS analysis, and (v) protein identification (Fig. 1).
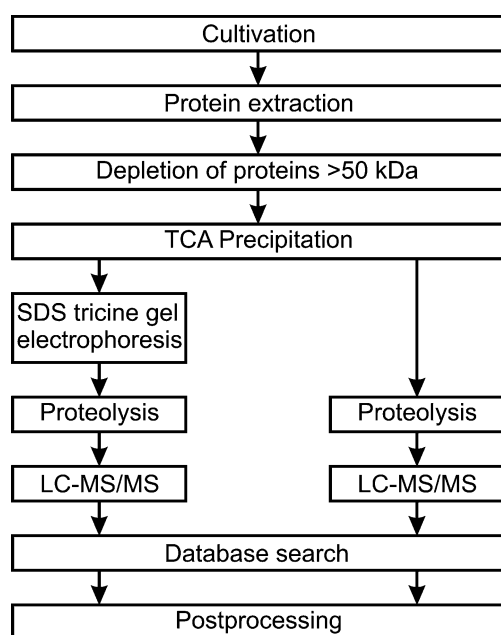
Finally, the challenge of the de novo annotation of open reading frames (ORF) coding for small proteins on a genome-wide scale is addressed with the software *RNA-code*.

Optimization steps

Different protein extraction methods

To estimate the influence of the cell disruption and protein extraction methods, two different lysis buffers (a slightly basic ammonia buffer containing 8 M urea and an acidic buffer containing 0.1% TFA) were applied as a variant of the method described in Klein et al. [20]. Similar protein amounts were obtained with both buffers, which could not be increased by the successive usage of both extraction buffers (data not shown). After the depletion of higher molecular weight proteins using centrifugal filtration (molecular weight cut-off: 50 kDa), high enrichment in proteins <30 kDa was observed, with a maximum at approximately 15 kDa in terms of quantity (Fig. 2) and number of identifications (Fig. 3). The total protein amount determined after depletion and precipitation was approximately 2% for urea and 1% for TFA extracts. Proteins were separated using 1D SDS tricine GE, and the LMW range of each lane was cut into nine slices. Proteins were digested in gel with endoprotease AspN or trypsin, and the resulting peptides were subsequently analyzed by LC/MS.

The analysis resulted in a total of 333 and 223 protein identifications for extractions with urea and TFA, respectively. Interestingly, only 148±13 proteins were detected using both protocols, which represents 44% of all detected proteins (Fig. 4a).

The importance of an efficient cell disruption and protein extraction has already been pointed out in other studies [18,
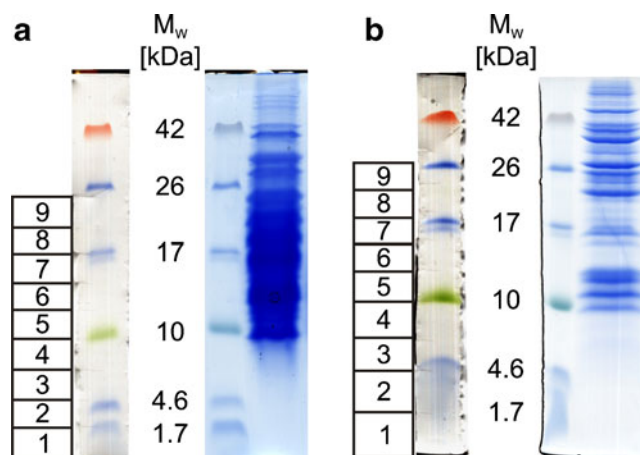

Fig. 1 Experimental workflow

Fig. 2 SDS tricine gel after protein extraction with urea lysis buffer (**a**) and 0.1% TFA (**b**) and subsequent depletion of high molecular weight proteins. Excised bands of the unstained gel part are numbered
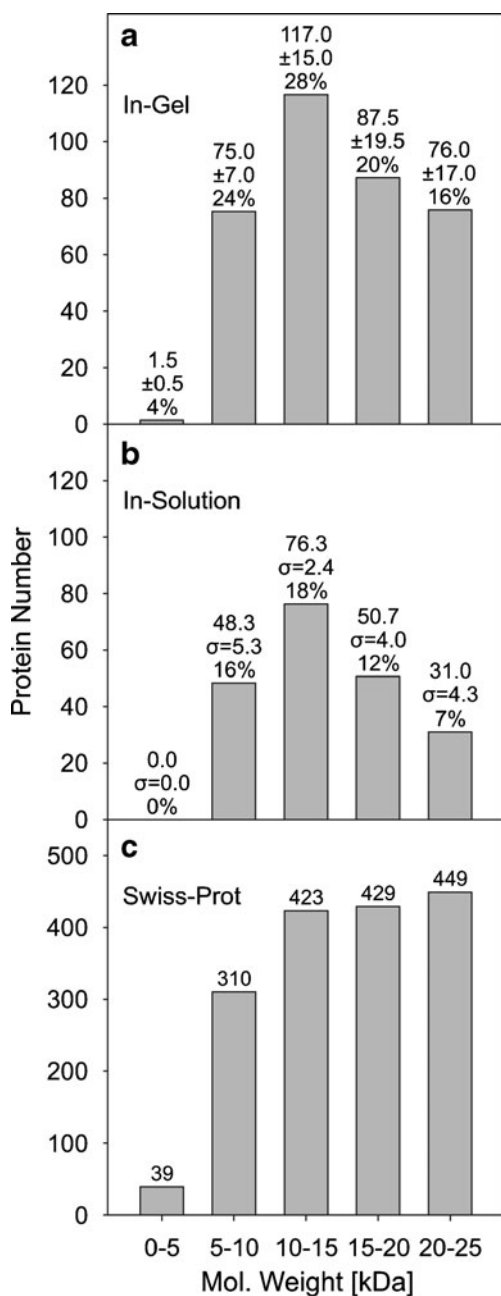
Fig. 3 Average mass distributions of the proteins identified using an in-gel (**a**) or in-solution (**b**) approach in comparison to the SwissProt protein database (**c**)

29]. Our results show that the choice of the extraction buffer can influence the number and type of identified proteins even more than the protease or the MS/MS fragmentation technique (discussed below).

For the proteins in the p*I* ranges of 5–7 and 11–14, the identification rate was higher with the urea than with the TFA lysis buffer (184 vs. 134 proteins, respectively, Fig. 5; Electronic supplementary material Figure S1). For very acidic proteins with a p*I* of <5, TFA lysis gives slightly better results than urea lysis (22 instead of 17 identified proteins).

Different protein separation methods

A 150 min gradient was used for the 1D-LC/MS analyses. However, a gel-based approach in which nine slices were analyzed by LC/MS using a 30 min gradient leads to a 49% increase (Fig. 3, Fig. 4b) in the identification rate. Thus, even though there are differences in terms of LC separation and measurement time, this indicates that investing time and effort in additional separation steps on the protein scale remains an efficient way of improving the proteome coverage. Nevertheless, some proteins may also be lost by additional separation steps. Eleven especially low-abundance (four proteins below 1000 copies/cell) or as-yet unquantified proteins (five proteins) were exclusively detected by the shorter LC/MS-based approach.

Proteolytic digestion

The possibility of increasing the protein identification rate as well as the average sequence coverage through the complementary application of more than one protease is a known strategy. Recently, Swaney and coworkers improved the coverage of the proteome of *Saccharomyces cerevisiae* by performing complementary proteolytic digestions with multiple enzymes and subsequently analyzing using LC/MS [19]. While the proteases trypsin, AspN, GluC, ArgC and LysC were used, the highest identification rate was obtained with trypsin. Nevertheless, the other proteases increased the identification rate by 18% (3908 instead of 3313 proteins) and—perhaps more importantly—the average sequence coverage increased from 24.5% to 43.4% as compared to that obtained with the exclusive use of trypsin.

In addition to trypsin, we used endoprotease AspN, which was predicted to create nearly the same number of proteolytic peptides in the molecular weight range 800–3,000 Da, and to present the highest orthogonality to trypsin in terms of sequence coverage for LMW proteins (Electronic supplementary material Table S4). Furthermore, the prediction showed that in a complementary analysis using both endoprotease AspN and trypsin, the number of unidentifiable LMW proteins would be reduced to 67 in comparison to the 233 not identified when using trypsin as the only protease. For unequivocal identification, at least three detectable proteolytic peptides were required in this in silico digestion (Electronic supplementary material Table S4).

In summary, 292.5±76.5 proteins could be identified with trypsin, and 163.5±9.5 (46%) of these could be verified using endoprotease AspN (Figs. 3 and 4c). The average sequence coverage of proteins identified by both proteases was increased from 48.0% to 63.7% by combining

Fig. 4 Influence of different protocol variations. Comparison of average protein identifications after **a** protein extraction with urea lysis buffer or 0.1% TFA, **b** digestion with the in-solution or the in-gel approach, **c** digestion with trypsin or AspN, **d** MS/MS fragmentation and detection by IT-CID or IT-ETcaD, and **e** MS/MS database search using the MASCOT or SEQUEST search engines
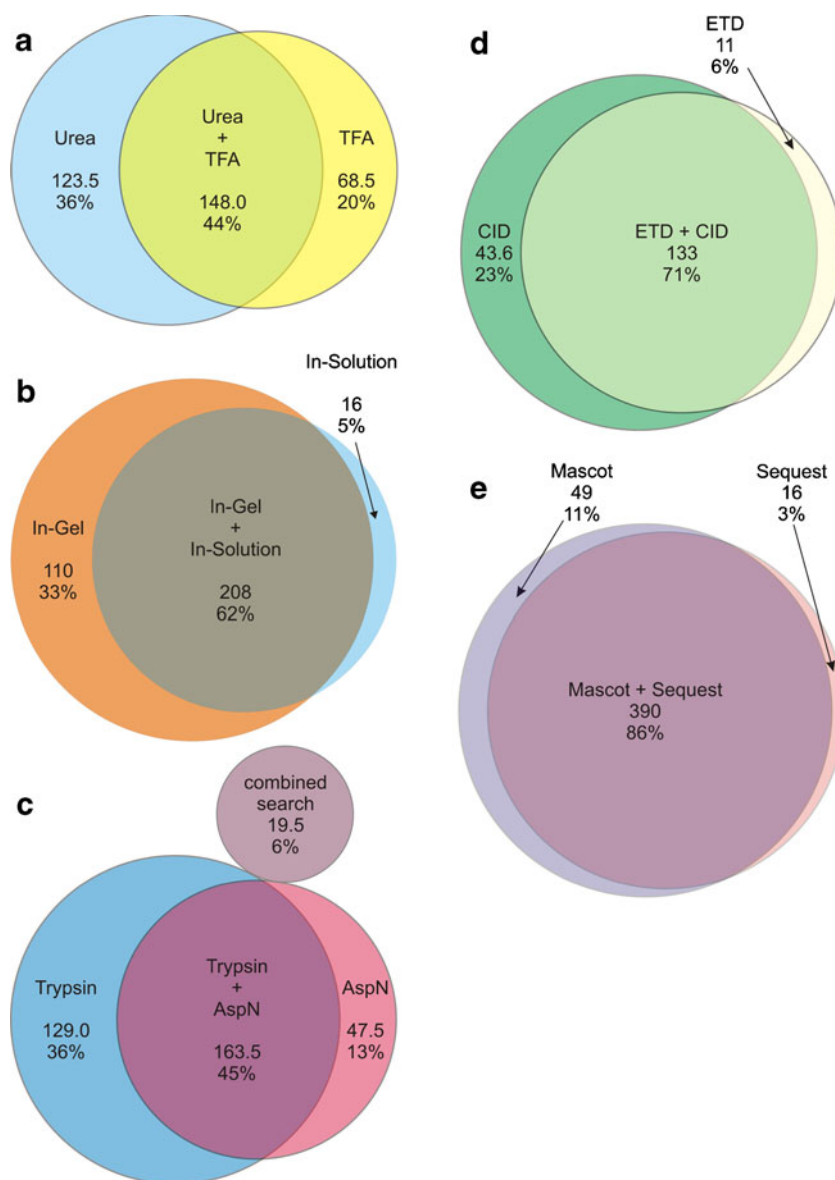
the results obtained using trypsin with those obtained using endoprotease AspN (Table. 5). Furthermore, 47.5±25.5 (13%) proteins could only be identified after proteolysis with endoprotease AspN. According to Ishihama et al. [9], 21 of the 63 additionally identified proteins have copy numbers per cell of below 1000, whereas 28 were not covered by this study. Performing a database search by combining the LC/MS results obtained through digestion with trypsin and endoprotease AspN yielded 19.5±9.5 (6%) additional protein identifications. The abundance of at least several of these proteins was very low (7 were determined to be present with less than 1100 copies/cell), whereas 22 were not yet quantified.

In contrast to tryptic peptides (except C-terminal peptides), which always possess a "mass spectrometry friendly" C-terminal charge due to the occurrence of a C-terminal arginine or lysine, this is not necessarily the case for proteolytic peptides derived via cleavage with endoprotease AspN. This resulted in decreased spectral quality and thus in lower average MASCOT scores (C-terminal arginine or lysine: both 39, for N-terminal aspartic acid and glutamic acid: 30 and 31) and slightly lower SEQUEST scores (for lysine and arginine: 3.3 and 3.1; for acid and glutamic acid: 3.0 and 3.0). The cleavage efficiency of endoprotease AspN was lower for glutamic than for aspartic acid (1586 instead of 205 identified peptides).

Variation of fragmentation technique

The fragments created by ETD, CID and HCD can either be detected with high sensitivity and a short measuring time in the linear iontrap (IT-ETD and IT-CID) or with high
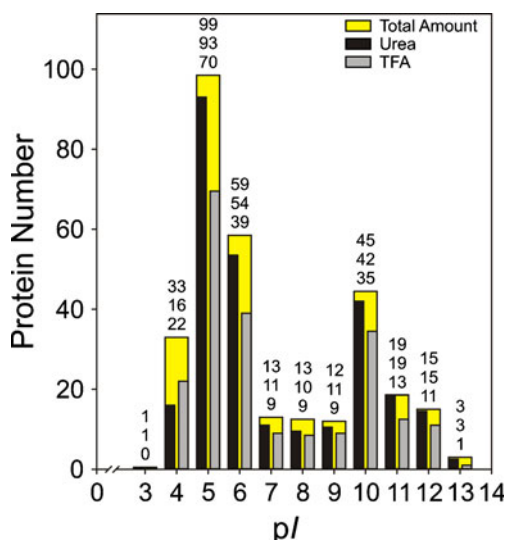
Fig. 5 p*I* distributions of the proteins identified with the in-gel approach after protein extraction with urea lysis buffer or 0.1% TFA in comparison with the total amount of identified proteins

Fig. 6 Comparison of different fragmentation methods after in in-solution proteolysis, as exemplified by the peptide DVFVHFSAIQTnGFK from the cold shock-like protein cspE (**a** IT-CID, **b** FT-CID, **c** IT-ETD, **d** FT-ETD, **e** FT-HCD). *n* denotes an Asn that was found to be deamidated

accuracy and resolution in the Orbitrap analyzer (Orbitrap-ETD, Orbitrap-CID and HCD).

The benefits of using different analyzer types for MS/MS measurements as well as the different fragmentation techniques ETD, CID and HCD were evaluated with biological triplicates.

Using the linear ion trap as the mass analyzer for MS/MS detection, the three methods (a) CID, (b) ETD and (c) CID combined with ETD by a data-dependent decision tree provided an average of 177 ($\sigma=19$), 144 ($\sigma=15$) and 160 ($\sigma=21$) protein identifications with very high confidence. The overlap between the IT-ETD and IT-CID results was 71%, whereas only 6% more identifications were gained by using IT-ETD (Fig. 4d). However, since IT-ETD confirmed 75% of the proteins identified by IT-CID, this complementary fragmentation technique represents a useful method of independent validation. Moreover, the average sequence coverage and the average number of identified peptides per protein were increased by 5.5% and 21.7%, respectively (Table. 5).

Comparing the two different mass analyzers for MS/MS fragment ions, the Orbitrap offers highly accurate fragment ion mass measurements as well as enhanced signal-to-noise ratios for highly abundant peptides (Fig. 6). In contrast, due to its lower speed and sensitivity, about 50% fewer MS/MS spectra could be recorded per run, resulting in about 15% of the unique peptides being identified. On average, MS/MS analysis of the fragments created by CID, HCD or ETD in the Orbitrap resulted in the identification of only 27, 23 and 25 LMW proteins, respectively. This is also consistent with a recent in-depth study by Kim and coworkers, who analyzed *E. coli* lysates by CID fragmentation in the LTQ Orbitrap using different conditions for MS and MS/MS
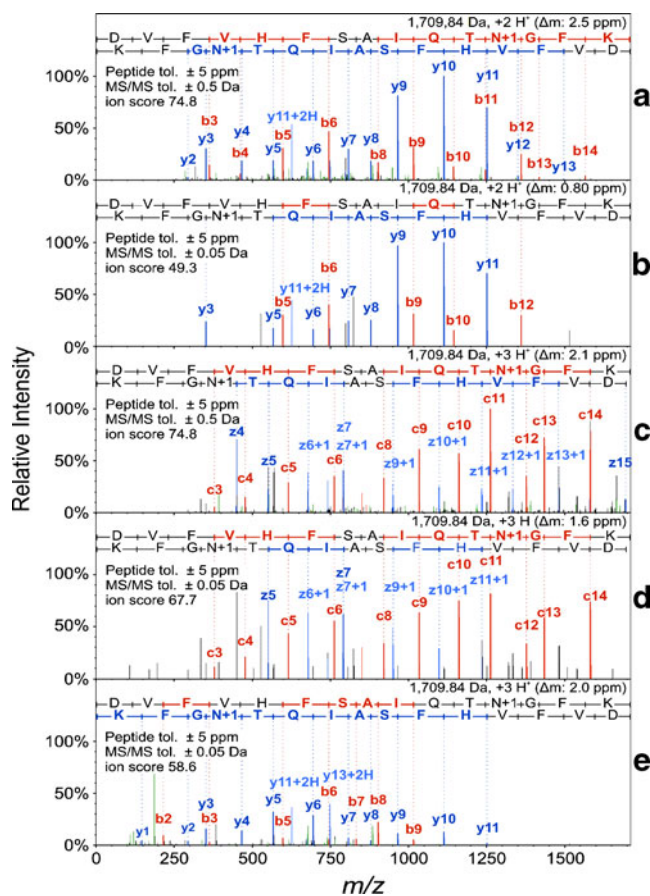
resolution [30]. However, the issue that the number of proteins identified is much lower due to the lower scanning speed and sensitivity of the techique may soon be overcome due to further improvements in the speed and sensitivity of the Orbitrap analyzer [31].

Influence of the MS analysis algorithm

There is still ongoing discussion about the quality of peptide MS/MS search engines [32, 33]. This issue is especially important here, due to the fact that the number of peptides per LMW protein formed by proteolysis is very limited. Additionally, the erroneous identification of a peptide could easily lead to wrong protein identification. Therefore, high sensitivity and accuracy is required during peptide identification. To address this issue with a special focus on LMW proteins, we performed searches with the two most widely used database search engines MASCOT and SEQUEST. After adjusting to 5% FDR using a decoy database, an overlap of 86% was observed (Fig. 4e). Here, MASCOT turned out to be more sensitive, resulting in the

unique identification of 49 unique proteins compared to the 16 discovered by SEQUEST. Furthermore, for the gel-based approach, the number of significant identifications performed by MASCOT, 1060±86 peptides (on average 5.4 peptides per protein), was higher than the 902±85 peptides (5.0 peptides per protein) identified with SEQUEST However, we decided to combine and re-evaluate the results obtained with both engines using SCAFFOLD in order to generate the final identification results.

Covered protein groups

According to the GO classification, the identified proteins were clustered using the GO terms "molecular function," "cell function," and "localization" [27]. Information about the copy number per cell was taken from Ishihama et al. [9].

Cellular localization of identified LMW proteins

With the protocol applied, we obtained good to excellent coverage for cytoplasmic (100 proteins, 45%), periplasmic (22 proteins, 52%) and ribosomal proteins (53 proteins, 98%). Not unexpectedly, the identification rate for inner membrane (43 proteins, 12%) and outer membrane proteins (12 proteins, 33%) was significantly lower (Table 2). However, it is possible to improve the coverage of membrane proteins by performing additional prefractionation [34, 35].

Protein abundance and molecular and cellular function

In order to estimate the copy numbers of a wide range of cytosolic proteins, Ishihama and coworkers [9] used label-free protein quantitation. The proteins identified in this and our study cover a dynamic range of six orders of magnitude. These proteins include highly abundant ribosomal proteins like the 50S ribosomal protein L33 (SwissProt entry: P0A7N9, 186,000,000 copies/cell) as well as rare proteins with less than 200 copies per cell such as Acyl-CoA thioesterase I (SwissProt entry: P0ADA1, 186 copies/cell). Furthermore, we identified about 100 proteins that are not

covered by the study of Ishihama et al. (Electronic supplementary material Table S5).

According to the GO annotations of *E. coli*, neither the biological processes associated with nor the molecular functions of 846 proteins are characterized. Interestingly, 579 (i.e., 68%) of these proteins possess a molecular weight of <25 kDa (Tables. 2, 3 and 4). In our study, we were able to identify 93 of these uncharacterized proteins. The coverage of such proteins by proteome studies will subsequently allow protein quantification, and thus may ultimately contribute to the elucidation of their functional roles.

Detection and evaluation of proteins predicted at the DNA or transcriptome level using *RNAcode*

Among the 1723 individually predicted proteins, there are 837 (49%) LMW proteins that have not yet been validated at the proteome level. Of those 837 LMW proteins, 96 were detected in our study. However, 91 of these were recently covered by Iwasaki et al. [10], whereas, to our knowledge, the existence of the five remaining proteins has never been established before.

Aside from all the experimental challenges involved, an additional reason for the underrepresentation of LMW proteins in proteome studies is probably the inherent difficulty of the annotation process, which results in an significant number of either dubious or missing protein predictions [14, 36, 37]. In order to improve the prediction and annotation of LMW proteins, we used the recently developed *RNAcode* algorithm [16]. *RNAcode* performs a comparison of homolog sequences that show evolutionary conservation and has already been applied to transcriptome data [17].

In the present study, we show how *RNAcode* can revise existing annotations and also estimate their specificity by performing a comparison with our proteome data. Of 1605 mapped LMW SwissProt protein loci, at least 70% of the sequences of 1401 overlapped with segments that gave high scores in *RNAcode*. Ninety-five percent of the proteins with either proteome or transcriptome evidence listed in the SwissProt database are positively classified by *RNAcode*

Table 2 Gene ontology annotation according to localization

Localization	Cytoplasm		Ribosome		Membrane		Periplasmic space		Cell projection/ flagellum		Extracellular		Cell wall/cell membrane		Other/not assigned	
Swissprot *E.coli* K12 <25 kDa	219		55		356		42		36		3		36		995	
In gel	101	46.1%	53	96.4%	43	12.1%	21	50.0%	2	5.6%	1	33.3%	11	30.6%	213	21.4%
In solution	63	28.8%	48	87.3%	23	6.5%	13	31.0%	1	2.8%	1	33.3%	5	13.9%	114	11.5%
In gel + in solution	110	50.2%	53	96.4%	47	13.2%	22	52.4%	2	5.6%	1	33.3%	11	30.6%	229	23.0%

Table 3 Gene ontology annotation according to biological process

Biological process	Antioxidant activity		Binding		Catalytic activity		Enzyme regulator activity		Molecular transducer activity		Structural molecule activity		Transcription regulator activity		Translation regulator activity		Other/not assigned	
Swissprot E.coli K12 <25 kDa	8		636		400		12		21		59		65		4		851	
In gel	7	87.5%	228	35.8%	121	30.3%	7	58.3%	2	9.5%	54	91.5%	25	38.5%	3	75.0%	131	15.4%
In solution	6	75.0%	147	23.1%	63	15.8%	4	33.3%	1	4.8%	49	83.1%	16	24.6%	3	75.0%	73	8.6%
In gel + in solution	7	87.5%	242	38.1%	128	32.0%	8	66.7%	3	14.3%	54	91.5%	28	43.1%	4	100%	141	16.6%

(Electronic supplementary material Table S6). This indicates that there is a strong enrichment of experimentally supported proteins in *RNAcode* predictions. Among the 455 proteins identified in this study, 449 (99%) show a clear evolutionary signal for conservation at the nucleic acid level. Proteome or transcriptome evidence is also reported in the SwissProt database for 81% (365/449) of these. Thus, the proteins identified in our study and the *RNAcode* predictions are highly correlated.

On the other hand, of the proteins not covered in our study or which had already been validated experimentally or by sequence homology according to the SwissProt database, only 68% were supported by *RNAcode* predictions (Electronic supplementary material Table S6). This difference suggests that many but probably not all of the as-yet unverified reading frames in the SwissProt database are real protein-coding segments. Interestingly, 229 high-scoring protein-coding segments detected with *RNAcode* do not overlap with annotated genes. Thus, the existence of LMW proteins which are not included in the current version of the SwissProt database was indicated by *RNAcode* analysis [16].

This analysis clearly shows that the existing SwissProt protein database can be improved, specifically with respect to evolutionary conservation, by the novel in silico approach. Furthermore, the results of our LMW proteome analysis are supported by other experimental data and they show a good correlation with the protein coding signals predicted by *RNAcode* too (Electronic supplementary material Table S6).

In this study, 54 proteins were identified which were only predicted according to EXPASY SwissProt database information (http://expasy.org/sprot/). Furthermore, five of the identified proteins (SwissProt entries P76549, P21418, P0A703, A5A614, and P0AEG8; Electronic supplementary material Tables S2 and S3) have not yet been validated according to the latest large-scale studies by Iwasaki et al. [10] and Ishihama et al. [9]. By applying *RNAcode*, the corresponding gene regions were predicted to code for these LMW proteins with high probability (Fig. 7).

Validation is crucial when claiming newly detected proteins. We analyzed the samples after extraction with urea or TFA lysis buffer and digestion with the endoproteases AspN and trypsin, which produce complementary peptides. This enabled us to unambiguously confirm the existence of all of them by multiple detection with FDR probabilities of below 0.05. For example, for the protein P0AEG8, identification is based on two tryptic peptides and four proteolytic peptides created by the endoprotease AspN, so the sequence coverage was increased to 65% (Fig. 7). Additionally, the predicted proteins were found in independently processed biological replicates.

Table 4 Gene ontology annotations according to molecular function

Molecular function	Cellular process		Developmental process		Interaction with cells and organisms		Localization		Metabolic process		Regulation		Reproduction		Response to stimulus		Other/not assigned	
Swissprot *E.coli* K12 <25 kDa	606		1		63		146		110		205		5		100		807	
In gel	206	34.0%	0	0.0%	6	9.5%	29	19.9%	43	39.1%	71	34.6%	1	20.0%	45	45.0%	148	18.3%
In solution	143	23.6%	0	0.0%	5	7.9%	19	13.0%	26	23.6%	49	23.9%	1	20.0%	32	32.0%	73	9.0%
In gel + in solution	218	36.0%	0	0.0%	6	9.5%	31	21.2%	44	40.0%	76	37.1%	1	20.0%	48	48.0%	160	19.8%

Perspectives on LMW proteome analysis

However, even these improved identification rates (especially in the molecular weight range of 5–15 kDa), compared to state of the art standard proteome studies (Fig. 8), of 62% for cytosolic proteins and 27% for all known LMW proteins (including membrane proteins) still leave some room for further improvement. Aside from aiming for increased coverage through the additional prefractionation of membrane proteins, our results indicate that improving protein and/or peptide separation leads to significantly higher identification rates as well as enhanced average sequence coverage.
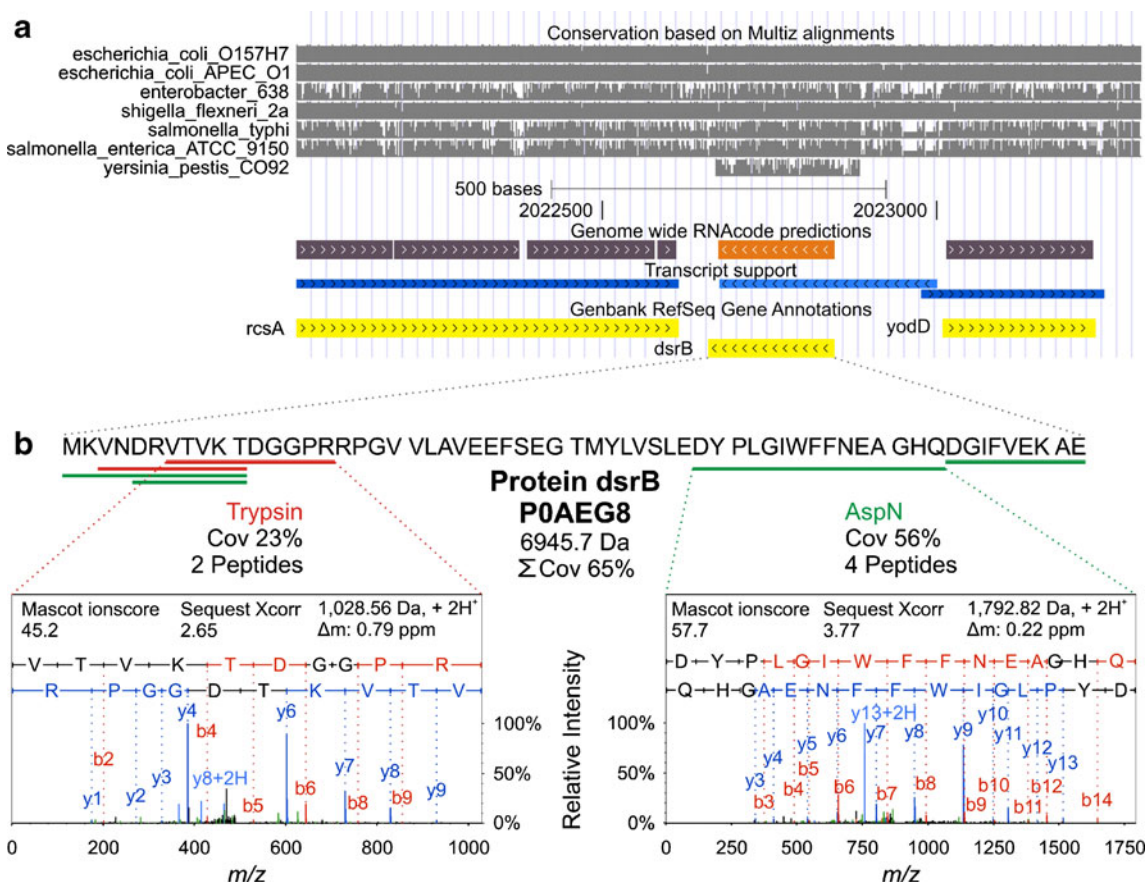

Fig. 7 Evaluation and validation of predicted proteins by **a** RNAcode and **b**. LC/MS/MS. **a** A UCSC screen shot of the genomic context around protein dsrB (Swiss Prot entry P0AEG8) is shown at the *top* with annotated protein coding genes (*yellow*), transcription units as defined by Cho et al. [41] (*blue*) and *RNAcode* high-scoring coding segments (*purple*). *Arrows within boxes* indicate the reading direction of the corresponding element. Marked in *light colors* are elements corresponding to protein dsrB. The *lower half* depicts the conservation of the *E. coli* region with respect to other enterobacteria. **b** Proteins were validated by LC/MS/MS analysis. Spectra and identification parameters of one of the peptides identified using the endoproteases trypsin or AspN are shown.
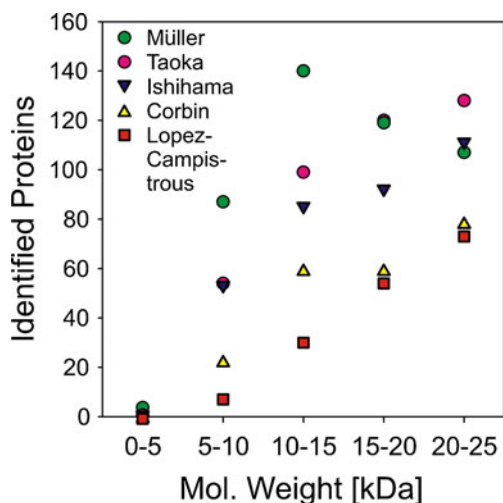
Fig. 8 Comparison of the total number of proteins identified here with the results of selected previous studies focusing on the coverage of the cytosolic proteome of *E. coli*

It was shown by Godoy et al. that near-complete proteome coverage is possible for yeast using *n*-dimensional protein and/or peptide separation prior to MS/MS analysis. However, these approaches are still very time intensive and require the analysis of several dozen proteolytic peptide fractions [38].

Recently, Iwasaki et al. used a non-commercially available 350 cm monolithic reversed-phase C_{18} column to achieve improved peptide separation for proteolytic peptide mixtures of whole *E. coli* cell lysates during a 41 h gradient. This approach allowed for the identification of 2602 proteins, of which 820 were LMW proteins (Table 1) [10]. However, even with this very powerful untargeted analysis, more than 50% of the LMW sub-proteome remained uncovered.

As a complement to the untargeted proteomics approaches, a targeted approach based on multiple reaction monitoring (MRM) has proven to be feasible for high-throughput proteomics studies [39]. The basic idea of this strategy is to optimize the detection of proteolytic peptides

and to develop a sensitive and specific mass spectrometric assay. In a first step, these assays are developed based on specific precursor/fragment ion pairs called MRM transitions as well as LC retention time information by analyzing synthesized peptides corresponding to a proteolytic protein fragment. In a second step, proteins from real samples are identified and quantified by analyzing the real proteolytic peptides using the optimized MRM transitions. Using this approach, even proteins with very low abundances could be detected with a high success rate. However, synthesizing several hundreds to thousands of artificial proteolytic peptides as well as establishing suitable MRM transitions are relatively time- and cost-intensive processes. Nevertheless, especially for very sensitive, specific, and reproducible analyses of limited numbers of proteins, this strategy may be the best method currently available [40].

Summary

In conclusion (see also Table 5), there are various tailor-made strategies that can be used for LMW proteome analyses which vary in their aims and the technical equipment employed:

- For higher sequence coverage, employing a combination of enzymes can significantly increase the number of unique peptides per protein.
- In order to increase the identification rate, the use of an acidic extraction buffer may prove to be beneficial. Furthermore, sequential extraction using different extraction buffers may improve the identification rates, even if the total amount of extracted protein is not increased significantly (data not shown).
- To enhance the robustness of identifications based on an increased number of unique MS/MS spectra, the use of additional enzymes or complementary fragmentation methods like ETD represent efficient options.
- An easy and—with respect to measuring time—neutral way to improve the sensitivity and accuracy of peptide

Table 5 Gains in identification rate, sequence coverage and identification robustness obtained by performing a combined analysis rather than the standard procedure alone

Standard	Option	Proteins	Coverage**	Unique peptides**	Unique spectra**
Urea	TFA	+25.2%	+5.9%	+19.3%	+21.7%
Trypsin	AspN	+16.2% (+22.9%*)	+15.7%	+74.6%	+78.2%
IT-CID	IT-ETD	+6.2%	+5.5%	+21.7%	+30.1%
IT-CID	FT-CID	+0%	+0.7%	+2.6%	+2.4%
MASCOT	SEQUEST	+3.6 %	+1.4%	+3.6%	+4.3%

* Combined identification using trypsin and AspN results in one search

** Related to proteins identified in both experiments

identification is to combine multiple MS analysis algorithms. This is especially important for the identification of LMW proteins, which relies on a very limited number of proteotypic peptides.

- In terms of the efficient use of measurement time, analyzing different preparations of the same sample instead of multiple replicates or using extremely long gradients could be advantageous, as this can increase the total number of proteins identified, the sequence coverage, and the number of peptides per protein.

In conclusion, this study can be used as a guideline to improve the coverage of cytosolic LMW proteins, especially in the molecular weight range of 5–20 kDa.

Furthermore, in this study we investigated an automated protein-coding gene annotation tool. We analyzed the accuracy of *RNAcode* prediction in comparison to SwissProt protein database entries and proteins that we had experimentally verified. We found that the predictions made by *RNAcode* are highly correlated with experimentally validated proteins. Hence, there are 229 high-scoring protein-coding segments that do not overlap with annotated genes and which indicate the existence of additional putative small proteins in *E. coli*.

References

1. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462
2. Han MJ, Lee SY (2006) The *Escherichia coli* proteome: past, present, and future prospects. Microbiol Mol Biol Rev 70:362–439
3. Link A, Robison K, Church G (1997) Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. Electrophoresis 18:1259–1313
4. Lopez-Campistrous A, Semchuk P, Burke L, Palmer-Stone T, Brokx SJ, Broderick G, Bottorff D, Bolch S, Weiner JH, Ellison MJ (2005) Localization, annotation, and comparison of the *Escherichia coli* K-12 proteome under two states of growth. Mol Cell Proteomics 4:1205–1209
5. Ihling C, Sinz A (2005) Proteome analysis of *Escherichia coli* using high-performance liquid chromatography and Fourier transform ion cyclotron resonance mass spectrometry. Proteomics 5:2029–2042
6. Gevaert K, Van Damme J, Goethals M, Thomas GR, Hoorelbeke B, Demol H, Martens L, Puype M, Staes A, Vandekerckhove J (2002) Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis: identification of more than 800 *Escherichia coli* proteins. Mol Cell Proteomics 1:896–903
7. Corbin RW, Paliy O, Yang F, Shabanowitz J, Platt M, Lyons CE, Root K, McAuliffe J, Jordan MI, Kustu S, Soupene E, Hunt DF (2003) Toward a protein profile of *Escherichia coli*: Comparison to its transcription profile. Proc Natl Acad Sci USA 100:9232–9237
8. Taoka M, Yamauchi Y, Shinkawa T, Kaji H, Motohashi W, Nakayama H, Takahashi N, Isobe T (2004) Only a small subset of the horizontally transferred chromosomal genes in *Escherichia coli* are translated into proteins. Mol Cell Proteomics 3:780–787
9. Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ, Frishman D (2008) Protein abundance profiling of the *Escherichia coli* cytosol. BMC Genomics 9:102
10. Iwasaki M, Miwa S, Ikegami T, Tomita M, Tanaka N, Ishihama Y (2010) One-dimensional capillary liquid chromatographic separation coupled with tandem mass spectrometry unveils the *Escherichia coli* proteome on a microarray scale. Anal Chem 82:2616–2620
11. Santos PM, Roma V, Benndorf D, von Bergen M, Harms H, Sa-Correia I (2007) Mechanistic insights into the global response to phenol in the phenol-biodegrading strain *Pseudomonas* sp. M1 revealed by quantitative proteomics OMICS. J Integr Biol 11:233–251
12. Benndorf D, Balcke GU, Harms H, von Bergen M (2007) Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. ISME J 1:224–234
13. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29
14. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE (2008) Small membrane proteins found by comparative genomics and ribosome binding site models. Mol Microbiol 70:1487–1501
15. Hardison RC (2003) Comparative genomics. PLoS Biol 1:e58
16. Washietl S, Findeiß S, Mueller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N (2010) RNAcode: robust prediction of protein coding regions in comparative genomics data. www.bioinf.uni-leipzig.de/Publications/PREPRINTS/10-001.pdf
17. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeisz S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, Stadler PF, Vogel J (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464:250–255
18. De Mey M, Lequeux GJ, Maertens J, De Muynck CI, Soetaert WK, Vandamme EJ (2008) Comparison of protein quantification and extraction methods suitable for *E. coli* cultures. Biologicals 36:198–202
19. Swaney DL, Wenger CD, Coon JJ (2010) Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J Proteome Res 9:1323–1329
20. Klein C, Aivaliotis M, Olsen JV, Falb M, Besir H, Scheffer B, Bisle B, Tebbe A, Konstantinidis K, Siedler F, Pfeiffer F, Mann M, Oesterhelt D (2007) The low molecular weight proteome of *Halobacterium salinarum*. J Proteome Res 6:1510–1518
21. Dai Y, Li L, Roser DC, Long SR (1999) Detection and identification of low-mass peptides and proteins from solvent suspensions of *Escherichia coli* by high performance liquid chromatography fractionation and matrix-assisted laser desorption/

ionization mass spectrometry. Rapid Commun Mass Spectrom 13:73–78

22. Harper RG, Workman SR, Schuetzner S, Timperman AT, Sutton JN (2004) Low-molecular-weight human serum proteome using ultrafiltration, isoelectric focusing, and mass spectrometry. Electrophoresis 25:1299–1306

23. Schagger H (2006) Tricine-SDS-PAGE. Nat Protoc 1:16–22

24. Elias JE, Haas W, Faherty BK, Gygi SP (2005) Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Meth 2:667–675

25. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74:5383–5392

26. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75:4646–4658

27. Bhatia VN, Perlman DH, Costello CE, McComb ME (2009) Software tool for researching annotations of proteins: open-source protein annotation software with data visualization. Anal Chem 81:9819–9823

28. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14:708–715

29. von der Haar T (2007) Optimized protein extraction for quantitative proteomics of yeasts. PLoS ONE 2:e1078

30. Kim M-S, Kandasamy K, Chaerkady R, Pandey A (2010) Assessment of resolution parameters for CID-based shotgun proteomic experiments on the LTQ-Orbitrap mass spectrometer. J Am Soc Mass Spectrom (in press)

31. Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, Lange O, Remes P, Taylor D, Splendore M, Wouters ER, Senko M, Makarov A, Mann M, Horning S (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Mol Cell Proteomics 8:2759–2769

32. Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 5:3475–3490

33. Price TS, Lucitt MB, Wu W, Austin DJ, Pizarro A, Yocum AK, Blair IA, FitzGerald GA, Grosser T (2007) EBP, a program for protein identification using multiple tandem mass spectrometry datasets. Mol Cell Proteomics 6:527–536

34. Molloy MP, Herbert BR, Slade MB, Rabilloud T, Nouwens AS, Williams KL, Gooley AA (2000) Proteomic analysis of the *Escherichia coli* outer membrane. Eur J Biochem 267:2871–2881

35. Masuda T, Saito N, Tomita M, Ishihama Y (2009) Unbiased quantitation of *Escherichia coli* membrane proteome using phase transfer surfactants. Mol Cell Proteomics 8:2770–2777

36. Jäger D, Sharma CM, Thomsen J, Ehlers C, Vogel J, Schmitz RA (2009) Deep sequencing analysis of the *Methanosarcina mazei* Gö1 transcriptome in response to nitrogen availability. Proc Natl Acad Sci 106:21878–21882

37. Basrai MA, Hieter P, Boeke JD (1997) Small open reading frames: beautiful needles in the haystack. Genome Res 7:768–771

38. de Godoy LMF, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, Walther TC, Mann M (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 455:1251–1254

39. Picotti P, Rinner O, Stallmach R, Dautel F, Farrah T, Domon B, Wenschuh H, Aebersold R (2009) High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. Nat Meth 7:43–46

40. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R (2009) Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. Cell 138:795–806

41. Cho B-K, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BO (2009) The transcription unit architecture of the *Escherichia coli* genome. Nat Biotech 27:1043–1049