Optimization of quantitative proteomic LC-MS analyses and proteomic insights into *Helicobacter pylori*

Dissertation
zur Erlangung des
Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät I – Biowissenschaften –

der Martin-Luther-Universität
Halle-Wittenberg,

vorgelegt

von Herrn Stephan Müller

geb. am 01.02.1982 in Scheßlitz

GutachterInnen:

Prof. Dr. Andrea Sinz

Prof Dr. Sacha Baginsky

Prof. Dr. Martin von Bergen

Verteidigt am 06.12.2013, Halle (Saale)

I. Contents

I. CONTENTS I

II. ABBREVIATIONS IV

III. SUMMARY VI

1 INTRODUCTION 1

1.1 Proteomics 1

1.2 *Helicobacter pylori* 3

1.3 Objectives and aims of this thesis 5

2 METHODS FOR IMPROVED IDENTIFICATION AND QUANTIFICATION RATES IN PROTEOMIC APPROACHES 6

2.1 Cell compartment fractionation 6

2.2 Gel-based protein fractionation 7
  2.2.1. Conventional SDS-PAGE 7
  2.2.2. Tricine SDS-PAGE 7
  2.2.3. Two-dimensional polyacrylamide gel electrophoresis 7
  2.2.4. Gel elution liquid fraction entrapment electrophoresis 8
  2.2.5. Applications of gel-based protein fractionation in shotgun proteomics 8

2.3 Liquid chromatography 9

2.4 Application of multiple proteases 11

2.5 Mass spectrometry 12
  2.5.1. Ionization techniques 12
  2.5.2. Mass analyzers 13
  2.5.3. Peptide fragmentation 17

2.6 Database search 22
  2.6.1. Database search engines 22
  2.6.2. Database refinement by proteogenomics 23

2.7 Quantitative Proteomics 24
  2.7.1. Isotope labeling techniques 24
  2.7.2. Label free quantification 27

3 STUDIES ON IDENTIFICATION AND QUANTIFICATION IMPROVEMENT IN PROTEOMIC APPROACHES 29

3.1 Cell fractionation - an important tool for compartment proteomics **29**

3.2 Optimization of parameters for coverage of low molecular weight proteins **39**

3.3 RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data **55**

3.4 Identification of new protein coding sequences and signal peptidase cleavage sites of *Helicobacter pylori* strain 26695 by proteogenomics **73**

3.5 Quantitative proteomics reveals altered expression of extracellular matrix related proteins of human primary dermal fibroblasts in response to sulfated hyaluronan and collagen applied as artificial extracellular matrix **90**

3.6 Stable isotope labeling by amino acids in cell culture based proteomics reveals distinct differences of protein expression between spiral and coccoid *Helicobacter pylori* **104**

3.6.1. Abstract 105
3.6.2. Introduction 105
3.6.3. Methods 107
3.6.4. Results 111
3.6.5. Discussion 117

4 DISCUSSION AND CONCLUSION 125

4.1 Improvement of identification rates and quantification rates in proteomics **125**

4.1.1. Enrichment and fractionation of LMW proteins 126
4.1.2. Application of multiple proteases 129
4.1.3. Application of different MS techniques 130
4.1.4. Application of multiple search engines 131

4.2 Protein database refinement **132**

4.3 SILAC in quantitative proteomics **134**

4.4 Influence of hyaluronan sulfation on primary dermal fibroblasts **135**

4.5 Quantitative proteomics of *H. pylori* by stable isotope labeling by amino acids in cell culture **136**

4.5.1. Choice of medium and labeled amino acids 136
4.5.2. Influence of morphology on the proteome of *H. pylori* 137
4.5.3. Opportunities of SILAC for further studies of *H. pylori* 138

4.6 Conclusion **138**

5 REFERENCES 140

6 APPENDIX IX

6.1 Scientific publications IX

6.2 Affirmation in lieu of an oath / Selbstständigkeitserklärung X

6.3 Curriculum vitae XI

6.4 Acknowledgment / Danksagung XIII

II. Abbreviations

2D:	two-dimensional
AC:	alternating current
ACN:	acetonitrile
aECM:	artificial extracellular matrix
amu:	atomic mass unit
BHI:	brain heart infusion
CagA:	cytotoxin associated gene A
CID:	collision-induced dissociation
COG:	cluster of orthologous group
Da:	Dalton
DC:	direct current
dFb:	dermal fibroblast
DIGE:	differential gel electrophoresis
DNA:	deoxyribonucleic acid
DTT:	dithiothreitol
ECM:	extracellular matrix
ENA:	epithelial derived neutrophile activating protein
ESI:	electrospray ionization
ELISA:	enzyme linked immunosorbent assay
ETD:	electron transfer dissociation
FA:	formic acid
FASP:	filter aided sample preparation
FC:	fold change
FCS:	fetal calf serum
FDR:	false discovery rate
FT:	fourier transformation
FWHM:	full width at half maximum
GAG:	glycosaminoglycan
GELFREE:	gel elution liquid fraction entrapment electrophoresis
GM-CSF:	granulocyte-monocyte colony stimulating factor
HA:	hyaluronan
HCD:	higher energy collision-induced dissociation
HPLC:	high performance liquid chromatography
HILIC:	hydrophilic interaction liquid chromatography
hsHA:	highly sulfated hyaluronan
ICAT:	isotope-coded affinity tag
ICPL:	isotope-coded protein label
ICR:	ion cyclotron resonance
IL:	interleukin
IMAC:	immobilized metal ion affinity chromatography
IR:	infra-red
ISD:	in source decay
IT:	ion trap
iTRAQ:	isobaric tags for relative and absolute quantitation
KEGG:	Kyoto encyclopedia of genes and genomes
LC:	liquid chromatography
LC-MS:	liquid chromatography coupled online to mass spectrometry
LIT:	linear ion trap
LMW:	low molecular weight
LPS:	lipopolysaccharide
LTQ:	linear trap quadrupole
MALT:	mucosa-associated lymphoid tissue

MALDI:	matrix assisted laser desorption ionization
MMP:	matrix metalloproteinase
MRM:	multiple reaction monitoring
MS:	mass spectrometry
MS/MS:	tandem mass spectrometry
MSIS:	mass selective instability scan
m/z:	mass to charge
MALT:	mucosa associated lymphoid tissue
MudPit:	multidimensional protein identification technology
NMR:	nuclear magnetic resonance
ORF:	open reading frame
PAGE:	polyacrylamide gel electrophoresis
pI:	isoelectric point
pH:	negative decimal logarithm of the hydrogen ion activity
PQD:	pulsed Q dissociation
PTM:	post translational modification
RAST:	rapid annotation using the subsystems technology
RF:	radio frequency
RNA:	ribonucleic acid
RP:	reversed phase
SCX:	strong cation exchange
SDS:	sodium dodecyl sulfate
SEC:	size exclusion chromatography
SILAC:	stable isotope labeling by amino acids in cell culture
TIMP:	tissue inhibitor of matrix metalloproteinase
TMT:	tandem mass tag
TNF:	tumor necrose factor
TOF:	time-of-flight
UHPLC:	ultra high performance liquid chromatography
UV:	ultra-violet
VacA:	vacuolating cytotoxin autotransporter A
WBA:	wideband activation

III. Summary

In recent years, proteomics has developed into one of the leading omics techniques in science. Proteomics is defined as "the analysis of the entire PROTEin complement expressed by a genOME, or by a cell or tissue type" [1]. Especially quantitative proteomic studies based on isotopic labeling techniques that investigate differences between biological samples on protein level are gaining popularity.

Here, methods for improved identification and quantification rates in proteomics as well as a non-targeted method for relative protein quantification of the major human pathogen *Helicobacter pylori* were developed. Firstly, proteomic methods were optimized in order to achieve the highest possible quantification rate. Besides subcellular fractionation [2] (chapter 3.1), the main focus was placed on improving the identification rates of low molecular weight (LMW) proteins below 25 kDa that are usually underrepresented in proteomic studies [3, 4] (chapter 3.2, 3.4). Secondly, the protein database quality is of decisive importance for proteomic studies. Therefore, the protein database of *H. pylori* strain 26695 was refined by proteogenomics [4] (chapter 3.4). Thirdly, a quantitative proteomic study based on stable isotope labeling by amino acids in cell culture (SILAC) was applied to study the effect of highly sulfated hyaluronan as artificial extracellular matrix for primary human dermal fibroblasts [5] (chapter 3.5). Finally, the gained knowledge of these studies was combined to establish a non-targeted quantitative proteomic method for *H. pylori*. This method was applied to investigate the influence of the cell morphology on protein level (chapter 3.6).

The first step was to optimize the identification rates for LMW proteins. These proteins are frequently lost during sample preparation such as gel destaining [6]. Additionally, proteolytic digestion of LMW proteins generates a low number of peptides compared to larger proteins. Moreover, LMW proteins like cytokines frequently have low abundances [7]. Hence, LMW proteins are harder to identify and quantify in proteomic studies.

Here, different enrichment and separation methods for LMW proteins were developed based on (i) centrifugal concentrators and subsequent tricine sodium dodecyl sulfate gel electrophoresis (SDS-PAGE) [3], (ii) size exclusion chromatography (SEC) [4] and (iii) gel elution liquid fraction entrapment electrophoresis (GELFREE) (chapter 3.6). Besides the enrichment of LMW proteins, multiple proteases were applied to increase the identification and quantification rates. The application of multiple proteases in separate proteolytic digestions creates a larger number of unique peptides [8]. The identification rates of small proteins benefit particularly from this due to the lower number of proteolytic peptides.

In the first study, precipitated LMW proteins of *E. coli* were subjected to either tricine SDS-PAGE fractionation with subsequent in-gel digestion or direct proteolysis by trypsin in solution [3] (chapter 3.2). The identification rate of LMW proteins (< 25 kDa) was increased by 49% (110 proteins) by tricine SDS-PAGE fractionation [3]. The protein identification rate for

the LMW proteome of *E. coli* was increased by 23% through the application of AspN in comparison to trypsin [3]. The second enrichment strategy by SEC followed by proteolysis with multiple proteases increased the protein identifications below 17 kDa by 18% in comparison to an extensive SDS-PAGE fractionation coupled to LC-MS after proteolytic digestion (GeLC-MS) (20 fractions) [4].

The second focus was placed on the optimization of the protein database quality for *H. pylori* strain 26695 (chapter 3.4). In MS-based proteomic studies, peptides and proteins are commonly identified by searching the MS data against a protein database. The protein sequences that are deposited in those databases are usually created on the basis of gene finding software. Typically, these tools have 300 nucleotides as a minimum length cut-off for open reading frames (ORF) to reduce the false discovery rates (FDRs) [9]. Hereby, proteins below 100 amino acids are frequently lacking in the annotations. Additionally, gene boundaries are also hard to detect.

Here, an in-depth proteomic study that covered 71% coverage of the predicted proteome of *H. pylori* strain 26695 was performed [4] (chapter 3.4). Based on this dataset, a proteogenomic study was performed to refine the protein database of *H. pylori* strain 26695. Therefore, a database was constructed of the NCBI (National Center for Biotechnology Information) database supplemented with the six-frame translation of the genome and protein coding region predictions by RNAcode [10] for *H. pylori* strain 26695.

In this study, four previously missing protein annotations were discovered and erroneous sequences for six additional proteins were corrected. Among the new identified proteins, the ferrous iron transport protein A, the lipopolysaccharide (LPS) biosynthesis protein HP0619 and the coiled-coil-rich protein HP0058 are of particular biological interest. Iron transport e.g. is essential for the survival of *H. pylori* in the stomach [11]. Additionally, the LPS biosynthesis pathway is supposed to be a drug target for the treatment of *H. pylori* infections [12]. Furthermore, the protein HP0058 is essential for the spiral shape and motility of *H. pylori* [13]. Moreover, signal peptidase cleavage sites for 63 proteins were identified by a database search that targets semi-specific cleaved peptides. *H. pylori* showed to have the motif LXA as the predominant signal peptidase recognition sequence at the N-terminal side of the cleavage position in contrast to other Gram-negative bacteria which mainly possess AXA [14].

In order to realize accurate relative quantification of hundreds of proteins of *H. pylori*, SILAC was tested and performed at first with dermal fibroblasts [5] (chapter 3.5). This study aimed to reveal differences in protein expression of primary human dermal fibroblasts (dFb) in response to sulfated hyaluronan applied as an artificial extracellular matrix (aECM) [5]. Sulfation of hyaluronan lead to reduced expression of several extracellular matrix (ECM) related proteins such as thrombospondin-1, collagen types I and XII, as well as the collagen degrading enzymes cathepsin K, matrix metalloproteinases MMP-2 and MMP-14. In addition, the tissue inhibitor of MMPs 2 (TIMP-2) was also found to be down-regulated.

Especially chronic skin wounds have a MMP-TIMP misbalance that may lead to fibrosis metastasis or tumor growth [15]. Several clinical products on the market aim to inhibit MMPs [16-19]. Collagen is excessively produced in hypertrophic scar formation [20]. Reduction of collagen type I and XII expression might indicate that sulfated hyaluronan positively regulates wound healing. In contrast to collagens type I and XII, the abundance of type VI was increased in response to sulfated hyaluronan. This ECM compound is important for the formation of an appropriate environment when the cell layer becomes confluent [21]. In conclusion, sulfated hyaluronan might improve the healing of skin wounds by modulation of the MMP-TIMP balance as well as the altered ECM production.

In the final study, the knowledge gained by the previous studies was combined to develop a quantitative proteomic approach based on SILAC for the major human pathogen *H. pylori* (chapter 3.6). The chemically defined Ham's F12 medium was chosen for this purpose since growth of *H. pylori* was reported without influencing the morphology [22, 23]. Incorporation of lysine and arginine was tested. Sufficient incorporation (> 95%) was only achieved for arginine. Lysine incorporation was to low (~ 80%) since *H. pylori* strain 26695 is a lysine autotroph. The experiment included (i) enrichment and fractionation of proteins below 50 kDa using the GELFREE device prior to proteolytic digestion, (ii) a GeLC-MS analysis with ten fractions, (iii) separate proteolytic digestions with trypsin and AspN and (iv) data analysis with the refined database for *H. pylori* strain 26695.

Here, the influence of morphological changes of *H. pylori* was investigated on protein level (chapter 3.6). *H. pylori* is a Gram-negative epsilon proteobacterium that colonizes the gastric mucosa of approximately 50% of mankind. It is responsible for severe diseases such as gastritis, peptic ulcers and gastric cancer. *H. pylori* occurs in three different morphologies: vital spiral cells, vital coccoid cells and damaged coccoid cells [24]. The coccoid morphology has shown to possess attenuated infectivity as well as colonization efficiency [25, 26]. Differences of protein expression between the two vital morphologies of *H. pylori* strain 26695 were studied. The comparison revealed significantly reduced expression of proteins that are associated with cell division, transcription, and translation processes as well as infectivity and colonization efficiency. Pathway analysis revealed that processes such as chemotaxis and the cytotoxin associated gene (cag) type four secretion are found to be down-regulated in coccoid cells. Additionally, the arginase *rocF* and the TNF-α inducing protein, that are involved in colonization and inflammation processes, show also reduced expression in the coccoid morphology.

In conclusion, methods for improved identification and quantification of proteins were combined with SILAC and a refined protein database for *H. pylori* strain 26695. This approach offers new possibilities for the investigation of *H.* pylori, such as studies on the influence of antibiotics. Additionally, infection processes could be investigated in co-cultures with human epithelial cells in SILAC media.

1 Introduction

1.1 Proteomics

Proteins are involved in many essential functions of life. Muscles or flagella e.g. are necessary for agitation, whereas proteins in bones and hair provide stability and shape. Additionally, proteins are involved in different processes such as enzymatic reactions, inter- and intracellular signal transduction, as well as immune reactions. The objective of proteomics is to study the proteome by means of qualitative analysis or quantification of changed protein expression. By definition proteomics is “the analysis of the entire PROTEin complement expressed by a genOME, or by a cell or tissue type” [1]. In recent years, large scale proteomic studies became one of the key research methods for biological processes. The elucidation of protein [27] or protein complex structures [28-30], the cell response to diseases [31] or certain stimuli such as toxic substances [32, 33], or the study of microbial decomposition of environmental pollutants [34] are some examples for the utilization of proteomic analyses.

Common identification of proteins or structural analyses is summarized under the term qualitative analysis. Quantitative proteomics examine changed protein expression of a cell-line or tissue sample in response to different stimuli like pharmaceuticals, chemicals, changed culture conditions etc. Protein quantification can be performed for instance by Enzyme Linked Immunosorbent Assay (ELISA), western blotting or mass spectrometry (MS).

Over the last two decades, MS has become the main method in proteomics. MS based proteomics is divided into top-down [35] and bottom-up approaches. In top-down proteomics, whole proteins or protein complexes are analyzed. Bottom-up analyses include protein extraction, optional chemical modification, and enzymatic digestion into peptides prior to MS analysis (Fig. 1-1).

Classical quantitative proteomics utilizes two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) [36]. In recent years, MS based quantification methods have become more popular. These are subdivided into label free and stable isotope labeling based quantification [37, 38]. Label free methods facilitate relative quantification of different biological samples by comparison of peptide signal intensities acquired in separate liquid chromatography coupled MS (LC-MS) analyses [39]. In contrast, differential stable isotope labeling comprises relative quantification of several biological samples within one LC-MS analysis [37, 38]. Differentially labeled peptides with the same amino acid sequence that contain different hydrogen, oxygen, nitrogen, carbon or sulfur isotopes possess the same ionization efficiency, as well as quasi the same chromatographic behavior. This enables relative quantification of theses peptides according to their peak intensities by LC-MS (chapter 2.7).

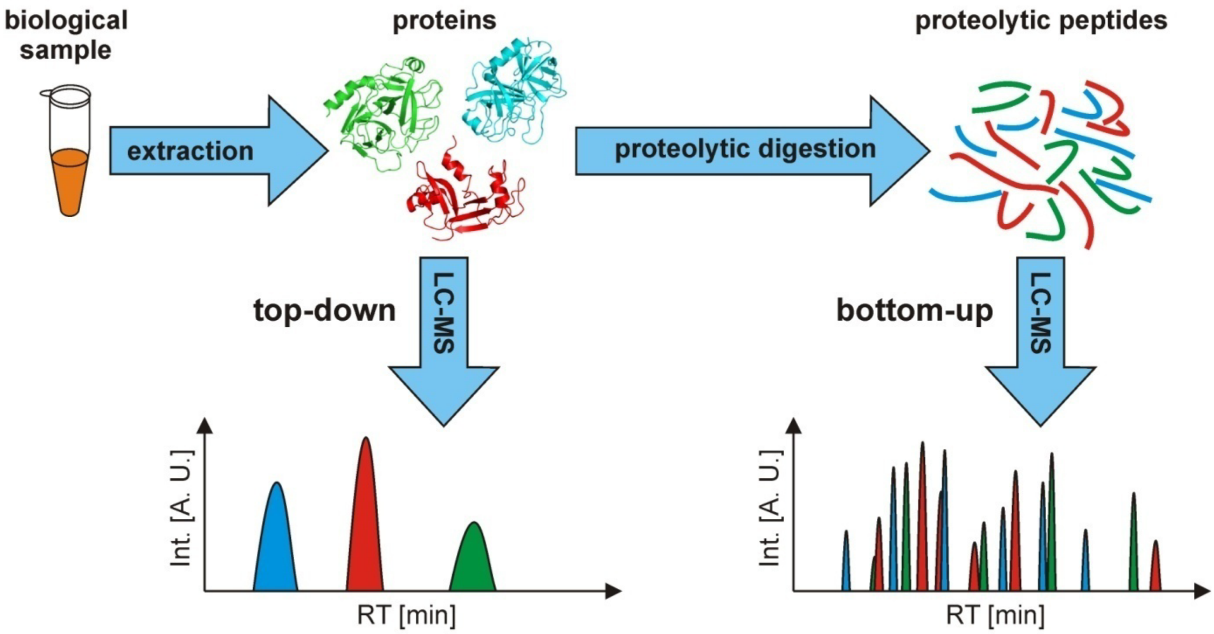


Fig. 1-1: Comparison of LC-MS based top-down and bottom-up proteomics analyses. Proteins are extracted from a biological sample. In a top-down approach, proteins are directly analyzed by LC-MS. In a bottom-up approach, proteins are subjected to proteolytic digestion by proteases such as trypsin prior to LC-MS analysis.

The ongoing development of MS and nano-flow ultra high performance liquid chromatography (nano-UHPLC) [40] continually increases the number of quantified proteins in proteomics experiments. Nevertheless, up to know, it is not possible to achieve a whole proteome coverage due to the enormous complexity, as well as the huge dynamic range in protein abundances. Human blood plasma e.g. exceeds ten orders of magnitude, whereas standard LC-MS analyses are able to cover two to four orders of magnitude [7, 41]. Especially low molecular weight (LMW) proteins are underrepresented in shot-gun experiments [42]. Small proteins are easily lost during the experimental workflow such as gel electrophoresis. Additionally, proteolytic digestion of LMW proteins generates less peptides. Hence, LMW proteins are harder to identify than larger proteins with a high number of proteolytic peptides. Moreover, LMW proteins such as interleukins are often low-abundant. Therefore, specific enrichment of LMW proteins has the ability to improve the proteome coverage and to provide additional identifications of proteins with important biologically relevance.

Conventional proteomic analyses heavily rely on the completeness and correctness of protein databases like UniProt [43]or NCBI [44]. The annotation of protein coding sequences on the basis of genomic data is usually performed with gene finding software such as GeneMARK [45], Glimmer [46], the Integrated Microbial Genome (IMG) system [47], or the Rapid Annotation using the Subsystems Technology (RAST) server [48]. However, the minimum length cut-off for open reading frames (ORF) of these tools is typically set to 300 nucleotides to keep the false discovery rate (FDR) low [9]. Hereby, LMW proteins below 100 amino acids are frequently lacking in the annotations. Moreover, there are certain exceptions to the common translation initiation model [49]. For example, leaderless transcripts are known for archea [50, 51], bacteria [52-54] as well as eukarya which possess leaderless mitochondrial mRNAs [55]. The assignment of the exact gene boundaries is also a typical error source. Additionally, the computational prediction of splice variants in eukaryotes is very challenging [56].

Bakke *et al.* [57] e.g. evaluated three automated genome annotation services for the GRAM negative bacterium *Halorhabdus utahensis*. The IMG system [47], RAST [48] and the J. Craig Venter Institute (JVCI) annotation service were compared comprehensively. RAST e.g. tends to annotate genes with alternative start codons other than ATG (39.0%) more often than IMG (14.3%) or JVCI (19.9%) [57]. A comparison of the gene predictions showed that the three tools share stop sites for 89.7% of all annotations whereas the overlap of genes that share exactly the same start and stop sites was only 47.7% [57]. Remarkably, genes with unique stop codons for one of these tools possess an average length between 250 and 500 bp for the three annotations [57]. This indicates that especially the correct annotation of LMW proteins is demanding.

1.2 *Helicobacter pylori*

In this thesis, the main studies are focused on the major human pathogen *Helicobacter pylori*. In 1906 Walter Krienitz reported the existence of spiral shaped bacteria in the stomach of a patient with a gastric carcinoma [58]. However, it took until 1983 before the scientific importance of this finding was noticed. Barry Marshall and Robin Warren re-discovered the bacterium *H. pylori* in the stomach of patients with chronic gastritis and peptic ulceration [59, 60] and were rewarded with the Nobel prize in physiology or medicine “for their discovery of the bacterium *Helicobacter pylori* and its role in gastritis and peptic ulcer disease” [61].

Nowadays, it is known that the Gram-negative epsilon proteobacterium *H. pylori* inhabits the stomach of about 50% of the human population [62, 63]. Prevalence rates of *H. pylori* in industrialized countries are much lower than in developing countries [62]. Nevertheless, the transmission routes are poorly understood. Although *H. pylori* was partly detected in the oral cavity of individuals [64-66], it is usually assumed that the main transmission route is orally through fecal matter [62, 63]. However, *H. pylori* is also able to survive in groundwater or in rivers, which hereby becomes a potential source for infections [63, 67-69].

Today, there is no doubt that *H. pylori* is the main reason for the development of gastric cancer and other diseases. However, only a small percentage of *H. pylori* carriers develop cancer or ulcer disease, whereas around 80% remain asymptomatic [70]. The risk of cancer development for *H. pylori* positive patients is estimated to be 1-2% whereas the risk to develop ulcers is approximately 10-20% [70].

Cancer development is closely related to inflammation [71, 72]. In particular, many proteins secreted by *H. pylori* including CagA [73], Tip-α [74] or VacA [75] are associated with an inflammatory response of the gastric mucosa. Gastric epithelial cells, which are infected by *H. pylori*, produce pro-inflammatory cytokines or chemokines like interleukine 1β (IL-1β), IL-6, IL-8, the tumor necrose factor α (TNF-α) the epithelial derived neutrophile activating protein 78 (ENA-78) and the granulocyte-monocyte colony stimulating factor (GM-CSF) [74, 76, 77]. Persistent inflammation can cause severe diseases like gastritis, dudedonal ulcer or gastric cancer in the worst case [78].

CagA, for example, is translocated into gastric epithelial cells by the type four secretion system of *H. pylori* [79]. Within the cells, the transcription factor NF-κB is activated by CagA and promotes the production of several pro-inflammatory cytokines such as TNF-α, IL-1 and IL-8 [72]. These cytokines are associated with cancer development [80, 81].

H. pylori forms three different morphologies with either spiral or coccoid cell shape that coexist in the gastric mucosa of infected patients [82]. The spiral morphology is vital, dividing and motile. The coccoid morphology is further subdivided into two subgroups [24]. This is a viable form with an intact cell structure and degenerative cells with disintegrated membrane structures which tend to form cell clusters [24] (Fig. 1-2). The transformation to the coccoid cell shape is promoted by nutritional deficiency, oxidative or acidic stress and antibiotics [83-86].

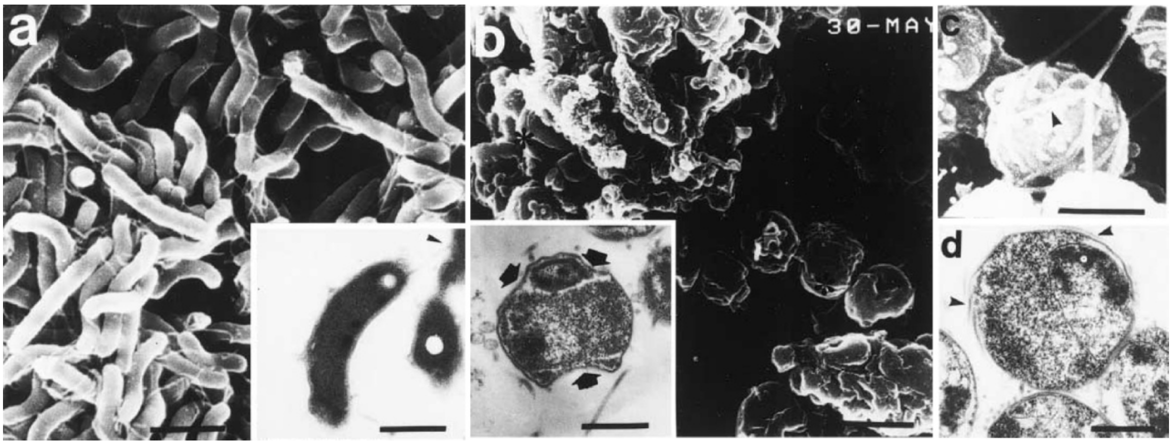


Fig. 1-2: Ultrastructure of H. pylori during culture. (**a**) Spiral forms. Flagella were seen on one side of the organisms (bar = 1.8 µm). [Inset] A spiral organism on the same culture day. A flagellum attached to the adjacent organism (arrowhead bar = 1.0 µm). (**b**) Type A coccoid forms. The surface was irregular and the organisms were clumped together (asterisks bar = 1.0 µm). [Inset] Type A coccoid form observed by transmission-electron microscopy. Arrows indicate hollows. The intracytoplasmic structure was obscure (bar = 0.5 µm). (**c**) Type B coccoid form. The surface was smooth and the flagella coiled about its own bodies (bar = 0.5 µm). (**d**) Type B coccoid form. The membranous structure was assumed to be firm. Arrowhead indicates the flagellum which coiled about its own body (bar = 0.5 µm). (**a**) the 1st day; (**b**) the2nd day; (**c**) and (**d**) the 3rd day. (Reprinted including figure captions with permission from [24]. Copyright (C) 2003, Elsevier)

Coccoid cells have an attenuated infectivity and colonization efficiency. For example, coccoid *H. pylori* failed to colonize the stomach mucosa of gnotobiotic piglets [25]. Additionally, coccoid cells generated a weaker inflammatory response in mice than the spiral morphology [26]. Inflammatory response of different adenocarcinoma cell lines to coccoid *H. pylori* was also attenuated [87, 88]. Incubation of gastric epithelial immortalized cells (GES-1) with coccoid *H. pylori* resulted in lower apoptosis rates as well as reduced production of chemokines and pro-inflammatory cytokines [89].

In contrast, spiral cells swim target-oriented to the antrum, which is the preferred site of infection. Their shape improves movement in viscous fluids like the gastric mucosa and enables them to target gastric epithelial cells [90].

Although many proteins produced by *H. pylori* are related to induction of inflammation and subsequently cancer development, only few proteomic studies have been carried out. *H. pylori* is one of the most intensively studied organisms. Over 28,000 publications which include the name *H. pylori* in the title are listed at the Web of Science®. However, if the results are additionally filtered for “proteomics” as topic, only 42 results remain (Feb 13th, 2013).

Nevertheless, proteomic studies are an indispensable tool for the investigation of biomolecular mechanisms of *H. pylori*. Several proteomic studies have provided insights into the response to acidic [91] or oxidative stress [92], the role of the ferric uptake regulator [93, 94], growth phase dependent changes [95-97] in the proteome, as well as pathomechanisms [98].

However, all these proteomic studies of *H. pylori* were based on comparative two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). This technique has certain advantages such as the identification of post-translational modified proteins, but requires good reproducibility of the protein separation. Additionally, 2D-PAGE experiments are labor-intensive and time-consuming. Frequently, several proteins are identified within one gel spot which complicates the assignment of the regulated proteins. Typically, only few hundred proteins are quantified and identified in 2D-PAGE analysis. To overcome these disadvantages, 2D-PAGE based quantitative proteomics is gradually replaced by isotope labeling techniques that facilitate relative quantification of hundreds to thousands of proteins by MS within one analysis (chapter 2.7).

1.3 Objectives and aims of this thesis

In this thesis, a high coverage non-targeted quantitative proteomic method for *H. pylori* was to be developed to investigate the influence of the cell morphology on protein level. Therefore, improved methods for the identification and quantification of peptides and proteins were established. These methods were combined and applied to establish a quantitative proteomics study for *H. pylori*. For this purpose, the focus was placed on

(i) enhancing the identification of low molecular weight proteins since these proteins are usually underrepresented in proteomic studies,

(ii) the database refinement by proteogenomics, as proteomic analyses strongly depend on the protein database quality, and

(iii) establishing a non-targeted quantitative proteomic analysis of *H. pylori* in combination with the developed methods and the refined database with the aim to unravel the impact of the cell morphology on the infectivity and the colonization efficiency.

2 Methods for improved identification and quantification rates in proteomic approaches

The improvement of identification and quantification rates in proteomic approaches mainly focuses on the extension of the dynamic concentration range. Therefore, fractionation cell compartments, proteins or peptides is widely applied prior to MS analysis. Additionally, the application of multiple proteases further improves identification and quantification rates in proteomics. Furthermore, the ongoing development of MS facilitates proteomic analyses with enhanced sensitivity and accuracy. Finally, database searches offer further potential for the optimization of MS data analysis. Here, major principles of sample preparation, MS, quantitative proteomics and data analysis will be discussed with primary focus on improving identification rates of LMW proteins.

2.1 Cell compartment fractionation

Different physical methods are available to separate individual cell compartments from each other. Prokaryotes do not possess compartmentalization with the exception of encapsulated enzymes [99, 100]. In contrast, eukaryotic cells are composed of different cell compartments which are in large part organelles with distinct biological functions. Separation of such organelles provides deeper insights into biological function of proteins such as signal transduction.

The most commonly used method for cell compartment fractionation utilizes ultracentrifugation to fractionate cell lysates into three fractions. The nucleic fraction is pelleted at 3000 × g [101]. The supernatant is subjected to ultracentrifugation at 100,000 × g to separate the membrane (pellet) and the cytosolic fraction (supernatant) [101]. The classical method for cell compartment fractionation uses gradient centrifugation [101, 102]. For this purpose, either a continuous or a discontinuous gradient is applied. Most commonly, discontinuous gradients based on different concentrations of sucrose are used for subcellular fractionation. During centrifugation, the different organelles are focused within the sucrose gradient at the position of equal density [102]. Hereby, the cytosol and organelles such as nuclei, mitochondria, plasma membranes, lysosomes, golgi apparatuses, and endoplasmic reticula are separable [101].

Alternatively, organelles can be separated by differential detergent fractionation. Proteins of different organelles are extracted subsequently by various detergent containing buffers [102, 103]. Generally, cytosolic, nuclear associated, membrane, and cytoskeletal proteins are extracted in separate fractions [103]. Matured kits for this method are commercially available from different manufacturers and have the advantage that they only require an ordinary bench top centrifuge instead of an ultracentrifuge.

2.2 Gel-based protein fractionation

2.2.1. Conventional SDS-PAGE

Sodium dodecyl sulfate gel electrophoresis (SDS-PAGE) is a standard method in biochemistry for high resolution separation of proteins [104]. SDS is applied to denaturalize the protein structures and to superimpose the charge state of the proteins. Hereby all proteins possess roughly the same weight to charge ratio. By application of a voltage across the gel, proteins migrate into the gel in the direction of the anode. The small pores of the polyacrylamide gel retard larger proteins more strongly than smaller ones. Hereby, proteins are separated according to their size. Proteins are commonly visualized in gels by Coomassie, silver or fluorescence [105].

2.2.2. Tricine SDS-PAGE

Conventional SDS-PAGE allows well resolved separation of a broad molecular weight range of proteins. Separation of proteins within a desired molecular weight range can be widely tuned by the choice of the acrylamide concentration. However, even if the acrylamide concentration is increased, proteins below 13 kDa are poorly resolved [106]. The reason for this is connected to the stacking behavior of small proteins. Proteins below 13 kDa migrate together with the SDS in the tris-glycine buffer system of Lämmli and get poorly resolved [106].

Schägger modified the system of Lämmli [104] for LMW proteins [106, 107]. Glycine in the cathode buffer was replaced by tricine and the acrylamide concentration was increased. Tricine SDS-PAGE facilitates high resolution separation of proteins and peptides down to 1 kDa. However, the upper stacking limit was reduced to 30 kDa which results in worse resolution of larger proteins. Schägger therefore recommends to use the system of Lämmli for proteins with a mass larger than 30 kDa [107]. Tricine SDS-PAGE should be used to separate proteins below 30 kDa [107].

2.2.3. Two-dimensional polyacrylamide gel electrophoresis

Two-dimensional PAGE (2D-PAGE) is a combination of isoelectric focusing (IEF) and SDS-PAGE. In the first dimension, proteins are separated on a gel strip with an immobilized pH gradient. A voltage is applied across the gel strip and the proteins are forced to travel to the position where the pH value is identical to their isoelectric point (pI), this pH value is where the net charge of the protein is zero. Afterwards, the IEF gel stripe is subjected to a SDS-PAGE to separate the proteins by size in the second dimension. 2D-PAGE enables separation of hundreds to thousands of proteins. Fluorescent dyes that are covalently tagged to proteins are utilized for relative quantification by two-dimensional differential gel electrophoresis (2D DIGE) [108]. [36]

2.2.4. Gel elution liquid fraction entrapment electrophoresis

Gel elution liquid fraction entrapment electrophoresis (GELFREE) is a technique that was originally applied to purify proteins. Recently, this method was established for the separation and fractionation of proteins. The system utilizes tube gels for SDS-PAGE separation of proteins. Unlike conventional SDS-PAGE, proteins are eluted at the end of the gel. The eluting fractions are trapped in a small volume (150 µL) against a membrane with a molecular weight cut off of 3 kDa. The separation efficiency of the system is dependent on the percentage of the applied gel. 12% gels for example are used to separate proteins from 3.5 to 50 kDa whereas 5% gels are used to resolve proteins between 75 and 500 kDa. This system benefits from parallel separation of a maximum of eight samples, a high loading capacity (up to 500 µg), and an improved protein recovery in the liquid phase. [109, 110]

2.2.5. Applications of gel-based protein fractionation in shotgun proteomics

The application of 2D DIGE compromises relative quantification of up to three samples on one gel [108]. An advantage of 2D DIGE compared to gel-free quantification techniques is the separation of protein isomers with different post translational modifications (PTM) like phosphorylations, acetylations or sulfations [111]. However, 2D DIGE has a limited dynamic range and the experiments are much more labor intensive than gel-free approaches. Additionally, 2D DIGE cannot compete with the amount of available data gained by MS based quantification techniques [112]. Especially, systems biology research projects require as much quantitative information as possible.

Gel based protein separation techniques are very popular in proteomics because of their high resolving power and orthogonality to liquid chromatography. Particularly, protein separation by SDS-PAGE prior to proteolytic digestion and LC-MS analysis, called GeLC-MS, is frequently used to increase the dynamic range. The gels are usually cut into several fractions and proteins are digested in the gel. Subsequent to proteolytic digestion, peptides are eluted from the gel pieces. The recovery strongly depends on the peptide sequence and varies between 70% and 90% compared to digestion in solution [113]. GeLC-MS is very popular because it is robust and offers high resolution separation of proteins that is orthogonal to reversed phase liquid chromatography (RP-LC). Furthermore, GeLC-MS facilitates efficient protein modification such as reduction and alkylation, deglycosylation or dephosphorylation within the gel. Reagents that are not compatible with LC-MS analysis can be easily removed.

However, Klein *et al.* [6] showed that especially LMW proteins partially elute during the extensive washing and destaining procedure of the in-gel digestion. Hence, this is one of the main reasons for the poor identification rates of LMW proteins in gel-based proteomics.

Conclusively, gel-based separation techniques offer high resolution separation of proteins. Modified protocols enable the separation of LMW proteins or relative protein quantification.

Gel-based separation can be easily combined with mass spectrometric analysis and offers further information such as the molecular weight of the fraction and the pI value of the proteins in 2D PAGE.

2.3 Liquid chromatography

Liquid chromatography (LC) is a technique for the separation of substance mixtures. The separation principle of LC is based on the distribution of different substances in the mobile and the stationary phases. The stationary phase is fixed in columns. The mobile phase moves through the column. It mediates interaction processes between stationary phase and analytes but it is also used to elute separated substances from the chromatography column. [114]

Different stationary phases are used for the separation of proteins or proteolytic peptides in proteomic experiments. The most common types are hydrophobic, ionic, hydrophilic, and affinity interaction, as well as size exclusion. These different LC techniques are used to reduce the sample complexity, in order to increase the sensitivity of MS analysis. Thus, the dynamic concentration range of MS analyses is extended.

Reversed phase liquid chromatography (RP-LC) was designated analog to polar “normal” phases such as silica gels, which were used at first for chromatography. Usually, alkane chains with 2-18 C-atoms are bound covalently to a solid support material such as silica gels. The non-polar character of the stationary phase increases with the length of the alkane chains. C4 or C8 stationary phases are used to separate proteins, whereas peptides are commonly separated on C18 columns. In the final stage of bottom-up proteomics experiments, peptides are typically separated on a C18 column that is directly coupled to an MS via an electrospray ionization (ESI) source.

The mobile phase consists of water and a non-polar solvent which is usually acetonitrile (CAN) or methanol with optional additives. Peptides or proteins are commonly separated and eluted by continuously increasing the percentage of the non-polar solvent. The separation behavior of RP-LC depends on the pH value of the mobile phase. For LC-MS applications, 0.1% formic acid (FA) is typically added to the mobile phase to ensure a pH of two and to provide protons for the peptide ionization. The application of two different pH values facilitates two-dimensional RP-RP separation of peptides [115]. For this purpose, peptides are separated at a basic pH value (e.g. ammonium formeate buffer pH = 10) in the first dimension and an acidic pH in the second dimension (e.g. 0.1% FA, pH = 2).

Ion chromatography is used to fractionate either proteins or peptides. It is subdivided into cation and anion exchange chromatography. Cation exchangers are negatively charged and interact with positively charged analytes, whereas positively charged anion exchangers bind negatively charged analytes. Ion exchangers are designed to retain their charge over a broad pH range. Analytes which are bound to an ion exchanger are usually eluted by a gradient of increasing sodium chloride concentration [114]. The chloride and sodium ions compete with the analytes for the charged binding groups of the ion exchanger. Especially strong cation exchange (SCX) chromatography is used to enrich post-translational modified peptides with phosphorylations or N-terminal acetylations [116].

Hydrophilic interaction chromatography (HILIC) is used to separate analytes according to their polarity. HILIC utilizes polar stationary phases such as silica gels modified with amide [117] or zwitterionic phases [118, 119]. The mobile phase contains a non-polar solvent like acetonitrile (ACN) or methanol mixed with low amounts of aqueous buffer [120]. The stationary phase exhibits a water rich layer whereas the mobile phase possesses a low water concentration. Analytes are distributed between the two phases but also interact directly with the stationary phase [121, 122]. Elution of analytes is carried out by increasing the amount of water in the mobile phase. HILIC is often used for the fractionation of peptides but it is also possible to couple it directly to MS analysis via an ESI source.

Size exclusion chromatography (SEC) separates proteins according to their size. The stationary phase is a material with defined pores. Smaller proteins penetrate further into the pores than larger ones. Thus, smaller proteins are retained stronger than larger proteins. The mobile phase has the task of transporting the proteins and preventing undesired interactions with the stationary phase. Therefore, salts, organic solvents and detergents are added to the mobile phase [123]. SEC of proteins represents an alternative to SDS-PAGE and provides complete orthogonality with RP-LC of proteolytic peptides.

Recently, monolithic columns have become increasingly popular. They consist of a continuous bed support with a porous structure that facilitates fast mass transfer with increased permeability and low backpressure at high flow rates [124, 125]. Long monolithic RP columns e.g. allow high resolution separation of peptides with extremely long gradients (up to 41 h) [126]. The possibility of immobilizing proteins on a monolithic support enables can be used to facilitate on-line proteolytic digestions [127] or affinity purification of peptides and proteins [128, 129]. Furthermore, monolithic columns can be synthesized with a large variety of interaction types such as RP, HILIC or ion exchange [124].

Affinity chromatography utilizes highly specific interactions between the analytes and the stationary phase to purify a specific compound or compound group out of a complex mixture of substances. Monoclonal antibodies are well suited to purify a certain protein [130]. Lectins are sugar-binding proteins which are used to bind distinct glycoproteins or glycopeptides [131]. Immobilized metal ion affinity chromatography (IMAC) is used to bind peptides or proteins containing polyhistidine tags [132] or phosphorylations [133, 134]. Especially monolithic columns are well-suited to immobilize antibodies, lectins, metal ions or avidine for the purification of biotinylated proteins or peptides [128, 129].

Frequently, multiple separation techniques are combined on protein or on peptide level to improve the separation and to increase the dynamic range. Strong cation exchangers (SCX) [135] and SEC [123, 136, 137] can be used to preseparate proteins prior to proteolytic digestion. Multidimensional peptide separation techniques typically use fractionation by SCX, HILIC or RP-LC (pH 10) in combination with RP-LC (pH 2) [115, 138, 139].

2.4 Application of multiple proteases

In bottom-up proteomics experiments, proteases are used to cleave proteins into peptides prior to MS analysis. The most commonly used protease for this purpose is trypsin. It hydrolyzes almost exclusively the peptide bonds that are located C-terminally to arginine and lysine [140]. The cleavage is inhibited if proline directly follows arginine or lysine on the carboxyl side, although it does occur to some extent [141]. Tryptic peptides have the advantage that the side chains of the C-terminal amino acids lysine and arginine are positively charged. Hereby, the ionization efficiency is improved. Additionally, tandem mass spectra of tryptic peptides contain both, N- and C-terminal fragment ions.

However there are a number of different proteases which are also useful in proteomic studies (Tab. 2-1). To improve the digestion efficiency, LysC is often used to perform a pre-digestion of proteins at denaturating conditions with up to 8 M urea prior to proteolysis with trypsin [142]. The endoproteinase LysN creates peptides that are preferably protonated at the N-termini. Fragmentation of these ions leads to enhanced intensities of N-terminal fragment ions [143, 144].

The optimal length of peptides for MS analysis is between seven and 35 amino acids [8]. The application of multiple proteases in separate digestion approaches provides more unique peptides with a suited length and offers increased peptide and protein identifications as well as higher protein sequence coverage [8].

Tab. 2-1: Frequently used proteases in proteomic studies. The pH optima and specificities are according to the manufactures information (Roche).

Protease	**pH optimum**	**Specificity**	**Cleavage side**
Trypsin	8.0	K, R	C-terminal
AspN	7.0-8.0	D (pH 7); D, E (pH 8)	N-terminal
GluC	4.0 and 7.8	E (pH 4); D, E (pH 7.8)	C-terminal
LysC	8.5-8.8	K	C-terminal
LysN	9.5	K	N-terminal
ArgC	7.2-8.0	R	C-terminal
Pepsin	1.8-2.2	Broad specificity; preferred hydrophobic and aromatic amino acids	Preferred C-terminal
Chymotrypsin	7.0-9.0	Y, F, W, L, M, A, D, E	C-terminal

2.5 Mass spectrometry

A mass spectrometer generally consists of three parts, an ion source, a mass analyzer, and an ion detector. Matrix assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) are the preferred methods for peptide and protein ionization. Time-of-flight (TOF), ion mobility, quadrupole, ion trap as well as fourier transformation (FT) are the most widely-used analyzers. Ion detection is commonly realized by electron multipliers, faraday cups or micro-channel plates.

All measurements reported in this thesis were performed on LTQ (linear trap quadrupole) hybrid Orbitrap mass spectrometers (LTQ Orbitrap XL ETD and LTQ OrbitrapVelos ETD). Therefore, the chapter mass spectrometry almost exclusively focuses on the working principle and the analytical capabilities of LTQ Orbitrap mass spectrometers. The LTQ Orbitrap XL ETD mass spectrometer is a combination of a linear ion trap (LIT) and an orbitrap mass analyzer with optional quadrupole like fragmentation and electron transfer dissociation (ETD) fragmentation capability (Fig. 2-1).

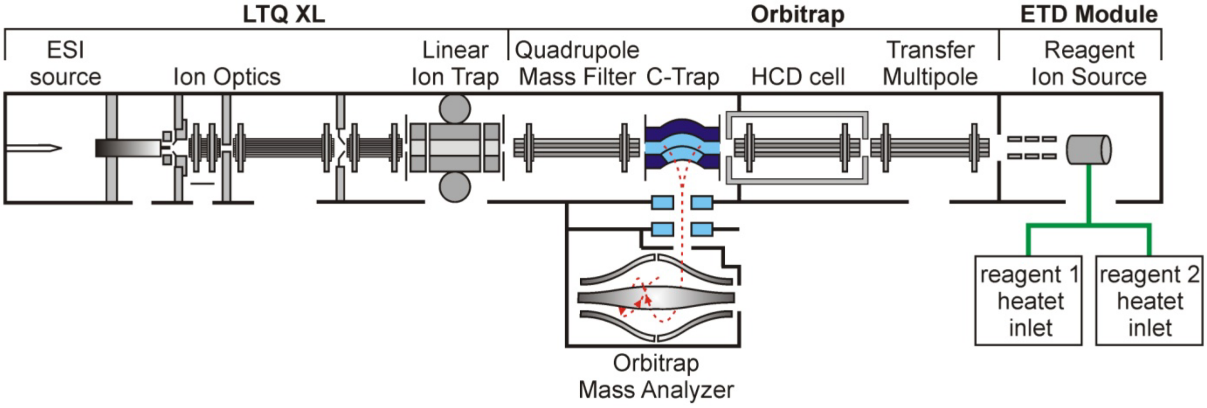


Fig. 2-1: Schematic of the LTQ Orbitrap XL ETD mass spectrometer. A LC is coupled to the mass spectrometer via an ESI source. Ion optics focus the ion beam and transfer the ions into the LIT for MS analysis. The LIT has two ion multipliers for ion detection (shown as circles). Alternatively, ions can be further transferred into the Orbitrap mass analyzer for high resolution, high mass accuracy scans. For this purpose, the ions are compressed within the C-trap in short ion packages which are injected into the Orbitrap for MS analysis. An octopole higher energy collision-induced dissociation (HCD) cell gives the opportunity for “quadrupole like fragmentation”. HCD spectra are recorded by the Orbitrap. An electron transfer dissociation (ETD) module enables ETD fragmentation within the LIT. ETD spectra can be recorded either in the LIT or the Orbitrap. Adapted and slightly modified from [145] with permission. Copyright © 2013 by American Society for Biochemistry and Molecular Biology.

2.5.1. Ionization techniques

In proteomics MALDI and ESI are almost exclusively used for ionization of proteins and peptides. Both techniques enable ionization of non-volatile molecules. LTQ Orbitrap instruments are equipped either with ESI or MALDI ion sources. Here, an ESI source was used to enable direct coupling of an LC with MS.

The MALDI technique was invented by Karas, Hillenkamp and coworkers in 1985 [146]. The principle of MALDI is based on the co-crystallization of analyte molecules with a chromophore carrying matrix. The chromophore enables absorption of ultra-violet (UV) or infrared light. UV MALDI is usually tuned to 337 nm: the emission wavelength of nitrogen lasers. Laser shots are absorbed by the matrix which leads to expansion and disorder to the crystal structure. The absorbed energy is immediately released by an explosive transfer of analyte and matrix molecules into the gas phase. Within the gas-phase radical matrix molecules transfer protons to analyte molecules. Subsequently, the analyte ions are accelerated by the application of high voltage and get analyzed by a mass analyzer.

In the 1980’s, John Fenn and coworkers developed the first ESI source coupling LC with MS [147, 148]. Nowadays nano-LC-ESI-MS is the most frequently used technique for proteomic analyses due to its high sample throughput and separation efficiency.

ESI is based upon desolvation of dissolved analyte ions at atmospheric pressure. The analyte containing liquid flows through a capillary. The electrical potential between the capillary and a counter-electrode disperses the liquid into small charged droplets which are accelerated in the direction of the MS orifice. The charge density of the droplets increases due to the evaporation of the solvent. Hereby, the electrostatic repulsion of the ions rises with the contraction of the droplets. When coulomb repulsion forces exceed the surface tension forces (Raleigh limit), the droplets are broken up explosively into smaller droplets. A row of successive decays finally results in completely desolved ions. [114]

The innovation of MALDI and ESI were major breakthroughs for analysis of non-volatile molecules. In 2002 Koichi Tanaka (MALDI-MS), John Fenn (ESI-MS) and Kurt Wüthrich (NMR spectroscopy) received the Noble Prize "for the development of methods for identification and structure analyses of biological macromolecules.”

MALDI- and ESI-MS have shown to provide complementary results. MALDI-MS tends to identify peptides that contain basic and aromatic amino acids [149], whereas ESI-MS enables enhanced ionization of nonpolar peptides [150]. Furthermore, tryptic peptides ending with lysine are favored by ESI whereas MALDI preferably ionizes peptides with a C-terminal arginine [151]. The application of MALDI- and ESI-MS for the quantitative analyses by isobaric tags, e.g., for relative and absolute quantitation (iTRAQ) showed a modest protein identification overlap between 50% [152] to 63% [153].

2.5.2. Mass analyzers

LTQ Orbitrap mass spectrometers combine the advantages of IT and FT mass analyzers. The FT analyzer is responsible for scans with high resolution and high mass accuracy whereas the IT analyzer facilitates fast scanning with high sensitivity.

The Orbitrap offers superior mass resolution of up to 450,000 in comparison to quadrupole ion traps. Additionally higher mass accuracy (down to 1-2 ppm) is achievable whereas LIT mass analyzers have a low mass accuracy in the range between 0.3 and 0.5 Da in normal scan mode. However, ion fragmentation experiments cannot be performed within the Orbitrap itself and scanning speed is rather low (Orbitrap XL approx. 1 Hz at R = 60,000) compared to LIT analyzers (LTQ Velos: up to 10 Hz at normal scan rate, 0.1 u FWHM, approx. R = 5000) [145].

Thermo Scientific launched a hybrid mass spectrometer called LTQ Orbitrap in 2005. Since that time Orbitrap mass spectrometers were further improved to achieve higher sensitivity and faster scan rates. In the latest version, the Orbitrap Fusion, the scanning speed was increased to a maximum of 15 Hz at a resolving power of 15,000.

Linear ion traps

The design of linear ion traps is a modification of the quadrupole mass analyzer that facilitates the trapping of ions. For this purpose the quadrupole is divided into three parts which are isolated at the end caps for electrical separation. Ions are trapped radially by a two-dimensional radio frequency (RF) field applied at the middle part and axially by direct current (DC) voltages applied at the two exterior sections which generate stopping potentials [154, 155]. In brief, a potential well is formed to confine ions within the linear ion trap.

The ion trap is filled with a low millibar pressure (e.g. LTQ Velos: 6.7 mbar [145]) of buffer gas such as Helium to increase the trapping efficiency [156, 157]. Ions are slowed down by collisions with gas molecules leading to more efficient trapping. The ion motion in linear ion traps is defined by Mathieu functions [155].

The dimensionless variables a_x, a_y, q_x and q_y (Eq. 2-1 - Eq. 2-2) are used to describe stable ion trajectories within the LIT. Ions which are confined in the LIT are oscillating with the frequencies ω_n that are dependent on β, a dimensionless parameter which is a function of a and q (Eq. 2-3). [155, 158]

$$a_x = -a_y = \frac{8 \cdot z \cdot e \cdot U}{m \cdot (x+y)^2 \cdot \Omega^2}$$

Eq. 2-1

$$q_x = -q_y = \frac{4 \cdot z \cdot e \cdot V}{m \cdot (x+y)^2 \cdot \Omega^2}$$

Eq. 2-2

$$\omega_n = (2 \cdot n + \beta) \cdot \frac{\Omega}{2} \text{with} 0 \leq \beta \leq 1, \; n = 0, \pm 1, \pm 2, \ldots$$

Eq. 2-3

z:	number of charges	e:	elementary charge ($1.602 \cdot 10^{-19}$ A·s) [A·s]
U:	amplitude of DC voltage [V]	V:	amplitude of RF voltage [V]
m:	mass of the ion [kg]	Ω:	angular frequency of RF voltage [s^{-1}]
x:	distance to the center in x direction [m]		
y:	distance to the center in y direction [m]		

For ion tapping RF (V) voltage, but not DC (U) voltage is applied whereas a stopping potential is applied to both end-cap electrodes [154]. The stopping potential is usually realized by grounding these electrodes [154]. Thereby all stable ions are located on the q_x axis (a_x=0, Fig.

2-2). Ions with higher *m/z* ratios are located closer to the origin. Ion ejection is performed either by mass selective instability scan (MSIS) or by resonance.

For MSIS, the RF voltage is ramped to raise q_x until ions become unstable by exceeding the critical value of 0.908 [154] (Fig. 2-2). Hereby, ions are ejected one after the other from smaller to higher *m/z* ratios. Resonance ejection is generally used to isolate a narrow range of *m/z* ratios. For this purpose a DC voltage is applied additionally to a RF voltage and ions are ejected according to their resonance conditions [154, 155].

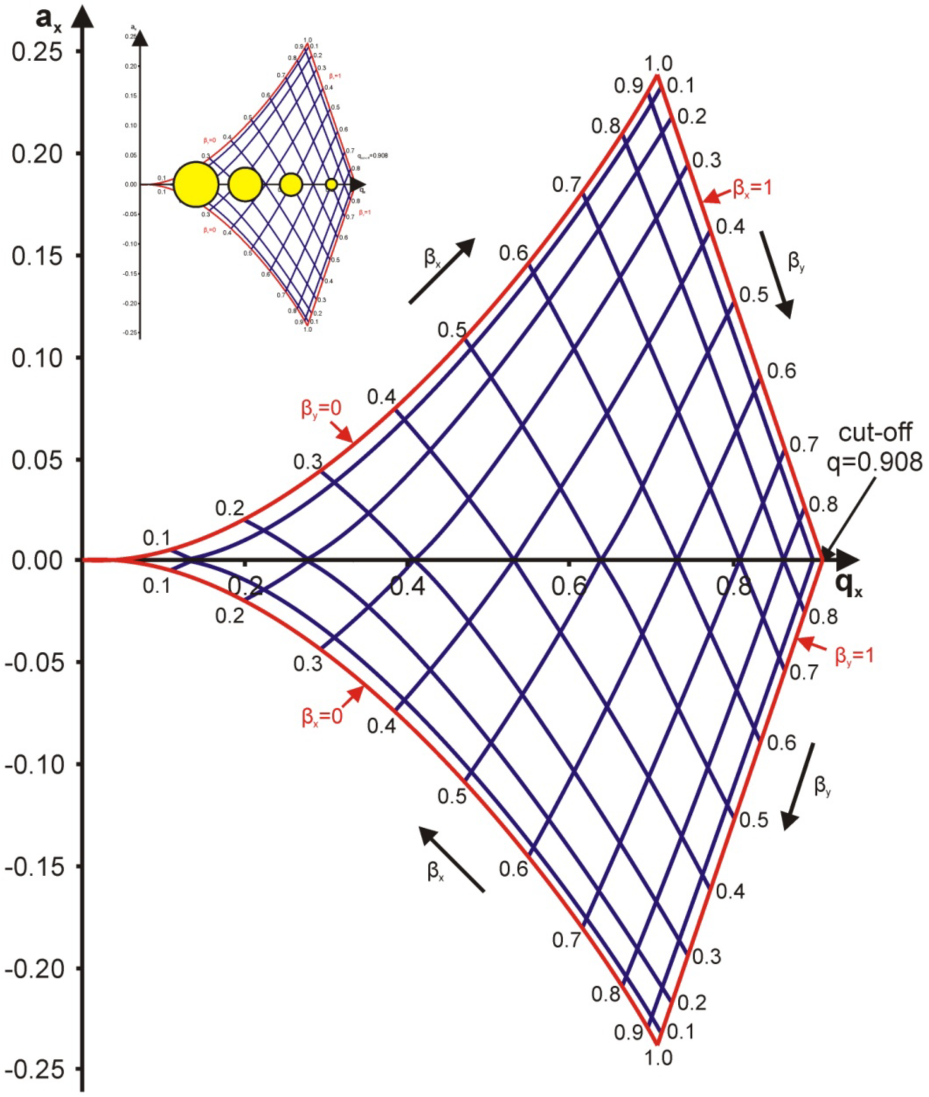


Fig. 2-2: Stability diagram of a LIT. The stability boundaries are indicated in red. Ions are trapped by application of RF voltage along the q_x axis. Ions with higher *m/z* ratio have lower q_x values as indicated as yellow dots in the scaled down diagram. The cut-off value for ion stability along the q_x axis is 0.908. Modified and reprinted with permission from [159]. Copyright © 2008 Elsevier B.V.

Orbitrap fourier transformation mass analyzers

In recent years FT analyzers have become very popular in proteomics. Among the different FT analyzers the Orbitrap, which was invented by Alexander Makarov in 2000 [160], is the most prevalent. The Orbitrap design is based on the Kingdon Trap [161], an ion trap which exploits a merely electrostatic field for ion trapping [162]. Thus, the Orbitrap does not require a superconducting magnet like FT ion cyclotron resonance (ICR) mass spectrometers. As a result of this, the Orbitrap has a smaller size and does not need liquid helium for cooling. Ions are trapped by a combined “quadro-logarithmic” electrostatic potential which is created by axially symmetric electrodes [162].

The ions are injected perpendicularly to the z-axis and laterally shifted from the center in z direction. Subsequently ions are confined on rotational trajectories around the inner electrode with oscillation in z direction. The electrostatic potential forces the ions to move along orbits (Fig. 2-3).

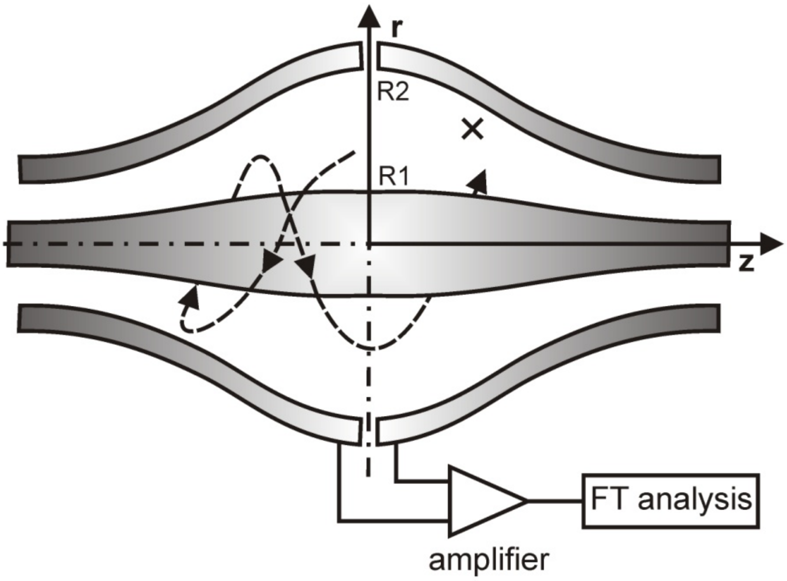


Fig. 2-3: Cross-section of an Orbitrap mass analyzer. r and z are the cylindrical coordinates with z=0 as symmetry plane. The maximum radius of the inner electrode is defined as R1. R2 is the maximum inner radius of the outer electrode. The cross indicates the position where the ions are injected perpendicular to the z-axis. The dashed line indicates a stable ion trajectory. The outer electrode is split in two parts at z=0. The image current of axial ion motion is detected and amplified. The mass spectrum is created after FT analysis of the signal. (Modified and reprinted with permission from [160]. Copyright © 2000 American Chemical Society)

The ion motion in z direction can be described as harmonic axial oscillation. The frequency of this oscillation is a function of the *m/z* ratio (Eq. 2-4) which is independent of the ion energy and position [160, 162].

$$\omega = \sqrt{\frac{k}{\frac{m}{z}}} \qquad \textbf{Eq. 2-4}$$

The axial ion motion induces a current (image current) which is detected by the split outer electrode (Fig. 2-3). Different ion species produce a superposition of the signal. To create a mass spectrum, the signal has to be amplified and processed by FT analysis [160]. The FT analysis transforms the signal into a frequency function that is finally converted into a mass spectrum.

2.5.3. Peptide fragmentation

Peptide fragmentation is used to obtain sequence information. There are different fragmentation techniques available in MS. The fragmentation pattern is dependent on the instrument type and the fragmentation method. Peptides usually fragmentize at their backbone. According to the nomenclature of Roepstorff *et al.* [163] peptide fragmentation is subdivided in a-x, y-b and c-z fragmentation (Fig. 2-4).

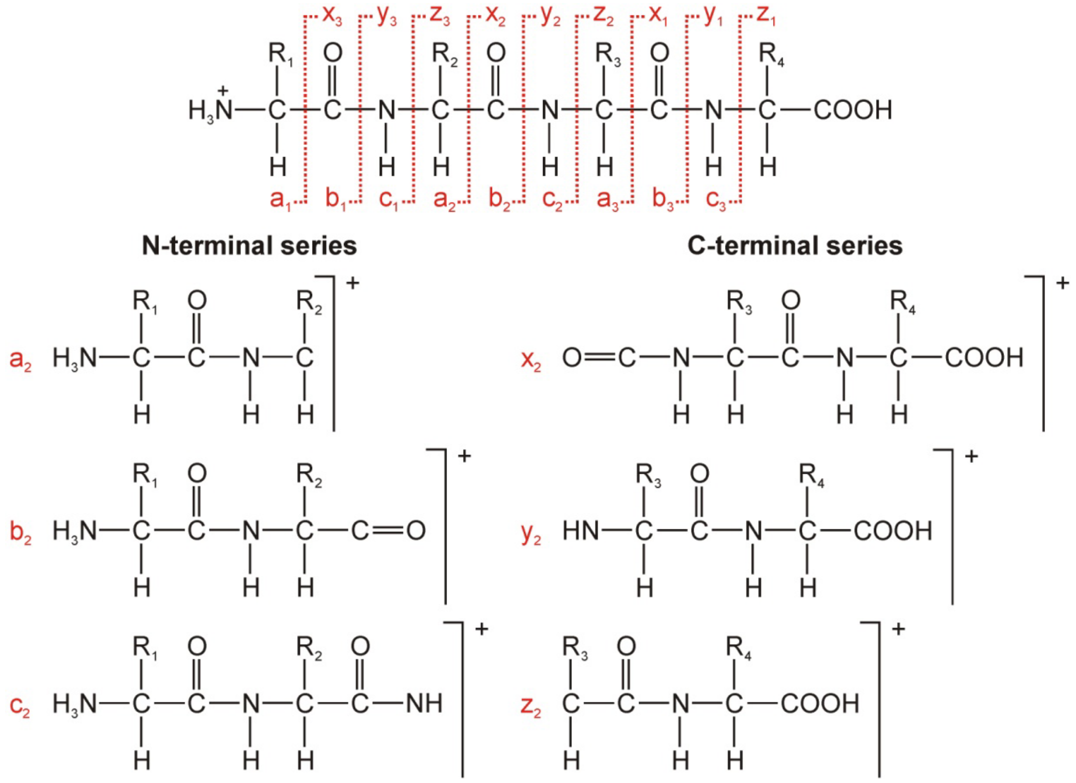


Fig. 2-4: Peptide fragmentation scheme according to Roepstorff et al. [163]. A peptide with four amino acids is shown. Three different cleavage points of the peptide backbone per peptide bond are possible. The N-terminal cleavage products are named a, b and z-ions whereas the C-terminal fragment ions are named x, y and z-ions. The two series are serially numbered starting at the N- or C-term, respectively. The fragment ions for cleavage at the second peptide bond are shown below the peptide.

Here, the possibilities of LTQ Orbitrap ETD hybrid mass spectrometers will be explained. Collision-induced dissociation (CID), pulsed-Q dissociation (PQD), electron transfer dissociation (ETD) and higher energy collision-induced dissociation (HCD) are available for ion fragmentation. HCD is performed in a separate octopole collision cell whereas the remaining three fragmentation techniques are carried out within the LIT (Fig. 2-1). The descriptions of the fragmentation techniques are according to the LTQ Orbitrap Velos Biotech Operations Training Course Manual [164].

The most widely used acquisition mode is based on simultaneous acquisition of MS scans within the Orbitrap and MS/MS scans within the LIT. A predefined number of precursor ions are chosen for subsequent tandem MS (MS/MS) experiments in each scan. While the next mass spectrum is acquired, MS/MS scans are performed in the LIT.

The LIT has two ion multipliers at both sides of the x-axis. As a consequence, all ions which would become unstable in y-direction could not be detected. To force the ions to oscillate in x-direction, an alternating current (AC) voltage is applied to the x-rods. The frequency of the AC voltage is kept constant during ion ejection while its amplitude is ramped together with the RF voltage. As a result, all ions become unstable in x-direction if the new stability limit of $q = 0.88$ is exceeded. [164]

Collision-induced Dissociation

For CID, a precursor ion is isolated in the LIT. All ions except the chosen *m/z* range ($q = 0.87$) are brought to resonance and get ejected. For fragmentation, the precursor ion is cooled down to $q = 0.25$ to facilitate trapping of fragment ions since smaller *m/z* ratios possess higher q values. [164]

The ion species of interest is excited by resonance conditions, but the applied AC voltage is much lower than the one applied for ion ejection. Consequently the chosen precursor ions oscillate with higher velocity, but it is still confined within the LIT. During oscillation precursor ions strike helium atoms, which are present in the trap from ion cooling and dissociate (see chapter 2.5.2) preferentially into y- and b-ions. The fragment ions are no longer in resonance due to their changed q-value which prevents consecutive fragmentations. Frequently neutral losses of water or ammonium of precursor ions occur upon collisions. The precursor ion minus these neutral losses can also be activated by CID to achieve a more complete fragmentation pattern (WideBandActivation^{TM} - WBA). The MS/MS spectrum is recorded by MSIS as described above. A CID fragmentation scheme is shown in Fig. 2-5. [164]

The smallest detectable ion in CID fragmentation is dependent on the activation q and the *m/z* of the precursor. As a rule of thumb, the detection limit is approximately 1/3 of the precursor *m/z* ratio (Eq. 2-5). [164]

$$m/z_{min} = m/z_{precursor} \cdot \frac{\mathrm{q_{activation}}}{0.908} \approx \frac{1}{3} \cdot m/z_{precursor} \qquad \textbf{Eq. 2-5}$$

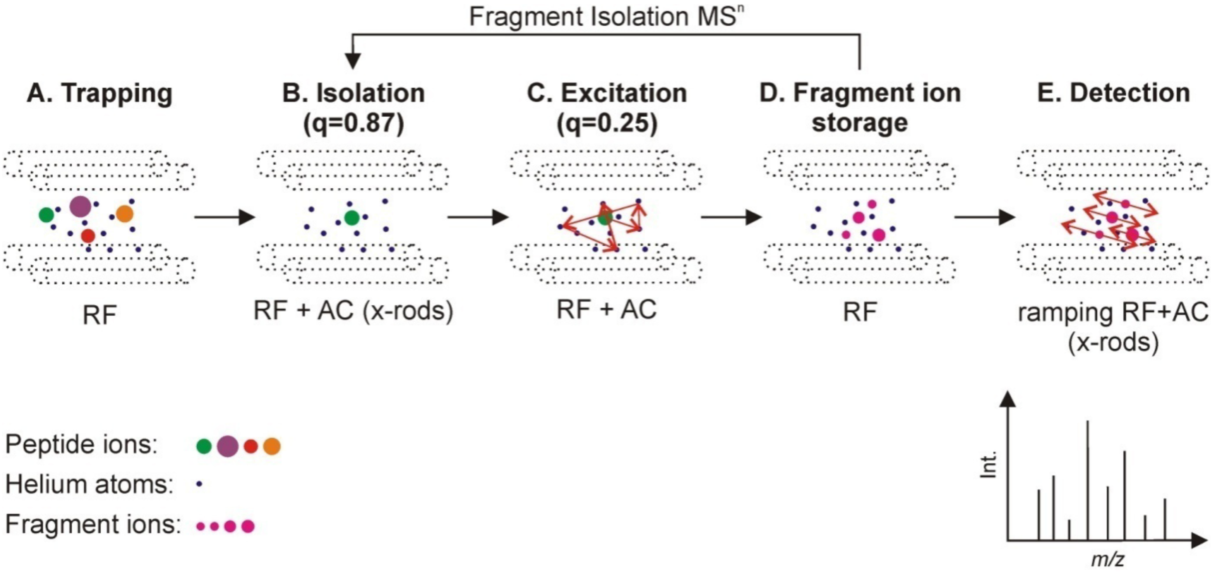


Fig. 2-5: CID fragmentation scheme. (**A**) All ions are trapped. (**B**) A selected precursor ion is isolated at q = 0.87. (**C**) The precursor ion is excited and collides with helium atoms. (**D**) The fragment ions are stored. It is possible to select a fragment ion and to perform MS^n experiments. (**E**) The fragment ions are scanned out in direction of the ion detectors and a fragment ion spectrum is recorded. The fragment ions can either be detected in the LIT or can be subjected to the Orbitrap for scanning.

Pulsed-Q Dissociation

Pulsed-Q Dissociation (PQD) was invented to overcome the low mass cut-off of CID in the LIT. In contrast to CID, the precursor ion is not cooled down to q = 0.25 and is activated at q = 0.87 with higher collisional energy instead. After a delay of 0.1 ms and before dissociation takes place, the q-value of the precursor ion is pulsed to the lowest obtainable value. As a result low *m/z* ratios of fragment ions are detectable, which would be lost by CID. However, fragment ions retain some of the energy which can lead to consecutive fragmentations. As a consequence, MS/MS spectra of PQD are very different compared to CID.

Higher energy Collision induced Dissociation

LTQ Orbitrap mass spectrometers offer a second CID method called higher energy collision-induced dissociation (HCD) that is similar to beam-type CID of triple quadrupole or quadrupole TOF (Q-TOF) instruments. Ions are transferred to the C-trap which is held at ground potential. A quadrupole mass filter enables precursor isolation. Subsequently, ions are injected with high velocity into the HCD octopole cell which is filled with $5 \cdot 10^{-3}$ mbar nitrogen. The ions hit nitrogen molecules with higher energy (~ 30-100 eV) than compared to CID in the LIT (multiple collisions at < 2eV) [165]. After fragmentation, the ions are transferred back into the C-trap for subsequent ejection into the Orbitrap. The MS/MS spectrum is acquired in the Orbitrap at high resolution. Therefore, there is no possibility of simultaneous acquisition of survey MS and HCD spectra.

HCD spectra are dominated by b- and y-ions like CID spectra, but the fragmentation pattern differs due to the application of higher energy. An advantage over CID in the LIT is that fragment ions with *m/z* ratios lower than 1/3 of the precursor are still detectable.

Electron Transfer Dissociation

A complementary fragmentation method to both CID and HCD is Electron Transfer Dissociation (ETD). It was developed by Syka *et al.* [166] in 2004. ETD is a radical-driven fragmentation and mainly produces c- and radical z-ions in contrast to b- and y-ions that are generated by CID and HCD [166]. Thus ETD provides additional information when it is combined ancillary to CID or HCD. A major advantage of ETD is that labile post translational modifications (PTMs) like phosphorylations [167], O-glycosylations [168] and N-glycosylations [169] are retained at the fragment ions.

For ETD, multiple charged analyte cations and radical anions have to be brought together for reaction. Fluoranthene radical anions are produced within the negative chemical ionization (NCI) source (Fig. 2-6, A). Fluoranthene has shown to offer the best electron transfer efficiency from the tested reagents by Hunt *et al.* [170]. The fluoranthene radical anions are subjected to the LIT where the ETD reaction with the peptides takes place (Fig. 2-6).

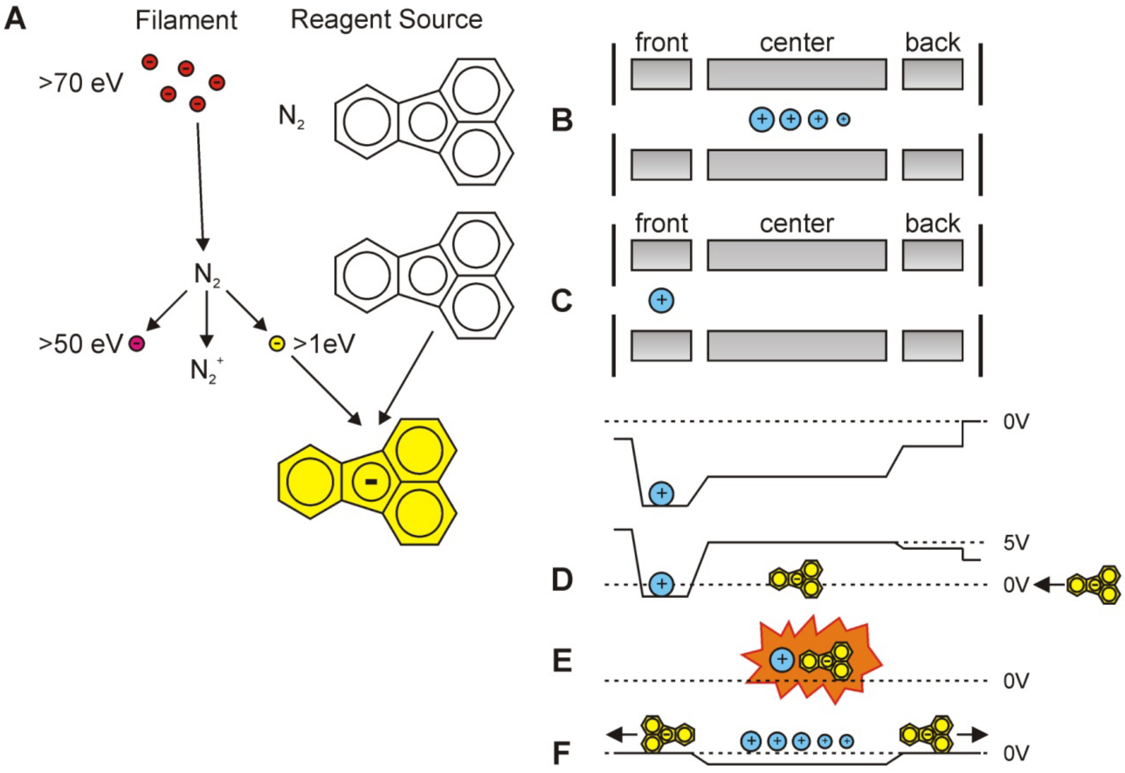


Fig. 2-6: Scheme of ETD fragmentation. (**A**) Fluoranthene is transported by nitrogen into the ion volume. Electrons (> 70 eV) produced by filament are also guided into the ion source. The electrons collide with nitrogen molecules and produce positive nitrogen ions, slowed down electrons (>50 eV) and thermal electrons (> 1 eV). The thermal electrons react with fluoranthene and produce fluoranthene radical ions which are transmitted to the LIT where the reaction takes place. (**B**) Protonated peptides are confined in the LIT. (**B**) A selected precursor ion is isolated and confined in the front section of the LIT by application of a DC offset voltage. (**C**) Fluoranthene radical ions from the negative chemical ion source are injected into the center of the LIT. (**E**) The positive precursor ions are transferred into the center of the LIT where the ETD reaction takes place. (**F**) Remaining fluoranthene radical anions are removed axially. The peptide fragment ions are either measured in the LIT or get axially ejected towards the C-trap to record the MS/MS spectrum within the Orbitrap. Adapted and slightly modified from [164] with permission from Thermo Scientific.

ETD has been shown to perform better than CID for the fragmentation of peptides with charge state three or higher [171]. However, ETD fragmentation of doubly charged peptides tends to result in very poor identification rates due to the charge state reduction during the ETD process. Charge state reduction of doubly charged peptides often leads to non-dissociative electron transfer. Additionally, fragment ion intensities are decreased for doubly charged peptides since either c- or radical z-type fragments remain uncharged after fragmentation. To overcome poor fragmentation efficiencies of peptides with charge state two, supplemental collision activation of charge reduced peptide ions was developed [172, 173]. The resulting fragment ion spectra consist of b-, c-, y- and radical z-type ions when high collisional energy is applied [172] whereas low-energy supplemental collisional activation generates nearly exclusively c- and radical z-type ions [173]. Additionally, c-1 radical ions and z+1 ions are produced due to a hydrogen transfer reactions [174].

Comparison of different fragmentation techniques

The most popular fragmentation technique in proteomics is CID. HCD and PQD offer the detectability of fragment ions lower than one third of the precursor m/z in contrast to CID. This is especially important for relative protein quantification by iTRAQ, where reporter groups with *m/z* between 113 and 121 [175] are used. ETD is most widely used for the analysis of labile PTMs [167-169]. In shotgun experiments, ETD can be used to confirm peptide identification derived by CID and to identify additional peptides. Molina *et al.* published a comparison of CID with ETD and alternating CID/ETD [176]. ETD increased the number of unique peptides by 7-8% and offered confirmation for 53% (ETD) and 71% (ETD/CID) of the peptide identifications by CID [176].

2.6 Database search

2.6.1. Database search engines

Shotgun proteomics strongly depends on database search algorithms for automated peptide identification. Popular search algorithms like Mascot [177], Sequest [178], X!Tandem [179], Andromeda [180] and OMSSA [181] use the same basic principle (Fig. 2-7).

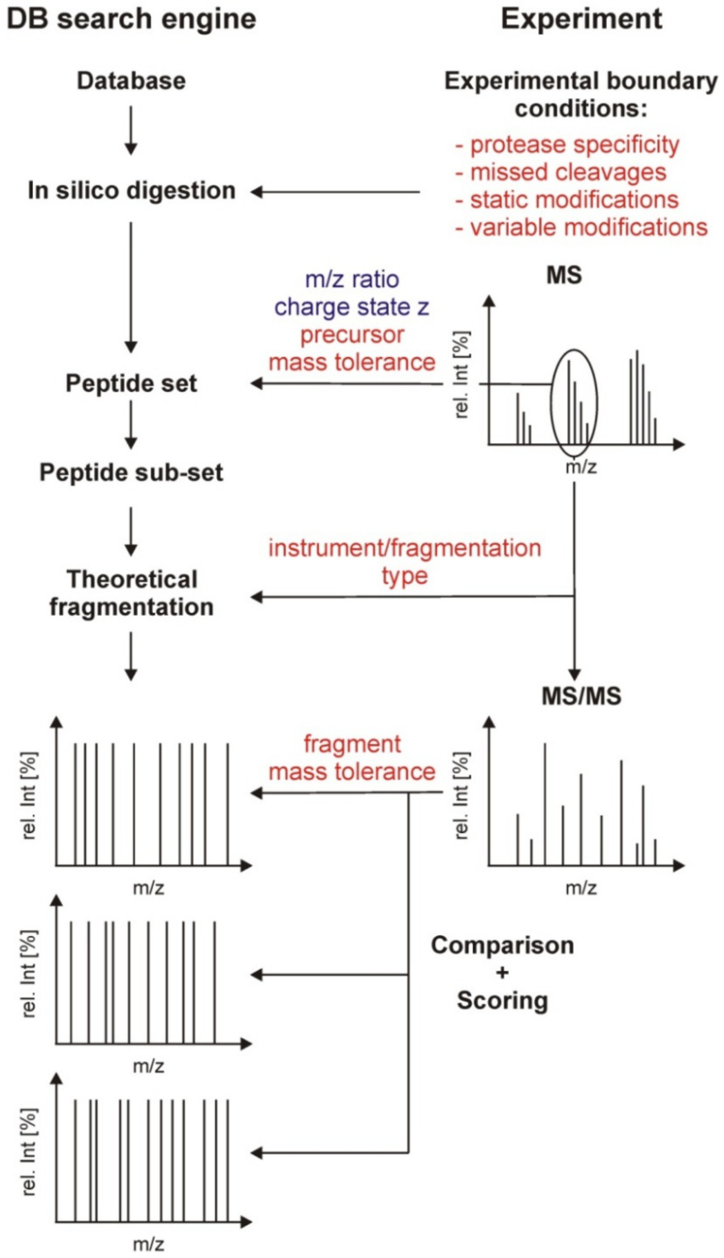


Fig. 2-7: Basic principle of database search engines. An in-silico digestion of all proteins in a given database is performed according to the user settings for protease specificity and allowed missed cleavages. Static modifications (e.g. carbamidomethylation of cysteines) are added to related peptides. For variable modifications peptides of all possible variations are considered. For example, if oxidation of methionine is defined as variable modification and a peptide contains two methionines, all possible variants with no, one and two oxidized methionines are considered. Hence, a peptide set is created for the database search. In a data dependent MS measurement, single precursor ions are selected for fragmentation. The monoisotopic *m/z* ratio of the precursor ion and its charge state is used to extract possible peptide candidates within a user defined precursor mass tolerance (peptide sub-set). A theoretical fragmentation of these peptides is calculated according to the instrument/fragmentation type (theoretical fragmentation). Finally, each theoretical fragment ion spectra of the peptide sub-set is compared to the experimentally derived MS/MS spectrum and receives a score. The best scored peptide is the output of the search engine for this individual MS/MS spectrum.

Other search engines use the interpretation of MS/MS spectra based on short sequence tags within the spectra by searching for consecutive fragment ion series [182]. Subsequent scoring of MS/MS spectra is restricted to peptides which include these sequence tags. The less frequently used search engines X!Hunter [183] and BiblioSpec [184] directly compare MS/MS spectra to spectrum library.

Sensitivity and accuracy differ depending on the scoring algorithm. Hence, the confidence and the quantity of peptide and protein identifications benefit from the usage of multiple search engines. Several software tools like Scaffold, MSblender [185], PeaksDB [186] and PeptideShaker [187-189] facilitate the integration of different search engine results into one data analysis. Additionally cumulative peptide and protein false discovery rates (FDR) are estimated by these tools.

Typically a reverse or random concatenated database is searched to estimate peptide and protein FDRs [190, 191]. These databases have the same amount of target (forward) and decoy (random or reverse) entries that enable statistical evaluation of database search results. It is advantageous to use reverse entries as decoys because amino acid compositions and sequence lengths of obtained decoy peptides are very similar to the target entries [191]. Commonly two different equations are used for FDR estimations in proteomics experiments (Eq. 2-6, Eq. 2-7).

[191]

$$FDR = \frac{2 \cdot ID_{decoy}}{ID_{target} + ID_{decoy}}$$

Eq. 2-6

[190]

$$FDR = \frac{ID_{decoy}}{ID_{target} + ID_{decoy}}$$

Eq. 2-7

2.6.2. Database refinement by proteogenomics

Proteomic studies are strongly dependent on the protein database quality. Conventional database searches are only able to identify peptide sequences that are part of the utilized protein database. Protein sequences are usually annotated computationally according to the genome of the investigated organism by gene-finding software such as GeneMARK [45], Glimmer [46], IMG [47], or RAST [48]. However, the prediction accuracy and completeness of these tools are often suboptimal.

In recent years, the combination of genomics and proteomics, called proteogenomics, has been used to refine protein databases. Proteogenomic studies are utilized for the confirmation and correction of existing protein annotations as well as the identification of new protein coding genes [192, 193].

Generally, a protein database for the investigated organism is constructed from the existing protein annotations and a six-frame translation of the genome. Instead of the genome, transcriptome data can be translated into protein sequences. This is particularly useful for eukaryotes, which possess a high content of non-coding DNA, to keep the database size manageable. Alternatively, manual annotations or alternative sequences can also be added to the database. Next, the MS data from proteomics experiments is searched against this combined database. [192, 193]

Peptides that match to the existing annotations confirm the predicted protein sequence entries. Peptides which are unique to the six-frame, transcriptome translation, or the database completion are used to identify new protein coding genes and incorrect annotations such as wrong translation initiation assignments. [192, 193]

2.7 Quantitative Proteomics

2.7.1. Isotope labeling techniques

In recent years, stable isotope labeling techniques for proteins or proteolytic peptides were developed to overcome the limitations of 2D-PAGE analysis in quantitative proteomics. All of these approaches have in common, that equal protein amounts of differentially labeled samples are mixed prior to MS analysis. Differentially labeled peptides with the same sequence co-elute during LC-MS analysis. Relative quantification is usually performed by comparing the signal intensities of these co-eluting peptides in the survey spectrum whereas chemical labeling with isobaric mass tags utilizes the intensities of reporter ions that are generated by CID fragmentation for quantification (Fig. 2-8).

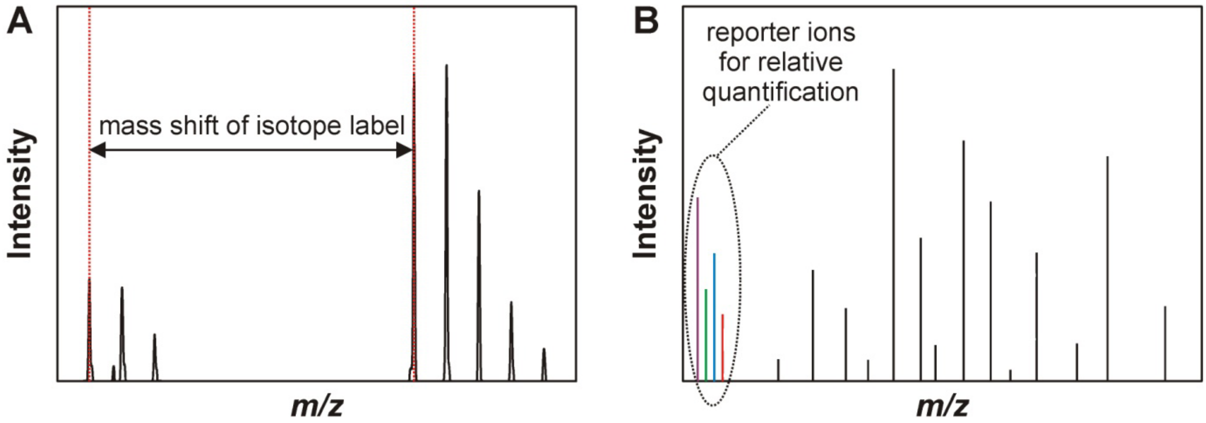


Fig. 2-8: MS based quantification with isotope labeling. (**A**) Relative quantification on the basis of differentially labeled peptides by comparing the individual intensities in the survey scan. The isotope pattern show the mass shift introduced by the isotope label. (**B**) Relative quantification on the basis of isobaric mass tags. The reporter ions are generated during peptide fragmentation. The reporter ion signal intensities are used for relative quantification whereas the other signals are used for peptide identification.

Metabolic labeling

Metabolic labeling is achieved during cell growth and division [37]. Protein labeling can be performed by growth on substrates fully labeled with stable isotopes like ^{15}N or ^{13}C [194-196]. Recently, the ^{36}S and ^{34}S stable isotope labeling of amino acids for quantification (SULAQ) was introduced [197, 198]. Here, cysteine and methionine residues are labeled metabolically with stable isotopes of sulfur (Fig. 2-9).

Stable isotope labeling by amino acids in cell culture (SILAC) [199] facilitates the incorporation of distinct isotopically labeled amino acids into the proteins. These amino acids are added to a chemically defined culture medium. The labeled amino acids should be essential for the studied organism to achieve quantitative incorporation into proteins. However, even though arginine is not an essential amino acids for humans, it is commonly used to label human cell lines and shows sufficient incorporation of greater than 95%.

Since labeling with heavy nitrogen, carbon, sulfur isotopes or SILAC is performed in cell cultures, there are special requirements for the culture medium. For SILAC, the medium has to lack the amino acids which are used for labeling. If nitrogen or carbon labeling is used, the labeled substrates should be the only metabolized N- or C-source of the organism. Therefore chemically defined growth substrates are indispensable for metabolic labeling. It is also possible to perform metabolic labeling of whole animals using isotope labeled feeding. However, those experiments are very cost-intensive and time-consuming.

Chemical labeling

Chemical labeling can be performed either at protein or at peptide level. The chemical label techniques isotope-coded affinity tag (ICAT) and isotope-coded protein label (ICPL) are commercially available for protein labeling [200]. The ICAT reagent includes a biotin tag that is bound covalently to the cysteine side chains of the proteins. After optional preseparation of proteins and enzymatic digestion, tagged peptides are enriched by avidin-biotin affinity purification [201]. Due to the low abundance of cysteines in proteins, quantification with ICAT is not very robust compared to other techniques [200] (Fig. 2-9).

With ICPL, proteins are labeled at primary amino groups which results in labeling of lysine residues and protein N-termini [202]. Up to four samples can be compared within one measurement. Due to the modification with ICPL, lysine sites are prevented from proteolytic digestion with trypsin. Hence, trypsin exclusively cleaves C-terminal from arginine resulting in longer peptides. To overcome the problem of long peptides, a two-stage digestion of trypsin and GluC is recommended to create shorter peptides which are more suitable for MS analysis. When using the 4-Plex labeling technique, two derivatives have deuterated labels. The deuteration produces retention time shifts of 10 to 20 seconds compared to the other derivatives [203]. However, the software tool ICPLQuant uses this feature for more reliable identification of ICPL multiplets [203].

ICPL can also be performed post-digestion [204]. Post-digest ICPL with trypsin generates smaller peptides with enhanced ionization efficiency compared to the standard ICPL approach. Moreover, peptides are labeled at lysine residues and their N-termini, which lead to more quantification features. [204]

Isobaric tags for relative and absolute quantitation (iTRAQ) [205] and tandem mass tags (TMT) [206] also use labeling of proteolytic peptides. TMT and iTRAQ labels are isobaric isomers that consist of a mass balance group and a reporter group [205, 206]. During MS/MS analysis, the mass balance group is released as a neutral fragment, whereas the differentially labeled reporter groups offer the information for relative quantification [205, 206]. The main advantage of iTRAQ and TMT compared to the other mentioned labeling techniques is that the sample complexity is not increased due to the isobaric label. However, when using ion-trap mass spectrometers, the quantification in the low *m/z* range of MS/MS spectra is challenging. TMT reagents are available at a maximum of 6-Plex [207], whereas iTRAQ facilitates relative quantification of up to eight samples within one measurement [175] (Fig. 2-9).

Enzymatic labeling

Enzymatic labeling by trypsin digestion in ^{18}O-labeled water is another possibility to label peptides [208]. Two labeled oxygen atoms are introduced at each peptide C-term during digestion by trypsin [208]. One labeled sample is mixed with a non-labeled reference post digestion. The mass difference of 4 amu of co-eluting peptides is used to relatively quantify changes of protein expression [209] (Fig. 2-9).

Spike in of labeled peptides

Relative quantification can also be performed by spiking isotopic labeled peptides into a sample. The so-called super-SILAC approach uses a SILAC-labeled peptide standard derived from cultured cell lines [210]. The applied cell lines should be related to the investigated tissue. Geiger *et al.* [210] for instance applied a SILAC-labeled mix of five cancer cell lines to investigate human tumor cells. Synthetic peptides with isotopic label can be used as well for relative or absolute quantification which is achieved by the addition of a defined amount of labeled peptides. Commonly, multiple reaction monitoring (MRM) is used for absolute quantification by spiked in peptides [211]. A drawback of this targeted approach is the focus on a set of chosen proteins.

In conclusion, chemical labeling such as ICAT, ICPL, TMT or iTRAQ are applicable for any protein sample. Especially for analysis of proteomics samples from animal experiments chemical labeling is the method of choice. On the other hand, metabolic labeling offers a more robust quantification workflow due to the labeling at an early state of the experiment. Therefore, subsequent separation and fractionation methods can be applied without influencing the quantification accuracy. Metabolic labeling techniques are very accurate (< 10% relative standard deviation, rsd) whereas TMT or iTRAQ possess medium accuracy (10-30% rsd) [38]. As a result of its robustness, easy handling, and well automated data processing SILAC is the method of choice among the different metabolic labeling techniques.

The choice of the labeling technique additionally influences the applicability of fractionation methods. Protein labeling allows the usage of several separation techniques on protein level whereas labeling after proteolytic digestion merely allows reduction of sample complexity by preseparation of peptides.

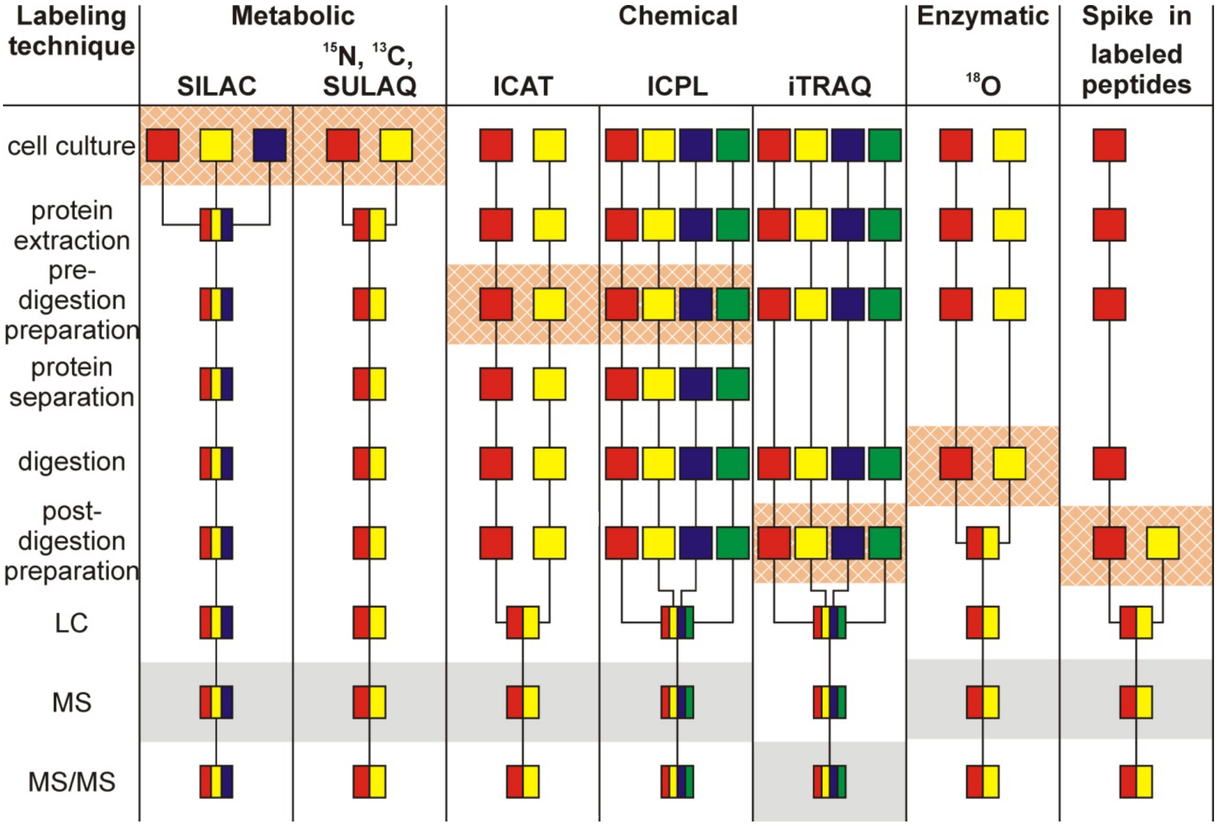


Fig. 2-9: Isotope labeling techniques for MS based proteomics. The labeling step is indicated by orange background with white line patterns. The quantification step is indicated by grey background. Modified and reprinted with permission from [37]. Copyright © 2007, Springer-Verlag.

2.7.2. Label free quantification

Label free quantification usually utilizes either spectral counting or ion intensity profiling. Spectral counting quantifies proteins according to the number of identified peptide MS/MS spectra. In order to compare protein abundance, spectral counting data can be normalized according to the length of the associated protein [212]. Spectral counting is an easy method to compare different datasets. However, quantitative accuracy is poor compared to other techniques and the results are strongly dependent on the measurement setup like dynamic exclusion of peptides for MS/MS acquisition [38, 39].

Label free relative quantification with ion intensities utilizes the peak volume from extracted ion chromatograms of identified peptides of multiple LC-MS/MS experiments [213]. The peptide identifications are matched between different LC-MS/MS experiments according to user defined *m/z* and retention time tolerances to increases the number of quantification features [213].

Comparison of ion intensities allows cost-effective relative protein quantification of all kinds of biological material. However, quantification accuracy of this method is 10% to 30% relative standard deviation (rsd) [38], whereas RSDs of metabolic labeling methods like SILAC can go below 10% [38, 214].

The accuracy of label free quantification by intensities strongly depends on the duty cycle of the precursor ion scans, because peak shapes are fitted according to the measurement points [38, 213]. In other words, the more data points (mass spectra) are acquired the better are the peak shapes. Furthermore, data processing for label free quantification such as data reduction, deisotoping, feature detection, and noise filtering has a strong influence on the results [213]. Additionally, reproducibility of LC-MS/MS runs have to be very high to obtain good results [38]. This is not exclusively a concern of the LC separation and the ESI quality, but also includes high reproducibility of cell compartment, protein or peptide fractionation methods.

3 Studies on identification and quantification improvement in proteomic approaches

3.1 Cell fractionation - an important tool for compartment proteomics

Maxie Rockstroh, **Stephan Müller**, Claudia Jende, Alexandra Kerzhner, Martin von Bergen, Janina Melanie Tomm.

Journal of Integrated OMICS, Vol 1, No 1 (2011), 135-143.
DOI: 10.5584/jiomics.v1i1.52

Abstract

In order to maximize coverage in proteome studies, a successful approach is the fractionation of cellular compartments. For providing evidence for the most reliable and efficient separation technique, we compared four different procedures for subcellular fractionation of Jurkat cells. The analysis of fractions by LTQ-Orbitrap yielded between 559 and 1195 unambiguously identified unique proteins. The assumed correct localization of the proteins was defined using Scaffold3 according to GO annotations, with the highest reliability (~80%) for the cytoplasmic fraction and the lowest (~20%) for the cytoskeletal fraction. This comparison revealed evidence for the efficiency of separating subcellular fractions and will thereby facilitate the decision on which procedure might be the best match to a specific research question and contribute to the emerging field of compartment proteomics.

Keywords

Subcellular compartments; Cellular fractionation; Protein localization; Mass spectrometry.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.52

Cell fractionation - an important tool for compartment proteomics

Maxie Rockstroh[1], Stephan A. Müller[1], Claudia Jende[1], Alexandra Kerzhner[1], Martin von Bergen[1,2], Janina M. Tomm* [1].

[1]Department of Proteomics, Helmholtz Centre for Environmental Research - UFZ, Permoser Str. 15, 04318 Leipzig, Germany; [2]Department of Metabolomics, Helmholtz Centre for Environmental Research - UFZ, Permoser Str. 15, 04318 Leipzig, Germany.

Received: 31 October 2010 **Accepted:** 22 November 2010 **Available Online**: 2 December 2010

ABSTRACT

In order to maximize coverage in proteome studies, a successful approach is the fractionation of cellular compartments. For providing evidence for the most reliable and efficient separation technique, we compared four different procedures for subcellular fractionation of Jurkat cells. The analysis of fractions by LTQ-Orbitrap yielded between 559 and 1195 unambiguously identified unique proteins. The assumed correct localization of the proteins was defined using Scaffold3 according to GO annotations, with the highest reliability (~80%) for the cytoplasmic fraction and the lowest (~20%) for the cytoskeletal fraction. This comparison revealed evidence for the efficiency of separating subcellular fractions and will thereby facilitate the decision on which procedure might be the best match to a specific research question and contribute to the emerging field of compartment proteomics.

Keywords: Subcellular compartments; Cellular fractionation; Protein localization; Mass spectrometry.

1. Introduction

In proteomics it is desired to obtain the largest possible coverage of the proteome of interest and especially to detect proteins of mediate or even minor abundance, too [1]. Beside the development of more and more sensitive mass spectrometers the most frequently applied approach for increased proteome coverage lies in the fractionation of the sample prior to analysis. This can be performed on the levels of subcellular compartments [2-4], proteins or peptides [5, 6] or a combination of different approaches [7]. The biologically most meaningful way is to separate subcellular compartments in order to preserve the linkage of proteins with the compartment in which they exert their activity. In many cases the biological relevance of a protein is closely linked to specific compartments and thereby it's influence on the whole phenotype of a cell.

Hence a great variety of methods for separating the subcellular compartments and subsequent proteome analysis have been developed (for review see [8]). Beside the coverage of the proteome, in praxis the hands-on time plays an important role for deciding in favor of a specific technique. Other criteria are reproducibility and in a few cases also high throughput capacity.

A well-established technique for separation of organelles is solely based on two different types of centrifugation, density velocity and density gradient centrifugation making use of differences in sedimentation coefficients and densities. With endpoint centrifugation, the membrane fraction of a broken cell can be obtained, regardless of the origin of the membrane [8]. Pellets resulting from a centrifugation scheme will stem mainly from the cytoplasmic membrane and only to lower percentages from organelles. A further sub-fraction that can be highly enriched by centrifugation contains the nuclei [9]. Due to their similarity in size but differences in density the remaining organelles like mitochondria, microsomes and lysosomes are often separated by density gradient centrifugation [10-12]. The centrifugation steps can be performed in buffers preserving protein structure and that are compatible with proteomic techniques like 2D-gel electrophoresis or LC-MS shotgun proteomics [8]. In summary, centrifugation schemes can be seen as recommended for enrichment of nuclei and membranes or for specific organelles like mitochondria, lysosomes and microsomes. Unfortunately, due to the

***Corresponding author:** Dr. Janina Tomm, UFZ Helmholtz Centre for Environmental Research, Department of Proteomics, Permoser Str. 15, 04318 Leipzig, Germany. Fax: +49-341-*2351787*. *Email* Address: Janina.Tomm@ufz.de.

nature of centrifugation, it is also time consuming and prevents high throughput.

In a more chemical orientated approach one can use a sequence of detergents with increasing solubilisation efficiency. Thereby a detergent like digitonin will be used to extract cytoplasmic proteins from a cell extract. The subsequent centrifugation will yield a highly enriched fraction of cytoplasmic proteins in the supernatant, whereas proteins from the pellet will be extracted by a stronger detergent like Triton X-100 [13]. There is a great variety in the sequence and choice of detergents described in other studies [14, 15]. Regrettably, this approach suffers from the wide variety of proteins and their interactions in turn leading to a modest specificity of extraction steps for subcellular compartments. Nevertheless, there are also some biologically highly relevant subcellular compartments like the proteome of the lipid rafts that can be extracted with high specificity [16].

In order to obtain high specificity and reproducibility while being cost- and time efficient, various combinations of physical and chemical methods using centrifugation and detergents have been developed. In addition, many protocols have been designed that lack ultracentrifugation and can be performed in volumes that are suitable for most widely distributed bench-top centrifuges, thereby increasing the high throughput capacity significantly.

Here we focused on the comparison of four different methods ranging from a rather simple separation into a soluble, mostly cytoplasmic fraction and an insoluble, mainly membranous fraction up to separation schemes leading to more than five different fractions. For three separations commercially available kits from Fermentas (ProteoJet Membrane extraction kit), Qiagen (Qproteome Cell Compartment Kit [17]) and Pierce (Subcellular Protein Fractionation Kit) were used. A fourth procedure was adapted from literature [18]. Hence we provide evidence for the decision on the most suitable separation for different purposes. It is noteworthy that the results might be cell line or tissue specific, so this has to be tested for the sample of choice. Here we focused on Jurkat cells, which serve as a cellular model for T helper-cells. They mimic important changes that also occur in native T-helper cells once they become stimulated. These processes lead to differential protein expression which has consequences in the cytoplasm, the nucleus and also in the membrane compartment.

With the development of shotgun mass spectrometry and data bases with predictions and reports on the subcellular distribution of proteins, a fast and reliable tool became available for testing the efficiency of the separation procedures. Again, in order to achieve optimal coverage and high reproducibility, a subfractionation was applied. The obtained fractions were applied to a SDS-gel and after a short run each lane was cut into three parts which were subjected to in-gel digestion. Measurement of the peptides by modern mass spectrometry revealed up to 670 proteins per fraction. For validating the results of subcellular fractionation approaches the number of several hundreds of proteins can be assumed to be sufficient to obtain a representative data set and for judging the success of the cellular fractionation.

In this study we provide evidence for the question which separation technique is the most favorable for a specific research question and approach. In addition to the achieved proteome coverage of subcellular compartments there are further requirements that need to be taken into account. For a specific research topic it might be helpful to use a combination of methods. The comparisons conducted here will help to facilitate proteomic research of subcellular compartments and organelles.

2. Material and methods

2.1 Cell culture

Jurkat T cells (clone E6-1, TIB-152, LGC Promochem, Wesel, Germany) were routinely maintained in RPMI-1640 medium (Biochrom AG., Berlin, Germany) containing 10% fetal bovine serum (Biochrom AG., Berlin, Germany), 1% L-Glutamine (Biochrom AG., Berlin, Germany), 1% streptomycin (100 mg/ml) / penicillin (100 U/ml) (PAA, Pasching, Austria) at an atmosphere of 5% CO_2, 95% humidity at 37 °C in a CO_2 incubator (MCO-18AIC, Sanyo Electric Co Ltd, Gunma-ken, Japan). Jurkat cells were cultured at 1 x 10^6 cells per ml medium. Cell viability and cell numbers were recorded by trypan blue exclusion.

2.2 Cell lysis and fractionation

All steps of the different fractionation methods were performed on ice using pre-chilled solutions unless noted otherwise. Centrifugation and incubation were carried out at 4 °C. If the composition of a buffer is not given, no further information was provided by the supplier. All fractions obtained were stored at -20 °C until further use. The fractionations were performed at least three times per method and the protein estimations were carried out in triplicates.

Method 1 (see also Fig. 1): Buffer 2 and 3 were supplemented with protease inhibitor solution (Roche, Mannheim, Germany) before use. Jurkat cells (5 x 10^6) were pelleted for 5 min at 250 x g and washed twice with 3 ml and 1.5 ml buffer 1, respectively. The cell pellet was resuspended in 1.5 ml buffer 2 by vortexing. The suspension was incubated for 10 min while continuously rocking. After 15 min centrifugation at 16,000 x g the supernatant 1 contained the cytosolic proteins. The pellet 1 was solved in 1 ml buffer 3 and the mixture was incubated for 30 min shaking at 1400 rpm in a thermomixer (Eppendorf, Hamburg, Germany). The suspension was centrifuged for 15 min at 16,000 x g. The supernatant 2 contained the membrane proteins, the cell debris containing pellet 2 was discarded. The protein determination for both fractions was carried out using the Bradford Quick Start Protein Assay according to the recommendations of the supplier (Bio-Rad Laboratories GmbH, München, Germany).

Method 2 (see also Fig. 1): All buffers were supplemented

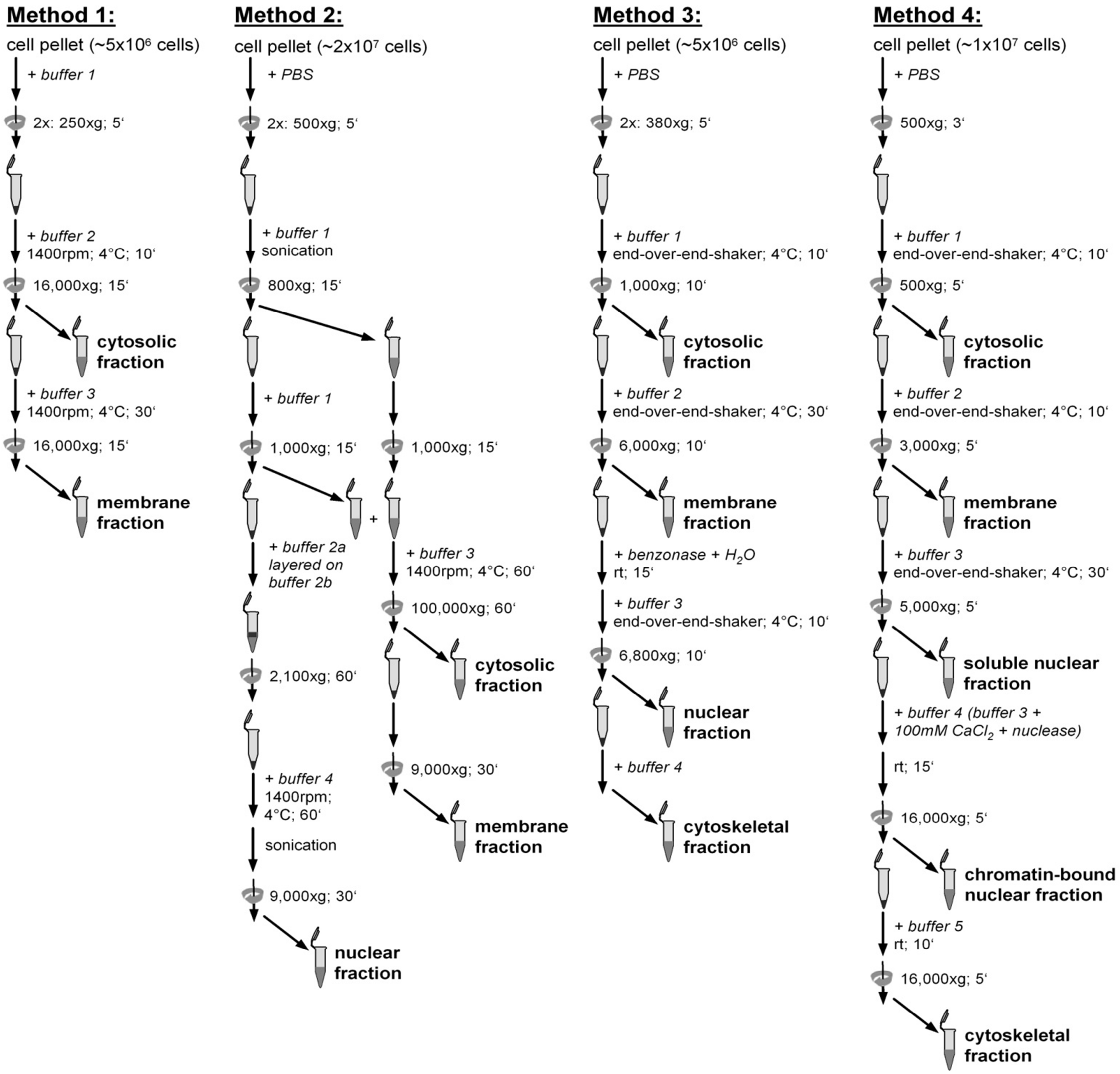


Figure 1. Schematic workflow. All centrifugation and incubation steps of the four different fractionation methods are shown (rpm is given for incubation in a thermomixer, x g for centrifugation).

with 1x protease inhibitor solution and 1 mM DTT directly before use. Jurkat cells (2 x 10^7) were washed twice with PBS and pelleted for 5 min at 300 x g. The cell pellet was resuspended in 1 ml buffer 1 (250 mM sucrose, 50 mM Tris-HCl, 5 mM $MgCl_2$) and cell lysis was performed by sonication on ice (3 times 10 s bursts with intensity ~40% and 30 s breaks). The suspension was centrifuged at 800 x g for 15 min and the pellet 1 was saved to isolate nuclei. The supernatant 1 was centrifuged again at 1,000 x g for 15 min. The obtained supernatant 2 was saved to isolate the cytosolic proteins, whereas pellet 2 was discarded.

The pellet 1 saved for isolation of the nuclei was dissolved in 1 ml buffer 1 and centrifuged at 1,000 x g for 15 min. The obtained supernatant 3 was added to the supernatant 2 for isolating cytosolic proteins and stored on ice until later. The pellet 3 was resuspended in 1ml buffer 2a (1 M sucrose, 50 mM Tris-HCl, 5 mM $MgCl_2$) and layered onto a 3 ml cushion of buffer 2b (2 M sucrose, 50 mM Tris-HCl, 5 mM $MgCl_2$). Afterwards centrifugation at 2,100 x g for 1 h was carried out. The pellet 4 was taken up in 500 µl buffer 4 (20 mM HEPES (pH 7.9), 1.5 mM $MgCl_2$, 0.5 M NaCl, 0.2 mM EDTA, 20% glycerol, 1% Triton X-100) and incubated 1 h shaking at 1400 rpm and 4 °C in a thermomixer. Afterwards the suspension was sonicated again on ice (3 times 10 s bursts with intensity of ~40% and 30 s breaks) and centrifuged at 9,000 x g for 30 min. The supernatant 5 contained the nuclear proteins.

The pooled supernatants 2 and 3 were centrifuged for 1 h at 100,000 x g in an ultracentrifuge. The supernatant 6 contained the cytosolic proteins. The pellet 6 was solved in 0.5 ml buffer 3 (20 mM Tris-HCl, 0.4 M NaCl, 15% glycerol, 1.5% Triton X-100), incubated 1 h shaking at 1400 rpm and 4 °C and centrifuged at 9,000 x g for 30 min. The supernatant 7 contained the membrane proteins. The Lowry-DC-Protein Assay (Bio-Rad Laboratories GmbH) was used to determine the protein content of all fractions obtained with method 2.

Method 3 (see also Fig. 1): All buffers were supplemented with protease inhibitor solution before use. Jurkat cells (5 x 10^6) in a 1.5 ml reaction tube were pelleted for 5 min at 380 x g and washed twice with 1 ml PBS. The cell pellet was mixed with 1 ml buffer 1 and incubated for 10 min on an end-over-end shaker. The lysate was centrifuged at 1,000 x g for 10 min. The supernatant 1 contained the cytosolic proteins. The pellet 1 was resuspended in 1 ml buffer 2 and incubated for 30 min on an end-over-end shaker and centrifuged at 6,000 x g for 10 min. The newly gained supernatant 2 contained primarily membrane proteins. The pellet 2 was mixed with 20 µl distilled water containing 35% benzonase by gently flicking the bottom of the tube. After 15 min incubation at room temperature 0.5 ml buffer 3 was added and the suspension incubated for 10 min on an end-over-end shaker. The insoluble material was pelleted by centrifugation at 6,800 x g for 10 min. The supernatant 3 contained the nuclear proteins. The pellet 3 contained primarily cytoskeletal proteins and was resuspended in 250 µl room temperatured buffer 4. The protein content of all fractions was determined using the BCA Protein Assay Macro Kit (SERVA Electrophoresis GmbH, Heidelberg, Germany).

Method 4 (see also Fig. 1): All buffers were supplemented with protease inhibitor solution before use. Jurkat cells (1 x 10^7) were washed with PBS and pelleted for 3 min at 500 x g in 1.5 ml reaction tubes. The cell pellet was solved in 1 ml buffer 1 and incubated for 10 min on an end-over-end shaker. The lysate was centrifuged at 500 x g for 5 min. The supernatant 1 contained the cytosolic proteins. The pellet 1 was mixed with 1 ml buffer 2, vortexed and incubated for 10 min on an end-over-end shaker. After centrifugation at 3,000 x g for 5 min, the obtained supernatant 2 contained primarily membrane proteins. The pellet 2 was dissolved in 0.5 ml buffer 3, vortexed and incubated for 30 min on an end-over-end shaker. Following centrifugation at 5,000 x g for 5 min the supernatant 3 contained soluble nuclear proteins. Buffer 4 was used at room temperature and prepared by adding 25 µl of 100 mM $CaCl_2$ and 15 µl of micrococcal nuclease to 0.5 ml buffer 3. 0.5 ml buffer 4 was added to the cell pellet 3, vortexed and incubated for 15 min at room temperature. The mixture was vortexed 15 s and centrifuged at 16,000 x g for 5 min. The supernatant 4 contained chromatin-bound nuclear proteins. The pellet 4 was resuspended with 0.5 ml buffer 5, vortexed and incubated for 10 min at room temperature. After centrifugation at 16,000 x g for 5 min the supernatant 5 contained the cytoskeletal proteins. The protein content of all fractions was determined using the BCA Protein Assay Macro Kit following the manufacturer's instructions (SERVA Electrophoresis GmbH, Heidelberg, Germany).

2.3 1D-gel electrophoresis

20 µg protein of each fraction were precipitated 15 min at -20 °C by addition of a 5-fold volume of ice cold acetone. The precipitates were centrifuged at 16,000 x g and 4 °C for 10 min and the supernatant was discarded. The dried pellets were dissolved in SDS-sample-buffer (62.5 mM Tris-HCl (pH 6,8), 10% glycerol, 2% SDS, 5% mercaptoethanol, 0.05% bromophenol blue) and separated by SDS-PAGE on a 4% stacking gel and 12% separation gel run according to standard laboratory procedures. For visual control of successful separation the gels were stained with Coomassie Brilliant Blue G250 after electrophoresis. For protein analysis and MS identification the proteins were allowed to enter only for about 2-3 cm into the gel and cut into 3 gel slices per sample after short staining with Coomassie solution.

2.4 Trypsin digestion and analysis by LC-MS/MS

The gel slices were destained with 50% methanol containing 5% acetic acid. After reduction with 10 mM DTT, proteins were alkylated with 100 mM iodoacetamide and then digested overnight at 37 °C using sequencing grade trypsin (Roche Applied Science, Mannheim, Germany). All membrane fraction containing gel slices were digested in a trypsin solution containing 30% methanol (except method 4). The resulting peptides were extracted two times from the gel with 5% formic acid and 50% acetonitrile. The combined extracts were evaporated, the residual peptides were dissolved in 0.1% FA and the solution was desalted by using C18-StageTips (ZipTipC18, Millipore Corporation, Billerica, MA, USA).

A nano-HPLC system (nanoAquity, Waters, Milford, MA, USA) coupled to a an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) via a nano electrospray ion source (TriVersa NanoMate, Advion, Ithaca, NY, USA) was used for LC/MS/MS analysis. Chromatography was performed with 0.1% formic acid in solvents A (100% water) and B (100% acetonitrile). Samples were injected on a trapping column (nanoAquity UPLC column, C18, 180 µm×20 mm, 5 µm, Waters) and washed with 2% acetonitrile containing 0.1% formic acid and a flow rate of 15 µl/min for 8 min. Peptides were separated on a C18 UPLC column (nanoAcquity UPLC column, C18, 75 µm×100 mm, 1.7 µm, Waters). Peptide elution was conducted using a gradient from 2 - 70% solvent B (0 min - 2%; 5 min - 6%; 45 min - 20%; 70 min - 30%; 75 min - 40%; 80 min - 70%) with a flow rate of 300 nl/min.

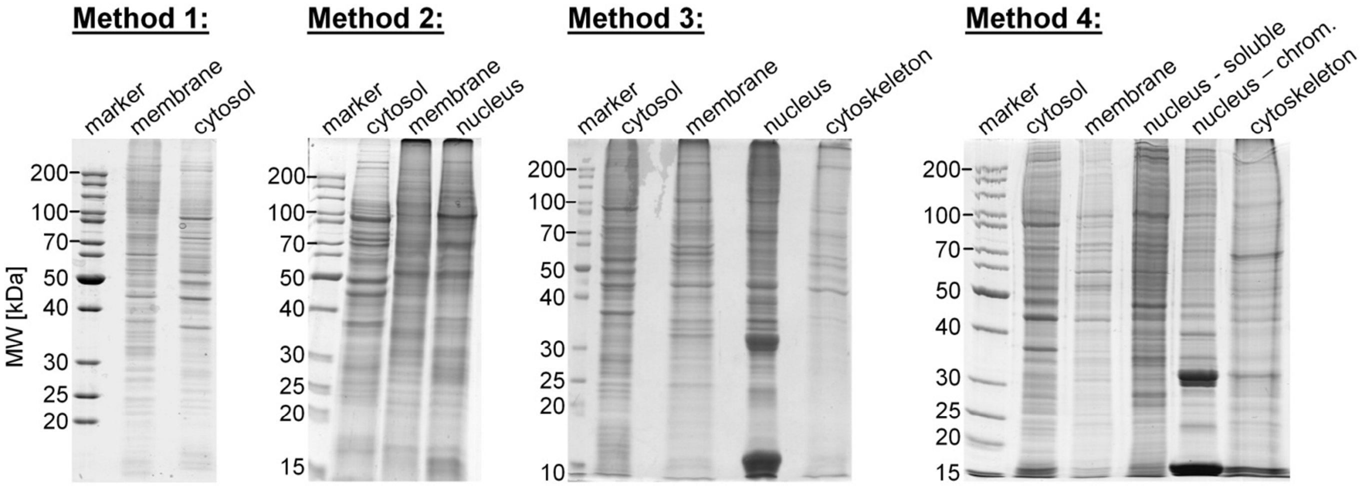


Figure 2. 1D-gels showing the different subcellular fractions. For initial evaluation of the fractions obtained by the four different methods, 20 µg of each protein fraction were separated in a 12% SDS-Gel and stained with colloidal Coomassie. The marker is located on the left hand side of each gel (nucleus-chrom. = chromatin-bound nuclear fraction).

Full scan MS spectra (from 400-1500 m/z, R = 60000) were acquired in positive ion mode in the LTQ-Orbitrap.

Peptide ions exceeding an intensity of 3000 were chosen for collision induced dissociation within the linear ion trap (isolation width 4 m/z, normalized collision energy35, activation time 30 ms, activation q = 0.25). For MS/MS acquisition, a dynamic precursor exclusion of 2 min was applied.

2.5 Data analysis of the mass spectrometric results

MS/MS samples were analyzed by Proteome Discoverer (version 1.0; Thermo Fisher Scientific, San Jose, CA, USA) using the MASCOT search algorithm (version 2.2.06; Matrix Science, London, UK) [19]. Mascot was set up to search a reverse concatenated database of all human proteins annotated in the SwissProt database (version 10/07/2010) assuming the digestion enzyme trypsin. Mascot was searched with a fragment ion mass tolerance of 0.5 Da and a parent ion tolerance of 5 ppm. Carbamidomethylation of cysteine was specified as a fixed modification. Oxidation of methionine and acetylation of the protein n-terminus were specified as variable modifications.

Scaffold 3 (version Scaffold 3_00_03, Proteome Software Inc., Portland, OR, USA) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they exceeded specific database search engine thresholds. Mascot identifications required at least ion minus identity scores of greater than -5 and ion scores of greater than 15. Protein identifications were accepted if they contained at least 2 identified peptides. Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. False discovery rate of proteins was determined to be lower than 0.2% for all samples. Gene ontology annotations were obtained from the EBI GO database (www.ebi.ac.uk/GOA/, version 10/08/2010).

3. Results and Discussion

3.1 Fractionation of Jurkat cells

The workflow of the four different methods used to fractionate Jurkat cells into several cellular compartments is shown schematically in Fig. 1. In method 1, 3 and 4 commercially available kits were used, whereas method 2 uses an adapted protocol from Nature Protocols [18]. All methods rely on cell lysis through sequential addition of different buffers to the cell pellets followed by incubation and centrifugation at different speeds. In method 2 sonication is additionally used to lyse the cells. From method 1 only two different fractions, cytosol and membrane, were obtained. In addition to the three fractions prepared with method 2 – cytosol, membrane and nucleus, a fourth cytoskeletal fraction can be separated with method 3. With method 4 even five different subcellular fractions can be isolated: cytosol, membrane, cytoskeleton, with the nuclear fraction further split into soluble and chromatin-bound nuclear fraction. Method 1 is least time consuming, with about 1.5 hours needed for the fractionation. In approximately 2 hours a fractionation with method 3 or 4 is completed. With at least 3.5 hours of work method 2 is the longest protocol of all four. In addition, method 2 is the most complicated protocol because there are two lines of work steps which have to be performed in parallel while all other methods require only one straight workflow. Moreover, an ultracentrifuge with acceleration up to 100,000 x g is needed for method 2, while a normal table-top centrifuge with up to 16,000 x g is sufficient for all other methods used. Nevertheless, all buffers for method 2 can be prepared in the lab and no expensive kit is needed and the largest number of protein identifications was obtained.

The total amount of obtained protein differed for the various methods (Tab. 1) from 0.78 mg to 3 mg per 1 x 10^7 cells, ranging between 0.5 and 1.57 mg for the cytoplasmic fraction and 0.08 to 0.92 mg for the nuclear fraction. This shows that

Table 1. Protein amounts obtained per 1 x 107 cells in each fraction.

Amount of protein obtained per 10^7 cells [mg]

Method Fraction	1	2	3	4
Cytosol	0.705	0.495	0.496	1.571
Membrane	0.630	0.208	0.135	0.362
Nucleus	–	0.079	0.183	–
Nucleus - soluble	–	–	–	0.520
Nucleus - chromatin-bound	–	–	–	0.401
Cytoskeleton	–	–	0.044	0.148
Total amount of protein	1.335	0.782	0.858	3.002

Table 2. Number of proteins identified in the subcellular fractions.

Method Fraction	1	2	3	4	Ø proteins identified / method
Cytosol	414	657	599	620	573
Membrane	249	458	352	523	396
Nucleus	-	603	258	-	431
Nucleus - soluble	-	-	-	670	670
Nucleus - chromatin-bound	-	-	-	370	370
Cytoskeleton	-	-	618	64	341
Total number of identified proteins	559	1231	1126	1195	1028

there is a rather wide variance in efficiency of the protein isolation. This should also to be taken into account when choosing the fractionation method combinable with the protein detection method used afterwards.

3.2 1D-gel electrophoresis

A first overview of the successful protein separation by the different fractionation methods was obtained by SDS-PAGE. All fractions gained using one method show clearly different band patterns, whereas the same subcellular fractions from different methods have some resemblance in their protein patterns (Fig. 2).

All cytosolic fractions show a comparable band pattern (e.g. five strong bands, of which one is at ~90 kDa, one slightly above 50 kDa, two between 40 and 50 kDa and one at ~38 kDa). Likewise the membrane fractions of method 1, 3 and 4 have a similar band pattern showing a more distinct band at approximately 60 kDa, whereas the separated membrane proteins of method 2 seem to run at slightly different heights. The nuclear fraction from method 2 has as well only partial similarities to the nuclear fractions of methods 3 and 4. The nuclear fraction from method 3 and the nuclear chromatin-bound fraction from method 4 show both two very prominent bands at ~15 and ~30 kDa. These bands are likely to represent histones. The soluble nuclear fraction from method 4 shares a stronger band at ~45 kDa with the nuclear fraction from method 3. As this band is also present in the chromatin-bound fraction, this protein might either be only loosely bound to the chromatin, or, more likely, is not completely separated from the chromatin-bound fraction.

3.3 Identification of proteins

The MS/MS data were analyzed by Proteome Discoverer using the MASCOT search algorithm. The MS/MS based peptide and protein identifications were validated by Scaffold 3. For evaluation of method 4 the two nuclear fractions were combined.

In the cytosolic fractions an average of 573 proteins was identified by all methods (Tab. 2). In the membrane fraction the amount of identified proteins varies a lot between the different methods. With method 1 only 249 proteins were found, whereas 523 proteins were identified with method 4. With method 2 more than the double amount of proteins (603) could be identified in the nucleus compared to method 3 (258). The two different nuclear fractions, soluble and chromatin-bound, obtained with method 4 yielded in 670 and 370 identified proteins, respectively, leading to 750 identified proteins for the nucleus in total (Fig. 3). The amount of cytoskeletal proteins identified with method 3 and 4 ranges from 64 proteins identified with method 4 and up to 618 with method 3. The total numbers of identified proteins were in the same range (between 1126 and 1231) for method 2, 3 and 4 while for method 1 only 559 proteins could be identified in total. Altogether, only the amount of identified proteins in the cytoplasmic and the membrane fractions are comparable within all methods. All methods differ significantly in the amount of proteins identified per fraction as well as in the amount of protein isolated in total.

3.4 Enrichment factor of different fractionation methods

To get a deeper insight into how efficiently each fractionation method worked out, the overlap and intersections in cytosolic, membrane and nuclear fraction were determined and plotted in venn diagrams (Fig. 3). For this aim the two nuclear fractions of method 4, soluble and chromatin-bound, were combined. The most proteins identified in two overlapping fractions were found in cytosol and membrane for method 1 and 3, whereas method 2 and 4 show the biggest overlap in the membrane and nuclear fraction.

Disregarding method 1, because it only yielded two frac-

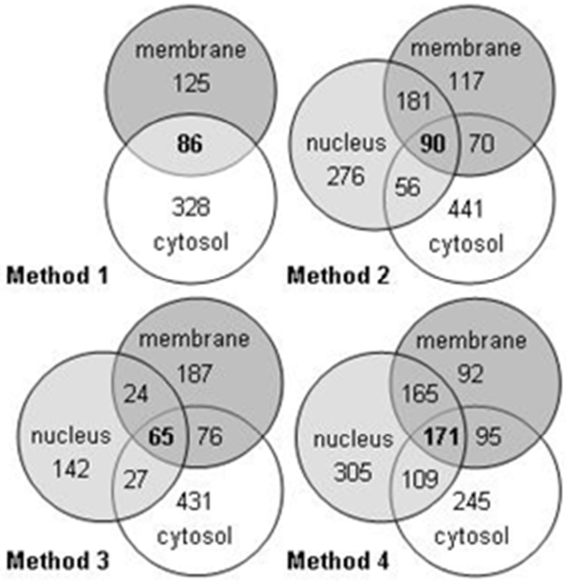


Figure 3. Overlap of proteins identified in the different subcellular fractions. For each of the fractionation methods used, a venn diagram was generated showing the overlap of the proteins identified in more than one fraction.

tions, the most proteins identified in only one fraction could be found with method 3 (80%). 68% of the identified proteins were found in only one fraction with method 2. Method 4 showed the smallest part of proteins identified in only one fraction (54%), while 46% of the identified proteins in this method were found in two or three of the fractions.

In this experiment Gene Ontology (GO) annotations were used by the evaluation program Scaffold 3 to analyze the subcellular localization of each protein identified in the samples. If the proteins identified in one fraction were supposed to be in that fraction according to the GO annotations, they were counted as proteins isolated in the ‘correct’ fraction. To compare how efficient each of the four fractionation methods fractionated the cells, the number of properly isolated proteins in each fraction was calculated. The percentage of the correctly separated proteins out of the total number of identified proteins in each fraction was calculated, too (Fig. 4). The cytosolic fraction was among all four methods the fraction with the most accurately isolated proteins (between 357 and 657 proteins) and comparable percentages about 80%. Between ~30 and 42% of the proteins found in the different membrane fractions where isolated correctly, leading to 74 till 188 isolated proteins in the ‘correct’ fraction in total. For the nuclear fraction 230 up to 345 nuclear proteins could be identified. The percentage of correct nuclear proteins from method 3 was very high with 90%, whereas method 4 showed a high amount of properly isolated proteins because of its two different nuclear fractions. Taking a closer look at transcription factors, there were 12 different ones detected using method 1 and 27 to 32 using method 2 to 4. With method 3 more appropriately isolated cytoskeletal proteins could be identified than with method 4, but the percentage is very low for both methods. The high false positive rate is likely due to the solubilisation of most of the proteins of the last cell pellet, where surely proteins of not completely dissolved membranes or other cellular compartments were inside.

3.5 Discrepancies between the predictions of the evaluation program and the measurements

The Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is a highly abundant protein, which accounts for 10 to 20% of the total cellular protein. It is commonly known as a glycolytic enzyme located in the cytoplasm with a key role in energy production [20]. By intensive research it became obvious that the GAPDH is in reality a multifunctional protein with diverse subcellular localizations in mammalian cells. The GAPDH can be found in the membrane, where it promotes endocytosis and membrane fusion and therefore vesicular secretory transport [21, 22]. Furthermore GAPDH is involved in the nuclear transport of RNA [23] and has the ability to activate the transcription in neurons [24]. Other functions in the nucleus are the assistance in DNA replication and DNA repair [25]. Due to the modulation of the cytoskeleton GAPDH can also be found in the cytoskeletal fraction [26, 27]. Thus the GAPDH can have not only a cytosolic, but also a membrane, nuclear and/or cytoskeletal localization.

According to the GO annotations the GAPDH is located only in the cytoplasm and membrane. This is contradictory to the various localizations described by the literature. In this experiment the GAPDH was found in all fractions obtained with method 2 and 4. With method 3 the enzyme was identified in the cytoplasmic, membrane and cytoskeletal fraction. For all of these three methods the localization in nucleus and

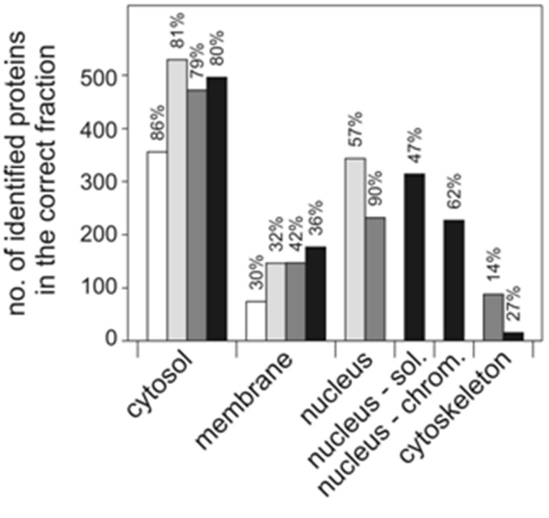


Figure 4. Evaluation of protein localization. For determination of the specificity of each method, the detected proteins in all fractions were analyzed in respect to their assumed localization according to GO terms using Scaffold 3. The bar chart shows the number of proteins identified in each fraction, which were expected to be in that cellular subfraction following Scaffold 3/GO annotations. On top of each bar the percentage of ‘correctly’ isolated proteins in the fractions is given (Method 1 = white bars, Method 2 = light grey bars; Method 3 = dark grey bars; Method 4 = black bars).

cytoskeletal fraction was validated as incorrect because of the incomplete GO annotations. So the GO annotations can only be used to get an overview of the subcellular localizations of a large dataset of proteins. If the localization of a distinct protein is of interest, then a literature search has to be made additionally.

3.6. Potential use of membrane proteins as markers for activation of Jurkat cells

Subcellular fractionation is an ideal tool to enrich and analyze different cellular compartments and low abundant proteins [28]. Due to the fractionation of the cells the less frequent membrane proteins, which otherwise are often covered by the numerous cytosolic proteins in MS measurement, can be identified and analyzed too. Surface proteins in the membrane are especially important for lymphocytes as they are needed for the recognition of antigens and cytokines and activation of other cells. Some of these surface proteins can be used as markers in the evaluation for different purposes. Activated lymphocytes express membrane proteins like CD25, CD69, CD71, and HLA-DR [29-32] which are absent or expressed only in low amounts on resting cells. These proteins are used as activation markers [33]. Similarly a number of known surface proteins like CD2, CD3 and CD5 were identified in the membrane fractions analyzed. In particular for CD2 and CD3 it is long known that they are involved in transmembrane signaling [34]. Despite the known marker, the analysis of the enriched membrane proteins gained by the subcellular fractionation could furthermore lead to the identification of new activation markers, when comparing the membrane proteome of resting and activated cells. Additionally, the identification and subcellular assignment of previously unknown proteins is conceivable. Newly identified membrane proteins may also be used to distinguish between the various T helper cell subpopulations and therefore assist in the process of revealing the different roles of T helper subsets.

4. Concluding remarks

The direct comparison between different methods allows an evidence-based decision on the method of choice for a specific research question. For some studies the mere separation of cytosolic and membrane proteins will be sufficient to perform subsequent analysis. Like for Western blotting method one provides a time-efficient solution of enrichment of certain proteins. When the analysis of the membrane fraction is of special interest the methods 2 or 4 might be favorable. If in the same instance also information about proteins with a nuclear localization it seems advisable to use method 4.

Acknowledgements

The work was supported by the German Research Council (Transregional Collaborative Research Centre 67, subproject Z4) and by the Helmholtz Impulse and Networking Fund through the Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE).

References

1. Nilsson, T., et al., Mass spectrometry in high-throughput proteomics: ready for the big time. Nat Methods. 7(9): p. 681-5.
2. Islinger, M., C. Eckerskorn, and A. Volkl, Free-flow electrophoresis in the proteomic era: a technique in flux. Electrophoresis. 31(11): p. 1754-63.
3. De Palma, A., et al., Extraction methods of red blood cell membrane proteins for Multidimensional Protein Identification Technology (MudPIT) analysis. J Chromatogr A. 1217(33): p. 5328-36.
4. Valot, B., S. Gianinazzi, and D.G. Eliane, Sub-cellular proteomic analysis of a Medicago truncatula root microsomal fraction. Phytochemistry, 2004. 65(12): p. 1721-32.
5. Liu, H., D. Lin, and J.R. Yates, 3rd, Multidimensional separations for protein/peptide analysis in the post-genomic era. Biotechniques, 2002. 32(4): p. 898, 900, 902 passim.
6. Kislinger, T., et al., Multidimensional protein identification technology (MudPIT): technical overview of a profiling method optimized for the comprehensive proteomic investigation of normal and diseased heart tissue. J Am Soc Mass Spectrom, 2005. 16(8): p. 1207-20.
7. Warren, C.M., et al., Sub-proteomic fractionation, iTRAQ, and OFFGEL-LC-MS/MS approaches to cardiac proteomics. J Proteomics. 73(8): p. 1551-61.
8. Michelsen, U. and J. von Hagen, Isolation of subcellular organelles and structures. Methods Enzymol, 2009. 463: p. 305-28.
9. Rio, D.C., et al., Preparation of cytoplasmic and nuclear RNA from tissue culture cells. Cold Spring Harb Protoc. 2010(6): p. pdb prot5441.
10. Sims, N.R. and M.F. Anderson, Isolation of mitochondria from rat brain using Percoll density gradient centrifugation. Nat Protoc, 2008. 3(7): p. 1228-39.
11. Kelson, T.L., J.R. Secor McVoy, and W.B. Rizzo, Human liver fatty aldehyde dehydrogenase: microsomal localization, purification, and biochemical characterization. Biochim Biophys Acta, 1997. 1335(1-2): p. 99-110.
12. Thiery, J., et al., Isolation of cytotoxic T cell and NK granules and purification of their effector proteins. Curr Protoc Cell Biol. Chapter 3: p. Unit3 37.
13. Ramsby, M.L., G.S. Makowski, and E.A. Khairallah, Differential detergent fractionation of isolated hepatocytes: biochemical, immunochemical and two-dimensional gel electrophoresis characterization of cytoskeletal and noncytoskeletal compartments. Electrophoresis, 1994. 15(2): p. 265-77.
14. Sawhney, S., R. Stubbs, and K. Hood, Reproducibility, sensitivity and compatibility of the ProteoExtract subcellular fractionation kit with saturation labeling of laser microdissected tissues. Proteomics, 2009. 9(16): p. 4087-92.
15. Churchward, M.A., et al., Enhanced detergent extraction for analysis of membrane proteomes by two-dimensional gel electrophoresis. Proteome Sci, 2005. 3(1): p. 5.
16. Solstad, T., et al., Quantitative proteome analysis of detergent-resistant membranes identifies the differential regulation of protein kinase C isoforms in apoptotic T cells. Proteomics. 10(15): p. 2758-68.
17. Wang, Y., et al., Cellular uptake of exogenous human PDCD5 protein. J Biol Chem, 2006. 281(34): p. 24803-17.
18. Cox, B. and A. Emili, Tissue subcellular fractionation and protein extraction for use in mass-spectrometry-based proteomics. Nat Protoc, 2006. 1(4): p. 1872-8.

19. Perkins, D.N., et al., Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 1999. 20(18): p. 3551-67.

20. Sirover, M.A., New nuclear functions of the glycolytic protein, glyceraldehyde-3-phosphate dehydrogenase, in mammalian cells. J Cell Biochem, 2005. 95(1): p. 45-52.

21. Tisdale, E.J., Glyceraldehyde-3-phosphate dehydrogenase is required for vesicular transport in the early secretory pathway. J Biol Chem, 2001. 276(4): p. 2480-6.

22. Glaser, P.E. and R.W. Gross, Rapid plasmenylethanolamine-selective fusion of membrane bilayers catalyzed by an isoform of glyceraldehyde-3-phosphate dehydrogenase: discrimination between glycolytic and fusogenic roles of individual isoforms. Biochemistry, 1995. 34(38): p. 12193-203.

23. Singh, R. and M.R. Green, Sequence-specific binding of transfer RNA by glyceraldehyde-3-phosphate dehydrogenase. Science, 1993. 259(5093): p. 365-8.

24. Morgenegg, G., et al., Glyceraldehyde-3-phosphate dehydrogenase is a nonhistone protein and a possible activator of transcription in neurons. J Neurochem, 1986. 47(1): p. 54-62.

25. Meyer-Siegler, K., et al., A human nuclear uracil DNA glycosylase is the 37-kDa subunit of glyceraldehyde-3-phosphate dehydrogenase. Proc Natl Acad Sci U S A, 1991. 88(19): p. 8460-4.

26. Fuchtbauer, A., et al., Actin-severing activity copurifies with phosphofructokinase. Proc Natl Acad Sci U S A, 1986. 83(24): p. 9502-6.

27. Huitorel, P. and D. Pantaloni, Bundling of microtubules by glyceraldehyde-3-phosphate dehydrogenase and its modulation by ATP. Eur J Biochem, 1985. 150(2): p. 265-9.

28. Huber, L.A., K. Pfaller, and I. Vietor, Organelle proteomics: implications for subcellular fractionation in proteomics. Circ Res, 2003. 92(9): p. 962-8.

29. Nakamura, S., et al., Human T cell activation. IV. T cell activation and proliferation via the early activation antigen EA 1. J Exp Med, 1989. 169(3): p. 677-89.

30. Ko, H.S., et al., Ia determinants on stimulated human T lymphocytes. Occurrence on mitogen- and antigen-activated T cells. J Exp Med, 1979. 150(2): p. 246-55..

31. Waldmann, T.A., The structure, function, and expression of interleukin-2 receptors on normal and malignant lymphocytes. Science, 1986. 232(4751): p. 727-32.

32. Neckers, L.M. and J. Cossman, Transferrin receptor induction in mitogen-stimulated human T lymphocytes is required for DNA synthesis and cell division and is regulated by interleukin 2. Proc Natl Acad Sci U S A, 1983. 80(11): p. 3494-8.

33. Caruso, A., et al., Flow cytometric analysis of activation markers on stimulated T cells and their correlation with cell proliferation. Cytometry, 1997. 27(1): p. 71-6.

34. Bagnasco, M., et al., Transmembrane signaling via both CD3 and CD2 human T cell surface molecules involves protein kinase-C translocation. Ric Clin Lab, 1989. 19(3): p. 221-9.

3.2 Optimization of parameters for coverage of low molecular weight proteins

Müller SA, Kohajda T, Findeiß S, Stadler PF, Washietl S, Kellis M, von Bergen M, Kalkhof S.

Anal Bioanal Chem. 2010 Dec;398(7-8):2867-81. Epub 2010 Aug 28.
DOI: 10.1007/s00216-010-4093-x

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 (solely, and the combined usage of trypsin and AspN as endoproteases) and protein separation (gel- or non-gelbased) were investigated. Compared to the standard procedure based solely on the use of urea lysis buffer, ingel 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, collisioninduced 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.

Keywords

LTQ Orbitrap; Nano-HPLC; Nano-ESI-MS; MS; Proteomics; Low molecular weight proteome; *Escherichia coli*

Anal Bioanal Chem (2010) 398:2867–2881
DOI 10.1007/s00216-010-4093-x

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**

Received: 3 June 2010 / Revised: 26 July 2010 / Accepted: 3 August 2010 / Published online: 28 August 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

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

Published in the special issue *Mass Spectrometry (DGMS 2010)* with Guest Editors Andrea Sinz and Jürgen Schmidt.

Electronic supplementary material The online version of this article (doi:10.1007/s00216-010-4093-x) contains supplementary material, which is available to authorized users.

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.

Keywords LTQ Orbitrap · Nano-HPLC · Nano-ESI-MS · MS · Proteomics · Low molecular weight proteome · *Escherichia coli*

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×*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 (nanoAcquity 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 MASCOT 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/MS^3 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).

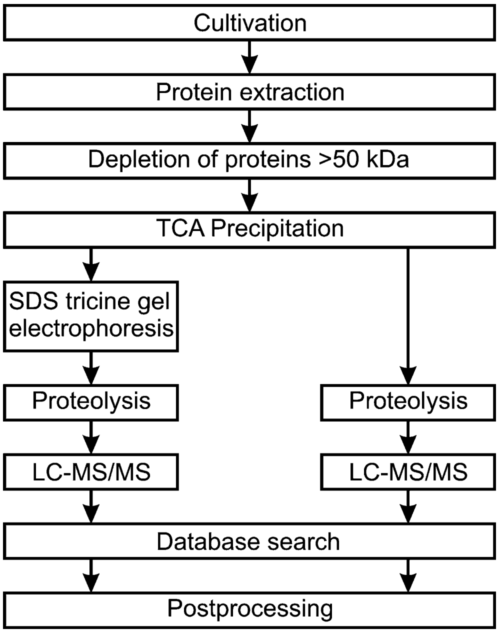


Fig. 1 Experimental workflow

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 *RNAcode*.

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,

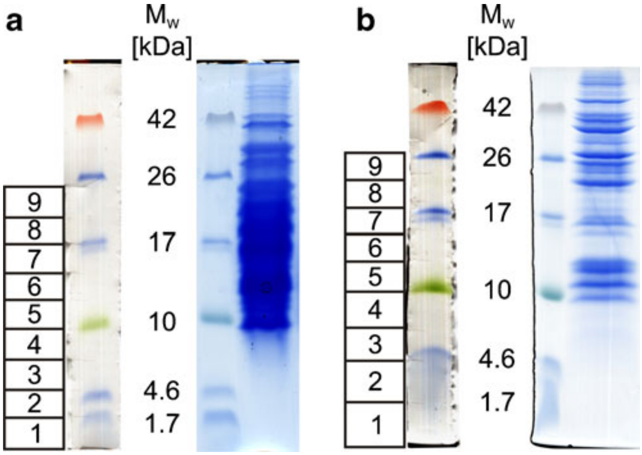


Fig. 2 SDS tricine gel after protein extraction with urea lysis buffer (**a**) and 0.1% TFA (**b**) and subsequent depletion of high molecular weight proteins. Excised bands of the unstained gel part are numbered

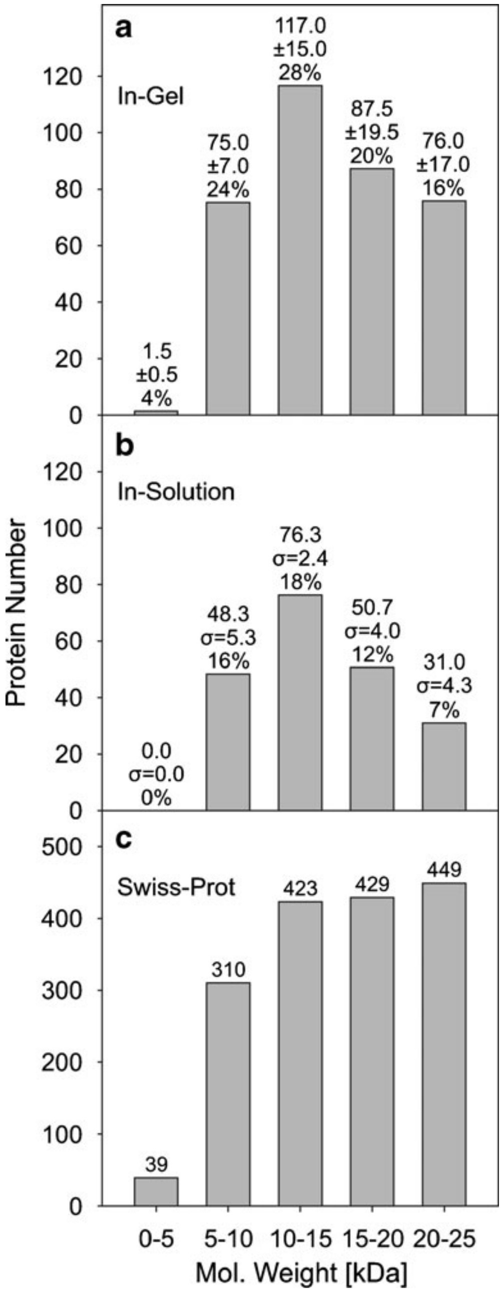


Fig. 3 Average mass distributions of the proteins identified using an in-gel (**a**) or in-solution (**b**) approach in comparison to the SwissProt protein database (**c**)

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 indentified 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

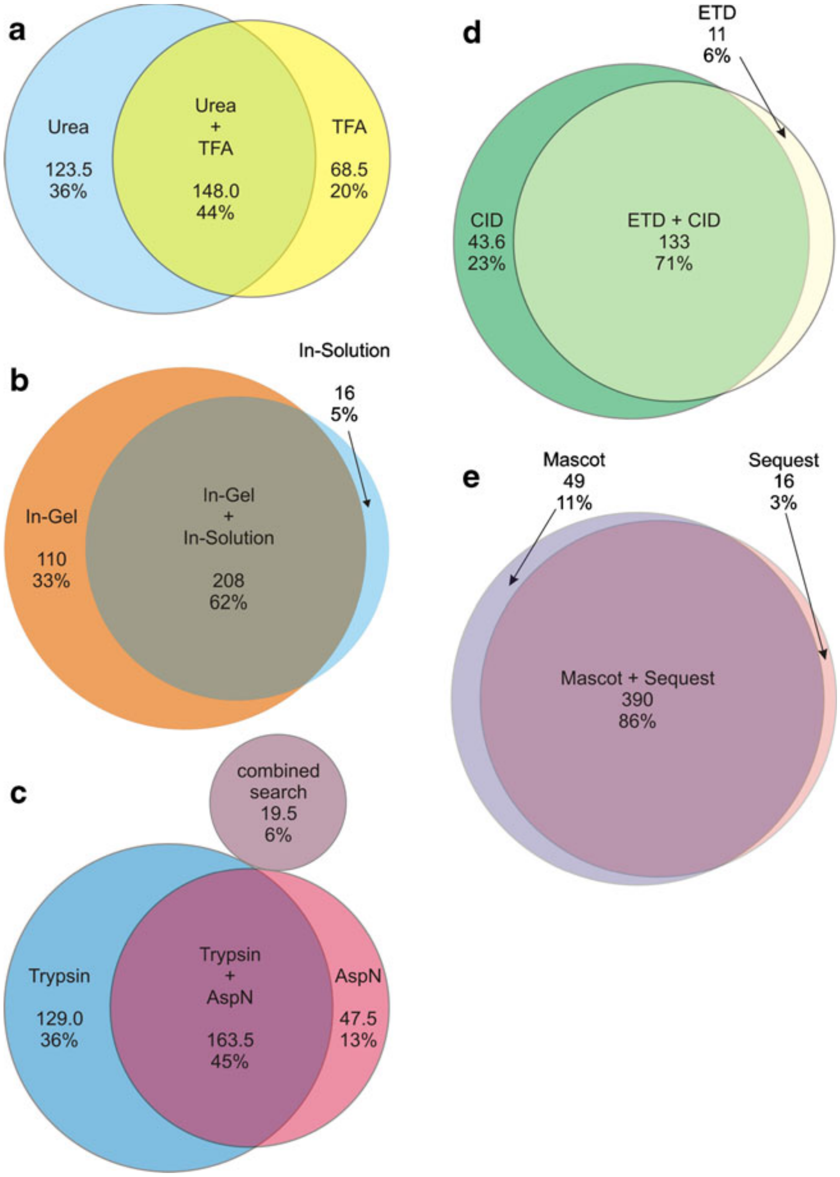


Fig. 4 Influence of different protocol variations. Comparison of average protein identifications after **a** protein extraction with urea lysis buffer or 0.1% TFA, **b** digestion with the in-solution or the in-gel approach, **c** digestion with trypsin or AspN, **d** MS/MS fragmentation and detection 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

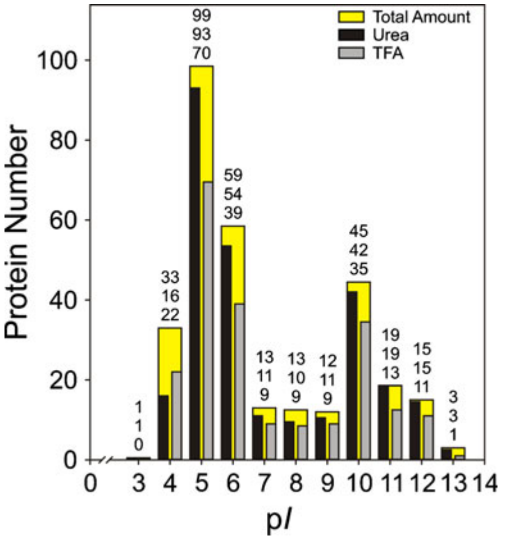


Fig. 5 p*I* distributions of the proteins identified with the in-gel approach after protein extraction with urea lysis buffer or 0.1% TFA in comparison with the total amount of identified proteins

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 (σ=19), 144 (σ=15) and 160 (σ=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].

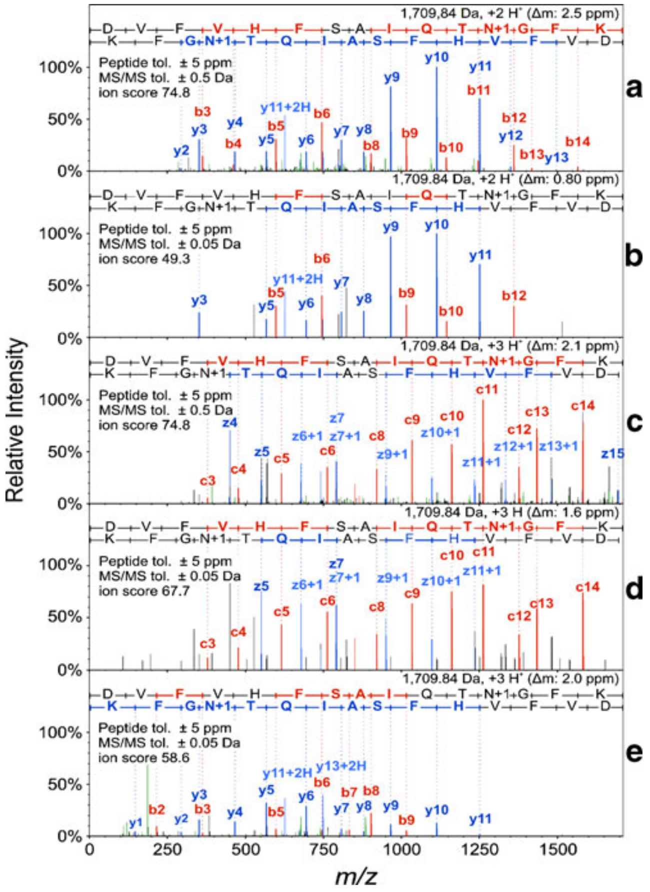


Fig. 6 Comparison of different fragmentation methods after in in-solution proteolysis, as exemplified by the peptide DVFVHFSAIQTnGFK from the cold shock-like protein cspE (**a** IT-CID, **b** FT-CID, **c** IT-ETD, **d** FT-ETD, **e** FT-HCD). *n* denotes an Asn that was found to be deamidated

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.

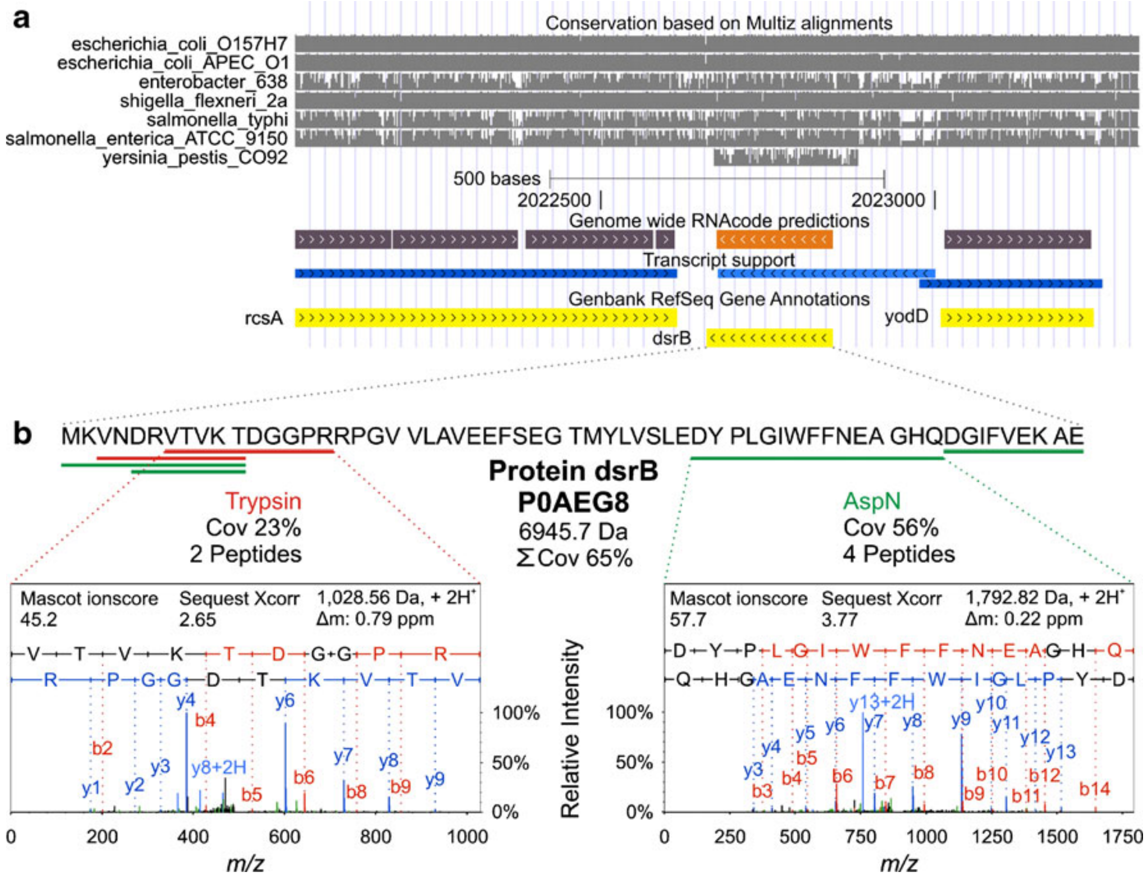


Fig. 7 Evaluation and validation of predicted proteins by **a** RNAcode and **b**. LC/MS/MS. **a** A UCSC screen shot of the genomic context around protein dsrB (Swiss Prot entry P0AEG8) is shown at the *top* with annotated protein coding genes (*yellow*), transcription units as defined by Cho et al. [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.

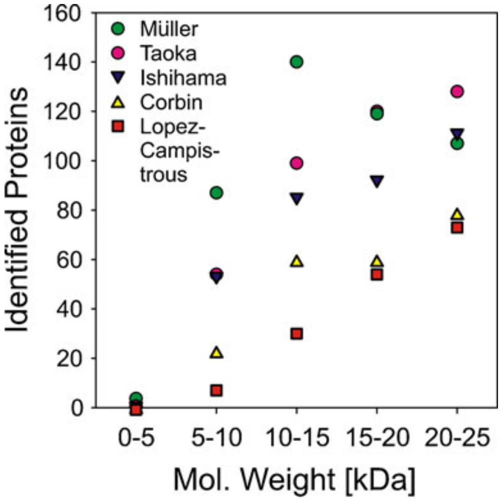


Fig. 8 Comparison of the total number of proteins identified here with the results of selected previous studies focusing on the coverage of the cytosolic proteome of *E. coli*

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*.

Acknowledgment Stephan A. Müller was supported by the German Research Council (Transregional Collaborative Research Centre 67, subproject Z4) and by the Helmholtz Impulse and Networking Fund through the Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE). Further financial support of the project was provided by the Helmholtz Alliance on Systems Biology and COST Action "Systems Chemistry" CM0703. Sven Findeiβ was supported by the DFG (grant No. STA 850/7-1 under the auspices of SPP-1258 "Small Regulatory RNAs in Prokaryotes"). Stefan Washietl was supported by an Erwin Schrödinger Fellowship from the Austrian Science Fund.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

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

3.3 RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data

Washietl S, Findeiss S, **Müller SA**, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N.

RNA. 2011 Apr;17(4):578-94. Epub 2011 Feb 28.
DOI: 10.1261/rna.2536111

Abstract

With the availability of genome-wide transcription data and massive comparative sequencing, the discrimination of coding from noncoding RNAs and the assessment of coding potential in evolutionarily conserved regions arose as a core analysis task. Here we present RNAcode, a program to detect coding regions in multiple sequence alignments that is optimized for emerging applications not covered by current protein gene-finding software. Our algorithm combines information from nucleotide substitution and gap patterns in a unified framework and also deals with real-life issues such as alignment and sequencing errors. It uses an explicit statistical model with no machine learning component and can therefore be applied ‘‘out of the box,’’ without any training, to data from all domains of life. We describe the RNAcode method and apply it in combination with mass spectrometry experiments to predict and confirm seven novel short peptides in Escherichia coli and to analyze the coding potential of RNAs previously annotated as ‘‘noncoding.’’ RNAcode is open source software and available for all major platforms at http://wash.github.com/rnacode.

Keywords

coding sequence; comparative genomics; small peptides; transcriptome

RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data

STEFAN WASHIETL,[1,2,8] SVEN FINDEIß,[3] STEPHAN A. MÜLLER,[4] STEFAN KALKHOF,[4] MARTIN VON BERGEN,[4] IVO L. HOFACKER,[2] PETER F. STADLER,[2,3,5,6,7] and NICK GOLDMAN[1]

[1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, United Kingdom
[2]Institute for Theoretical Chemistry, University of Vienna, A-1090 Wien, Austria
[3]Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center for Bioinformatics, University of Leipzig, D-04107 Leipzig, Germany
[4]Department of Proteomics, Helmholtz Centre for Environmental Research, 04318 Leipzig, Germany
[5]Max Planck Institute for Mathematics in the Sciences, D-04103 Leipzig, Germany
[6]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, 04103 Leipzig, Germany
[7]Santa Fe Institute, Santa Fe, New Mexico 87501, USA

ABSTRACT

With the availability of genome-wide transcription data and massive comparative sequencing, the discrimination of coding from noncoding RNAs and the assessment of coding potential in evolutionarily conserved regions arose as a core analysis task. Here we present RNAcode, a program to detect coding regions in multiple sequence alignments that is optimized for emerging applications not covered by current protein gene-finding software. Our algorithm combines information from nucleotide substitution and gap patterns in a unified framework and also deals with real-life issues such as alignment and sequencing errors. It uses an explicit statistical model with no machine learning component and can therefore be applied "out of the box," without any training, to data from all domains of life. We describe the RNAcode method and apply it in combination with mass spectrometry experiments to predict and confirm seven novel short peptides in *Escherichia coli* and to analyze the coding potential of RNAs previously annotated as "noncoding." RNAcode is open source software and available for all major platforms at http://wash.github.com/rnacode.

Keywords: coding sequence; comparative genomics; small peptides; transcriptome

INTRODUCTION

Distinguishing protein-coding from non-protein-coding sequence is the first and most crucial step in genome annotation. While the coding regions are subsequently investigated for properties of their protein products, a completely different toolkit is applied to the nucleic acid sequences of the noncoding regions. The quality of the analysis of coding potential therefore also affects the annotation of putative noncoding RNA (ncRNA) genes.

Discrimination between coding and noncoding regions poses technical as well as biological challenges not addressed by standard gene finders (Dinger et al. 2008). Ironically, investigators interested in noncoding RNAs hence have repeatedly implemented their own custom solutions to detect coding regions (see, e.g., Mourier et al. 2008; Shi et al. 2009). The *tarsal-less* gene in *Drosophila melanogaster* (also known as *polished-rice* in *Trilobium*) illustrates some of these challenges (Rosenberg and Desplan 2010). The transcript lacks a long open reading frame (ORF) and was originally annotated as noncoding RNA. Later it was found to produce several short, independently translated peptides of 11–32 amino acids (Galindo et al. 2007; Kondo et al. 2007) with a regulatory role in epidermal differentiation (Kondo et al. 2010). How many such short functional peptides may be hidden among RNAs remains an open question (Rosenberg and Desplan 2010).

The detection of protein-coding genes in genomic DNA data is a well-studied problem in computational biology (Burge and Karlin 1998). Using machine learning techniques, sophisticated models of genes have been built that can be used to annotate whole genomes (Brent 2008) and that have been constantly improved over the years (Flicek 2007;

---

[8]**Present address:** MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, MA 02139, USA.

Reprint requests to: Stefan Washietl, MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, MA 02139, USA; e-mail: wash@mit.edu; fax: (617) 253-6652.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2536111.

RNA (2011), 17:578–594. Published by Cold Spring Harbor Laboratory Press. Copyright © 2011 RNA Society.

Brent 2008). Regular community meetings demonstrate a density of high-quality software not usually seen in other fields (Guigó et al. 2006; Coghlan et al. 2008). New types of high-throughput data, such as genome-wide transcription maps, massive comparative sequencing, and meta-genomics studies, however, have led to new challenges beyond classical gene finding. Many transcripts are found that do not overlap known or predicted genes (Carninci et al. 2005; The ENCODE Project Consortium 2007). Statistical methods are necessary to assess the coding potential of this "black matter" transcription (Frith et al. 2006). Similarly, comparative sequencing has revealed a plethora of evolutionarily conserved regions without other annotation (Siepel et al. 2005). A reliable analysis of the coding potential of these regions is an essential step preceding any downstream analysis.

Evolutionary analysis has previously proved useful for de novo detection of coding regions. Various algorithms have been developed to predict coding potential in pairwise alignments (Badger and Olsen 1999; Rivas and Eddy 2001; Mignone et al. 2003; Nekrutenko et al. 2003), and the power of multi-species comparison for the purpose of coding region prediction was demonstrated impressively in yeast (Kellis et al. 2003), human (Clamp et al. 2007), and more recently in 12 drosophilid genomes (Stark et al. 2007; Lin et al. 2008). There is no doubt that these types of analysis are powerful and useful additions to classical gene finders.

In this study, we introduce "RNAcode," a program to detect protein-coding regions in multiple sequence alignments. The initial motivation was to use RNAcode in combination with the widely adopted structural RNA gene-finding program RNAz (Washietl et al. 2005). Similar in spirit to the program QRNA (Rivas and Eddy 2001), the goal is to produce more accurate annotations of ncRNAs by combining information from explicit models for structural RNAs and protein-coding RNAs. The direct identification of conserved regions as protein coding can reduce the number of false-positive ncRNA predictions, which is still the main problem in large-scale screens (Washietl et al. 2007).

More generally, RNAcode was designed to fill a specific gap in the current repertoire of comparative sequence analysis software. It provides the following features for which, to our knowledge, no other program is available: (1) RNAcode relies on evolutionary signatures only and is based on a direct statistical model. No machine learning or training is involved, and it can thus be applied in a generic way to data from all species. (2) It makes use of all evolutionary signatures that are known to be relevant rather than focusing on one particular feature. (3) It predicts local regions of high coding potential together with an estimate of statistical significance in the form of an intuitive *P*-value. (4) RNAcode deals with real-life issues such as sequencing and alignment errors. (5) It is provided as a robust, platform-independent, and easy-to-use C-implementation that is applicable to the analysis of selected regions and that can be integrated in annotation pipelines of larger scale.

ALGORITHM

Evolutionary changes in the nucleotide sequence of coding genes typically preserve the encoded protein. This type of negative (stabilizing) selection leads to frequent synonymous and conservative amino acid mutations, insertions/deletions preserving the reading frame, and the absence of premature stop codons. Our algorithm integrates this information in a unified scoring scheme. It takes as input a multiple nucleotide sequence alignment including a "reference" sequence, which is the one we wish to search for potential coding regions, and predicts local segments that show statistically significant protein-coding potential. Figure 1 shows an overview of the algorithm that is described in more detail in the following sections. First, we introduce a scoring scheme that acts on pairwise alignments and considers amino acid substitutions and gap patterns. Second, we describe how maximum scoring regions under this scheme can be computed for a multiple alignment by considering all pairwise combinations of a reference sequence to the other sequences in the alignment. Third, we indicate how assessment of the statistical significance of these regions can be performed.

Amino acid substitutions

Consider two aligned nucleotide triplets *a* and *b* that correspond to two potential codons. To see if they encode synonymous or biochemically similar amino acids, we can translate the triplets and use amino acid similarity matrices such as the widely used BLOSUM series of matrices (Henikoff and Henikoff 1992). Let A_a and A_b be the translated amino acids of the triplets *a* and *b*, respectively, and $S(A_a, A_b)$ their BLOSUM score. In absolute terms, this score is of little value: Highly conserved nucleotide sequences will get high amino acid similarity scores upon translation even when noncoding.

We need to ask, therefore, what is the expected amino acid similarity score assuming that the two triplets evolve under some noncoding (neutral) nucleotide model. Deviations from this expectation will be evidence of coding potential. To this end, we estimate a phylogenetic tree for the input alignment using a maximum-likelihood method under the well-known HKY85 nucleotide substitution model (Hasegawa et al. 1985). Furthermore, we note that two aligned triplets can have zero, one, two, or three differing positions, i.e., they can have a Hamming distance $h(a, b) \in \{0,1,2,3\}$. It is straightforward to calculate the expected score for a given protein matrix, a parametrized HKY85 background model, and a given Hamming distance *x*:

$$\langle s \rangle_{h=x} = \sum_{\substack{a,b \\ h(a,b)=x}} S(A_a, A_b)\pi_{a_1}\pi_{a_2}\pi_{a_3}\, Prob(a \rightarrow b \mid t). \qquad (1)$$

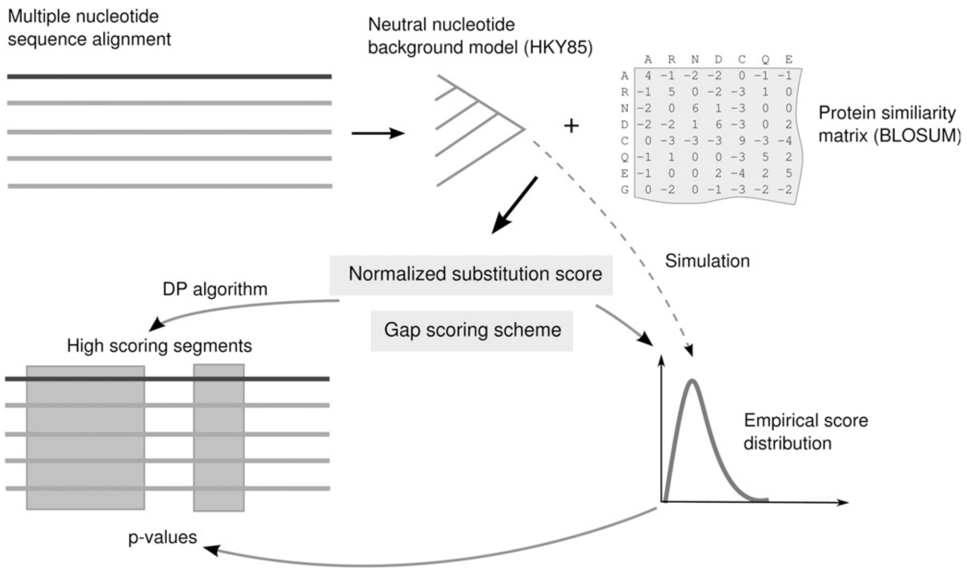


FIGURE 1. Overview of the RNAcode algorithm. First, a phylogenetic tree is estimated from the input alignment including a reference sequence (darker line) under a noncoding (neutral) nucleotide model. From this background model and a protein similarity matrix, a normalized substitution score is derived to evaluate observed mutations for evidence of negative selection. This substitution score and a gap scoring scheme are the basis for a dynamic programming (DP) algorithm to find local high-scoring coding segments. To estimate the statistical significance of these segments, a background score distribution is estimated from randomized alignments that are simulated along the same phylogenetic tree. The parameters of the extreme value distributed random scores are estimated and used to assign *P*-values to the observed segments in the native alignment.

Here a_1, a_2, and a_3 denote the first, second, and third nucleotide in triplet a; π is the stationary frequency in the HKY85 model; and $\text{Prob}(a \rightarrow b|t)$ is the probability that triplet a changes to b after some time t. The analytic expression for this probability is given by Hasegawa et al. (1985). The pairwise evolutionary distance t between two sequences is calculated as the sum of all branch lengths separating the two sequences in the estimated phylogenetic tree.

Put in simple terms, the score $\langle s \rangle$ is the average score over all possible pairs weighted by the probability of observing such a pair under our background assumption. We condition on the observed Hamming distance $h(a, b)$ because this reduces the effect of implicit information on average amino acid frequencies contained in the BLOSUM matrix, and was found to give better results. We can use this expected score $\langle s \rangle$ to normalize our observed scores s arriving at the final protein-coding score σ for an aligned triplet:

$$\sigma = s - \langle s \rangle. \tag{2}$$

To illustrate this with an example, consider the aligned triplets GAA and GAT. The triplets encode glutamic acid and aspartic acid, respectively, and score $s = +3$ in the BLOSUM62 matrix. Furthermore, assume that under some background model, the expected score for pairs with one difference is $\langle s \rangle_{h=1} = -1$. The overall score is thus $\sigma = 3 - (-1) = +4$. The positive score reflects the conservative mutation between the biochemically similar amino acids. A synonymous mutation usually gives the strongest support for negative selection. Since it also gives the highest scores in any protein matrix, there is no need to treat it differently from conservative mutations, and we can score both types of mutations using the same rules. Under this simple scoring scheme, the average triplet score in a coding alignment under negative selection will be positive, while in noncoding alignments, it will be 0 on average. We found that the HKY85 substitution model accurately models noncoding regions for this particular purpose (see the Results section).

Reading frames and gaps

It is straightforward to score an alignment that does not contain gaps. The alignment can simply be translated in all reading frames and the resulting triplets assigned a substitution score σ as described above. Real alignments, however, usually contain gaps. For the purpose of finding coding regions, gap patterns contribute valuable information (Kellis et al. 2004). Negative selection not only acts on the type of amino acid but also on the reading frame that is generally preserved when insertions/deletions occur. Our algorithm incorporates this information into the scoring scheme and, in addition, also deals with practical problems that occur in real-life data such as alignment and sequencing errors. Figure 2 shows some selected gap patterns to illustrate the basic principles. A more formal specification of the algorithm can be found in the Appendix.

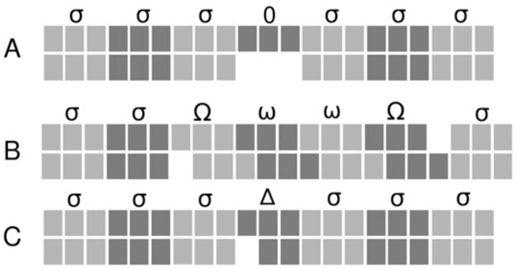


FIGURE 2. Examples of typical gap patterns and scoring paths in a pairwise alignment assumed to be coding. Nucleotides are shown as blocks, codons as three consecutive blocks of the same shading. (*A*) A gap of length three does not change the reading frame and in-frame-aligned codons are scored with the normalized substitution score σ. (*B*) A single gap destroys the reading frame but gets corrected downstream by another gap. The triplets that are out-of-phase because of this obvious alignment error are penalized by the two frameshift penalties Ω and ω. (*C*) A single gap that, in principle, destroys the reading frame is interpreted as a sequence error. Penalized by a high negative score Δ, this frameshift is ignored, and downstream codons are considered to be in-phase.

In real coding regions we will frequently encounter gap lengths that are multiples of three that do not break the coding frame (Fig. 2A). We treat this kind of gap neutrally and give it a score of 0. The aligned triplets before and after the gap are in the same phase and thus can be assigned a score of σ.

Any gap not a multiple of three will result in a frameshift and the sequences are out-of-phase. We assign a penalty score $\Omega < 0$ for the frameshift event and each subsequent aligned triple that is out-of-phase receives an additional smaller penalty $\omega < 0$. Changing the frame back is also penalized, again by Ω (Fig. 2B). The basic idea is that noncoding regions have many frameshifts, and long stretches in the same frame are rare. In contrast, coding regions should not have any frameshifts at all. In real data frameshifts can also be observed in coding regions because of alignment errors. However, they usually get reverted soon by another gap. Consequently, only relatively short regions are out-of-frame.

Gaps in coding regions that are not a multiple of three can also be the result of sequence errors. This is particularly problematic for low-coverage sequencing. In order not to miss substantial parts of true coding regions that appear to be out-of-frame because of a single sequence error, we allow change of the phase and penalize this event with a negative score Δ (Fig. 2C). Clearly, this event should be rare and hence the penalty must be high; the condition $\Delta < 2\Omega$ must be met at least, or otherwise a sequence error event would always be chosen as a more favorable explanation than the frameshifting gaps in the optimization algorithm described below.

Stop codons

Under normal conditions, a reading frame cannot go beyond a stop codon. To reflect this in our algorithm, stop codons in the reference sequence get a score of $-\infty$. We allow relaxation of this for stop codons in the other sequences because if they are of low quality, erroneous stop codons might be observed. These should not automatically destroy a potentially valid coding region but rather be penalized with a relatively large negative score.

Calculating the optimal score for a pairwise alignment

Using the scoring scheme introduced above, we need to find the interpretation of a given alignment as aligned codons in a particular reading frame, out-of-frame codons, and sequence errors that maximizes the score. This is achieved by a dynamic programming algorithm that is described in full detail in the Appendix.

Finding maximum scoring segments in a multiple alignment

To find regions of high coding potential in a multiple sequence alignment, we first consider the pairwise combinations of the reference sequence with each other sequence. In these pairwise alignments, we calculate the optimal score of each alignment block delimited by two columns i and j using the dynamic programming algorithm. Once the maximum scores have been found for each pairwise alignment, we take the average of all pairs and store the optimal scores for the blocks between any two columns i and j of the multiple alignment in a matrix S_{ij} (for details, see Appendix). In this matrix, we identify maximal scoring segments, i.e., segments with a positive score that cannot be improved by elongating the segment in any direction. This approach is meaningful because in noncoding regions the average substitution score is ≈ 0 and gaps can only contribute negative scores.

Statistical evaluation

To assess the statistical significance of high scoring segments, we empirically estimate the score distribution of neutral alignments conditional on the phylogeny derived from the alignment under consideration. Again, we use the phylogenetic tree estimated under the HKY85 model as our null model. We simulate neutral alignments along this tree and calculate high-scoring segments in exactly the same way as for the native alignment. The score distribution follows an extreme value distribution, and we found that it is well approximated by the Gumbel variant with two free parameters (see the Results section). Fitting this distribution allows us to calculate a *P*-value for every high-scoring segment actually observed. This *P*-value expresses the probability that a segment with equal or higher score would be found in the given alignment by chance.

RESULTS

Classification accuracy

We tested RNAcode on six different comparative test sets. These test sets were created from genome-wide alignments (Blanchette et al. 2004; Schneider et al. 2006; Kuhn et al. 2009) typical of those that are widely used for comparative analysis today. The set consisted of alignments of *E. coli* with nine enterobacteria, *Methanocaldococcus jannaschii* with 10 methanogen Archaea, *Saccharomyces cerevisiae* with six other *Saccharomyces* strains, *Drosophila melanogaster* with 11 drosophilid species and three other insects, *Caenorhabditis elegans* with five other nematode species, and *Homo sapiens* aligned to 16 vertebrate genomes. From these alignments, we extracted both annotated coding regions/exons and randomly chosen regions without coding annotation. We then calculated the maximum coding potential score and its associated *P*-value for each alignment. We did not include explicit information on the reading direction, i.e., the coding regions were randomly either in forward or reverse complement direction and both directions were scored.

A typical score distribution (Fig. 3A) shows that random noncoding regions generally do not contain maximal scoring segments with scores higher than 15, whereas coding regions show a wide range of maximal scoring segments of much higher scores. The score efficiently discriminates coding and noncoding regions. Receiver operating curves (ROC) show the sensitivity and specificity of the classification at different score cutoffs (Fig. 3B). In general, we observe the area under the curves (AUC) of the ROCs to be close to 1, i.e., close to perfect discrimination. Usually, the high specificity range (Fig. 3B, insets) is of particular interest for large-scale analysis. At a false-positive rate of 0.05%, for example, we can detect ∼90% of coding regions in all six test sets.

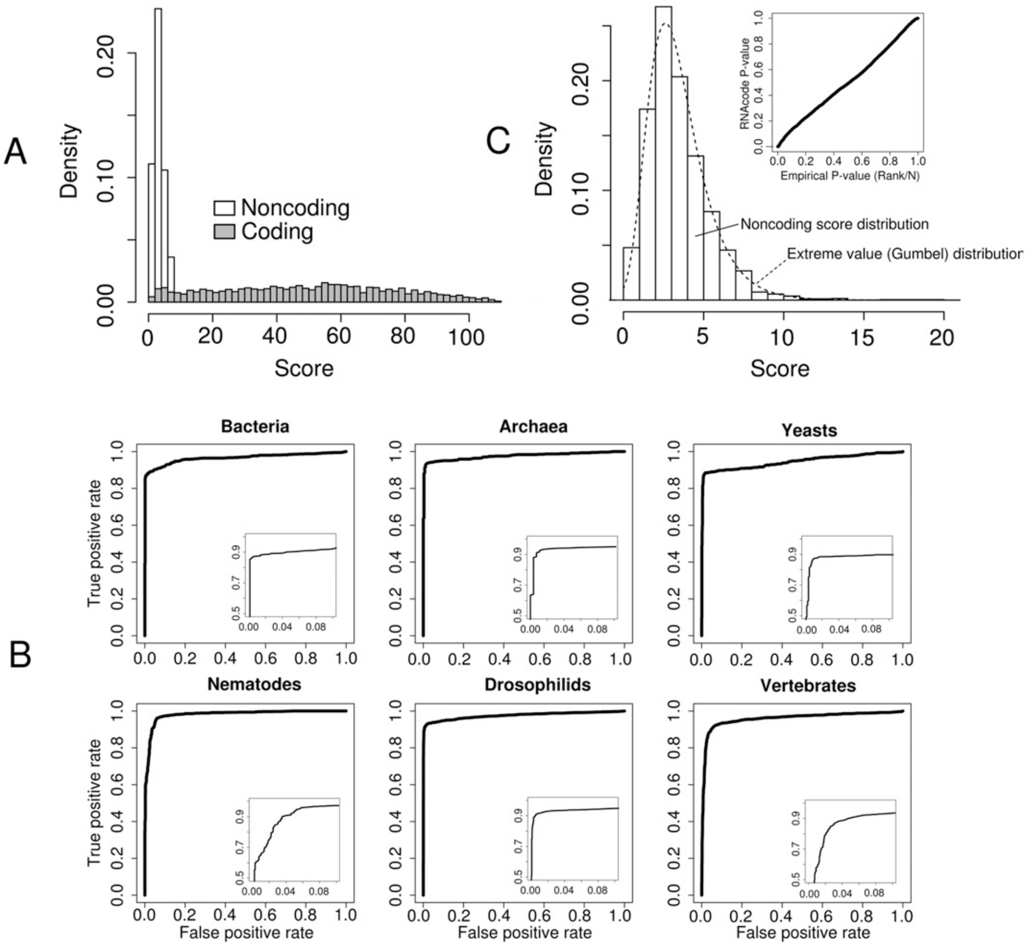


FIGURE 3. RNAcode results on comparative test sets from various species. (*A*) Score distributions of annotated coding regions and randomly chosen noncoding regions in the *Drosophila* test set. (*B*) ROC curves for all six test sets. The full curve for all ranges of sensitivity/specificity from 0 to 1 is shown in the main diagrams. (*Insets*) The high specificity rate with false positive rates from 0 to 0.1. (*C*) Score distribution of noncoding alignments. The same distribution of the *Drosophila* test set as shown in *A* is shown in more detail. The fitted Gumbel distribution is shown as dotted line. (*Upper right* diagram) Comparison of the calculated *P*-values (via simulation and fitting of the Gumbel distribution) to the empirical *P*-values, i.e., the actual observed frequencies in the test set.

Accuracy of *P*-value estimates

The fact that the amino acid similarity scores used in our scoring scheme are adjusted by the expected score under a neutral null model ensures that the RNAcode score is properly normalized with respect to base composition and sequence diversity (phylogeny). In other words, the RNAcode score is independent of sequence conservation and GC content. Unlike other abstract classifiers, it is therefore possible to interpret and compare scores in absolute terms. However, even more important is an accurate estimate of the statistical significance of a prediction. Similar to the well-known statistics of local alignments (e.g., BLAST), RNAcode scores follow an extreme value distribution (Fig. 3C). This allows us to calculate *P*-values (see the section "Statistical Evaluation").

To test the accuracy of this approach, we compared *P*-values calculated by this procedure to empirically determined *P*-values on a set of noncoding *Drosophila* alignments. To this end, we calculated the *P*-value for each alignment in the set and compared each to the proportion of alignments with better scores than the given one (Fig. 3C, inset). The excellent agreement of the *P*-values calculated by RNAcode and the actual observed frequencies confirms that the Gumbel distribution is an accurate approximation of the background scores. In addition, it also confirms that the HKY85 nucleotide substitution model and our simulation procedure accurately model real noncoding data.

Influence of parameter choice

The frameshift penalties in our algorithm are user-definable parameters. We found that the algorithm is relatively robust with respect to the particular choice of these parameters. Three different sets of parameters gave almost identical results (Supplemental Fig. 1). However, ignoring information from gap patterns altogether by setting all penalties to a neutral value of zero leads to a drop in classification performance. This shows that gap patterns do, indeed, hold relevant information for classification although most information is contained in the substitution score, a result that is consistent with previous reports (Lin et al. 2008).

Comparison to other comparative metrics

To further evaluate the performance of our new approach, we have created a more extensive data set that systematically covers alignments with varying numbers of sequences and different conservation levels (see Materials and Methods). On this data set, we have compared the RNAcode substitution score to two other commonly used metrics that are based on evolutionary signatures.

The ratio of nonsynonymous (*dN*) to synonymous substitutions (*dS*) gives information on the type of selection acting on a protein-coding sequence (Yang and Nielsen 2000). A low *dN/dS* ratio indicates negative selection, which was found to be a reliable way to detect coding regions in pairwise (Nekrutenko et al. 2003) and multiple alignments (Lin et al. 2008). The structure of the genetic code leads to a periodic pattern of evolutionary rates (Bofkin and Goldman 2007), another characteristic of protein-coding regions that was applied, for example, to assess the coding potential of unannotated transcripts in *S. cerevisiae* (David et al. 2006) and in human in the ENCODE pilot project (The ENCODE Project Consortium 2007).

We calculated the *dN/dS* ratio for all alignments in our data sets using a maximum likelihood method (Yang and Nielsen 2000). To quantify the substitution rate periodicity, we re-implemented a likelihood test described previously (Materials and Methods) (The ENCODE Project Consortium 2007). In essence, it compares a null model with equal rates for each nucleotide position to an alternative model allowing for a periodic pattern "...ABCABCABC..." of rates. It thus captures the periodicity of the codons without the need to explicitly determine the reading direction or frame.

We found that the RNAcode substitution score consistently outperforms the *dN/dS* ratio and the periodicity score (Fig. 4). The difference is particularly pronounced for alignments of low sequence conservation. These alignments presumably contain more conservative amino acid substitutions, which RNAcode—in contrast to the *dN/dS* ratio—can take advantage of. Interestingly, the fact that the *dN/dS* ratio and the periodicity score are calculated over a phylogenetic tree for the complete alignment does not lead to better performance than the RNAcode score, which is calculated from pairwise comparisons.

Influence of alignment properties

The performance of RNAcode depends on the evolutionary information contained in the alignment. The results shown in Figure 4 illustrate this dependency in terms of alignment size and sequence diversity. In the extreme case of pairwise alignments with very low sequence diversity (90%–100% mean pairwise sequence identity), the classification performance is relatively poor (AUC $<$ 0.9). Adding more sequences ($N = 4$) and higher sequence diversity (identities below 90%) leads to much better performance (AUC $\approx$ 0.99). Adding even more sequences ($N = 8$) results in further improvement and almost perfect discrimination. We conclude that alignments with as few as four sequences that are <90% identical will give satisfactory results in practical applications of RNAcode.

The alignment method used might affect performance. All tests in this study were run on genome-wide alignments generated by MultiZ (Blanchette et al. 2004). We found that re-aligning with other commonly used alignment programs did not change our results (Supplemental Fig. 2).

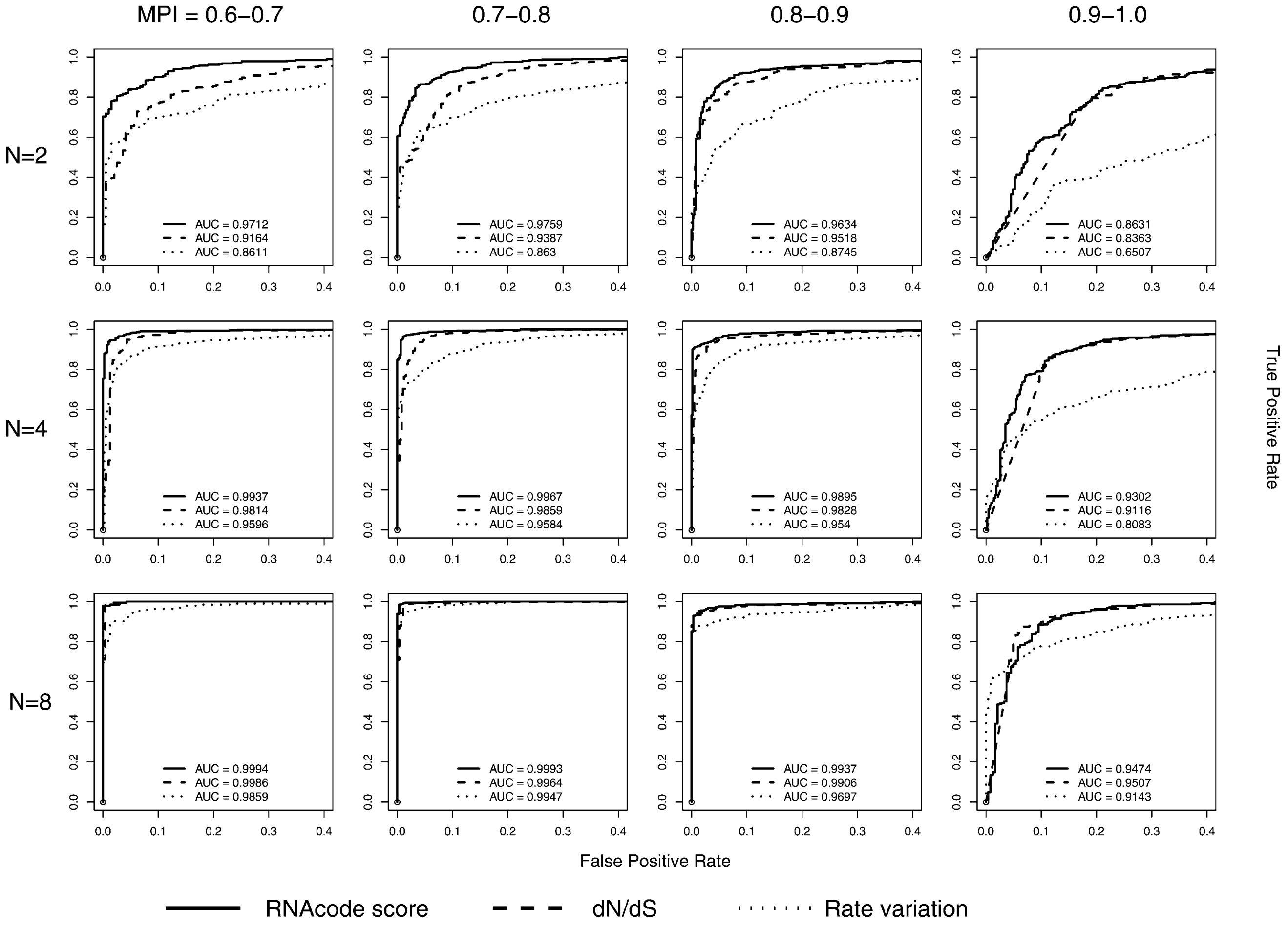


FIGURE 4. Comparison of the RNAcode substitution score with other comparative metrics. The ROC curves show the classification performance of the *dN/dS* ratio, substitution rate variation, and the average substitution score σ used by RNAcode. Results are shown for alignments of length 30 from vertebrates, archaebacteria, yeasts, and drosophilid species grouped by the number of sequences in the alignment (*N*) and the mean pairwise sequence identity (MPI). The area under the ROC curve (AUC) as a measure for classification performance is shown for all methods and sets.

Automatic annotation of *Drosophila* genome

The main purpose of RNAcode is to classify conserved regions of unknown function, to discriminate coding from noncoding transcripts, and to analyze the coding potential in non-standard genes (e.g., short ORFs or dual-function RNAs; see below for examples). RNAcode's algorithm is built on a direct statistical model that deliberately ignores any species-specific information and does not resort to machine learning. RNAcode is thus not optimized for the genome-wide annotation of protein-coding genes in well-known model organisms. However, to demonstrate that RNAcode is also efficient for this purpose and to study our algorithm in direct comparison to today's best gene finders, we automatically annotated chromosome 2L ($\approx$23 Mb) of the *D. melanogaster* genome. We ran RNAcode with standard parameters and a *P*-value cutoff of 0.001 on MultiZ alignments available at the UCSC Genome Browser and compared the results to FlyBase (Drysdale and FlyBase Consortium 2008) annotation. Of the 10,535 annotated coding exons in FlyBase, 9245 overlapped (by at least one nucleotide) with an RNAcode prediction (sensitivity 87.8%). In total, RNAcode predicts 13,166 high-scoring coding regions with $p < 0.001$. Of these, 12,207 had overlap with one of the annotated exons, i.e., 959 were false positives (specificity: 92.7%). This result is surprisingly close to the currently best "full" gene finders. In the same overlap statistics, CONTRAST (Gross et al. 2007) achieves 91.0%/97.0% (sensitivity/specificity) and NSCAN (Gross and Brent 2006) 91.8%/97.2%. These algorithms can take advantage of species-specific features such as splice site signals, codon usage, exon length distributions, etc., information that is not available when studying non-model organisms or atypical genes (see below for examples). Our results show that evolutionary events alone hold a considerable amount of information and that RNAcode efficiently makes use of it.

Novel peptides in *E. coli*

The *E. coli* genome was one of the first completely sequenced genomes and is generally well annotated. However, even in this compact and extensively studied genome, the protein annotation is far from perfect. Protein gene annotation is largely based on compositional analysis and homology with known protein domains. The statistical power of these criteria is limited for small proteins. Standard gene-finding software is usually run with an arbitrary cutoff of 40–50 amino acids to avoid an excess of false positives and suffers from the lack of training data of verified short peptides.

Here, we attempted to produce a set of predictions based on evolutionary signatures only. We created alignments of the *E. coli* reference strain K12 MG1655 to 53 other completely sequenced enterobacteria strains including *Erwinia*, *Enterobacter*, and *Yersinia* (see Materials and Methods) (Supplemental Table 1). A screen of these alignments with RNAcode and a *P*-value cutoff of 0.05 resulted in 6542 high-scoring coding segments. We discarded all predictions that overlapped annotated proteins. For the remaining RNAcode predictions, we tried to identify a complete ORF (starting with AUG and ending in a stop codon) in the *E. coli* reference sequence (see Materials and Methods). This step is necessary because the boundaries of high-scoring segments usually do not correspond exactly to the ORF (a main problem here is the relatively short alignment blocks produced by MultiZ, which do not always cover an ORF over its full length). This procedure gave 35 potential new protein-coding genes between 11 and 73 amino acids in length (see Supplemental Table 2).

To assess the quality of these predictions, we first looked at the overall sensitivity of our screen on already annotated proteins. Of the 4267 RefSeq proteins, 3987 overlapped with a RNAcode prediction (sensitivity 93.4%). Hemm et al. (2008) revisited the annotation of small proteins in *E. coli* and found 18 novel examples using a combination of different bioinformatics and experimental methods. In a set of 18 new and 42 literature-curated proteins between 16 and 50 amino acids compiled by Hemm et al. (2008), 30 (50.0%) overlap with RNAcode predictions. These results show that our screen not only gives almost perfect results on typical *E. coli* proteins, but also recovers a substantial fraction of small proteins that are particularly difficult to detect. Moreover, our final list of 35 candidates for novel proteins is rather short and shows the high specificity in this screen.

For additional support, we compared our list of predicted candidates with publicly available transcriptome data (Tjaden et al. 2002; Cho et al. 2009). These data sets cover a broad range of experimental conditions and therefore reflect a comprehensive genome-wide transcription map of *E. coli*. Eight candidates (23%) overlap with regions that show clear evidence for transcription (Supplemental Table 2).

To further substantiate our predictions, we used mass spectrometry (MS) as a direct experimental test for the existence of the novel peptides in *E. coli* cells. MS is particularly well suited to screen simultaneously for a large set of proteins without resorting to cloning or recombinant expression (Aebersold and Mann 2003). Many, but by no means all, proteins of an organism are expressed and detectable under the actual applied conditions by current MS-based proteomics. Detecting small peptides in complex protein mixtures is particularly challenging for various reasons. Compared to the overall protein expression level, short peptides often show low abundance, they are easily lost using standard proteomic protocols, and only a limited number of proteolytic peptides can be obtained (Klein et al. 2007). To meet these challenges, we developed a protocol that is specifically optimized for small proteins by avoiding sample loss by a simple extraction method and a combined purification and enrichment step using filtration (Müller et al. 2010; Materials and Methods). In order to improve the reliability of our results, we applied two different buffer systems for extractions, and for an improved coverage of peptides, we used two different proteases. This strategy led to an increased detection rate as well as to higher confidence in the hits by confirmation in independent experiments.

Using this protocol, we were able to identify 455 small molecular weight proteins (MW < 25 kDa) representing 27% of the 1672 known *E. coli* proteins below this size listed in the SWISS-PROT protein database (UniProt Consortium 2010). In a search against the list of 35 newly predicted proteins, we obtained evidence for the expression of seven candidates (20%) (Supplemental Table 3). For the rest of the candidates, we cannot distinguish whether they are false-positive RNAcode predictions or false negatives in the MS experiment. However, considering that the success rate of the MS experiments is roughly the same on known and predicted proteins (27% and 20%, respectively), we would expect a good fraction of our candidates to be true proteins not detectable by this particular growth conditions and MS approach.

Although it is not possible to give a conclusive statement on all predictions without additional experiments, compelling evidence from evolutionary analysis, transcriptomics data, and the MS experiments strongly suggests that several of the candidates are bona fide proteins. Figure 5 shows two examples in more detail. In both cases, RNAcode reported short but statistically highly significant ($p \approx 10^{-8}$ and $p \approx 10^{-6}$, respectively) signals between two well-annotated proteins. The loci overlap with transcribed regions as determined by Cho et al. (2009). In addition, our MS experiments detected several proteolytic fragments that can be assigned to these proteins.

The coding potential of "noncoding" RNAs

In addition to assisting and complementing classical protein gene annotation strategies, a major area of application of RNAcode is the functional classification of individual

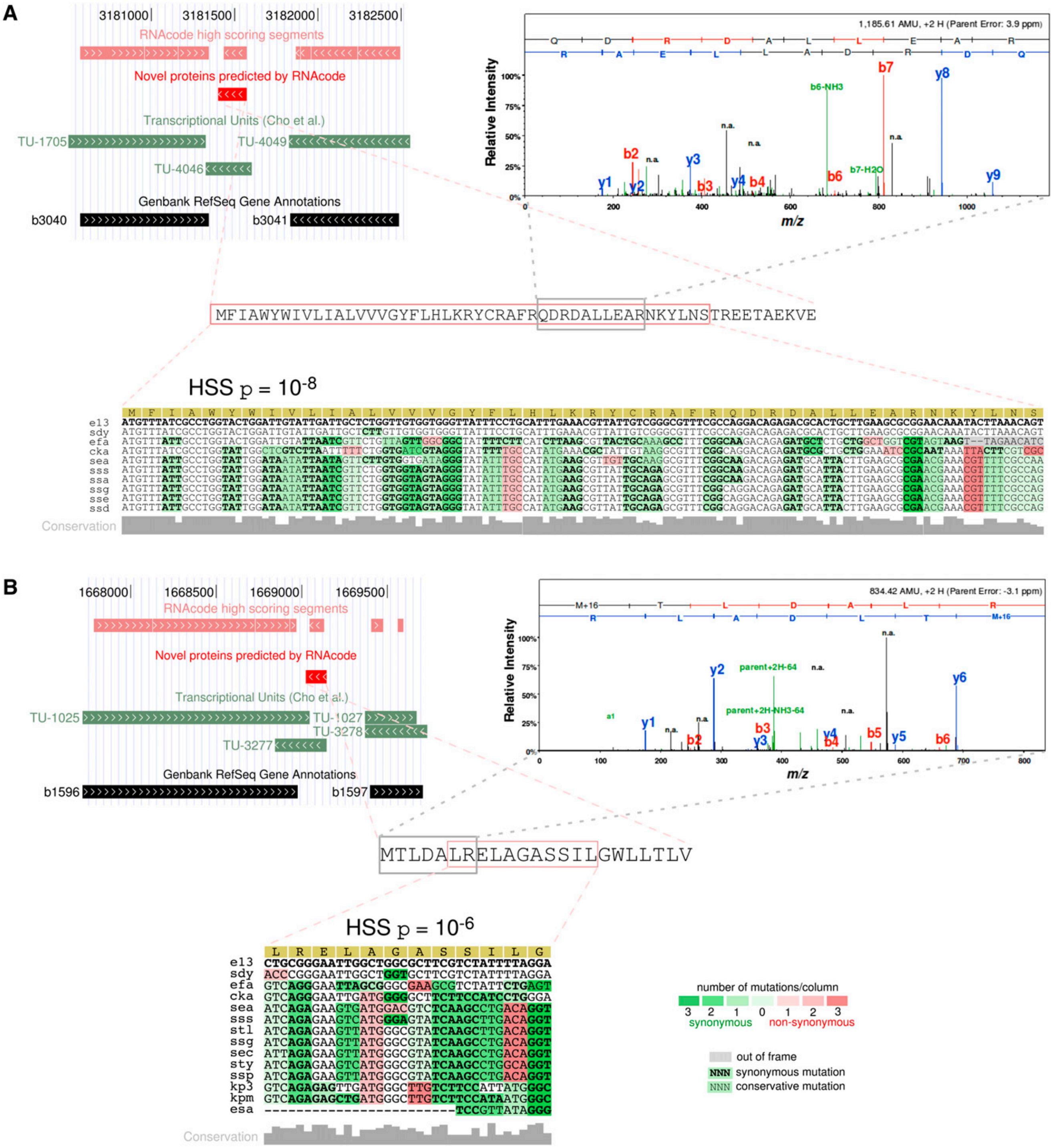


FIGURE 5. Examples of novel short proteins in *Escherichia coli*. Sequence, genomic context, the high-scoring RNAcode segment, and fragment ion mass spectra are shown. Genome browser screenshots were made at http://archaea.ucsc.edu (Schneider et al. 2006). Arrows within annotated elements indicate their reading direction. The shading of mutational patterns was directly produced by the RNAcode program. The full species names for the abbreviations can be found in Supplemental Table 1. The mass spectra are shown for two selected proteolytic peptides, which were scored with 80% probability and used in combination with the detection of additional peptides to confirm the expression of the candidates (for details, see Supplemental Table 3). The proteins shown in *A* and *B* correspond to candidates 28 and 19, respectively, listed in Supplemental Tables 2 and 3.

conserved or transcribed regions. As an illustrative example, we analyzed the bacterial RNA C0343, which is listed in the Rfam database (Gardner et al. 2009) as noncoding RNA (ncRNA) of unknown function. The RNA originally detected by Tjaden et al. (2002) is also detected as transcript in the study of Cho et al. (2009) (Fig. 6). In our screen of the *E. coli* genome, we found a high-scoring coding segment with $p \approx 10^{-9}$ overlapping the C0343 ncRNA. The prediction corresponds to a potential ORF encoding 57 amino acids (Fig. 6A; candidate 8 in Supplemental Table 2). Analysis of the secondary structure using RNAz (Gruber et al. 2010) does not give any evidence for a functional RNA. Given the strong coding signal, we conclude that the "noncoding RNA" C0343 is, in fact, a small protein. This is also confirmed by our MS experiments that detected proteolytic fragments of this protein in *E. coli* cells (Supplemental Table 3).

To test RNAcode on another example from Rfam, we analyzed RNAIII, an ncRNA known to regulate the expression of many genes in *Staphylococcus aureus* (Boisset et al. 2007). In addition to its role as regulatory RNA, the RNAIII transcript also contains an ORF coding for the 26-amino-acid-long delta-haemolysin gene (*hld*). We ran RNAcode with standard parameters on the Rfam seed alignment. It reports one high-scoring segment below a *P*-value cutoff of 0.05, which corresponds to the *hld* gene (Fig. 6B). The annotated alignment shows that the ORF is highly conserved with only few mutations. Nevertheless, these few mutations are sufficient to yield a statistically significant signal that allows RNAcode to locate the correct ORF. Again, we also ran RNAz on the alignment, which reports a conserved RNA secondary structure with a probability of 0.99. The combination of RNAcode and RNAz clearly shows the dual function of RNAIII. This example demonstrates how RNAcode can assist the classification of ncRNAs in particular for non-standard and ambiguous cases (Dinger et al. 2008).

As another example, we analyzed the SR1 RNA of *Bacillus subtilis* that was originally found by Licht et al. (2005) (Fig. 6C). Although the investigators noticed a potential short ORF in the transcript, the corresponding peptide could not be detected. Further experiments (Heidrich et al. 2006, 2007) clearly showed a function of SR1 in the arginine catabolism pathway by RNA/RNA interaction with the *ahrC* mRNA, thus confirming its nature as functional noncoding RNA. Using RNAcode, we found clear evolutionary evidence for a well-conserved small peptide deriving from SR1 ($p \approx 10^{-12}$), arguing for a role as dual-function RNA. Only recently, Gimpel et al. (2010) showed that the *gapA* operon is regulated by a short peptide encoded in SR1, which exactly corresponds to the high-scoring coding segment found by RNAcode (Fig. 6C).

Finally, we analyzed the *tarsal-less* gene mentioned in the Introduction (Galindo et al. 2007; Kondo et al. 2007). The small peptides produced by this unusually organized polycistronic gene were overlooked originally, and it was thought to be noncoding. Analysis using RNAcode predicts three significant high-scoring coding segments (*P*-values = 2.4×10^{-5}, 5.5×10^{-5}, 0.010) in this transcript, covering one known peptide and partially covering a second. Using a relaxed *P*-value cutoff, four of the five known peptides are identified (Supplemental Fig. 3).

Implementation and performance

RNAcode is implemented in ISO C. The program takes an alignment in either CLUSTAL W format or MAF format (popularized through the UCSC Genome Browser). It outputs relative coordinates and/or genomic coordinates of predicted coding regions, the raw score, and the *P*-value in either a human readable tabular format or as standard GTF annotation format. In addition, RNAcode offers an option to generate color annotations of the alignment. This kind of visualization helps to quickly identify mutational patterns, which allows visual discrimination between alignments of high and low coding potential. RNAcode produces publication-quality vector graphics in Postscript (EPS) format (see, e.g., Figs. 5, 6). To generate the color annotated images, it is not enough to know just the region and score of the high-scoring segments, but we also have to infer the state path that led to this prediction. Therefore, we have also implemented the backtracking step for the dynamic programming algorithm. In addition to the mutation patterns, this allows annotation of regions that are likely to be out-of-phase and the location of potential sequence errors inferred by the algorithm.

The dynamic programming algorithm used to score an alignment of N sequences with n columns requires $\mathcal{O}(N \cdot n^2)$ CPU time and memory. Large genomic alignments are therefore broken up into windows of several hundred nucleotides in length in practical applications (see Materials and Methods). There is nothing to be gained by feeding RNAcode with alignment windows that are longer than actual contiguous pieces of coding sequence.

The analysis of 1 Mb of *Drosophila* MultiZ alignments with up to 12 species (10,426 alignment blocks) took 2 h and 6 min on a single Pentium 4 CPU running at 3.2 GHz. This includes calculation of *P*-values with 100 randomizations for all predictions. However, it is generally not of interest to calculate exact *P*-values for hits that are clearly not statistically significant. Therefore, we added an option to stop the sampling procedure as soon as too many of the randomizations score better than the original alignment (e.g., for 1000 randomizations and a significance level of $p < 0.05$, the sampling would stop after 50 random alignments with a better score than the native alignment). Depending on the density of coding regions in the input alignments, this simple heuristic can speed up the process considerably. Using this option, the 1 Mb of fly alignments could be scored in 1 h and 4 sec without any loss in sensitivity or specificity.

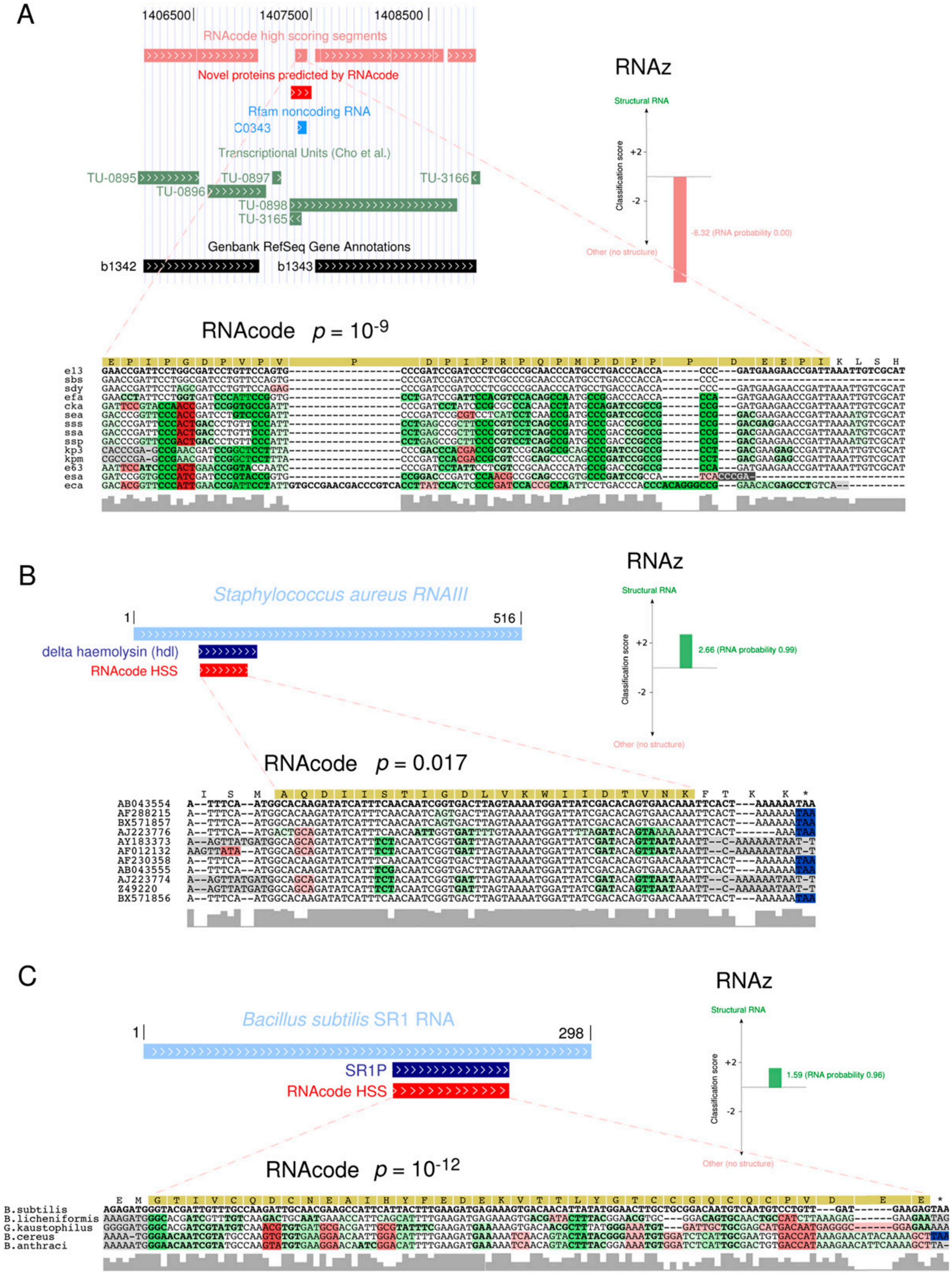


FIGURE 6. Examples of ambiguities between the coding and noncoding nature of three RNAs. (*A*) The RNA C0343 from *E. coli* is listed as a noncoding RNA in Rfam. However, it overlaps with an RNAcode-predicted coding segment. While there is no evidence for a RNA secondary structure according to the RNAz classification value, the highly significant RNAcode prediction and MS experiments suggest that C0343 is an mRNA and not an ncRNA. (*B*) RNAIII of *Staphylococcus aureus* (Rfam RF00503) contains a short ORF of a hemolysin gene. RNAcode predicts the open reading frame at the correct position, while RNAz clearly detects a structural signal. These results are consistent with the well-established dual nature of this molecule. (*C*) The *Bacillus subtilis* RNA SR1 is known to have function on the RNA level by targeting an mRNA. RNAcode detects a short ORF that was shown by Gimpel et al. (2010) to produce a small peptide and is thus another example of a dual-function RNA.

DISCUSSION

We have introduced RNAcode as a comparative genomics tool for the identification of protein-coding regions. Inspired by our own experiences in analysis of comparative sequence data in the context of ncRNA annotation, the design emphasized practicability and robustness and focused on the single task of discriminating protein-coding from noncoding regions. RNAcode therefore is not a gene-finder. By design, it neither uses nor predicts any features related to transcript structure such as splice sites, processing sites, or termination signals. Its direct statistical model is based on universal evolutionary signatures of coding sequence only. RNAcode is therefore a true ab initio approach that can be applied to data from all living species. In fact, it does not need any information on the source of its input data, facilitating, e.g., the application to meta-genomics data (Meyer et al. 2009; Shi et al. 2009).

We evaluated a variety of alternative possible metrics and algorithms, but found that pairwise BLOSUM-derived substitution scores together with the relatively simple gap scoring scheme presented was the most efficient solution. We were surprised that this algorithm also outperformed more sophisticated phylogenetic models acting on the whole tree. An exact dynamic programming scheme is used to determine high-scoring coding blocks in the input alignment in a way that is robust against sequence and alignment errors.

Although we do not include any species-specific features such as codon usage or splicing signals, the approach shows remarkable accuracy. Without any training or specifically optimizing the parameters, RNAcode could successfully discriminate between coding and noncoding regions in vertebrates, insects, nematodes, yeasts, bacteria, and even archaea that show a highly biased GC content. We also showed that it can reproduce accurately the current annotation in *D. melanogaster* and identified novel peptides in *E. coli* that have previously evaded annotation in this intensively studied organism. Case studies on individual examples of ncRNAs showed that RNAcode can help to identify mis-annotated ncRNAs and, in combination with RNAz, can identify dual-function RNAs.

The high discrimination performance in combination with accurate *P*-values, visualization, and the readily available open source implementation make RNAcode, we hope, an attractive and easy-to-use solution for many different applications in comparative genomics.

MATERIALS AND METHODS

Implementation details

To estimate the phylogenetic tree for the null model, we use a maximum likelihood implementation provided by PHYML (Guindon and Gascuel 2003). To simulate random alignments along this tree, we use code from Seq-Gen (Rambaut and Grassly 1997).

As a technical detail, we note that our simulation procedure does not simulate gap patterns. Instead, we simulate the alignments without gaps and introduce the original gap patterns afterward. The *P*-values for true coding regions are thus conservative because we use the coding gap pattern also for the background. There are algorithms to simulate the evolution of insertions and deletions. However, it is hard to estimate realistic parameters for these models, and thus we chose this conservative approach that has been successfully used in other applications (Goldman et al. 1998; Gesell and Washietl 2008).

We used the versions of the BLOSUM matrices that are provided with the EMBOSS package (Rice et al. 2000). The current implementation of RNAcode includes the EMBOSS62 and the EMBOSS90 matrices.

For fitting the extreme value parameters to the empirical score distributions, we used an implementation from Sean Eddy's HMMER package (http://hmmer.janelia.org).

Alignment data and benchmarks

Multiple sequence alignments were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu; http://archaea.ucsc.edu). We used the following assemblies, alignments, reference annotations, and (if applicable) selected chromosomes, respectively: *H. sapiens*: hg18, multiz18, UCSC Genes, chr22; *D. melanogaster*: dm3, multiz15, FlyBase Genes (version 5.12), chr2L; *C. elegans*: ce6, multiz6, WormBase Genes (version WS190), chr5; *S. cerevisiae*: sacCer1, multiz7, SGD Genes (version from 01/30/2009), chr4; *E. coli*: eschColi_K12, multizEnterobacteria, GenBank RefSeq; *M. jannaschii*: methJann1, multizMethanococcus, GenBank RefSeq. All data from UCSC were downloaded around the middle of 2009.

To generate the positive test set of known exons, we first extracted alignment blocks corresponding to the annotated exons in the reference annotation. If an exon was covered by several blocks, these were merged. If the resulting alignment was longer than 200 columns, we only used the first 200 columns. As negative control, we selected a comparable number of random blocks that do not overlap annotated coding exons or repeats.

For the tests shown in Figure 4, we selected from the complete set of coding exons a balanced subset of alignments of varying window length (30 nt, 60 nt, 90 nt), varying number of sequences ($N = 2, 4, 8$), and mean pairwise identity (60%–100%). We discarded alignment windows that contained gaps and stop codons in any of the sequences so that they could be directly analyzed using PAML. It is unclear how to handle frameshifts and internal stop codons when calculating a phylogenetic model using PAML, which is not gene-finding software per se. By limiting the analysis to in-frame-aligned sense codons, we ensure a fair comparison to RNAcode that can take advantage of information in gap patterns and stop codons. To calculate the *dN/dS* ratio, we used the codeml program with the default codon model ("model 0"). The periodicity score is calculated as the log-likelihood ratio between two models. As the null model, we used an HKY nucleotide substitution model ("model 4" in PAML's baseml) with equal rates for each site. The alternative model considers three rate classes in a periodic pattern "...ABCABCABC...". The maximum likelihood tree under this model was calculated using the partition model functions of baseml. We used the option "Mgene = 0" keeping all other parameters (κ and π) of the HKY model constant in all three rate classes. The results in Figure 4 are shown for length = 30; sets of length 60 and 90 show qualitatively similar results but saturate earlier to perfect discrimination (data not shown).

E. coli screen

For the screen of novel proteins in the *E. coli* genome, we generated multiple sequence alignments of our own because we noticed that the available alignments at UCSC missed many known coding regions. Moreover, we wanted to improve the evolutionary signal by adding additional species. We used the MultiZ alignment pipeline to align 54 species available from GenBank (Supplemental Table 1).

We then screened the alignments using the default parameters of RNAcode and a *P*-value cutoff of 0.05. This resulted in 20,528 high-scoring coding segments. This number is much higher than the actual number of ORFs mainly because the MultiZ alignments of such a high number of species fragmented the ORFs into relatively small blocks. We combined high-scoring coding segments if they were closer than 15 nt apart and in the same frame, yielding 6542 regions. We discarded all regions that overlapped with an annotated ORF, leaving 229 regions. For these regions, we inferred potential ORFs starting with an ATG and ending in a canonical stop codon. If we did not find an ORF within the RNAcode high-scoring segment, we extended the prediction by 51 nt upstream and downstream and repeated the search. We found 35 loci with a potential ORF (Supplemental Table 2).

Transcriptomics data

The analysis of Cho et al. (2009) represents a comprehensive transcription map for *E. coli*. The corresponding supplemental data were downloaded from http://systemsbiology.ucsd.edu/publication and the Gene Expression Omnibus web page http://www.ncbi.nlm.nih.gov/geo/. The data were converted into BED and WIG formatted files and loaded as custom tracks into the UCSC for visualization and comparison to the novel predicted proteins.

Mass spectrometry experiments

Cell growth

E. coli strain K12 cells were grown in LB medium to stationary phase. One liter of fresh medium was inoculated with 100 mL of a starter culture grown under the same conditions. Cells were collected by centrifugation (10 min, 8000*g*, 4°C).

Protein preparation

Cells were resuspended in urea lysis buffer (40 mL, 8 M urea, 10 mM DTT, 1 M NaCl, 10 mM Tris/HCl at pH 8.0) (Klein et al. 2007) or acidic lysis buffer (40 mL, 0.1% TFA) (Dai et al. 1999) and disrupted using ultrasonication (5 min, 50% duty cycle, Branson Sonifier 250; Emerson, USA). Cell debris was removed by centrifugation (15 min, 10,000*g*, 4°C). High-molecular-weight proteins were depleted by centrifugation through a filter membrane (cutoff molecular weight 50 kDa, Pall Macrosep 50K; Pall Life Science, USA) (Harper et al. 2004). The flow-through was split into aliquots of 1200 μL. Where TFA was used for cell lysis, the samples were titrated to neutral pH by adding NH_4HCO_3 (final concentration 250 mM), and protein disulfide bonds were reduced by adding DTT (10 mM). Cysteine alkylation was conducted by adding 2-iodoacetamide (51.5 mM) and incubation for 45 min at room temperature in the dark.

Gel electrophoresis

Prior to protein separation by 1D gel electrophoresis, the proteins were desalted and concentrated by TCA precipitation (final concentration 20% [w/v]). The protein pellet was redissolved with SDS loading buffer (2% [w/v] SDS, 12% [w/v] glycerol, 120 mM 1,4-dithiothreitol, 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 at pH 8.25). Gel electrophoresis was performed according to Schägger (2006) (with slight modifications). In brief, a 20% T, 6% C separation gel combined with a 4% T, 3% C stacking was used. As protein marker, a prestained low-molecular-weight protein standard (molecular weight range 1.7 kDa–42 kDa, multicolor low-range protein ladder; Fermentas, Germany) was applied. For each cell lysis experiment, eight aliquots were used, of which two were stained with colloidal Coomassie, two were stored as a reserve, and four were used for further analysis. Nine gel slices per lane were excised between 1 and 25 kDa and used for in-gel digestion.

Protein digestion

The gel slices were washed twice with water for 10 min and once with NH_4HCO_3 (10 mM). The low-molecular-weight proteins were digested by adding modified porcine trypsin (100 ng; Sigma-Aldrich) or endoprotease AspN (100 ng; Sigma-Aldrich) in NH_4HCO_3 (10 mM, 30 μL volume). Digestion was performed overnight at 37°C. The supernatant and the solutions from two subsequent gel elution steps (first elution step 40% [v/v] acetonitril, second elution step 80% [v/v]) were collected and united. The samples were dried using vacuum centrifugation.

Mass spectrometry

For validation of the existence of the predicted protein by mass spectrometry, an unbiased bottom-up approach and a targeted analysis were applied. Peptides were reconstituted in 0.1% formic acid. Samples were injected by the autosampler and concentrated on a trapping column (nanoAcquity 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 4 min, the peptides were eluted onto the separation column (nanoAcquity UPLC column, C18, 75 μm × 250 mm, 1.7 μm; Waters). Chromatography was performed with 0.1% formic acid in solvents A (100% water) and B (100% ACN). Peptides were eluted over 90 min with an 8%–40% solvent B gradient using a nano-HPLC system (nanoAcquity; Waters) coupled to an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific). For an unbiased analysis, continuous scanning of eluted peptide ions was carried out between m/z 350 and 2000, automatically switching to CID-MS/MS mode upon detection of ions exceeding an intensity of 2000. For CID-MS/MS measurements, a dynamic precursor exclusion of 3 min was applied. For a targeted analysis, a scan range of m/z = 400–1800 was chosen. CID-MS/MS measurements were triggered if a precursor of a given inclusion list was measured with an error of <20 ppm. The inclusion lists contained all theoretically proteolytic peptides within a molecular weight range of 600 Da to 4000 Da of all predicted proteins considering methionine oxidation, cysteine carbamidomethylation, and up to one (for trypsin) or three (for AspN) proteolytic miscleavages.

Data analysis

Raw spectra were analyzed with ProteomeDiscoverer 1.0 software (Thermo Fisher Scientific, USA). Mascot (Perkins et al. 1999), Sequest (Yates et al. 1995), and X!Tandem (Craig and Beavis 2004) searches were conducted on a protein sequence database, which contains all sequences predicted by RNAcode (RNAcode database) as well as on an extended SWISS-PROT database containing protein sequences predicted by RNAcode and all validated proteins of Hemm et al. (2008). The searches were performed tolerating up to one proteolytic missed cleavage, a mass tolerance of 7 ppm for precursor ions, 0.5 Da for MS/MS product ions allowing for methionine oxidation (optional modification), and cysteine carbamidomethylation (fixed modification). Scaffold (version Scaffold_2_06_00; Proteome Software Inc.) was used to validate MS/MS-based peptide and protein identifications. Peptide identifications were accepted if they could be established at >50% probability as specified by the Peptide Prophet algorithm (Keller et al. 2002). Protein identifications were categorized to be unambiguously identified if they could be established at >99% probability and contained at least two identified peptides that had to achieve a score higher than 80%. Less stringent evidence for proteins was assigned if two peptides were observed with at least one peptide scored higher than 80% and the protein identification probability exceeds 90%. Protein probabilities were assigned by the Protein Prophet algorithm (Nesvizhskii et al. 2003). Additionally, the fragment spectra were checked manually.

Availability

RNAcode is open source software released under the GNU general public license version 3.0. The latest version is available at http://wash.github.com/rnacode.

The package includes a "Getting Started" guide that describes all steps involved in using RNAcode, including obtaining an alignment for analyses that start with a single sequence that is to be assessed for coding potential.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article. Additional data files can be downloaded from http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/RNAcode.

ACKNOWLEDGMENTS

This work was supported in part by grants from the Wellcome Trust (grant 078968), the Deutsche Forschungsgemeinschaft (grant No. STA 850/7-1 under the auspices of SPP-1258, "Small Regulatory RNAs in Prokaryotes," as well as the SFB Transregional Collaborative Research Centre 67: "Functional Biomaterials for Controlling Healing Processes in Bone and Skin—From Material Science to Clinical Application"), and the Austrian GEN-AU project "Noncoding RNA." S.W. was supported by a GEN-AU mobility fellowship sponsored by the Bundesministerium für Wissenschaft und Forschung.

APPENDIX: DYNAMIC PROGRAMMING ALGORITHM

In the following, we formally describe the algorithms implemented in RNAcode. The core algorithm is a dynamic programming algorithm to find the optimal score for a pairwise alignment from all possible interpretations of the aligned sites as in-frame codons, out-of-frame codons, or sequence errors (cf. Fig. 2). The scores from pairwise alignments are then combined to find optimal scoring segments in a multiple alignment.

We start from a fixed multiple sequence alignment $\mathbb{A}$ and assume that the first row is the *reference sequence*. The projected pairwise alignment of the reference sequence with sequence k is denoted by $\mathbb{A}^k$. Now consider a position i in the reference sequence. It corresponds to a uniquely determined alignment column $\alpha(i)$, which, in turn, determines i_k, the last position of sequence k that occurs in or before alignment column $\alpha(i)$.

Suppose i is a third codon position. Then the alignment block $\mathbb{A}\,[\alpha(i-3)+1, \alpha(i)]$ corresponds to the (potential) codon ending in i. We define a score:

$$\sigma_i^k = \text{score}\left(\mathbb{A}^k[\alpha(i-3)+1, \alpha(i)]\right). \tag{3}$$

In the ungapped case, σ_i^k is the normalized BLOSUM score that was introduced in the main text. Let g_i^k denote the number of gaps in sequence k in this block. We observe that sequences 1 (reference) and k stay in-frame if and only if $g_i^k - g_i^1 \equiv 0$, mod 3. Otherwise, the two sequences change their phase within this interval. The local shift in frame between sequence k and the reference sequence is therefore:

$$z_i^k = \begin{cases} 0 & \text{if } g_i^k - g_i^1 \equiv 0 \mod 3 \\ +1 & \text{if } g_i^k - g_i^1 \equiv 1 \mod 3 \\ -1 & \text{if } g_i^k - g_i^1 \equiv 2 \mod 3 \end{cases} \tag{4}$$

As discussed in the main text, alignment errors or sequence errors may destroy coherence between aligned codons and give $z_i^k \neq 0$. Therefore, we introduce the penalties (negative scores) Ω for switching from in-frame to out-of-frame or back, as well as ω for every out-of-frame codon in between, and Δ for silently changing the phase and assuming subsequent codons are still in-frame (sequencing error). All penalties are negative; in particular, $\frac{1}{2}\Delta<\Omega<\omega<0$. Furthermore, we set $\sigma_i^k = -\infty$ if $z_i^k \neq 0$ to mark the fact that we lose coherence of the frame and force the algorithm to select a frameshift or sequence error penalty and not a substitution score that would be meaningless for out-of-frame triples.

Having defined all possible states and the associated scores, we now describe a dynamic programming algorithm to calculate the optimal score for a pairwise alignment. Let $S_{b,i}^{0,k}$ be the optimal score of the pairwise alignment $\mathbb{A}^k[\alpha(b), \alpha(i)]$ subject to the condition that i is a third codon position and sequence k ends in-frame, i.e., also with a third codon position. Analogously, we define $S_{b,i}^{+,k}$ and $S_{b,i}^{-,k}$ for those alignments where sequence k ends

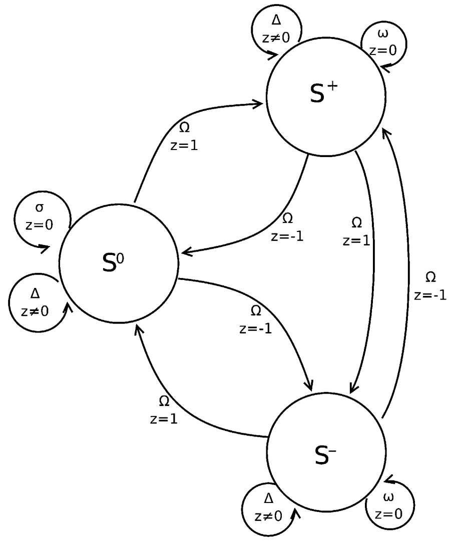


FIGURE 7. Finite state automaton representing the scoring of pairwise alignments. The three states correspond to the relative phases of the sequences. Insertions and deletions with $z \neq 0$ lead to local changes in-phase that are penalized by Ω. Extension in each of the two out-of-frame states S^+ and S^- is penalized by ω. In/dels interpreted as sequencing errors or true frameshifts are penalized by Δ.

in the first and second codon position, respectively. Clearly, we initialize $S^{\chi,k}_{b,b} = 0$ for $\chi \in \{0,+,-\}$.

The entries in these matrices satisfy the following recursions:

$$S^{0,k}_{b,i} = \begin{cases} S^{0,k}_{b,i-3} + \sigma^k_i & \text{if } z^k_i = 0 \\ \max\begin{cases} S^{0,k}_{b,i-3} + \Delta, \\ S^{-k}_{b,i-3} + \Omega \end{cases} & \text{if } z^k_i = +1 \\ \max\begin{cases} S^{0,k}_{b,i-3} + \Delta, \\ S^{-k}_{b,i-3} + \Omega \end{cases} & \text{if } z^k_i = -1 \end{cases} \tag{5}$$

The expressions for the two out-of-frame scores are analogous. We show only one of them explicitly:

$$S^{+,k}_{b,i} = \begin{cases} S^{+,k}_{b,i-3} + \omega & \text{if } z^k_i = 0 \\ \max\begin{cases} S^{0,k}_{b,i-3} + \Omega \\ S^{+k}_{b,i-3} + \Delta \end{cases} & \text{if } z^k_i = +1 \\ \max\begin{cases} S^{+,k}_{b,i-3} + \Delta \\ S^{-k}_{b,i-3} + \Omega \end{cases} & \text{if } z^k_i = -1 \end{cases} \tag{6}$$

A state diagram corresponding to the above algorithm is shown in Figure 7. As presented here, the algorithm assumes that any sequence errors (penalized by Δ) occur in sequence k, not in the reference.

Now we determine the optimal score S_{bi} of the multiple alignment $\mathbb{A}[\alpha(b), \alpha(i)]$, subject to the condition that b is a first codon position and i is a third codon position.

$$S_{bi} = \max\begin{cases} \sum\limits_{k>1} \max\limits_{\chi\in\{0,+,-\}} S^{\chi,k}_{b,i} \\ S_{b,i-1} + \Delta \\ S_{b,i-2} + \Delta \end{cases} \tag{7}$$

The second and third terms here correspond to frameshifts in the reference sequence.

It is easy now to determine the best scoring segment(s) of $\mathbb{A}$ from the maximal entries in the matrix (S_{bi}). If we were to score only pairwise alignments, it would be possible to use a local alignment-like algorithm that does not keep track of the beginning of the segment, b. In the multiple alignment, however, the individual pairwise alignments are constrained by the requirement that a coding segment starts in the same column for all sequences, forcing us to keep track of b explicitly. The algorithm scales as $\mathcal{O}(N \cdot n^2)$ in time and space, where n is the length of the reference sequence and N the number of rows in the alignment.

Received November 10, 2010; accepted January 12, 2011.

REFERENCES

Aebersold R, Mann M. 2003. Mass spectrometry-based proteomics. *Nature* **422:** 198–207.

Badger JH, Olsen GJ. 1999. CRITICA: Coding region identification tool invoking comparative analysis. *Mol Biol Evol* **16:** 512–524.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14:** 708–715.

Bofkin L, Goldman N. 2007. Variation in evolutionary processes at different codon positions. *Mol Biol Evol* **24:** 513–521.

Boisset S, Geissmann T, Huntzinger E, Fechter P, Bendridi N, Possedko M, Chevalier C, Helfer AC, Benito Y, Jacquier A, et al. 2007. *Staphylococcus aureus* RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. *Genes Dev* **21:** 1353–1366.

Brent MR. 2008. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat Rev Genet* **9:** 62–73.

Burge CB, Karlin S. 1998. Finding the genes in genomic DNA. *Curr Opin Struct Biol* **8:** 346–354.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.

Cho BK, 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 Biotechnol* **27:** 1043–1049.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci* **104:** 19428–19433.

Coghlan A, Fiedler TJ, McKay SJ, Flicek P, Harris TW, Blasiar D, nGASP Consortium, Stein LD. 2008. nGASP—the nematode genome annotation assessment project. *BMC Bioinformatics* **9:** 549. doi: 10.1186/1471-2105-9-549.

Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20:** 1466–1467.

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.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci* **103:** 5320–5325.

Dinger ME, Pang KC, Mercer TR, Mattick JS. 2008. Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. *PLoS Comput Biol* **4:** e1000176. doi: 10.1371/journal.pcbi.1000176.

Drysdale R; FlyBase Consortium. 2008. FlyBase: A database for the *Drosophila* research community. *Methods Mol Biol* **420:** 45–59.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Flicek P. 2007. Gene prediction: compare and CONTRAST. *Genome Biol* **8:** 233. doi: 10.1186/gb-2007-8-12-233.

Frith MC, Bailey TL, Kasukawa T, Mignone F, Kummerfeld SK, Madera M, Sunkara S, Furuno M, Bult CJ, Quackenbush J, et al. 2006. Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA Biol* **3:** 40–48.

Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. 2007. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* **5:** e106. doi: 10.1371/journal.pbio.0050106.

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, et al. 2009. Rfam: updates to the RNA families database. *Nucleic Acids Res* **37:** D136–D140.

Gesell T, Washietl S. 2008. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics* **9:** 248. doi: 10.1186/1471-2105-9-248.

Gimpel M, Heidrich N, Mäder U, Krügel H, Brantl S. 2010. A dual-function sRNA from *B. subtilis*: SR1 acts as a peptide encoding mRNA on the *gapA* operon. *Mol Microbiol* **76:** 990–1009.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149:** 445–458.

Gross SS, Brent MR. 2006. Using multiple alignments to improve gene prediction. *J Comput Biol* **13:** 379–393.

Gross SS, Do CB, Sirota M, Batzoglou S. 2007. CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol* **8:** R269. doi: 10.1186/gb-2007-8-12-r269.

Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. 2010. RNAz 2.0: Improved noncoding RNA detection. *Pac Symp Biocomput* **15:** 69–79.

Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, et al. 2006. EGASP: the human ENCODE genome annotation assessment project. *Genome Biol* **7:** S2.1–S31.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696–704.

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.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22:** 160–174.

Heidrich N, Chinali A, Gerth U, Brantl S. 2006. The small untranslated RNA SR1 from the *Bacillus subtilis* genome is involved in the regulation of arginine catabolism. *Mol Microbiol* **62:** 520–536.

Heidrich N, Moll I, Brantl S. 2007. In vitro analysis of the interaction between the small RNA SR1 and its primary target *ahrC* mRNA. *Nucleic Acids Res* **35:** 4331–4346.

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.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci* **89:** 10915–10919.

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.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Kellis M, Patterson N, Birren B, Berger B, Lander ES. 2004. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J Comput Biol* **11:** 319–355.

Klein C, Aivaliotis M, Olsen JV, Falb M, Besir H, Scheffer B, Bisle B, Tebbe A, Konstantinidis K, Siedler F, et al. 2007. The low molecular weight proteome of *Halobacterium salinarum*. *J Proteome Res* **6:** 1510–1518.

Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, Kageyama Y. 2007. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat Cell Biol* **9:** 660–665.

Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S, Payre F, Kageyama Y. 2010. Small peptides switch the transcriptional activity of shavenbaby during *Drosophila* embryogenesis. *Science* **329:** 336–339.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, et al. 2009. The UCSC Genome Browser database: Update 2009. *Nucleic Acids Res* **37:** D755–D761.

Licht A, Preis S, Brantl S. 2005. Implication of CcpN in the regulation of a novel untranslated RNA (SR1) in *Bacillus subtilis*. *Mol Microbiol* **58:** 189–206.

Lin MF, Deoras AN, Rasmussen MD, Kellis M. 2008. Performance and scalability of discriminative metrics for comparative gene identification in 12 *Drosophila* genomes. *PLoS Comput Biol* **4:** e1000067. doi: 10.1371/journal.pcbi.1000067.

Meyer MM, Ames TD, Smith DP, Weinberg Z, Schwalbach MS, Giovannoni SJ, Breaker RR. 2009. Identification of candidate structured RNAs in the marine organism '*Candidatus* Pelagibacter ubique'. *BMC Genomics* **10:** 268. doi: 10.1186/1471-2164-10-268.

Mignone F, Grillo G, Liuni S, Pesole G. 2003. Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. *Nucleic Acids Res* **31:** 4639–4645.

Mourier T, Carret C, Kyes S, Christodoulou Z, Gardner PP, Jeffares DC, Pinches R, Barrell B, Berriman M, Griffiths-Jones S, et al. 2008. Genome-wide discovery and verification of novel structured RNAs in *Plasmodium falciparum*. *Genome Res* **18:** 281–292.

Müller SA, Kohajda T, Findeiß S, Stadler PF, Washietl S, Kellis M, von Bergen M, Kalkhof S. 2010. Optimization of parameters for coverage of low molecular weight proteins. *Anal Bioanal Chem* **398:** 2867–2881.

Nekrutenko A, Chung WY, Li WH. 2003. ETOPE: Evolutionary test of predicted exons. *Nucleic Acids Res* **31:** 3564–3567.

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

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20:** 3551–3567.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* **13:** 235–238.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16:** 276–277.

Rivas E, Eddy SR. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2:** 8. doi: 10.1186/1471-2105-2-8.

Rosenberg MI, Desplan C. 2010. Molecular biology. Hiding in plain sight. *Science* **329:** 284–285.

Schägger H. 2006. Tricine-SDS-PAGE. *Nat Protoc* **1:** 16–22.

Schneider KL, Pollard KS, Baertsch R, Pohl A, Lowe TM. 2006. The UCSC archaeal genome browser. *Nucleic Acids Res* **34:** D407–D410.

Shi Y, Tyson GW, DeLong EF. 2009. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459:** 266–269.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450:** 219–232.

Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E, Rosenow C. 2002. Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res* **30:** 3732–3738.

UniProt Consortium. 2010. The universal protein resource (UniProt) in 2010. *Nucleic Acids Res* **38:** D142–D148.

Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci* **102:** 2454–2459.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, et al. 2007. Structured RNAs in the encode selected regions of the human genome. *Genome Res* **17:** 852–864.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17:** 32–43.

Yates JR III, Eng JK, McCormack AL, Schieltz D. 1995. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* **67:** 1426–1436.

3.4 Identification of new protein coding sequences and signal peptidase cleavage sites of *Helicobacter pylori* strain 26695 by proteogenomics

Stephan A. Müller, Sven Findeiß, Sandy R. Pernitzsch, Dirk K. Wissenbach, Peter F. Stadlerb, Ivo L. Hofacker, Martin von Bergen, Stefan Kalkhof

J Proteomics. 2013 Jun 28;86:27-42. Epub 2013 May 9.
DOI:10.1016/j.jprot.2013.04.036

Abstract

Correct annotation of protein coding genes is the basis of conventional data analysis in proteomic studies. Nevertheless, most protein sequence databases almost exclusively rely on gene finding software and inevitably also miss protein annotations or possess errors. Proteogenomics tries to overcome these issues by matching MS data directly against a genome sequence database. Here we report an in-depth proteogenomics study of *Helicobacter pylori* strain 26695. MS data was searched against a combined database of the NCBI annotations and a six-frame translation of the genome. Database searches with Mascot and X! Tandem revealed 1115 proteins identified by at least two peptides with a peptide false discovery rate below 1%. This represents 71% of the predicted proteome. So far this is the most extensive proteome study of Helicobacter pylori. Our proteogenomic approach unambiguously identified four previously missed annotations and furthermore allowed us to correct sequences of six annotated proteins. Since secreted proteins are often involved in pathogenic processes we further investigated signal peptidase cleavage sites. By applying a database search that accommodates the identification of semi-specific cleaved peptides, 63 previously unknown signal peptides were detected. The motif LXA showed to be the predominant recognition sequence for signal peptidases.

Keywords

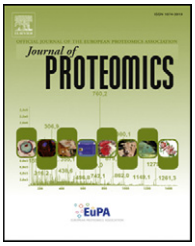
Proteogenomics; Proteomics; Coding sequence; Helicobacter pylori; Signal peptide

JOURNAL OF PROTEOMICS 86 (2013) 27–42

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot



Identification of new protein coding sequences and signal peptidase cleavage sites of *Helicobacter pylori* strain 26695 by proteogenomics

Stephan A. Müller[a], *Sven Findeiß*[b,c], *Sandy R. Pernitzsch*[d], *Dirk K. Wissenbach*[e], *Peter F. Stadler*[b,f,g,h,i], *Ivo L. Hofacker*[b,c], *Martin von Bergen*[a,e,j], *Stefan Kalkhof*[a,*]

[a] *Department of Proteomics, UFZ, Helmholtz-Centre for Environmental Research Leipzig, 04318 Leipzig, Germany*
[b] *Institute for Theoretical Chemistry, University of Vienna, A-1090 Wien, Austria*
[c] *Bioinformatics and Computational Biology research group, University of Vienna, A-1090 Wien, Austria*
[d] *Research Center for Infectious Diseases (ZINF), University of Würzburg, Würzburg, Germany*
[e] *Department of Metabolomics, UFZ, Helmholtz-Centre for Environmental Research Leipzig, 04318 Leipzig, Germany*
[f] *Bioinformatics Group, Department of Computer Science, University Leipzig, 04107 Leipzig, Germany*
[g] *RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, 04103 Leipzig, Germany*
[h] *Santa Fe Institute, Santa Fe, 87501 NM, USA*
[i] *Max-Planck-Institute for Mathematics in Sciences, 04103 Leipzig, Germany*
[j] *Aalborg University, Department of Biotechnology, Chemistry and Environmental Engineering, Sohngaardsholmsvej 49, DK-9000 Aalborg, Denmark*

ARTICLE INFO

Article history:
Received 30 October 2012
Accepted 26 April 2013
Available online 9 May 2013

Keywords:
Proteogenomics
Proteomics
Coding sequence
Helicobacter pylori
Signal peptide

ABSTRACT

Correct annotation of protein coding genes is the basis of conventional data analysis in proteomic studies. Nevertheless, most protein sequence databases almost exclusively rely on gene finding software and inevitably also miss protein annotations or possess errors. Proteogenomics tries to overcome these issues by matching MS data directly against a genome sequence database. Here we report an in-depth proteogenomics study of *Helicobacter pylori* strain 26695. MS data was searched against a combined database of the NCBI annotations and a six-frame translation of the genome. Database searches with Mascot and X! Tandem revealed 1115 proteins identified by at least two peptides with a peptide false discovery rate below 1%. This represents 71% of the predicted proteome. So far this is the most extensive proteome study of *Helicobacter pylori*. Our proteogenomic approach unambiguously identified four previously missed annotations and furthermore allowed us to correct sequences of six annotated proteins. Since secreted proteins are often involved in pathogenic processes we further investigated signal peptidase cleavage sites. By applying a database search that accommodates the identification of semi-specific cleaved peptides, 63 previously unknown signal peptides were detected. The motif LXA showed to be the predominant recognition sequence for signal peptidases.

Biological significance

The results of MS-based proteomic studies highly rely on correct annotation of protein coding genes which is the basis of conventional data analysis. However, the annotation of protein coding sequences

---

[*] *Corresponding author at:* Department of Proteomics, UFZ, Helmholtz-Centre for Environmental Research, Permoserstr. 15, 04318 Leipzig, Germany. Tel.: +49 341 2351354; fax: +49 341 2351786.
E-mail address: stefan.kalkhof@ufz.de (S. Kalkhof).

1874-3919/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jprot.2013.04.036

in genomic data is usually based on gene finding software. These tools are limited in their prediction accuracy such as the problematic determination of exact gene boundaries. Thus, protein databases own partly erroneous or incomplete sequences. Additionally, some protein sequences might also be missing in the databases.

Proteogenomics, a combination of proteomic and genomic data analyses, is well suited to detect previously not annotated proteins and to correct erroneous sequences. For this purpose, the existing database of the investigated species is typically supplemented with a six-frame translation of the genome. Here, we studied the proteome of the major human pathogen *Helicobacter pylori* that is responsible for many gastric diseases such as duodenal ulcers and gastric cancer. Our in-depth proteomic study highly reliably identified 1115 proteins (FDR < 0.01%) by at least two peptides (FDR < 1%) which represent 71% of the predicted proteome deposited at NCBI.

The proteogenomic data analysis of our data set resulted in the unambiguous identification of four previously missed annotations, the correction of six annotated proteins as well as the detection of 63 previously unknown signal peptides. We have annotated proteins of particular biological interest like the ferrous iron transport protein A, the coiled-coil-rich protein HP0058 and the lipopolysaccharide biosynthesis protein HP0619. For instance, the protein HP0619 could be a drug target for the inhibition of the LPS synthesis pathway.

Furthermore it has been proven that the motif "LXA" is the predominant recognition sequence for the signal peptidase I of *H. pylori*. Signal peptidases are essential enzymes for the viability of bacterial cells and are involved in pathogenesis. Therefore signal peptidases could be novel targets for antibiotics. The inclusion of the corrected and new annotated proteins as well as the information of signal peptide cleavage sites will help in the study of biological pathways involved in pathogenesis or drug response of *H. pylori*.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The first DNA-based genome was sequenced by Frederick Sanger in 1977 [1]. At the start of this development, genome sequencing was restricted to rather small genomes. Further developments such as computer-based alignment of shotgun fragments [2] and the polymerase chain reaction [3] rendered genome sequencing into a well automated and cost-effective high-throughput method. Hence, hundreds of additional genomes will be sequenced and have to be analyzed within the next years.

Annotation of protein coding sequences in genomic data is usually based on gene finding software such as IMG [4], RAST [5], Glimmer [6], or GeneMark [7]. These tools are limited in their prediction accuracy. For example, it is typically problematic to determine exact gene boundaries. This limitation can be partially overcome by the use of additional information such as regulatory motifs like ribosome binding sites, which are normally located in vicinity of open reading frames. However, many exceptions to the classical translation initiation model are known [8]. The previously underestimated number of leaderless mRNAs in various species is only one example [9–11]. Beyond annotation problems, there is also the problem of missing functional information. Although substantial effort is spent on functional assignment, even for the model organism *Helicobacter pylori* 26695 about 33% of protein coding genes still belong to the class of hypothetical proteins [12]. Furthermore, most tools use a minimum open reading frame length cutoff, typically of 300 nucleotides, in order to keep the false discovery rate low [13]. As a consequence, short protein coding genes with less than 100 amino acids that are expressed and functional are lacking in the annotation [14]. In eukaryotes, additionally the prediction of alternative splice variants for commonly used software packages is challenging. Furthermore, the results of standard gene annotation algorithms differ from each other [15]. Dependent on the method, automatic predictions that differ by the limitations of the applied approach, protein sequences are deposited in databases such as NCBI or UniProt. These problems create the need for improving the existing protein coding gene annotations [16,17]. A complementary approach to commonly used protein coding gene annotation methods is applied by the software RNAcode [17]. It neither relies on splicing, on training data nor on species specific gene features such as open reading frame detection or sequence motifs necessary for ribosome binding. RNAcode simply analyzes a multiple alignment of nucleotide sequences by means of a statistical framework that compares nucleotide variation and the implied amino acid variation in all six possible reading frames to detect high scoring segments in which synonymous substitutions or insertions or deletions that preserve the reading frame, i.e., typical for conserved protein coding regions, are overrepresented.

Comparative genomic studies of different *H. pylori* strains already investigated differences of current coding sequence annotations [12,18,19]. Medigue et al. [18] identified putative DNA sequencing errors which result in missing or erroneous protein annotations. On the other hand, Boneca et al. [12] focused on the functional annotation and reported length differences of existing coding sequences of the strains 26695 and J99. The sources of size variation were classified due to nucleotide insertions/deletions, different start or stop codons, intragenic frame-shifts, slipped-strand mispairing mechanisms originated from homopolymeric repeats as well as pseudogenes. Moreover, Sharma et al. [20] provided a list of re-annotated protein coding genes based on transcriptome data. However, only minor parts of these results were used to improve protein databases.

High quality protein databases are the fundament of proteomic studies. Missing annotations or erroneous annotated protein sequences lead to decreased protein identification rates in classical shotgun proteomic studies that exclusively rely on database searches of MS data. The combination of proteomics and genomics, called proteogenomics, has been proven to be well suited for confirming predicted genes, correct starting and stop sites of genes and in identifying new genes and splicing variants. [21–30].

In a typical proteogenomic approach, an existing protein sequence database is complemented by a six-frame translation of the whole genome to generate a comprehensive database. Transcriptome data can also be used to improve and extend the database [20]. In particular, database refinement for eukaryotes benefits from transcriptome data due to the inclusion of additional splice variants [28]. The identification of peptides supporting unique sequences within the six-frame translation is of great interest. Peptides located at the N- or C-terminal of an annotation can be used to correct the translation start and stop sites, while novel genes can be found as peptide sequences mapping to intergenic regions [21,24]. Peptides within annotated intronic regions can be used to identify new exons in eukaryotes. Novel splice variants can be identified either by exon–exon spanning peptides or by fragments that map to intergenic regions and which are subsequently connected to an existing gene [28,31].

The ongoing development of MS has made it possible to acquire spectra with high resolution, high mass accuracy and fast scanning speed [32]. The introduction of nano-UHPLC [33,34], multidimensional LC [35] as well as the application of ultra-long gradients [36] or long monolithic columns [37] for peptide separation enable LC–MS/MS analyses to dig deeper into the proteome. Cell compartment [38,39] or protein fractionation [40–42] prior to proteolytic digestion, as well as the application of multiple proteases [42,43] are widely used strategies to further improve the proteome coverage.

As a consequence of this development whole proteomes can be nearly completely covered in proteomic studies [44,45]. Recently, Nagaraj et al. [46] identified 10,255 proteins encoded by 9207 genes using a human cancer cell line. For this approach, three different proteases and fractionation on the protein and peptide level prior to LC–MS/MS analysis were applied. Comparison with transcriptome data (16,846 transcripts, 11,936 genes) derived from RNA-Seq [47,48] proved the high coverage. This project demonstrates that nowadays even coverage of complex proteomes such as the one expressed in human of up to 77% is achievable by shotgun proteomics using extensive fractionation and subsequent state of the art mass spectrometric analysis.

Here, we present the results of an in-depth proteome study of *H. pylori* strain 26695. We combined a GeLC–MS procedure and an offline 2D-LC–MS approach using size exclusion chromatography (SEC) of proteins focused on low molecular weight (MW) proteins of less than 25 kDa in the first dimension. Overall, 1115 proteins or 71% of the predicted proteome deposited at NCBI were identified based on at least two peptides with a false discovery rate (FDR) below 1%, respectively. Furthermore, proteogenomic analysis revealed ten proteins with either none (four) or incomplete (six) annotation. These protein coding sequence corrections were partially confirmed by comparison of MS/MS spectra with ^{13}C- and ^{15}N-labeled synthetic peptides. Additionally, 63 previously unknown signal peptide sequences could be annotated by MS/MS spectra with a search strategy allowing for semi-specific cleaved peptides and revealed the predominant recognition motif LXA for signal peptidases. The results of this study are deposited at http://www.bioinf.uni-leipzig.de/publications/supplements/12-023/ and are linked to the UCSC microbial genome browser [49].

2. Materials and methods

2.1. Cell culture

H. pylori strain 26695 from cryostock was grown on GC-Agar plates (Oxoid) supplemented with 10% heat-inactivated donor horse serum (Biochrom AG), 1% vitamin mix, 10 μg ml^{-1} vancomycin, 5 μg ml^{-1} trimethoprim and 1 μg ml^{-1} nystatin. After incubation for 1–2 days in anaerobic jars under microaerophilic conditions (CampyGen bags from Oxoid (CN0025A) providing atmosphere of 10% CO_2 and 6% O_2), bacteria were restreaked to fresh plates. For liquid culture, bacteria were harvested from plate and resuspended to a final $OD_{600\ nm}$ of 0.02 per ml in 50 ml Brain Heart Infusion medium (BHI) supplemented with 10% FCS and the same antibiotics as described above. Bacteria were grown under agitation at 140 rpm in jars under microaerophilic conditions (same conditions like above) to the transition from exponential to stationary phase. For the proteomic analysis, *H. pylori* cells were collected by centrifugation (4000× *g*, 10 min, 4 °C) and washed twice with ice-cold PBS prior to protein extraction and pre-separation. Two biological replicates were used for the proteomic analysis.

2.2. Protein extraction and preseparation

Cells were lysed in a urea buffer as previously described [42]. Cell debris and undissolved material were removed by centrifugation (10 min, 16,000 ×*g*, 18 °C). Protein concentrations were measured with the Bradford QuickStart assay (Biorad, Hercules, CA, USA). An amount of 60 μg protein per biological replicate was precipitated with acetone. The resulting protein pellets were redissolved in 20 μl Lämmli-buffer and subjected to 1-D-SDS PAGE (12% separation gel, 4% stacking gel). The gel was fixed in fixing-solution for 1 h (50% methanol, 10% acetic acid, 100 mM ammonium acetate) and stained with Coomassie (0.025% Coomassie G250 in 10% acetic acid).

SEC was used to enrich and preseparate the low MW proteome of *H. pylori*. Cell lysates were filtered with 0.2 μm syringe filter (VWR, Germany). SEC was performed on a HPLC system (Prominence, Shimadzu, Japan) with a Biosep S-2000 SEC column (ID 4.6 mm, length 30 cm, Phenomenex, USA). Separation was carried out isocratic at 20 °C and at a flow of 0.35 ml/min of mobile phase (50 mM phosphate buffer pH = 7, 25% v/v acetonitrile (ACN), 100 mM NaCl, 2 M urea, 5 mM DTE). 100 μl cell lysate (protein conc. about 1 mg/ml) was injected per run. Eight fractions, each one minute sampling time, were collected automatically after a dead time of 9 min (Waters fraction collector III, Waters, Milford, MA, USA). 16 runs were pooled to achieve a valuable amount of protein for subsequent analysis.

The last four fractions, representing proteins below 25 kDa, were used for further analysis. ACN was removed by vacuum centrifugation (Concentrator plus, Eppendorf, Hamburg, Germany) and sample volume was reduced to 50%. Samples were concentrated and cleaned by C-18 spin columns (Pepclean C-18 Spin Columns, Pierce, USA) according to the manufacture's instruction with slight modifications. In brief, elution of proteins was carried out in four stages with increasing ACN content (30%, 50%, 70%, 90% ACN supplied with 0.1% formic acid). The protocol was repeated once again with the flow through of the first binding step. The combined eluates of each SEC fraction were dried by vacuum centrifugation for further usage.

2.3. *Proteolytic digestion*

The protein lanes of the 1-D-SDS PAGE were cut into 20 slices of equal size. In-gel digestion with trypsin was performed as previously described [50]. Peptide eluates were dried in a vacuum centrifuge and redissolved in 0.1% formic acid.

Concentrated and dried SEC fractions were redissolved in 6 M urea containing 100 mM NH_4HCO_3. Samples were titrated with 1 M NH_4HCO_3 to a pH of 8. Cysteines were alkylated using DTT (2 μmol, 37 °C, 30 min) and IAA (8 μmol, room temperature, in the dark). Excess of IAA was removed by the addition of DTT (4 μmol). 10 μg of each protein fraction was separately digested with trypsin, LysC and AspN (sequencing grade, Roche, Mannheim, DE) with an enzyme to protein weight ratio of approximately 1:20. Protein digestion was stopped by the addition of formic acid (final concentration 1% (v/v)). Proteolytic peptides were dried by vacuum centrifugation and resuspended in 0.1% formic acid.

2.4. *LC–MS/MS analysis*

LC–MS/MS analysis was carried out on a nano-HPLC system (nanoAquity, Waters, Milford, MA, USA) coupled online to a LTQ Orbitrap XL ETD mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) via a chip-based nano-ESI source (TriVersa NanoMate, Advion, Ithaca, NY, USA). Peptide solutions were injected on trapping column (nanoAquity UPLC column, C18, 180 μm × 20 mm, 5 μm, Waters) and washed for 8 min with 2% (v/v) ACN containing 0.1% (v/v) formic acid with a flow of 15 μl/min. After washing, peptides were separated on a nano-UPLC column (nanoAcquity UPLC column, C18, 75 μm × 150 mm, 1.7 μm, Waters). Peptides were eluted by a gradient from 2 to 40% (v/v) ACN containing 0.1% (v/v) formic acid (2 min, 2%; 7 min, 6%; 105 min, 20%; 148 min, 30%; 191 min, 40%) with a flow of 300 nl/min.

Peptides were ionized by the nano-ESI source with a voltage of 1.7 kV in positive ion mode. MS analysis switched automatically between full scan MS mode (*m/z* 400–1400, R = 60,000, orbitrap analyzer) and acquisition of fragment ion spectra (linear ion trap analyzer). Peptide ions with intensities above 3000 counts were chosen for collision induced dissociation within the linear ion trap (isolation width 4 amu, normalized collision energy 35%, activation time 30 ms, activation Q 0.25). Formerly selected precursor ions were dynamically excluded for 5 min.

Additionally, retention time dependent exclusion lists were used for the measurement of SEC samples. Separate exclusion lists were created for the two biological samples as well as for the different proteases. Therefore a database search against a NCBI database containing all proteins of *H. pylori* strain 26695 (NC_000915; 03.03.2011) with Proteome Discoverer (version 1.0; Thermo Fisher Scientific, San Jose, CA, USA) using the Mascot (version 2.3.01; Matrix Science, London, UK) search algorithm was performed. A precursor ion tolerance of 5 ppm and a fragment ion tolerance of 0.5 Da were defined. Carbamidomethylation of cysteines was specified as fixed modification whereas oxidation of methionines was adjusted as variable modification. Peptides exceeding an ion score of 20 were excluded by *m/z* values with a deviation of ±10 ppm and a retention time window of ±5 min. The measurements were started with the fractions of the highest MW.

Additionally, we integrated MS data published by Jungblut et al. [51] to further complement and validate our results. This dataset was obtained by MALDI-MS measurements of 2-DE separated proteins (710 spots) and by high-throughput using the GeLC–MS approach for different samples.

2.5. *Database construction*

The *H. pylori* genome and all annotated protein sequences have been downloaded from NCBI (NC_000915; 03.03.2011). In order to generate a comprehensive database for the subsequent analysis the annotated protein sequences were concatenated with a six-frame translation of the complete genome. For each frame nucleotide triplets are translated into the corresponding amino acid. If a triplet contains non-canonical nucleotides, i.e. other than A, C, G and T, it is translated into X. The one-letter code X is replaced by all 20 canonical amino acids in database searches to test all possibilities. Peptides containing more than one X are discarded for database searches. The amino acid chain is terminated if a triplet encodes a canonical stop codon. All chains shorter than six amino acids are rejected.

2.6. *Initial database search*

The spectrum files from our experiments were recalibrated using the "first search" option of Maxquant 1.1 (version 1.1.1.25, Max Planck Institute of Biochemistry, Munich, Germany) with the NCBI database of *H. pylori* strain 26695 (NC_000915; 03.03.2011). Resulting apl files were converted into mgf file format. Database searches were performed with the Mascot (version 2.3.01, Matrixscience, London, UK) and the X! Tandem (The GPM, thegpm.org; version CYCLONE (2010.12.01.1)) search engines against a reverse concatenated NCBI database of *H. pylori* strain 26695 (NC_000915; 03.03.2011) complemented with a six-frame translation of the genome (131,190 target and 131,190 decoy entries).

Mascot and X! Tandem were searched with a precursor tolerance of 5 ppm and a fragment ion mass tolerance of 0.5 Da. Carbamidomethylation of cysteines was specified as a fixed modification. Oxidation of methionine was defined as a variable modification. For AspN digestions, pyroglutamate formation of glutamic acid and glutamine at the peptide N-terminus was specified as additional variable modifications. Two missed cleavages were allowed for trypsin and LysC, whereas three were set for AspN.

Scaffold (version 3.4.9, Proteome Software Inc., Portland, OR, USA) was used to validate MS/MS based peptide and protein identifications. Protein and peptide FDRs were calculated according to Käll et al. [52].

Peptide identifications required at least Mascot ion scores greater than both the associated identity scores and 25 or X! Tandem - Log(Expect Scores) scores greater than 1.95. Protein identifications were accepted if they contained at least two unique peptides in a single experiment. Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony.

Database search of the integrated dataset was done according to the recommendations in the supplementary material of Jungblut et al. [51] except for missed cleavage limits were set to two. Peptide and protein identifications were filtered according to the same thresholds applied to our data.

2.7. *Identification and validation of erroneous and new protein annotations*

Peptides which could not be matched to the NCBI database but to the six-frame translation were used for further analysis. The peptide localization was mapped and visualized using the UCSC microbial genome browser [49,53]. Additionally, a BLAST search with standard parameter settings against the NCBI reference sequence database of *H. pylori* (taxID 210) was performed to identify similar proteins in other strains. The genome location together with the information of the BLAST search was used to classify the peptides into N-terminal elongations, truncated sequences due to DNA sequencing errors of existing protein annotations and regions without protein annotations. Thereby, possible DNA sequencing errors as well as wrong annotated translation start sites are detectable.

DNA sequencing errors in genes inevitably lead to erroneous protein annotations. These errors are also part of the six-frame translation. Hence, DNA sequencing errors can only be corrected by proteogenomics if the true sequences are included into the protein database. Peptides which were matched to previously untranslated regions at the 3’ or 5’ end were searched by BLAST against *H. pylori* species (taxID 210) to get a list of protein sequences which include these sequences. The derived protein sequences offer new targets for a second database search to validate the supposed DNA sequencing errors and to correct the resulting erroneous protein sequences.

Furthermore, detected translation start sites were corrected and also added to the database. This database supplementation opens the possibility to identify peptides matching to the new annotated protein N-termini. The plain search against the six-frame translation does not offer the possibility to identify new protein N-termini since peptides have to be specifically cleaved in conventional databases searches. With this supplemented database, a second search with identical settings was performed to gain additional peptide identifications to proof our results.

2.8. *Confirmation of peptides for protein re-annotation*

Synthetic peptides with isotopic label at the C-terminal amino acid (^{13}C and ^{15}N) were ordered (Thermo Scientific, Ulm, Germany) to confirm peptide identifications, which were used for re-annotation of protein coding sequences. Fragment ion spectra of peptides were measured by direct infusion at the same instrument configuration with identical settings for CID according to the shotgun experiments.

Using these spectra a reference spectrum library was generated using NIST MS Search 2.0 (National Institute of Standards and Technology, Gaithersburg, MD). Match scores, reverse match scores and probability (%) scores were calculated for each of the identified peptides by comparing the corresponding MS/MS spectra with the reference library using NIST MS Search 2.0 identify search.

2.9. *Identification and filtering of signal peptide annotations*

For identification of signal peptides of annotated proteins, an additional database search was set up using the dedicated proteases with semi-proteolytic cleavage option. Here, specific cleavage of either the peptide N- or the C-terminus serves as a sufficient identification criterion. The precursor mass tolerance was reduced to 3 ppm since more than 95% of the previous identified peptides were found in this range. Thereby, the tremendous growth of search space for semi-proteolytic database searches should be limited. FDRs of semi-proteolytic peptides were adjusted for all experiments to less than 1% using thresholds for the delta mascot ion score and the X! Tandem – Log(Expect Scores). Additionally, spectra quality of remaining semi-proteolytic peptides was inspected manually.

Semi-proteolytic peptides with non-specific N-terminal cleavage were considered to be candidates cleaved by signal peptidases if no further peptide belonging to the same protein was identified N-terminal to their peptide loci. The minimum length of a signal peptide was defined to be seven amino acids. Potential signal peptides were additionally filtered according to the known characteristics of bacterial signal peptides [54] to distinguish signal peptidase cleavages from other proteolytic products.

The signal peptide structure is defined by

(i) a positively charged region near the N-term
(ii) followed by a hydrophobic region and
(iii) a three amino acid long signal peptidase recognition sequence.

The calculated net charge for the N-region from amino acids −15 to −21 relatively to the cleavage site had to be larger than zero. Thus, for calculation, lysine, arginine and the protein N-term were assumed to be positively charged, whereas aspartic and glutamic acids were expected to be negatively charged.

The GRAVY (grand average of hydropathy) score according to Kyte and Doolittle [55] for the hydrophobic region from amino acids −6 to −14 had to be larger than one. The recognition sequence was not used as a filtering criterion since it might differ to the motif reported for Gram-negative bacteria. Resulting signal peptidase cleavage sites were compared with computational signal peptide predictions of PerdiSi [56] and SignalP [57] with standard settings for Gram-negative bacteria.

2.10. *Peptide mapping and visualization*

Identified peptides were mapped to the *H. pylori* genome using tblastn with an e-value of 10^4, word size 2 and the low complexity filter turned off. Perfect and full length sequence matches were used. For a peptide with no perfect match, the maximum number of mismatches was set to the number of leucines and isoleucines as well as the number of X (see 2.5) in the peptide sequence. With this setting the best fit for the peptide to the DNA sequence was selected. The peptides were visualized in the UCSC microbial genome browser [49,53]. Note that each peptide might have multiple mappings. An UCSC track for each experiment has been compiled and can be visualized using the data sets and links available at http://www.bioinf.uni-leipzig.de/publications/supplements/12-023/. Multiple mappings are reflected in the UCSC tracks by the gray intensity of the mapped peptides. Each peptide initially receives a score of 1000 which is divided by the number of mappings. Thus, the score of a peptide with four genomic mappings is 250 which is displayed in light gray whereas a unique mapped peptide has a score of 1000 and a dark gray shading. Furthermore, the experiment and the number of mappings for each peptide are indicated in the sequence identifier (peptide ID:#mappings:experiment).

2.11. *RNAcode screen*

The Multiz pipeline [58] was used to generate genome wide alignments of 22 epsilon proteobacteria (Supplementary Table 1). Alignments were scanned for protein coding potential regions using RNAcode [17] with a p-value cutoff of 0.05 and the –stop-early and –best-only options. High scoring segments in the same reading frame and not more than 15 nucleotides apart were combined. This resulted in 3458 high scoring segments. Intergenic segments were screened for open reading frames. If the segment did not contain a complete open reading frame with a minimum length of 10 amino acids it was extended by 51 nucleotides in each direction. This resulted in 18 short protein coding gene predictions not yet contained in the published gene annotations.

2.12. *Submission to PRIDE and UniProtKB*

For PRIDE [59] (http://www.ebi.ac.uk/pride) submission, we carried out an additional database search with Mascot and X! Tandem using the SearchGUI [60]. Therefore we searched against a NCBI database of *H. pylori* strain 26695 complemented with the sequence corrections, signal peptide cleavage sites and missing annotations identified in our study. Search configurations were identical to those described in the initial database search. For pride xml export we used the software PeptideShaker (http://code.google.com/p/peptide-shaker/). The complete experimental data set is accessible on the PRIDE [59] web service.

3. Results

3.1. *Proteome analysis of **H. pylori** strain 26695*

We analyzed cell lysates from *H. pylori* by GeLC–MS and offline 2D-LC–MS to achieve broad coverage of the proteome. Furthermore we integrated the results published by Jungblut et al. [51]. Mascot and X! Tandem were used to search spectra against a compiled database including (i) the NCBI database of *H. pylori* strain 26695 and (ii) a six-frame translation of the genome. The database was concatenated with the same number of reverse entries to approximate and control the FDR (see Fig. 1 for an overview of the method). Peptide identification lists with according FDR calculations as well as a protein identification table are available in the supplementary material (Supplementary Material 2 and 3).

Peptide FDRs of all samples were calculated to be lower than 0.3% in our dataset (Supplementary Material 3). For GeLC–MS analysis two independent biological replicates were separated by 12% SDS-PAGE and analyzed by LC–MS/MS after in-gel digestion with trypsin. The database search revealed 1091 protein identifications according to the NCBI part of the database (replicate I: 1018, replicate II: 1061) by at least two peptides and covers 69% of the predicted proteome. The two replicates show a protein identification overlap of 91% which demonstrates a good reproducibility.

Additionally, SEC was used to enrich proteins with a MW below 25 kDa in order to cover small open reading frames. Four fractions were prepared, aliquoted and proteins were separately digested by endoproteases trypsin, LysC and AspN. Overall 385 proteins (24% proteome coverage) were identified by this 2D-LC–MS approach.

LysC provided the best results with 368 protein identifications (replicate I: 323, replicate II: 339) followed by trypsin with 291 (I: 252, II: 270) and AspN with 142 (I: 133, II: 93). This approach was focused on the identification of low MW proteins, showing 30% proteome coverage below 20 kDa. In comparison to the GeLC–MS approach, 24 additional proteins could be identified which have all a MW below 17 kDa. This represents an increase of 18% for this MW range.

Overall, we discovered 1115 proteins in our dataset by at least two peptides and a peptide FDR lower than 1%. This corresponds to a *H. pylori* proteome coverage of 71%.

In the re-analyses of the most comprehensive proteome dataset that has been published so far for *H. pylori* strain 26695 (Jungblut et al. [51]), 549 proteins corresponding to 35% of the proteome were identified. In comparison to our results only one additional protein (gi 15645950) was identified. In contrast to our dataset, peptide FDRs of this dataset were higher than 1% for two fractionations (pellet fraction: FDR 1.1%, startline fraction: 3.1%).

As a complementary gene prediction approach we used RNAcode [17]. Our analysis gave 3485 high scoring segments of which 89% (3106/3485) was found in-frame with annotated coding sequences. The screen has a sensitivity of 90.1% since 1420 of 1576 annotated CDS were recovered by at least one overlapping RNAcode hit. These results are highly similar to our previous analysis in *Escherichia coli* [17]. A more detailed look shows that 1238 CDS are only recovered by RNAcode hits within the gene boundaries. The remaining 182 CDS represent

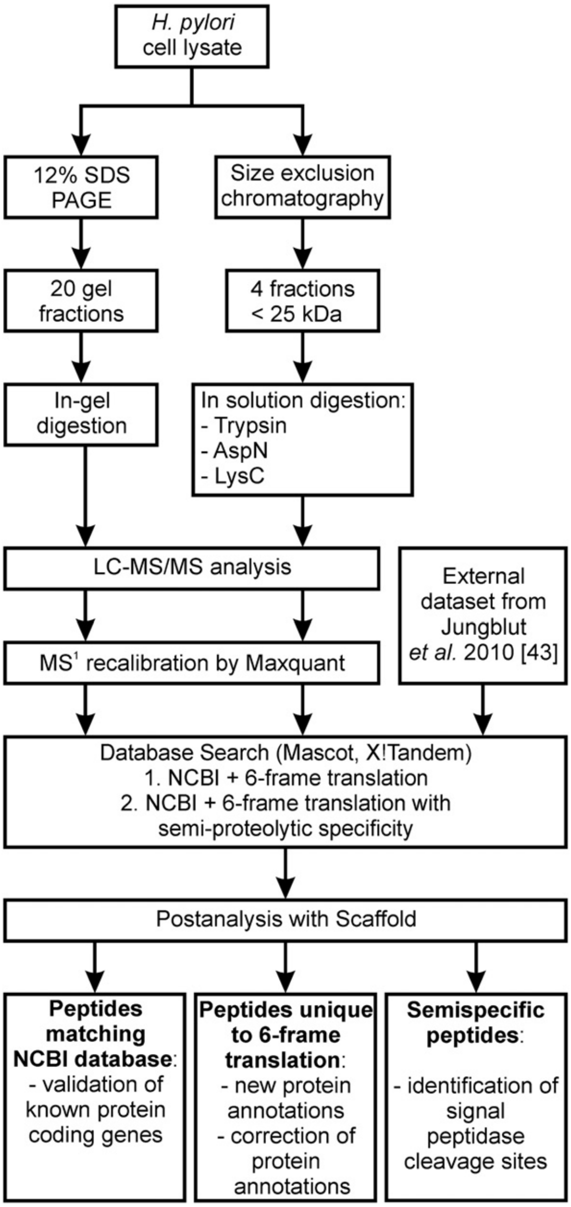

H. pylori cell lysate
12% SDS PAGE
Size exclusion chromatography
20 gel fractions
4 fractions < 25 kDa
In-gel digestion
In solution digestion:
- Trypsin
- AspN
- LysC
LC-MS/MS analysis
External dataset from Jungblut et al. 2010 [43]
MS¹ recalibration by Maxquant
Database Search (Mascot, X!Tandem)
1. NCBI + 6-frame translation
2. NCBI + 6-frame translation with semi-proteolytic specificity
Postanalysis with Scaffold
Peptides matching NCBI database:
- validation of known protein coding genes
Peptides unique to 6-frame translation:
- new protein annotations
- correction of protein annotations
Semispecific peptides:
- identification of signal peptidase cleavage sites

Fig. 1 – Experimental workflow of the proteogenomic analysis. Proteins extracted from *Helicobacter pylori* cell lysates were separated by 12% SDS-PAGE and size exclusion chromatography. Gel fractions were digested by trypsin whereas trypsin, AspN and LysC were separately applied to SEC fractions. Samples were analyzed by LC–MS/MS. MS1 data was recalibrated using Maxquant. At this point the dataset of Jungblut et al. [51] was integrated. A database search against a reverse concatenated database of the NCBI entries and the six-frame translation was performed. Additionally a database search with semi-proteolytic specificity was made. After post-analysis with Scaffold peptides were mapped to the NCBI database. Peptides which were unique to the six-frame translation were subjected to further analyses to discover new and to correct existing protein annotations. Semispecific peptides were used to identify signal peptidase cleavage sites.

candidates with erroneous annotated gene boundaries. Of these, 60 CDS were recovered by gene boundary overlapping RNAcode hits only and 122 CDS have both types of hits those within the annotated CDS and those overhanging the gene boundaries. In this study RNAcode predictions were used to support the experimentally identified annotation errors.

3.2. *Refinement of protein annotations by proteogenomics*

For the identification of novel protein sequences, searches against a reverse concatenated database including the NCBI database of *H. pylori* strain 26695 and a six-frame translation of the genome were performed. Out of the 21915 peptides being identified, 21,774 could be mapped to the 1576 existing protein coding annotations. However, 57 peptides (0.3%) were unique to the six-frame translation and match to unique locations in the genome.

Peptides that are unique to the six-frame translation were classified according to their genomic location. Additionally, a BLAST analysis against the NCBI reference sequence database was applied to determine similar proteins in other *H. pylori* strains (e.g. similar proteins from *H. pylori J*99 for HP1186 and HP0694). Both protein sequences from other strains derived by BLAST as well as sequences with new translation start sites were added to the existing database for an additional search. With this strategy, we were able to identify additional peptides which validate presumed DNA sequencing errors. These sequencing errors result in frame-shift errors which lead to erroneous truncated protein annotations. The peptides which were used for identification of new or correction of existing protein annotations are shown in Supplementary Table 2. The following refinements of protein annotation were submitted to the UniProt database to ensure public availability.

3.2.1. *Identification of missing protein annotations*

We could identify four missing protein annotations. Three proteins were missing due to DNA sequencing errors that resulted in frame-shifts within a protein coding sequence. The ferrous iron transporter protein A gene was simply missing in the annotation by Tomb et al. [61].

Seven different peptides were identified for the coding region HP0058 (Supplementary Table 2) which was not annotated in the NCBI protein database of *H. pylori* strain 26695. Already, Medigue et al. [18] reported that this region contains an authentic frame-shift and is not the result of a sequencing artifact. The contingency gene of this hypothetical protein was identified by GeneMark [62,63]. Interestingly, Specht et al. [64] also reported that two cytosines were missing at the genomic position 62,013. The corrected protein sequence comprises 400 amino acids and a molecular weight (MW) of 46 kDa. Our results provide experimental evidence of this prediction. Peptides were identified in fractions 12 and 13 (45–57 kDa) of both in-gel digestion replicates supporting this MW.

The annotation for the hypothetical protein HP0744 was also missing in the NCBI protein database of strain 26695. We identified nine different peptides that could be mapped to this region (Supplementary Table 2). Peptides belonging to this region were identified in the same gel fraction (fraction 11, 35–45 kDa) of both biological replicates. Again, Medigue et al. [18] published that this region has an authentic frame-shift and could code for a protein. Indeed, seven peptides are located on frame −1 whereas four peptides are located on frame −2. An insertion of one nucleotide at the stop codon can correct the frame-shift, so both parts of HP0744 are on frame −1.

Furthermore, the gene HP0619 was not part of the NCBI reference protein database. We identified five peptides on frame +2 and nine peptides on frame +1 in this region (Supplementary Table 2). Once more, a frame-shift error was predicted in this region [18]. In fact, the previously assigned stop codon can be converted to a leucine codon by insertion of a thymine at nucleotide position 665045. Hereby, the frame-shift error is corrected and all identified peptides are located on the same frame. Oleastro et al. [65] identified homology between the glycosyltransferase jhp0563 of strain J99 and HP0619 of strain 26695. Transcription of these genes was validated for both strains by reverse transcription PCR analysis [65].

Three different peptides identified a new protein coding gene (DNA 0100057) in the intergenic region of the ORFs HP0585 and HP0586 (Fig. 2). BLAST analysis revealed that this region encodes for the ferrous iron transport protein A in four other *H. pylori* strains (Lithuania75, SNT49, G27 and ELS37) with 100% identity and an expectation value of 5×10^{-29}. Conclusively, the ferrous iron transporter protein A gene has been missed during annotation by Tomb et al. [61]. However, the identical protein sequence was already predicted by an unpublished observation made by Medigue and Bocs (gi 13431987, P57798.1). Recently, the sequence was also submitted by the Research Institute for Physico-Chemical Medicine Moscow to the NCBI database and was inserted as a provisional entry (not yet published).

3.2.2. *Identification of erroneously annotated translation start sites*

In addition to missing protein annotations, we could also identify four protein annotations with an extended sequence at the protein N-termini. The misannotations of translation start sites for two proteins were due to frame-shift errors which are a result of DNA sequencing errors, whereas the other two protein starts were simply wrongly annotated.

We identified a peptide within the intergenic region of HP1433 and HP1434 which are both encoded on the minus strand (Fig. 3). It is in frame with the downstream gene HP1433 and there is no stop codon in between these sequences. In conclusion, the hypothetical protein HP1433 (gi 15646042) has a wrong start codon assignment. Protein annotations in other *H. pylori* strains include the identified peptide within the annotated sequence which contradicts the current annotation. Additionally, the new start site is supported by a highly significant RNAcode prediction (p-value of 1.1×10^{-14}). The extended protein sequence has 893 amino acids and a MW of 104 kDa. In line, all peptides belonging to HP1433 and the peptide for the start site correction were identified in fractions 17–20 (100–300 kDa) supporting the MW. Based on these findings we suggest a re-annotation of HP1433 in *H. pylori* strain 26695.

Moreover, the protein start for S-ribosylhomocysteinase (HP0105) was erroneously annotated. We identified one peptide upstream of the previous coding sequence annotation (Supplementary Table 2). BLAST analysis showed that *H. pylori* strain XZ274 has another translation start site annotated for this protein which includes the peptide sequence upstream HP0105. In the second database search including all three possible start codons for HP0105, we identified three additional peptides which confirm the new translation start (methionine codon ATG at nucleotide position 113295). The UniProt database had already included the corrected start site inferred by homology.

Two peptides were identified between the protein coding regions HP0760 and HP0761 (Supplementary Table 2) which neither match to the same frame of HP0760 (phosphodiesterase) nor HP0761 (hypothetical protein). BLAST analysis showed that both peptides match perfectly to phosphodiesterase of many other *H. pylori* strains (e.g. P12, Lithuania75). We conclude that the protein coding region HP0760 was truncated due to a frame-shift error as suggested by Medigue et al. [18]. In contrast to the NCBI reference database, the sequence was already corrected at UniProt according to homology comparison.

Seven different peptides give evidence for a wrongly annotated translation start site of HP0564 (gi 15645189) (Supplementary Table 2). The supposed correction is strengthened by two peptides which overlap with the previously annotated protein start. Additionally, the start codon of the gene HP0564 is annotated as GTG which is usually coding for valine. However, GTG is translated into methionine when it is a start codon. The two peptides, which are N-terminal extended over the previously annotated start, show that the triplet GTG is translated into valine at this position and thus increase the confidence of the start site correction. For further validation, we included sequences with different start sites to our database search. Thereby, we could identify the N-terminus in both biological replicates of the AspN digestion of SEC fractions (Supplementary Table 2).

3.2.3. *Identification of erroneously annotated translation termination due to frame-shift errors*

Protein annotations for HP1186 and HP0694 are found to be truncated at the C-terminus because of DNA sequencing errors resulting in frame-shifts. Five different peptides downstream of the gene HP1186 coding for carbonic anhydrase (gi 15645800) were identified in different samples (Supplementary Table 2). Additionally, one of these peptides could also be identified in the dataset of Jungblut et al. [51]. Protein BLAST analysis of the identified peptides resulted in 100% identity matches to the carbonic anhydrase of other strains like J99 (gi 15612177) suggesting a DNA sequencing error (Supplementary Fig. 1). The second database search including the protein sequence from strain J99 identified an additional peptide which is located upstream relatively to the identified peptides. This suggests a re-annotation of the 3′ end of HP1186 according to the previously reported frame-shift error for HP1186 [12,18]. Indeed, there were two errors in the DNA sequence. At position 1256328 a thymine was missing whereas adenine at position 1256383 has to be deleted. This explains why the peptides found downstream of the gene HP1186 are on the same frame.

The corrected protein sequence of HP1186 comprises 247 amino acids and has a MW of 28 kDa. All peptides were identified in fractions 6 or 7 of the in-gel digestion corresponding to a MW of 20 to 25 kDa. A putative signal peptidase cleavage site after the first 18 amino acids (ΔMW 1848 Da) predicted by PerdiSi [56] and SignalP [57] could be a reasonable explanation for the mass difference.

The predicted coding region of the hypothetical protein HP0694 (gi 15645317) is also wrongly annotated due to a DNA sequencing error downstream of the annotated C-terminus.

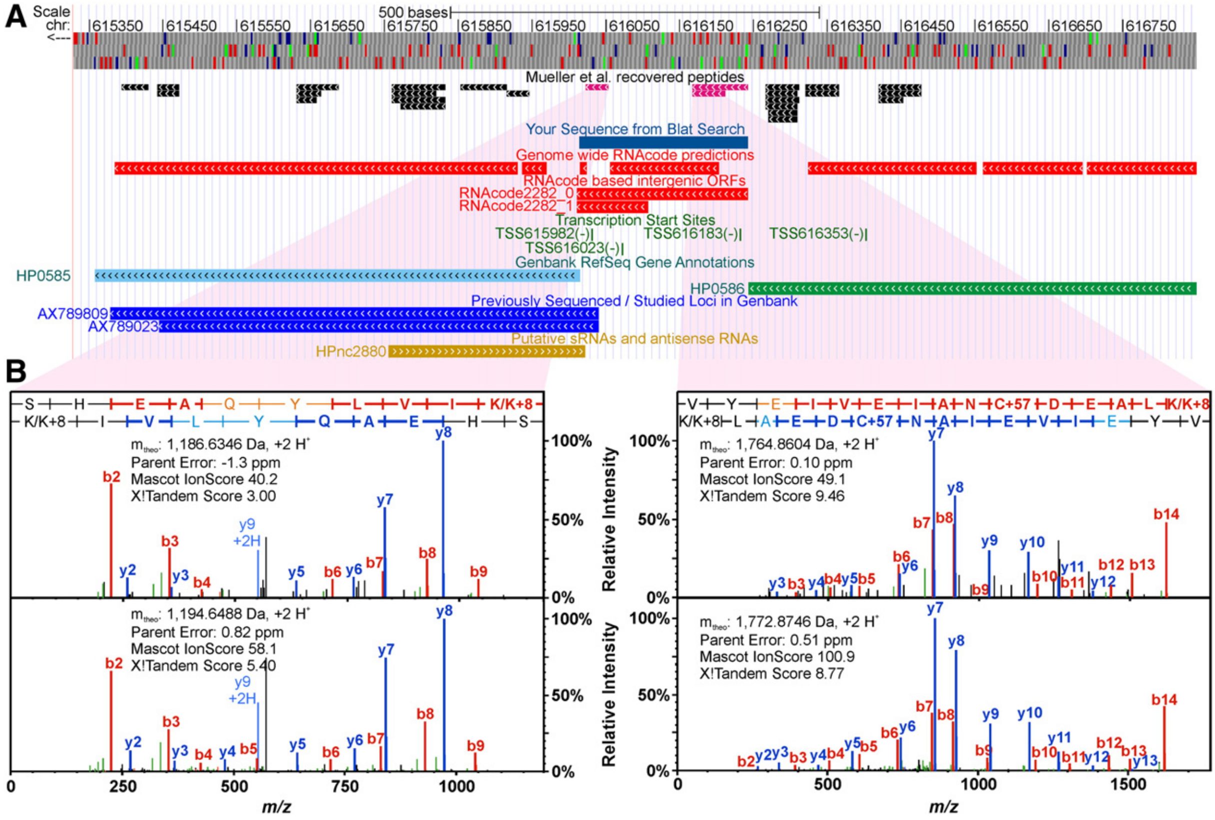


Fig. 2 – (A) Three peptides (magenta) mapped into the intergenic region of HP0585 (endonuclease III) and HP0586 (hypothetical protein). In addition two RNAcode predictions are found at this locus which can be extended to ORFs. Note that RNAcode2282_1 is a sub-region of RNAcode2282_0 and together with the protein expression data the longer ORF is most plausible. A sequence search against the NCBI refseq database matches with up to 100% identity to the ferrous iron transporter protein A annotated in various *Helicobacter pylori* strains. The possible independent expression of the homolog in the studied strain is further supported by the annotated transcription start TSS16353. (B) Confirmation of two identified peptides by comparison of the MS/MS spectra of the experiment (upper spectra) and the corresponding synthesized peptide (lower spectra) containing 6 × ^{13}C 2 × ^{15}N-labeled lysines.

The peptide VAFTITDISK belongs to a region next to the 3′ end of HP0694 (Supplementary Table 2). Protein BLAST of this peptide revealed 100% identity with outer membrane proteins of other strains (e.g. J99; gi 15611701). An additional database search including these protein sequences succeeded in additional peptide identifications (Fig. 4). Moreover, all peptides belonging to this protein were identified in fraction 9 (approx. 26–29 kDa) of the in gel digestion. The discrepancy between the theoretical (38 kDa) and the experimental derived MW can be partly explained by signal peptide cleavage after amino acid 17 which was predicted by PerdiSi [56] and SignalP [57]. These findings strongly indicate a sequencing error that results in a pre-major stop due to a frame-shift error [18] for the predicted coding region of HP0694. Manual inspection of the DNA sequence revealed two sequencing errors in this region. Firstly, the stop codon for HP0694 has to be converted in an arginine codon (AGG) by deletion of a thymine at position 745343. Secondly, an adenine has to be inserted at position 745389.

3.2.4. *Validation of novel and corrected protein annotations*

To validate the peptide identifications leading to corrected protein annotations of *H. pylori* strain 26695, we ordered 12 heavy peptides labeled with ^{15}N and ^{13}C isotopes at the C-terminal amino acid. Tandem MS spectra of the synthetic peptides were acquired using direct infusion. Comparison of MS/MS spectra of the biological samples with the corresponding synthetic peptides correlates well for all tested peptides and further validates the above described revised gene annotations (Figs. 2–4, Supplementary Figs. 9–22). The reverse match score as well as the correlation probability of NIST MS search are listed in supplementary Table 2.

Furthermore, we identified transcripts for all newly annotated proteins in a whole transcriptome analysis from *H. pylori* 26695 based on high-throughput sequencing approach of cDNA libraries (RNA-Seq) (S. Pernitzsch and C. M. Sharma, unpublished data, Supplementary Method 1, Supplementary Figs. 2–5). The RNAseq data from *H. pylori* strain 26695 confirmed transcription for the intergenic region of HP0585 and HP0586 as well as for the coding regions HP0619, HP0744 and HP0058.

3.3. *Identification of signal peptides*

The export of secreted proteins as well as proteins which are located in the inner or outer membrane or the periplasm

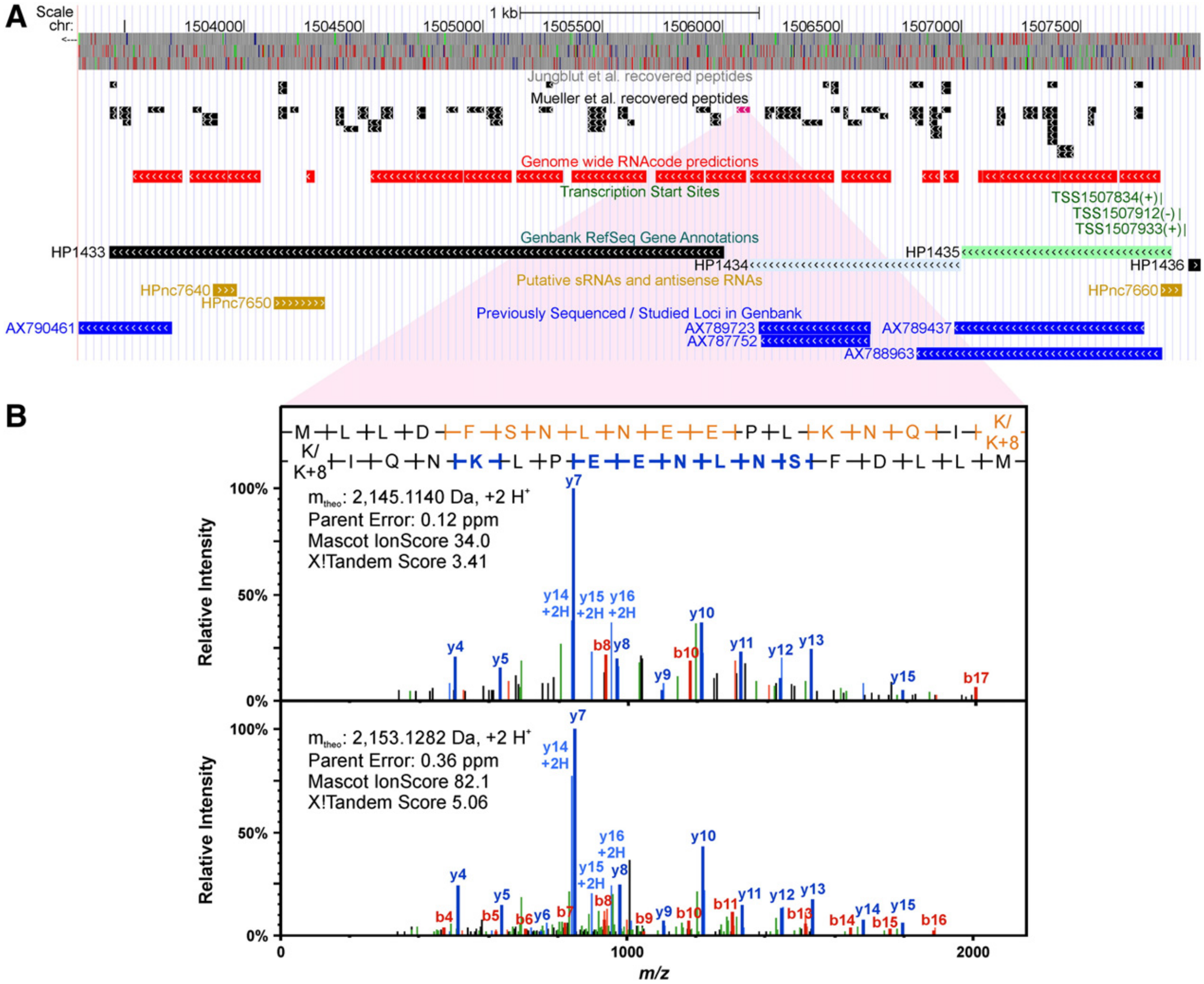


Fig. 3 – (A) Genomic location of HP1433 a hypothetical protein which is encoded in an operon together with the formyltetrahydrofolate hydrolase (HP1434) and the protease IV (HP1435). The operon is transcribed from the transcription start site TSS1507912 which is located upstream of the gene HP1435. Beside putative anti-sense RNAs (HPnc yellow) several previously studied loci are annotated (blue). The latter correspond to a protein–protein interaction study. The RNAcode predictions (red) together with the identified peptides (black) in combination with the magenta colored peptide suggest the HP1433 start codon position correction directly downstream to the HP1434 stop codon. (B) Confirmation of this peptide by comparison of the MS/MS spectra of the experiment (upper spectrum) and the corresponding synthesized peptide (lower spectrum) containing 6 × ^{13}C 2 × ^{15}N-labeled lysines.

usually requires a N-terminal signal sequence which is removed by signal peptidases [54]. Signal peptide cleavage leads to new protein N-termini. After enzymatic digestion in proteomic studies, peptides of new protein N-termini have a specifically cleaved C-terminus but a non-specifically cleaved N-terminus according to the used protease. Thus, peptides near the protein N-termini with non-specific cleaved N-terminus were considered to be potentially cleaved by a signal peptidase. Signal peptide candidates were identified by a database search allowing for semi-specific peptides.

Overall, 72 candidates were identified with a FDR below 1% of which 63 fulfilled our filtering criteria for signal peptide identification (Supplementary Material 4). Thirty eight signal peptide sequences were identified in more than one sample. The analysis of the dataset from Jungblut et al. [51] provided an independent validation of eight signal peptides and the identification of one additional sequence. The structure of the identified signal peptide sequences is illustrated in Fig. 5A and B with a sequence logo graphic [66]. Leucine (75%) is predominately localized at the −3 position relative to the cleavage site. The −1 position is mainly alanine (84%).

A search for signal peptide sequences for *H. pylori* 26695 in the UniProt database revealed only one experimentally validated signal peptide for the Cytochrome c-553 (HP1227). Computational tools such as PerdiSi [56] and SignalP [57] provide 191 and 182 significant predictions, respectively (Fig. 5C). However, only 28 of the experimentally validated signal peptides were supported by significant predictions of at least one algorithm (Fig. 5C).

In order to improve the prediction accuracy, we lowered the predefined thresholds for the significance scores of both tools (PerdiSi Score > 0.2, SignalP Dmaxcut > 0.3) and added our filtering criteria according to the signal peptide structure. Furthermore, we restricted the amino acids at the −1 to −3

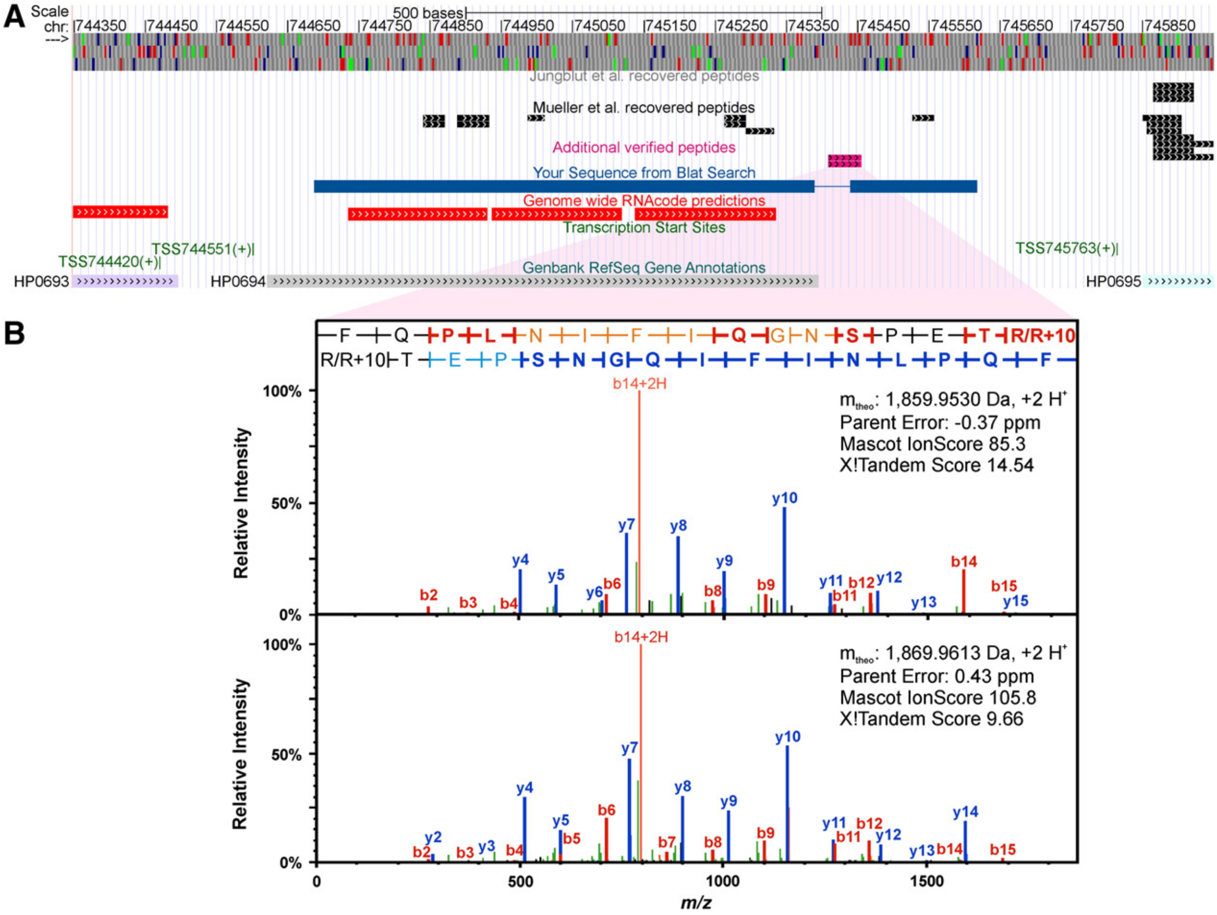


Fig. 4 – (A) Genomic location of HP0694 (hypothetical protein). HP0694 has two alternative transcription start sites (TSS744420 and TSS744551, green). The RNAcode prediction (red) resamples the annotated open reading frame. One peptide was identified in between the genes HP0694 and HP0695. The BLAST search of this peptide matched perfectly to the protein sequence gi 15611701 annotated in *Helicobacter pylori* strain J99. This indicates a genomic sequencing error (thin line within the blue box). An additional peptide (magenta) can be identified if the corrected DNA sequence is used in the database. It was found in two biological replicates. (B) Confirmation of this peptide by comparison of the MS/MS spectra of the experiment (upper spectrum) and the corresponding synthesized peptide (lower spectrum) containing 6 × ^{13}C 2 × ^{15}N-labeled lysines.

positions according to our findings. The amino acid that occupies the position −1 had to be either A, V, Y G, S or L whereas the position −3 was restricted to either L, A, I, V, S or C, respectively. Additionally, the position −2 must not be proline.

The number of significantly reported signal peptidase cleavage sites was slightly increased by the new criteria. Remarkably, the application of new significance criteria for the prediction tools provided support for 17 additional signal peptides. The overlap of significant predictions of PerdiSi and SignalP was increased by 42% to 139 (Fig. 5D).

4. Discussion

The human pathogen *H. pylori* is a Gram-negative Epsilonproteobacterium which has been associated with many gastric diseases like gastritis, duodenal ulcers as well as gastric cancer. It colonizes about half of the human's population, but approximately 80% of the infected individuals are asymptomatic [67,68]. The complete genome sequencing of the strain 26695 [61] in 1997 provides a fundamental basis for studying *H. pylori* on the genome, transcriptome and proteome levels. Proteomic studies of *H. pylori* are an inherent part of basic research of this pathogen. During the last years, proteomic studies offered further insights into the adaption to acidic [69] or oxidative stress [70,71] as well as pathogenic mechanisms [72,73].

Nevertheless, proteomic studies are strongly dependent on the protein database quality. Different genome studies already showed that there might be discrepancies in coding sequence annotation of different *H. pylori* strains as a result of DNA sequencing errors or erroneous predictions [12,18,19]. However, this data is solely based on bioinformatics and not validated by biological experiments. Here, we show that proteogenomics offers the opportunity to identify new protein coding genes and to correct erroneous protein annotations on the basis of experimental study results. This also includes the detection and correction of DNA sequencing errors that result in frame-shifts.

However, proteogenomic studies require a high proteome and protein sequence coverage of MS data. Our study revealed

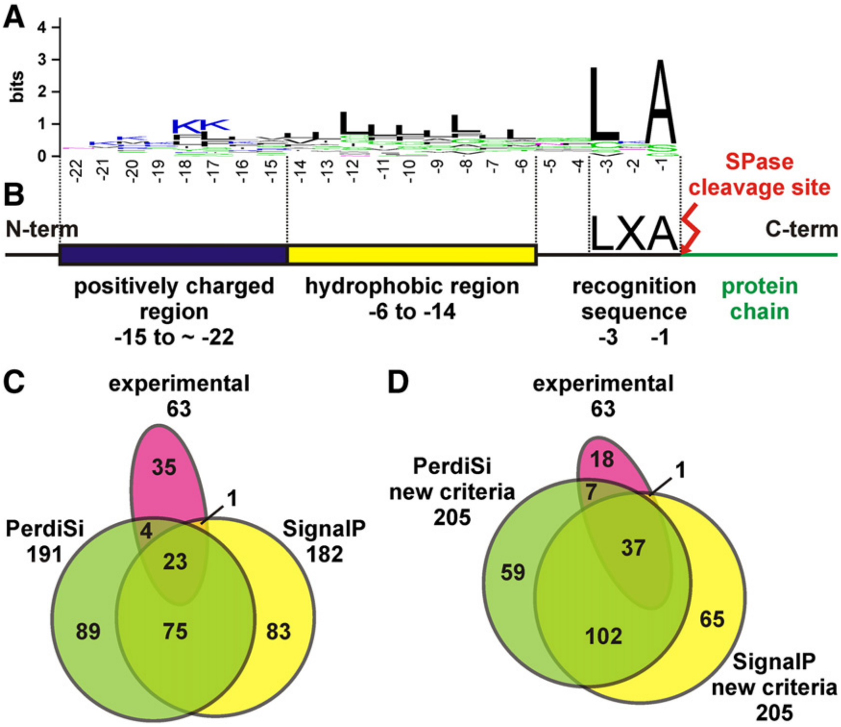


Fig. 5 – Comparison of identified signal peptide sequences with software predictions. (A) Sequence logo of the experimentally identified signal peptides (sequence logo graphic was created with the web-based tool WebLogo, version 2.8.2, [66]). The hydrophobic region and the positively charged N-terminal region are clearly identifiable. The predominant SPase recognition sequence is LXA for *Helicobacter pylori*. (B) Schematic signal peptide structure for *H. pylori*. The positively charged region is between amino acids −15 and −22 relatively to the SPase cleavage site whereas the hydrophobic region is between amino acids −6 and −14. The predominant recognition sequence LXA is presented for the −3 to −1 positions. (C) Comparison of experimentally derived signal peptides with significant predictions from PerdiSi and SignalP. (D) Comparison of experimentally derived signal peptides with predictions from PerdiSi and SignalP after adaption of the significance criteria to the signal peptide structure of *H. pylori*.

1115 proteins representing 71% of the annotated proteome with average protein sequence coverage of 49%. A similar proteogenomic study of *Pseudomonas fluorescens* Pf0-1 covered 66% of the annotated and identified 16 new ORFs [74] which is comparable to our results. However we still miss 29% of the proteome either due to false annotations or experimental limitations such as the detectable minimum protein weight of approximately 5 kDa in our approach. The latter might also be a reason why we miss the recently discovered short transcripts harboring conserved open reading frames [20] even though we already significantly increased the coverage of low molecular weight proteins due to a SEC based enrichment strategy.

Our dataset allowed us to unambiguously correct six protein annotations (HP1433, HP0105, HP0760, HP0564, HP1186, HP0694) and to discover four proteins which were not part of the NCBI reference sequence database (HP0058, HP0744, HP0619, intergenic region HP0585–0586—ferrous iron transport protein A). Five of these protein annotations were additionally validated by comparing MS/MS spectra of biological samples with synthetic peptides. Furthermore, seven of the new annotated respectively corrected protein annotations are supported by significant RNAcode predictions. We also show that proteogenomics has the ability to identify and correct DNA sequencing errors. Three previously missing annotations as well as three erroneous annotations were the result of DNA sequencing errors. Thus, the application of proteomics in combination with comparative genome analysis offers new information which cannot be gained by one of these techniques alone.

Finally, all newly identified proteins were also found in a whole transcriptome analysis of *H. pylori* strain 26695 that was grown under comparable conditions (S. Pernitzsch and C. M. Sharma, unpublished data, Supplementary Figs. 2–5).

Remarkably, the new annotated and corrected proteins are supposed to be of high interest for further studies. For example, the protein which is located between HP0585 and HP0586 is similar to the ferrous iron transport protein A of other *H. pylori* strains. Iron transport is essential for the survival of *H. pylori* in the stomach [75]. Iron is transported into the cell and stored by ferritin to prevent iron scarcity [76]. Velayudhan et al. [77] investigated the role of the ferric iron transporter B for iron uptake and virulence. However, they did not study the influence of the transporter A because it was missing in the annotations of Tomb et al. [61]. Furthermore, transcription of the infection related gene *vacA* is up-regulated under iron deficient conditions [78].

In addition, the previously missing annotation for HP0619 which codes for a putative lipopolysaccharide (LPS) biosynthesis protein might be a drug target candidate for the inhibition of the LPS biosynthesis pathway [79]. The protein HP0058 which was not annotated in the NCBI database has shown to be important for the morphology and motility of *H. pylori* [64]. This coiled-coil-rich protein forms filamentous structures which are essential for the helical shape of *H. pylori* [64]. HP0058 deletion mutants are straight shaped and exhibit reduced motility in soft agar assays [64] which might indicate attenuated colonization efficiency.

Furthermore, we investigated signal peptide cleavage sites of the annotated proteins. *H. pylori* encodes for two different signal peptidases (SPases I and II) [54]. Here we demonstrate that high accurate MS allows the identification of signal peptide sequences in a shotgun approach. Nevertheless, database searches with semi-proteolytic specificity require a careful adjustment of FDRs since the search space increases exponentially. Our FDRs were adjusted to less than 1% using only semi-proteolytic peptides which results in more restrictive but much more significant signal peptide candidate identifications.

Additional filtering criteria were applied according to the known signal peptide structure of bacteria [54] and resulted in 63 significant signal peptides out of 77 candidates. The dataset of Jungblut et al. [51], offered only one additional signal peptide compared to our data. Since this data was not acquired by high accurate MS, the quantity of identifications is lower compared to the 62 signal peptides of our dataset.

Signal peptide candidates which did not fulfill our criteria might be produced by side-specificity of utilized proteases or could be cleavage products of other proteases. Our criteria may lead to higher false negative rates but improve the confidence of our results. For example, the doubtful assignment for the uncharacterized protein HP0659 with a signal peptide length of 103 was sorted out due to both thresholds for the hydrophobic and positively charged region.

Signal peptidases from Gram-negative bacteria require more or less conserved amino acids at the −1 and −3 positions relative to the cleavage site [54]. We showed that the predominant recognition sequence for the signal peptidases of *H. pylori* is LXA. Nevertheless, alanine, isoleucine, valine, serine and cysteine were also detected at the −3 position, whereas glycine, serine, valine, leucine and threonine are also suitable at the −1 position. Since no cysteines were found on the +1 position, we consider that all identified cleavage sites are targeted by the signal peptidase I.

We compared our results with those derived by two different signal peptide prediction tools. However, only 44% of our findings were supported by significant predictions. The low overlap results from either non-significant scoring by these tools or erroneous cleavage site predictions. The prediction algorithms of PerdiSi and SignalP were trained with datasets of experimentally validated signal peptides from Gram-negative bacteria [56,57]. The moderate prediction accuracy could be a result of the lack of experimentally determined cleavage sites as well as the missing subdivision according to phylogeny. This may lead to algorithms which are very strongly oriented towards well studied bacteria such as *E. coli*. This hypothesis is substantiated by the fact that the predominant signal peptidase recognition sequence is thought to be AXA for Gram-negative bacteria, whereas our data suggest rather LXA for *H. pylori*. Indeed, the signal peptides of *E. coli* and *H. pylori* show clear differences (Supplementary Fig. 23). The length and position of the hydrophobic as well as the predominant signal peptidase recognition sequence (AXA) are different for *E. coli*.

In order to increase the confidence of these tools for *H. pylori*, we applied our filtering criteria with additional restriction of amino acids for the −3 and −1 positions according to our findings and lowered the individual scoring thresholds. Hereby, we improved the support for our data to 71% and increased the overlap of the SignalP and PerdiSi from 98 to 139 predictions (Fig. 5C and D). However, correctness of signal peptidase cleavage site predictions can only be improved by modification of the individual algorithms. Therefore, we encourage the scientists that work on signal peptide prediction tools to use our findings to enhance the prediction accuracy of these tools.

To our knowledge no other study has investigated the specificity of the signal peptidases of *H. pylori*. Signal peptidases are essential enzymes for the viability of bacterial cells [54,80] and are involved in pathogenesis [81,82] Therefore signal peptidases could be novel targets for antibiotics [80]. Additionally, inclusion of signal peptides into the database could increase peptide and protein identifications of future proteome studies.

Both signal peptidase cleavage sites, corrected and missing protein annotations were submitted to the UniProt protein database. For visualization of our data, we also offer custom tracks for the UCSC microbial genome browser to support further proteome and transcriptome studies (http://www.bioinf.uni-leipzig.de/publications/supplements/12-023/).

In conclusion, using proteogenomic approaches for protein coding sequence annotations will help to improve and complete protein databases.

This approach is easily adaptable to other bacterial species. For eukaryotes, the database construction has to be slightly modified. A direct translation of eukaryotic DNA sequences would lead to a tremendous increase of the protein database sizes due to high content of non-coding regions. Instead of the DNA, a translation of mRNA transcripts into protein sequences has to be performed. For this purpose, one has to utilize for instance freely available transcriptome datasets. Furthermore, high quality proteomic datasets which are deposited at PRIDE [83] provide the possibility to carry out proteogenomic analyses without extensive measurements.

Some might argue that the utilization of six-frame translation databases in proteomic studies would solve the problem of erroneous and missing annotations. However, the database size increases approximately six-fold for bacteria like *H. pylori* with a small genome and a high amount of protein coding content. Other organisms and especially eukaryotes have large amounts of non-coding DNA. A six-frame translation of the human DNA generates a database which is larger than the whole UniProt database. Additionally, the search space increases exponentially when variable modifications are used. This leads to higher FDRs and increased processing time. Moreover, biological information of identified proteins is usually retrieved by accession numbers of publicly available databases such as NCBI or UniProt. A plain search against the genome would need additional extensive data processing to gain further biological information. Furthermore, identification of peptides which cover the N-termini of proteins is limited due to the fact that the start and end positions of gene products are often not exactly detected by a six-frame translation. Therefore database searches against six-frame translations are impracticable for conventional proteomic studies.

We expect that further proteomic studies will strongly benefit from proteogenomics because of their dependency on the protein database quality. Here, we showed that even protein databases of well-studied organisms like the investigated *H. pylori* strain 26695 are not error free. Proteins of particular biological interest like the ferrous iron transport protein A, the coiled-coil-rich protein HP0058 and the lipopolysaccharide biosynthesis protein HP0619 were actually missing in the annotations. Database entries for these proteins might be important to study biological pathways involved in pathogenesis or drug response. Our approach additionally demonstrates that frame-shift errors, which are a result of inaccurate DNA sequencing, can be identified and corrected by proteogenomics. Therefore, we highly recommend the application of proteogenomics within new genome sequencing projects to generate more accurate protein coding sequence annotations and to increase the experimental support of predicted protein coding genes.

Acknowledgment

We thank C. M. Sharma (Research Center for Infectious Diseases (ZINF), Würzburg) for providing cell pellets of *H. pylori* strain 26695, sharing unpublished RNA-Seq data and critical comments on the manuscript. We thank Konrad U. Förstner (Research Center for Infectious Diseases (ZINF), Würzburg) for help with analysis of RNA-Seq data. Additionally, we thank Andrea Tanzer for the support to setup the UCSC track hub. We also thank Sven Haange for critical comments on the manuscript. This work was funded by the Deutsche Forschungsgemeinschaft (SFB/TR67 and STA 850/10-2). It was also kindly supported by Helmholtz Impulse and Networking Fund through Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2013.04.036.

REFERENCES

[1] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 1977;265:687–95.

[2] Staden R. A strategy of DNA sequencing employing computer programs. Nucleic Acids Res 1979;6:2601–10.

[3] Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 1986;51(Pt. 1):263–73.

[4] Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 2012;40:D115–22.

[5] Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 2008;9:75.

[6] Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 2007;23:673–9.

[7] Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 2005;33:W451–4.

[8] Malys N, McCarthy JE. Translation initiation: variations in the mechanism can be anticipated. Cell Mol Life Sci 2011;68: 991–1003.

[9] Schmidtke C, Findeiss S, Sharma CM, Kuhfuss J, Hoffmann S, Vogel J, et al. Genome-wide transcriptome analysis of the plant pathogen *Xanthomonas* identifies sRNAs with putative virulence functions. Nucleic Acids Res 2012;40:2020–31.

[10] Moll I, Grill S, Gualerzi CO, Bläsi U. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. Mol Microbiol 2002;43:239–46.

[11] Jäger D, Sharma CM, Thomsen J, Ehlers C, Vogel J, Schmitz RA. Deep sequencing analysis of the *Methanosarcina mazei* Gö1 transcriptome in response to nitrogen availability. Proc Natl Acad Sci 2009;106:21878–82.

[12] Boneca IG, Hd Reuse, Epinat JC, Pupin M, Labigne A, Moszer I. A revised annotation and comparative analysis of *Helicobacter pylori* genomes. Nucleic Acids Res 2003;31:1704–14.

[13] Dinger ME, Pang KC, Mercer TR, Mattick JS. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol 2008;4:e1000176.

[14] Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. Small membrane proteins found by comparative genomics and ribosome binding site models. Mol Microbiol 2008;70: 1487–501.

[15] Bakke P, Carney N, Deloache W, Gearing M, Ingvorsen K, Lotz M, et al. Evaluation of three automated genome annotations for *Halorhabdus utahensis*. PLoS One 2009;4:e6291.

[16] Warren AS, Archuleta J, Feng WC, Setubal JC. Missing genes in the annotation of prokaryotic genomes. BMC Bioinformatics 2010;11:131.

[17] Washietl S, Findeiss S, Muller SA, Kalkhof S, von Bergen M, Hofacker IL, et al. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA 2011;17:578–94.

[18] Medigue C, Rose M, Viari A, Danchin A. Detecting and analyzing DNA sequencing errors: toward a higher quality of the *Bacillus subtilis* genome sequence. Genome Res 1999;9:1116–27.

[19] Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 1999;397:176–80.

[20] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 2010;464:250–5.

[21] Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief Funct Genomic Proteomic 2008;7: 50–62.

[22] Bindschedler LV, McGuffin LJ, Burgis TA, Spanu PD, Cramer R. Proteogenomics and in silico structural and functional annotation of the barley powdery mildew *Blumeria graminis* f. sp. hordei. Methods 2011;54:432–41.

[23] Borchert N, Dieterich C, Krug K, Schutz W, Jung S, Nordheim A, et al. Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models. Genome Res 2010;20:837–46.

[24] Bringans S, Hane JK, Casey T, Tan KC, Lipscombe R, Solomon PS, et al. Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen *Stagonospora nodorum*. BMC Bioinformatics 2009;10: 301.

[25] Christie-Oleza JA, Pina-Villalonga JM, Bosch R, Nogales B, Armengaud J. Comparative proteogenomics of twelve *Roseobacter* exoproteomes reveals different adaptive strategies amongst these marine bacteria. Mol Cell Proteomics 2011;11(2).

[26] Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R, et al. Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proc Natl Acad Sci U S A 2009;106:16428–33.

[27] Helmy M, Tomita M, Ishihama Y. OryzaPG-DB: rice proteome database based on shotgun proteogenomics. BMC Plant Biol 2011;11:63.

[28] Renuse S, Chaerkady R, Pandey A. Proteogenomics. Proteomics 2011;11:620–30.

[29] Sarwal MM, Sigdel TK, Salomon DR. Functional proteogenomics—embracing complexity. Semin Immunol 2011;23:235–51.

[30] Vergara D, Tinelli A, Martignago R, Malvasi A, Chiuri VE, Leo G. Biomolecular pathogenesis of borderline ovarian tumors: focusing target discovery through proteogenomics. Curr Cancer Drug Targets 2010;10:107–16.

[31] Krug K, Nahnsen S, Macek B. Mass spectrometry at the interface of proteomics and genomics. Mol Biosyst 2011;7: 284–91.

[32] Scigelova M, Hornshaw M, Giannakopulos A, Makarov A. Fourier transform mass spectrometry. Mol Cell Proteomics 2011:10.

[33] Plumb R, Castro-Perez J, Granger J, Beattie I, Joncour K, Wright A. Ultra-performance liquid chromatography coupled to quadrupole-orthogonal time-of-flight mass spectrometry. Rapid Commun Mass Spectrom 2004;18:2331–7.

[34] Sandra K, Moshir M, D'Hondt F, Verleysen K, Kas K, Sandra P. Highly efficient peptide separations in proteomics Part 1. Unidimensional high performance liquid chromatography. J Chromatogr B Analyt Technol Biomed Life Sci 2008;866:48–63.

[35] Sandra K, Moshir M, D'Hondt F, Tuytten R, Verleysen K, Kas K, et al. Highly efficient peptide separations in proteomics. Part 2: bi- and multidimensional liquid-based separation techniques. J Chromatogr B Analyt Technol Biomed Life Sci 2009;877:1019–39.

[36] Kocher 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 2012;7:882–90.

[37] Iwasaki M, Miwa S, Ikegami T, Tomita M, Tanaka N, Ishihama Y. One-dimensional capillary liquid chromatographic separation coupled with tandem mass spectrometry unveils the *Escherichia coli* proteome on a microarray scale. Anal Chem 2010;82:2616–20.

[38] Rockstroh M, Müller S, Jende C, Kerzhner A, von Bergen M, Tomm JM. Cell fractionation—an important tool for compartment proteomics; 2010.

[39] Lee YH, Tan HT, Chung MCM. Subcellular fractionation methods and strategies for proteomics. Proteomics 2010;10:3935–56.

[40] Doucette AA, Tran JC, Wall MJ, Fitzsimmons S. Intact proteome fractionation strategies compatible with mass spectrometry. Expert Rev Proteomics 2011;8:787–800.

[41] Tran JC, Doucette AA. Gel-eluted liquid fraction entrapment electrophoresis: an electrophoretic method for broad molecular weight range proteome separation. Anal Chem 2008;80:1568–73.

[42] Müller SA, Kohajda T, Findeiss S, Stadler PF, Washietl S, Kellis M, et al. Optimization of parameters for coverage of low molecular weight proteins. Anal Bioanal Chem 2010;398: 2867–81.

[43] Swaney DL, Wenger CD, Coon JJ. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J Proteome Res 2010;9:1323–9.

[44] Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, et al. The complete genome and proteome of *Mycoplasma* mobile. Genome Res 2004;14:1447–61.

[45] Nagaraj N, Kulak NA, Cox J, Neuhauser N, Mayr K, Hoerning O, et al. System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top orbitrap. Mol Cell Proteomics 2012:11.

[46] Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, et al. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol 2011:7.

[47] Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, et al. Annotating genomes with massive-scale RNA sequencing. Genome Biol 2008:9.

[48] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–8.

[49] Chan PP, Holmes AD, Smith AM, Tran D, Lowe TM. The UCSC Archaeal Genome Browser: 2012 update. Nucleic Acids Res 2012;40:D646–52.

[50] Müller SA, van der Smissen A, von Feilitzsch M, Anderegg U, Kalkhof S, von Bergen M. Quantitative proteomics reveals altered expression of extracellular matrix related proteins of human primary dermal fibroblasts in response to sulfated hyaluronan and collagen applied as artificial extracellular matrix. J Mater Sci Mater Med 2012;23:3053–65.

[51] Jungblut PR, Schiele F, Zimny-Arndt U, Ackermann R, Schmid M, Lange S, et al. *Helicobacter pylori* proteomics by 2-DE/MS, 1-DE-LC/MS and functional data mining. Proteomics 2010;10: 182–93.

[52] Käll L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res 2008;7:40–4.

[53] Schneider KL, Pollard KS, Baertsch R, Pohl A, Lowe TM. The UCSC Archaeal Genome Browser. Nucleic Acids Res 2006;34:D407–10.

[54] Paetzel M, Karla A, Strynadka NCJ, Dalbey RE. Signal peptidases. Chem Rev 2002;102:4549–79.

[55] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–32.

[56] Hiller K, Grote A, Scheer M, Munch R, Jahn D. PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res 2004;32:W375–9.

[57] Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 2011;8:785–6.

[58] Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 2004;14:708–15.

[59] Vizcaíno JA, Côté R, Reisinger F, Barsnes H, Foster JM, Rameseder J, et al. The Proteomics Identifications database: 2010 update. Nucleic Acids Res 2010;38:D736–42.

[60] Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L. SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X! Tandem searches. Proteomics 2011;11:996–9.

[61] Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 1997;388:539–47.

[62] Borodovsky M, Mills R, Besemer J, Lomsadze A. Prokaryotic gene prediction using GeneMark and GeneMark.hmm.In: Andreas D, Baxevanis, et al, editors. Current protocols in bioinformatics/editoral board; 2003 [Chapter 4:Unit4 5].

[63] Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 1998;26:1107–15.

[64] Specht M, Schatzle S, Graumann PL, Waidner B. *Helicobacter pylori* possesses four coiled-coil-rich proteins that form

extended filamentous structures and control cell shape and motility. J Bacteriol 2011;193:4523–30.

[65] Oleastro M, Monteiro L, Lehours P, Megraud F, Menard A. Identification of markers for *Helicobacter pylori* strains isolated from children with peptic ulcer disease by suppressive subtractive hybridization. Infect Immun 2006;74:4064–74.

[66] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14:1188–90.

[67] Malfertheiner P, Megraud F, O'Morain CA, Atherton J, Axon ATR, Bazzoli F, et al. Management of *Helicobacter pylori* infection—the Maastricht IV/Florence Consensus Report. Gut 2012;61:646–64.

[68] Suzuki R, Shiota S, Yamaoka Y. Molecular epidemiology, population genetics, and pathogenic role of *Helicobacter pylori*. Infect Genet Evol 2012;12:203–13.

[69] Shao C, Zhang Q, Tang W, Qu W, Zhou Y, Sun Y, et al. The changes of proteomes components of *Helicobacter pylori* in response to acid stress without urea. J Microbiol 2008;46: 331–7.

[70] Zeng H, Guo G, Mao XH, Tong WD, Zou QM. Proteomic insights into *Helicobacter pylori* coccoid forms under oxidative stress. Curr Microbiol 2008;57:281–6.

[71] Chuang MH, Wu MS, Lin JT, Chiou SH. Proteomic analysis of proteins expressed by *Helicobacter pylori* under oxidative stress. Proteomics 2005;5:3895–901.

[72] Akada JK, Aoki H, Torigoe Y, Kitagawa T, Kurazono H, Hoshida H, et al. *Helicobacter pylori* CagA inhibits endocytosis of cytotoxin VacA in host cells. Dis Model Mech 2010;3:605–17.

[73] Lahner E, Bernardini G, Santucci A, Annibale B. *Helicobacter pylori* immunoproteomics in gastric cancer and gastritis of the carcinoma phenotype. Expert Rev Proteomics 2010;7: 239–48.

[74] Kim W, Silby MW, Purvine SO, Nicoll JS, Hixson KK, Monroe M, et al. Proteomic detection of non-annotated protein-coding genes in *Pseudomonas fluorescens* Pf0-1. PLoS One 2009;4:e8455.

[75] Tsugawa H, Suzuki H, Matsuzaki J, Hirata K, Hibi T. FecA1, a bacterial iron transporter, determines the survival of *Helicobacter pylori* in the stomach. Free Radic Biol Med 2012;52: 1003–10.

[76] Kusters JG, van Vliet AHM, Kuipers EJ. Pathogenesis of *Helicobacter pylori* infection. Clin Microbiol Rev 2006;19:449–90.

[77] Velayudhan J, Hughes NJ, McColm AA, Bagshaw J, Clayton CL, Andrews SC, et al. Iron acquisition and virulence in *Helicobacter pylori*: a major role for FeoB, a high-affinity ferrous iron transporter. Mol Microbiol 2000;37:274–86.

[78] Szczebara F, Dhaenens L, Armand S, Husson MO. Regulation of the transcription of genes encoding different virulence factors in *Helicobacter pylori* by free iron. FEMS Microbiol Lett 1999;175:165–70.

[79] Sarkar M, Maganti L, Ghoshal N, Dutta C. In silico quest for putative drug targets in *Helicobacter pylori* HPAG1: molecular modeling of candidate enzymes from lipopolysaccharide biosynthesis pathway. J Mol Model 2012;18:1855–66.

[80] Paetzel M, Dalbey RE, Strynadka NCJ. The structure and mechanism of bacterial type I signal peptidases—a novel antibiotic target. Pharmacol Ther 2000;87:27–49.

[81] Ollinger J, O'Malley T, Ahn J, Odingo J, Parish T. Inhibition of the sole type I signal peptidase of *Mycobacterium tuberculosis* is bactericidal under replicating and nonreplicating conditions. J Bacteriol 2012;194:2614–9.

[82] Schallenberger MA, Niessen S, Shao C, Fowler BJ, Romesberg FE. Type I signal peptidase and protein secretion in *Staphylococcus aureus*. J Bacteriol 2012;194:2677–86.

[83] Vizcaíno JA, Côté RG, Csordas A, Dianes JA, Fabregat A, Foster JM, et al. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res 2013;41: D1063–9.

3.5 Quantitative proteomics reveals altered expression of extracellular matrix related proteins of human primary dermal fibroblasts in response to sulfated hyaluronan and collagen applied as artificial extracellular matrix

Stephan A. Müller, Anja van der Smissen, Margarete von Feilitzsch, Ulf Anderegg, Stefan Kalkhof, Martin von Bergen.

J Mater Sci Mater Med. 2012 Dec; 23(12):3053-65.
DOI: 10.1007/s10856-012-4760-x

Abstract

Fibroblasts are the main matrix producing cells of the dermis and are also strongly regulated by their matrix environment which can be used to improve and guide skin wound healing processes. Here, we systematically investigated the molecular effects on primary dermal fibroblasts in response to high-sulfated hyaluronan [HA] (hsHA) by quantitative proteomics. The comparison of non- and highsulfated HA revealed regulation of 84 of more than 1,200 quantified proteins. Based on gene enrichment we found that sulfation of HA alters extracellular matrix remodeling. The collagen degrading enzymes cathepsin K, matrix metalloproteinases-2 and -14 were found to be down-regulated on hsHA. Additionally protein expression of thrombospondin-1, decorin, collagen types I and XII were reduced, whereas the expression of trophoblast glycoprotein and collagen type VI were slightly increased. This study demonstrates that global proteomics provides a valuable tool for revealing proteins involved in molecular effects of growth substrates for further material optimization.

Keywords

Hyaluronan; extracellular matrix; proteomics; sulfation; glycosaminoglycan

J Mater Sci: Mater Med (2012) 23:3053–3065
DOI 10.1007/s10856-012-4760-x

Quantitative proteomics reveals altered expression of extracellular matrix related proteins of human primary dermal fibroblasts in response to sulfated hyaluronan and collagen applied as artificial extracellular matrix

Stephan A. Müller · Anja van der Smissen · Margarete von Feilitzsch · Ulf Anderegg · Stefan Kalkhof · Martin von Bergen

Received: 31 May 2012 / Accepted: 27 August 2012 / Published online: 19 September 2012
© The Author(s) 2012. This article is published with open access at Springerlink.com

Abstract Fibroblasts are the main matrix producing cells of the dermis and are also strongly regulated by their matrix environment which can be used to improve and guide skin wound healing processes. Here, we systematically investigated the molecular effects on primary dermal fibroblasts in response to high-sulfated hyaluronan [HA] (hsHA) by quantitative proteomics. The comparison of non- and high-sulfated HA revealed regulation of 84 of more than 1,200 quantified proteins. Based on gene enrichment we found that sulfation of HA alters extracellular matrix remodeling. The collagen degrading enzymes cathepsin K, matrix metalloproteinases-2 and -14 were found to be down-regulated on hsHA. Additionally protein expression of thrombospondin-1, decorin, collagen types I and XII were reduced, whereas the expression of trophoblast glycoprotein and collagen type VI were slightly increased. This study demonstrates that global proteomics provides a valuable tool for revealing proteins involved in molecular effects of growth substrates for further material optimization.

1 Introduction

The skin is the largest organ of the human body. It has many essential functions like body temperature regulation, oxygen uptake, pathogen defense and fluid loss prevention. Thus dermal wounds can cause severe health problems by the restriction of these functions. The therapeutic band width of skin wound treatment includes dressing with autografts, allografts, xenografts or tissue-engineered skin substitutes (TESS). TESS have been proven to be a good alternative to conventional treatment by grafting of skin wounds [1]. Clinical products from different companies are extensively reviewed by Eisenbud et al. [2] and Damanhuri et al. [3], while Metcalfe and Ferguson [4] have reviewed developments of bioengineered artificial skin. The usage of cell-free scaffolds as matrix supports for self-regeneration of skin is an alternative to skin biopsies and dermal cell culturing. Especially cell-free scaffolds based on biodegradable substances like polylactides, collagens and/or glycosaminoglycans (GAGs) which mimic the extracellular matrix (ECM) are good alternatives to conventional skin grafting [5–8].

A promising approach for the development of new artificial ECMs (aECMs) for wound healing of skin tissue

Electronic supplementary material The online version of this article (doi:10.1007/s10856-012-4760-x) contains supplementary material, which is available to authorized users.

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

S. A. Müller · A. van der Smissen · M. von Feilitzsch ·
U. Anderegg · S. Kalkhof · M. von Bergen
Collaborative Research Center (SFB-TR67),
Matrixengineering, Leipzig, Germany

A. van der Smissen · M. von Feilitzsch · U. Anderegg
Department of Dermatology Venerology and Allergology,
Leipzig University, 04103 Leipzig, Germany

M. von Bergen
Department of Metabolomics, UFZ, Helmholtz-Centre for
Environmental Research Leipzig, 04318 Leipzig,
Germany

is the integration of chemically modified natural ECM components. In particular sulfated GAG have been supposed to improve wound healing of skin tissue by the interaction of negatively charged sulfate groups with cytokines, growth factors and dermal cells [9, 10].

Sulfated derivatives of GAGs mimic the behavior of heparin, the most biological active natural GAG compound which plays an important role in wound healing [11]. Heparin interacts with a huge variety of different proteins, like growth factors FGFs (fibroblast growth factors)-1, -2 and -7 [12] or cytokines such as platelet factor 4 [13], interleukin 8 (IL-8) [10, 14] or interferon gamma [15]. Heparin further binds to adhesion proteins like selectins [16], the heparin-binding growth associated molecule [13] and fibronectin [17]. Protein binding to heparin promotes different functions like protection from proteolysis (i.e. FGFs-1, -2 and -7) [12, 18] or modification of biological activity shown for transforming growth factor β1 (TGF-β1) [13]. Thus heparin and other sulfated GAG have an influence on key processes of wound healing like inflammation, cell proliferation or cell–matrix interactions [13]. Most interactions between sulfated GAG and proteins are governed by negatively charged sulfate groups which form ionic bonds with basic amino acid residues [10, 12, 13, 15].

Hence, cell studies with sulfated GAG can provide valuable information for the engineering of new skin substitutes. We have chosen hyaluronan (HA) to investigate the effect of chemical sulfation. HA is the most suited GAG for this study since naturally HA does not contain sulfate groups. It has a regular sequence of alternating units of *N*-acetylglucosamine and glucuronic acid and is not covalently linked to proteins. Additionally, HA can be chemically modified without loss of structure [11]. Since our research focus is on acquiring knowledge about the influence of synthetized aECMs for improved wound healing of skin tissue we have chosen dermal fibroblasts (dFbs) as model cells for investigation of our modified aECM. They are crucial for wound healing of skin tissue and strongly regulated by their surrounding ECM [19]. The previous work of van der Smissen et al. [20] showed that sulfated GAGs improved initial cell adhesion and proliferation of dFbs in a sulfation dependent manner. By testing a few selected mRNA of involved key proteins the expression levels of collagen type I α chain, HA synthase 2 and matrix metalloproteinase-1 (MMP-1) were found to be significantly reduced on high-sulfated GAGs, whereas low-sulfated GAG derivatives only slightly changed the mRNA expression of these components.

On the basis of these data [20], the influence of HA sulfation on the expression of other proteins by a non-targeted approach is of great interest since this will allow detecting so far unrecognized signaling pathways in response to the tested biomaterials. We analyzed the influence of aECMs consisting of collagen type I mixed with HA or its high-sulfated derivative (hsHA) on protein level. For that reason, we have chosen stable isotope labeling by amino acids in cell culture (SILAC) which is a well-established method enabling accurate relative quantification of thousands of proteins in an untargeted approach [21, 22]. As long as primary cells can be cultivated for a sufficient time to obtain quantitative isotope labeling, SILAC provides superior protein coverage and better quantitative reproducibility in comparison to the usage of cells or organs from different individuals or label free quantification [23]. Especially relative quantification to a control of the same donor within one measurement reduces variability.

Global analyses provide a broader overview and higher protein coverage than targeted experiments. Computational analyzes of regulated proteins according to databases like PANTHER (Protein ANalysis THrough Evolutionary Relationships) [24] reveal protein cluster enriched according to their molecular functions and biological processes. Bioinformatics tools like DAVID [25] additionally calculate enrichment factors and determine statistical significance of these clusters.

While these approaches are limited to detecting known pathways the global approach also offers the chance to unravel so far unknown proteins or complexes that might also be pivotal to the process of interest. In order to extract this potential from the wealth of raw data gathered through omics approaches it is necessary to build up cell type and research specific databases. More specifically the effects of aECM on different cell types involved in wound healing should be summarized in a database allowing a focused comparison with future data.

A generally important aspect of global analysis is the assumption that the conditions do not cause an overall extreme stress to the cells, since then the effects would reflect all but not dominantly specifically mechanism about the subtle changes occurring during adaptation. The general effects can be monitored by the amount of overall changes and as a valid assumption the significantly ($P < 0.05$) changed proteins should not exceed 5–10 %.

In this study over 2,000 proteins were unambiguously identified and the gene enrichment process revealed that HA sulfation affects predominantly ECM remodeling by simultaneously down-regulation of ECM degenerating proteins like MMPs-2 and -14 as well as cathepsin K (catK). Additionally, other ECM proteins including decorin, thrombospondin-1 (TSP-1), and collagen types I, VI and XII are regulated. Beside this detailed information on coordinated ECM remodeling the summary of affected pathways and molecular functions allows to build a database for monitoring of aECM caused effects on fibroblasts.

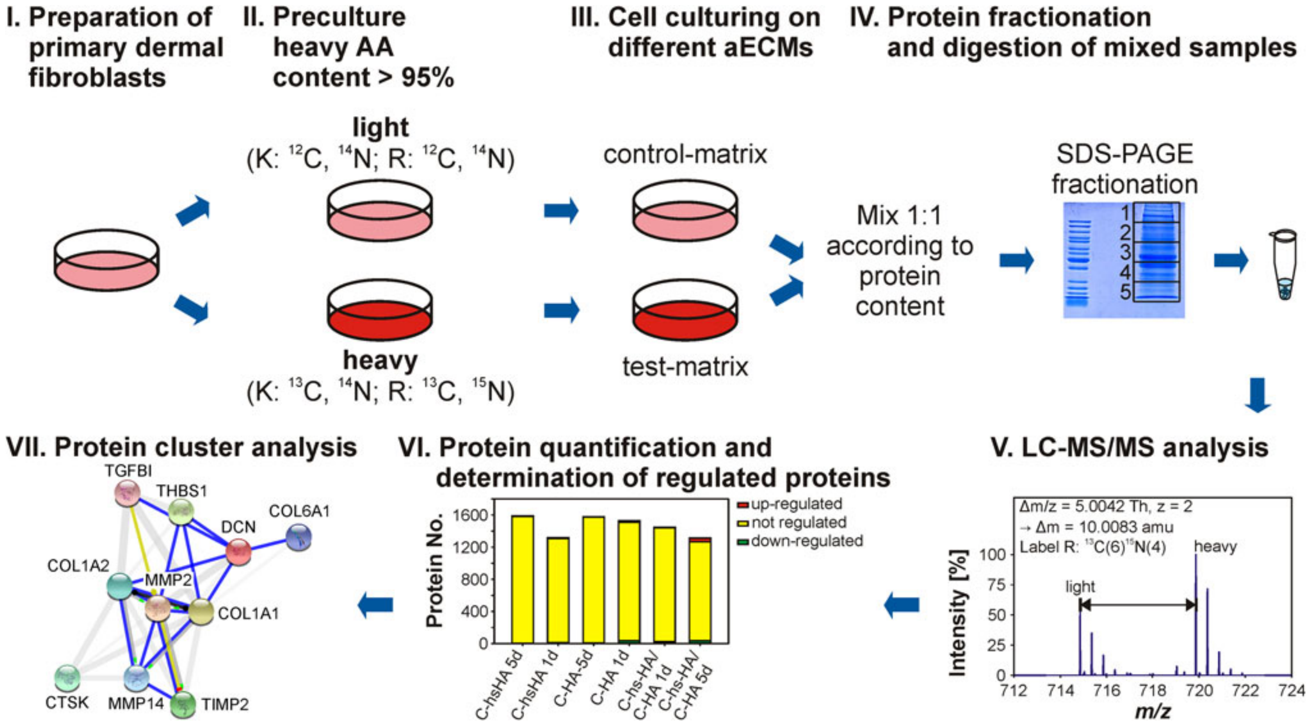

I. Preparation of primary dermal fibroblasts
II. Preculture heavy AA content > 95%
III. Cell culturing on different aECMs
IV. Protein fractionation and digestion of mixed samples
light
(K: 12C, 14N; R: 12C, 14N)
heavy
(K: 13C, 14N; R: 13C, 15N)
control-matrix
test-matrix
Mix 1:1 according to protein content
SDS-PAGE fractionation
1
2
3
4
5
V. LC-MS/MS analysis
Δm/z = 5.0042 Th, z = 2
→ Δm = 10.0083 amu
Label R: 13C(6)15N(4)
light
heavy
Intensity [%]
m/z
VI. Protein quantification and determination of regulated proteins
up-regulated
not regulated
down-regulated
Protein No.
C-hsHA 5d
C-hsHA 1d
C-HA 5d
C-HA 1d
C-hs-HA/C-HA 1d
C-hs-HA/C-HA 5d
VII. Protein cluster analysis
TGFBI
THBS1
COL6A1
DCN
COL1A2
MMP2
COL1A1
CTSK
MMP14
TIMP2

Fig. 1 Experimental workflow. **I** Primary dermal fibroblasts are prepared from healthy female donors. **II** Primary dermal fibroblasts are precultured either in light medium (L) or heavy medium (H) containing isotopically labeled lysine and arginine until heavy amino acid content is larger than 95 % in the proteins. **III** Cells are cultured on different aECMs. Control-matrix and test-matrix have different isotope labeling. **IV** After culturing for 1 respectively 5 days, cells are harvested, lysed and mixed 1:1 according to their protein content. Proteins are fractionated by SDS-PAGE and digested in-gel by trypsin. **V** Peptides are analyzed by LC–MS/MS. **VI** MS data is processed by Maxquant. Pairs of light and heavy labeled peptides enable relative protein quantification. Regulated proteins are determined. **VII** Proteins considered to be regulated are subjected to bioinformatics tools like DAVID and PANTHER for cluster analysis according to biological processes and molecular function (cluster diagram was made on http://string-db.org/)

2 Materials and methods

2.1 Sample preparation

The study was conducted according to Declaration of Helsinki Principles (1975) and was approved by the local ethics committee (065-2009).

Primary human dFbs from healthy breast skin were isolated as previously described [26] by dispase II (Roche Diagnostics GmbH, Mannheim, Germany) mediated removal of epidermal sheet and digestion of the dermal compartment with collagenase (Sigma-Aldrich Chemie GmbH, Steinheim, Germany). Cell suspension was passed through 70 µM filters (BD Biosciences, Bedford, MA, USA) to remove tissue debris. In total four biological replicates deriving from different donors were applied in this study.

Cells were cultured with Dulbecco's Modified Eagle Medium (DMEM, Biochrom AG, Berlin, Germany) supplemented with 10 % fetal calf serum (FCS, Biochrom AG, Berlin, Germany) and 1 % penicillin/streptomycin (PAA Laboratories GmbH, Pasching, Austria) at 37 °C, 5 % CO_2 until confluence. For experiments cells between passages 2–8 were used [20].

An overview of the experimental workflow after isolation of primary dFb is shown in Fig. 1. For isotope labeling dFb were cultivated in SILAC DMEM (Pierce SILAC Protein Quantitation Kit—DMEM, Pierce Biotechnology, Rockford, USA) containing either 0.798 mmol/l heavy $^{13}C^{14}N$ lysine and 0.398 mmol/l heavy $^{13}C^{15}N$ arginine (heavy medium) or $^{12}C^{14}N$ lysine and $^{12}C^{14}N$ arginine (light medium) supplemented with 10 % dialyzed FCS for 10 days on polystyrene (PS) culture plates with medium change every 2 days.

4.0×10^5 (24 h exposure) and accordingly 1.5×10^5 cells (5 days exposure) were transferred to 75 cm^2 cell culture flasks coated with different aECMs consisting of rat tail collagen type I (C) (BD Bioscience, Heidelberg, Germany) and HA (Aqua Biochem, Dessau, Germany) or hsHA (provided by Innovent e.V., Jena, Germany) described by van der Smissen et al. [20] and incubated for 1 or 5 days. At the time point 5 days the monolayer appeared with a donor dependent confluence of 70–100 %. One day incubation was meant to determine immediate cell responses to hsHA,

Table 1 aECMs, according abbreviations and the applied culture medium and incubation time

aECM (abbreviation)	SILAC labeling medium	Incubation time (days)
PS control matrix replicates 1 + 2	Light medium	1
		5
	Heavy medium	1
		5
Collagen type I (C) control matrix replicates 3 + 4	Light medium	1
		5
	Heavy medium	1
		5
Collagen type I/hyaluronan (C-HA)	Light medium	1
		5
Collagen type I/C-hsHA	Heavy medium	1
		5

whereas 5 days of incubation should reflect changes in the proteome of almost confluent grown dFbs. The appropriate culture variations are listed in Table 1. These variations offer the comparison of light and heavy labeled cells after the cultivation on the different aECMs.

Proteomic analysis was carried out on the basis of cell lysates. Therefore, fibroblasts were harvested after days 1 or 5 post seeding by addition of 0.25 % EDTA (Sigma, St. Louis, MO, USA) in 1× PBS (PAA) to prevent damage of integrins by trypsin. The cell pellet was stored on ice and washed three times with cold 1× PBS before the final centrifugation for 6 min, 12,000 rpm at 4 °C. The supernatant was discarded and cell pellet immediately frozen at −80 °C until further use.

Harvested cells were disrupted in 100 µl lysis buffer containing 6 M urea, 2 M thiourea and 100 mM ammonium bicarbonate by vortexing for 3 min. Cell debris and undissolved material were removed by centrifugation (16,000×*g*, 10 min, 18 °C). Protein concentration of the supernatants was measured using Quick Start Bradford Protein Assay (Biorad, Hercules, CA) with bovine serum albumin as reference. Samples gained from the different aECMs were combined at 1:1 (w/w) protein ratio with the appropriate control (PS or C).

2.2 SDS-PAGE and in-gel digestion of proteins

In order to increase the amount of quantified proteins, samples were fractionated using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). For the gel separation, 15 µg protein of each sample were mixed 3:1 with 4× Laemmli sample buffer (12 % [w/v] SDS, 6 % [v/v] *β*-mercaptoethanol, 30 % [w/v] glycerol, 150 mM Tris–HCl [pH 7.0], 0.04 % [w/v] bromphenol blue) and incubated 1 h at 37 °C. Protein separation was performed by 12 % SDS-PAGE with a 4 % stacking gel. Gel electrophoresis was stopped after proteins entered approximately 3 cm in the gel. The Coomassie staining procedure was performed according to Müller et al. [27].

The protein lanes were cut in five equal gel slices. In-gel digestion of protein was performed similar to Mörbt et al. [28] with 100 ng trypsin per slice (trypsin sequencing grade from bovine pancreas, Roche, Mannheim, Germany). Samples were concentrated by vacuum centrifugation and reconstituted with 0.1 % (v/v) formic acid after tryptic digestion.

2.3 Liquid chromatography tandem MS analysis

Tryptic peptides from in-gel digestion were separated by nano-high performance liquid chromatography (nano-HPLC) prior to mass spectrometry (MS) analysis to increase the number of quantified peptides and corresponding proteins. Liquid chromatography tandem MS analysis was performed according to Müller et al. [27] with some slight modifications. Peptides were analyzed with a nano-HPLC system (nanoAquity, Waters, Milford, MA, USA) coupled online with an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) via a nano-electrospray ion source (TriVersa NanoMate, Advion, Ithaca, NY, USA). Samples were injected on a trapping column (nanoAquity UPLC column, C18, 180 µm × 20 mm, 5 µm, Waters) and washed with 2 % acetonitrile containing 0.1 % formic acid and a flow rate of 15 µl/min for 8 min. A C18 UPLC column (nanoAcquity UPLC column, C18, 75 µm × 150 mm, 1.7 µm, Waters) was used for peptide separation. Peptides were eluted using a gradient from 2 to 85 % acetonitrile, 0.1 % formic acid (0 min, 2 %; 2 min, 2 %; 7 min, 6 %; 55 min, 20 %; 73 min, 30 %; 91 min, 40 %; 94 min, 85 %) with a flow rate of 300 nl/min and a column temperature of 40 °C.

MS analysis was performed with a spray voltage of 1.8 kV in positive ion mode. The mass spectrometer automatically switched between full scan MS mode (from 400 to 1,400 m/z, $R = 60{,}000$) and MS^2 acquisition. Peptide ions exceeding an intensity of 5,000 counts were fragmented within the linear ion trap by collision induced dissociation (isolation width 4 m/z, normalized collision energy 35, activation time 30 ms, activation Q 0.25). A dynamic precursor exclusion of 3 min for tandem MS measurements was applied.

2.4 Data analysis

Protein identification and relative quantification was carried out with the software MaxQuant [29] (version 1.2.0.18, Max Planck Institute of Biochemistry, Munich, Germany). Peptides with the same sequence but different labeling states elute at the same retention time. Heavy to light peptide pairs can be detected by their distinct mass shifts according to the labeling with heavy arginine and lysine. MaxQuant uses the intensity of heavy and light labeled peptide pairs to calculate relative peptide abundances. The derived peptide intensity ratios belonging to the same protein are the basis for relative protein quantification.

Within the MaxQuant workflow, database searching was carried out by the Andromeda search engine [30] against a reverse concatenated IPI human database (version 3.68) including a contaminant list. Recalibration of precursor masses by the option "first search" with a 20 ppm mass tolerance against the human first search database provided by MaxQuant.org. Trypsin with maximum two missed cleavages was set as protease. Carbamidomethylation of cysteine was specified as fixed modification, and oxidation of methionine and acetylation of the protein N-terminal were defined as variable modifications. A peptide mass tolerance of 6 ppm was applied. For tandem MS identification six top peaks per 100 Da were chosen and searched with a fragment ion mass tolerance of 0.5 Da.

Peptide and protein false discovery rates were limited to 1 %. Protein identification required at least two unique peptides. The minimal peptide length was set to six amino acids. For protein quantification, the minimal peptide ratio count was set to 2. The option "match between runs" was used for samples measured within the same batch. Re-quantification of proteins was also applied.

Proteins with a $\log_2$ fold change (FC) above 0.5 or below −0.5 were considered to be up- respectively down-regulated. Furthermore, only proteins showing in at least three out of four replicates regulation in the same direction and an average FC of all replicates fulfilling the criteria for regulation were considered as significantly regulated.

For identification of significantly regulated clusters of functionally related regulated proteins the web-based bioinformatics tool DAVID [24] was used. The list of regulated proteins was subjected to DAVID, whereas all identified proteins served as background for cluster analysis. Protein clustering was performed according to biological and molecular function derived from the PANTHER classification system [24].

2.5 Control experiments for significance estimation of regulation thresholds

Experiments with primary cells often show large variation between different donors. Additionally, technical variance is another error source. With regard to these issues, we tested the significance of our regulation thresholds with two control experiments. Each control experiment was performed in triplicates. The first experiment was to evaluate the labeling effect of the SILAC experiments. Therefore, protein samples of the same donor from cells grown on light and heavy medium with collagen type I as matrix were mixed 1:1 (w/w) according to their protein content. The second control experiment examined the donor effect and included protein samples from three donors. Therefore, heavy and light labeled protein samples of different donors were mixed 1:1 (w/w) according to their protein content (donor A heavy + donor B light, donor B heavy + donor C light, donor A light + donor B heavy). Further treatment and measurement was similar to the other samples. Proteins with a $\log_2$ FC larger than 0.5 or lower than −0.5 were defined as regulated.

2.6 Western blot and zymography

Data analysis with MaxQuant and the bioinformatics tool DAVID resulted in a set of regulated protein clusters. Selected proteins belonging to regulated clusters were chosen for further confirmation by western blotting or zymography. Western blots of cell lysates were performed with antibodies against MMP-14, TSP-1, collagen types I and VI (α chain 1). The enzymatic activity of MMP-2 in the culture supernatant was tested by gelatine zymography [26] to investigate whether altered MMP-2 expression leads to activity changes.

3.5×10^5 cells were seeded on aECM provided in petri dishes (94 mm diameter) and incubated for 72 h with DMEM/10 % FCS, another 24 h with DMEM/0 % FCS to generate serum free supernatants and additional 24 h with DMEM/10 % FCS to gain an incubation time of 5 days in total. Samples from six different donors were applied for validation by western blotting and zymography.

Cell extracts were prepared by detaching cells with 0.05 % trypsin/0.02 % EDTA (Biochrom, Berlin, Germany) and cooled lysis of cell pellets with RIPA-buffer (50 mM HEPES, 150 mM NaCl, 5 mM EDTA, 1 mM

EGTA, 1 % Triton X-100, 0.1 % SDS, 1 % deoxycholate, 1 mM dithiothreitol [Roth, Karlsruhe, Germany; Serva, Heidelberg, Germany; Sigma, Taufkirchen, Germany]). Protein lysates were separated by SDS-PAGE with appropriate SDS gels (Amersham ECL gels, GE Healthcare, München, Germany) and blotted on OPTITRAN BAS83 membrane. Primary antibodies for MMP-14 (rabbit-anti-human, clone ID: EP1264Y, Epitomics, Burlingame, USA), TSP-1 (rabbit-anti-human, Abcam, Cambridge, United Kingdom), collagen type VI α chain 1 (rabbit-anti-human, Atlas Antibodies, Stockholm, Sweden), collagen type I α 1 (rabbit-anti-human, Sigma) and GAPDH (mouse-anti-human, Merck-Millipore, Darmstadt, Germany) were combined with IRDye 680RD goat-anti-rabbit or IRDye 680RD goat-anti-rabbit (LI-COR, Lincoln, USA) as secondary antibodies.

Cell-free supernatants were concentrated by ultrafiltration using vivaspin six columns (GE Healthcare) for MMP-2 gelatine zymography [26]. An amount of 5 μg of concentrated supernatant was diluted in a sample buffer (0.3 M Tris–HCl pH 8.8, 4 % saccharose, 10 % SDS and 0.1 % bromphenol blue), applied to a 10 % SDS-gel containing, 0.1 % gelatine, and was electrophoretically separated. After electrophoresis, gels were washed in 2.5 % Triton X-100 for 30 min and were incubated overnight at room temperature in a development buffer containing 0.05 M Tris–HCl, pH 8.0, 8 mM $CaCl_2$. MMP-2 associated gelatine digestion was visualized as white bands in the gel after staining with 0.1 % Coomassie blue R250 and clearing with 7.5 % acetic acid. MMP-2 activity was quantified by densitometric measuring (Intas, Göttingen, Germany). The absolute integrated area under the peak was determined.

3 Results

3.1 Significance estimation of regulation thresholds

In order to estimate the effects of technical variance during cell culture (labeling effect) and the biological variance caused by different donors (donor effect), we set up two control experiments. Samples from three different donors were used for the significance estimation.

To investigate the labeling effect, cells from three different donors were split up and cultivated in either heavy (containing $^{13}C^{14}N$ lysine and $^{13}C^{15}N$ arginine) or light SILAC medium. Heavy and light stable isotope labeled cells of the same donor were lysed, mixed and analyzed. Analogously the donor effect was determined by mixing differentially labeled samples of the different donors. Between 600 and 900 proteins were quantified by MaxQuant. Analysis of labeling effect resulted in average 0.6 % of all identified proteins fulfilling the up-regulation threshold, whereas 5.9 % pass the threshold for down-regulation. This is clearly showing that the abundance of light labeled proteins is overestimated during protein quantification process even so typical contaminants such as keratins, trypsin as well as rat collagen type I, which was used as aECM component, were defined as contaminants and thus discarded during the quantification process.

Six proteins are found to be regulated in all three replicates with a $\log_2$ FC less than -0.5. Namely, two Ras-related proteins (RAB2, RAB5), histone H1.2, dermcidin and collagen type I α chains 1 and 2 are fulfilling the threshold in all samples. The fact that all of these proteins are showing a higher abundance of light labeled protein in this control experiment indicates that these proteins can be classified as contaminants. Dermcidin for example is a 91 amino acid long antimicrobial peptide secreted by perspiratory glands which can occur as a contaminant. Even rat collagen type I was already inserted to the contaminant list of MaxQuant, the abundance of light labeled human collagen type I is higher than the heavy labeled counterpart in this control experiment. Only unique peptides were accepted for calculation of heavy to light ratios. Which means that collagen type I contamination has to stem from another source than the applied aECM.

The donor effect was estimated by measuring a mixture of heavy and light control samples of different donors. On average 12.9 % of all identified proteins show a $\log_2$ FC less than -0.5, and 8.2 % have a FC larger than 0.5. To evaluate whether this donor effect is random or not, only proteins which had the same direction of FC in all replicates were used for further analysis. Only 0.1 % of proteins identified in all replicates fulfilled the threshold criteria and have the same direction of FC. This demonstrates that variability of protein abundance by different donors is exclusively a random effect.

To estimate the false positive rate (FPR) of regulated proteins in our SILAC experiment, we used a stochastic equation based on combinatorics. The FPR is calculated by summing up, that three out of four or four out of four measurements are representing a regulation by chance in the same direction. As probability for false positive up- or down-regulation we took the experimental values derived from the donor effect measurements as it showed the largest variability ($p = 8.2$ %, $q = 12.9$ %).

$$\mathrm{FPR} = \binom{4}{3} \cdot p^3 \cdot (1-p) + \binom{4}{4} \cdot p^4 + \binom{4}{3} \cdot q^3 \cdot (1-q) + \binom{4}{4} \cdot q^4 < 1\,\% \qquad (1)$$

With Eq. 1 a FPR lower than 1 % was calculated for the chosen protein regulation thresholds demonstrating high significance.

Table 2 Protein quantifications on C-HA and C-hsHA at 1 and 5 days post seeding

	C-HA 1 days	C-HA 5 days	C-hsHA 1 days	C-hsHA 5 days	C-hsHA/C-HA 1 days[a]	C-hsHA/C-HA 5 days[a]
Total protein quantifications	2262	2150	2244	2224	2109	1885
Proteins quantified in ≥3 replicates	1589	1318	1575	1529	1448	1213
	70 %	61 %	70 %	69 %	69 %	64 %
Down-regulated	0	13	7	36	24	38
Up-regulated	7	12	1	18	9	46
Regulated proteins (%)	0.44	1.90	0.51	3.53	2.28	6.92

[a] Values for C-hsHA/C-HA are calculated by measured ratios of C-HA and C-hsHA at corresponding time points

3.2 Classification of quantified proteins

In the main proteomic experiments cellular response to the different aECMs (C-HA and C-hsHA) after different incubation times (1 or 5 days) was investigated. Overall 2,419 proteins were quantified. Between 61 and 70 % of these proteins were quantified in at least three out of four biological replicates. Cell compartment classification of the identified proteins was done according to gene ontology (GO) annotations using the software STRAP (Software Tool for Rapid Annotation of Proteins) [31] (Supplementary Table 1). Most proteins were assigned to the cytoplasm (35 %) followed by the nucleus (34 %) and the plasma membrane (15 %). Since we found only a few regulated proteins for C-HA (1.9 %) but a more prominent effect for C-hsHA (3.6 %), we calculated the ratio between C-hsHA and C-HA at corresponding time points (Table 2). Thus we eliminated the effect of the control matrix by dividing the ratios of C-hsHA and C-HA. This is supported by the normal distribution around zero in the density plot of $\log_2$ FCs for C-hsHA related to C-HA at day 5 post seeding (Fig. 2a). The fraction of regulated protein was between 2 (C-hsHA/C-HA day 1) and 6.7 % (C-hsHA/C-HA 5 days) (Table 2).

Proteins fulfilling the regulation thresholds were clustered using the web-based tool DAVID [25] according to their molecular functions respectively their biological process using the PANTHER GO database [24]. The cluster analysis revealed one significant cluster (enrichment score >1.5) for the comparison of C-hsHA and C-HA at day 5 post seeding. Ten regulated proteins were associated to the ECM and cell adhesion (MF00178, MF00179, BP00124). Based on the low number of regulated proteins on day 1 post seeding, no significant clusters could be determined. Classification and clustering of regulated proteins according to the PANTHER database is shown in Fig. 2.

3.3 Effects of HA sulfation on the expression of ECM and cell adhesion related proteins

Bioinformatics analysis with DAVID shows regulation of 10 proteins associated with ECM (PANTHER cluster: MF00178, MF00179, BP00124) at day 5 (Fig. 2) according to HA sulfation. We manually added catK to the ECM cluster since it is an important protein for collagen degradation [32].

The regulated proteins MMP-14, collagen types I, VI and TSP-1 were chosen to confirm the SILAC results by western blotting (Fig. 3; Supplementary Table 2). MMPs-2, -14, collagen type VI and TSP-1 showed the same regulation as revealed by SILAC analysis. On the other hand, collagen type I western blots could not verify down-regulation of collagen type I on C-hsHA after 5 days of exposure.

Additionally, MMP-2 zymography was performed to measure the relative activity in the culture supernatants. Both, protein expression determined by SILAC and MMP-2 activity in the culture supernatant are diminished by HA sulfation.

4 Discussion

Previous investigations indicated that matrices with sulfated GAGs modulate cellular responses like cell adhesion, cell proliferation or matrix production [20]. In this study, we set up a SILAC experiment to extend knowledge about protein regulation caused by sulfation of HA with an untargeted approach. We focused on fibroblasts since these cells are crucial for wound closure and synthesis of new tissue.

We used primary dFb from healthy individuals in our experiments to examine effects on the proteome as close as possible to the in vivo situation. This is indispensable if the results should be referred to the original cell metabolism. For example Pan et al. [33] showed that a hepatoma cell line had up-regulated cell-cycle associated functions and down-regulation of drug metabolism compared to their cognate primary cells. Contrary to our results, Abatangelo et al. [34] reported that soluble hsHA (substitution degree 3) had no growth promoting effect on a mouse fibroblast cell line (NTC L929). This result might also be caused by the usage of an immortalized cell line. However, a clear

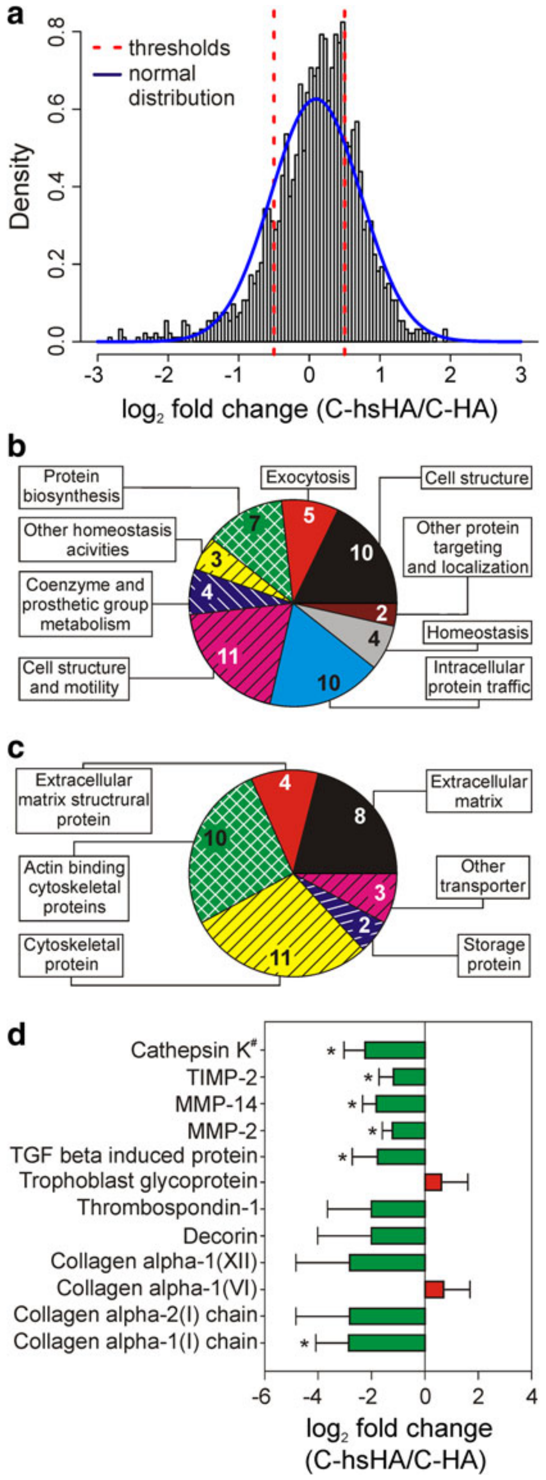


Fig. 2 Cluster analysis of proteins regulated by HA sulfation at day 5 post seeding. **a** The $\log_2$ FC between the matrices C-hsHA and C-HA is plotted against the density. FCs show a normal distribution around zero. **b** Clustering of proteins regulated by HA sulfation according to PANTHER biological processes. **c** Clustering of proteins regulated by HA sulfation according to PANTHER molecular functions. **d** FCs of proteins clustered by DAVID according to PANTHER biological processes and molecular function (*MF00178* ECM, *MF00179* ECM structural protein, *BP00124* Cell adhesion). **T* test *P* value <0.05. $^{\#}$catK was added manually to the cluster according to its collagen degrading function in the lysosomes

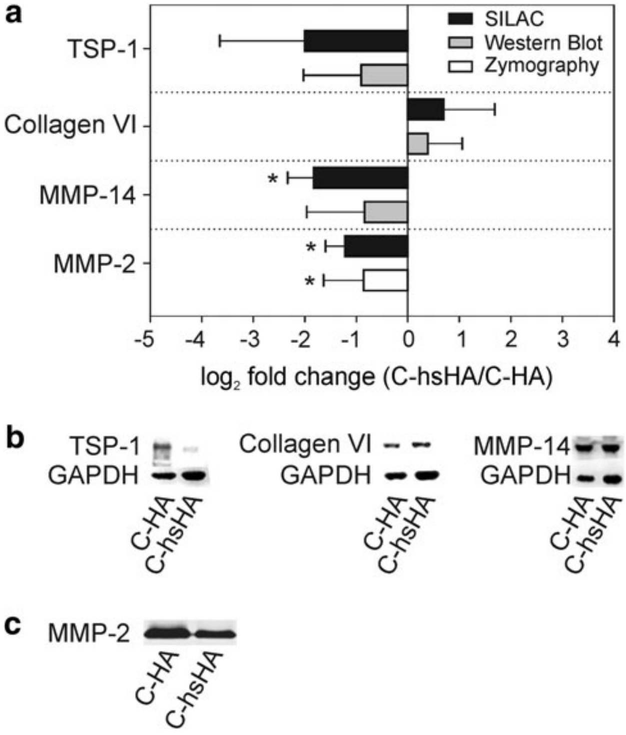


Fig. 3 Validation of selected proteins regulated by HA sulfation at day 5 post seeding by western blotting and zymography. **a** Comparison of $\log_2$ FC values derived by SILAC, western blotting and zymography. **T* test *P* value <0.05. **b** Representative western blots. **c** Representative MMP-2 zymography of culture supernatant

drawback of experiments with primary cells lies in their higher biological variance compared to cell lines.

SILAC is a well-established method to relatively quantify the abundance of proteins in a shotgun approach. It is well suited for experiments with primary cells because control and treated sample from the same donor are compared within one measurement. Therefore, the effect of the donor is minimized. Nevertheless, experiments with primary cells cause high variance of results. In order to cope with this, we applied an extended set of controls for the labeling as well as the donor effect. The results are showing that for primary dFb, SILAC can be used to investigate changes in the proteome with a FPR lower than 1 % with the applied criteria.

As expected the applied aECMs showed good biocompatibility which is in line with toxicity studies for sulfated HA [34]. Our previous results already showed that sulfation of HA increases cell adhesion and proliferation [20]. The good biocompatibility is reflected by the fact that there were no significantly regulated clusters detectable after

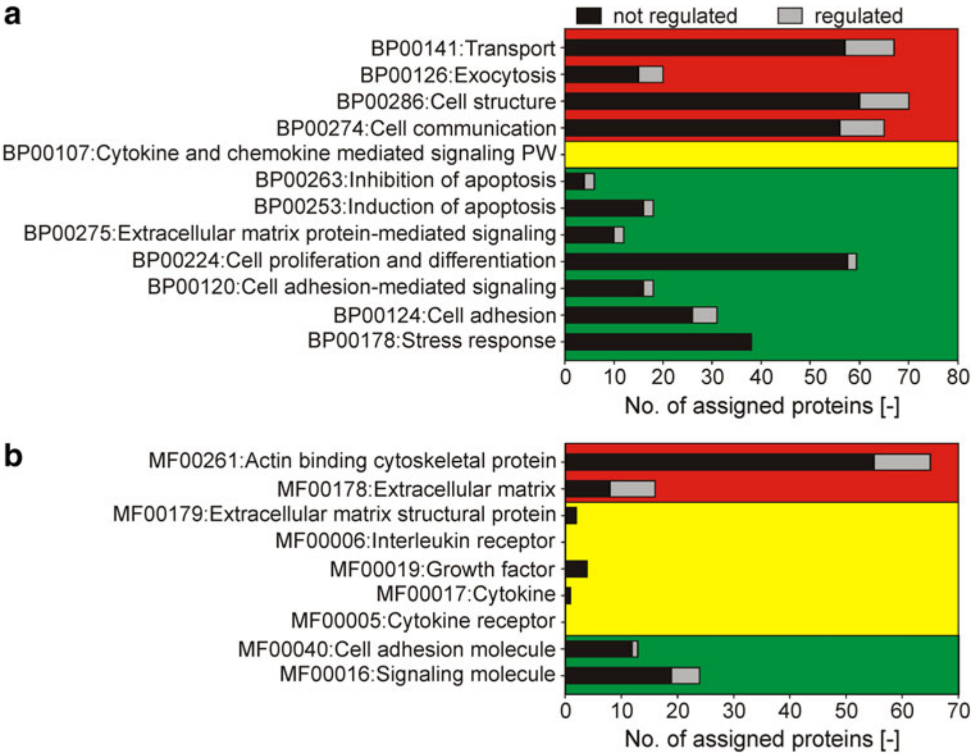


Fig. 4 Comparison of regulated and non-regulated relevant protein clusters by HA sulfation according to PANTHER **a** biological processes and **b** molecular function at day 5 post seeding. The *bars* indicate the number of identified proteins which were regulated (*gray*) or not regulated (*black*) for each protein cluster. The two *graphs* are divided in three *boxes*. Regulated protein clusters are in the *upper boxes* (*red*). Protein clusters with less than five proteins are in the *middle boxes* (*yellow*). Protein clusters which are not regulated and include more than five protein identifications are in the *lower boxes* (*green*) (Color figure online)

24 h of culture. In order to allow future work to focus on the really relevant pathways and molecular functions in terms of effects of aECM, we summarized the results of gene enrichment analyses in Fig. 4. The figure shows relevant protein clusters with information about significant enrichment of regulated proteins. The diagram highlights that neither apoptosis nor stress response are regulated by HA sulfation. Thus any cell activation or danger programs are excluded for the application of C-hsHA. However, proteins in those relevant clusters could be selected and used for fast and reliable detection with targeted approaches like selected reaction monitoring [35], western blotting or enzyme linked immunosorbent assay [36].

We focused on the significant clusters to show relevant effects caused by HA sulfation. The gene enrichment analyzes resulted in a clear enrichment in terms of cell adhesion and regulation of the ECM. The biochemical relationship between members of this cluster are shown in Fig. 5a.

Collagen type I, which is the main type in the dermal ECM [37], is the first member of the regulated protein cluster (Fig. 2). Collagen type I gives tensile strength to skin and bone tissue [38]. It replaces destroyed dermal tissue and is deposited mainly by myofibroblasts upon stimulation by TGF-β [39]. Western blotting could not confirm decreased collagen type I expression in this study. Nevertheless, the previous study of van der Smissen et al. [20] support the results derived by SILAC.

Collagen type XII protein expression is also reduced in response to hsHA. It is localized at the surface of collagen fibrils and acts as a bridge between them [40]. Increased expression of collagen type XII by dFbs is known to promote collagen type I gel contraction [41]. Thereby deformability is decreased and migration of dFb into the ECM is inhibited [40]. dFbs produce more collagen type XII when they grow on attached compared to floating collagen type I gels [42], but the underlying mechanism is not discovered by now.

On the other hand, cells on C-hsHA express higher levels of collagen type VI. This ECM compound is known to be produced by dFb when they get confluent to generate an appropriate cell environment [43].

TSP-1 is also down-regulated for C-hsHA. The expression of TSP-1 is increased in response to tissue damage, inflammation, or growth factors like platelet derived growth factor, TGF-β and basic FGF [44, 45]. Freshly synthetized TSP-1 gets integrated in the ECM or binds to the cell surface, where it is quickly internalized and degraded [46]. TSP-1 has the ability to activate TGF-β and to inhibit angiogenesis [45, 47, 48]. It is also known to influence adhesion, migration, cytoskeletal organization and apoptosis of cells by interaction with different cell receptors [45]. Thereby the mode of TSP-1 action strongly depends on the cell type and its cell surface receptors. For example smooth muscle cell migration is induced [45], while essential signal cascades like the extracellular signal-

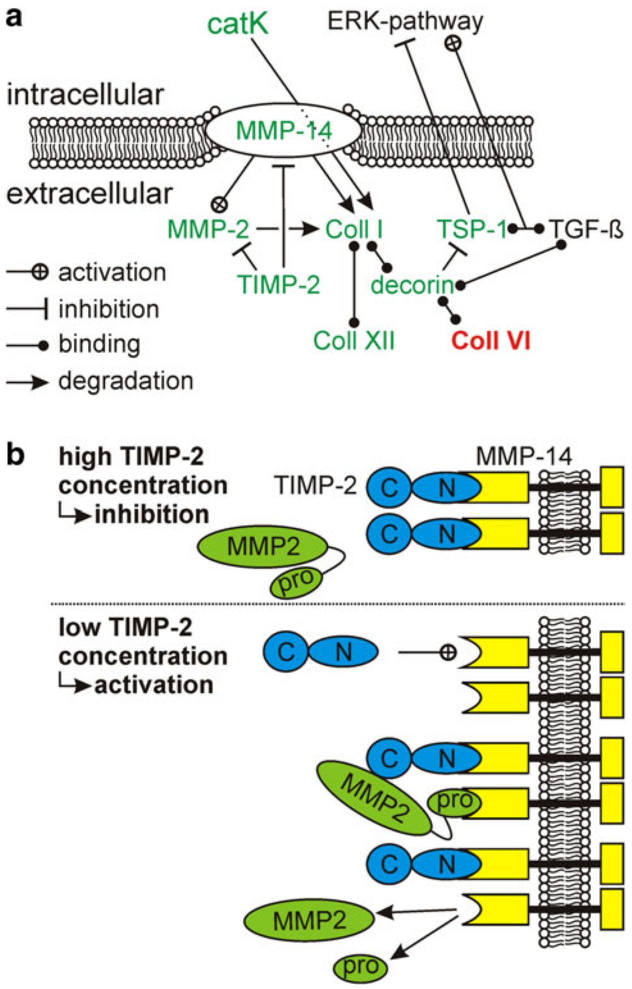


Fig. 5 Biological processes and the relationship between proteins in the regulated ECM associated cluster. **a** Scheme of protein relationships in a biochemical context. The connection *lines* between the different proteins indicate activation, inhibition, binding, or degradation of associated proteins or pathways. Proteins, which were found to be down-regulated on C-hsHA are *green*, whereas up-regulated proteins are marked *bold red*. **b** Influence of TIMP-2 on activation of MMP-2 according to the proposed mechanism by Nagase et al. [61]. High concentrations of TIMP-2 inhibit proMMP-2 conversion by blocking the active site of MMP-14. On the other hand, low concentrations of TIMP-2 are required for MMP-2 activation. TIMP-2 binds to MMP-14 with its N-terminal domain. In a second step proMMP-2 is recruited by MMP-14 bound TIMP-2. Closely located free MMP-14 binds proMMP-2 and cleaves the propeptide to activate MMP-2 (Color figure online)

regulated kinase (ERK) pathway are inhibited by TSP-1 [49]. The ERK pathway includes a phosphorylation cascade of different proteins in response to growth factors, cytokines or hormones. It controls different cell functions like cell proliferation, differentiation and apoptosis [50]. Aberrant activation of the ERK pathway is present in many cancers [51].

In our study, TSP-1 abundance was lower for the C-hsHA matrix which had pro-proliferative properties on dFb in a previous study [20]. Hence, TSP-1 can promote proliferation when TGF-β is bound whereas unbound TSP-1 reduces proliferation by inhibition of the ERK pathway (Fig. 5a). C-hsHA strongly binds TGF-β [52] and thereby prevents TGF-β-signaling in fibroblasts grown on C-hsHA (Anderegg U, personal communication). Therefore, TSP-1 might be less effective on C-hsHA in addition to its decreased expression observed here.

Additionally, decorin was found to be down-regulated for C-hsHA. This proteoglycan with attached chondroitin and dermatan sulfate chains interacts with many proteins of the regulated ECM cluster. Two different binding sites related to collagen fibrils enable decorin to bridge collagen types I and VI [53] (Fig. 5a). Decorin has also the ability to bind to collagen type XII [54]. It is essential for ECM cross-linking since decorin deficient mice produce abnormally fused collagen bundles which lead to increased skin fragility [55]. On the other hand, cell attachment to TSP-1 is inhibited by decorin through binding to its cell adhesive site [56] (Fig. 5a). Decorin is also important for binding different growth factors like FGF-2 with its sulfated GAG chains [57].

The MMPs-2 and -14 (also named MT1-MMP) and their inhibitor tissue inhibitor of metalloproteinases 2 (TIMP-2) build a complex regulation network for collagen degradation during wound healing. Besides collagen type I, membrane bound MMP-14 has a huge variety of different substrates including laminin, lumican, integrin αV, transglutaminase, CD44H, syndecan 1 and IL-8 [58]. Collagen fibers are degraded by MMP-14 in short fragments which are further degraded intracellular by phagocytosis involving catK [32, 59].

MMP-2 is secreted in its inactive form proMMP-2 and gets activated by MMP-14 (Fig. 5b) [60, 61]. Lee et al. [59] showed that MMP-14 but not MMP-2 is necessary for phagocytosis of collagen type I. Indeed, MMP-2 is able to cleave interstitial but not helical collagen type I [62]. Thus MMP-14 is the key enzyme for collagen phagocytosis. TIMP-2 is an inhibitor of both MMPs-2 and -14. Interestingly, activation of proMMP-2 by MMP-14 is enhanced by a low amount of TIMP-2, whereas higher concentrations lead to inhibition of MMP-14 [63] (Fig. 5b). Additionally, blocking of TIMP-2 by an antibody abrogates MMP-2 activation [63, 64]. Moreover, HA has also the ability to induce proMMP-2 activation [65]. Sulfated HA might not have the ability to induce proMMP-2 activation, which results in lower abundance of active MMP-2 for cells grown on C-hsHA.

CatK is also related to ECM degradation processes due to its ability to degrade collagens, elastins and proteoglycans [66]. Collagens are degraded after endocytosis in the lysosomes where catK is highly expressed [32]. CatK is usually not expressed in healthy skin, while its expression is induced by inflammation or in scar formation [32, 66]. For example it is up-regulated in synovial fibroblasts, which are key players in rheumatic arthritis because of their cartilage degrading activity [66]. In our experiment MMPs-2, -14, TIMP-2 and catK are down-regulated when comparing the aECMs C-hsHA and C-HA. Furthermore previous results showed, that MMP-1 is significantly down-regulated on mRNA level for C-hsHA [20] suggesting altogether that matrix remodeling is diminished by hsHA. This hypothesis is strengthened by the down-regulation of collagen types I and XII expression. Cells growing on non-sulfated matrix might degrade the provided aECM and build up their own matrix according to their requirements.

Interestingly, therapeutic wound dressings which result in an reduced ECM degradation or direct inactivation of MMPs are known to improve healing of chronic skin wounds since disorders in the MMP–TIMP balance can lead to fibrosis, metastasis or tumor growth [37]. There are several clinical products on the market, which target MMPs to rebalance the wound environment and to improve healing of chronic wounds. Promogran® for example consists of oxidized regenerated cellulose and collagen which binds and inactivates MMPs [67]. The product Fibracol® also reduces the activity of MMPs by competitive inhibition with collagen [68]. A formulation of metal ions and citric acid is used in DerMax® wound dressings to reduce oxygen free radicals and MMP-2 activity [69, 70].

In conclusion, introduction of sulfate groups in HA of growth substrates influences the expression of MMPs and other ECM related proteins which are involved in ECM remodeling by dFbs. These effects occur without induction of stress, promising good biocompatibility of hsHA. Especially, considering the described positive effects on healing of chronic wounds by inhibition of MMPs along with increased proliferation [20] and the low cellular stress level further encourages the application of hsHA as an appropriate therapeutic agent in wound dressings.

Our study shows that quantitative proteomics is a valuable tool for unbiased evaluation of aECM effects. It can be used to preselect suited aECM prior to animal testing. Moreover, the untargeted protein analysis provides a set of biological markers and pathways for further detailed investigations. Thereby animal experiments can be reduced to promising aECMs for clinical application. Nevertheless, in vitro experiments cannot completely simulate the situation in vivo. Ultimately further investigations of aECMs in animal experiments are indispensable to proof their influence on wound healing and long term effects.

Acknowledgments This work was funded by the Deutsche Forschungsgemeinschaft (SFB/TR67). It was also kindly supported by Helmholtz Impulse and Networking Fund through Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE). We would like to thank Matthias Schnabelrauch and Stephanie Moeller from Innovent e.V., Jena for the synthesis of sulfated HA. Furthermore, we would like to thank Vera Hintze and Dieter Scharnweber from the Max Bergmann Center for Biomaterials, Technical University Dresden for providing the aECM coated PS plates.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

1. Dieckmann C, Renner R, Milkova L, Simon JC. Regenerative medicine in dermatology: biomaterials, tissue engineering, stem cells, gene transfer and beyond. Exp Dermatol. 2010;19(8):697–706. doi:10.1111/j.1600-0625.2010.01087.x.
2. Eisenbud D, Huang NF, Luke S, Silberklang M. Skin substitutes and wound healing: current status and challenges. Wounds Compend Clin Res Pract. 2004;16(1):2–17.
3. Damanhuri M, Boyle J, Enoch S. Advances in tissue-engineered skin substitutes. Wounds Int. 2011;2(1):27–34.
4. Metcalfe AD, Ferguson MWJ. Bioengineering skin using mechanisms of regeneration and repair. Biomaterials. 2007;28(34):5100–13. doi:10.1016/j.biomaterials.2007.07.031.
5. Gravante G, Delogu D, Giordan N, Morano G, Montone A, Esposito G. The use of hyalomatrix PA in the treatment of deep partial-thickness burns. J Burn Care Res. 2007;28(2):269–74. doi:10.1097/bcr.0b013e318031a236.
6. Wainwright DJ. Use of an acellular allograft dermal matrix (alloderm) in the management of full-thickness burns. Burns. 1995;21(4):243–8. doi:10.1016/0305-4179(95)93866-i.
7. Kumbar SG, Nukavarapu SP, James R, Nair LS, Laurencin CT. Electrospun poly(lactic acid-co-glycolic acid) scaffolds for skin tissue engineering. Biomaterials. 2008;29(30):4100–7. doi:10.1016/j.biomaterials.2008.06.028.
8. Brown-Etris M, Cutshall WD, Hiles MC. A new biomaterial derived from small intestine submucosa and developed into a wound matrix device. Wounds Compend Clin Res Pract. 2002;14(4):150–66.
9. Hintze V, Miron A, Moeller S, Schnabelrauch M, Wiesmann HP, Worch H, et al. Sulfated hyaluronan and chondroitin sulfate derivatives interact differently with human transforming growth factor-β1 (TGF-β1). Acta Biomater. 2012;8(6):2144–52. doi:10.1016/j.actbio.2012.03.021.
10. Pichert A, Samsonov SA, Theisgen S, Thomas L, Baumann L, Schiller J, et al. Characterization of the interaction of interleukin-8 with hyaluronan, chondroitin sulfate, dermatan sulfate and their sulfated derivatives by spectroscopy and molecular modeling. Glycobiology. 2012;22(1):134–45. doi:10.1093/glycob/cwr120.
11. Barbucci R, Benvenuti M, Casolaro M, Lamponi S, Magnani A. Sulfated hyaluronic-acid as heparin-like material: physicochemical and biological characterization. J Mater Sci Mater Med. 1994;5(11):830–3. doi:10.1007/bf00213143.
12. Ye S, Luo Y, Lu W, Jones RB, Linhardt RJ, Capila I, et al. Structural basis for interaction of FGF-1, FGF-2, and FGF-7 with different heparan sulfate motifs. Biochemistry. 2001;40(48):14429–39.
13. Capila I, Linhardt RJ. Heparin-protein interactions. Angew Chem Int Ed Engl. 2002;41(3):391–412.

14. Ramdin L, Perks B, Sheron N, Shute JK. Regulation of interleukin-8 binding and function by heparin and alpha 2-macroglobulin. Clin Exp Allergy. 1998;28(5):616–24.
15. Fernandez-Botran R, Romanovskis P, Sun X, Spatola AF. Linear basic peptides for targeting interferon-γ-glycosaminoglycan interactions: synthesis and inhibitory properties. J Pept Res. 2004; 63(2):56–62. doi:10.1111/j.1399-3011.2003.00107.x.
16. Simonis D, Christ K, Alban S, Bendas G. Affinity and kinetics of different heparins binding to P- and L-selectin. Semin Thromb Hemost. 2007;33(5):534–9. doi:10.1055/s-2007-982085.
17. Calaycay J, Pande H, Lee T, Borsi L, Siri A, Shively JE, et al. Primary structure of a DNA- and heparin-binding domain (Domain III) in human plasma fibronectin. J Biol Chem. 1985; 260(22):12136–41.
18. Luo Y, Lu W, Mohamedali KA, Jang JH, Jones RB, Gabriel JL, et al. The glycine box: a determinant of specificity for fibroblast growth factor. Biochemistry. 1998;37(47):16506–15. doi:10.1021/bi9816599.
19. Nolte SV, Xu W, Rennekampff HO, Rodemann HP. Diversity of fibroblasts—a review on implications for skin tissue engineering. Cells Tissues Organs. 2008;187(3):165–76.
20. van der Smissen A, Hintze V, Scharnweber D, Moeller S, Schnabelrauch M, Majok A, et al. Growth promoting substrates for human dermal fibroblasts provided by artificial extracellular matrices composed of collagen I and sulfated glycosaminoglycans. Biomaterials. 2011;32(34):8938–46.
21. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002;1(5):376–86.
22. Ong SE, Mann M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). Nat Protoc. 2006;1(6): 2650–60. doi:10.1038/nprot.2006.427.
23. Greco TM, Seeholzer SH, Mak A, Spruce L, Ischiropoulos H. Quantitative mass spectrometry-based proteomics reveals the dynamic range of primary mouse astrocyte protein secretion. J Proteome Res. 2010;9(5):2764–74. doi:10.1021/pr100134n.
24. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res. 2003;31(1):334–41.
25. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57. doi:10.1038/nprot.2008.211.
26. Saalbach A, Klein C, Schirmer C, Briest W, Anderegg U, Simon JC. Dermal fibroblasts promote the migration of dendritic cells. J Investig Dermatol. 2010;130(2):444–54. doi:10.1038/jid.2009.253.
27. Müller SA, Kohajda T, Findeiss S, Stadler PF, Washietl S, Kellis M, et al. Optimization of parameters for coverage of low molecular weight proteins. Anal Bioanal Chem. 2010;398(7–8): 2867–81. doi:10.1007/s00216-010-4093-x.
28. Mörbt N, Mögel I, Kalkhof S, Feltens R, Röder-Stolinski C, Zheng J, et al. Proteome changes in human bronchoalveolar cells following styrene exposure indicate involvement of oxidative stress in the molecular-response mechanism. Proteomics. 2009; 9(21):4920–33. doi:10.1002/pmic.200800836.
29. Cox J, Matic I, Hilger M, Nagaraj N, Selbach M, Olsen JV, et al. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. Nat Protoc. 2009;4(5): 698–705. doi:10.1038/nprot.2009.36.
30. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10(4):1794–805. doi:10.1021/pr101065j.
31. Bhatia VN, Perlman DH, Costello CE, McComb ME. Software tool for researching annotations of proteins: open-source protein annotation software with data visualization. Anal Chem. 2009; 81(23):9819–23. doi:10.1021/ac901335x.
32. Quintanilla-Dieck MJ, Codriansky K, Keady M, Bhawan J, Runger TM. Expression and regulation of cathepsin K in skin fibroblasts. Exp Dermatol. 2009;18(7):596–602. doi:10.1111/j.1600-0625.2009.00855.x.
33. Pan CP, Kumar C, Bohl S, Klingmueller U, Mann M. Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. Mol Cell Proteomics. 2009;8(3):443–50. doi:10.1074/mcp.M800258-MCP200.
34. Abatangelo G, Barbucci R, Brun P, Lamponi S. Biocompatibility and enzymatic degradation studies on sulphated hyaluronic acid derivatives. Biomaterials. 1997;18(21):1411–5.
35. Maiolica A, Junger MA, Ezkurdia I, Aebersold R. Targeted proteome investigation via selected reaction monitoring mass spectrometry. J Proteomics. 2012;. doi:10.1016/j.jprot.2012.04.048.
36. Clark MF, Lister RM, Bar-Joseph M. ELISA techniques. In: Arthur Weissbach HW, editor. Methods in enzymology. London: Academic Press; 1986. p. 742–66.
37. Schultz GS, Ladwig G, Wysocki A. Extracellular matrix: review of its roles in acute and chronic wounds. World Wide Wounds. 2005. http://www.worldwidewounds.com/2005/august/Schultz/Extrace-Matric-Acute-Chronic-Wounds.html.
38. Buehler MJ. Nature designs tough collagen: explaining the nanostructure of collagen fibrils. Proc Natl Acad Sci USA. 2006; 103(33):12285–90. doi:10.1073/pnas.0603216103.
39. Vedrenne N, Coulomb B, Danigo A, Bonté F, Desmoulière A. The complex dialogue between (myo)fibroblasts and the extracellular matrix during skin repair processes and ageing. Pathol Biol (Paris). 2012;60(1):20–7. doi:10.1016/j.patbio.2011.10.002.
40. Akutsu N, Milbury CM, Burgeson RE, Nishiyama T. Effect of type XII or XIV collagen NC-3 domain on the human dermal fibroblast migration into reconstituted collagen gel. Exp Dermatol. 1999;8(1):17–21. doi:10.1111/j.1600-0625.1999.tb00343.x.
41. Nishiyama T, McDonough AM, Bruns RR, Burgeson RE. Type XII and XIV collagens mediate interactions between banded collagen fibers in vitro and may modulate extracellular matrix deformability. J Biol Chem. 1994;269(45):28193–9.
42. Fluck M, Giraud MN, Tunc V, Chiquet M. Tensile stress-dependent collagen XII and fibronectin production by fibroblasts requires separate pathways. Biochim Biophys Acta Mol Cell Res. 2003;1593(2–3):239–48. doi:10.1016/s0167-4889(02)00394-4.
43. Hatamochi A, Aumailley M, Mauch C, Chu ML, Timpl R, Krieg T. Regulation of collagen VI expression in fibroblasts. Effects of cell density, cell–matrix interactions, and chemical transformation. J Biol Chem. 1989;264(6):3494–9.
44. Dameron KM, Volpert OV, Tainsky MA, Bouck N. Control of angiogenesis in fibroblasts by p53 regulation of thrombospondin-1. Science. 1994;265(5178):1582–4.
45. Chen H, Herndon ME, Lawler J. The cell biology of thrombospondin-1. Matrix Biol. 2000;19(7):597–614. doi:10.1016/s0945-053x(00)00107-4.
46. McKeown-Longo PJ, Hanning R, Mosher DF. Binding and degradation of platelet thrombospondin by cultured fibroblasts. J Cell Biol. 1984;98(1):22–8.
47. Adams JC. Thrombospondin-1. Int J Biochem Cell Biol. 1997; 29(6):861–5.
48. Lawler J. The functions of thrombospondin-1 and-2. Curr Opin Cell Biol. 2000;12(5):634–40. doi:10.1016/s0955-0674(00)00143-5.
49. Wang XQ, Lindberg FP, Frazier WA. Integrin-associated protein stimulates alpha2beta1-dependent chemotaxis via Gi-mediated inhibition of adenylate cyclase and extracellular-regulated kinases. J Cell Biol. 1999;147(2):389–400.

50. Peyssonnaux C, Eychene A. The Raf/MEK/ERK pathway: new concepts of activation. Biol Cell. 2001;93(1–2):53–62. doi:10.1016/S0248-4900(01)01125-X.
51. Wong KK. Recent developments in anti-cancer agents targeting the Ras/Raf/MEK/ERK pathway. Recent Pat Anti-Cancer Drug Discov. 2009;4(1):28–35.
52. Hempel U, Hintze V, Möller S, Schnabelrauch M, Scharnweber D, Dieter P. Artificial extracellular matrices composed of collagen I and sulfated hyaluronan with adsorbed transforming growth factor β1 promote collagen synthesis of human mesenchymal stromal cells. Acta Biomater. 2012;8(2):659–66. doi:10.1016/j.actbio.2011.10.026.
53. Nareyeck G, Seidler DG, Troyer D, Rauterberg J, Kresse H, Schonherr E. Differential interactions of decorin and decorin mutants with type I and type VI collagens. Eur J Biochem/FEBS. 2004;271(16):3389–98. doi:10.1111/j.1432-1033.2004.04273.x.
54. Font B, Eichenberger D, Rosenberg LM, van der Rest M. Characterization of the interactions of type XII collagen with two small proteoglycans from fetal bovine tendon, decorin and fibromodulin. Matrix Biol J Int Soc Matrix Biol. 1996;15(5):341–8.
55. Danielson KG, Baribault H, Holmes DF, Graham H, Kadler KE, Iozzo RV. Targeted disruption of decorin leads to abnormal collagen fibril morphology and skin fragility. J Cell Biol. 1997;136(3):729–43.
56. Merle B, Malaval L, Lawler J, Delmas P, Clezardin P. Decorin inhibits cell attachment to thrombospondin-1 by binding to a KKTR-dependent cell adhesive site present within the N-terminal domain of thrombospondin-1. J Cell Biochem. 1997;67(1):75–83. doi:10.1002/(sici)1097-4644(19971001)67:1<75:aid-jcb8>3.0.co;2-t.
57. Seidler DG, Dreier R. Decorin and its galactosaminoglycan chain: extracellular regulator of cellular function? IUBMB Life. 2008;60(11):729–33. doi:10.1002/iub.115.
58. Sato H, Takino T, Miyamori H. Roles of membrane-type matrix metalloproteinase-1 in tumor invasion and metastasis. Cancer Sci. 2005;96(4):212–7. doi:10.1111/j.1349-7006.2005.00039.x.
59. Lee H, Overall CM, McCulloch CA, Sodek J. A critical role for the membrane-type 1 matrix metalloproteinase in collagen phagocytosis. Mol Biol Cell. 2006;17(11):4812–26. doi:10.1091/mbc.E06-06-0486.
60. Nagase H. Matrix metalloproteinases. A mini-review. Contrib Nephrol. 1994;107:85–93.
61. Nagase H. Cell surface activation of progelatinase A (proMMP-2) and cell migration. Cell Res. 1998;8(3):179–86. doi:10.1038/cr.1998.18.
62. Seltzer JL, Eisen AZ. Native Type I collagen is not a substrate for MMP2 (gelatinase a). J Investig Dermatol. 1999;112(6):993.
63. Strongin AY, Collier I, Bannikov G, Marmer BL, Grant GA, Goldberg GI. Mechanism of cell surface activation of 72-kDa type IV collagenase. Isolation of the activated form of the membrane metalloprotease. J Biol Chem. 1995;270(10):5331–8.
64. Yoshizaki T, Sato H, Furukawa M. Recent advances in the regulation of matrix metalloproteinase 2 activation: from basic research to clinical implication (review). Oncol Rep. 2002;9(3):607–11.
65. Isnard N, Robert L, Renard G. Effect of sulfated GAGs on the expression and activation of MMP-2 and MMP-9 in corneal and dermal explant cultures. Cell Biol Int. 2003;27(9):779–84. doi:10.1016/s1065-6995(03)00167-7.
66. Hou WS, Li ZQ, Gordon RE, Chan K, Klein MJ, Levy R, et al. Cathepsin K is a critical protease in synovial fibroblast-mediated collagen degradation. Am J Pathol. 2001;159(6):2167–77. doi:10.1016/s0002-9440(10)63068-4.
67. Cullen B, Smith R, McCulloch E, Silcock D, Morrison L. Mechanism of action of PROMOGRAN, a protease modulating matrix, for the treatment of diabetic foot ulcers. Wound Repair Regen. 2002;10(1):16–25. doi:10.1046/j.1524-475X.2002.10703.x.
68. Shi L, Ermis R, Kiedaisch B, Carson D. The effect of various wound dressings on the activity of debriding enzymes. Adv Skin Wound Care. 2010;23(10):456–62. doi:10.1097/01.ASW.0000383224.64524.ae.
69. van den Berg AJ, Halkes SB, van Ufford HC, Hoekstra MJ, Beukelman CJ. A novel formulation of metal ions and citric acid reduces reactive oxygen species in vitro. J Wound Care. 2003;12(10):413–8.
70. Karim RB, Brito BLR, Dutrieux RP, Lassance FP, Hage JJ. MMP-2 assessment as an indicator of wound healing: a feasibility study. Adv Skin Wound Care. 2006;19(6):324–7.

3.6 Stable isotope labeling by amino acids in cell culture based proteomics reveals distinct differences of protein expression between spiral and coccoid *Helicobacter pylori*

Stephan A. Müller[a], Sandy Pernitzsch[b], Sven-Bastiaan Haange[a], Jürgen Vogel[b], Cynthia Sharma[b], Martin von Bergen[a,c,d] and Stefan Kalkhof[a]. [In Preparation]

Affiliations

[a]Department of Proteomics, UFZ, Helmholtz-Centre for Environmental Research Leipzig, 04318 Leipzig, Germany

[b]Institute for Molecular Infection Biology (IMIB), University of Würzburg, Würzburg, Germany

[c] Department of Metabolomics, UFZ, Helmholtz-Centre for Environmental Research Leipzig, 04318 Leipzig, Germany

[d]Aalborg University, Department of Biotechnology, Chemistry and Environmental Engineering, Aalborg University, Sohngaardsholmsvej 49, DK-9000 Aalborg, Denmark

Correspondence

Dr. Stefan Kalkhof,
Department of Proteomics
UFZ, Helmholtz-Centre for Environmental Research
Permoserstr. 15
04318 Leipzig, Germany

Key words

Helicobacter pylori, SILAC, cell morphology, quantitative proteomics

3.6.1. Abstract

About 50% of mankind is infected by the carcinogenic, Gram-negative ε-proteobacterium *Helicobacter pylori*. Especially duodenal ulcers and stomach cancer are connected to these infections. *H. pylori* has two viable morphological stages, spiral and coccoid forms. The spiral morphology is infectious whereas coccoid shaped cells show no or strongly reduced infectivity as well as attenuated host colonization efficiency. Here, we investigated relative changes in protein expression between the spiral and the viable coccoid morphologies. For this purpose, we established stable isotope labeling by amino acids in cell culture (SILAC) for the *H. pylori* strain 26695 and applied this method to identify 72% and to relatively quantify 47% of its proteome. Our results show, that crucial processes such as chemotaxis and the cytotoxin associated gene type four secretion apparatus are down-regulated in coccoid cells. Additionally, cell division, transcriptional and translational processes are also inhibited. Furthermore, the proteins arginase and the TNF-α inducing protein that are involved in colonization and inflammation processes are also down-regulated. However, the vacuolating autotransporter A and several outer membrane proteins have shown to be up-regulated in coccoid cells. This newly established method for relative protein quantification of *H. pylori* samples offers new possibilities to study the impact of antibiotics or pathways which are regulated during the infection process.

3.6.2. Introduction

The major human pathogen *Helicobacter pylori* is a gram-negative bacterium that colonizes the stomach of about half the human population. It has the ability to survive in the acidic environment of the stomach. Supported by chemotaxis, *H. pylori* cells swiftly swim to more neutral pH of the gastric mucosa [215]. The urea channel UreI is used to transport urea into the environment in response to acidic conditions [216]. Subsequently, the urease (UreA, UreB) of *H. pylori* converts urea into carbon dioxide and ammonia to partially neutralize the acidic environment [217]. Essential for the survival of *H. pylori* in the stomach, the intracellular urease stabilizes the pH of the cytoplasm in response to strong acids [218].

The cork-screw like shape allows spiral *H. pylori* cells to penetrate the viscous mucosa that protects the gastric epithelial cells from acid [219]. Adhesins such as BabA and OipA promote adherence to gastric epithelial cells [220]. The preferred binding site of *H. pylori* cells is in close proximity to the tight junctions of the epithelial cells in order to have optimal access to nutrients that are released by gastric epithelial cells [221]. Tight junctions are the major barrier that separates the stomach content as well as pathogens from the underlying tissue. They are based on integral membrane proteins such as occludin, claudins and junctional adhesion molecules that connect the cells to each other [222]. Additionally, these proteins are coupled to the actin cytoskeleton via scaffolding proteins [222].

H. pylori deregulates the cell junctions by the virulence factors cytotoxicity-associated immunodominant antigen CagA, the vacuolating cytotoxin autotransporter VacA and the serine protease HtrA. CagA is translocated by the type four secretion system into epithelial cells and disrupts junctions of claudin-4 by activation of the Rho kinase [223]. Secreted VacA as well as ammonia produced by urease reduce the transepithelial electric resistance of gastric epithelial cells [224, 225]. The serine proteases HtrA is translocated into epithelial cells and cleaves E-cadherin, the major protein of adherence junctions [226]. Additionally, the expression of claudins and E-cadherin are reduced in infected epithelial cells [227, 228].

Infected gastric epithelial cells produce several pro-inflammatory cytokines and chemokines such as interleukin 8 (IL-8), IL1-β, IL-6, epithelial derived neutrophil activating protein 78 (ENA-78), tumor necrose factor α (TNF-α) and the granulocyte-monocyte colony stimulating factor (GM-CSF) [74, 76, 77, 229, 230]. The production of these substances is triggered by virulence factors such as CagA or the tumor necrose factor α inducing protein (Tip-α) [73, 230-232]. Persistent infections accompanied by gastric inflammation can cause severe diseases like gastritis, peptic ulcer, mucosa-associated lymphoid tissue (MALT) lymphoma and gastric adenocarcinoma [233, 234].

Three different morphologies of *H. pylori* have been found in gastric biopsies: spiral viable cells and coccoid forms that are further subdivided into viable but under standard conditions non-cultivable and degenerative cells. Saito *et al.* [24] describes the viable coccoid morphology as cells with intact cell wall structures and flagella coiled around their bodies. The degenerative phenotype is characterized by disintegrated membrane structures and cell clustering [24].

In vivo, spiral and coccoid forms of *H. pylori* coexist [235]. The conversion from the spiral to the coccoid morphology can be triggered by starvation, oxidative or acidic stress and antibiotics but also prolonged *in vitro* culturing [83-86]. It is controversially discussed whether coccoid cells are viable or not [236]. However, many studies have proven coccoid *H. pylori* to be biologically active [24, 85, 237-240]. The coccoid morphology, e.g., showed to retain protein expression activity [240]. Even though the infectivity is strongly reduced compared to the spiral morphology, the protein content is not altered by the transformation from the spiral to the coccoid cell shape [95].

In animal experiments, coccoid cells were unable to colonize the stomach mucosa of gnotobiotic piglets [25]. Furthermore, coccoid cells induced less inflammation response in mice [26]. Gastric epithelial immortalized (GES-1) cells stimulated with coccoid *H. pylori* showed lower apoptosis rates and reduced production of pro-inflammatory cytokines and chemokines compared to infection with the spiral form [89]. Different gastric adenocarcinoma cell lines also showed less inflammatory response to coccoid *H. pylori* [87, 88]. However, the documented effects are dependent on the multiplicity of infection.

Therefore, distinct protein expression differences can be expected between both viable morphologies. This was already studied by Bumann *et al.* [95]. In this study, 27 proteins were observed to be differentially expressed between the two morphologies of *H. pylori*. Further studies include investigations on altered protein expression in response to oxidative stress [83, 92], mice colonization [241], gastric epithelial cell apoptosis [98], growth conditions [96], acidic stress [91], and iron uptake [93, 94]. However, all of these proteomic studies were performed by comparative two-dimensional gel electrophoresis (2D-PAGE) [91-96, 98, 241-243].

In recent years, the development of different isotope labeling techniques enabled high throughput relative shotgun quantification of proteins belonging to different cells states within one analysis. Metabolic isotope labeling was first applied by ^{13}C or ^{15}N [195, 244] labeling before stable isotope labeling by amino acids in cell culture (SILAC) was developed [199]. The application SILAC offers direct metabolic labeling of distinct amino acids [199] and has proven to be a technique with low relative standard deviation (< 10%) [38, 214]. Hereby, relative quantification is obtained by the abundance of differentially labeled proteolytic peptides that co-elute during LC-MS/MS analysis.

Typically, stable isotope labeled lysine and arginine are used for SILAC to ensure labeling of all tryptic peptides (except the ones originating from the protein C-terminus). However, cells have to be cultivable in a chemically defined, minimal medium and complete incorporation of labeled amino acids has to be assured. Specific labeling of selected amino acids, namely cysteine and methionine, can also be achieved by growth in the presence of isotopically labeled sulfur [197].

In this study, we established SILAC as a general method to analyze protein expression changes of *H. pylori*. We were able to quantify 47% of the *H. pylori* proteome and investigated distinct differences in protein expression between the spiral and the coccoid morphology. Additionally, we used the HPnc5490 sRNA deletion mutant as internal control. This mutant is known to regulate the transcription of the chemotaxis receptor *tlpB* by trans antisense interaction [245]. Our study reveals regulation of proteins involved in processes like colonization and inflammation promotion of gastric epithelial cell and infectivity.

3.6.3. Methods

Incorporation of isotopically labeled amino acids

The incorporation of stable isotope labeled lysine and arginine was tested before starting the main experiment. *H. pylori* strain 26695 was cultured in Ham's F12 medium (without arginine and lysine, Biosera, UK) supplemented with isotopically labeled arginine ($6 \times {}^{13}C$) and lysine ($6 \times {}^{13}C$, $2 \times {}^{15}N$) (Cambridge Isotope Laboratories, USA) to test the incorporation of these amino acids. Cells were cultivated for more than six cell divisions with one intermediate medium exchange to prevent nutrient deficiency. Isotope labeled amino acid incorporation efficiency was determined after tryptic digestion by LC-MS/MS analysis.

Cell culture

The main experiment was only performed with stable isotope labeled arginine since incorporation of labeled lysine was not sufficient. *H. pylori* strain 26695 was cultured in Ham's F12 medium (without arginine, Biosera, UK) supplemented with either "light", "heavy" or "medium" isotopically labeled arginine (Cambridge Isotope Laboratories, USA) and 5% (v/v) dialyzed fetal calf serum (FCS) (Thermo Scientific, USA) according to Tab. 3-1. Four biological replicates were used for this study.

A preparatory cell culture with the appropriate medium for five cell doublings was applied to reach full incorporation of labeled amino acids in the proteins. Main cultures were started with an optical density of 0.02 at 600 nm. Cells were cultured at 37 °C, 5% O_2 and 10% CO_2 while shaking at 140 rpm. Cultures with medium and heavy labeling were stopped after 8 h showing only spiral morphology. Morphology transformation to coccoid shape was examined after 48 h. Light labeled cells were cultured for 72 h to attain coccoid morphology.

Tab. 3-1: Isotopic label of different cell cultures

Morphology / cell type	**Labeling**	**Designation**
Coccoid	Arginine (6 × ^{12}C, 4 × ^{14}N)	Light
Spiral	Arginine (6 × ^{13}C, 4 × ^{14}N)	Medium
HPnc5490 sRNA deletion mutant	Arginine (6 × ^{13}C, 4 × ^{15}N)	Heavy

Cell harvesting and lysis

Cells were harvested by centrifugation (5000×g), washed twice with 4 °C cold PBS and stored at -80 °C until further usage. Cell pellets were resuspended in lysis buffer (4% w/v SDS, 100 mM Tris/HCl pH 7.6, 0.1 M DTT) and incubated for 3 min at 95°C. For more efficient lysis and cleavage of DNA, cells were further disrupted by ultrasonification. Cell debris and undissolved material was removed by centrifugation (16000×g, 18 °C, 5 min). The protein concentration of each sample was determined by Pierce 660 nm assay (Thermo Fisher Scientific, USA) using bovine serum albumin as standard (Sigma Aldrich, Germany). Afterwards, the corresponding samples of each biological replicate with heavy, medium and light arginine labeling were mixed 1:1:1 according to the protein content.

SDS-PAGE

For 1-D SDS-PAGE, 50 µg total protein of each biological replicate was concentrated using centrifugal filtration devices (Vivacon 500, Sartorius Stedim Biotech GmbH, Germany) with a MW cut-off of 10 kDa. Protein separation by 1-D SDS-PAGE was performed as previously described [4]. Each line was cut in ten gel slices. The slices were divided into two parts and were subjected to different reaction tubes for trypsin and AspN digestions.

Gel elution liquid fraction entrapment electrophoresis

To increase the coverage of low molecular weight (LMW) proteins, an additional protein fractionation was carried out by gel elution liquid fraction entrapment electrophoresis (GELFREE) on a GELFREE® 8100 Fractionation System (Expedeon, USA) with a 12% tris acetate cartridge kit (Expedeon, USA). A protein amount of 200 µg was subjected to GELFREE separation per biological sample. Five fractions in the molecular weight range between 0 and 50 kDa were collected to increase the coverage of low molecular weight proteins (Tab. 3-2).

Tab. 3-2: Program for the GELFREE 12% tris acetate cartridge kit separation.

Fraction	**1**	**2**	**3**	**4**	**5**
t [min]	59	75.6	93.9	112.2	130.5
Buffer Exchange	X	-	X	-	-
Voltage [V]	50	50	85	85	85
MW range	0-10	10-20	20-30	30-40	40-50

Proteolytic digestion

Two different endoproteases, namely trypsin and AspN, were applied for proteolytic digestion of fractions obtained by 1-D SDS-PAGE as well as GELFREE separation. Reduction and alkylation of proteins for in-gel digestion were performed as previously described [246]. In-gel digestions were conducted by addition of either trypsin from bovine pancreas (100 ng per slice, Roche, Germany) or Asp-N from *Pseudomonas fragi* (100 ng per slice, Roche, Germany) and incubation overnight at 37 °C. Digestions were stopped by addition of formic acid (final concentration 1% (v/v)). Peptides were eluted twice with 50% (v/v) acetonitrile containing 0.1% (v/v) formic acid. Eluates were combined with the supernatant and dried by vacuum centrifugation. Samples were reconstituted with 0.1% (v/v) formic acid for LC-MS/MS analysis.

The fractions derived from the GELFREE separation were digested using the filter assisted sample preparation (FASP) protocol [142] with minor modifications. Briefly, approximately 5 µg protein per fraction was used. Proteolytic digestion was performed after alkylation by addition of either 150 ng trypsin or 150 ng AspN and incubation overnight at 37 °C. Eluted peptides were concentrated using ZipTips (Merck Millipore, Germany) according to the protocol of the manufacturer. Samples were dried by vacuum centrifugation and reconstituted with 0.1% (v/v) formic acid.

LC-MS/MS analysis

Peptides were separated on a nano-HPLC system (nanoAquity, Waters, Milford, MA, USA) coupled online with an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA). Peptides were washed with 2% acetonitrile containing 0.1% formic acid and a flow rate of 15 µL/min for 5 min on a trapping column (nanoAquity UPLC column, C18, 180 µm×20 mm, 5 µm, Waters). Peptide separation was performed using a gradient from 2-

40% acetonitrile, 0.1% formic acid on a C18 column (nanoAcquity UPLC column, C18, 75 µm×150 mm, 1.7 µm, Waters) with a flow rate of 300 nl/min and a column temperature of 40 °C. Fractions derived from in-gel digestion were separated with a gradient of 94 min (2 min, 2%; 7 min, 6%; 55 min 20%; 91 min, 40%; 94 min, 80%), whereas GELFREE fractions were separated using a gradient of 154 min (3 min, 2%; 11 min, 6%; 90 min 20%; 150 min, 40%; 154 min, 80%).

The mass spectrometer automatically switched between full scan MS mode (*m/z* 300-1600, R = 60000) and tandem MS acquisition. Peptide ions exceeding an intensity of 2000 counts were fragmented within the LIT by CID (isolation width 3 *m/z*, normalized collision energy 35%, activation time 10 ms, activation Q 0.25). A dynamic precursor exclusion of 2 min for MS/MS measurements was applied.

Data analysis

Peptide identification and relative protein quantification was carried out by Maxquant [199, 247] (version 1.2.2.5, Max Planck Institute of Biochemistry, Munich, Germany). Peptide and protein identification was performed by Andromeda [180] using a concatenated database containing forward and reverse entries of all proteins of *H. pylori* strain 26695 from NCBI refined by results of a proteogenomic analysis [4]. Precursor masses were recalibrated by the option “first search” using a peptide mass tolerance of 20 ppm. The main search was performed with a peptide mass tolerance of 6 ppm and a fragment mass tolerance of 0.5 Da. Two proteolytic missed cleavages were allowed. For samples digested with trypsin, carbamidomethylation of cysteine was defined as fixed modification, whereas oxidation of methionine was set as variable modification. For endoprotease AspN digestions, pyro-glu modification of glutamic acid and glutamine at the peptide N-terminus were additionally specified as variable modifications. AspN was specified to cleave at the N-terminal side of aspartic acid and glutamic acid. An FDR of 1% was applied for peptide and protein identifications. Two unique peptides were necessary for protein identifications. For relative protein quantification, the required minimum ratio count was set to two.

Only proteins which were identified in at least three out of four biological replicates were considered for statistical analysis. A fold change (FC, $\log_2$ of protein ratio) of ± 0.5 was set as regulation threshold for proteins. A heteroscedastic, two-sided student t-test was applied to distinguish significant protein regulation ($\alpha = 5\%$). Proteins were defined as significantly regulated between the different cell states if they fulfilled both thresholds, an average FC exceeding ± 0.5 and a t-test p-value lower than 0.05. All quantified proteins were loaded into the kyoto encyclopedia of genes and genomes (KEGG) system [248, 249] using the software tool KEGGArray [250] for pathway and functional analysis. Furthermore, the Clusters of Orthologous Groups of proteins (COGs) database was used to functionally classify proteins [251].

3.6.4. Results

Establishment of SILAC for H. pylori

We have chosen the chemically defined Ham's F-12 medium for our SILAC study since it permits growth of *H. pylori* without influencing the morphology [22, 23]. Growth characteristics and cell morphology were not altered by the SILAC medium. Incorporation of isotopically labeled lysine and arginine was tested before starting the main experiment. Proteins were fully labeled with arginine but not with lysine. Even at four fold lysine concentration, lysine incorporation was below 80% after six cell divisions. Therefore, we decided to use only isotopically labeled arginine in our study.

Protein identifications and quantifications

Overall, 1143 proteins, representing 72% of the proteome of *H. pylori*, were identified by at least two unique peptides. Within this set, 743 proteins (47% proteome coverage) were quantified in at least three out of four biological replicates. Comparison of the spiral and the coccoid morphology showed significant expression differences (t-test p-value < 0.05, average fold change > 0.5 or < -0.5) for 162 proteins of which 74% displayed a higher expression in spiral cells (Fig. 3-1 A; Tab. 3-3). Only 32 proteins fulfilled the regulation thresholds when comparing spiral wild type cells with the ΔHPnc5490 sRNA mutant (Fig. 3-1 B; Tab. 3-3).

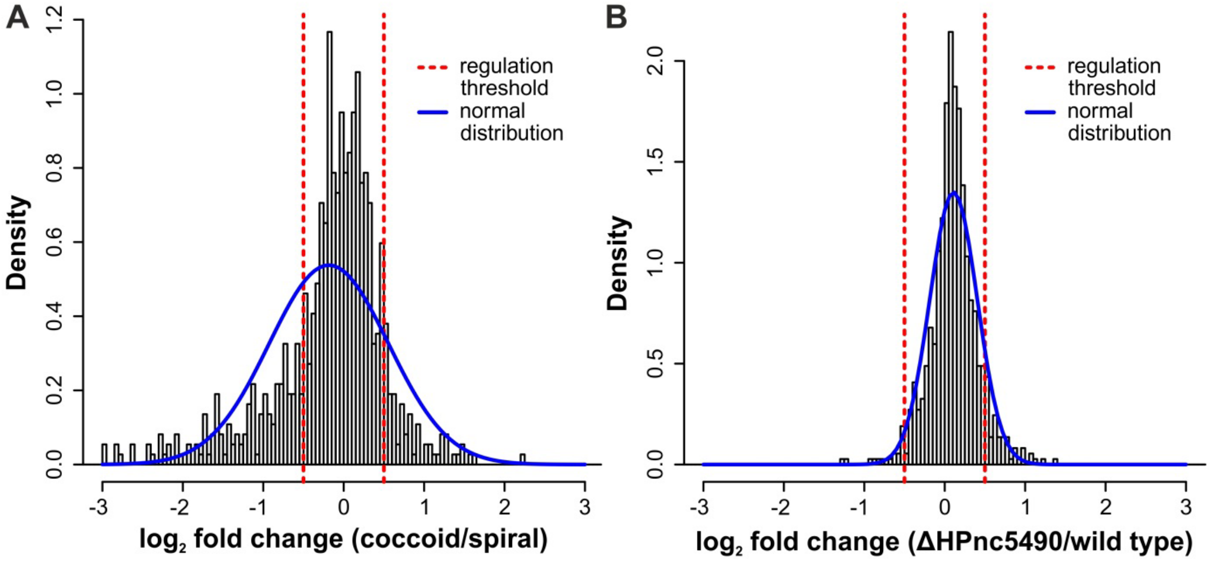


Fig. 3-1: Ratio blots of the coccoid and spiral morphology of *H. pylori* strain 26695 (**A**) and the HPnc5490 sRNA deletion mutant and the wild type (**B**). The distribution of the regulation according to the morphology is shifted towards lower expression for the coccoid morphology. The expression values are widely distributed dependent on the morphology whereas the ΔHPnc5490 sRNA mutant ratio blot is narrow distributed.

Tab. 3-3: Protein identifications and quantifications between the different cell states.

	Spiral vs. coccoid	**ΔHPnc5490 sRNA deletion mutant vs. spiral**
Protein identifications	1143	1143
Quantified proteins in ≥ 3 biological replicates	743	743
Regulated proteins	162 (22.8%)	32 (4.3%)
Higher expression in spiral cells	120 (16.2%)	13 (1.8%)
Higher expression in coccoid cells / HPnc5490 sRNA deletion mutant	42 (5.6%)	19 (2.6%)

Since only arginine was used in this study for isotopic labeling, only arginine containing peptides can be utilized for relative quantification. AspN was applied as additional protease to trypsin to increase the number of unique peptides for protein identifications and quantifications (Fig. 3-2). The application of AspN only slightly increased the number of protein identifications on average by 4.7% (50.5), but strongly supported the quantification by 65.3% (6829.8) additional unique peptides. Thus, the number of protein quantifications was increased on average by 16.7% (111.0) in comparison to trypsin. In summary, AspN provided 17.9% (113) additional protein quantifications in at least three out of four biological samples.

The application of the GELFREE separation for proteins below 50 kDa provided on average 11.4% (110.3) additional protein identifications and 20.3% (131.0) quantifications. The number of quantified proteins in at least three out of four biological samples was increased by 22.4% (136) by the GELFREE separation.

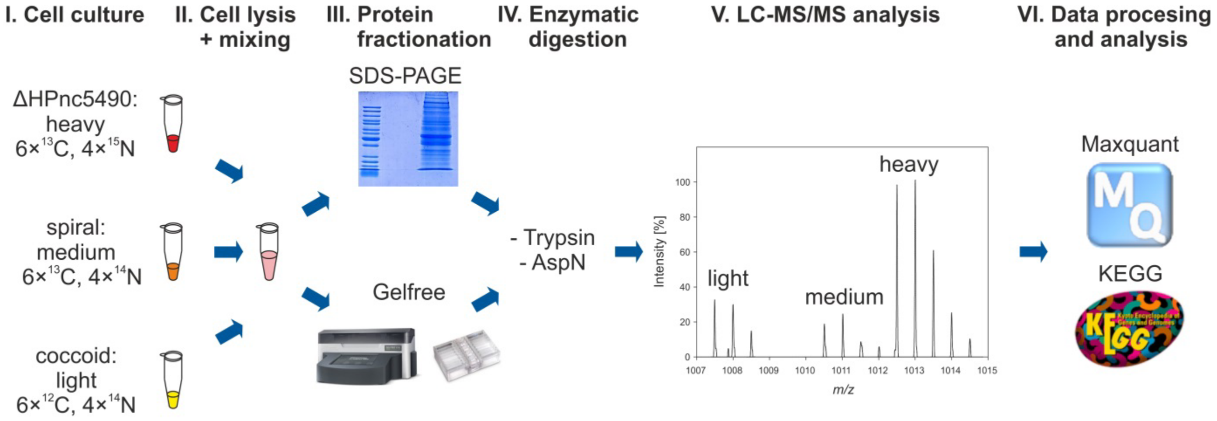


Fig. 3-2: Experimental workflow of the SILAC experiment. (**I**) Starter culture is grown until 95% incorporation of the labeled amino acids is reached. The main culture is performed in the same media. (**II**) After cell harvesting and lysis, cell lysates are mixed 1:1:1 according to their protein content. (**III**) Proteins are separated by SDS-PAGE and GELFREE fractionation. (**IV**) The gained fractions are digested separately with trypsin and AspN. (**V**) Samples are analyzed by LC-MS/MS. (**VI**) Protein identification and quantification is performed by Maxquant. Pathway and functional analysis is carried out with KEGG.

Differences in protein expression of the wild type compared to the HPnc5490 sRNA mutant

The HPnc5490 sRNA is predicted to interact with the 5'UTR of the *tlpB* (HP0103) mRNA which encodes for one of the four chemotaxis receptors. TlpB is believed to play a role in pH sensing, quorum sensing and pH taxis [245]. Indeed, TlpB was found to be 9.5-fold up-regulated (HP0103, $\log_2FC = 3.24$, p-value = 1.6E-04) in the ΔHPnc5490 sRNA mutant compared to the wild type. The protein $CheV_2$ (HP0616, $\log_2$ FC = 0.51, p-value = 1.0E-03) which is also involved in chemotaxis was also significantly up-regulated in the mutant. Remarkably, the arginase RocF (HP1399, $\log_2$ FC = -1.41, p-value = 8.4E-05) that is crucial for the buffering of the acidic environment in the stomach, showed to be down-regulated in the mutant. No accumulation of functionally related proteins was observed among the significantly regulated proteins for the HPnc5490 deletion mutant.

Differences between spiral and coccoid morphology

Overall, 162 proteins were found to be significantly differentially expressed in coccoid cells compared to the spiral morphology. The regulated proteins were classified into functional groups according to KEGG database [249] and COG identifiers [251] (Fig. 3-3, Tab. 3-4). Most groups showed to be down-regulated in coccoid cells. Several proteins involved in cell division and transcription and translation were lower expressed in coccoid cells. Additionally, several proteins related to chemotaxis or infectivity were found to be lower expressed. Remarkably, numerous outer membrane proteins were higher expressed in coccoid cells.

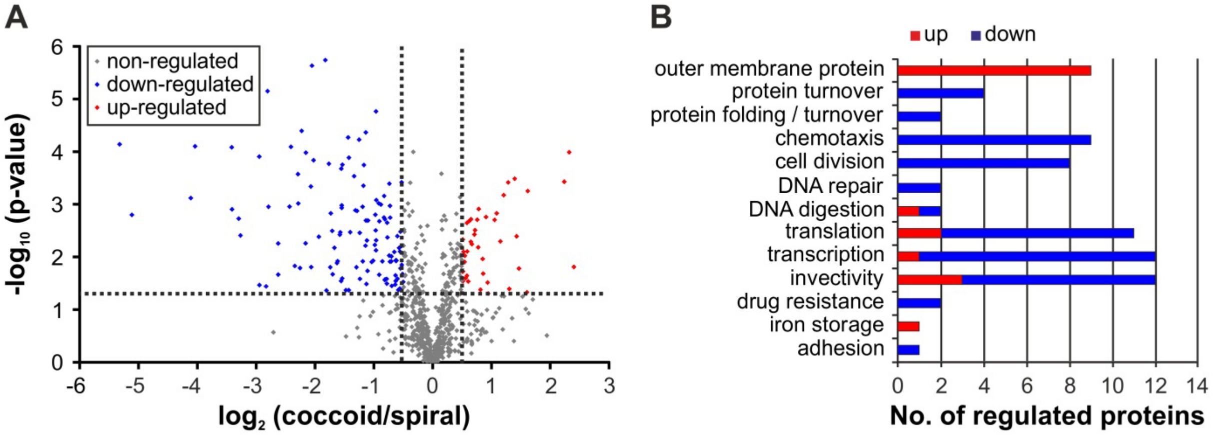


Fig. 3-3: (**A**) Vulcano blot of relatively quantified proteins between the coccoid and the spiral morphology of *H. pylori*. The dotted lines indicate thresholds set for regulation ($FC < -0.5$ and $FC > 0.5$) and the significance (t-test $p\text{-value} < 0.05$). Significantly higher expressed proteins in coccoid cells are indicated in red whereas lower expressed proteins are marked blue. (**B**) Classification of regulated proteins. Red bars indicate higher expression for the coccoid morphology whereas blue bars indicate lower expression.

Expression of several proteins involved in DNA replication showed to be attenuated in coccoid cells. The NAD dependent RNA ligase LigA (HP0615) and the DNA polymerase subunit α (HP1460) were also significantly down-regulated in coccoid cells. The cell division protein FtsZ (HP0979) and the plasmid replication-partition related protein exhibited the most pronounced down-regulation (Tab. 3-3). The protein FtsZ is essential for the production of a new cell wall during cell division [252].

Additionally, the expression of many transcriptional regulators, as well as the DNA-directed RNA polymerase subunit ω (HP0776), was reduced in coccoid cells (Tab. 3-3). In contrast, the adenine/cytosine DNA methyltransferase was higher expressed in the coccoid morphology.

Interestingly, proteins involved in chemotaxis as well as the flagellar assembly of *H. pylori* were also lower expressed in coccoid cells. The chemotaxis response regulators CheV$_2$ (HP0616) and CheV$_3$ (HP0393) were down-regulated in the coccoid morphology. Moreover, the flagellar motor switch proteins FliN/FliY (HP1030) and FliG (HP0352) were found to be lower expressed. In addition, the hook basal-body proteins FliE (HP1557) and FlgG (HP1585), the MS-ring FliF (HP0351) as well as the motor switch protein G (HP0352) were less abundant in coccoid cells.

Different cytotoxicity-associated gene (cag) pathogenicity island proteins and the cytotoxicity-associated immunodominant antigen CagA (also named Cag26) itself were found to be down-regulated in coccoid cells. Namely, Cag6 (HP0526), Cag14 (HP0535), Cag22 (HP0543) and Cag26 / CagA (HP0547) were significantly lower expressed in coccoid cells. CagA was 5.5 times higher expressed in the spiral morphology. The cag proteins are part of the type IV secretion system that translocates CagA into the epithelial cells of the stomach mucosa during infection. On the other hand expression of the vacuolating cytotoxin autotransporter VacA (HP0887), the virulence associate protein VapD (HP0315), and the cell adhesion protein OipA (HP0638) were significantly increased in coccoid cells.

Additionally, the infection-related proteins arginase RocF (HP1399) and tumor necrose factor α (TNF-α) inducing protein (Tip-α, HP0596) were significantly down-regulated in coccoid cells. The abundance of two urease accessory proteins UreE (HP0070) and UreG (HP0068), which are meaningful for pH adaption of the environment, were also found to be reduced.

Tab. 3-4: Classification of regulated proteins between the coccoid and spiral morphology of *H. pylori*

gi accession	protein description	HP No.	log$_2$ FC	T-test p-value	Functional class
15646069	DNA polymerase III subunit alpha	HP1460	-2,95	1,23E-04	cell division
15645752	plasmid replication-partition related protein	HP1138	-1,82	1,82E-06	cell division
15645594	cell division protein FtsZ	HP0979	-1,67	6,68E-03	cell division
15645240	NAD-dependent DNA ligase LigA	HP0615	-1,55	1,17E-03	cell division
15645128	DNA gyrase subunit B	HP0501	-1,54	2,56E-02	cell division
15645753	SpoOJ regulator (Soj) / ATPases involved in chromosome partitioning	HP1139	-0,97	3,12E-03	cell division
15645657	response regulator for cell division	HP1043	-0,96	1,72E-05	cell division
15645796	Predicted ATPase implicated in cell cycle control	HP1182	-0,96	8,33E-04	cell division
15646192	flagellar basal body rod protein FlgG	HP1585	-2,95	3,43E-02	chemotaxis
15645644	flagellar motor switch protein FliY	HP1030	-1,77	1,70E-04	chemotaxis
15646164	flagellar hook-basal body protein FliE	HP1557	-1,56	2,86E-02	chemotaxis
15644980	flagellar motor switch protein G FliG	HP0352	-1,44	5,37E-05	chemotaxis
15645649	flagellar biosynthesis regulator FlhF	HP1035	-1,42	1,30E-04	chemotaxis
15645241	chemotaxis protein CheV2	HP0616	-0,98	1,58E-03	chemotaxis
15645021	chemotaxis protein CheV3	HP0393	-0,81	1,92E-03	chemotaxis
15645370	flagellar protein FlaG	HP0751	-0,81	2,66E-02	chemotaxis
15644979	flagellar MS-ring protein FliF	HP0351	-0,78	3,22E-02	chemotaxis
15646091	Exodeoxyribonuclease 7 small subunit	HP1482	-0,96	4,82E-03	DNA digestion
15645227	endonuclease III	HP0602	1,09	1,48E-03	DNA digestion
15645673	Holliday junction DNA helicase B	HP1059	-2,29	2,67E-04	DNA repair
15645541	recombination protein RecR	HP0925	-1,94	4,14E-03	DNA repair
15645238	ABC transporter, ATP-binding protein	HP0613	-1,55	2,10E-04	drug resistance
15644863	putative beta-lactamase	HP0235	-0,72	1,79E-03	drug resistance
15645161	cag pathogenicity island protein Cag14	HP0535	-2,63	2,15E-02	infectivity
15646009	arginase RocF	HP1399	-2,41	8,11E-05	infectivity
15645173	Cytotoxicity-associated immunodominant antigen CagA / Cag 26	HP0547	-2,35	1,48E-02	infectivity
15645169	cag pathogenicity island protein Cag22	HP0543	-2,05	2,32E-06	infectivity
15645152	cag pathogenicity island protein Cag6	HP0526	-1,53	1,77E-04	infectivity
15645221	TNF-α inducing protein Tip-α	HP0596	-1,34	2,92E-04	infectivity
15644698	urease accessory protein UreG	HP0068	-0,86	1,21E-02	infectivity
15644700	urease accessory protein UreE	HP0070	-0,65	8,02E-03	infectivity
15644804	cell binding factor 2	HP0175	-0,51	2,12E-02	infectivity
15645505	vacuolating cytotoxin autrotransporter	HP0887	1,05	1,98E-03	infectivity
15645262	outer membrane protein (Omp13) / OipA	HP0638	1,21	6,69E-04	infectivity / OMP
15644943	virulence associated protein D (VapD)	HP0315	1,60	4,69E-02	infectivity
15645277	nonheme iron-containing ferritin (Pfr)	HP0653	2,23	3,72E-04	iron storage
15646005	outer membrane protein (Omp30)	HP1395	0,51	1,19E-02	OMP
15644882	outer membrane protein (Omp8)	HP0254	0,51	4,65E-03	OMP
15645539	outer membrane protein (Omp22)	HP0923	0,58	2,27E-02	OMP
15645739	peptidoglycan associated lipoprotein precursor (Omp18)	HP1125	0,62	2,07E-03	OMP
15644757	outer membrane protein (Omp4)	HP0127	0,66	5,68E-03	OMP
15645770	outer membrane protein (Omp25)	HP1156	0,86	2,07E-02	OMP

gi accession	protein description	HP No.	$\log_2$ FC	T-test p-value	Functional class
15645329	outer membrane protein (Omp15)	HP0706	0,91	1,74E-03	OMP
15645100	outer membrane protein (Omp11)	HP0472	1,39	3,26E-04	OMP
15644740	co-chaperone and heat shock protein (GrpE)	HP0110	-0,83	6,92E-04	protein folding / turnover
15645638	co-chaperone-curved DNA binding protein A (CbpA)	HP1024	-0,55	7,94E-03	protein folding / turnover
15644666	ATP-dependent Clp protease (ClpA)	HP0033	-3,41	1,25E-03	protein turnover
15644665	hypothetical protein HP0032 /predicted ClpS protease	HP0032	-1,81	4,34E-02	protein turnover
15645984	ATP-dependent protease ATP-binding subunit	HP1374	-1,57	3,41E-03	protein turnover
15645989	ATP-dependent protease (Lon)	HP1379	-0,82	2,75E-02	protein turnover
15645175	transcription termination factor Rho	HP0550	-2,79	1,12E-03	transcription
15645485	transcription elongation factor GreA	HP0866	-1,40	3,37E-03	transcription
15645395	DNA-directed RNA polymerase subunit omega	HP0776	-1,31	1,30E-03	transcription
15644718	RNA polymerase sigma factor RpoD	HP0088	-1,24	3,49E-03	transcription
15645469	type I restriction enzyme M protein (HsdM)	HP0850	-0,98	5,90E-03	transcription
15644795	response regulator (OmpR)	HP0166	-0,88	1,01E-03	transcription
15645285	ribonuclease H	HP0661	-0,78	2,61E-02	transcription
15645518	hydrogenase expression/formation protein (HypB)	HP0900	-0,60	4,03E-03	transcription
15645635	response regulator	HP1021	-0,53	6,16E-03	transcription
15645951	nickel responsive regulator	HP1338	-0,53	3,82E-04	transcription
15644685	adenine/cytosine DNA methyltransferase	HP0054	2,40	1,55E-02	transcription
15645682	ribosomal protein L11 methyltransferase	HP1068	-2,81	7,11E-06	translation
15644897	putative ATP-binding protein	HP0269	-1,73	1,48E-03	translation
15645412	peptide deformylase / tRNA	HP0793	-1,30	1,57E-02	translation
15646061	tRNA modification GTPase TrmE	HP1452	-1,11	1,84E-02	translation
15645661	ribosome-binding factor A	HP1047	-1,10	2,02E-03	translation
15644646	hypothetical protein HP0013 / predicted t-RNA	HP0013	-0,64	3,34E-02	translation
15646056	50S ribosomal protein L34	HP1447	-0,58	4,26E-02	translation
15645267	glutamyl-tRNA synthetase	HP0643	-0,55	3,30E-03	translation
15645176	50S ribosomal protein L31	HP0551	-0,52	1,41E-02	translation
15645811	30S ribosomal protein S12	HP1197	0,56	2,51E-02	translation
15646040	ribosomal RNA small subunit methyltransferase A	HP1431	0,59	7,84E-03	translation
15646066	hypothetical protein HP1457 / putative collagen binding protein	HP1457	-1,18	4,45E-04	adhesion

3.6.5. Discussion

Recommendations for SILAC set-up for *H. pylori*

The application of SILAC for quantitative proteomics is rather uncommon for bacteria. The main reason for this is the autotrophy of bacteria to many amino acids. However, some studies successful utilized SILAC for bacteria [253, 254]. SILAC offers robust relative quantification due to the early stage of labeling. Other labeling techniques for proteomic studies include chemical derivatization of proteins or peptides which is an additional error source. Furthermore, MS analyses of SILAC samples do not require special MS method adjustment like for example iTRAQ. Not least, data processing is well automated for SILAC experiments [255]. Hence, with exception of the culture medium, no changes have to be accomplished in comparison to standard bottom-up protocols of a proteome analysis.

Here, SILAC was established for *H. pylori* strain 26695. To our knowledge, this is the first publication about a SILAC study of *H. pylori*. Growth of *H. pylori* was shown in chemically defined Ham's F12 medium without influencing the morphology. The incorporation of stable isotope labeled lysine and arginine was tested. Complete labeling with arginine was achieved after five cell doublings. However, we could not force *H. pylori* to stop lysine synthesis. Even after raising the concentration of lysine four fold to the Ham's F12 medium recipe, only 80% incorporation was gained. *H. pylori* has shown to be auxotroph for arginine but has the ability to produce lysine [23, 256, 257].

A possibility to include lysine for SILAC analysis of *H. pylori* would be to create a lysine deficient strain of *H. pylori*. Therefore, a deletion mutant of the gene for diaminopimelate decarboxylase (*dapE,* HP0290) could be utilized [258]. However, such a mutation could possibly lead to undesired side effects. Alternatively, the application of other isotopically labeled amino acids for which *H. pylori* shows deficiency [23, 256, 257], would be another opportunity to improve the quantification rates. *H. pylori* is most likely leucine deficient since no leucine synthetase is reported. Additionally, *H. pylori* showed no growth in the absence of leucine [23, 256, 257]. Therefore, we recommend using isotopically labeled leucine in addition to arginine. However, the custom Ham's F12 SILAC medium we used in our experiment only permitted the addition of lysine and arginine.

Another strategy, called SULAQ, would be to grow *H. pylori* in the presence of isotopically labeled sulfur [197, 198] to label methionine and cysteine in combination with SILAC. However, the basis of this technique is that the examined organism is able to synthesize these sulfur containing amino acids. Three different studies on the nutritional requirements of *H. pylori* have shown that this organism does not survive without methionine [23, 256, 257]. Nevertheless, one of these studies reported growth of *H. pylori* in the absence of cysteine [256] whereas Testerman *et al.* were not able to substitute magnesium sulfate for cysteine [23]. The existence of a cysteine synthetase (HP0107) suggests the possibility of sulfur labeling. After careful testing of cysteine synthesis in the presence of different sulfur sources, it could be worth considering sulfur labeling in combination with isotopically labeled arginine and/or leucine for quantitative proteomics.

Improvement of the protein quantification rate

In order to improve the protein quantification rate, we used endoprotease AspN as an additional protease. The usage of AspN was preferred over the application of ArgC due to the higher orthogonality to trypsin. ArgC would produce a huge amount of identical peptides compared to tryptic digestion. Additionally, ArgC creates longer, harder identifiable peptides than AspN. The application of AspN in addition to trypsin increased the number of quantified proteins by 17.9% (113 in at least three out of four replicates). Additionally, the quantification accuracy and reliability could be enhanced by the identification of 65.3% (6829.8) additional unique peptides.

Furthermore, the GELFREE protein separation was used to increase the coverage of proteins below 50 kDa. Hereby, 22.4% (136, in at least three out of four replicates) additional proteins could be quantified. Our approaches clearly demonstrate that protein fractionation of LMW as well as the application of AspN as additional protease significantly increases the protein identification rates. Additionally, the accuracy of protein quantifications is improved by the larger number of quantification features per protein.

Evaluation of SILAC for *H. pylori*

This study included four biological replicates in a triple SILAC approach. Regulation significance was evaluated by statistical heteroscedastic, two-sided student t-test ($\alpha < 0.05$). Additionally the minimal required average FC was set to 0.5 for both, up- and down-regulation. Besides spiral and coccoid cells, the HPnc5490 knock-out mutant was used as third cell stage to evaluate the performance of SILAC for *H. pylori*. The target protein TlpB has shown to be 9.5-fold up-regulated. No functionally related clusters with accumulation of significantly regulated proteins could be identified for the HPnc5490 knock-out mutant. In contrast, coccoid and spiral cells showed to have more pronounced protein expression differences. Our study revealed significant regulation of cell division, transcriptional and translational processes as well as chemotaxis and infectivity related proteins. Here we discuss quantitative differences in the proteome of the spiral and coccoid morphology.

Effects on cell division

Several proteins related to cell division showed to be down-regulated in coccoid cells. Remarkably, important proteins for DNA replication were down-regulated (Fig. 3-4). Among these proteins, the NAD-dependent DNA ligase LigA (HP0615) plays a critical role in the joining of Okazaki fragments during DNA replication, but also in DNA recombination and repair mechanisms [259-262]. The DNA polymerase III holoenzyme was also down-regulated (Fig. 3-4). Moreover, the lower expressed protein FtsZ is known to be essential for the synthesis of a new cell wall during cell division [252]. In summary, cell division is strongly reduced when *H. pylori* differentiates to its coccoid morphology.

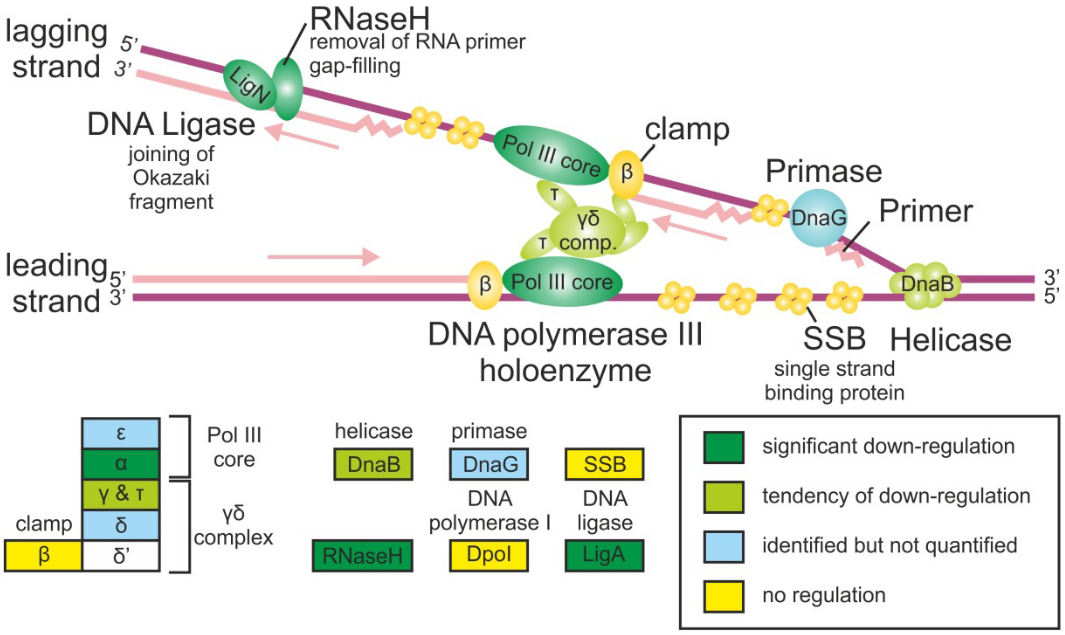


Fig. 3-4: Regulated proteins associated with DNA replication. Protein regulation is shown with respect to the coccoid morphology. The DNA polymerase III core α unit was significantly down-regulated. Expression of the DNA ligase LigA as well as the RNAseH was also lower in coccoid cells. These findings suggest that DNA replication is diminished in the coccoid morphology of *H. pylori*. Modified according to KEGG pathway map hpy03030 (DNA replication) [248, 249].

Effects on transcription and translation

Transcriptional and translational processes were also found to be reduced in the coccoid morphology. Among the transcription regulators, only the adenine/cytosine DNA methyltransferase (HP0054) was up-regulated in coccoid cells. Most likely, transcription of genes is down-regulated based on DNA methylation. Another possible explanation would be that *H. pylori* focuses more on DNA maintenance than on DNA replication and cell division. Additionally, proteins involved in protein folding and turnover such as chaperones and Clp proteases were also found to be down-regulated. This finding suggests that not only protein expression, but also degradation is generally reduced in coccoid cells. Hereby, *H. pylori* possibly establishes the basis to survive in its dormant cell stage for a long time period.

Regulation of outer membrane proteins

Outer membrane proteins are the only functional group that was found to be higher expressed in coccoid cells. Many outer membrane proteins are known to be involved in cell adhesion [263]. Potentially, coccoid *H. pylori* cells attach more strongly to host cells. It has been shown that coccoid H. pylori is able to survive in the palatine tonsils of humans [64]. This may be connected to the improved adhesion ability.

Regulation of chemotaxis associated proteins

Different proteins related to chemotaxis and the flagellar assembly were also down-regulated. *H. pylori* has four different chemotaxis receptors. The chemoreceptor *tlpA* (HP0099) binds arginine [264], TlpB (HP0103) recognizes low pH [215], TlpD (HP0599) is a soluble receptor for nutrients (intracellular energy levels) [265], whereas the function of TlpC (HP0082) is unknown [264]. The response regulators CheW (HP0391), $CheV_1$, $CheV_2$ and $CheV_3$ bind to these receptors and attract the phosphokinase CheA (HP0392). CheA has the ability to autophosphorylate. CheA-P transfers a phosphate group to the response regulator CheY (HP1067) [264]. When CheY is not phosphorylated, *H. pylori* cells swim only in one direction [266]. Phosphorylated CheY-P leads to direction changes or tumbling by signal transduction of CheY-P to the flagellar assembly proteins FliM and FliN [266]. The kinase CheZ (HP0170) is responsible for dephosphorylation of CheY-P to stop the direction change signal [264, 267].

Binding of ligands to one of the chemotaxis receptors can either lead to phosphorylation of CheY and subsequent direction changes (TlpB) or to a stop of the signal cascade and swimming in one direction (TlpA, TlpD) [215]. Deletion mutants of *cheW*, *cheV₁* and *cheV₂* also swim only in one direction [264]. Conclusively, these response regulators are necessary for CheY phosphorylation. Loss of $CheV_3$ on the other hand leads to repetitive direction changes [268].

In our study chemotaxis response regulators $CheV_2$ and $CheV_3$ were down-regulated. Additionally, CheA, CheW and CheY show the tendency of down-regulation though they do not fulfill the thresholds for significant regulation. The expression of all four chemotaxis receptors remains unaffected by the morphology.

Moreover, the flagellar motor switch proteins FliN/FliY and FliG are lower expressed. FliN is a very important protein for chemotaxis as it transduces the signal for direction changes from CheY-P [266]. Other proteins in the flagellar assembly, like the flagellar hook-basal body protein FliE or the flagellar basal body rod protein FlgG are also down-regulated (Fig. 3-5). In conclusion, signal transduction for chemotactic behavior is inhibited and the flagellar assembly seems to be partly decomposed. These findings suggest that coccoid *H. pylori* cells are most likely a non-chemotactic phenotype.

Actually, chemotaxis plays a crucial role in the colonization and infection of the stomach mucosa by *H. pylori* [25, 215, 265, 269-273]. Deletion mutants of *cheY* and *cheA* failed to colonize the gastric mucosa of mice [273]. Eaton *et al.* [25] reported that flagellin (FlaA or FlaB) deficient *H. pylori* strains were motile and had morphologically normal flagella but could not persist longer than ten days in the stomach of gnotobiotic piglets. Motile but non-chemotactic *ΔcheY* mutants lost the ability to colonize the stomach of Mongolian gerbils [270]. Deletion mutants of the pH chemoreceptor *tlpB*, on the other hand, were able to colonize gerbil stomachs but generated significantly reduced inflammation compared to the wild type.

Terry *et al.* [269] showed that non-chemotactic deletion mutants either of *cheA*, *cheW* or *cheY* were able to infect the stomach of FVB/N mice. However, the 50% infection dose was increased and *ΔcheW* mutants did only colonize the stomach corpus but not the antrum [269]. Additionally, *ΔcheW* mutants did not reach the infection level of the wild type strain before six months [269]. Briefly summarized, chemotaxis is important for colonization efficiency, especially of the antrum. Furthermore, inflammatory response of the host is attenuated when the stomach mucosa is infected by non-chemotatic phenotypes. However, effects are strongly dependent on the animal model system.

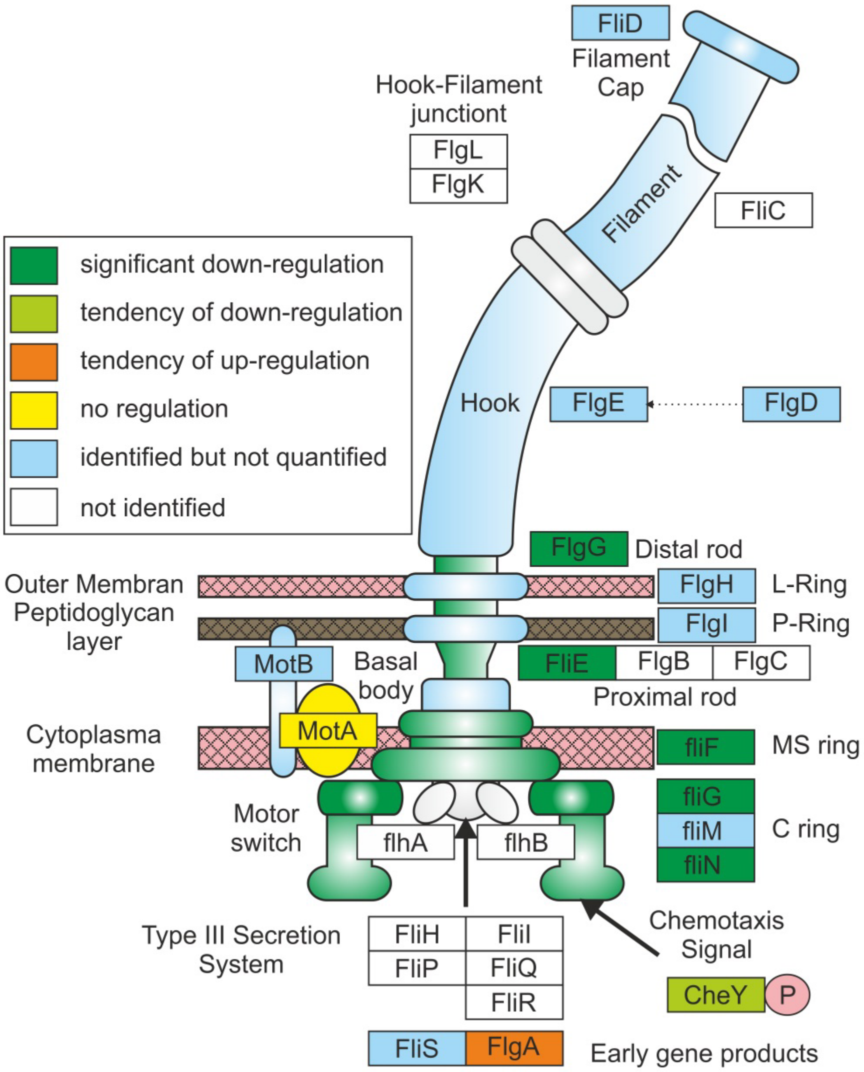


Fig. 3-5: Regulated proteins associated with the flagellar assembly of *H. pylori*. Protein regulation is shown with respect to the coccoid morphology. Modified according to KEGG pathway map hpy02040 (flagellar assembly) [248, 249].

Regulation of pathogenicity related proteins

Most interestingly, coccoid cells exhibited lower abundances of several *cag* pathogenicity island proteins which are responsible for CagA translocation into host epithelial cells. CagA activates NFκ-B in infected cells and thereby promotes production of pro-inflammatory interleukin 8 (IL-8), epithelial derived neutrophil activating protein78 (ENA-78), tumor necrose factor α (TNF-α), and the granulocyte-monocyte colony stimulating factor (GM-CSF) [76, 77]. A persistent inflammation can cause peptic ulcer or stomach cancer in the worst case.

In our study Cag6, Cag14, Cag22 and Cag26/CagA were significantly down-regulated in coccoid cells. Cag14 and Cag22 showed to have no effect on infectivity, whereas *Δcag22* mutants exhibited significantly reduced translocation of CagA into epithelial cells [79]. *CagA* deletion mutants showed no infectivity [79]. In conclusion, infectivity of coccoid cells is attenuated due to diminished translocation of CagA accompanied with its non-chemotatic phenotype (Fig. 3-6).

The pathogenicity related enzyme arginase RocF was also found to be significantly down-regulated in coccoid cells. It hydrolyzes arginine to generate urea and ornithine [274]. Urea can be catabolized by its urease to produce carbon dioxide and ammonia. Thereby, *H. pylori* is able to neutralize the acidic environment in the stomach [217] (Fig. 3-6). The pH optimum of the *H. pylori* arginase at 6.1 (activity down to pH 5.5) emphasizes its exceptional role for acidic resistance [217]. In addition to urea generation, the *H. pylori* arginase also inhibits T-cell proliferation and reduces the expression of the T-cell receptor ζ-chain [275]. Furthermore, nitric oxide generation of macrophages, which requires arginine as substrate, is inhibited by the *H. pylori* arginase [276].

Moreover, we also found that the protein abundance of two urease accessory proteins UreE (HP0070) and UreG (HP0068) significantly decreased. Volland *et al.* [277] have shown that the *ΔureE* mutant had strongly reduced urease activity whereas the *ΔureG* mutant completely lost its urease activity. This indicates that urease activity might be strongly reduced for the coccoid morphology. Apparently, infectivity as well as colonization ability of coccoid *H. pylori* is also decreased by the reduced ability to adapt to acidic stress and the diminished capability to modulate immune response of the host.

Tip-α is another infection related protein which exhibited lower expression in coccoid cells of *H. pylori*. Tip-α forms homo-dimers that are secreted into the environment by the type two secretion system [278]. Subsequently, it is shuttled by membrane located nucleolin into the cytoplasm of stomach epithelial cells (Fig. 3-6) [229]. Translocated into the cells, Tip-α induces gene expression of TNF-α, IL-6 and several chemokines by activation of NFκ-B [229, 230]. Incubation of immortalized human gastric epithelial mucosa cells (GES-1) and a gastric cancer cell line (SGC7901) with Tip-α promoted the expression of TNF-α, IL-1β, and IL-8 [74] (Fig. 3-6). These proteins play key roles in inflammatory response and tumor promotion.

Different human gastric cancer cell lines exhibit nucleolin on the cell surface whereas normal epithelial mouse cells of the glandular stomach had no significant amounts on the cell surface [279]. Moreover, it has been shown that *H. pylori* isolates derived from patients with gastric cancer produced significantly higher amounts of Tip-α than those from patients with chronic gastritis [232]. Furthermore, Tip-α levels of patients which later developed cancer were also increased [230, 232]. Nucleolin is a possible drug target. The DNA aptamer AS1411, which is in clinical trials for the treatment of renal cancer and myeloid leukemia [280, 281], inhibits growth of stomach cancer cells by induction of S-phase arrest [279]. Interestingly, colonization efficiency of the murine mucosa by a mouse-adapted strain of *H. pylori* significantly decreased for *tip-α* knock-out mutants [282]. Eventually, reduced levels of Tip-α and arginase could be jointly responsible for attenuated colonization efficiency of coccoid *H. pylori*.

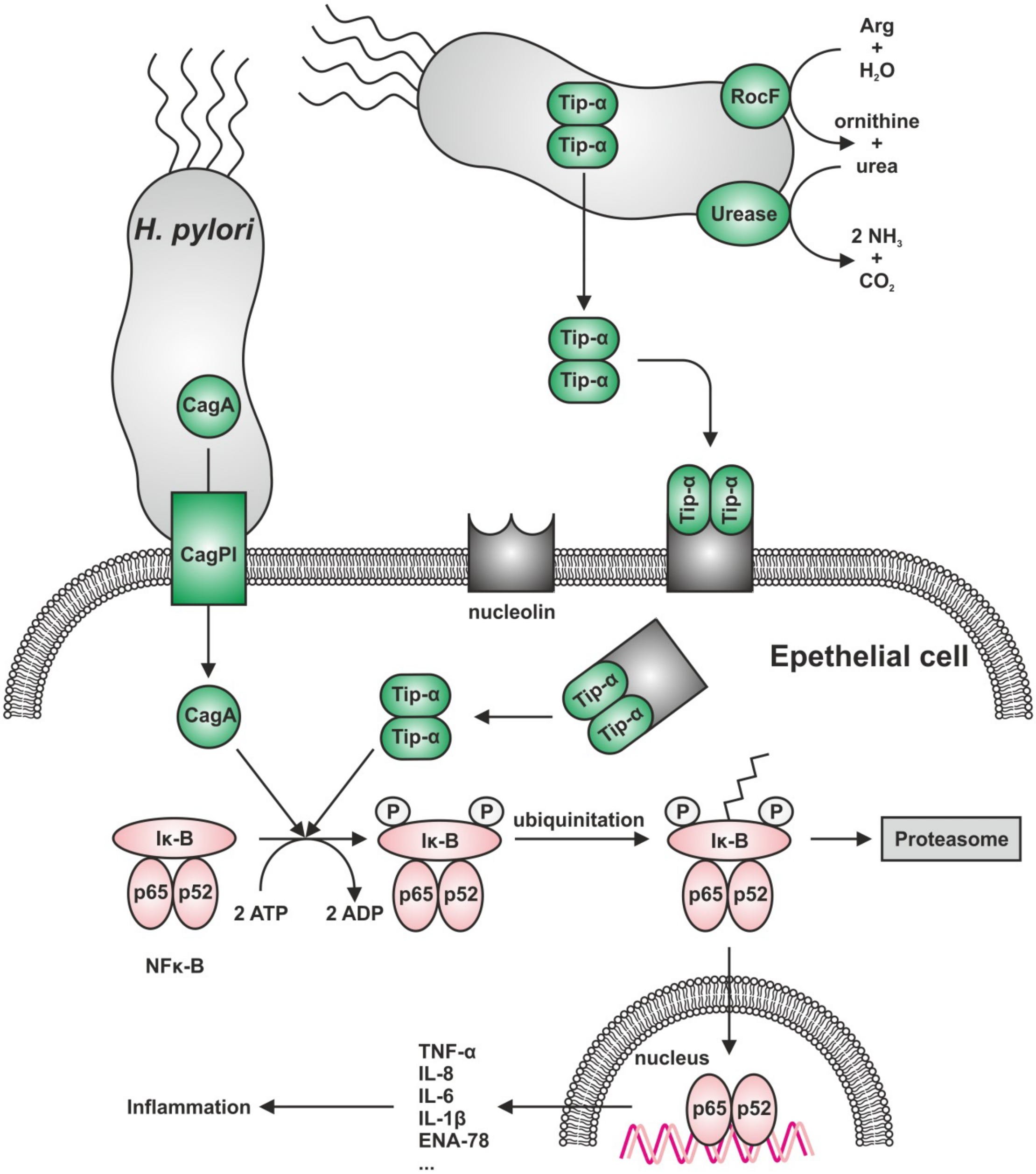


Fig. 3-6: Inflammation cascade in response to *H. pylori* infection. CagA is translocated into the epithelial cells by the type IV secretion system which includes several cagPI proteins. Tip-α is shuttled by nucleolin into the epithelial cells. Both CagA and Tip-α activate NFκ-B which gets phosphorylated. The phosphorylated Iκ-B subunit is subsequently ubiquinylated and digested by the proteasome. The released p65-p52 complex enters the nucleus and promotes the transcription of several pro-inflammatory genes. The arginase RocF catalyzes the degradation of arginine to ornithine and urea. Urea is further converted to ammonia and carbon dioxide by different ureases. Green colored proteins are down-regulated in coccoid cells. These results suggest that the inflammation cascade in gastric epithelial cells is strongly attenuated when *H. pylori* occurs in its coccoid morphology.

Conclusions and future perspectives

In conclusion, we show that SILAC is well suited to investigate changes in protein expression of *H. pylori*. The SILAC approach for *H. pylori* allows high proteome coverage and excellent quantification accuracy, especially between biological replicates of up to three different treatments. Thus, this method enables new possibilities in the research of *H. pylori*. In order to obtain the highest possible quantification rate and accuracy, we recommend enrichment of low molecular weight proteins and the application of multiple proteases. The established SILAC method for *H. pylori* could be further improved by the application of isotopic labeled leucine in addition to arginine. As a result, the number of protein quantifications as well as the quantification accuracy could be increased.

Our study illustrates that crucial processes for cell division, the infectivity and colonization efficiency of *H. pylori* are diminished in its coccoid phenotype. There is strong evidence that the flagellar assembly of the coccoid morphology is partly degraded. Additionally, down-regulation of several proteins involved in chemotaxis suggests that the coccoid morphology of *H. pylori* is a non-chemotatic phenotype with reduced ability for colonization and infection of gastric epithelial cells. Reduced expression of the arginase RocF and several cag pathogenicity island proteins including CagA elucidate the strongly decreased infectivity of coccoid *H. pylori*. Lower Tip-α expression in coccoid cells can also be associated with its reduced colonization efficiency. Furthermore, diminishment of essential cell functions like DNA replication and transcription exhibit the loss of cell growth of coccoid cells.

Based on the established SILAC protocol, one might also think about co-cultures with human epithelial cells. This would allow relative quantification of both the host and the pathogen proteome within one experiment. Consequently, this would give the opportunity to study the mutual influences of host-cells and *H. pylori* on proteome level within one shotgun approach. The influence of new drugs for *H. pylori* eradication could also be tested by SILAC studies to reveal the underlying effects.

4 Discussion and conclusion

Different methods for improved identification and quantification rates in proteomics were developed. These methods represent the pillars upon which the quantitative proteomic study of *H. pylori* is based (Fig. 4-1):

(i) The improvement of the coverage in proteomic studies with focus on LMW proteins [3]

(ii) The refinement of protein databases by RNAcode predictions [10] and proteogenomics [4]

(iii) The utilization of SILAC for quantitative proteomics [5]

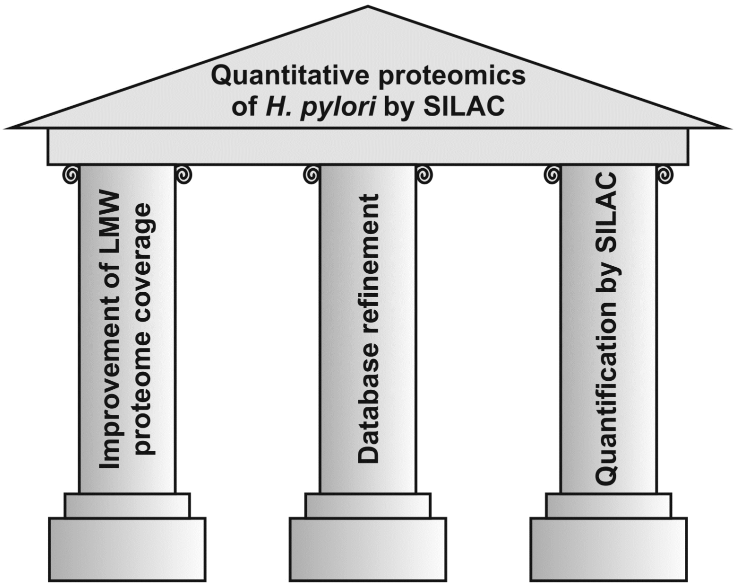


Fig. 4-1: Systematic structure of the thesis.

4.1 Improvement of identification rates and quantification rates in proteomics

The greatest challenge of proteomics is the immense dynamic concentration range of proteins in biological samples. The human plasma e.g. has a dynamic range of more than ten orders of magnitude [7]. Albumin is the most abundant protein in human plasma with 35-50 mg/ml whereas the concentration of interleukins and chemokines is commonly below 10 pg/ml [7]. LC-MS/MS typically cover a dynamic range of two to four orders of magnitude [7, 41]. Therefore, fractionation of proteins and/or peptides prior to LC-MS analysis is essential for proteomics.

Here, the focus was placed on the enrichment, fractionation and improved MS-based identification of LMW proteins. Biologically important proteins such as interleukines and chemokines e.g. have an average MW of 16 kDa (human according to UniProt database). Additionally, LMW proteins are harder to detect by LC-MS/MS analysis due to the lower number of peptides that are generated by proteolytic digestion. Hence, it is reasonable to enhance the detection of LMW proteins by specific enrichment and fractionation.

4.1.1. Enrichment and fractionation of LMW proteins

Three different strategies were developed to improve the coverage of LMW proteins:

(i) Depletion of proteins larger than 50 kDa with subsequent precipitation and separation on a tricine SDS-PAGE [3] (chapter 3.2)
(ii) Size exclusion chromatography for combined enrichment and fractionation of LMW proteins [4] (chapter 3.3)
(iii) SDS-PAGE elution fractionation of proteins below 50 kDa (chapter 3.6)

Tricine SDS-PAGE

The first strategy was based on the protocol of Klein *et al.* [6] with slight modifications. Briefly, filters with a molecular weight cut-off of 50 kDa instead of 100 kDa were used [3]. The precipitated samples were subjected to either direct in-solution digestion or 20% tricine SDS-PAGE fractionation with subsequent in-gel digestion. The tricine buffer system enables separation of proteins below 13 kDa that would migrate together with SDS in the tris-glycine buffer system [106, 107].

In-solution digested samples were analyzed by a 150 min gradient (2-40% ACN; run time 170 min) whereas a 30 min gradient (2-40% ACN; run time 50 min) was applied for the nine gel fractions. The gel-based approach with trypsin identified on average 221 proteins below 25 kDa whereas 172 protein identifications were received on average by the shotgun LC-MS approach with CID. Thus, 28% additional proteins could be identified by the gel-based fractionation. Nevertheless, for an optimized identification of LMW proteins, measurement time still has to be increased from 170 min to 450 min.

However, the tricine SDS-PAGE strategy has some drawbacks. The LMW enrichment and fractionation is hard to reproduce due to many manual steps. Large sample amounts were necessary to gain enough material for subsequent SDS tricine PAGE. Only 1-2% of the original protein amount were recovered after enrichment and precipitation. Hence, 1-2 mg proteins are necessary to gain a reasonable amount for a GeLC-MS analysis. It can be assumed that filtration and precipitation lead to severe sample losses. Moreover, peptide recovery after proteolytic in-gel digestions have shown to vary between 70% and 90% [113]. Protein losses during tricine SDS-PAGE might also be a problem [6]. Additionally, sample preparation and fractionation is very time-consuming.

SEC fractionation

To overcome the limitations of LMW enrichment with subsequent protein separation on a tricine SDS-PAGE, enrichment and fractionation were combined in one step by using SEC [4]. The Phenomenex Biosep S-2000 (ID 4.6 mm, length 30 cm) SEC column was used since it was designed for the separation of proteins between 0.5 and 100 kDa under denaturating conditions. SEC is known to have high protein recovery rates. It has been shown that protein recovery of more than 90% can be achieved by SEC [283, 284]. Four protein fractions of a *H. pylori* cell lysate below 25 kDa were collected by SEC. These fractions were subjected to proteolysis by trypsin, AspN and LysC. This method permitted the identification of 18% additional proteins below 17 kDa in comparison to an extensive fractionated GeLC-MS approach with 20 fractions [4].

GELFREE separation

The SDS-PAGE gel elution fractionation with the GELFREE device in combination with filter aided sample preparation (FASP) [142] for proteolysis was chosen as final strategy for the quantitative proteomic study of *H. pylori*. This device enables high reproducible enrichment and fractionation of up to eight samples in one run. The 12% cartridges facilitate protein separation between 10 and 50 kDa. All proteins below 10 kDa elute with the sample breakthrough in the first fraction. A method with five fractions was developed and the efficiency was verified by the analysis of a proteolytic *E. coli* digestion. The five GELFREE fractions were analyzed after tryptic FASP digestion by a 110 min LC-MS method, whereas a 220 min LC-MS method was applied for the non-fractionated reference sample. Overall, 419 proteins could be identified below 25 kDa which is comparable to the tricine SDS-PAGE strategy [3] (best replicate: 369 proteins for tryptic digestion, nine fractions, overall measuring time 450 min). The fractionation gained in 86% more protein identifications below 50 kDa (+149% < 25 kDa) than the in-solution sample.

The GELFREE device offers fractionation and enrichment of LMW proteins in parallel for up to eight samples within 131 min. Therefore, this method is less time-consuming and more reproducible than the SEC fractionation and the tricine SDS-PAGE strategy. Therefore, the decision was made to use the GELFREE device for the quantitative proteomic study of *H. pylori*. The application of GELFREE separation of LMW proteins offered on average 110.3 (+11.4%) additional protein identifications and 131 (+20.3%) quantifications (Tab. 4-1). The number of quantified proteins in at least three out of four biological samples was increased by 136 (+22.4%). Among them, 65 quantified proteins were below 25 kDa, representing an improvement of 49.2% for this LMW range.

Tab. 4-1: Improvement of the SILAC analysis of *H. pylori* by the additional application of AspN. The values are averaged over four biological replicates (except last row).

	GeLC-MS	GeLC-MS + GELFREE	Gain	Relative gain
Unique peptides	13518.5	17282.8	3764.3	+27.8%
Ratio counts	12435.25	20637.5	8202.3	+66.0%
Protein identifications	968.5	1078.8	110.3	+11.4%
Average sequence coverage	34.5%	42.7%	8.2%	+23.7%
Protein quantifications	644.8	775.8	131.0	+20.3%
Quantified proteins in at least 3 out of 4 biological samples	607	743	136.0	+22.4%

Comparison of LMW protein enrichment and fractionation methods

In conclusion, it has been proven that enrichment and fractionation of LMW proteins offers significantly increased protein identifications and quantifications. Tricine SDS-PAGE fractionation shows the highest resolving power among the three tested strategies in the LMW range. Nevertheless, the separation is hard to reproduce, and sample preparation is time-consuming. Additionally, proteins might be lost during SDS-PAGE and in-gel digestion. The GELFREE approach is the best strategy in means of reproducibility and speed. Therefore, it is the most appropriate enrichment and fractionation method for large scale proteomic studies. However, the prefabricated cartridges only facilitate separation of proteins larger than 10 kDa. For high resolution separation below 10 kDa, one might apply the first fraction from GELFREE separation onto tricine SDS-PAGE.

Protein fractionation is always a compromise of time effort, robustness, and available sample amount. Here, the GELFREE separation has shown to the best enrichment method for LMW proteins. However, the increased identification rates of 86% have to be bought by increasing the analysis time to 250%. Additionally, at least 50 µg of protein have to be applied to gain reasonable amounts for the analysis of LMW proteins. Alternatively, automated multi-dimensional LC separation of peptides can be used to enhance the identification and quantification rates [138, 285]. The application of 2D RP-RP LC, e.g., with different pH values for both dimensions has shown to be a robust method. Yang *et al.* [286] have shown that this method is able to increase peptide identifications by 1.8 fold and protein identifications by 1.6 fold. These 2D-RP-RP systems are commercially available (Waters ,UK) and show excellent reproducibility. Ultra-long monolithic columns are another possibility to increase the identification rates due to the fast mass transfer and low-backpressure. Iwasaki *et al.* [126] were able to identify 2602 proteins of *E. coli* (60% of the proteome) from 4 µg protein sample by application of a 41 h long gradient on a 350 cm long monolithic capillary column.

4.1.2. Application of multiple proteases

A further possibility for increasing identification and quantification rates in proteomic studies is the application of multiple proteases. Swaney *et al.* [8] evaluated the application of five different proteases for the same biological sample. It was shown that the separate application of AspN in addition to trypsin increased the number of protein identifications about 15% whereas cumulative protein sequence coverage was nearly 60% higher [8]. A comparison of technical replicates with replicates digested by different proteases revealed that the application of multiple proteases performed significantly better [8].

In this project, the advantage from the application of AspN [3] (chapter 3.6) or AspN and LysC [4] in addition to trypsin was used to increase the number of unique peptides, protein identifications as well as quantifications. A statistical evaluation of different commercially available proteases has shown that the application of AspN in addition to trypsin provides the highest number of unique peptides with suitable length for MS analysis [3]. Therefore AspN was chosen as the best proteases for the completion of trypsin. AspN on average increased the number of unique peptides by 75% and offered 23% additional protein identifications (67) for the LMW proteome of *E. coli* [3].

Three different proteases were used to increase the number of unique peptides as well as the proteome coverage for the proteogenomic analysis of *H. pylori* strain 26695 [4]. Among the three proteases, LysC performed best with on average 2345.5 unique peptides and 331 protein identifications (Tab. 4-2). Trypsin offered on average 2368.5 unique peptides and 262 protein identifications whereas AspN permitted the identification of 610 unique peptides and 113 proteins (Tab. 4-2). Related to trypsin, LysC and AspN provided 1312.5 (+55.4%) and 606 (+25.6%) additional unique peptides as well as 4.5 (+1.7%) and 87.5 (+33.5%) extra protein identifications, respectively (Tab. 4-2).

Tab. 4-2: Impact of the application of AspN and LysC in addition to trypsin. The values are averaged on two biological replicates.

	Trypsin	**AspN**	**AspN gain**	**AspN relative gain**	**LysC**	**LysC gain**	**LysC relative gain**
Unique Peptides	2368.5	610.0	+606.0	+25.6%	2345.5	+1312.5	+55.4%
Protein identifications	261.0	113.0	+4.5	+1.7%	331.0	+87.5	+33.5%

The superior performance of LysC over AspN might be correlated to the better ionization efficiency of peptides with a C-terminal lysine. Additionally, LysC is more robust than AspN or trypsin in respect to detergents and salts.

In the SILAC study of *H. pylori*, only arginine labeling was feasible due to non-sufficient lysine incorporation. Therefore, only arginine containing peptides grant quantitative information. Commonly, minimum two quantification features of different unique peptides are recommended for relative quantification by SILAC. A second protease had to be chosen to increase the number of unique peptides in respect to detect additional arginine-containing unique peptides which should not be redundant to those derived from tryptic proteolysis.

A statistical evaluation of AspN and ArgC was performed. ArgC would offer exclusively arginine-containing peptides, whereas AspN creates only a minimum number of similar of peptides compared to trypsin. The detectable mass range for peptides was defined from 600 to 3000 Da. An *in-silico* proteolytic digestion of the whole proteome of *H. pylori* strain 26695 was carried out and the number of arginine-containing peptides was calculated. AspN and ArgC offer 10151 and 8118 detectable arginine-containing peptides, respectively. This statistical evaluation is in accordance to the study of Swaney *et al.* [8] in which AspN outperformed ArgC by 2.6 fold when a CID-based LC-MS/MS method was applied. In comparison to a tryptic digestion, the number of additional unique arginine-containing detectable peptides is 10081 for AspN but only 5295 for ArgC.

Therefore, AspN was applied for all fractions in addition to trypsin to increase the number of protein quantifications. The application of AspN gained on average in the additional identification of 6829.8 (+65.3%) unique peptides and 50.5 (+4.7%) proteins (Tab. 4-3). This has been the basis for the quantification of 113 (+17.9%) additional proteins in at least three out of four biological samples (Tab. 4-3). In the LMW range below 25 kDa, the number of proteins quantified in at least 3 replicates was even increased by 30.3% (+47). Furthermore, the detection of more quantification features increases the accuracy of the relative protein quantification.

Tab. 4-3: Improvement of the SILAC analysis of *H. pylori* by the additional application of AspN. The values are averaged over four biological replicates (except last row).

	Trypsin	**Trypsin + AspN**	**Gain**	**Relative gain**
Unique peptides	10453.0	17282.8	6829.8	+65.3%
Ratio counts	12691.5	20637.5	7946.0	+62.6%
Protein identifications	993.8	1078.8	85.0	+8.6%
Average sequence coverage	32.7%	42.7%	10.0%	+30.6%
Protein quantifications	664.8	775.8	111.0	+16.7%
Quantified proteins in at least 3 out of 4 biological samples	630	743	113.0	+17.9%

4.1.3. Application of different MS techniques

Complementary MS techniques are known to improve the number of peptide identifications. Here, different fragmentation techniques combined with different mass analyzers were compared. Collision-induced dissociation with spectrum acquisition within the IT performed best [3]. The best results were obtained by ETD fragmentation as a complementary method in combination with CID. Both fragmentation methods showed an overlap of 71% [3]. The number of unique peptide and protein identifications was increased by 21.7% and 6.2%, respectively [3]. However, ETD and CID can be combined in different ways within on measurement.

Alternating CID and ETD acquisition of the same precursor ions can offer confirmations for peptide identifications [176]. The application of data dependent decision tree for CID and ETD according to the charge state and *m/z* value of the peptide ions might also improve the number of identifications. The usefulness of ETD also strongly depends on the mass distribution of proteolytic peptides. Typically, ETD performs better for larger peptides with higher charge states. It has been shown that ETD works better than CID for proteolytic digestions with AspN and LysC which tend to create larger peptides [8]

The impact of nanoLC-MALDI-MS in combination with nanoLC-ESI-MS was also evaluated. However, in contrast to the literature [149, 151-153, 287] the number of protein identifications were only increased by 2% using LC-MALDI-MS (results not shown). The very small improvement of LC-MALDI-MS might be related to the larger number of fragment ion spectra, as well as the superior mass accuracy of the LTQ Orbitrap XL mass spectrometer compared to the Ultraflex III MALDI TOF/TOF (Bruker Daltonics, US). Additionally, LC-MALDI-MS analyses took 12 h per fraction whereas LC-ESI-MS analyses only require 2-4 h per run. As a conclusion all analyses were exclusively performed by LC-ESI-MS due to its higher sample throughput and the modest identification gain of LC-MALDI-MS.

4.1.4. Application of multiple search engines

The application of multiple search engines for the data analysis is an effective method to increase the number of peptide identifications as well as their confidence. The utilization of freely available search engines like OMSSA or X!Tandem is a cost-effective possibility. Software packages such as Scaffold (Proteome Discovery, US), OpenMS or peptide shaker (http://peptide-shaker.googlecode.com) facilitate the integration of multiple search engine results into one analysis with the estimation of FDRs.

Here, Mascot was used in combination with Sequest [3] or X!Tandem [4]. Sequest offered 3.6% additional unique peptides as well as 3.6% more protein identifications [3]. In the proteogenomic analysis of *H. pylori*, X!Tandem was applied as second search engine in addition to Mascot, since this database search engine is directly integrated in Scaffold [4]. Therefore, this extensive databases search approach could be performed more comprehensively in less time. Overall, 3215 from 21915 identified unique peptides (+17.2%) were only identified by X!Tandem. Summing up all samples, the application of X!Tandem in addition to Mascot provided 188 (+16.5%) additional proteins with at least two unique peptides.

In conclusion, the application of multiple search engines is a time–effective method to increase the number of peptide and protein identifications. Here, the data suggests that X!Tandem performs better than Sequest. However, both search engines were applied to different data sets and the cumulative FDRs were adjusted in combination with Mascot. Nevertheless, proteomic studies always benefit from the utilization of different search engines.

However, both SILAC studies were solely based on database searches against Andromeda because Maxquant is restricted to this search engine. Andromeda has shown to perform as well as Mascot [180]. Additionally, Maxquant allows recalibration of the precursor masses according to high-scoring peptide identifications that are used as internal standards [180]. Hereby, the precursor mass tolerance can be minimized for the database search which leads to more accurate protein quantifications.

4.2 Protein database refinement

Protein databases are the basis for the analysis of MS-based proteomic studies. However, the protein sequences stored in these databases are commonly derived from gene finding software predictions on the basis of genomic data. These software tools differ in their prediction accuracy and the precision of gene boundaries.

In a study of Bakke *et al.* [57], three different automatic annotation tools were compared and showed notable differences in terms of unique gene annotations and start codon assignments. Additionally, according to the typical minimum length cut-off for ORF prediction of 300 bp [9], LMW proteins are often lacking in the annotations. Exceptions of the classical translation initiation such as leaderless mRNAs [49-55] also contribute to incomplete or erroneous gene annotation.

The database for the proteogenomic study was constructed from the NCBI database of *H. pylori* strain 26695, a six-frame translation of its genome and 18 RNAcode predictions. RNAcode neither utilizes training data sets nor species specific gene features such as open reading frame detection or ribosome binding sites to predict protein coding genes [10]. It is based on evolutionary changes in the DNA sequence [288] such as mutations, deletions or insertions that preserve the reading frame [10]. The algorithm scores segments of multiple nucleotide sequence alignments according to evolutionary changes and reports a p-value that is assigned by parameters of the extreme value distribution from randomized alignments [10].

It has been shown in the study on optimization of parameters for coverage LMW proteins, that RNAcode predictions correlate highly with proteomic data (99%) [3]. Additionally, RNAcode is well suited to validate new or corrected protein sequences of proteogenomic experiments, since predictions are not based on complete ORFs [10]. Thus, it does also predict protein coding sequence fragments which might be a result of DNA sequencing errors. This could be shown for the DNA sequencing error of the carbonic anhydrase (HP1186). The extension of the previously annotated protein sequence is supported by the RNAcode prediction 1369_0 (HP1186, supplementary figure 1, [4]). The gene that codes for the ferrous iron transporter protein A was also confirmed by RNAcode predictions (Fig. 2, [4]). Additionally, the newly identified proteins HP0619 and HP0744, as well as the corrected sequences for the proteins HP0564 and HP0760, were supported by RNAcode predictions (Suppl. Figures 2-8, [4]). Briefly summarized, the proteogenomic study of *H. pylori* unambiguously identified four proteins that were lacking in the NCBI database, and corrected the sequences for six additional proteins.

Additionally, 63 signal peptide cleavage sites were identified by a database search that permits semi-specific cleaved peptides. Signal peptide cleavage sites were validated by known characteristics of bacterial signal peptides [14], a positively charged N-terminal region that is followed by a hydrophobic region and a peptide recognition sequence of three amino acids. For Gram-negative bacteria, AXA is reported to be the predominant recognition sequence [14]. However, the predominant motif for *H. pylori* was shown to be LXA (62%) in this study, whereas only 11% of the identified signal peptides have the motif AXA.

Signal peptidase cleavage sites could also be detected by selective enrichment of N-terminal peptides prior to MS analysis. The strategy of McDonald *et al.* [289] utilizes acetylation of all primary amines which includes lysine residues as well as protein N-termini. After proteolysis, N-termini of proteolytic peptides are labeled with a biotin tag. Streptavidin is used for negative enrichment of peptides derived from protein N-termini, since these peptides are not biotinylated. Schepmoes *et al.* [290] use a similar strategy. Primary amines are acetylated in the first step. Peptides that contain free amines after proteolytic digestion are removed by amine-reactive silica-bond succinic anhydride beads. In contrast, Xu and Jeffrey [291] used a positive enrichment method for N-terminal peptides using Edmann chemistry. All amines are blocked by phenyl isothiocyanate. The first amino acid is cleaved off by addition of TFA. The generated free amine at the second amino acid is modified with a biotin tag. This enables selective enrichment of N-terminal peptides by avidin beads after proteolysis. Nevertheless, the proteogenomics study of *H. pylori* showed that identification of signal peptide cleavage sites is also possible by an un-targeted approach using high quality MS data.

Signal peptide cleavage identifications provide evidence for potentially secreted or membrane bound proteins. Additionally, signal peptides are essential for bacteria [14, 292] and are also involved in pathogenesis [293, 294]. Hence, bacterial signal peptidases are supposed to be novel targets for antibiotics [292].

Conclusively, the protein database for *H. pylori* strain 26695 was refined by the identification of new proteins and the correction of protein sequences as well as the investigation of signal peptidases. The refined database will help future proteomic studies on *H. pylori* strain 26695 to gain more information from MS data. Here, the refined database was used to increase the quality of the quantitative proteomic study on *H. pylori* strain 26695 (chapter 3.6). Furthermore, the elucidated signal peptidase specificity might help to design new drugs for the treatment of *H. pylori* infections. The developed strategy is easily applicable for the proteogenomic analysis of freely available proteomic datasets. However, this study shows that high quality MS data sets are necessary for proteogenomics.

4.3 SILAC in quantitative proteomics

A huge variety of different quantification methods is available in proteomics (chapter 2.7). In this study, a quantification method was to be selected that possesses high accuracy and that facilitates fractionation on protein level for LMW enrichment.

Label free quantification is universally applicable in quantitative proteomics. It facilitates relative quantification of all kinds of samples including those from animal experiments. However, sample fractionation and LC-MS/MS analysis have to be very reproducible to achieve quantitative results with high accuracy [39, 213]. Label free approaches have an accuracy of 10-30% rsd compared to less than 10% for metabolic labeling methods [38]. Additionally, label free approaches are more prone to errors since data analysis depends strongly on precise algorithms for peak picking, feature detection and normalization [213, 295]. Furthermore, the available LTQ Orbitrap instruments in this study usually have a cycle time of approximately three seconds at a survey scan resolution of 60,000 when six MS/MS scans are performed per cycle. As a result, the number of data points per peak are not sufficient for label free quantification. Therefore, it was decided rather to use a labeling than a label free method for quantitative proteomics to improve the accuracy as well as the reproducibility.

Fractionation of post digest ICPL, iTRAQ and TMT labeling are meant to be performed at peptide level after mixing of the differentially labeled samples. Separation of proteins prior to proteolytic digestion and chemical labeling has to be very reproducible to minimize the error. Since LMW protein enrichment and separation was optimized here, chemical labeling on peptide level was not considered in this study.

Labeling of proteins is more suited for protein fractionation. The differentially labeled biological samples are mixed at protein level prior to fractionation. Thus, reproducibility of fractionation has no effect on the quantification results. Metabolic labeling such as ^{15}N or SILAC and the chemical labeling methods ICAT and ICPL are available for protein labeling. Here, metabolic labeling is the most accurate MS based quantification method due to the early stage of sample combination [37]. However, metabolic ^{15}N labeling has shown to alter metabolite and protein levels of *E. coli* [296]. SILAC introduces less heavy isotopes than ^{15}N or ^{13}C labeling which probably leads to minor stable isotope effects. Additionally, the data analysis of SILAC studies is well automated by different software tools such as Maxquant [255], Thermo Proteome Discoverer (Thermo Scientific, US) or Mascot Distiller (Matrix Science, UK). Thus, SILAC is the quantification method that meets the chosen requirements best. Nevertheless, SILAC requires complete incorporation of labeled amino acids into proteins and cells have to be grown in minimal medium.

4.4 Influence of hyaluronan sulfation on primary dermal fibroblasts

The Transregio Collaborative Research Centre TRR67 investigates artificial extracellular matrices (aECMs) for improved wound healing of skin and bone tissue. Within the project, the main focus was placed on the effect of glycosaminoglycan (GAG) sulfation. Especially hyaluronan is well-suited to study the effects of chemically sulfated GAGs [297] since it

(i) has no sulfate groups,
(ii) possesses a regular structure of alternating N-acetylglucosamine and glucuronic acid units
(iii) is not covalently linked to proteins
(iv) is easily chemically modifiable without destroying its structure

Here, the influence of highly sulfated hyaluronan (hsHA) provided as extracellular matrix (ECM) on human primary dermal fibroblasts (dFb) was investigated on protein level by SILAC [5].

Since primary cells show high biological relevance in comparison with immortalized cell lines, the significance thresholds were evaluated by a control experiment. For this purpose, differentially labeled samples from different donors were mixed and analyzed. A $\log_2$ fold change of ± 0.5 was set as regulation threshold. The measured variation was used to estimate the false positive rate of this study. With the threshold that single proteins had to be regulated at least in three out of four biological replicates in the same direction, a false positive rate of less than 1% was assumed. Additionally, a cluster analysis with PANTHER (Protein Networks and Pathway Analysis) [298] and DAVID (Database for Annotation, Visualization and Integrated Discovery) [299] was used to determine effects based on rather clusters of proteins than individual proteins. Furthermore, regulation of single proteins was validated by western blotting.

Ten proteins associated to the ECM showed to be significantly regulated ([5], Fig. 2). Most interestingly, the ECM degrading enzymes cathepsin K (catK), matrix metalloproteinases 2 and 14 as well as the tissue inhibitor of MMPs 2 (TIMP-2) were found to down-regulated in response to hsHA. In line with these results, it has been shown that osteogenic-differentiated human mesenchymal stromal cells also have reduced expression and activity of MMP-2 in response to hsHA [300]. Chronic skin wounds have misbalance of MMPs and TIMPs that may cause fibrosis metastasis or tumor growth [15]. The inhibition of MMPs is a common strategy to treat chronic skin wounds [16-19]. Therefore, the application of hsHA might be a promising approach for the treatment of chronic skin wounds.

Furthermore, the expression of collagens type I and XII was reduced by hsHA. Especially collagen I is excessively produced in hypertrophic scar formation [20]. Collagen VI expression was increased by hsHA. This compound is produced when cells become confluent to provide an appropriate ECM environment [21]. This indicates that scar formation might be reduced in response to hsHA.

4.5 Quantitative proteomics of *H. pylori* by stable isotope labeling by amino acids in cell culture

The development of SILAC for quantitative proteomics of *H. pylori* is challenging. The SILAC study on the influence of hyaluronan sulfation was the basis for the development of a quantitative proteomics method for *H. pylori*. The acquired knowledge about the experimental design setup as well as the data analysis of SILAC studies was used to design a more complex SILAC study.

Bacterial SILAC studies are rarely used in proteomic research. Some are reported for *Bacilus subtilis* [254], *E. coli* [301], *Bifidobacterium longum* [302] and *Salmonella serovars* [253]. The greatest problem for the setup of SILAC for bacteria is amino acid autotrophy of many bacterial species. Therefore, it is hard to achieve a sufficient incorporation of isotopic labeled amino acids into proteins. Typically, growth substrates with differential isotopes of nitrogen [303-305], carbon [244] or sulfur [197, 198] are used to perform metabolic labeling of bacteria. The differential isotopes are incorporated via amino acid synthesis. Alternatively, chemical isotope labeling can be applied for quantitative proteomics of bacteria [306-309]. However, most quantitative bacterial proteomic studies are based on 2D PAGE [83, 95, 241, 308, 310, 311].

No SILAC study of *H. pylori* is published so far. Several challenges had to be solved to establish SILAC for this organism. *H. pylori* is usually cultured in brain heart infusion (BHI) medium which consist of extracts from boiled bovine or porcine brains and hearts. Therefore, a chemically defined medium had to be found that permits growth of *H. pylori* and enables the supplementation with specific isotope labeled amino acids. Additionally, this medium should have no influence on the morphology of *H. pylori*. Furthermore, incorporation of isotope labeled amino acids had to be tested.

4.5.1. Choice of medium and labeled amino acids

The Ham’s F-12 medium supplemented with FCS showed to permit growth of *H. pylori* without influencing the morphology in previous studies [22, 23]. In this study, Ham’s F12 medium was supplemented with 5% dialyzed FCS to prevent undesired introduction of free amino acids into the culture medium. As described in the literature, this medium promoted growth of *H. pylori* and had no effect on its morphology. This was also the case for the Ham’s F12 SILAC medium which was supplemented with dialyzed FCS.

The next step was to test the incorporation of selected stable isotope labeled amino acids. For this purpose, arginine and lysine were chosen. The utilization of stable isotope labeled lysine and arginine has the advantage that all tryptic peptides, except the ones derived from the protein C-termini, are differentially labeled. *H. pylori* is auxotroph for arginine but not for lysine. Nevertheless, lysine was tested for incorporation since Ham's F12 medium was commercially available without lysine and arginine. Additionally, it was thought that *H. pylori* would stop synthesis in the presence of freely available lysine.

The incorporation of lysine and arginine was tested for up to seven cell doublings. Arginine incorporation was sufficient (> 95%) after four cell doublings. However, incorporation of stable isotope labeled lysine was not sufficient. Even when increasing the lysine concentration four fold to the original recipe, a maximal incorporation of 80% was achieved. It has been shown that incomplete amino acid incorporation can be mathematically corrected [312]. However, such an approach is labor-intensive and error-prone. Therefore, the main experiment was only performed with labeled arginine.

4.5.2. Influence of morphology on the proteome of *H. pylori*

A triplex SILAC design with light (6 ^{12}C, 4 ^{14}N), medium (6 ^{13}C, 4 ^{14}N) and heavy (6 ^{13}C, 4 ^{15}N) labeled arginine was designed. Spiral (medium) *H. pylori* cells were compared with the coccoid form (light) as well as the HPnc5490 sRNA deletion mutant (heavy).

The coccoid morphology showed to have reduced expression of proteins related to cell division, transcription and translation. This observation is in accordance with the termination of cell growth when *H. pylori* becomes coccoid. Proteins involved in chemotaxis and the flagellar assembly were also lower expressed by the coccoid form. These findings suggest that coccoid *H. pylori* loses the ability to target the antrum by chemotaxis and flagellar motion which consistently leads to decreased host colonization efficiency. Additionally, several proteins of the type four secretion system, as well as the virulence factors CagA, RocF and Tip-α, were found to be significantly down-regulated in coccoid cells. The reduced expression of major virulent factors might explain the attenuated infectivity. In contrast, the vacuolating cytotoxin VacA as well as the infectivity related adherence factor OipA were more abundant in coccoid cells. A possible explanation would be that coccoid *H. pylori* stores VacA and secrets it when there are suitable conditions for the retransmission into its virulent form. It is also known that VacA is stronger expressed in response to iron deficiency [313]. Furthermore, several outer membrane proteins showed to be higher expressed in coccoid cells. This might explain adherence to the palatine tonsils [64].

In conclusion, it has been shown that infectivity and colonization efficiency are attenuated in coccoid *H. pylori* cells due to down-regulation of important proteins involved in chemotaxis and infection processes. The finding that several outer membrane proteins are up-regulated for coccoid cells might indicate that these cells are rather a dormant cell stage than a preliminary stage of cell death.

4.5.3. Opportunities of SILAC for further studies of *H. pylori*

The application SILAC for *H. pylori* that was developed in this thesis offers new possibilities for the research of *H. pylori*. Especially SILAC co-cultures with gastric epithelial cells could be a promising approach to reveal major mechanisms of the pathogenesis of *H. pylori*. The effect of distinct proteins could be further evaluated by RNA interference for the knockdown of single genes in combination with SILAC. Additionally, the effect of different stimuli such as oxidative or acidic stress, as well as the treatment of *H. pylori* with antibiotics could be tested with high proteome coverage. Investigations on antibiotic resistance of *H. pylori* could be another interesting project.

SILAC could be also combined with transcriptome or metabolome studies to monitor biological reactions in more detail. It also enables the application of hyperplexing in combination with TMT or iTRAQ labels. Recently, Dephoure and Gygi showed the application of a 3-plex SILAC approach in combination with a 6-plex TMT labeling [314]. Hereby, protein abundance changes could be examined for 18 samples simultaneously. This new developed SILAC method for *H. pylori* could be further optimized by the utilization of isotopic labeled leucine in combination with arginine. Consequently, the higher amount of quantification features would lead to more protein quantifications, as well as improved quantification accuracy.

4.6 Conclusion

In conclusion, it has been shown that enrichment and fractionation of LMW proteins offers significantly increased identification and quantification rates. The application of the GELFREE system in combination with the FASP method for in-solution digestion demonstrated the best performance. It offered the best identification rates, the highest reproducibility, as well as the easiest applicability among the three developed strategies for improved LMW protein coverage. This makes the GELFREE LMW protein strategy feasible for large scale proteomic studies.

The applications of multiple proteases, as well as the MS data analysis by multiple search engines, provide further improvements for peptide and protein identification rates. The utilization of an additional protease doubles the measurement time whereas more extensive data analysis solely increases the computing time. Therefore, multiple proteases should be applied if the best possible protein sequence coverage should be achieved and measuring time is not limited. In contrast, the application of multiple search engines offers general but smaller advantages for every proteomic project.

The objective of the proteogenomic analysis was to refine the protein database of *H. pylori*. It has been shown that existing database entries could be corrected and new protein sequences were identified. Besides protein sequence annotation refinements, signal peptide cleavage sites for 63 proteins and the predominant recognition sequence for signal peptidase I were determined. Signal peptide cleavage plays an important role in bacterial pathogenesis through its contribution in secretion of virulent proteins. Therefore, the results might have an impact on pathogenesis research of *H. pylori*.

In this thesis, SILAC was chosen to be the best suited method for quantitative proteomic studies. SILAC demonstrates the best accuracy and enables fractionation on protein and peptide level. In the first SILAC study, the effects of high-sulfated hyaluronan on primary dFb were evaluated. High-sulfated hyaluronan has shown to be a promising artificial ECM for improved wound healing of skin tissue. It modulates the ECM production of dFb and reduces the production of MMP-2 and MMP-14 that are known to be highly expressed in chronic skin wounds.

Based on the knowledge gained in this study, SILAC was established for relative protein quantification of *H. pylori*. An appropriate culture medium was found for this analysis and isotope labeled lysine as well as arginine incorporation was tested. Arginine provided a complete incorporation, whereas no sufficient incorporation could be achieved for lysine. Methods that were developed in the previous studies were applied for the SILAC study of *H. pylori*. Selective enrichment and separation of LMW proteins and the proteolysis with AspN in addition to trypsin were applied to increase the number of protein identifications. The established SILAC workflow allowed the identification of 1143 proteins of which 743 proteins were quantified. This represents 72% and 47% of the *H. pylori* proteome, respectively.

In a first application, it was shown that major differences between the spiral and coccoid morphology of *H. pylori* could be evaluated. Down-regulation of several proteins involved in pathogenicity, chemotaxis, cell division, as well as transcription, indicate the attenuated infectivity and colonization efficiency of coccid *H. pylori*. The SILAC workflow for *H. pylori* could now be applied to comprehensively study the effect of different kind of stimuli such as antibiotics, as well as acidic or oxidative stress.

5 References

[1] Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, et al. Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. Electrophoresis. 1995;16:1090-4.

[2] Rockstroh M, Müller S, Jende C, Kerzhner A, Bergen Mv, Tomm JM. Cell fractionation - an important tool for compartment proteomics2010.

[3] Müller SA, Kohajda T, Findeiss S, Stadler PF, Washietl S, Kellis M, et al. Optimization of parameters for coverage of low molecular weight proteins. Anal Bioanal Chem. 2010;398:2867-81.

[4] Müller SA, Findeiß S, Pernitzsch SR, Wissenbach DK, Stadler PF, Hofacker IL, et al. Identification of new protein coding sequences and signal peptidase cleavage sites of Helicobacter pylori strain 26695 by proteogenomics. J Proteomics. 2013;86:27-42.

[5] Müller SA, van der Smissen A, von Feilitzsch M, Anderegg U, Kalkhof S, von Bergen M. Quantitative proteomics reveals altered expression of extracellular matrix related proteins of human primary dermal fibroblasts in response to sulfated hyaluronan and collagen applied as artificial extracellular matrix. Journal of materials science Materials in medicine. 2012;23:3053-65.

[6] Klein C, Aivaliotis M, Olsen JV, Falb M, Besir H, Scheffer B, et al. The low molecular weight proteome of Halobacterium salinarum. Journal of proteome research. 2007;6:1510-8.

[7] Anderson NL, Anderson NG. The Human Plasma Proteome: History, Character, and Diagnostic Prospects. Molecular & Cellular Proteomics. 2002;1:845-67.

[8] Swaney DL, Wenger CD, Coon JJ. Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics. Journal of proteome research. 2010;9:1323-9.

[9] Dinger ME, Pang KC, Mercer TR, Mattick JS. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS computational biology. 2008;4:e1000176.

[10] Washietl S, Findeiss S, Muller SA, Kalkhof S, von Bergen M, Hofacker IL, et al. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA. 2011;17:578-94.

[11] Tsugawa H, Suzuki H, Matsuzaki J, Hirata K, Hibi T. FecA1, a bacterial iron transporter, determines the survival of Helicobacter pylori in the stomach. Free Radical Biology and Medicine. 2012;52:1003-10.

[12] Sarkar M, Maganti L, Ghoshal N, Dutta C. In silico quest for putative drug targets in *Helicobacter pylori* HPAG1: molecular modeling of candidate enzymes from lipopolysaccharide biosynthesis pathway. Journal of Molecular Modeling. 2012;18:1855-66.

[13] Specht M, Schätzle S, Graumann PL, Waidner B. Helicobacter pyloriPossesses Four Coiled-Coil-Rich Proteins That Form Extended Filamentous Structures and Control Cell Shape and Motility. Journal of Bacteriology. 2011;193:4523-30.

[14] Paetzel M, Karla A, Strynadka NCJ, Dalbey RE. Signal peptidases. Chem Rev. 2002;102:4549-79.

[15] Schultz GS, Ladwig G, Wysocki A. Extracellular matrix: review of its roles in acute and chronic wounds World Wide Wounds. 2005.

[16] Cullen B, Smith R, McCulloch E, Silcock D, Morrison L. Mechanism of action of PROMOGRAN, a protease modulating matrix, for the treatment of diabetic foot ulcers. Wound Repair Regen. 2002;10:16-25.

[17] Karim RB, Brito BLR, Dutrieux RP, Lassance FP, Hage JJ. MMP-2 Assessment as an Indicator of Wound Healing: A Feasibility Study. Adv Skin Wound Care. 2006;19:324-7.

[18] Shi L, Ermis R, Kiedaisch B, Carson D. The Effect of Various Wound Dressings on the Activity of Debriding Enzymes. Adv Skin Wound Care. 2010;23:456-62.

[19] van den Berg AJ, Halkes SB, van Ufford HC, Hoekstra MJ, Beukelman CJ. A novel formulation of metal ions and citric acid reduces reactive oxygen species in vitro. Journal of wound care. 2003;12:413-8.

[20] Harrop AR, Ghahary A, Scott PG, Forsyth N, Uji-Friedland RTA, Tredget EE. Regulation of Collagen Synthesis and mRNA Expression in Normal and Hypertrophic Scar Fibroblasts in Vitro by Interferon-γ. Journal of Surgical Research. 1995;58:471-7.

[21] Hatamochi A, Aumailley M, Mauch C, Chu ML, Timpl R, Krieg T. Regulation of collagen VI expression in fibroblasts. Effects of cell density, cell-matrix interactions, and chemical transformation. The Journal of biological chemistry. 1989;264:3494-9.

[22] Testerman TL, McGee DJ, Mobley HL. Helicobacter pylori growth and urease detection in the chemically defined medium Ham's F-12 nutrient mixture. J Clin Microbiol. 2001;39:3842-50.

[23] Testerman TL, Conn PB, Mobley HL, McGee DJ. Nutritional requirements and antibiotic resistance patterns of Helicobacter species in chemically defined media. J Clin Microbiol. 2006;44:1650-8.

[24] Saito N, Konishi K, Sato F, Kato M, Takeda H, Sugiyama T, et al. Plural transformation-processes from spiral to coccoid Helicobacter pylori and its viability. J Infect. 2003;46:49-55.

[25] Eaton KA, Suerbaum S, Josenhans C, Krakowka S. Colonization of gnotobiotic piglets by Helicobacter pylori deficient in two flagellin genes. Infect Immun. 1996;64:2445-8.

[26] She FF, Lin JY, Liu JY, Huang C, Su DH. Virulence of water-induced coccoid Helicobacter pylori and its experimental infection in mice. World journal of gastroenterology : WJG. 2003;9:516-20.

[27] Kalkhof S, Haehn S, Paulsson M, Smyth N, Meiler J, Sinz A. Computational modeling of laminin N-terminal domains using sparse distance constraints from disulfide bonds and chemical cross-linking. Proteins: Structure, Function, and Bioinformatics. 2010;78:3409-27.

[28] Stengel F, Aebersold R, Robinson CV. Joining Forces: Integrating Proteomics and Cross-linking with the Mass Spectrometry of Intact Complexes. Molecular & Cellular Proteomics. 2012;11.

[29] Serpa JJ, Parker CE, Petrotchenko EV, Han J, Pan J, Borchers CH. Mass spectrometry-based structural proteomics. Eur J Mass Spectrom (Chichester, Eng). 2012;18:251-67.

[30] Temme C, Zhang L, Kremmer E, Ihling C, Chartier A, Sinz A, et al. Subunits of the Drosophila CCR4-NOT complex and their roles in mRNA deadenylation. RNA. 2010;16:1356-70.

[31] Domanski D, Percy AJ, Yang J, Chambers AG, Hill JS, Freue GVC, et al. MRM-based multiplexed quantitation of 67 putative cardiovascular disease biomarkers in human plasma. Proteomics. 2012;12:1222-43.

[32] Dautel F, Kalkhof S, Trump S, Michaelson J, Beyer A, Lehmann I, et al. DIGE-Based Protein Expression Analysis of B[a]P-Exposed Hepatoma Cells Reveals a Complex Stress Response Including Alterations in Oxidative Stress, Cell Cycle Control, and Cytoskeleton Motility at Toxic and Subacute Concentrations. Journal of proteome research. 2010;10:379-93.

[33] Yu LR. Pharmacoproteomics and toxicoproteomics: the field of dreams. J Proteomics. 2011;74:2549-53.

[34] Taubert M, Vogt C, Wubet T, Kleinsteuber S, Tarkka MT, Harms H, et al. Protein-SIP enables time-resolved analysis of the carbon flux in a sulfate-reducing, benzene-degrading microbial consortium. ISME J. 2012;6:2291-301.

[35] Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, et al. Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature. 2011;480:254-8.

[36] Westermeier R, Schickle H. The current state of the art in high-resolution two-dimensional electrophoresis. Archives of physiology and biochemistry. 2009;115:279-85.

[37] Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem. 2007;389:1017-31.

[38] Schulze WX, Usadel B. Quantitation in mass-spectrometry-based proteomics. Annual review of plant biology. 2010;61:491-516.

[39] Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, et al. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nature biotechnology. 2010;28:83-9.

[40] Plumb R, Castro-Perez J, Granger J, Beattie I, Joncour K, Wright A. Ultra-performance liquid chromatography coupled to quadrupole-orthogonal time-of-flight mass spectrometry. Rapid Communications in Mass Spectrometry. 2004;18:2331-7.

[41] Zubarev RA. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics. 2013;13:723-6.

[42] Finamore F, Pieroni L, Ronci M, Marzano V, Mortera SL, Romano M, et al. Proteomics investigation of human platelets by shotgun nUPLC-MSE and 2DE experimental strategies: a comparative study. Blood Transf. 2010;8:S140-S8.

[43] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004;32:D115-9.

[44] Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005;33:D501-4.

[45] Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. 2005;33:W451-4.

[46] Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007;23:673-9.

[47] Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 2012;40:D115-22.

[48] Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC genomics. 2008;9:75.

[49] Malys N, McCarthy JE. Translation initiation: variations in the mechanism can be anticipated. Cellular and molecular life sciences : CMLS. 2011;68:991-1003.

[50] Hering O, Brenneis M, Beer J, Suess B, Soppa J. A novel mechanism for translation initiation operates in haloarchaea. Molecular Microbiology. 2009;71:1451-63.

[51] Benelli D, Maone E, Londei P. Two different mechanisms for ribosome/mRNA interaction in archaeal translation initiation. Mol Microbiol. 2003;50:635-43.

[52] Zheng X, Hu G-Q, She Z-S, Zhu H. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. BMC genomics. 2011;12:361.

[53] Vesper O, Amitai S, Belitsky M, Byrgazov K, Kaberdina Anna C, Engelberg-Kulka H, et al. Selective Translation of Leaderless mRNAs by Specialized Ribosomes Generated by MazF in Escherichia coli. Cell. 2011;147:147-57.

[54] Moll I, Grill S, Gualerzi CO, Bläsi U. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. Molecular Microbiology. 2002;43:239-46.

[55] Christian BE, Spremulli LL. Preferential Selection of the 5′-Terminal Start Codon on Leaderless mRNAs by Mammalian Mitochondrial Ribosomes. Journal of Biological Chemistry. 2010;285:28379-86.

[56] Bonizzoni P, Rizzi R, Pesole G. Computational methods for alternative splicing prediction. Briefings in Functional Genomics & Proteomics. 2006;5:46-51.

[57] Bakke P, Carney N, Deloache W, Gearing M, Ingvorsen K, Lotz M, et al. Evaluation of three automated genome annotations for Halorhabdus utahensis. PloS one. 2009;4:e6291.

[58] Krienitz W. Ueber das Auftreten von Spirochäten verschiedener Form im Mageninhalt bei Carcinoma ventriculi. Dtsch med Wochenschr. 1906;32:872-.

[59] Marshall BJ, Warren JR. Unidentified Curved Bacilli on Gastric Epithelium in Active Chronic Gastritis. Lancet. 1983;1:1273-5.

[60] Marshall BJ, Warren JR. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet. 1984;1:1311-5.

[61] Jonsson R. The Nobel prize in physiology or medicine 2005. Scandinavian journal of immunology. 2005;62:497.

[62] Muhammad JS, Zaidi SF, Sugiyama T. Epidemiological ins and outs of helicobacter pylori: a review. JPMA The Journal of the Pakistan Medical Association. 2012;62:955-9.

[63] Azevedo NF, Huntington J, Goodman KJ. The Epidemiology of Helicobacter pylori and Public Health Implications. Helicobacter. 2009;14:1-7.

[64] Kusano K, Inokuchi A, Fujimoto K, Miyamoto H, Tokunaga O, Kuratomi Y, et al. Coccoid Helicobacter pylori exists in the palatine tonsils of patients with IgA nephropathy. J Gastroenterol. 2010;45:406-12.

[65] Burgers R, Schneider-Brachert W, Reischl U, Behr A, Hiller KA, Lehn N, et al. Helicobacter pylori in human oral cavity and stomach. European journal of oral sciences. 2008;116:297-304.

[66] Young KA, Allaker RP, Hardie JM. Morphological analysis of Helicobacter pylori from gastric biopsies and dental plaque by scanning electron microscopy. Oral microbiology and immunology. 2001;16:178-81.

[67] Voytek MA, Ashen JB, Fogarty LR, Kirshtein JD, Landa ER. Detection of Helicobacter pylori and fecal indicator bacteria in five North American rivers. Journal of water and health. 2005;3:405-22.

[68] Fujimura S, Kato S, Watanabe A. Water source as a Helicobacter pylori transmission route: a 3-year follow-up study of Japanese children living in a unique district. J Med Microbiol. 2008;57:909-10.

[69] Nurgalieva ZZ, Malaty HM, Graham DY, Almuchambetova R, Machmudova A, Kapsultanova D, et al. Helicobacter pylori infection in Kazakhstan: effect of water source and household hygiene. The American Journal of Tropical Medicine and Hygiene. 2002;67:201-6.

[70] Kusters JG, van Vliet AHM, Kuipers EJ. Pathogenesis of Helicobacter pylori Infection. Clinical Microbiology Reviews. 2006;19:449-90.

[71] Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. Cell. 2010;140:883-99.

[72] Karin M, Greten FR. NF-[kappa]B: linking inflammation and immunity to cancer development and progression. Nature reviews Immunology. 2005;5:749-59.

[73] Kumar Pachathundikandi S, Brandt S, Madassery J, Backert S. Induction of TLR-2 and TLR-5 expression by Helicobacter pylori switches cagPAI-dependent signalling leading to the secretion of IL-8 and TNF-alpha. PloS one. 2011;6:e19614.

[74] Tang CL, Hao B, Zhang GX, Shi RH, Cheng WF. Helicobacter pylori tumor necrosis factor-alpha inducing protein promotes cytokine expression via nuclear factor-kappaB. World journal of gastroenterology : WJG. 2013;19:399-403.

[75] Radin JN, González-Rivera C, Ivie SE, McClain MS, Cover TL. Helicobacter pylori VacA Induces Programmed Necrosis in Gastric Epithelial Cells. Infection and Immunity. 2011;79:2535-43.

[76] Foryst-Ludwig A, Naumann M. p21-activated kinase 1 activates the nuclear factor kappa B (NF-kappa B)-inducing kinase-Ikappa B kinases NF-kappa B pathway and proinflammatory cytokines in Helicobacter pylori infection. The Journal of biological chemistry. 2000;275:39779-85.

[77] Rieder G, Einsiedl W, Hatz RA, Stolte M, Enders GA, Walz A. Comparison of CXC chemokines ENA-78 and interleukin-8 expression in Helicobacter pylori-associated gastritis. Infect Immun. 2001;69:81-8.

[78] Fuentes-Panana E, Camorlinga-Ponce M, Maldonado-Bernal C. Infection, inflammation and gastric cancer. Salud Publica Mexico. 2009;51:427-33.

[79] Fischer W, Puls J, Buhrdorf R, Gebert B, Odenbreit S, Haas R. Systematic mutagenesis of the Helicobacter pylori cag pathogenicity island: essential genes for CagA translocation in host cells and induction of interleukin-8. Mol Microbiol. 2001;42:1337-48.

[80] Hannelien V, Karel G, Jo VD, Sofie S. The role of CXC chemokines in the transition of chronic inflammation to esophageal and gastric cancer. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2012;1825:117-29.

[81] Chu W-M. Tumor necrosis factor. Cancer letters. 2013;328:222-5.

[82] Goldstein NS. Chronic Inactive Gastritis and Coccoid Helicobacter pylori in Patients Treated for Gastroesophageal Reflux Disease or With H pylori Eradication Therapy. American Journal of Clinical Pathology. 2002;118:719-26.

[83] Zeng H, Guo G, Mao X, Tong W, Zou Q. Proteomic Insights into Helicobacter pylori Coccoid Forms Under Oxidative Stress. Current Microbiology. 2008;57:281-6.

[84] Mizoguchi H, Fujioka T, Kishi K, Nishizono A, Kodama R, Nasu M. Diversity in Protein Synthesis and Viability ofHelicobacter pylori Coccoid Forms in Response to Various Stimuli. Infection and Immunity. 1998;66:5555-60.

[85] Nilius M, Strohle A, Bode G, Malfertheiner P. Coccoid like forms (CLF) of Helicobacter pylori. Enzyme activity and antigenicity. Zentralblatt fur Bakteriologie : international journal of medical microbiology. 1993;280:259-72.

[86] Mouery K, Rader BA, Gaynor EC, Guillemin K. The Stringent Response Is Required for Helicobacter pylori Survival of Stationary Phase, Exposure to Acid, and Aerobic Shock. Journal of Bacteriology. 2006;188:5494-500.

[87] Cole SP, Cirillo D, Kagnoff MF, Guiney DG, Eckmann L. Coccoid and spiral Helicobacter pylori differ in their abilities to adhere to gastric epithelial cells and induce interleukin-8 secretion. Infect Immun. 1997;65:843-6.

[88] Osaki T, Yamaguchi H, Taguchi H, Fukada M, Kawakami H, Hirano H, et al. Interleukin-8 induction and adhesion of the coccoid form of Helicobacter pylori. J Med Microbiol. 2002;51:295-9.

[89] Liu ZF, Chen CY, Tang W, Zhang JY, Gong YQ, Jia JH. Gene-expression profiles in gastric epithelial cells stimulated with spiral and coccoid Helicobacter pylori. J Med Microbiol. 2006;55:1009-15.

[90] Sycuro LK, Wyckoff TJ, Biboy J, Born P, Pincus Z, Vollmer W, et al. Multiple Peptidoglycan Modification Networks Modulate <italic>Helicobacter pylori's</italic> Cell Shape, Motility, and Colonization Potential. PLoS Pathog. 2012;8:e1002603.

[91] Shao C, Zhang Q, Tang W, Qu W, Zhou Y, Sun Y, et al. The changes of proteomes components of Helicobacter pylori in response to acid stress without urea. J Microbiol. 2008;46:331-7.

[92] Chuang MH, Wu MS, Lin JT, Chiou SH. Proteomic analysis of proteins expressed by Helicobacter pylori under oxidative stress. Proteomics. 2005;5:3895-901.

[93] Lee HW, Choe YH, Kim DK, Jung SY, Lee NG. Proteomic analysis of a ferric uptake regulator mutant of Helicobacter pylori: regulation of Helicobacter pylori gene expression by ferric uptake regulator and iron. Proteomics. 2004;4:2014-27.

[94] Choi YW, Park SA, Lee HW, Lee NG. Alteration of growth-phase-dependent protein regulation by a fur mutation in Helicobacter pylori. FEMS Microbiol Lett. 2009;294:102-10.

[95] Bumann D, Habibi H, Kan B, Schmid M, Goosmann C, Brinkmann V, et al. Lack of stage-specific proteins in coccoid Helicobacter pylori cells. Infect Immun. 2004;72:6738-42.

[96] Uwins C, Deitrich C, Argo E, Stewart E, Davidson I, Cash P. Growth-induced changes in the proteome of Helicobacter pylori. Electrophoresis. 2006;27:1136-46.

[97] Choi YW, Park SA, Lee HW, Kim DS, Lee NG. Analysis of growth phase-dependent proteome profiles reveals differential regulation of mRNA and protein in Helicobacter pylori. Proteomics. 2008;8:2665-75.

[98] Gonzalo CR, Adriana VB, Guillermo MH, Xochitl VM, Ruben AGG, Yolanda LV. Comparative Proteomic Analysis of Helicobacter pylori-Expressed Proteins in Gastric Epithelial Cell Apoptosis. Curr Proteomics. 2009;6:187-97.

[99] Cheng S, Liu Y, Crowley CS, Yeates TO, Bobik TA. Bacterial microcompartments: their properties and paradoxes. BioEssays : news and reviews in molecular, cellular and developmental biology. 2008;30:1084-95.

[100] Sutter M, Boehringer D, Gutmann S, Gunther S, Prangishvili D, Loessner MJ, et al. Structural basis of enzyme encapsulation into a bacterial nanocompartment. Nat Struct Mol Biol. 2008;15:939-47.

[101] Huber LA, Pfaller K, Vietor I. Organelle proteomics - Implications for subcellular fractionation in proteomics. CircRes. 2003;92:962-8.

[102] Lee YH, Tan HT, Chung MCM. Subcellular fractionation methods and strategies for proteomics. Proteomics. 2010;10:3935-56.

[103] Ramsby M, Makowski G. Differential detergent fractionation of eukaryotic cells. Cold Spring Harbor protocols. 2011;2011:prot5592.

[104] Laemmli UK. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature. 1970;227:680-5.

[105] Granvogl B, Plöscher M, Eichacker L. Sample preparation by in-gel digestion for mass spectrometry-based proteomics. Anal Bioanal Chem. 2007;389:991-1002.

[106] Schagger H, von Jagow G. Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. Analytical biochemistry. 1987;166:368-79.

[107] Schagger H. Tricine-SDS-PAGE. Nature protocols. 2006;1:16-22.

[108] Van den Bergh G, Arckens L. Fluorescent two-dimensional difference gel electrophoresis unveils the potential of gel-based proteomics. Current Opinion in Biotechnology. 2004;15:38-43.

[109] Tran JC, Doucette AA. Gel-eluted liquid fraction entrapment electrophoresis: an electrophoretic method for broad molecular weight range proteome separation. Anal Chem. 2008;80:1568-73.

[110] Tran JC, Doucette AA. Multiplexed size separation of intact proteins in solution phase for mass spectrometry. Anal Chem. 2009;81:6201-9.

[111] Lilley KS, Friedman DB. All about DIGE: quantification technology for differential-display 2D-gel proteomics. Expert Review of Proteomics. 2004;1:401-9.

[112] Kolkman A, Dirksen EHC, Slijper M, Heck AJR. Double Standards in Quantitative Proteomics: Direct Comparative Assessment of Difference in Gel Electrophoresis and Metabolic Stable Isotope Labeling. Molecular & Cellular Proteomics. 2005;4:255-66.

[113] Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protocols. 2007;1:2856-60.

[114] Lottspeich F, Engels J, Zettelmeier Lay S. Bioanalytik (2nd edition): Spektrum Akademischer Verlag; 2006.

[115] Gilar M, Olivova P, Daly AE, Gebler JC. Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. Journal of Separation Science. 2005;28:1694-703.

[116] Mohammed S, Heck AJR. Strong cation exchange (SCX) based analytical methods for the targeted analysis of protein post-translational modifications. Current Opinion in Biotechnology. 2011;22:9-16.

[117] Yoshida T. Peptide Separation in Normal Phase Liquid Chromatography. Analytical Chemistry. 1997;69:3038-43.

[118] Hägglund P, Bunkenborg J, Elortza F, Jensen ON, Roepstorff P. A New Strategy for Identification of N-Glycosylated Proteins and Unambiguous Assignment of Their Glycosylation Sites Using HILIC Enrichment and Partial Deglycosylation. Journal of proteome research. 2004;3:556-66.

[119] Sonnenschein L, Seubert A. Separation of inorganic anions using a series of sulfobetaine exchangers. Journal of Chromatography A. 2011;1218:1185-94.

[120] Naidong W. Bioanalytical liquid chromatography tandem mass spectrometry methods on underivatized silica columns with aqueous/organic mobile phases. J Chromatogr B Analyt Technol Biomed Life Sci. 2003;796:209-24.

[121] Boersema P, Mohammed S, Heck AR. Hydrophilic interaction liquid chromatography (HILIC) in proteomics. Anal Bioanal Chem. 2008;391:151-9.

[122] Hemström P, Irgum K. Hydrophilic interaction chromatography. Journal of Separation Science. 2006;29:1784-821.

[123] Ihling C, Sinz A. Proteome analysis of Escherichia coli using high-performance liquid chromatography and Fourier transform ion cyclotron resonance mass spectrometry. Proteomics. 2005;5:2029-42.

[124] Liang Y, Zhang L, Zhang Y. Recent advances in monolithic columns for protein and peptide separation by capillary liquid chromatography. Anal Bioanal Chem. 2013;405:2095-106.

[125] Sproß J, Sinz A. A Capillary Monolithic Trypsin Reactor for Efficient Protein Digestion in Online and Offline Coupling to ESI and MALDI Mass Spectrometry. Analytical Chemistry. 2010;82:1434-43.

[126] Iwasaki M, Miwa S, Ikegami T, Tomita M, Tanaka N, Ishihama Y. One-dimensional capillary liquid chromatographic separation coupled with tandem mass spectrometry unveils the Escherichia coli proteome on a microarray scale. Anal Chem. 2010;82:2616-20.

[127] Spross J, Brauch S, Mandel F, Wagner M, Buckenmaier S, Westermann B, et al. Multidimensional nano-HPLC coupled with tandem mass spectrometry for analyzing biotinylated proteins. Anal Bioanal Chem. 2013;405:2163-73.

[128] Pfaunmiller EL, Paulemond ML, Dupper CM, Hage DS. Affinity monolith chromatography: a review of principles and recent analytical applications. Anal Bioanal Chem. 2013;405:2133-45.

[129] Sproß J, Sinz A. Monolithic media for applications in affinity chromatography. Journal of Separation Science. 2011;34:1958-73.

[130] Harper RA, Pierce J, Savage CR. PURIFICATION OF HUMAN EPIDERMAL GROWTH-FACTOR BY MONOCLONAL-ANTIBODY AFFINITY-CHROMATOGRAPHY. Method Enzymol. 1987;146:3-11.

[131] Fanayan S, Hincapie M, Hancock WS. Using lectins to harvest the plasma/serum glycoproteome. Electrophoresis. 2012;33:1746-54.

[132] Kuo W-H, Chase H. Exploiting the interactions between poly-histidine fusion tags and immobilized metal ions. Biotechnol Lett. 2011;33:1075-84.

[133] Thingholm T, Jensen O. Enrichment and Characterization of Phosphopeptides by Immobilized Metal Affinity Chromatography (IMAC) and Mass Spectrometry. In: Graauw M, editor. Phospho-Proteomics: Humana Press; 2009. p. 47-56.

[134] Cheung R, Wong J, Ng T. Immobilized metal ion affinity chromatography: a review on its applications. Appl Microbiol Biotechnol. 2012;96:1411-20.

[135] Barnea E, Sorkin R, Ziv T, Beer I, Admon A. Evaluation of prefractionation methods as a preparatory step for multidimensional based chromatography of serum proteins. Proteomics. 2005;5:3367-75.

[136] Lecchi P, Gupte AR, Perez RE, Stockert LV, Abramson FP. Size-exclusion chromatography in multidimensional separation schemes for proteome analysis. Journal of biochemical and biophysical methods. 2003;56:141-52.

[137] Gordon SM, Deng J, Lu LJ, Davidson WS. Proteomic characterization of human plasma high density lipoprotein fractionated by gel filtration chromatography. Journal of proteome research. 2010;9:5239-49.

[138] Sandra K, Moshir M, D'hondt F, Tuytten R, Verleysen K, Kas K, et al. Highly efficient peptide separations in proteomics: Part 2: Bi- and multidimensional liquid-based separation techniques. Journal of Chromatography B. 2009;877:1019-39.

[139] Wang H, Hanash S. Multi-dimensional liquid phase based separations in proteomics. Journal of Chromatography B. 2003;787:11-8.

[140] Olsen JV, Ong S-E, Mann M. Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. Molecular & Cellular Proteomics. 2004;3:608-14.

[141] Rodriguez J, Gupta N, Smith RD, Pevzner PA. Does trypsin cut before proline? Journal of proteome research. 2008;7:300-5.

[142] Wisniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nature methods. 2009;6:359-62.

[143] Hohmann L, Sherwood C, Eastham A, Peterson A, Eng JK, Eddes JS, et al. Proteomic Analyses Using Grifola frondosa Metalloendoprotease Lys-N. Journal of proteome research. 2009;8:1415-22.

[144] Taouatas N, Drugan MM, Heck AJR, Mohammed S. Straightforward ladder sequencing of peptides using a Lys-N metalloendopeptidase. Nat Meth. 2008;5:405-7.

[145] Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, et al. A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed. Molecular & Cellular Proteomics. 2009;8:2759-69.

[146] Karas M, Bachmann D, Hillenkamp F. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. Analytical Chemistry. 1985;57:2935-9.

[147] Whitehouse CM, Dreyer RN, Yamashita M, Fenn JB. Electrospray interface for liquid chromatographs and mass spectrometers. Analytical Chemistry. 1985;57:675-9.

[148] Fenn J, Mann M, Meng C, Wong S, Whitehouse C. Electrospray ionization for mass spectrometry of large biomolecules. Science. 1989;246:64-71.

[149] Stapels MD, Barofsky DF. Complementary Use of MALDI and ESI for the HPLC-MS/MS Analysis of DNA-Binding Proteins. Analytical Chemistry. 2004;76:5423-30.

[150] Cech NB, Enke CG. Relating electrospray ionization response to nonpolar character of small peptides. Anal Chem. 2000;72:2717-23.

[151] Stapels MD, Cho JC, Giovannoni SJ, Barofsky DF. Proteomic analysis of novel marine bacteria using MALDI and ESI mass spectrometry. Journal of biomolecular techniques : JBT. 2004;15:191-8.

[152] Kobayashi D, Kumagai J, Morikawa T, Wilson-Morifuji M, Wilson A, Irie A, et al. An Integrated Approach of Differential Mass Spectrometry and Gene Ontology Analysis Identified Novel Proteins Regulating Neuronal Differentiation and Survival. Molecular & Cellular Proteomics. 2009;8:2350-67.

[153] Yang Y, Zhang S, Howe K, Wilson DB, Moser F, Irwin D, et al. A comparison of nLC-ESI-MS/MS and nLC-MALDI-MS/MS for GeLC-based protein identification and iTRAQ-based shotgun quantitative proteomics. Journal of biomolecular techniques : JBT. 2007;18:226-37.

[154] March RE. An introduction to quadrupole ion trap mass spectrometry. J Mass Spectrom. 1997;32:351-69.

[155] Douglas DJ, Frank AJ, Mao D. Linear ion traps in mass spectrometry. Mass Spectrometry Reviews. 2005;24:1-29.

[156] Douglas DJ, French JB. Collisional focusing effects in radio frequency quadrupoles. Journal of the American Society for Mass Spectrometry. 1992;3:398-408.

[157] Kim T, Tolmachev AV, Harkewicz R, Prior DC, Anderson G, Udseth HR, et al. Design and implementation of a new electrodynamic ion funnel. Anal Chem. 2000;72:2247-55.

[158] Dawson PH. Quadrupole Mass Spectrometry and Its Applications (AVS Classics in Vacuum Science and Technology): Springer; 2008.

[159] Krishnaveni A, Kumar Verma N, Menon AG, Mohanty AK. Numerical observation of preferred directionality in ion ejection from stretched rectilinear ion traps. International Journal of Mass Spectrometry. 2008;275:11-20.

[160] Makarov A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. Analytical Chemistry. 2000;72:1156-62.

[161] Kingdon KH. A Method for the Neutralization of Electron Space Charge by Positive Ionization at Very Low Gas Pressures. Physical Review. 1923;21:408-18.

[162] Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R. The Orbitrap: a new mass spectrometer. J Mass Spectrom. 2005;40:430-43.

[163] Roepstorff P, Fohlman J. Letter to the editors. Biological Mass Spectrometry. 1984;11:601-.

[164] Thermo Scientific. LTQ Orbitrap Velos Biotech Operations - Training Course Manual. 2010.

[165] Frese CK, Altelaar AFM, Hennrich ML, Nolting D, Zeller M, Griep-Raming J, et al. Improved Peptide Identification by Targeted Fragmentation Using CID, HCD and ETD on an LTQ-Orbitrap Velos. Journal of proteome research. 2011;10:2377-88.

[166] Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci U S A. 2004;101:9528-33.

[167] Wiesner J, Premsler T, Sickmann A. Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. Proteomics. 2008;8:4466-83.

[168] Hanisch F-G. O-Glycoproteomics: Site-Specifi c O-Glycoprotein Analysis by CID/ETD Electrospray Ionization Tandem Mass Spectrometry and Top-Down Glycoprotein Sequencing by In-Source Decay MALDI Mass Spectrometry. In: McGuckin MA, Thornton DJ, editors. Mucins: Humana Press; 2012. p. 179-89.

[169] Scott NE, Parker BL, Connolly AM, Paulech J, Edwards AV, Crossett B, et al. Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by CID, higher energy collisional dissociation, and electron transfer dissociation MS applied to the N-linked glycoproteome of Campylobacter jejuni. Molecular & cellular proteomics : MCP. 2011;10:M000031-MCP201.

[170] Hunt DF, Coon JJ, Syka JE, Marto JA. Electron Transfer Dissociation for Biopolymer Sequence Analysis. In: FOUNDATION UOVP, editor. US patent 201201840422005.

[171] Good DM, Wirtala M, McAlister GC, Coon JJ. Performance Characteristics of Electron Transfer Dissociation Mass Spectrometry. Molecular & Cellular Proteomics. 2007;6:1942-51.

[172] Campbell JL, Hager JW, Le Blanc JCY. On Performing Simultaneous Electron Transfer Dissociation and Collision-Induced Dissociation on Multiply Protonated Peptides in a Linear Ion Trap. Journal of the American Society for Mass Spectrometry. 2009;20:1672-83.

[173] Swaney DL, McAlister GC, Wirtala M, Schwartz JC, Syka JE, Coon JJ. Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. Anal Chem. 2007;79:477-85.

[174] Chalkley RJ, Medzihradszky KF, Lynn AJ, Baker PR, Burlingame AL. Statistical analysis of Peptide electron transfer dissociation fragmentation mass spectrometry. Anal Chem. 2010;82:579-84.

[175] Pichler P, Koecher T, Holzmann J, Mazanek M, Taus T, Ammerer G, et al. Peptide Labeling with Isobaric Tags Yields Higher Identification Rates Using iTRAQ 4-Plex Compared to TMT 6-Plex and iTRAQ 8-Plex on LTQ Orbitrap. Analytical Chemistry. 2010;82:6549-58.

[176] Molina H, Matthiesen R, Kandasamy K, Pandey A. Comprehensive Comparison of Collision Induced Dissociation and Electron Transfer Dissociation. Analytical Chemistry. 2008;80:4825-35.

[177] Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;20:3551-67.

[178] Eng J, McCormack A, Yates J. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry. 1994;5:976-89.

[179] Fenyö D, Beavis RC. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. Analytical Chemistry. 2003;75:768-74.

[180] Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. Journal of proteome research. 2011;10:1794-805.

[181] Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, et al. Open Mass Spectrometry Search Algorithm. Journal of proteome research. 2004;3:958-64.

[182] Edwards NJ. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol. 2007;3.

[183] Craig R, Cortens JC, Fenyo D, Beavis RC. Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. Journal of proteome research. 2006;5:1843-9.

[184] Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ. Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. Analytical Chemistry. 2006;78:5678-84.

[185] Kwon T, Choi H, Vogel C, Nesvizhskii AI, Marcotte EM. MSblender: A Probabilistic Approach for Integrating Peptide Identifications from Multiple Database Search Engines. Journal of proteome research. 2011;10:2949-58.

[186] Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, et al. PEAKS DB: De Novo sequencing assisted database search for sensitive and accurate peptide identification. Molecular & Cellular Proteomics. 2011.

[187] Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics. 2011;11:996-9.

[188] Barsnes H, Vaudel M, Colaert N, Helsens K, Sickmann A, Berven F, et al. compomics-utilities: an open-source Java library for computational proteomics. BMC bioinformatics. 2011;12:70.

[189] Vaudel M, Barsnes H, Martens L. http://code.google.com/p/peptide-shaker/. 2012.

[190] Käll L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. Journal of proteome research. 2008;7:40-4.

[191] Elias JE, Haas W, Faherty BK, Gygi SP. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Meth. 2005;2:667-75.

[192] Renuse S, Chaerkady R, Pandey A. Proteogenomics. Proteomics. 2011;11:620-30.

[193] Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief Funct Genomic Proteomic. 2008;7:50-62.

[194] Krijgsveld J, Ketting RF, Mahmoudi T, Johansen J, Artal-Sanz M, Verrijzer CP, et al. Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics. Nat Biotech. 2003;21:927-31.

[195] Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. Accurate quantitation of protein expression and site-specific phosphorylation. Proceedings of the National Academy of Sciences. 1999;96:6591-6.

[196] Snijders APL, de Vos MGJ, Wright PC. Novel Approach for Peptide Quantitation and Sequencing Based on 15N and 13C Metabolic Labeling. Journal of proteome research. 2005;4:578-85.

[197] Jehmlich N, Kopinke F-D, Lenhard S, Vogt C, Herbst F-A, Seifert J, et al. Sulfur-36S stable isotope labeling of amino acids for quantification (SULAQ). Proteomics. 2012;12:37-42.

[198] Herbst F-A, Taubert M, Jehmlich N, Behr T, Schmidt F, von Bergen M, et al. Sulfur-34S stable isotope labeling of amino acids for quantification (SULAQ34) of proteomic changes in Pseudomonas fluorescens during naphthalene degradation. Molecular & Cellular Proteomics. 2013.

[199] Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Molecular & cellular proteomics : MCP. 2002;1:376-86.

[200] Kellermann J, Lottspeich F. Isotope-Coded Protein Label. In: Marcus K, editor. Quantitative Methods in Proteomics: Humana Press; 2012. p. 143-53.

[201] Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nature biotechnology. 1999;17:994-9.

[202] Schmidt A, Kellermann J, Lottspeich F. A novel strategy for quantitative proteomics using isotope-coded protein labels. PROTEOMICS. 2005;5:4-15.

[203] Brunner A, Keidel EM, Dosch D, Kellermann J, Lottspeich F. ICPLQuant - A software for non-isobaric isotopic labeling proteomics. Proteomics. 2010;10:315-26.

[204] Leroy B, Rosier C, Erculisse V, Leys N, Mergeay M, Wattiez R. Differential proteomic analysis using isotope-coded protein-labeling strategies: comparison, improvements and application to simulated microgravity effect on Cupriavidus metallidurans CH34. Proteomics. 2010;10:2281-91.

[205] Wiese S, Reidegeld KA, Meyer HE, Warscheid B. Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. Proteomics. 2007;7:340-50.

[206] Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, et al. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. Analytical Chemistry. 2003;75:1895-904.

[207] Dayon L, Hainard A, Licker V, Turck N, Kuhn K, Hochstrasser DF, et al. Relative Quantification of Proteins in Human Cerebrospinal Fluids by MS/MS Using 6-Plex Isobaric Tags. Analytical Chemistry. 2008;80:2921-31.

[208] Stewart II, Thomson T, Figeys D. 18O Labeling: a tool for proteomics. Rapid Communications in Mass Spectrometry. 2001;15:2456-65.

[209] Yao XD, Freas A, Ramirez J, Demirev PA, Fenselau C. Proteolytic O-18 labeling for comparative proteomics: Model studies with two serotypes of adenovirus. Analytical Chemistry. 2001;73:2836-42.

[210] Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. Super-SILAC mix for quantitative proteomics of human tumor tissue. Nature methods. 2010;7:383-5.

[211] Kirkpatrick DS, Gerber SA, Gygi SP. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. Methods. 2005;35:265-73.

[212] Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. Statistical Analysis of Membrane Proteome Expression Changes in Saccharomyces cerevisiae. Journal of proteome research. 2006;5:2339-47.

[213] Cappadona S, Baker PR, Cutillas PR, Heck AJ, van Breukelen B. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. Amino acids. 2012;43:1087-108.

[214] Ong SE, Kratchmarova I, Mann M. Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). Journal of proteome research. 2003;2:173-81.

[215] Croxen MA, Sisson G, Melano R, Hoffman PS. The Helicobacter pylori chemotaxis receptor TlpB (HP0103) is required for pH taxis and for colonization of the gastric mucosa. J Bacteriol. 2006;188:2656-65.

[216] Bury-Mone S, Skouloubris S, Labigne A, De Reuse H. The Helicobacter pylori UreI protein: role in adaptation to acidity and identification of residues essential for its activity and for acid activation. Mol Microbiol. 2001;42:1021-34.

[217] McGee DJ, Zabaleta J, Viator RJ, Testerman TL, Ochoa AC, Mendz GL. Purification and characterization of Helicobacter pylori arginase, RocF: unique features among the arginase superfamily. European Journal of Biochemistry. 2004;271:1952-62.

[218] Stingl K, Uhlemann EM, Schmid R, Altendorf K, Bakker EP. Energetics of Helicobacter pylori and its implications for the mechanism of urease-dependent acid tolerance at pH 1. J Bacteriol. 2002;184:3053-60.

[219] Sycuro LK, Pincus Z, Gutierrez KD, Biboy J, Stern CA, Vollmer W, et al. Peptidoglycan Crosslinking Relaxation Promotes Helicobacter pylori's Helical Shape and Stomach Colonization. Cell. 2010;141:822-33.

[220] Guruge JL, Falk PG, Lorenz RG, Dans M, Wirth HP, Blaser MJ, et al. Epithelial attachment alters the outcome of Helicobacter pylori infection. Proc Natl Acad Sci U S A. 1998;95:3925-30.

[221] Tan S, Tompkins LS, Amieva MR. Helicobacter pylori usurps cell polarity to turn the cell surface into a replicative niche. PLoS Pathog. 2009;5:e1000407.

[222] Wroblewski LE, Peek RM, Jr. "Targeted disruption of the epithelial-barrier by Helicobacter pylori". Cell Commun Signal. 2011;9:29.

[223] Lapointe TK, O'Connor PM, Jones NL, Menard D, Buret AG. Interleukin-1 receptor phosphorylation activates Rho kinase to disrupt human gastric tight junctional claudin-4 during Helicobacter pylori infection. Cell Microbiol. 2010;12:692-703.

[224] Papini E, Satin B, Norais N, de Bernard M, Telford JL, Rappuoli R, et al. Selective increase of the permeability of polarized epithelial cell monolayers by Helicobacter pylori vacuolating toxin. J Clin Invest. 1998;102:813-20.

[225] Wroblewski LE, Shen L, Ogden S, Romero-Gallo J, Lapierre LA, Israel DA, et al. Helicobacter pylori dysregulation of gastric epithelial tight junctions by urease-mediated myosin II activation. Gastroenterology. 2009;136:236-46.

[226] Hoy B, Lower M, Weydig C, Carra G, Tegtmeyer N, Geppert T, et al. Helicobacter pylori HtrA is a new secreted virulence factor that cleaves E-cadherin to disrupt intercellular adhesion. EMBO Rep. 2010;11:798-804.

[227] Fedwick JP, Lapointe TK, Meddings JB, Sherman PM, Buret AG. Helicobacter pylori activates myosin light-chain kinase to disrupt claudin-4 and claudin-5 and increase epithelial permeability. Infect Immun. 2005;73:7844-52.

[228] Chan AO, Lam SK, Wong BC, Wong WM, Yuen MF, Yeung YH, et al. Promoter methylation of E-cadherin gene in gastric mucosa associated with Helicobacter pylori infection and in gastric cancer. Gut. 2003;52:502-6.

[229] Watanabe T, Tsuge H, Imagawa T, Kise D, Hirano K, Beppu M, et al. Nucleolin as cell surface receptor for tumor necrosis factor-alpha inducing protein: a carcinogenic factor of Helicobacter pylori. Journal of cancer research and clinical oncology. 2010;136:911-21.

[230] Suganuma M, Watanabe T, Yamaguchi K, Takahashi A, Fujiki H. Human gastric cancer development with TNF-alpha-inducing protein secreted from Helicobacter pylori. Cancer letters. 2012;322:133-8.

[231] Jang JY, Yoon HJ, Yoon JY, Kim HS, Lee SJ, Kim KH, et al. Crystal structure of the TNF-alpha-Inducing protein (Tipalpha) from Helicobacter pylori: Insights into Its DNA-binding activity. Journal of molecular biology. 2009;392:191-7.

[232] Suganuma M, Yamaguchi K, Ono Y, Matsumoto H, Hayashi T, Ogawa T, et al. TNF-alpha-inducing protein, a carcinogenic factor secreted from H. pylori, enters gastric cancer cells. International journal of cancer Journal international du cancer. 2008;123:117-22.

[233] Crowe SE. Helicobacter infection, chronic inflammation, and the development of malignancy. Curr Opin Gastroenterol. 2005;21:32-8.

[234] Montecucco C, Rappuoli R. Living dangerously: how Helicobacter pylori survives in the human stomach. Nat Rev Mol Cell Biol. 2001;2:457-66.

[235] Bai H, Li Q, Liu X, Li Y. Characteristics and Interactions of Helicobacter pylori and H. pylori-Infected Human Gastroduodenal Epithelium in Peptic Ulcer: A Transmission Electron Microscopy Study. Dig Dis Sci. 2010;55:82-8.

[236] Kusters JG, Gerrits MM, Van Strijp JA, Vandenbroucke-Grauls CM. Coccoid forms of Helicobacter pylori are the morphologic manifestation of cell death. Infection and Immunity. 1997;65:3672-9.

[237] Li N, Han L, Chen J, Lin X, Chen H, She F. Proliferative and apoptotic effects of gastric epithelial cells induced by coccoid Helicobacter pylori. Journal of Basic Microbiology. 2012:n/a-n/a.

[238] Bode G, Mauch F, Malfertheiner P. The coccoid forms of Helicobacter pylori. Criteria for their viability. Epidemiology and infection. 1993;111:483-90.

[239] Hua J, Ho B. Is the coccoid form of Helicobacter pylori viable? Microbios. 1996;87:103-12.

[240] Oliver JD. Recent findings on the viable but nonculturable state in pathogenic bacteria. FEMS microbiology reviews. 2010;34:415-25.

[241] Zhang MJ, Zhao F, Xiao D, Gu YX, Meng FL, He LH, et al. Comparative proteomic analysis of passaged Helicobacter pylori. J Basic Microbiol. 2009;49:482-90.

[242] Jungblut PR, Bumann D, Haas G, Zimny-Arndt U, Holland P, Lamer S, et al. Comparative proteome analysis of Helicobacter pylori. Mol Microbiol. 2000;36:710-25.

[243] Jungblut PR, Schiele F, Zimny-Arndt U, Ackermann R, Schmid M, Lange S, et al. Helicobacter pylori proteomics by 2-DE/MS, 1-DE-LC/MS and functional data mining. Proteomics. 2010;10:182-93.

[244] Cargile BJ, Bundy JL, Grunden AM, Stephenson JL. Synthesis/Degradation Ratio Mass Spectrometry for Measuring Relative Dynamic Protein Turnover. Analytical Chemistry. 2003;76:86-97.

[245] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, et al. The primary transcriptome of the major human pathogen Helicobacter pylori. Nature. 2010;464:250-5.

[246] Morbt N, Mogel I, Kalkhof S, Feltens R, Roder-Stolinski C, Zheng J, et al. Proteome changes in human bronchoalveolar cells following styrene exposure indicate involvement of oxidative stress in the molecular-response mechanism. Proteomics. 2009;9:4920-33.

[247] Mann M. Functional and quantitative proteomics using SILAC. Nat Rev Mol Cell Biol. 2006;7:952-8.

[248] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27-30.

[249] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109-14.

[250] Kotera M, Hirakawa M, Tokimatsu T, Goto S, Kanehisa M. The KEGG Databases and Tools Facilitating Omics Analysis: Latest Developments Involving Human Diseases and Pharmaceuticals. In: Wang J, Tan AC, Tian T, editors. Next Generation Microarray Bioinformatics: Humana Press; 2012. p. 19-39.

[251] Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research. 2000;28:33-6.

[252] de Boer P, Crossley R, Rothfield L. The essential bacterial cell-division protein FtsZ is a GTPase. Nature. 1992;359:254-6.

[253] Feng Y, Chien KY, Chen HL, Chiu CH. Pseudogene Recoding Revealed from Proteomic Analysis of Salmonella Serovars. Journal of proteome research. 2012;11:1715-9.

[254] Soufi B, Kumar C, Gnad F, Mann M, Mijakovic I, Macek B. Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) Applied to Quantitative Proteomics of Bacillus subtilis. Journal of proteome research. 2010;9:3638-46.

[255] Cox J, Matic I, Hilger M, Nagaraj N, Selbach M, Olsen JV, et al. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. Nature protocols. 2009;4:698-705.

[256] Reynolds DJ, Penn CW. Characteristics of Helicobacter pylori growth in a defined medium and determination of its amino acid requirements. Microbiology. 1994;140 (Pt 10):2649-56.

[257] Nedenskov P. Nutritional requirements for growth of Helicobacter pylori. Appl Environ Microbiol. 1994;60:3450-3.

[258] Karita M, Etterbeek ML, Forsyth MH, Tummuru MK, Blaser MJ. Characterization of Helicobacter pylori dapE and construction of a conditionally lethal dapE mutant. Infect Immun. 1997;65:4158-64.

[259] Gottesman MM, Hicks ML, Gellert M. Genetics and function of DNA ligase in Escherichia coli. Journal of molecular biology. 1973;77:531-47.

[260] Lindahl T, Barnes DE. Mammalian DNA ligases. Annual review of biochemistry. 1992;61:251-81.

[261] Shrivastava N, Nag JK, Misra-Bhattacharya S. Molecular Characterization of NAD$^+$-Dependent DNA Ligase from *Wolbachia* Endosymbiont of Lymphatic Filarial Parasite *Brugia malayi*. PloS one. 2012;7:e41113.

[262] Mills SD, Eakin AE, Buurman ET, Newman JV, Gao N, Huynh H, et al. Novel Bacterial NAD+-Dependent DNA Ligase Inhibitors with Broad-Spectrum Activity and Antibacterial Efficacy In Vivo. Antimicrobial Agents and Chemotherapy. 2011;55:1088-96.

[263] Odenbreit S, Swoboda K, Barwig I, Ruhl S, Borén T, Koletzko S, et al. Outer Membrane Protein Expression Profile in Helicobacter pylori Clinical Isolates. Infection and Immunity. 2009;77:3782-90.

[264] Lertsethtakarn P, Ottemann KM, Hendrixson DR. Motility and Chemotaxis in Campylobacter and Helicobacter. Annual Review of Microbiology. 2011;65:389-410.

[265] Rolig AS, Shanks J, Carter JE, Ottemann KM. Helicobacter pylori Requires TlpD-Driven Chemotaxis To Proliferate in the Antrum. Infection and Immunity. 2012;80:3713-20.

[266] Sarkar MK, Paul K, Blair D. Chemotaxis signaling protein CheY binds to the rotor protein FliN to control the direction of flagellar rotation in Escherichia coli. Proc Natl Acad Sci U S A. 2010;107:9370-5.

[267] Terry K, Go AC, Ottemann KM. Proteomic mapping of a suppressor of non-chemotactic cheW mutants reveals that Helicobacter pylori contains a new chemotaxis protein. Molecular Microbiology. 2006;61:871-82.

[268] Colland F, Rain JC, Gounon P, Labigne A, Legrain P, De Reuse H. Identification of the Helicobacter pylori anti-sigma28 factor. Mol Microbiol. 2001;41:477-87.

[269] Terry K, Williams SM, Connolly L, Ottemann KM. Chemotaxis Plays Multiple Roles during Helicobacter pylori Animal Infection. Infection and Immunity. 2005;73:803-11.

[270] McGee DJ, Langford ML, Watson EL, Carter JE, Chen Y-T, Ottemann KM. Colonization and Inflammation Deficiencies in Mongolian Gerbils Infected by Helicobacter pylori Chemotaxis Mutants. Infection and Immunity. 2005;73:1820-7.

[271] Eaton KA, Morgan DR, Krakowka S. Motility as a Factor in the Colonization of Gnotobiotic Piglets by Helicobacter-Pylori. J Med Microbiol. 1992;37:123-7.

[272] McColm AA. Nonprimate Animal Models of H. pylori Infection. Methods in molecular medicine. 1997;8:235-51.

[273] Foynes S, Dorrell N, Ward SJ, Stabler RA, McColm AA, Rycroft AN, et al. Helicobacter pylori possesses two CheY response regulators and a histidine kinase sensor, CheA, which are essential for chemotaxis and colonization of the gastric mucosa. Infect Immun. 2000;68:2016-23.

[274] Kim S, Sierra R, McGee D, Zabaleta J. Transcriptional profiling of gastric epithelial cells infected with wild type or arginase-deficient Helicobacter pylori. BMC Microbiology. 2012;12:175.

[275] Zabaleta J, McGee DJ, Zea AH, Hernández CP, Rodriguez PC, Sierra RA, et al. Helicobacter pylori Arginase Inhibits T Cell Proliferation and Reduces the Expression of the TCR ζ-Chain (CD3ζ). The Journal of Immunology. 2004;173:586-93.

[276] Gobert AP, McGee DJ, Akhtar M, Mendz GL, Newton JC, Cheng Y, et al. Helicobacter pylori arginase inhibits nitric oxide production by eukaryotic cells: a strategy for bacterial survival. Proc Natl Acad Sci U S A. 2001;98:13844-9.

[277] Voland P, Weeks DL, Marcus EA, Prinz C, Sachs G, Scott D. Interactions among the seven Helicobacter pylori proteins encoded by the urease gene cluster. American Journal of Physiology - Gastrointestinal and Liver Physiology. 2003;284:G96-G106.

[278] Suganuma M, Kuzuhara T, Yamaguchi K, Fujiki H. Carcinogenic role of tumor necrosis factor-alpha inducing protein of Helicobacter pylori in human stomach. Journal of biochemistry and molecular biology. 2006;39:1-8.

[279] Watanabe T, Hirano K, Takahashi A, Yamaguchi K, Beppu M, Fujiki H, et al. Nucleolin on the Cell Surface as a New Molecular Target for Gastric Cancer Treatment. Biological and Pharmaceutical Bulletin. 2010;33:796-803.

[280] Mongelard F, Bouvet P. AS-1411, a guanosine-rich oligonucleotide aptamer targeting nucleolin for the potential treatment of cancer, including acute myeloid leukemia. Current opinion in molecular therapeutics. 2010;12:107-14.

[281] Ireson CR, Kelland LR. Discovery and development of anticancer aptamers. Molecular Cancer Therapeutics. 2006;5:2957-62.

[282] Godlewska R, Pawlowski M, Dzwonek A, Mikula M, Ostrowski J, Drela N, et al. Tip-alpha (hp0596 gene product) is a highly immunogenic Helicobacter pylori protein involved in colonization of mouse gastric mucosa. Curr Microbiol. 2008;56:279-86.

[283] Stopher DA, Gage R. Determination of a new antifungal agent, voriconazole, by multidimensional high-performance liquid chromatography with direct plasma injection onto a size-exclusion column. Journal of Chromatography B: Biomedical Sciences and Applications. 1997;691:441-8.

[284] Rea JC, Lou Y, Cuzzi J, Hu Y, de Jong I, Wang YJ, et al. Development of capillary size exclusion chromatography for the analysis of monoclonal antibody fragments extracted from human vitreous humor. Journal of Chromatography A. 2012;1270:111-7.

[285] Xie F, Smith RD, Shen Y. Advanced proteomic liquid chromatography. Journal of Chromatography A. 2012;1261:78-90.

[286] Yang F, Shen Y, Camp DG, Smith RD. High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. Expert Review of Proteomics. 2012;9:129-34.

[287] Bodnar WM, Blackburn RK, Krise JM, Moseley MA. Exploiting the complementary nature of LC/MALDI/MS/MS and LC/ESI/MS/MS for increased proteome coverage. Journal of the American Society for Mass Spectrometry. 2003;14:971-9.

[288] Jukes TH, Cantor CR. Evolution of Protein Molecules: Academy Press; 1969.

[289] McDonald L, Robertson DH, Hurst JL, Beynon RJ. Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. Nature methods. 2005;2:955-7.

[290] Schepmoes AA, Zhang Q, Petritis BO, Qian WJ, Smith RD. N-terminal enrichment: developing a protocol to detect specific proteolytic fragments. Journal of biomolecular techniques : JBT. 2009;20:263-5.

[291] Xu G, Jaffrey SR. N-CLAP: global profiling of N-termini by chemoselective labeling of the alpha-amine of proteins. Cold Spring Harbor protocols. 2010;2010:pdb prot5528.

[292] Paetzel M, Dalbey RE, Strynadka NCJ. The structure and mechanism of bacterial type I signal peptidases - A novel antibiotic target. Pharmacol Ther. 2000;87:27-49.

[293] Ollinger J, O'Malley T, Ahn J, Odingo J, Parish T. Inhibition of the Sole Type I Signal Peptidase of Mycobacterium tuberculosis Is Bactericidal under Replicating and Nonreplicating Conditions. Journal of Bacteriology. 2012;194:2614-9.

[294] Schallenberger MA, Niessen S, Shao C, Fowler BJ, Romesberg FE. Type I Signal Peptidase and Protein Secretion in Staphylococcus aureus. Journal of Bacteriology. 2012;194:2677-86.

[295] Filiou MD, Martins-de-Souza D, Guest PC, Bahn S, Turck CW. To label or not to label: applications of quantitative proteomics in neuroscience research. Proteomics. 2012;12:736-47.

[296] Filiou MD, Varadarajulu J, Teplytska L, Reckow S, Maccarrone G, Turck CW. The N-15 isotope effect in Escherichia coli: A neutron can make the difference. Proteomics. 2012;12:3121-+.

[297] Barbucci R, Benvenuti M, Casolaro M, Lamponi S, Magnani A. Sulfated hyaluronic acid as heparin-like material: physicochemical and biological characterization. J Mater Sci: Mater Med. 1994;5:830-3.

[298] Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res. 2003;31:334-41.

[299] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols. 2008;4:44-57.

[300] Kliemt S, Lange C, Otto W, Hintze V, Moller S, von Bergen M, et al. Sulfated hyaluronan containing collagen matrices enhance cell-matrix-interaction, endocytosis, and osteogenic differentiation of human mesenchymal stromal cells. Journal of proteome research. 2013;12:378-89.

[301] Soares NC, Spät P, Krug K, Macek B. Global Dynamics of the Escherichia coli Proteome and Phosphoproteome During Growth in Minimal Medium. Journal of proteome research. 2013;12:2611-21.

[302] Couté Y, Hernandez C, Appel RD, Sanchez J-C, Margolles A. Labeling of Bifidobacterium longum Cells with 13C-Substituted Leucine for Quantitative Proteomic Analyses. Applied and Environmental Microbiology. 2007;73:5653-6.

[303] Khan Z, Amini S, Bloom J, Ruse C, Caudy A, Kruglyak L, et al. Accurate proteome-wide protein quantification from high-resolution 15N mass spectra. Genome Biol. 2011;12:R122.

[304] Li L, Li Q, Rohlin L, Kim U, Salmon K, Rejtar T, et al. Quantitative Proteomic and Microarray Analysis of the Archaeon Methanosarcina acetivorans Grown with Acetate versus Methanol. Journal of proteome research. 2006;6:759-71.

[305] Pereira-Medrano AG, Margesin R, Wright PC. Proteome characterization of the unsequenced psychrophile Pedobacter cryoconitis using 15N metabolic labeling, tandem mass spectrometry, and a new bioinformatic workflow. Proteomics. 2012;12:775-89.

[306] Kuhn K, Baumann C, Tommassen J, Prinz T. TMT Labelling for the Quantitative Analysis of Adaptive Responses in the Meningococcal Proteome. In: Christodoulides M, editor. Neisseria meningitidis: Humana Press; 2012. p. 127-41.

[307] Lee JY, Pajarillo EAB, Kim MJ, Chae JP, Kang D-K. Proteomic and Transcriptional Analysis of Lactobacillus johnsonii PF01 during Bile Salt Exposure by iTRAQ Shotgun Proteomics and Quantitative RT-PCR. Journal of proteome research. 2012;12:432-43.

[308] Matallana-Surget S, Joux F, Wattiez R, Lebaron P. Proteome analysis of the UVB-resistant marine bacterium Photobacterium angustum S14. PloS one. 2012;7:e42299.

[309] Qiao J, Wang J, Chen L, Tian X, Huang S, Ren X, et al. Quantitative iTRAQ LC–MS/MS Proteomics Reveals Metabolic Responses to Biofuel Ethanol in Cyanobacterial Synechocystis sp. PCC 6803. Journal of proteome research. 2012;11:5286-300.

[310] Colquhoun DR, Hartmann EM, Halden RU. Proteomic Profiling of the Dioxin-Degrading Bacterium Sphingomonas wittichii RW1. Journal of Biomedicine and Biotechnology. 2012;2012:9.

[311] Ramachandran B, Dikshit KL, Dharmalingam K. Recombinant E. coli expressing Vitreoscilla haemoglobin prefers aerobic metabolism under microaerobic conditions: a proteome-level study. J Biosci. 2012;37:617-33.

[312] Liao L, Park SK, Xu T, Vanderklish P, Yates JR. Quantitative proteomic analysis of primary neurons reveals diverse changes in synaptic protein content in fmr1 knockout mice. Proceedings of the National Academy of Sciences. 2008;105:15281-6.

[313] Szczebara F, Dhaenens L, Armand S, Husson MO. Regulation of the transcription of genes encoding different virulence factors in Helicobacter pylori by free iron. FEMS Microbiol Lett. 1999;175:165-70.

[314] Dephoure N, Gygi SP. Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast. Science signaling. 2012;5:rs2.

6 Appendix

6.1 Scientific publications

Optimization of parameters for coverage of low molecular weight proteins

Müller SA, Kohajda T, Findeiß S, Stadler PF, Washietl S, Kellis M, von Bergen M, Kalkhof S.

Anal Bioanal Chem. 2010 Dec;398(7-8):2867-81. Epub 2010 Aug 28.
DOI: 10.1007/s00216-010-4093-x

Cell fractionation - an important tool for compartment proteomics

Maxie Rockstroh, **Stephan Müller**, Claudia Jende, Alexandra Kerzhner, Martin von Bergen, Janina Melanie Tomm.

Journal of Integrated OMICS, Vol 1, No 1 (2011), 135-143.
DOI: 10.5584/jiomics.v1i1.52

RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data

Washietl S, Findeiss S, **Müller SA**, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N.

RNA. 2011 Apr;17(4):578-94. Epub 2011 Feb 28.
DOI: 10.1261/rna.2536111

Identification of new protein coding sequences and signal peptidase cleavage sites of *Helicobacter pylori* strain 26695 by proteogenomics

Stephan A. Müller, Sven Findeiß, Sandy R. Pernitzsch, Dirk K. Wissenbach, Peter F. Stadlerb, Ivo L. Hofacker, Martin von Bergen, Stefan Kalkhof
J Proteomics. 2013 Jun 28;86:27-42. Epub 2013 May 9.
DOI:10.1016/j.jprot.2013.04.036

Quantitative proteomics reveals altered expression of extracellular matrix related proteins of human primary dermal fibroblasts in response to sulfated hyaluronan and collagen applied as artificial extracellular matrix

Stephan A. Müller, Anja van der Smissen, Margarete von Feilitzsch, Ulf Anderegg, Stefan Kalkhof, Martin von Bergen. J Mater Sci Mater Med. 2012 Dec; 23(12):3053-65.
DOI: 10.1007/s10856-012-4760-x

6.2 Affirmation in lieu of an oath / Selbstständigkeitserklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit mit dem Titel „Optimization of quantitative proteomic LC-MS analyses and proteomic insights into *Helicobacter pylori*“ selbstständig angefertigt habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Schriften oder Medien entnommen sind, sind als solche kenntlich gemacht.

Ich habe keine vergebliche Promotionsversuche unternommen. Die Dissertation habe ich in der gegenwärtigen bzw. in einer anderen Fassung keiner anderen Fakultät vorgelegt.

Ort, Datum

Unterschrift

6.3 Curriculum vitae

Berufstätigkeit

Seit September 2013:	Wissenschaftlicher Mitarbeiter am Deutschen Zentrum für neurodegenerative Erkrankungen (DZNE) in der Abteilung Neuroproteomik
Juni 2013	Wissenschaftlicher Mitarbeiter an der Medizinischen Universität Wien in der Core Facility Proteomics
Februar – Mai 2013	Wissenschaftlicher Mitarbeiter am Department Proteomics, Arbeitsgruppe Functional Genomics, Helmholtz-Zentrum für Umweltforschung (UFZ) Leipzig
Juli 2009 – Dezember 2012	Wissenschaftlicher Mitarbeiter / Doktorand am Department Proteomics, Arbeitsgruppe Massenspektrometrie, Helmholtz-Zentrum für Umweltforschung (UFZ) Leipzig
November 2008 – Juni 2009	Wissenschaftlicher Mitarbeiter in der Arbeitsgruppe Bioanalytik am Lehrstuhl für Bioverfahrenstechnik der Universität Erlangen-Nürnberg

Studium

Oktober 2002 – September 2008	Studium des Chemie- und Bioingenieurwesens an der Friedrich-Alexander Universität Erlangen-Nürnberg Studienrichtung Technische Chemie - Biotechnologie
August – September 2008	Studentische Hilfskraft in der Arbeitsgruppe Bioanalytik am Lehrstuhl für Bioverfahrenstechnik der Universität Erlangen-Nürnberg
Februar – Juli 2008	Diplomarbeit „Disulfide Bond Mapping – LC-MALDI MS gestützte Strukturaufklärung von Proteinen“ im Bereich Bioanalytik am Lehrstuhl für Bioverfahrenstechnik der Universität Erlangen-Nürnberg
November 2006 – Februar 2007	Studienarbeit „Kultivierung scherkraftsensitiver Flagellate“ in der Arbeitsgruppe Marine Biotechnologie am Lehrstuhl für Bioverfahrenstechnik der Universität Erlangen-Nürnberg

Studienbegleitende Tätigkeiten

Januar – Juni 2007	Praktikum im Bereich Schnellpyrolyse von Biomasse im „Biolique“-Forschungsvorhaben am Institut für Technische Chemie, Bereich Chemisch-Physikalische Verfahren (ITC-CPV) des Forschungszentrums Karlsruhe
Juli – August 2004	Praktikum im Bereich Anodisierung von Aluminiumteilen und Abwasserreinigung bei der FTE automotive GmbH in Ebern
Oktober 2002 – November 2008	Nebenjob bei IWS Industrie-Wartungssysteme GmbH & Co. KG

Schulische Ausbildung

1992 – 2001	Friedrich-Rückert-Gymnasium Ebern Abschluss mit allgemeiner Hochschulreife
1988 – 1992	Grundschule Reckendorf

Wehrdienst

September 2001 – Juni 2002

Ort, Datum

Unterschrift

6.4 Acknowledgment / Danksagung

Zu allererst möchte ich mich bei meinen Eltern bedanken, die mich das ganze Studium und während der Doktorarbeit großartig unterstützt haben. Ein besonderer Dank gilt meiner Freundin Julia, die mich und meine Launen, besonders in schwierigen Zeiten während der Doktorarbeit, ertragen und mich jederzeit seelisch und moralisch unterstützt hat.

Des Weiteren möchte ich Dr. Stefan Kalkhof meinen großen Dank aussprechen. Er hat mich während meiner Doktorarbeit stets mit Engelsgeduld betreut. Stefan hat mich in meiner Doktorandenzeit immer vorbildlich unterstützt und Mut zugesprochen. Darüber hinaus möchte ich mich bei Prof. Dr. Martin von Bergen für die Möglichkeit meine Doktorarbeit an seinem Department anzufertigen und seine Unterstützung bedanken.

Ferner möchte ich Frau Prof. Dr. Sinz für die universitäre Betreuung meiner Dissertation und die kritischen Diskussionen danken. Dies ist keine Selbstverständlichkeit, da man als externer Doktorand nur Arbeit aber kaum Nutzen bringt.

Außerdem möchte ich mich bei meinen Kooperationspartner bedanken. Ohne sie wären die vielen interdisziplinären Projekte nicht möglich gewesen. Hier gilt mein Dank insbesondere Sven Findeiß. Er war immer bereit neue Ideen mit zu entwickeln und hat stets geduldig seine bioinformatischen Ansätze erklärt. Bei Sandy Pernitzsch und Cynthia Sharma möchte ich mich für die großartige Unterstützung in den *Helicobacter pylori* Projekten bedanken. Anja van der Smissen und PD Dr. Ulf Anderegg danke ich für die gute Zusammenarbeit im Fibroblastenprojekt.

Nicht zuletzt möchte ich mich ganz herzlich bei meinen Kollegen an den Departments Proteomics und Metabolomics bedanken. Die gute Kollegialität und stete Hilfsbereitschaft haben einem das Leben als Doktorand viel einfacher gemacht. Ich habe während meiner Zeit am UFZ nicht nur gute Kollegen sondern auch gute Freunde gewonnen.

Bei Jacqueline möchte ich mich für die Hilfe bei der Probenvorbereitung und ihre herzliche Art bedanken. Steffi, Franziska, Fernando und Tommy danke ich für das gute Arbeitsklima in der Gruppe Massenspektrometrie, sowie die guten Gespräche und gegenseitige Hilfestellung. Zudem möchte ich mich bei den Teilnehmern der „wissenschaftlichen Doktorandenrunde“ um 16 Uhr in der Kaffeeküche bedanken. Hier konnte man nicht nur seine wissenschaftlichen Fragen sondern auch Sorgen loswerden.

Besonders möchte ich mich bei Sven Haange bedanken. Er hat mir als Spezialist für bakterielle Proteomics immer mit gutem Rat zur Seite gestanden. Zudem möchte ich mich bei ihm für seine Englischkorrektur bedanken. Recht herzlich danke ich zudem meiner angeheirateten Cousine Amy Welsh für die Rechtschreibkorrektur meiner Arbeit.