Title

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

Authors

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

Affiliations

[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] Bioinformatics Group, Department of Computer Science, University Leipzig, 04107 Leipzig, Germany

[e] RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, 04103 Leipzig, Germany

[f] Santa Fe Institute, Santa Fe, 87501 New Mexico, USA

[g] Max-Planck-Institute for Mathematics in Sciences, 04103 Leipzig, Germany

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

Correspondence

Dr. Stefan Kalkhof,
Department of Proteomics,
UFZ, Helmholtz-Centre for Environmental Research,
Permoserstr. 15,
04318 Leipzig,
Germany
Email: stefan.kalkhof@ufz.de
Phone: +49-341-2351354
Fax: +49-341-2351786

Abstract

The 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 include erroneous or miss protein annotations. 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 *H. pylori*. Our proteogenomics 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 allowing for semi-specific cleavage of the proteases, 72 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

1 Introduction

The first DNA-based genome was sequenced by Frederick Sanger in 1977 [1]. At the starting point of this development, genome sequencing was restricted to rather small genomes. Further developments like computer-based alignment of shotgun fragments [2] and the polymerase chain reaction [3] transformed genome sequencing to a well automated and cost-effective high-throughput method. A major breakthrough in genome sequencing was reached, when the first individual human genome was fully sequenced by the genome project of Craig Venter in 2007 [4]. Nowadays, many companies are providing whole genome sequencing. Therefore, genome sequencing is easily accessible for the science community and is no longer a bottleneck in research. Hence, hundreds of additional genomes will be sequenced and have to be analyzed within the next years.

The annotation of protein coding sequences in genomic data is usually based on gene finding software like IMG [5], RAST [6], Glimmer [7], or GeneMark [8]. These tools are limited in their prediction accuracy. 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, e.g. ribosome binding site, which are located in vicinity of open reading frames. However, many exceptions to the classical translation initiation model are known [9]. The previously underestimated number of leaderless mRNAs in various species is only one example [PMID: 11849551, PMID: 19996181, PMID: 22080557]. Another common problem is the large amount of so called hypothetical genes that have been predicted but for which no function is assigned to so far. Although big effort is spend on functional assignment, even for the model organism *Helicobacter pylori* 26695 about 33% of protein coding genes still belong to this class [10]. Furthermore, most tools use a minimum open reading frame length cut off in order to keep the false discovery rate low. As a consequence, short protein coding genes that are expressed and functional are lacking in the annotation [11]. In eukaryotes additionally the prediction of alternative splice variants for commonly used software packages is challenging. Additionally, the results of common gene annotation algorithms differ from each other [12]. Dependent on the used method automatic predictions which differ by the limitations of the applied approach protein sequences are deposited in databases like NCBI or UniProt. All these issues still keep researches working on the improvement of existing protein coding gene annotation [13, 14]. A complementary approach to commonly used protein coding gene annotation methods is RNAcode [14]. It does neither rely on species specific gene features like open reading frame detection or sequence motifs necessary for ribosome binding or splicing nor on trainings data. RNAcode simply analyzes by a statistical framework nucleotide and amino acid variations in multiple sequence alignments and thereby detects high scoring segments with features, e.g. synonymous substitutions and indels that preserve the reading frame, typical for conserved protein coding regions.

Comparative genomics studies of different *H. pylori* strains already investigated differences of current coding sequence annotations [10, 15, 16]. Medigue *et al.* [15] identified putative DNA sequencing errors which result in missing or erroneous protein annotations. Boneca *et al.* [10] on the other hand focused on 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 frameshifts, slipped-strand misspairing mechanisms and pseudogenes. However, the results of these studies were only used to some extent to improve protein databases.

High quality protein databases are the fundament of proteomics studies. Missing annotations or erroneous annotated protein sequences will not be covered since protein identification in classical shotgun proteomics only 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. [17-26].

In a typical proteogenomics approach, an existing protein sequence database is complemented by a six frame translation of the whole genome to generate a comprehensive data source. Transcriptome data can also be used to improve and extend the database in particular in eukaryotes due to the inclusion of additional splice variants [24]. The identification of peptides supporting unique sequences within the six frame translation is of great interest. Peptides located 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 [17, 24]. Peptides within annotated intronic regions can be used to identify new exons. Novel splice variants can be identified either by exon-exon spanning peptides or by fragments that map to intergenic regions and are subsequently connected to an existing gene [24, 27]. ,

The ongoing development of MS has made it possible to acquire MS/MS spectra with high resolution, high mass accuracy and fast scanning speed [28]. The invention of nano-UHPLC [29, 30], multidimensional LC [31] as well as the application of ultra-long gradients [32] or long monolithic columns [33] for peptide separation enables LC-MS/MS analysis to dig deeper into the proteome. Cell compartment [34, 35] or protein fractionation [36, 37] prior to proteolytic digestion are widely used strategies to further improve proteome coverage.

As a consequence of this development whole proteomes can be nearly completely covered in proteomics studies [38, 39]. Recently, Nagaraj *et al.* [40] identified 10255 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 (16846 transcripts, 11936 genes) derived from RNA-Seq [41, 42] showed 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.

In this study 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 respectively 71% of the predicted proteome were identified based on at least two peptides with a false discovery rate (FDR) below 1%. Furthermore proteogenomic analysis revealed ten proteins with either none (4) or incomplete (6) 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, 72 previously unknown signal peptide sequences could be annotated by MS/MS spectra with a search strategy allowing for semi-specific enzyme cleavage 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 genome browser [43].

2 Materials and methods

2.1 Cell Culture

Helicobacter pylori strain 26695 was streaked out from cryostock 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 providing atmosphere of 10% CO_2 and 6% O_2), bacteria were harvested and restreaked to fresh plates. For liquid culture, 50 ml Brain Heart Infusion medium (BHI from BD, supplemented with 10% FCS and same antibiotics as mentioned above) were inoculated with harvested cells from plate at $OD_{600\ nm}$ of 0.02 per ml. Bacteria were grown shaking at 140 rpm in jars under microaerophilic conditions (same conditions like above) and harvested at the transition from exponential to stationary growth phase. Finally, *H. pylori* cells were collected by centrifugation (4000xg, 10 min, 4 °C) and washed two times with ice-cold PBS prior to protein extraction and pre-separation. Two biological replicates were used for proteome analysis.

2.2 Protein extraction and preseparation

Cells were lysed in a urea buffer as previously described [44]. Cell debris and undissolved material were removed by centrifugation (10 min, 16000×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 were 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).

To enrich and preseparate the low MW proteome of *H. pylori* SEC was used. 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 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) were 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 further 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, the 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 and destained, reduced, alkylated with iodoacetamide and digested with trypsin as previously described [34]. 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. The 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 every protein fraction were separately digested with trypsin, LysC and AspN (sequencing grade, Roche, Mannheim, DE) with an enzyme to protein weight ratio of approx. 1:20. The 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 2-40% (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=60000, 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 depended 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 fragment ion tolerance of 0.5 Da were defined. Carbamidomethylation of cysteines and oxidation of methionines were specified as fixed respectively 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.

2.5 Additionally, we integrated MS data published by Jungblut *et al.* [45] 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.

2.6 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 proteins 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.

2.7 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 (262166 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, pyro-glu modification of glutamic acid and glutamine at the peptide N-terminus were 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.* [46].

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.* [45] except missed cleavage limits were set to two. Peptide and protein identifications were filtered according to the same thresholds applied to our data.

2.8 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 genome browser [43, 47]. 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 were 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 could be detected. Similar sequences of other strains were included to the database to identify possible sequencing errors. Furthermore, detected translation start sites were corrected and also added to the database. With this database, a third search with identical settings was performed to gain additional peptide identifications to proof our results.

2.9 Confirmation of peptides for protein re-annotation

Synthetic peptides with isotopical labeled 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 adjustments according to the shotgun experiments.

Using these spectra a reference spectra 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^2 spectra with the reference library using NIST MS Search 2.0 identify search.

2.10 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 semi-proteases, meaning only either the N- or the C-terminal cleavage is required to be specific as identification criteria. 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-term were considered to be cleaved by signal peptidases if no other peptide belonging to the same protein was identified N-terminal to their peptide loci. Additionally, the minimum length of a signal peptide was considered to be 7 amino acids. Resulting signal peptidase cleavage sites were compared with computational signal peptide predictions of PerdiSi [48] and SignalP [49].

2.11 Peptide mapping and visualization

Identified peptides were mapped to the *H. pylori* genome using tblastn with an evalue of 10000 and the low complexity filter turned off. Perfect and full length sequence matches were used. If no such match has been found the maximal number of mismatches has been set to the number of Leucine and Isoleucine amino acids in the sequence is used and the best matching position is selected. The peptides were visualized in the UCSC browser [43, 47]. 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 athttp://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 gets initially 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 is indicated in the sequence identifier (peptide ID:#mappings:experiment).

2.12 RNAcode screen

The Multiz pipeline [50] was used to generate genome wide alignments of 22 epsilon proteobacteria (Supplemental Table 1). Alignments were scanned for protein coding potential regions using RNAcode [14] 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 has been extended by 51 nucleotides in each direction. This resulted in 18 short protein coding gene predictions not yet contained in the published gene annotations.

2.13 Submission to Pride and UniProtKB

For PRIDE [51] (http://www.ebi.ac.uk/pride) submission, we made an additional database search with Mascot and X!Tandem using the SearchGUI [52]. Therefore we searched against a NCBI database of *H. pylori* strain 26695 complemented with the sequence corrections, signal peptide cleavage sites and missing annotations with the same configurations as described in materials and methods. For pride xml export we used the software PeptideShaker (http://code.google.com/p/peptide-shaker/). The complete experimental data set accessed on the PRIDE web service.

3 Results

3.1 Proteome analysis of *Helicobacter pylori* strain 26695

To achieve broad coverage of the *Helicobacter pylori (HP)* proteome, we analyzed cell lysates by GeLC-MS and offline 2D-LC-MS. Furthermore we integrated the results published by Jungblut *et al.* [45]. Mascot and TandemX! were used to search spectra against a compiled database including (i) the NCBI data 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 Figure for an overview of the method). Peptide identification lists with according FDR calculations as well as a protein identification table are in the supplementary material.

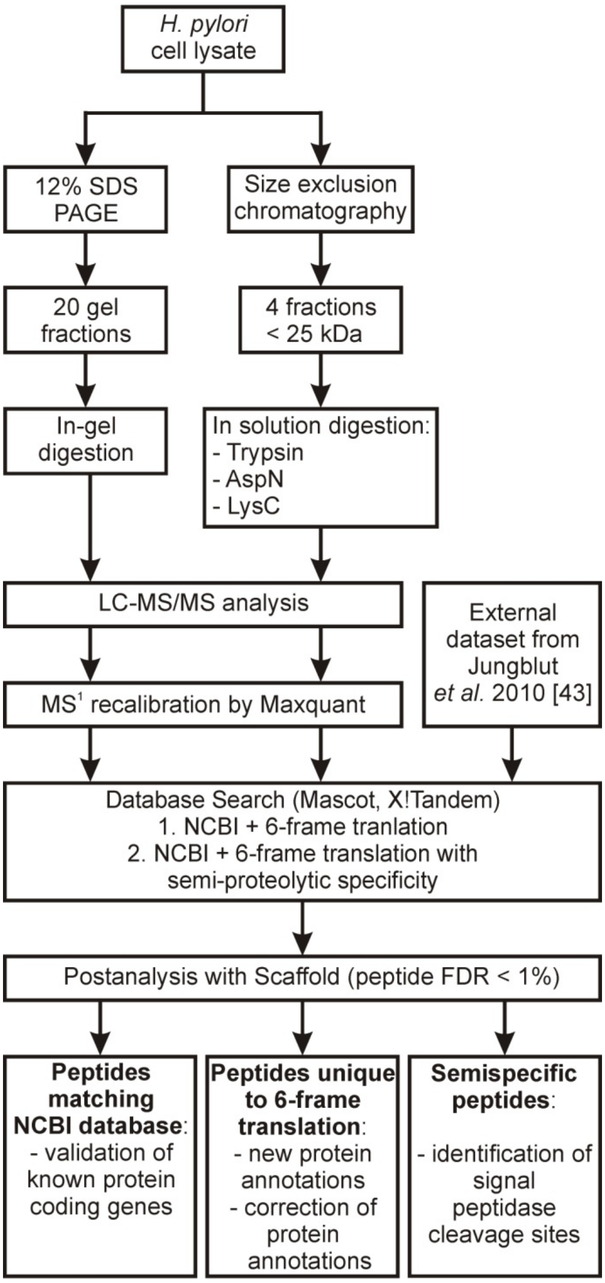
H. pylori cell lysate
12% SDS PAGE
Size exclusion chromatography
20 gel fractions
4 fractions < 25 kDa
In-gel digestion
In solution digestion:
- Trypsin
- AspN
- LysC
LC-MS/MS analysis
External dataset from Jungblut et al. 2010 [43]
MS1 recalibration by Maxquant
Database Search (Mascot, X!Tandem)
1. NCBI + 6-frame tranlation
2. NCBI + 6-frame translation with semi-proteolytic specificity
Postanalysis with Scaffold (peptide FDR < 1%)
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

Figure : Experimental workflow of the proteogenomic analysis. Proteins extracted from *H. pylori* cell lysates were separated by 12% and size exclusion chromatography. Gel fractions were digested by trypsin wherease trypsin, AspN and LysC were separately applied to size exclusion chromatography (SEC) fractions. Samples were analyzed by LC-MS/MS. MS^1 data was recalibrated using Maxquant. At this point the dataset of Jungblut *et al.* [45] was integrated. A database search against a reverse concatenated database of the NCBI entries and the six-frame translation was performed. Additionally a database serach with semi-proteolytic specificity was made. After postanalysis with Scaffold applying a peptide FDR of 1%, 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.

Peptide FDRs of all samples were calculated to be lower than 0.3% in our dataset (peptide identification supplement). 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 an identification overlap of 91% demonstrating good reproducibility.

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 identification of low MW proteins, showing 30.0% 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 proteome coverage of 71%.

In the Re-Analyses of the most comprehensive proteome dataset being published (Jungblut *et al.* [45]) 549 proteins corresponding to 35% of the proteome were identified. In comparison to our results only one additional protein (gi 15645950) could be identified. In contrast to our dataset, peptide FDRs of this dataset was higher than 1% for two fractionations (pellet fraction: FDR 1.1%, startline fraction: 3.1%).

3.2 Refinement of protein annotations by Proteogenomics

For identification of novel protein sequences, we searched against a reverse concatenated database including the NCBI database of *H. pylori* strain 26695 and a six-frame translation. Of the 21915 peptides being identified, 21717 could be mapped to the 1576 existing protein coding annotations. However, 198 peptides (0.9 %) were unique to the six-frame translation.

Those peptides were classified according to their genome location. Additionally, a BLAST analysis against the NCBI reference sequence database was applied to determine similar proteins in other *H. pylori* strains. 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 in the regions of interest to further validate our results. Hereby, sequencing errors resulting in frame shifts became obvious. The peptides which were used for identification of new respectively the existing protein coding sequences are shown in Supplement table 1. The following refinements of protein annotation were submitted to the UniProt database to make them publicly available.

3.2.1. Identification of missing protein annotations

We could identify four missing protein annotations. Three proteins were missing due to DNA sequencing errors resulting in frame shifts within a protein coding sequence. The fourth protein was simply missed during annotation [53].

Seven different peptides could be identified for the coding region HP0058 (Supplementary table 2) which was not annotated in the NCBI protein database. Medigue *et al.* [15] already 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 [54, 55]. Our results provide the experimental confirmation of this prediction. The protein sequence comprises 389 amino acids and a molecular weight (MW) of 45 kDa. Peptides were identified in fractions 12 and 13 (45-57 kDa) of both in-gel digestion replicates supporting this MW.

The protein coding annotation for the gene HP0744, was also missing in the NCBI protein database. We identified nine different peptides in this region (Supplementary table 2). Peptides for this region were identified in the same gel fraction (fraction 11, 35-45 kDa) of both biological replicates. Again, Medigue *et al.* [15] 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. Insertion of one nucleotide at the stop codon can correct the frame shift, so both parts of HP0744 are on frame -1.

As third finding, the gene HP0619 was not annotated as a protein coding sequence. We identified five peptides on frame +2 and nine peptides on frame +1 in this region (Supplement table 2). Once more, Medigue *et al.* [15] indicated a frame shift error in this region. In fact, by the change of the stop codon on frame +2 into a leucine code (according to the BLAST result gi 7465084) by addition of a thymine (at nucleotide position 665045) can correct the frame shift error.

Three different peptides identified a new protein coding gene (DNA 0100057) between the genes HP0585 and HP0586 (Figure). BLAST analysis matched the sequence to the ferrous iron transport protein A of foru different *H. pylori* strains (Lithuania75, SNT49, G27 and ELS37) with 100% identity and a expect value of 5×10^{-29}. Conclusively the ferrous iron transporter protein gene has been missed during annotation by Tomb et al. [53]. However, the identical protein sequence was already predicted by an unpublished observation of Medigue and Bocs (gi 13431987, P57798.1), but it was missing in the NCBI reference sequence database of strain 26695.

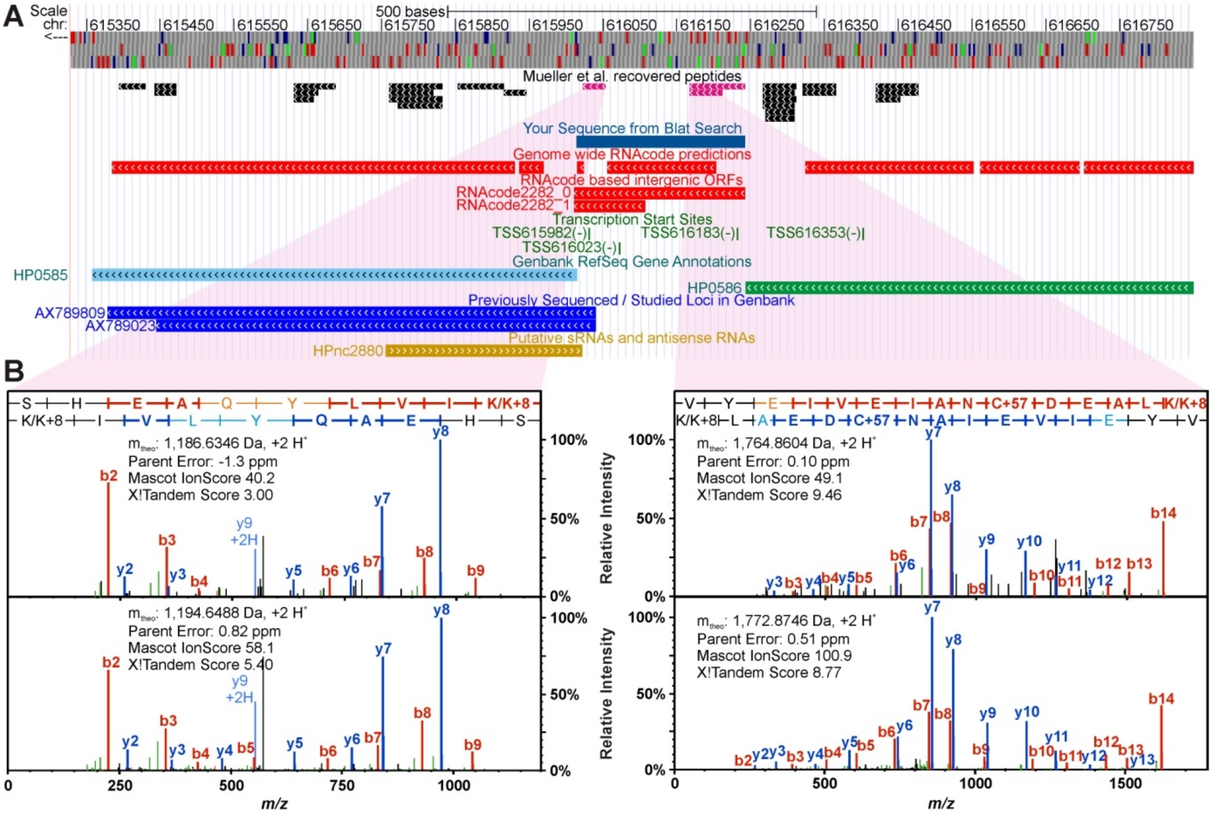

Figure : **(A)** Three peptides (magenta) mapped into the intergenic region of HP0585 (endonuclease III) and HP0586 (hypothetical protein). In addition two RNAcode predictions are found at this locus which can be extended to ORFs. Note that RNAcode2282_1 is a sub-region of RNAcode2282_0 and together 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 CID spectra of the experiment (upper spectra) and the corresponding synthesized peptide (lower spectra) containing $^{13}C_6{}^{15}N_2$-labelled lysines.

3.2.2. Identification of erroneously annotated translation start sites

We could detect four protein annotations with an extended sequence at the protein N-term. Translation start sites for two proteins were wrong due to frame shift errors which are a result of DNA sequencing errors. 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 (Figure). It is on the same frame as the downstream gene HP1433 and there is no stop codon between these sequences. In conclusion, the hypothetical protein HP1433 (gi 15646042) has a wrong start codon assignment. Protein annotations in other *H. pylori* strains annotate this protein including the identified peptide 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 that particular gene.

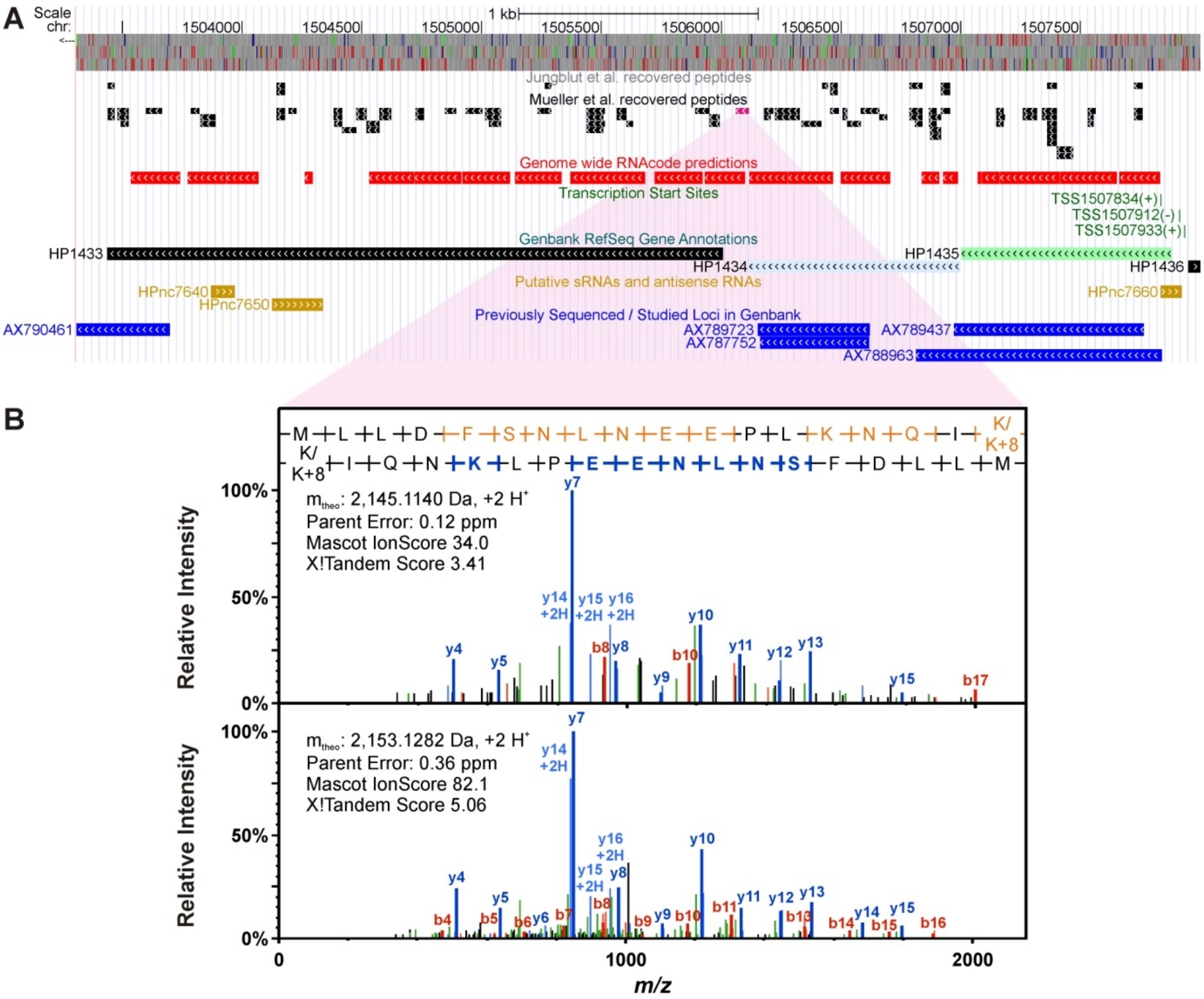

Figure : **(A)** Genomic location of HP1433 a hypothetical protein which is encoded in an operon together with the formyltetrahydrofolate hydrolase (HP1434) and the protease IV (HP1435). The operon is transcribed from the transcription start site TSS1507912 which is located upstream of the HP1435 gene. Beside putative anti-sense RNAs (HPnc yellow) several previously studied loci are annotated (blue). The latter correspond to a protein-protein interactions 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 CID spectra of the experiment (upper spectrum) and the corresponding synthesized peptide (lower spectrum) containing $^{13}C_6{}^{15}N_2$-labelled lysines.

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. In an additional 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. The two peptides are neither on the same frame as HP0760 (phosphodiesterase) nor HP0761 (hypothetical protein) (Supplementary table 2). BLAST analysis showed that both peptides match perfectly to phosphodiesterase of many other helicobacter strains. We conclude that the protein coding region HP0760 was truncated due to a frame shift error as suggesded by Medigue *et al.* [15]. 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 the gene 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. When GTG is a start codon, it is translated to methionine. The two peptides, which are N-terminal extended over the previously annotated start, show that this triplet is translated into valine at this position and thus increase 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.* [45]. 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 figure 1). The second database search including this protein sequence identified an additional peptide which is located upstream related 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 [10, 15]. 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 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 [48] and SignalP [49] could be a reasonable explanation for this mass difference.

The predicted coding region HP0694 (gi 15645317) is also wrongly annotated due to a DNA sequencing error downstream of the annotated C-terminus. The peptide VAFTITDISK belongs to a region next to the 3' end of the gen HP0694 (Supplementary table 2). Protein BLAST of this peptide revealed 100% identity with outer membrane proteins of other strains (e.g. strain J99; gi 15611701). An additional database search including this protein sequence succeeded in additional peptide identifications (Figure). 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 [48] and SignalP [49]. These findings strongly indicate a sequencing error resulting in a pre-major stop due to a frame shift error [15] for the predicted coding region 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.

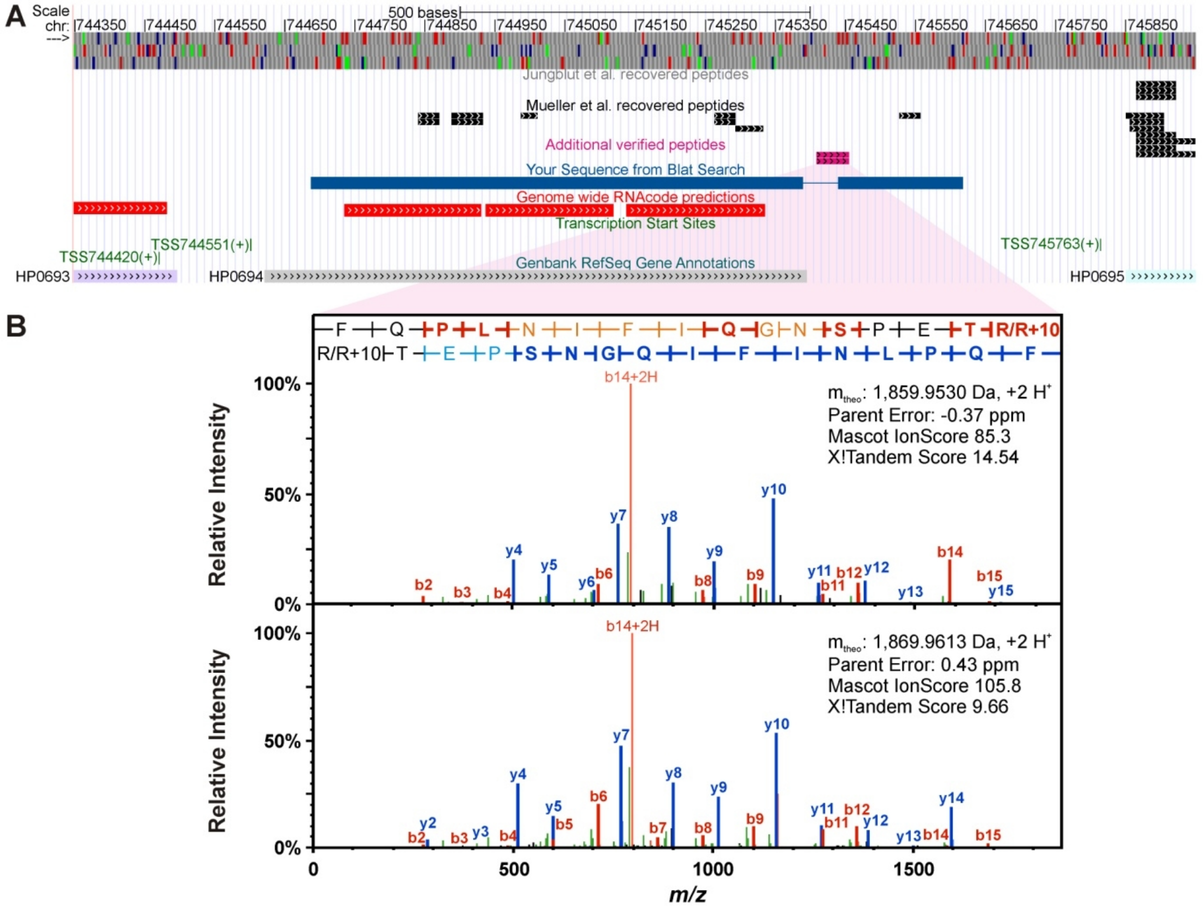

Figure : (A) Genomic location of HP0694 (hypothetical protein). HP0694 has two alternative transcription start sites (TSS744420 and TSS744551, green). The RNAcode prediction (red) resamples the annotated open reading frame. One peptide was identified in between the genes HP0694 and HP0695. The BLAST 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 data base. It was found in two biological replicates. (B) Confirmation of this peptide by comparison of the CID spectra of the experiment (upper spectrum) and the corresponding synthesized peptide (lower spectrum) containing $^{13}C_6{}^{15}N_2$-labelled lysines.

3.2.4. Validation of new identified 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 correlate well for all tested peptides and further validates the above described revised gene annotations (Figure 2-4, Supplementary figures 2-15). The reverse match score as well as the correlation probability of NIST MS search are listed in supplementary table 2.

3.3 Identification of signal peptides

Signal peptide cleavage leads to a new protein start. After enzymatic digestion, peptides of new protein N-termini have a specifically cleaved C-terminus but a non-specifically cleaved N-terminus according to the used protease. Therefore, peptides near the protein N-termini with non-specific cleaved N-terminus were considered to be cleaved by a signal peptidase. Signal peptide sequences were identified by a database search allowing semi-specific peptides.

We checked for signal peptide sequences, but neither the UniProt nor the NCBI database have signal peptides included for *H. pylori* strain 26695. Computational tools like PerdiSi [48] and SignalP [49] provide 191 and 182 significant predictions, respectively.

Overall, we identified 72 previously unknown signal-peptide sequences with our dataset. 49 signal peptide sequences were identified in more than one sample.. Analysis of the dataset from Jungblut *et al.* [45] offered validation of seven signal peptides and identification of three new sequences. As shown in Figure , comparison of the identified signal peptide sequences with predictions from the web-based tools PerdiSi [48] and SignalP [49] predict LXA and AXA as the dominant recognition sequence for signal peptidases. Our data shows that leucine (65.3%) is predominately localized at the -3 position relative to the cleavage site. The -1 position is mainly alanine (77.0%). About 64.0% of the identified signal peptides were validated by PerdiSi [48] and/or SignalP [49] by significant (40, 53.3%) or non-significant predictions (8, 10.7%) whereas 27 (36.0%) signal peptides were not predicted at all. Subcellular localization of the proteins with identified signal peptides was predicted by SOSUI-GramN [56]. Predicted localizations of proteins with signal peptide cleavage are shown in Figure , D. Most proteins cleaved by signal peptidases are considered to be located in the inner- and outer-membrane.

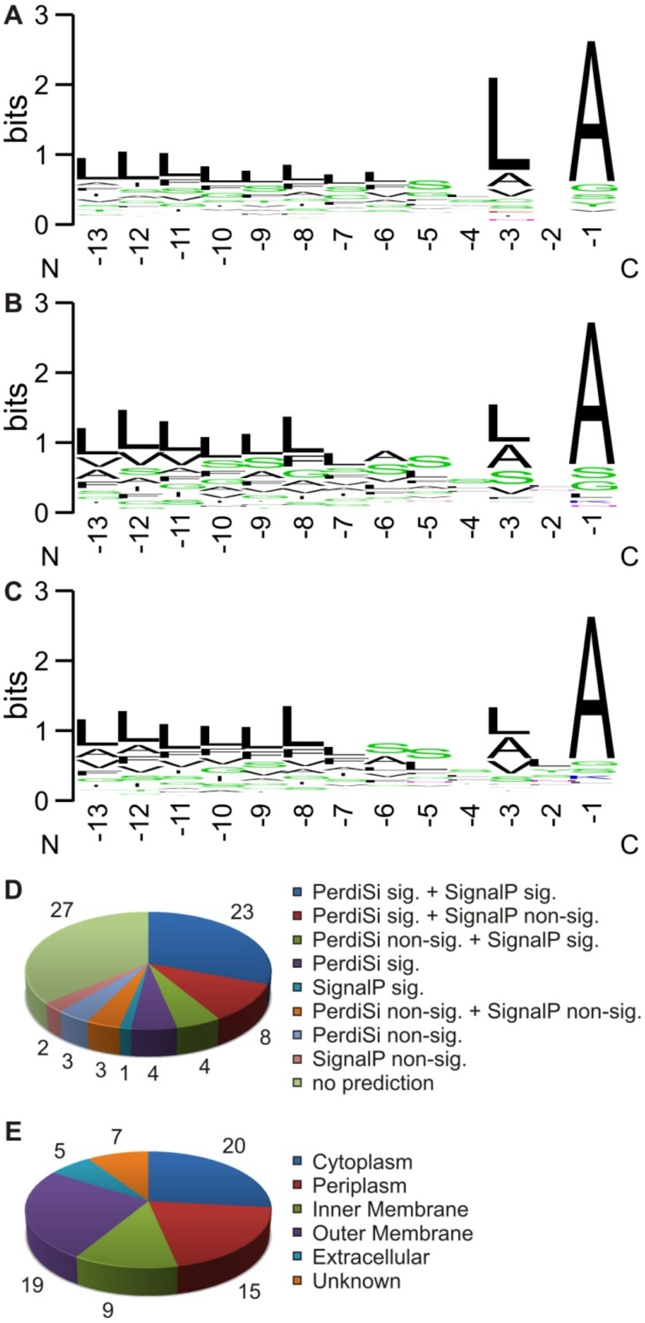

Figure : Comparison of identified signal peptide sequences with software predictions and protein localization. (**A-C**) Sequence logo of the identified signal peptides (**A**), significant signal peptide predictions derived from Perdisi (**B**) and signal peptide predictions derived from SignalP (sequence logos were created with the web-based tool WebLogo, Version 2.8.2). (**D**) Comparison of identified signal peptide sequences with software predictions of Perdisi and SignalP (sig.: significant) (**E**) Prediction of subcellular protein localization according to SOSUI-GramN.

4 Discussion

H. pylori is a Gram-negative human pathogen responsible for many gastric diseases like gastric and duodenal ulcers as well as gastric cancer. About 50% of the world's population is infected with *H. pylori*, but approximately 80% of the individual carriers are asymptomatic [57, 58]. The complete sequencing of the strain 26695 [53] in 1997 facilitates studies of *H. pylori* on genome, transcriptome and proteome level. Proteomics studies of *H. pylori* are an inherent part of basic research of this pathogen. For example response to acidic [59] or oxidative stress [60, 61] as well as pathogenic mechanisms [62, 63] were investigated in proteomics studies. Those studies are strongly dependent on the protein database quality. Proteogenomics is well suited to improve database quality and a high proteome and protein sequence coverage of MS data is required.

Our study revealed 1115 proteins representing 71% of the annotated proteome with an average protein sequence coverage of 49%. A transcriptome study of *H. pylori* strain 26695 in 2010 showed that 88% of the annotated protein coding genes were expressed [64]. Since our samples were taken from the same strain, we consider that around 81% of all expressed proteins were covered with our approach. A similar proteogenomics study of *Pseudomonas fluorescens* Pf0-1 covered 66% of the annotated and identified 16 new ORFs [65] which is comparable to our results. However we still miss 28% either due to false annotations or experimental limitations such as the detectable minimum protein weight of approximately 5 kDa with our approach. The latter might also be a reason why we miss the recently discovered short transcripts harboring conserved open reading frames [64].

Comparative genome studies already showed that there might be discrepancies in coding sequence annotation of different *H. pylori* strains as result of either DNA sequencing errors or erroneously predictions [10, 15, 16]. However, most of this data is solely based on bioinformatics and not validated by biological experiments. Proteogenomics fills the gap of predictions and reality. Our dataset allowed us to unambiguously correct six protein annotations (HP1433, HP0105, HP0760, HP0564, HP1186, HP0694) and to discover four new protein annotations (HP0058, HP0744, HP0619, intergenic region HP0585-0586 – ferrous iron transport protein A). Five of these protein annotations were additionally validated by comparison of MS/MS spectra of our biological sample with those from synthetic peptides. Furthermore, X 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 erroneously 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.

Two of the new annotated proteins might be interesting for further studies. The protein which is located between HP0585 and HP0586 is similar to the ferrous iron transport protein A in other strains. This protein might be interesting in future studies because iron transport is essential for the survival of H. pylori in the stomach [66]. The previously missing annotation for the protein coding region HP0619 codes for a putative lipopolysaccharide (LPS) biosynthesis protein. This protein might be drug target for inhibition of the LPS biosynthesis pathway [67].

Furthermore, we investigated signal peptide cleavage sites of the annotated proteins. *H. pylori* is coding for two different signal peptidases [68]. 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 resulting in more restrictive but much more significant signal peptide identifications.

Based on the reanalyzed and integrated dataset from Jungblut *et al.* [45], we could validate five signal peptide sequences by an independent experiment and identify three additional ones. Since this data was not acquired by high accurate MS, the quantity of identifications is lower compared to the 72 signal peptides in our dataset. Signal peptidases from gram negative bacteria require more or less conserved amino acids at the -1 and -3 positions relative to the cleavage site [68]. We showed that the predominant recognition sequence for the signal peptidases of *H. pylori* is LXA. Nevertheless other aliphatic amino acids like alanine or valine can replace leucine at the -3 position, whereas glycine, serine 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 signal peptidase I. To our knowledge no other study has investigated the specifity of the signal peptidases of H. pylori. Signal peptidases are essential enzymes for the viability of bacterial cells [68, 69]and are involved in pathogenesis [70, 71] Therefore signal peptidases could be novel targets for antibiotics [69]. Additionally, inclusion of signal peptides to the database could increase peptide and protein identifications of future proteome studies.

Both, signal peptidases cleavage sites, corrected and missing protein annotations were submitted to the UniProt protein database as well as in the UCSC genome browser to support further proteome and transcriptome studies.

In conclusion, using proteogenomics approaches for protein coding sequence annotations will help to improve and complete protein databases. We expect that further proteomics 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. 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 S. R. Pernitzsch and C. M. Sharma (Research Center for Infectious Diseases Research, Würzburg) for providing cell pellets of *H. pylori* strain 26695 and critical comments on the manuscript. 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).

Appendix. Supplementary material

5 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] Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007;5:e254.

[5] 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.

[6] 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.

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

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

[9] Malys N, McCarthy JE. Translation initiation: variations in the mechanism can be anticipated. Cellular and molecular life sciences : CMLS. 2011;68:991-1003.

[10] Boneca IG, Reuse Hd, 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

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

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

[26] 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.

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

[28] Scigelova M, Hornshaw M, Giannakopulos A, Makarov A. Fourier Transform Mass Spectrometry. Molecular & Cellular Proteomics. 2011;10.

[29] 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.

[30] 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.

[31] 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.
[32] 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.
[33] 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.
[34] Rockstroh M, Müller S, Jende C, Kerzhner A, Bergen Mv, Tomm JM. Cell fractionation - an important tool for compartment proteomics2010.
[35] Lee YH, Tan HT, Chung MCM. Subcellular fractionation methods and strategies for proteomics. Proteomics. 2010;10:3935-56.
[36] Doucette AA, Tran JC, Wall MJ, Fitzsimmons S. Intact proteome fractionation strategies compatible with mass spectrometry. Expert Rev Proteomics. 2011;8:787-800.
[37] 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.
[38] 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.
[39] 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. Molecular & Cellular Proteomics. 2012;11.
[40] 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.
[41] 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.
[42] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods. 2008;5:621-8.
[43] Chan PP, Holmes AD, Smith AM, Tran D, Lowe TM. The UCSC Archaeal Genome Browser: 2012 update. Nucleic Acids Res. 2012;40:D646-52.
[44] 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.
[45] 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.
[46] 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.
[47] Schneider KL, Pollard KS, Baertsch R, Pohl A, Lowe TM. The UCSC Archaeal Genome Browser. Nucleic Acids Res. 2006;34:D407-10.
[48] 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-W9.
[49] Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Meth. 2011;8:785-6.
[50] 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.
[51] 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-D42.

[52] 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.

[53] 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.

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

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

[56] Imai K, Asakawa N, Tsuji T, Akazawa F, Ino A, Sonoyama M, et al. SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in gram-negative bacteria. Bioinformation. 2008;2:417-21.

[57] 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.

[58] Suzuki R, Shiota S, Yamaoka Y. Molecular epidemiology, population genetics, and pathogenic role of Helicobacter pylori. Infection, Genetics and Evolution. 2012;12:203-13.

[59] 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.

[60] 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.

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

[62] 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. Disease Models & Mechanisms. 2010;3:605-17.

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

[64] 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.

[65] Kim W, Silby MW, Purvine SO, Nicoll JS, Hixson KK, Monroe M, et al. Proteomic Detection of Non-Annotated Protein-Coding Genes in <italic>Pseudomonas fluorescens</italic> Pf0-1. PLoS One. 2009;4:e8455.

[66] 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.

[67] 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.

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

[69] 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.

[70] 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.

[71] 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.