High Accuracy Mass Spectrometry in Refinement of Genome Annotation

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Karsten Krug
aus Staßfurt

Tübingen
2013

Tag der mündlichen Qualifikation: 23.07.2013

Dekan: Prof. Dr. Wolfgang Rosenstiel

1. Bericherstatter: Prof. Dr. Boris Macek

2. Berichterstatter: Prof. Dr. Olaf Rieß

3. Berichtertatter: Prof. Dr. Matthias Selbach

Danksagung

Mein ganz besonderer Dank gilt meinen Eltern, Heidemarie und Werner Krug, die mir ermöglicht haben, meinen Weg zu gehen, und die mich dabei die vielen Jahre unterstützt haben.

Prof. Dr. Boris Macek möchte ich für seine intensive Betreuung dieser Arbeit, für seine ständige Diskussionsbereitschaft sowie für die Bereitstellung einer hervorragenden Arbeitsumgebung am Proteome Center Tübingen danken.

Mein Dank gilt ebenfalls Prof. Dr. Olaf Rieß für seine Bereitschaft, diese Arbeit zu betreuen.

Allen gegenwärtigen und ehemaligen Mitarbeitern des Proteome Centers Tübingen danke ich für das freundliche Arbeitsklima, insbesondere danke ich Alejandro Carpy, Dr. Boumediene Soufi, Dr. Mirita Franz-Wachtel, Ulrike Grammig, Katarina Matic, Dr. Sven Nahnsen, Dr. Nicole Sessler, Sasa Popic, Vaishnavi Ravikumar, Dr. Nelson C. Soares, Gesa Behrends, Philip Spät, Dr. Ana Velic, Johannes Madlung, Silke Wahl und Irina Droste-Borel.

Ulrike Grammig danke ich weiterhin für ihre allgegenwärtige Unterstützung bei allen organisatorischen Angelegenheiten.

Allen Co-Autoren der Manuskripte, welche zu dieser Arbeit beigetragen haben, möchte ich für die angenehme und erfolgreiche Zusammenarbeit danken, insbesondere danke ich Dr. Nadine Borchert, Dr. Christoph Dieterich, Dr. Amit Sinha und Prof. Dr. Ralf J. Sommer.

Abschließend danke ich Prof. Dr. Alfred Nordheim, der mir der Anstellung die Möglichkeit eröffnet hat, am Proteome Center Tübingen zu arbeiten und zu forschen.

List of publications related to this thesis

Teile der vorliegenden Arbeit wurden bereits veröffentlicht oder wurden zur Veröffentlichung eingereicht:

1. **Karsten Krug**, Sven Nahnsen, and Boris Macek:
   Mass spectrometry at the interface of proteomics and genomics.
   Mol. BioSyst. 2011, 7, 284-291. doi:10.1039/c0mb00168f
   Reproduced by permission of The Royal Society of Chemistry

2. **Karsten Krug**, Alejandro Carpy, Gesa Behrends, Katarina Matic, Nelson C. Soares, and Boris Macek:
   Deep Coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments.
   Under revision in *Mol Cell Proteomics*

3. Nadine Borchert*, Christoph Dieterich*, **Karsten Krug***, Wolfgang Schütz, Stephan Jung, Alfred Nordheim, Ralf J. Sommer, and Boris Macek:
   Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models.
   Genome Res. 2010, 20, 837-846. doi: 10.1101/gr.103119.109

4. Nadine Borchert*, **Karsten Krug***, Florian Gnad, Amit Sinha, Ralf J. Sommer, and Boris Macek:
   Phosphoproteome of *Pristionchus pacificus* provides insights into architecture of signaling networks in nematode models.
   Mol Cell Proteomics 11, 1631-1639. doi: 10.1074/mcp.M112.022103

* = equal contribution

Table of contents

List of abbreviations ... 3

Summary ... 4

Zusammenfassung ... 6

1. Introduction ... 8
  - 1.1. Advances in genomics ... 9
  - 1.2. Genome annotation ... 11
  - 1.3. Mass spectrometry-based proteomics ... 13
    - 1.3.1. Sample preparation ... 15
    - 1.3.2. LC-MS/MS ... 16
    - 1.3.3. Data processing and database search ... 19
    - 1.3.4. Quantitative proteomics ... 23
  - 1.4. Proteogenomics - Mass spectrometry at the interface of proteomics and genomics 24
    - 1.4.1. Historical overview ... 25
    - 1.4.2. Challenges in proteogenomic experiments ... 26
  - 1.5. Phosphoproteomics ... 29
    - 1.5.1. Ser/Thr/Tyr Phosphorylation ... 30
    - 1.5.2. Phosphopeptide enrichment ... 31
    - 1.5.3. MS analysis of phosphopeptides ... 32
    - 1.5.4. Identification of phosphorylation sites ... 33
  - 1.6. Aims of the thesis ... 34

2. Results ... 35
  - 2.1. Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments ... 35
    - 2.1.1. Synopsis: ... 35
    - 2.1.2. Contributions: ... 36
  - 2.2. Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models. ... 37

2.2.1. Synopsis 37
2.2.2. Contributions 38
2.3. Phosphoproteome of Pristionchus pacificus provides insights into architecture of signaling networks in nematode models. 39
2.3.1. Synopsis 39
2.3.2. Contributions 40

3. Conclusions 41

References 43

Manuscripts related to this thesis 48
i. Mass spectrometry at the interface of proteomics and genomics 49
ii. Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments 57
iii. Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models 84
iv. Phosphoproteome of *Pristionchus pacificus* Provides Insights into Architecture of Signaling Networks in Nematode Models 97

Curriculum vitae 109
Complete list of publications 110

List of abbreviations

CID	collision-induced dissociation
Da	Dalton
EST	expressed sequence tag
FDR	false discovery rate
HMM	hidden Markov model
HCD	higher energy collisional dissociation
HPLC	high-performance liquid chromatography
LC	liquid chromatography
m/z	mass-to-charge ratio
MMA	mixture-model approach
MQ	MaxQuant
MS	mass spectrometry
MS/MS	tandem mass spectrometry
NGS	next generation sequencing
ORF	open reading frame
PEP	posterior error probability
ppm	parts per million
PSM	peptide spectrum match
PTM	posttranslational modification
SNP	single nucleotide polymorphism
TDA	target-decoy approach
Th	Thomson
TPP	Trans-Proteomic Pipeline

Summary

Major improvements in DNA sequencing technologies during the last decade gave rise to “next generation sequencing” (NGS) technology, that enables routine sampling of entire genomes and transcriptomes of various organisms; however, the annotation of the raw genome sequence remains a big challenge for *ab initio* gene prediction programs. Experimental evidence of gene expression at the RNA and protein level can be used to train the machine learning algorithms and greatly improves accuracy of the resulting gene predictions. While NGS can provide gene expression data at the transcript level, translational evidence of genes on a large scale can only be addressed using mass spectrometry (MS)-based proteomics. Moreover, this technology is an indispensable tool to study regulatory post translational protein modifications (PTMs) such as phosphorylation. In this work I studied to what extent high accuracy MS-based proteomics can contribute to refining genome sequencing data, which is in focus of a fast-evolving research field termed “proteogenomics”. I first addressed the main parameters of a simple proteogenomic experiment, such as the actual false discovery rate of protein database search and sequence coverage of a bacterial genome using state-of-the-art MS technology. To that end I used a comprehensive proteome dataset of the model gram negative bacterium *Escherichia coli*, comprising its complete expressed proteome in exponential growth, and applied this approach to its well characterized genome. This analysis demonstrated a substantial underestimation of the false discovery rate in a commonly used proteogenomics workflow and pointed to the need for further improvement of sequence coverage in shotgun proteomic experiments. I further demonstrated the utility of proteogenomics in annotation of protein coding regions of a complex, eukaryotic genome on the example of *Pristionchus pacificus*, a model nematode increasingly used in evolutionary biology. The application led to the identification of several thousand novel peptide sequences that were used, together with transcriptomic data, to refine the existing genome annotation. Finally, I studied functional aspects of the refined *P. pacificus* proteome by using data from an in-depth phosphoproteomic study which enabled me to describe functional categories of

detected *P. pacificus* phosphoproteins, to define its kinome and to perform a comparative analysis with a recent phosphoproteomics study of the model nematode *Caenorhabditis elegans.* Taken together, this work demonstrates the value of high accuracy MS based proteomics in refinement of genome sequencing data.

Zusammenfassung

Im Verlauf des letzten Jahrzehnts führten wesentliche Verbesserungen der Techniken zur DNA Sequenzierung zu einer neuen Generation von Sequenzierungstechnologien („next generation sequencing“, NGS), welche eine routinemäßige Sequenzierung ganzer Genome und Transkriptome verschiedenster Organismen ermöglichte. Die Annotation der Genomsequenz stellt nach wie vor eine Herausforderung für Programme zur *ab initio* Genvorhersage dar, welche auf Algorithmen des maschinellen Lernens basieren. Experimentelle Bestätigung von Genexpression auf RNA- und Proteinebene kann dazu verwendet werden, die Genauigkeit der Genvorhersagen enorm zu verbessern. Während NGS Technologien Genexpressionsdaten auf der Ebene der Transkription generieren, kann die Bestätigung der Translation global nur mittels Massenspektrometrie (MS)-basierter Proteomik analysiert werden. Darüber hinaus stellt diese Technologie ein unverzichtbares Werkzeug zur Untersuchung regulatorischer, posttranslationaler Proteinmodifikationen (PTM), wie zum Beispiel Phosphorylierung, dar. In dieser Arbeit untersuche ich, in welchem Umfang hochgenaue, MS-basierte Proteomik zur Verbesserung der Annotation von genomischen Sequenzierdaten beitragen kann, welches im Fokus einer sich rasant entwickelten Forschungszweigs namens „Proteogenomik“ steht. Zuerst untersuche ich grundlegende Parameter eines einfachen proteogenomischen Experimentes, wie zum Beispiel die eigentliche Fehlerrate (false discovery rate, FDR) und Sequenzabdeckung eines bakteriellen Genoms mittels modernster MS Technologie gewonnener Daten. Hierzu verwende ich einen umfassenden Proteomdatensatz des gram-negativen Modelbakteriums *Escherichia coli,* bestehend aus allen exprimierten Proteinen der exponentiellen Wachstumsphase, und wende diesen auf das sehr gut charakterisierte Genom des Bakteriums an. Dieser Versuch zeigte eine erhebliche Unterschätzung der Fehlerrate (FDR) einer häufig verwendeten Vorgehensweise, und deutete auf die Notwendigkeit hin, die Sequenzabdeckung MS-basierter Proteomik zu verbessern. Des Weiteren demonstriere ich den Nutzen eines proteogenomischen Experiments bei der Annotation Protein kodierender Bereiche eines komplexen, eukaryotischen Genoms am

Beispiel des Fadenwurms *Pristionchus pacificus*, welcher vermehrt als Modellorganismus in der Evolutionsbiologie verwendet wird. Das Experiment führte zur Identifikation mehrerer Tausend, bisher unbekannter Peptidsequenzen. Diese wurden zusammen mit Transkriptionsdaten dazu verwendet, die existierende Annotation des Genoms zu verbessern. Abschließend betrachte ich die verbesserte Annotation des *P. pacificus* Proteoms, um dessen funktionelle Aspekte zu untersuchen. Dazu verwende ich Daten eines MS-basierten Experiments zur globalen Identifikation von Proteinphosphorylierungsstellen, um die phosphorylierten Proteine funktionell zu chrakterisieren, das Kinom des Organismus zu bestimmen und die gewonnenen Ergebnisse mit einer jüngst veröffentlichten Studie des Phosphoproteoms des Modellorganismus *Caenorhabditis elegans* zu vergleichen. Zusammengenommmen demonstriert diese Arbeit den Nutzen hochgenauer MS-basierter Proteomik in der Verbesserung von Genomsequenzierungsdaten

1. Introduction

During the last decades methods and technologies used in bioscience have advanced dramatically and extended the scope of biological studies to a system-wide analysis within a single experiment. Currently, high-throughput analytical methods perform measurements in a massively parallel manner can routinely be used to study how biological information is transmitted from genes to their products and how it influences major processes in the cell. The technologies to sequence DNA, RNA and proteins have undergone a revolution in terms of quantity and quality of the generated data. The 'next-generation sequencing' (NGS) technologies for DNA and RNA sequencing comprise the methods that evolved after the classical Sanger method, which is referred to as the 'first generation' sequencing. The NGS technology enabled single laboratories to generate data that previously required concerted effort of large-scale sequencing centers and led to the deciphering of genomes of many different organisms. Furthermore, this technology allowed the analysis of gene expression by sequencing gene transcripts (RNA-seq) and complemented the already mature microarray technology for gene expression analysis. On the other hand mass spectrometry (MS) emerged as method of choice to study gene expression at the protein level. Due to recent developments in the MS technology as well as the optimization of biochemical protocols used for efficient sample preparation, MS-based proteomics is approaching the routine identification of nearly all proteins expressed at the point of analysis together with their post translational modifications. Both, NGS and proteomics technologies require efficient bioinformatic tools for the analysis of the gathered data in order to assemble the correct genome sequence and to identify the correct protein sequences, respectively. The assembled genome sequence has to be subsequently annotated. One of the first steps in the annotation process is to determine the positions of genes on the genome, which is commonly done by *ab inito* gene prediction programs. The resulting databases contain the predicted set of gene sequences of an organism. These sequences are typically *in-silico* translated into amino acid sequences and used to search the acquired MS data in order to identify the corresponding proteins. This database-driven approach is well established in proteomics but restricts the identification to proteins that are present in the corresponding databases. In cases where there is only insufficient gene annotation available, e.g. an early draft annotation of a newly sequenced organism, proteomics data can be directly applied to the raw genome sequence and the resulting data can be used to refine the existing annotation draft. This approach, often referred to as proteogenomics, is the focus of this thesis.

Here, I will first give a brief introduction on recent advances in genomics and outline the principle behind *ab initio* gene prediction. I will then introduce the experimental workflow of a typical MS based proteomics experiment and review the challenges that specifically occur in a proteogenomics experiment. I will explain MS-based analysis of protein phosphorylation in the last part of the introduction before I summarize the results of the manuscripts associated to this thesis and give detailed description of my contribution to these studies.

1.1. Advances in genomics

The automated Sanger sequencing for genome analysis dominated the field of genomics for almost two decades and led to numerous groundbreaking studies such as the completion of the first human genome sequence (Lander et al. 2001; Venter et al. 2001; Collins et al. 2004). This method, considered as the ‘first-generation’ method, is based on the principle of dideoxynucleotide chain termination reaction introduced with the classical Sanger method (Sanger et al. 1977) but using differentially dye-labeled dideoxynucleotides that enabled an automated read-out of the nucleotide sequence within a single reaction. However, this approach was very expensive, time consuming and labor intensive leading to the necessity to develop more efficient sequencing strategies. During the last decade technical advances have led to improved sequencing methods that are referred to as ‘next-generation sequencing’ (NGS) and that can inexpensively produce large volumes of sequencing data. Several different sequencing technologies are available and excellently reviewed in the literature (Ansorge 2009; Metzker 2010). The most frequently used platforms are Solexa (Illumina), 454 FLX

(Roche), and to a lesser extent SOLiD (ABI). These platforms rely on diverse sequencing protocols each having very specific assets and drawbacks that make them differentially suitable for specific biological applications (**Table 1**). The general strategy common across these different platforms starts with the construction of an appropriate DNA library by random fragmentation of starting DNA and ligation of resulting fragments with custom linkers. The sequencing reactions rely either on reversible terminator sequencing (Illumnia), pyrosequencing (Roche), or sequencing by ligation (ABI) and are performed step by step together with the detection of the sequenced nucleotide. The identification of the sequenced nucleotides is based on the detection of a fluorescence dye or emitted light, and thus requires sensitive imaging technology. Typically, the DNA libraries are amplified in order to obtain sufficient light signal intensity for reliable detection. The nucleotide sequence can then be reconstructed from the recorded imaging signals.

TABLE 1: OVERVIEW OF THE MOST FREQUENTLY USED NGS PLATFORMS. DATA IS TAKEN FROM (METZKER 2010)

Platform	Run time (days)	Acquired data per run (Gb)	Read length (bp)	Pros	Cons
SOLiD	7-14	30-50	50	Inherent error correction	Long run times
Solexa	4-9	18-35	75 or 100	Most widely used platform	Low multiplexing capabilities of samples
454 FLX	0.35	0.45	330 (average)	Longer reads improve mapping in repetitive regions; fast run times	High reagent costs; high error rates in homopolymer repeats

All three technologies produce millions of short sequence reads ranging from several tens of base pairs (bp) to several hundred bp. Therefore, the first step in the analysis of NGS data is to assemble sequences into longer contigs and, ultimately, into the complete genome sequence. Typically, the reads are assembled by aligning them to a references genome (re-sequencing) or *de novo* assembly which is necessary in case of a completely unknown genome and which is highly demanding in terms of efficient algorithms and hardware resources. (Zhang et al. 2011).

The development of efficient DNA enrichment technologies enabled the sequencing of targeted regions of a genome (genomic capture). Exome capture allows the analysis of the complete protein-coding region of the genome (Teer and Mullikin 2010). Total DNA is fragmented and applied to probes complementary to the desired DNA sequences. The target DNA fragments hybridize to the probes, the non-targeted sequences are washed away, and the enriched DNA is subsequently eluted for sequencing. Targeted analysis and sequencing of gene transcripts (RNA-seq) revolutionized the field of transcriptomics by providing a far more precise measurement of levels of transcripts and their isoforms than other methods (Wang et al. 2009b). Protocols to isolate different RNA populations implement poly(A) selection to enrich poly(A+)-transcripts (mRNA, microRNA, snoRNAs), or the depletion of ribosomal RNA to extract other non-coding and protein-coding RNA that do not possess poly(A) tails, including histone mRNAs, tRNAs and certain small RNAs (Cui et al. 2010). Longer RNAs are fragmented and the resulting fragments are used to construct cDNA libraries for sequencing. The data derived from transcriptomic studies using NGS technology provide experimental evidence on gene expression in high-throughput and greatly facilitate the discrimination between coding and non-coding parts of the genome (Nagalakshmi et al. 2008).

1.2. Genome annotation

Methods to find genes in genome sequence have evolved since the early days of genetics. The determination of gene positions along the DNA sequence is usually the first step in the annotation of a newly sequenced genome. Classical gene finding approaches comprised sequencing of randomly chosen cDNA clones, searching genomic regions that are similar to proteins in databases, and manual *de novo* annotation by human curators (Brent 2005). Such experiments were extremely painstaking and expensive and were not capable to find genes in a genome sequence at global scale and reasonable expense. Presently, gene finding has turned into a computational problem enabling an automated and inexpensive annotation of a genome sequence which is required to keep up with the vast amount of raw sequencing data produced by NGS technologies. Computational approaches can be divided into homology-based and *ab initio* methods, although many gene finding programs are hybrids (Knapp and Chen 2007). The underlying algorithms of *ab intio* gene finders recognize specific structural patterns such as start/stop codons and promoter sequences to define the positions of genes along the DNA sequence. Due to the relatively simple gene structure, gene prediction is straightforward and existing gene finding tools achieve an accuracy of over 90% on both levels of sensitivity and specificity in prokaryotic and some simple eukaryotic organisms (Knapp and Chen 2007). Eukaryotic genomes pose a much greater challenge for gene finding tools due to the much larger size and complex gene structure consisting of coding sequences (exons) intervened by non-coding sequences (introns). The majority of gene finders is based on hidden Markov models (HMM) that take a sequence of inputs (DNA nucleotides) and a set of classes and assign a class (exons, introns, TATA boxes, etc.) to each individual input. The gene finders are trained on a specific genome, like the human genome, that should be taken into account when applied to new genomes (Korf 2004). Thus, any gene finder has specific patterns of inaccuracies, and their performance can be evaluated at the level of nucleotides, exons, and whole genes. At the nucleotide level, the nucleotide might be incorrectly predicted to be in a coding region or vice versa. At the exon level, several other scenarios might occur. An exon can be only partially correct or overlapping, completely missed, or incorrectly predicted. At the level of whole genes, the gene finder might miss a gene completely. Popular implementations of HMM gene finders are Augustus (Stanke and Waack 2003), GlimmerHMM (Majoros et al. 2004) and SNAP (Korf 2004), their performance besides other tools are reviewed in (Knapp and Chen 2007). Typical values for the fraction of correctly predicted exons of the three gene predictors varied between 35-61% when applied to a human test dataset (FSH298) consisting of 298 genes with 2555 coding exons (Knapp and Chen 2007). The subsequent *in silico* translation of resulting gene predictions represent the theoretical proteome of an organism. These databases provide an essential resource for the proteomics community as these databases are typically used to search the acquired mass spectra in order to identify the correct amino acid sequences of the analyzed proteins.

1.3. Mass spectrometry-based proteomics

Mass spectrometers enable the analysis of the elemental composition of molecules by measuring their mass-to-charge ratio. MS instruments typically consist of an ion source, a mass analyzer and a detector. Analyzed molecules are first ionized to produce charged ions which are separated in a mass analyzer according to their mass-to-charge ratios and subsequently detected, resulting in the mass spectra of the analyzed molecules (Steen and Mann 2004). The use of this technology to analyze biopolymers emerged during 1980s and was revolutionized by invention of two soft ionization methods, matrix-assisted laser desorption/ionization (MALDI)(Hillenkamp et al. 1991) and electrospray (ESI)(Fenn et al. 1989), which were awarded the Nobel Prize in Chemistry in 2002. Mass spectrometry gradually displaced the classical technique for protein analysis, known as Edman degradation, due to higher speed and sensitivity and enabled global analysis of proteins present in a cell, tissue or an organism, which today is termed proteomics (Steen and Mann 2004). Since that time the technology improved constantly and today several different MS platforms and instruments are in routine use for proteomics applications (Ahmed 2008). The different MS instruments can be roughly classified according to their achieved accuracy of measured mass-to-charge ratios that is expressed in parts per million (ppm) relative to the m/z ratio of the measured ion. For example, a peptide ion of 1000 Da measured with 1 ppm accuracy results in an absolute mass error of 0.001 Da. Low accuracy instruments (ion traps, triple quadrupoles) enable high speed acquisition, but the resulting data suffers from poor resolution and mass accuracy, which is typically 200 -500 ppm.

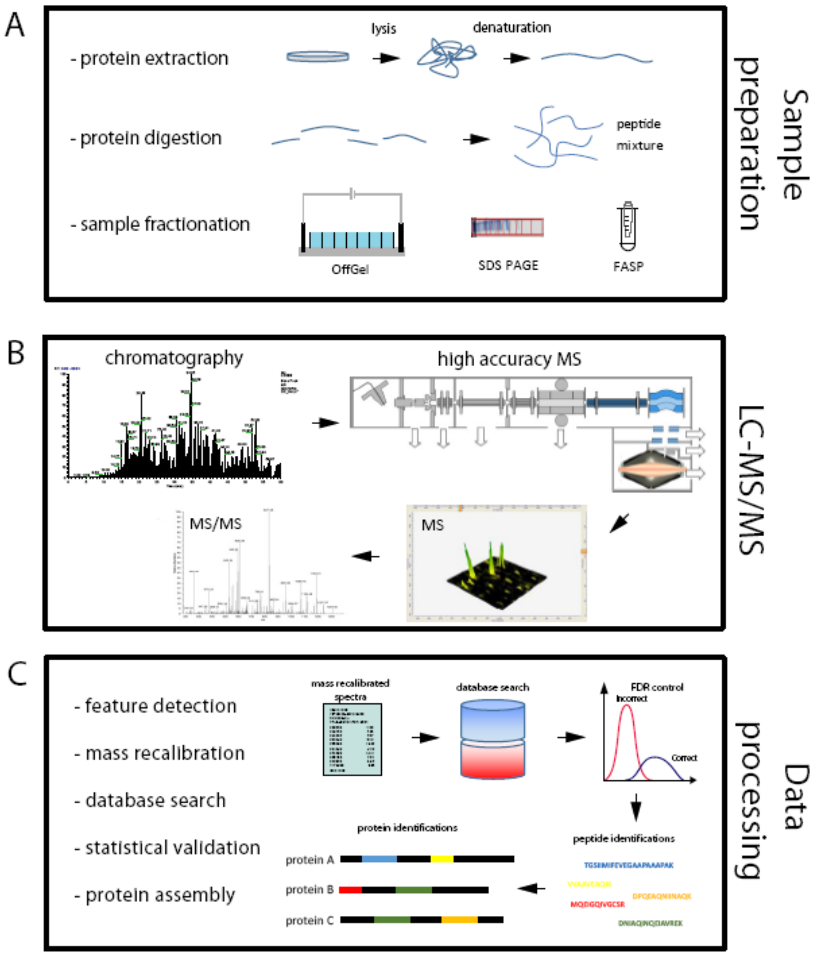


Figure 1: Qualitative shotgun proteomics workflow for protein identification. A) The efficient extraction of the protein content from a cell lysate is the first crucial step in order to achieve comprehensive proteome coverage. Proteins are digested in-solution and the resulting complex peptide mixture is fractionated by OffGel isoelectric focusing. Alternatively, proteins are separated on a 1D-gel followed by subsequent in-gel digestion. The filter-aided sample preparation (FASP) combines the advantages of the in-solution and in-gel digestion workflows and presents an attractive method for comprehensive proteome analysis. **B)** The peptide fractions are further separated by liquid chromatography (LC). The sample is loaded onto a nano-HPLC column that is directly coupled to an electrospray ionization source. Peptides elute according to their hydrophobicity, become ionized and enter the mass spectrometer (MS) in which their mass-to-charge ratio are measured at high accuracy in the MS scan. The most abundant peptides are fragmented and resulting fragment ions are detected in a MS/MS scan. **C)** Sophisticated algorithms had to be developed that are capable to detect and extract tremendous number of peptide features in a single LC-MS/MS run. Peptide masses acquired at high mass accuracy are recalibrated to enable a narrow database search tolerance and to increases sensitivity and specificity of database search. The proportion of false positive identifications is typically controlled by target-decoy search strategy. Identified peptide sequences are assembled into proteins and protein groups and the results of data processing are usually presented in a data sheet format that can be further investigated by the user.

High accuracy instruments, such as quadrupole time-of-flight (Q-TOF), Fourier transform ion cyclotron resonance (FT-ICR), or orbitrap instruments typically have a very good resolution and can achieve sub-ppm accuracy, but are characterized by relatively low acquisition speed resulting in lower proteome coverage. Newer generations of MS instruments try to circumvent this problem by combining a low- and high accuracy mass analyzer. Arguably the most successful hybrid instrument to date is the LTQ-Orbitrap, which combines a high accuracy orbitrap mass analyzer with a fast and sensitive ion trap mass analyzer. This configuration is commonly used in peptide-based "shotgun" experiments that aim to achieve as comprehensive proteome coverage as possible and which I will elaborate hereafter.

1.3.1. Sample preparation

The typical "shotgun" proteomics workflow starts with a cell culture or tissue of interest that needs to be lysed in a way that enables the highest possible retrieval of its protein content. Ideally, proteins should be extracted and solubilized in a buffer containing a potent denaturating agent and a detergent and digested in solution using a sequence-specific protease such as trypsin. Peptides are much easier to analyze than intact proteins that might not be soluble under the same conditions and get modified and processed in a way that affects their mass. Furthermore, peptides are easier to ionize and mass spectrometers are usually more sensitive in mass ranges resulting from peptides (up to ~20 amino acids) than in the mass range from whole proteins. In order to reduce the complexity of the resulting peptide mixture, the sample is separated into several fractions that are analyzed separately by MS. Peptides can be separated according to their isoelectric point by the application of OffGel electrophoresis (Hubner et al. 2008). Alternatively, the protein extract can be separated by 1D SDS-PAGE subsequently in-gel digested (Shevchenko et al. 2006) which extracts membrane proteins more efficiently. The filter-aided sample preparation (FASP) method combines the advantages of the 1D gel-based workflow and in-solution digestion (Manza et al. 2005; Wisniewski et al. 2009) making it especially attractive for comprehensive proteome analysis.

1.3.2. LC-MS/MS

Even if separated previously, the resulting peptide mixtures give rise to thousands of peptides upon digestion. Simultaneous ionization of such a complex peptide mixture negatively influences the dynamic range of the measurement and therefore leads to lower identification rates. To further reduce the complexity of the resulting mixture, the peptides are separated by high-performance liquid chromatography (HPLC). The sample is injected onto nano-HPLC column (inner diameter 25-75 µm) and eluted using a solvent gradient of increasing organic content to separate peptides according to their hydrophobicity. Very hydrophilic peptides start to elute immediately while extremely hydrophobic peptides are retained for a longer period on the column and elute at a later time point. After elution the peptides are ionized by electrospray ionization (ESI). Briefly, the peptides flow through a needle, at the tip the liquid vaporizes and the peptides become ionized by a strong electric potential (Steen and Mann 2004). The resulting ions enter the mass spectrometer through a transfer capillary and reach a vacuum system in which they are pulled by an electric field to the mass analyzer (**Figure 2A**). The mass-to-charge ratios of intact peptide (precursor) ions eluting from HPLC column at a specific time point are measured in a "full" or MS scan performed at high resolution and accuracy in the orbitrap mass analyzer. Briefly, the analyzer measures the axial frequency of the peptide ions spinning around a spindle like electrode (Hu et al. 2005). Application of the Fourier transform algorithm converts the image current into mass spectra. Each recorded peptide ion signal appears as a cluster of isotope peaks, caused by the natural occurrence of ^{13}C isotope in about 1% of all carbon atoms (**Figure 2B**). The isotope peaks are separated by 1 Da, which can be easily resolved by the orbitrap analyzer. Depending on the charge state (z) of the peptide ion, the isotope peaks are separated by 1/z Th. For example, if the isotope peaks are separated by 0.5 Th, the charge state of the peptide cluster must be 2.

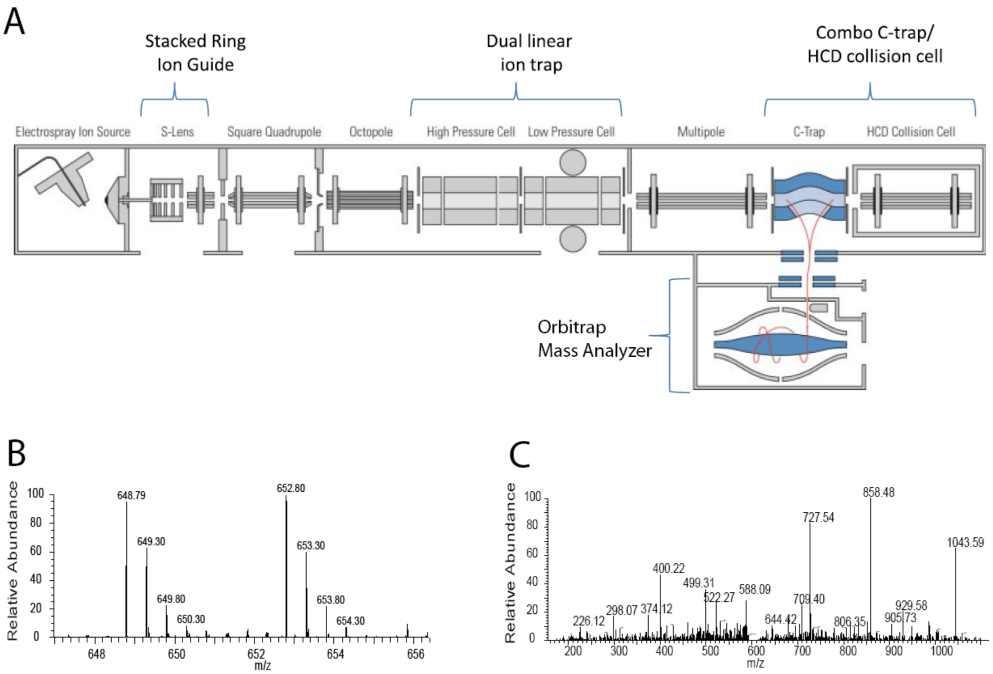


Figure 2: Hybrid MS instruments for 'high-low' sequencing strategy A) Schematic of the LTQ Orbitrap-Velos instrument. The figure has been adopted an modified from (Olsen et al. 2009). **B)** Isotope clusters of two peptide ions (SILAC pair, see paragraph 1.3.4), measured at high mass accuracy and resolution in the Orbitrap mass analyzer. The leftmost peaks of the clusters (648.79 m/z and 652.8 m/z) represent the population of peptide ions that contain only ^{12}C atoms, whereas The adjacent peaks result from peptides ion containing one ^{13}C atom (649.3 m/z and 653.3 m/z). The distance between two isotope peaks is 0.5 Th indicating doubly charged peptide ions. The two isotope cluster represent two versions of the same peptide, the unlabeled (light) version and a stable isotope labeled (heavy) version. In this particular case the mass difference between the two peptides is 8 Da. The signal intensity provides quantitative information on the relative abundance of the peptides, in this case both versions are equally abundant resulting in an 1:1 ratio. **C)** MS/MS spectrum of a fragmented peptide ion. The most abundant peptide ions are isolated, fragmented, and resulting fragments are detected at low accuracy and high speed in the linear ion trap.

To retrieve information about the amino acid sequence, the most abundant precursor ions are fragmented in the collision cell of the instrument. Although there are several different ways to fragment peptides (Ma and Johnson 2012), the most frequently used fragmentation method in proteomics relies on the collision of precursor ions with an inert gas such as helium. This process termed collision-induced dissociation (CID) preferentially results in the cleavage of the peptide bonds primarily producing two types of ions designated as b-ions for N-terminal and y-ions for C-terminal types. Other ion types might occur depending on the cleavage position on the peptide backbone (**Figure**

3), as well as many additional fragment ions such as internal ions, immonium ions or neutral loss ions. The mass-to-charge ratios of resulting fragment ions are measured in an MS/MS (or tandem MS) scan that is usually performed with high speed and high sensitivity, but low resolution in the ion trap mass analyzer (**Figure 2C)**. This so-called "high-low" strategy provides the best tradeoff between accurately measured peptide mass-to-charge ratios that directly influences a key parameter in database search (see below), and high sequencing speed required for comprehensive proteome coverage. Newer generations of orbitrap instruments (**Figure 2A**) show great promise for in-depth and accurate proteome analysis by combining a faster and more sensitive dual-pressure ion trap with a high accuracy orbitrap mass analyzer (Olsen et al. 2009), which - in the latest configuration level of the instrument - has been replaced by a high-field orbitrap that enables a higher resolution at increased sequencing speed (Michalski et al. 2012). In addition, an improved higher energy collision-induced dissociation (HCD) collision cell enables the acquisition of both MS and MS/MS spectra at high speed and mass accuracy. The routine application of the "high-high" strategy in the future will further improve the quality of acquired MS data and facilitate the data processing.

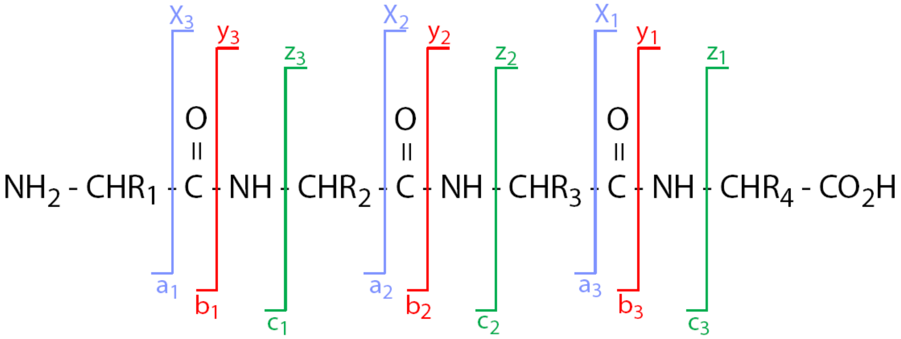


Figure 3 Nomenclature of peptide fragmentation ions. Breakage of the bond between the carbonyl C and Cα produces a and x ions, breakage of the peptide bond results in b and y ions while the breakage of the Cα-N bond produces c and z ions.

1.3.3. Data processing and database search

The data derived from a single LC-MS run consists of several tens of thousands peptide mass spectra that require automated data processing methods. Several methods have been proposed to assign the amino acid sequence to a mass spectrum such as d*e novo* sequencing approaches (Allmer 2011), correlation to spectral libraries (Lam et al. 2007) or searching protein sequence databases. Despite clear difficulties, the latter is by far the best established method within the proteomics community. Specialized software converts the MS and MS/MS spectra into a format suitable for database search (peak list). The latest generation of processing software makes use of the high mass accuracy of modern mass spectrometers to identify spectral features used for mass recalibration to compensate for drifts in instrument calibration. Mass recalibration performed in this way can improve the mass accuracy by 5-10 fold (Cox and Mann 2008).

Database search

Extracted and recalibrated peak lists are submitted to a database search engine, which identifies peptides by searching the measured mass spectra against a protein database that typically comprises the translated gene sequences derived from *ab initio* gene prediction programs (see section 1.2). The protein sequences are digested *in silico* and the measured peptide mass spectra are correlated to theoretical mass spectra resulting from candidates that have a peptide mass within a user specified tolerance (precursor mass tolerance) (**Figure 4A**). The number of potential peptide candidates defines the search space which directly affects sensitivity and specificity of database search. The size of the search space primarily depends on the mass accuracy of the MS instrument, the size of the database, number of missed cleaved peptides, and presence of post translational protein modifications (**Figure 4B**). The degree of similarity between the observed mass spectrum and each theoretical candidate spectrum is measured by the search score that depends on the applied search engine. Typically only the top scoring peptide-spectrum-match (PSM) is suggested by the search engine as correct amino acid sequence. The type of search score depends on the scoring function of the search engine uses to compare the observed to the theoretical spectra. Probability based search engines like Mascot (Matrix Science) and Andromeda (Cox et al. 2011) calculate the probability that the observed PSM is a random event.

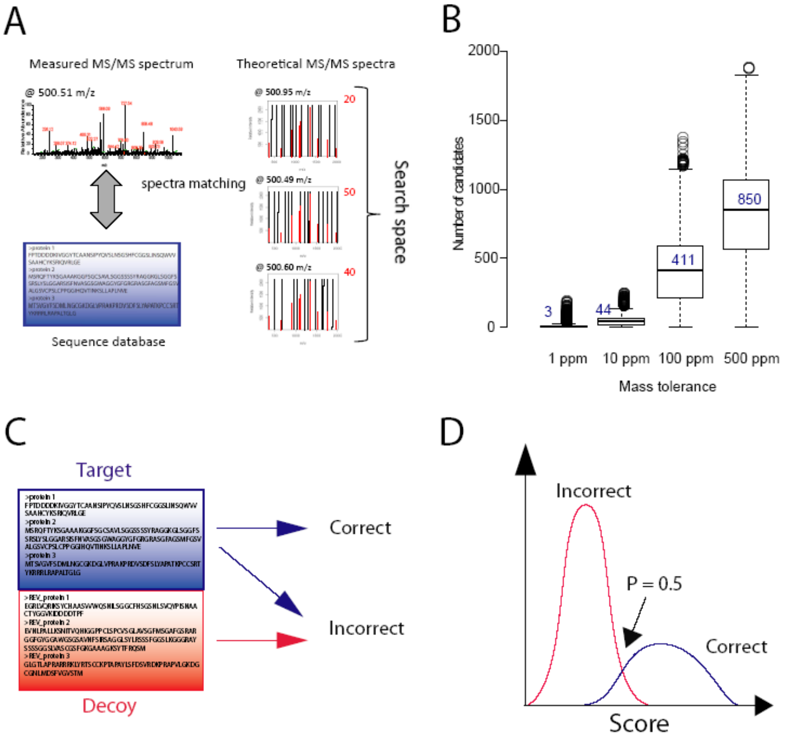


Figure 4: The principle of database search for peptide sequence assignment to MS/MS spectra. **A)** The measured spectrum is correlated to a number of theoretical spectra derived from a sequence database that contains all theoretical protein sequences of a specific organism. The degree of similarity between observed and theoretical spectra is measured in a search score, the number of possible candidate sequences defines the search space. **B)** Relationship of the search space and mass tolerance in database search. High mass accuracy instruments enable the use of narrow mass tolerances (typically around 10 ppm) which correspond to a median search space of 44 peptides in this particular example. Low accuracy data requires mass tolerances of >500 ppm resulting in an almost 20-fold larger search space (median: 850 peptide candidates). **C)** The target-decoy search strategy controls the proportion of false positive identification in a collection of PSMs **D)** Mixture model methods calculate a posterior probability for each individual PSM, which then can also be used to estimate the FDR.

For example, the algorithm implemented in Andromeda calculates the probability of observing at least *k* out of *n* matches by chance, where *k* is the number of matches of measured fragment ions to theoretical fragment ions within a specified tolerance (fragment ion tolerance) and *n* depicts the total number of theoretical fragment ions (Cox et al. 2011). These probabilities (*P*) are often reported as $-10\ \log_{10} P$ representing the actual PSM score. Further scoring schemes that have been described in the literature include cross correlation (for example SEQUEST (Eng et al. 1994)) and dot product (for example TANDEM (Craig and Beavis 2004)) calculations of tandem mass spectra.

Estimation of false discovery rates

Large-scale shotgun experiments produce matches of hundreds of thousands mass spectra to a database and the respective search engines return a match for almost every input spectrum of which only a fraction is true. In an ideal experiment the result of a database search would be a list of all peptides that have been sequenced in the mass spectrometer (true positives). However, in a real experiment the list contains peptides that were not sequenced (false positives), and leaves out peptides that were sequenced (false negatives). The fraction of true positive PSMs among all PSMs returned by the particular search engine expresses the sensitivity, whereas the fraction of true negatives (i.e. peptide sequences that were truly not identified) among all peptides that were not sequenced is termed specificity. Both quantities have to be maximized to enable high identification and low error rates at the same time. Therefore, the statistical validation of PSMs has become a crucial task in every proteomics experiment. The most commonly used and accepted confidence measures are the false discovery rate (FDR) as global property of a collection of PSMs, and the posterior error probability (PEP) for individual PSMs (Kall et al. 2008). The target-decoy search strategy (Elias and Gygi 2007) and mixture model methods (Keller et al. 2002) are two complementary and well established approaches to assign these confidence measures to database search results.

By using the target-decoy search strategy the database containing the sequences of interest (target) is complemented with a database of the same size containing definite incorrect protein sequences (decoy) (**Figure 4C**). The decoy database can be constructed by randomization or reversal of the target sequences; the effect of different strategies for decoy sequence generation was discussed elsewhere (Bianco et al. 2009; Wang et al. 2009a). The basic assumption of the target-decoy approach is that false matches from the target database and matches to decoy peptides follow the same distribution, given an equal size of the target and decoy databases. Any PSM to the decoy database is per definition a false positive identification and the overall number of decoy PSMs (N_{decoy}) can be used to estimate the expected proportion of false positive identifications among the PSMs resulting from the target database (N_{target}) by applying following formula:

$$FDR = \frac{2\,N_{decoy}}{N_{target} + N_{decoy}}$$

Mixture model approaches explain the distribution of database search scores for all PSMs in the dataset as a mixture of correct and incorrect PSMs (**Figure 4D**). The score distributions of the two populations are estimated from the data using the expectation maximization (EM) algorithm. Once the two distributions are know the posterior error probability (PEP), i.e. the probability that the score of a particular PSM belongs to the distribution of incorrect PSMs, can be calculated according to the Bayes' theorem. Alternatively, posterior probabilities (PP) for each PSM can be reported that are essentially the compliment of PEPs and therefore defined as 1-PEP. The FDR of a collection of PSMs at a specified posterior probability threshold (PP_T) can then be calculated as reported in (Nesvizhskii 2010)

$$FDR(PP_T) = \frac{\sum_{\{i|PP_i \geq PP_T\}}(1 - PP_i)}{\sum_{\{i|PP_i \geq PP_T\}} 1}$$

Protein assembly

Since the ultimate goal of a typical proteomics experiment is to identify proteins, the identified peptide sequences have to be assembled into proteins. This represents a non-trivial and challenging task, especially in proteomes of higher eukaryotes where alternative splicing of multi-exon genes is a common process. The same peptide sequence can be present in several different proteins or different protein isoforms leading to ambiguous protein identifications, a circumstance that is known as the protein inference problem (**Figure 5**). One way to partially circumvent this problem is to join and report proteins sharing a set of distinct peptide sequences as single protein group (Nesvizhskii and Aebersold 2005). Once the peptide sequences identified at a specific FDR are assembled into protein groups, the FDR has to be recomputed at the protein level to address the amplification of error rates starting at PSM levels to nonredundant peptide sequences to proteins. The different aspects of protein inference are covered in detail elsewhere (Nesvizhskii and Aebersold 2005)

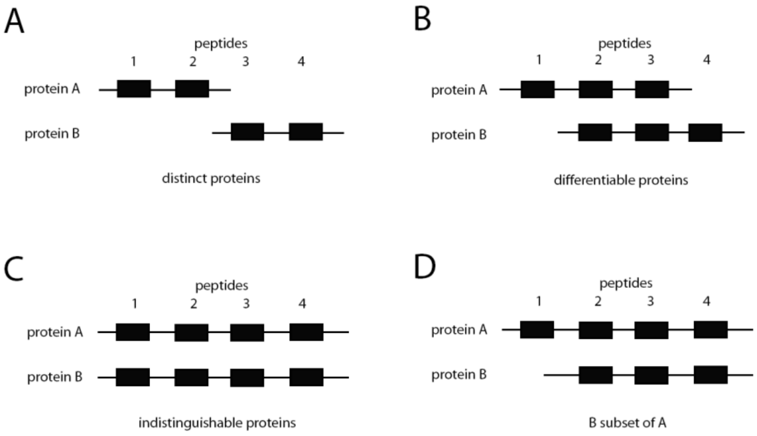


FIGURE 5: THE PROTEIN INFERENCE PROBLEM – FOUR BASIC GROUPING SCENARIOS. FIGURE ADAPTED AND MODIFIED FROM (NESVIZHSKII AND AEBERSOLD 2005). A) FOUR DISTINCT PEPTIDES ALLOW IDENTIFICATION OF TWO DISTINCT PROTEINS. **B)** THE PRESENCE OF ONE UNIQUE PEPTIDE PER PROTEIN (1, 4) ENABLES THE DISCRIMINATION OF THE TWO PROTEINS. **C)** THE TWO PROTEINS SHARE THE SAME SET OF PEPTIDES AND CANNOT BE DISTINGUISHED. **D)** PROTEIN B IS A SUBSET OF PROTEIN A.

1.3.4. Quantitative proteomics

Two general strategies are used for MS-based protein quantification:1) introduction of a stable isotope mass label that allows the discrimination between peptides from different experimental conditions in single LC-MS/MS run, or 2) the use of normalized peptide ion signals recorded in the mass spectrometer to compare protein abundance between different LC-MS/MS runs (“label-free” quantification). The latter require efficient normalization algorithms since the ion signals are extremely variable between peptides due to different ionization efficiency resulting from the diversity of chemical structures. Examples of label-free quantitation approaches are protein abundance index (PAI) (Rappsilber et al. 2002), spectral counting (Zhang et al. 2006; Asara et al. 2008), iBAQ (intensity based absolute quantification) (Schwanhausser et al. 2011) , or the label-free algorithm implemented in the MaxQuant software.

The incorporation of stable isotope atoms such as carbon (^{13}C), nitrogen (^{15}N), and deuterium (^{2}H), into proteins of one or more samples enables an accurate way to perform relative quantification of protein expression (Ong et al. 2002). The labeled sample is mixed with the unlabeled sample in equal amounts and the mixture is subjected LC-MS/MS analysis. The chemical and physical properties of labeled and unlabeled peptides are exactly the same, except for the mass difference introduced by 'heavy' stable isotopes, resulting in the appearance of two signals (doublets) of each peptide ion in the mass spectrometer that can be directly compared. The isotope label can be introduced chemically by adding a specific mass tag, or metabolically by incorporation of the heavy isotopes present in e.g. cell culture medium. Stable isotope labeling by amino acids in cell culture (SILAC) presents a frequently used metabolic labeling method, in which an essential amino acid is replaced by its isotope labeled counterpart during growth in cell culture. Cell lysates from different SILAC encoded populations are mixed prior to any protein separation and digestion minimizing the quantitation errors, and analyzed together in a single LC-MS/MS run. Quantitative proteomics was not in focus of this thesis and is extensively reviewed elsewhere (Aebersold and Mann 2003; Ong and Mann 2005; Choudhary and Mann 2010).

1.4. Proteogenomics - Mass spectrometry at the interface of proteomics and genomics

Combining MS based proteomics and genomics is not a new concept. Already in 1995, the application of MS/MS spectra to nucleotide sequences was demonstrated (Yates et al. 1995) and since then several groups applied this concept to find new genes and to improve genome annotations in various organisms using different types of mass spectrometers by searching the acquired peptide mass spectra against the genome that was translated into all six reading frames (**Figure 6A**). This strategy enables the identification of unpredicted open reading frames and the refinement of existing gene models in terms of protein start and stop positions, exon-intron structure as well as their exact boundaries (**Figure 6B**). The detected peptide sequences are used as extrinsic evidence for retraining of gene finding algorithms discussed above in order to refine existing gene models.

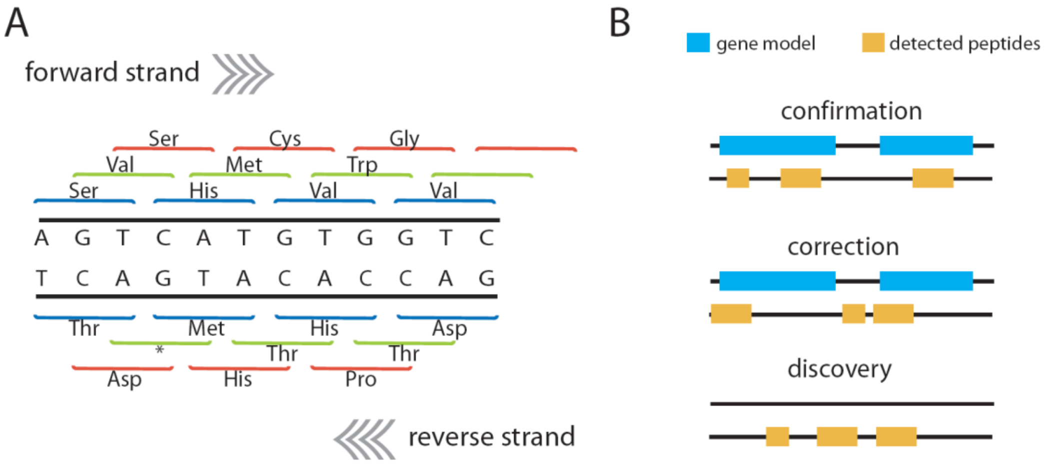


FIGURE 6: SEARCHING PROTEOGENOMIC DATABASES TO REFINE EXISTING GENOME ANNOTATIONS. A) THE TYPICAL PROTEOGENOMIC DATABASE CONSISTS OF THE TRANSLATION OF THE NUCLEOTIDE SEQUENCE INTO ALL SIX READING FRAMES, THREE ON THE FORWARD AND THREE ON THE REVERSE STRAND, RESPECTIVELY. THIS DATABASE ENABLES THE IDENTIFICATION OF PEPTIDES THAT ARE NOT PRESENT IN THE CORRESPONDING PROTEOME DATABASE THAT IS TYPICALLY DERIVED BY TRANSLATION OF THE EXISTING GENOME ANNOTATION **B)** THE PEPTIDES DETECTED BY SEARCHING A SIX-FRAME DATABASE PROVIDE EVIDENCE FOR THE CORRECT ANNOTATION OF EXISTING GENE MODELS, OR CAN BE USED TO CORRECT EXISTING GENE MODELS, AND CAN DETECT GENES THAT ARE ENTIRELY MISSING IN THE EXISTING GENOME ANNOTATION.

1.4.1. Historical overview

Early proteogenomic studies predominantly involved the use of low accuracy mass spectrometers, often in combination with 2D gel separation of proteins. For example, Link *et al.* used triple-quadrupole and ion trap MS in combination with 2D PAGE to analyze abundant proteins in *Haemophilus influenza*, and identified several proteins that were not previously annotated (Link et al. 1997). Neubauer *et al*. used triple quadrupole MS and searched expressed sequence tag (EST) databases with MS data to detect components of the mammalian spliceosome that were not contained in a comprehensive protein database (Neubauer et al. 1998) and Jungblut *et al.* demonstrated the expression of six genes in *Mycobacterium tuberculosis* not predicted by genomic approaches using 2D PAGE and MALDI-QTOF MS (Jungblut et al. 2001). Soon after the first draft of the human genome became available, Choudhary *et al.*

investigated the feasibility of searching the MS/MS spectra against a six-frame translation of all 23 human chromosomes and compared it to the search against protein and EST database (Choudhary et al. 2001). More recent studies mainly employed ion trap mass spectrometers but with application of advanced, shotgun proteomics approaches instead of 2D PAGE protein separation. Oshiro *et al.* used a combination of expression profiling and ion trap MS to verify independent transcription and translation of genes in *Saccharomyces cerevisiae* (Oshiro et al. 2002) whereas Jaffe *et al.* used ion trap MS and introduced the concept of a proteogenomic map to predict ORFs in *Mycoplasma pneumoniae* based on expressed protein-based evidence (Jaffe et al. 2004). As the field of MS-based proteomics was developing, several public repositories of MS/MS data became available and were used in proteogenomics projects. Fermin *et al.* used MS data from the Human Proteome Organization (HUPO) Plasma Project (Cottingham 2006) consisting mostly of ion trap data to search against a six-frame translation of the human genome in order to find novel blood proteins (Fermin et al. 2006). Tanner *et al.* used ion trap data as well as MS data obtained from the PeptideAtlas (Desiere et al. 2006) to identify novel and extended genes, alternative splicing events and variant alleles of coding SNPs in human genome (Tanner et al. 2007). Recent applications of mass spectrometry to genomics included large-scale, mostly ion trap based proteogenomics studies of model organisms such as *Shewanella oneidensis* (Gupta et al. 2007) several species of the *Mycobacterium genus* (de Souza et al. 2008; de Souza et al. 2009; Gallien et al. 2009) *Toxoplasma gondii* (Xia et al. 2008), *Arabidopsis thaliana* (Baerenfaller et al. 2008; Castellana et al. 2008) and *Caenorhabditis elegans* (Merrihew et al. 2008). Recently, data derived from proteogenomic studies were integrated in the Encyclopedia of DNA Elements (ENCODE) to facilitate the generation of high-quality functional annotation of the human genome (Rosenbloom et al. 2012).

1.4.2. Challenges in proteogenomic experiments

Besides general challenges arising in any shotgun proteomics experiment such as proteome coverage, dynamic range and sequencing speed of mass spectrometers, challenges specifically concerning proteogenomic experiments primarily involve computational aspects such as the construction appropriate databases, searching the acquired mass spectra against these databases and statistical scoring and validation of resulting PSMs. These computational aspects were reviewed in (Castellana and Bafna 2010).

Database construction

To enable the identification of all possible gene products from MS data, the genome sequence has to be translated *in silico* into all six reading frames leading to an at least six-fold increase in database entries. Therefore, the search space is noticeably larger compared to standard proteome databases, which negatively influences sensitivity and specificity of the database search and therefore leads to decreased identification rates. This circumstance will be even more pronounced, if target-decoy databases are used to control the FDR resulting in an additional doubling of the search space. The construction of an appropriate proteogenomic database is relatively straightforward for microbial genomes due to their small size and relatively simple genome structure. Proteogenomic databases of eukaryotes are far more complex due to the exon-intron gene structure and alternative splicing which leads to a disproportional increase in the size of six-frame databases. Moreover, any peptide sequence spanning a splice junction is not covered by regular six-frame databases and thus cannot be identified using a standard search engine as mentioned above. One approach to identify intron-split peptides described in (Allmer et al. 2004) combines *de novo* MS/MS sequencing and classical search engines like Sequest or Mascot. *De novo* deduced peptide sequences are aligned to the genomic sequence to assemble a database of possible peptides that match a particular mass spectrum. This database can be used by MS/MS search engines to identify and validate peptides that span an intron–exon boundary.

Database search

Once an appropriate database has been assembled, it is queried by the acquired mass spectra. As previously mentioned, the search space is significantly larger compared to standard organism-specific protein databases. High accuracy MS has the intrinsic potential to decrease the search space by enabling the use of narrow precursor mass tolerances in database search. **Figure 7** illustrates the increase of the search space based on a simple prokaryotic genome (*Escherichia coli* K12). Even in this simplified example it is noticeable that high measurement accuracy greatly limits the number of candidate peptide sequences that ‘‘compete’’ for an MS spectrum acquired at a given mass accuracy. Lower number of candidate peptide sequences leads to a smaller search space and this in turn leads to an increased sensitivity and specificity of the database search. However, the actual mass tolerance during database search usually exceeds the achieved measurement mass accuracy by several folds to provide ample number of candidate peptide sequences and prevent ‘‘forcing’’ the search engine to report a

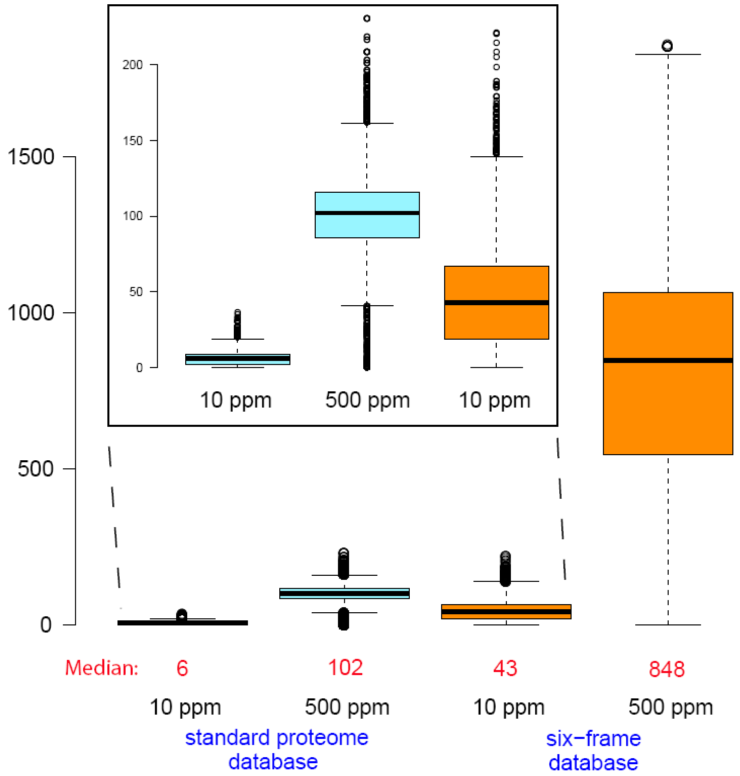


Figure 7: Boxplot representation of the database search space resulting from a protein database and six-frame translation as a function of mass tolerance. The search space of a standard protein database is depicted in cyan, the search space resulting from the translation of the corresponding genome into six reading frames is shown in orange.

particular peptide. Various bioinformatics strategies have been applied to reduce the search space in proteogenomic experiments. Most approaches incorporate auxiliary information into database search, such as knowledge about the pI of analyzed peptides (Sevinsky et al. 2008), or d*e novo* deduced short peptide sequence tags (Kuster et al. 2001).

Validation of search results

Besides the increased search space, the use of six-frame databases is made difficult by the fact that about 80% of database entries are spurious protein sequences resulting from reading frames that are never transcribed. These sequences are in fact decoy sequences and therefore should be considered in any FDR calculation. However, it is usually not known *a priori* which frame at a certain locus is used to translate the open reading frame into a protein sequence. In case of target-decoy databases the result is an unequal size of target and decoy sequence that introduces a bias in a decoy based FDR estimation. Many proteogenomic studies did not go beyond the standard procedure to control the FDR of six-frame database searches and it was stated that in those cases the actual error rate may not be known (Nesvizhskii 2010). The general consensus in the community is to carefully examine the database search results, e.g. the identification of unpredicted ORFs should be further validated by manual investigation of the corresponding MS/MS spectra and the expression of corresponding genomic region can be proved by orthogonal methods such as RT-PCR. Furthermore, a complementing shotgun transcriptomics experiment can be performed in order to increase the reliability of any novel peptide identified in six-frame database searches.

1.5. Phosphoproteomics

The MS based proteomics workflow can be modified to enable the global and *in vivo* analysis of posttranslational modifications (PTMs) of proteins, such as phosphorylation, acetylation, and ubiquitination, primarily due to the application of efficient enrichment strategies and the optimization of fragmentation and acquisition methods for the analysis of modified peptides. Published studies predominantly focused on phosphorylation of serine, threonine and tyrosine (Ser/Thr/Tyr) residues, which was the first protein modification applicable to large scale MS analysis. In this section, I will give a brief introduction of the MS-based workflow for the global analysis of Ser/thr/Tyr phosphorylation; a very detailed overview about the principles and applications of phosphoproteomics can be found in (Macek et al. 2009).

1.5.1. Ser/Thr/Tyr Phosphorylation

The reversible phosphorylation of proteins is involved in almost every known cellular signaling pathway and presents the most common and important signal transduction event in eukaryotic as well as prokaryotic systems. Protein kinases transfer a phosphate group from adenosine triphosphate (ATP) on a protein, while phosphatases catalyze the removal the phosphate by hydrolysis. Phosphorylation can occur on several amino acids; the most prominent type of phosphorylation in eukaryotic systems involves serine, threonine and tyrosine phosphorylation, in which the phosphate group is bound to the hydroxyl group of the amino acids. It has been shown that Ser/Thr/Tyr phosphorylation is not exclusive to eukaryotic systems, but also plays a key role in prokaryotic signal transduction, besides the canonical Asp/His phosphorylation characteristic for two-component signaling systems in bacteria (Deutscher and Saier 2005). The interplay between kinases and phosphatases presents a switch to alter protein activity of its substrates by, e.g. inducing a conformational change or creating a docking site for other signaling proteins. This activation or deactivation of certain proteins directs the signal propagation inside the cell. The complement of protein kinases encoded in a genome is termed kinome and can be classified into orthologous kinase groups with conserved functions across different species and kinase families within a specific lineage (Manning et al. 2002). The classification is primary based on sequence similarity of the catalytic kinase domain and the taxonomy of Hanks and Hunter (Hanks and Hunter 1995). Eukaryotic protein kinases (ePKs) make up a large superfamily of protein kinases sharing a conserved catalytic domain and are divided into two main subgroups - serine/threonine kinases and tyrosine kinases. Atypical protein kinases (aPKs) lack sequence similarity to ePKs, but are known or predicted to have an enzymatic activity.

Most of the kinase groups and families are conserved throughout metazoans, many of them are also conserved in yeasts. The number of kinases encoded in a genome differs across the species ranging from ~130 (*S. cerevisiae*) to ~ 500 (human). The study and comparison of kinomes across species provides valuable insights into the evolution and architecture of protein kinase signaling pathways (Manning et al. 2002).

TABLE 2: CLASSIFICATION OF PROTEIN KINASES. THE TABLE WAS COMPILED USING THE INFORMATION PROVIDED ON HTTP://KINASE.COM/WIKI.

	Group	Description
ePK	AGC	Contains core intracellular signaling kinases
	CMGC	Diversity funcions in cell cycle control, MAPK signaling and splicing
	CAMK	Calcium and calmodulin-dependent kinases
	CK1	Casein kinase 1, small and ancient family
	Ste	Consists of three main families that activate the MAPK family
	TK	Tyrosine kinase; phosphorylate almost exclusively on tyrosine residues
	TKL	Tyrosine kinase-like; most similar to tyrosine kinases
	RGC	Receptor guanylate cyclase
	Other	Several families that do not fit other ePK groups
aPK	aTYPICAL	Diverse group with no structural similarity to ePKs

1.5.2. Phosphopeptide enrichment

Due to the substoichiometric nature of the modification, phosphorylated peptides represent only a small proportion of all peptides in the cell lysate and therefore have to be biochemically enriched before they can be efficiently analyzed in the mass spectrometer (Macek et al. 2009). Among several strategies that exist to specifically enrich phosphopeptides, two widely used affinity based enrichment methods are strong cation exchange (SCX) chromatography and titanium dioxide (TiO_2) enrichment. The combination of both enrichment steps provides a robust and efficient strategy for the large-scale analysis of phosphorylated peptides.

Strong cation exchange (SCX) chromatography

The principle of SCX chromatography is based on the difference in the solution charge states of phosphorylated and nonphosphorylated peptides. The solution charge state of a tryptic peptide at pH 2.7 is +2 because C-terminal lysine or arginine and the N-terminal amino group are protonated. If the same peptide carries a negatively charged phosphate group, the net charge will be reduced by one. Therefore, phosphorylated peptides can be enriched by their decreased net charge (Macek et al. 2009). Phosphopeptides are separated on an analytical column using a linear salt gradient resulting in several SCX fractions. Due to a zero or negative net charge of multiply phosphorylated peptides, they are not retained on the SCX column. Thus the unbound material ("flow-through") has to be analyzed separately.

Titanium dioxide (TiO_2) enrichment

TiO_2 spheres covalently bind the phosphate group of the modified peptides, but also glutamic and aspartic acid. In order to circumvent the binding of nonphosphorylated peptides that are rich in these acidic residues, 2,5-dihydroxy benzoic acid (DHB) is usually used as a competitive binder in the buffer. DHB is bound by TiO_2 spheres with higher affinity than unphosphorylated peptides but with lower affinity than phosphopeptides. Thus, TiO_2 enrichment relies on the competitive binding of phosphopeptides, and DHB which has to be removed from the sample prior to LC-MS/MS analysis in order to avoid its binding to the separation column.

1.5.3. MS analysis of phosphopeptides

The challenge to analyze phosphorylated peptides in a mass spectrometer is the low coverage of sequence-specific fragment ions compared to the fragmentation of unmodified peptides (Macek et al. 2009). Using collision induced dissociation as described above the fragmentation generally occurs at the bonds containing the lowest energy; in case of phosphorylated serine and threonine peptides this is the particularly labile O-phosphate bond. The result of this fragmentation is a prominent neutral loss of phosphoric acid (97.97 Da) from the phosphopeptide ion, often without any further sequence specific b- and y-fragment ions. To circumvent this problem the linear ion trap consecutively dissociates ion species that result from neutral losses upon CID, which are positioned at -97.97 Th, -48.99 Th, or -32.66 Th away from singly, doubly, or triply charged precursor ions, respectively. This so called multi stage activation (MSA) strategy is routinely used to analyze protein phosphorylation on LTQ-based hybrid instruments.

1.5.4. Identification of phosphorylation sites

To enable the identification of phosphorylated peptides the mass of the phosphate group has to be defined as variable modification during database search. The search engine adds the mass of the phosphate group to every serine, threonine, and tyrosine contained in the candidate peptide sequence to the search space and therefore is able to identify the MS/MS spectrum of a phosphorylated peptide. Determination of the exact position of a phosphorylation site within the peptide sequence is often difficult, especially for multiply phosphorylated peptides. For example, a peptide with consecutive serines, threonines, or tyrosines requires identified fragment ions between each of them in order to unambiguously assign the phosphorylation sites. A computational approach that addresses the phosphorylation site localization after peptide identification is integrated into the MaxQuant framework (Cox et al. 2011). The identified MS/MS spectrum is compared to all theoretical spectra in which the phosphorylation is placed at each possible position. Each position is scored, the scores are normalized and transformed into a probability (‘localization probability’) providing a statistical means for assigning phosphorylation to individual sites.

1.6. Aims of the thesis

In this thesis I use data derived from mass spectrometry-based proteomics to improve the annotation of genome sequencing data and address general properties of the proteogenomic approach. Specific aims of the thesis are:

1) Assessment of general properties of a typical, shotgun proteomics-based proteogenomics experiment using the small and well characterized genome of *E. coli*.
    a. Determination of sensitivity, specificity, accuracy, and false discovery rate of a proteogenomic experiment based on two well established and complementary MS data processing frameworks, MaxQuant (MQ) and Trans-Proteomic Pipeline (TPP).
    b. Refinement of the *E. coli* K12 genome annotation.
    c. Determination of genome sequence coverage by shotgun proteomics data in a typical bacterial proteomics dataset.
2) Application of proteogenomics to a complex genome assembly of a eukaryotic organism using the example of the model nematode *P. pacificus.*
    a. Refinement of the existing genome annotation and assembly of the *P. pacificus* protein database based on the refined annotation.
    b. Analysis of general properties of the *P. pacificus* proteome and comparison to other model nematodes, in particular *C. elegans*.
    c. Creation of the first comprehensive catalog of experimentally confirmed expressed proteome of *P. pacificus*.
3) Functional annotation of the *P.pacificus* proteome with emphasis on processes related to protein phosphorylation and signal transduction.
    a. Characterization of the *P. pacificus* phosphoproteome using a qualitative phosphoproteomic dataset and its comparison to *C. elegans*.
    b. Characterization of the theoretical kinome of *P. pacificus.*
    c. Global functional annotation of the theoretical *P. pacificus* proteome.

2. Results

2.1. Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments

Karsten Krug, Alejandro Carpy, Gesa Behrends, Katarina Matic, Nelson C. Soares, Boris Macek

2.1.1. Synopsis:

The model Gram-negative bacterium *Escherichia coli* is one of the most intensively studied organisms. Many fundamental molecular processes are universal throughout the natural world and are best understood in *E. coli*. Especially the K-12 strain is a preferred model in biochemical genetics and molecular biology and was the earliest organism suggested as a candidate for whole genome sequencing (Blattner et al. 1997). Since then, its genome has been re-sequenced (Hayashi et al. 2006) and its features have been extensively studied (Karp et al. 2007). The *E. coli* chromosome consists of 4.6 Mb and encodes about 4,500 genes of which ~4,300 genes are protein coding (Hayashi et al. 2006). All protein entries contained in the UniProt database (*E. coli* K-12 reference proteome) are assigned with status 'reviewed' indicating that there is experimental evidence of protein existence for every database entry. The high quality of genome annotation makes *E. coli* an ideal model to assess general properties of simple microbial proteogenomics experiments. In this study we performed a comprehensive analysis of the *Escherichia coli* proteome using high accuracy LTQ-Orbitrap MS and map the corresponding MS/MS spectra onto a six-frame translation of the *E. coli* genome. We assumed complete annotation of the *E. coli* genome and regard all six frame-specific (novel) PSMs as false positive identifications. This enabled us to assess the sensitivity, specificity, accuracy and actual false discovery rate in a typical bacterial proteogenomic dataset. To increase reliability of our results we used two complementary computational frameworks for processing and statistical assessment of MS/MS data: MaxQuant and Trans-Proteomic Pipeline. We showed that the posterior error probability distribution of novel hits is almost identical to that of reversed (decoy) hits, pointing to substantial underestimation of FDR even in “simple” proteogenomic experiments obtained by high accuracy MS. The use of a small and well annotated bacterial genome enabled us to address genome coverage achieved in state-of-the-art bacterial proteomics: identified peptide sequences mapped to all estimated expressed *E. coli* proteins, but covered 27.5% of the total genome sequence. Our results pointed to the necessity for further technological and bioinformatic improvements in proteogenomic strategies.

2.1.2. Contributions:

Alejandro Carpy, Katarina Matic, and Gesa Behrends prepared the samples; Alejandro Carpy performed the MS measurements. I performed the processing of MS raw files and complete down-stream analysis of the retrieved results. In particular I constructed appropriate proteome databases and performed the data processing using MaxQuant and Trans-Proteomic Pipeline. I developed and applied a computational pipeline to map the identified peptide sequences onto the genome sequence as well as the protein sequences contained in the *E. coli* database. Based on the results derived from the pipeline I developed a strategy to assess several parameters of both data processing workflows such as the actual false discovery rate. Furthermore, I performed the genome coverage analysis of MS data. I prepared all figures and wrote the manuscript with the help of Prof. Dr. Boris Macek. In total, my contribution to this manuscript was about 75%.

2.2. Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models.

Nadine Borchert, Christoph Dieterich, Karsten Krug, Wolfgang Schütz, Stephan Jung, Alfred Nordheim, Ralf J. Sommer, Boris Macek

2.2.1. Synopsis

Pristionchus pacificus is a nematode which is increasingly used as model organism in developmental biology. In comparison to the classical nematode model, *Caenorhabditis elegans*, *P. pacificus* occupies a completely different niche and has a necromenic lifestyle in association to beetles. While C*. elegans* was the first multicellular organism having its genome sequenced (Consortium 1998), the genome of *P. pacificus* was sequenced recently using the whole genome shotgun method (Dieterich et al. 2008) and revealed the same number of chromosomes (5+1) but a substantially larger genome sizes of 169 Mb than *C. elegans* (100 Mb). The application of the SNAP gene finder revealed 29,000 protein-coding genes of which ~11,000 are restricted to *P. pacificus* and have no homologs in other species (‘pioneer or orphan genes’). The aim of this study was to refine the genome annotation of *P. pacificus* by performing transcriptome and proteome analysis. The proteome analysis on an LTQ-Orbitrap mass spectrometer detected 27,000 non-redundant peptide sequences from more than 4,000 proteins. Identified ESTs and detected peptide sequences were used to retrain the SNAP gene prediction algorithm to refine the initial genome annotation, which led to a decrease in the number of previously predicted protein-coding genes from 29,000 to 24,000 and refinement of numerous gene models while the number of pioneer genes only slightly decreased. Some of the corresponding proteins appear to be products of highly homologous genes, pointing to their common origin. We show that >50% of all pioneer genes are transcribed under standard culture conditions and that pioneer proteins significantly contribute to a unimodal distribution of predicted protein sizes in *P. pacificus*, which has an unusually low median size of 240 amino acids (26.8 kDa). In contrast, the predicted proteome of *C. elegans* follows a distinct bimodal protein size distribution, with significant functional differences between small and large protein populations. Combined, these results provide the first catalog of the expressed genome of *P. pacificus*, refinement of its genome annotation, and the first comparison of related nematode models at the proteome level.

2.2.2. Contributions

The samples for proteomics analysis were prepared by Nadine Borchert and measured by Stephan Jung. The genome refinement using transcriptomics and proteomics data was performed by Christoph Dieterich. I contributed to the proteomics part of this study in which I was primarily involved in data analysis. In particular I constructed the proteogenomic database based on the *P. pacificus* genome assembly and performed the processing of MS raw data and subsequent downstream analysis of the results. I re-processed the acquired MS raw files using the refined genome annotation to provide the catalog of expressed proteins. I performed the comparative protein size analysis of *P. pacificus* and the other model nematodes, Gene Ontology analysis, and the analysis of pioneer proteins. I was actively involved in the preparation of figures, supplementary data, and the methods section of the manuscript. My contribution to this work was about 33%.

2.3. Phosphoproteome of Pristionchus pacificus provides insights into architecture of signaling networks in nematode models.

Nadine Borchert, Karsten Krug, Florian Gnad, Amit Sinha, Ralf J. Sommer, Boris Macek

2.3.1. Synopsis

The analysis of the predicted proteomes of *C. elegans* and *P. pacificus* revealed distinct proteome structure in terms of protein size distribution. While the *P.pacificus* proteome is characterized by a unimodal protein size distribution, the *C. elegans* protein size distribution is characterized by two modes with distinct protein function of small and large protein populations. In particular, the population of large proteins was enriched in functions related to protein phosphorylation, signal transduction, and ion transport. To gain insight into the architecture of signal transduction networks in model nematodes, we performed a large-scale qualitative phosphoproteome analysis of *P. pacificus*. Using two-stage enrichment of phosphopeptides from a digest of *P. pacificus* proteins and their subsequent analysis via high accuracy MS, we detected and localized 6,809 phosphorylation events on 2,508 proteins. We compared the detected *P. pacificus* phosphoproteome to the recently published phosphoproteome of *C. elegans*. The overall numbers and functional classes of phosphoproteins were similar between the two organisms. Interestingly, the products of orphan genes were significantly underrepresented among the detected *P. pacificus* phosphoproteins. We defined the theoretical kinome of *P. pacificus* and compared it to that of *C. elegans*. While tyrosine kinases were slightly underrepresented in the kinome of *P. pacificus*, all major classes of kinases were present in both organisms. Application of our kinome annotation to a recent transcriptomic study of dauer and mixed stage populations showed that Ser/Thr and Tyr kinases show similar expression levels in *P. pacificus* but not in *C. elegans*. This study presents the first systematic comparison of phosphoproteomes and kinomes of two model nematodes and, as such, will be a useful resource for comparative studies of their signal transduction networks.

2.3.2. Contributions

The samples were prepared by Nadine Borchert who also performed the two-stage phosphopeptide enrichment and the MS measurements with the help of Johannes Madlung. I performed the data processing of MS raw files and subsequent Gene Ontology enrichment analysis of the detected phosphoproteome. I used a recently published phosphoproteome dataset of *C. elegans* (Zielinska et al. 2009) to perform a comparative analysis of the phosphoproteomes of the two nematodes. I used the psiPRED software to determine secondary protein structures of both proteomes. I developed and applied a computational pipeline to predict the theoretical kinome of the two nematodes. I further annotated the *P. pacificus* proteome with Gene Ontology terms, protein family and pathway information (Pfam, Interpro, KEGG). I was actively involved in the preparation of figures, supplementary data, and the methods section of the manuscript. In total, my contribution to this work was about 50%.

3. Conclusions

In this thesis I studied several different aspects of the application of high mass accuracy shotgun proteomics data to refinement of genome annotation. Based on the obtained results I conclude the following points:

1) The small and extensively studied genomes of model bacteria are well suited to assess general properties of a typical proteogenomics workflow; assuming a complete and correct genome annotation, these models can be used to evaluate the performance of different data processing strategies.
    a. Sensitivity, specificity, accuracy, and false discovery rate (FDR) markedly differ between the two MS data processing workflows, MaxQuant (MQ) and Trans-Proteomic Pipeline (TPP). The TPP workflow demonstrated the highest specificity and low false discovery rates, while the MQ workflow demonstrated the best tradeoff between high sensitivity and acceptable FDR.
    b. Despite the high quality of genome annotation of *E. coli* it was still possible to reveal several annotation errors. Nine novel peptides were identified by both workflows that point to wrongly annotated protein termini, or the absence of the corresponding protein in the database of the particular *E. coli* K12 strain.
    c. Although the dataset comprised almost all expressed proteins, the identified peptide sequences covered 27.5 % of the raw genome sequence and 30% of protein coding sequence. Those regions were identified with a median of seven MS/MS scan events per nucleotide. This points to the limitation of MS in genome annotation alone, where comprehensive several fold genome coverage is desired, and which is routinely achieved by NGS technologies.
2) The genome annotation of *P. pacificus* could be markedly improved using proteomics and transcriptomics data.

    a. The shotgun proteomics approach identified 2,700 novel peptide sequences. Together with transcriptomics data the refinement of genome annotation lead to a decrease in the number of predicted genes from 29,000 to 24,000.
    b. The theoretical proteomes of *P. pacificus* and *C. elegans* show distinct distributions of protein sizes. In *P. pacificus* the distribution is unimodal, whereas in *C. elegans* the distribution is bimodal with functional differences between the two protein populations, in particular regarding protein phosphorylation.
    c. In total, 4,029 *P. pacificus* proteins were detected by MS providing the first global proteome catalog of this nematode. Among all identified proteins about 10 % were products of pioneer genes, which is in contrast to transcriptomics data, where >50% of pioneer genes were found to be transcribed.

3) The *P. pacificus* genome was functionally annotated using a qualitative phosphoproteomic dataset.
    a. The detected phosphoproteome of *P. pacificus* comprised 6,800 phosphorylation sites on 2,500 proteins. Compared to *C. elegans* the same functional classes of proteins are phosphorylated, but the relative frequencies of phosphorylated serines, threonines, and tyrosines are markedly different.
    b. The kinome of *P. pacificus* consists of 368 kinases which is a 11% smaller kinome than in *C. elegans,* but all major kinase groups are present.
    c. About 73% of the refined theoretical proteome could be annotated by at least one functional term.

References

Aebersold R, Mann M. 2003. Mass spectrometry-based proteomics. *Nature* **422**(6928): 198-207.

Ahmed FE. 2008. Utility of mass spectrometry for proteome analysis: part I. Conceptual and experimental approaches. *Expert Rev Proteomics* **5**(6): 841-864.

Allmer J. 2011. Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev Proteomics* **8**(5): 645-657.

Allmer J, Markert C, Stauber EJ, Hippler M. 2004. A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases. *FEBS Lett* **562**(1-3): 202-206.

Ansorge WJ. 2009. Next-generation DNA sequencing techniques. *New biotechnology* **25**(4): 195-203.

Asara JM, Christofk HR, Freimark LM, Cantley LC. 2008. A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen. *Proteomics* **8**(5): 994-999.

Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S. 2008. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* **320**(5878): 938-941.

Bianco L, Mead JA, Bessant C. 2009. Comparison of Novel Decoy Database Designs for Optimizing Protein Identification Searches Using ABRF sPRG2006 Standard MS/MS Data Sets. *Journal of Proteome Research* **8**(4): 1782-1791.

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, ColladoVides J, Glasner JD, Rode CK, Mayhew GF et al. 1997. The complete genome sequence of Escherichia coli K-12. *Science* **277**(5331): 1453-&.

Brent MR. 2005. Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Res* **15**(12): 1777-1786.

Castellana N, Bafna V. 2010. Proteogenomics to discover the full coding content of genomes: A computational perspective. *J Proteomics*.

Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP. 2008. Discovery and revision of Arabidopsis genes by proteogenomics. *Proc Natl Acad Sci U S A* **105**(52): 21034-21038.

Choudhary C, Mann M. 2010. Decoding signalling networks by mass spectrometry-based proteomics. *Nature reviews Molecular cell biology* **11**(6): 427-439.

Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS. 2001. Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **1**(5): 651-667.

Collins FS, Lander ES, Rogers J, Waterston RH, Conso IHGS. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011): 931-945.

Consortium TCeS. 1998. Genome sequence of the nematode C-elegans: A platform for investigating biology. *Science* **282**(5396): 2012-2018.

Cottingham K. 2006. HUPO Plasma Proteome Project: challenges and future directions. *J Proteome Res* **5**(6): 1298.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**(12): 1367-1372.

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**(4): 1794-1805.

Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**(9): 1466-1467.

Cui P, Lin QA, Ding F, Xin CQ, Gong W, Zhang LF, Geng JN, Zhang B, Yu XM, Yang J et al. 2010. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics* **96**(5): 259-265.

de Souza GA, Malen H, Softeland T, Saelensminde G, Prasad S, Jonassen I, Wiker HG. 2008. High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using Mycobacterium tuberculosis as an example. *BMC Genomics* **9**: 316.

de Souza GA, Softeland T, Koehler CJ, Thiede B, Wiker HG. 2009. Validating divergent ORF annotation of the Mycobacterium leprae genome through a full translation data set and peptide identification by tandem mass spectrometry. *Proteomics* **9**(12): 3233-3243.

Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. 2006. The PeptideAtlas project. *Nucleic Acids Res* **34**(Database issue): D655-658.

Deutscher J, Saier MH, Jr. 2005. Ser/Thr/Tyr protein phosphorylation in bacteria - for long time neglected, now well established. *Journal of molecular microbiology and biotechnology* **9**(3-4): 125-131.

Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, Fulton L, Fulton R, Godfrey J, Minx P et al. 2008. The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. *Nat Genet* **40**(10): 1193-1198.

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**(3): 207-214.

Eng JK, Mccormack AL, Yates JR. 1994. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. *J Am Soc Mass Spectr* **5**(11): 976-989.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**(4926): 64-71.

Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ. 2006. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol* **7**(4): R35.

Gallien S, Perrodou E, Carapito C, Deshayes C, Reyrat JM, Van Dorsselaer A, Poch O, Schaeffer C, Lecompte O. 2009. Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Res* **19**(1): 128-135.

Gupta N, Tanner S, Jaitly N, Adkins JN, Lipton M, Edwards R, Romine M, Osterman A, Bafna V, Smith RD et al. 2007. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res* **17**(9): 1362-1377.

Hanks SK, Hunter T. 1995. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **9**(8): 576-596.

Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, Ohtsubo E, Baba T, Wanner BL, Mori H et al. 2006. Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Molecular Systems Biology* **2**: 2006 0007.

Hillenkamp F, Karas M, Beavis RC, Chait BT. 1991. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal Chem* **63**(24): 1193A-1203A.

Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R. 2005. The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry : JMS* **40**(4): 430-443.

Hubner NC, Ren S, Mann M. 2008. Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. *Proteomics* **8**(23-24): 4862-4872.

Jaffe JD, Berg HC, Church GM. 2004. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4**(1): 59-77.

Jungblut PR, Muller EC, Mattow J, Kaufmann SHE. 2001. Proteomics reveals open reading frames in Mycobacterium tuberculosis H37Rv not predicted by genomics. *Infect Immun* **69**(9): 5905-5907.

Kall L, Storey JD, MacCoss MJ, Noble WS. 2008. Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res* **7**(1): 40-44.

Karp PD, Keseler IM, Shearer A, Latendresse M, Krummenacker M, Paley SM, Paulsen I, Collado-Vides J, Gama-Castro S, Peralta-Gil M et al. 2007. Multidimensional annotation of the Escherichia coli K-12 genome. *Nucleic Acids Res* **35**(22): 7577-7590.

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**(20): 5383-5392.

Knapp K, Chen YPP. 2007. An evaluation of contemporary hidden Markov model genefinders with a predicted exon taxonomy. *Nucleic Acids Research* **35**(1): 317-324.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59.

Kuster B, Mortensen P, Andersen JS, Mann M. 2001. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1**(5): 641-650.

Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R. 2007. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**(5): 655-667.

Lander ES Consortium IHGS Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Link AJ, Hays LG, Carmack EB, Yates JR. 1997. Identifying the major proteome components of Haemophilus influenzae type-strain NCTC 8143. *Electrophoresis* **18**(8): 1314-1334.

Ma B, Johnson R. 2012. De novo sequencing and homology searching. *Mol Cell Proteomics* **11**(2): O111 014902.

Macek B, Mann M, Olsen JV. 2009. Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu Rev Pharmacol Toxicol* **49**: 199-221.

Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**(16): 2878-2879.

Manning G, Plowman GD, Hunter T, Sudarsanam S. 2002. Evolution of protein kinase signaling from yeast to man. *Trends in biochemical sciences* **27**(10): 514-520.

Manza LL, Stamer SL, Ham AJ, Codreanu SG, Liebler DC. 2005. Sample preparation and digestion for proteomic analyses using spin filters. *Proteomics* **5**(7): 1742-1745.

Merrihew GE, Davis C, Ewing B, Williams G, Kall L, Frewen BE, Noble WS, Green P, Thomas JH, MacCoss MJ. 2008. Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. *Genome Res* **18**(10): 1660-1669.

Metzker ML. 2010. Applications of Next-Generation Sequencing Sequencing Technologies - the Next Generation. *Nat Rev Genet* **11**(1): 31-46.

Michalski A, Damoc E, Lange O, Denisov E, Nolting D, Muller M, Viner R, Schwartz J, Remes P, Belford M et al. 2012. Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol Cell Proteomics* **11**(3): O111 013698.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881): 1344-1349.

Nesvizhskii AI. 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics* **73**(11): 2092-2123.

Nesvizhskii AI, Aebersold R. 2005. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **4**(10): 1419-1440.

Neubauer G, King A, Rappsilber J, Calvio C, Watson M, Ajuh P, Sleeman J, Lamond A, Mann M. 1998. Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat Genet* **20**(1): 46-50.

Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, Lange O, Remes P, Taylor D, Splendore M et al. 2009. A dual pressure linear ion trap - Orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics*: M900375-MCP900200.

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**(5): 376-386.

Ong SE, Mann M. 2005. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **1**(5): 252-262.

Oshiro G, Wodicka LM, Washburn MP, Yates JR, 3rd, Lockhart DJ, Winzeler EA. 2002. Parallel identification of new genes in Saccharomyces cerevisiae. *Genome Res* **12**(8): 1210-1220.

Rappsilber J, Ryder U, Lamond AI, Mann M. 2002. Large-scale proteomic analysis of the human spliceosome. *Genome Res* **12**(8): 1231-1245.

Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, Cline MS, Karolchik D, Barber GP, Clawson H et al. 2012. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Research* **40**(D1): D912-D917.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**(12): 5463-5467.

Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* **473**(7347): 337-342.

Sevinsky JR, Cargile BJ, Bunger MK, Meng F, Yates NA, Hendrickson RC, Stephenson JL, Jr. 2008. Whole genome searching with shotgun proteomic data: applications for genome annotation. *J Proteome Res* **7**(1): 80-88.

Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M. 2006. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* **1**(6): 2856-2860.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**: ii215-225.

Steen H, Mann M. 2004. The ABC's (and XYZ's) of peptide sequencing. *Nature reviews Molecular cell biology* **5**(9): 699-711.

Tanner S, Shen ZX, Ng J, Florea L, Guigo R, Briggs SP, Bafna V. 2007. Improving gene annotation using peptide mass spectrometry. *Genome Res* **17**(2): 231-239.

Teer JK, Mullikin JC. 2010. Exome sequencing: the sweet spot before whole genomes. *Human molecular genetics* **19**(R2): R145-151.

Venter JC Adams MD Myers EW Li PW Mural RJ Sutton GG Smith HO Yandell M Evans CA Holt RA et al. 2001. The sequence of the human genome. *Science* **291**(5507): 1304-1351.

Wang G, Wu WW, Zhang Z, Masilamani S, Shen RF. 2009a. Decoy Methods for Assessing False Positives and False Discovery Rates in Shotgun Proteomics. *Analytical Chemistry* **81**(1): 146-159.

Wang Z, Gerstein M, Snyder M. 2009b. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**(1): 57-63.

Wisniewski JR, Zougman A, Nagaraj N, Mann M. 2009. Universal sample preparation method for proteome analysis. *Nat Methods* **6**(5): 359-362.

Xia D, Sanderson SJ, Jones AR, Prieto JH, Yates JR, Bromley E, Tomley FM, Lal K, Sinden RE, Brunk BP et al. 2008. The proteome of Toxoplasma gondii: integration with the genome provides novel insights into gene expression and annotation. *Genome Biol* **9**(7): R116.

Yates JR, 3rd, Eng JK, McCormack AL. 1995. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem* **67**(18): 3202-3210.

Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF. 2006. Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* **5**(11): 2909-2918.

Zhang WY, Chen JJ, Yang Y, Tang YF, Shang J, Shen BR. 2011. A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. *PLoS One* **6**(3).

Zielinska DF, Gnad F, Jedrusik-Bode M, Wisniewski JR, Mann M. 2009. Caenorhabditis elegans has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins. *J Proteome Res* **8**(8): 4039-4049.

Manuscripts related to this thesis

i. **Karsten Krug**, Sven Nahnsen, and Boris Macek:
Mass spectrometry at the interface of proteomics and genomics.
Mol. BioSyst. 2011, 7, 284-291. doi:10.1039/c0mb00168f

ii. **Karsten Krug**, Alejandro Carpy, Gesa Behrends, Katarina Matic, Nelson C. Soares, and Boris Macek:
Deep Coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments.
Under revision in *Mol Cell Proteomics*

iii. Nadine Borchert*, Christoph Dieterich*, **Karsten Krug***, Wolfgang Schütz, Stephan Jung, Alfred Nordheim, Ralf J. Sommer, and Boris Macek:
Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models.
Genome Res. 2010, 20, 837-846. doi: 10.1101/gr.103119.109

iv. Nadine Borchert*, **Karsten Krug***, Florian Gnad, Amit Sinha, Ralf J. Sommer, and Boris Macek:
Phosphoproteome of *Pristionchus pacificus* provides insights into architecture of signaling networks in nematode models.
Mol Cell Proteomics 11, 1631-1639. doi: 10.1074/mcp.M112.022103

* = equal contribution

Mass spectrometry at the interface of proteomics and genomics†

Karsten Krug, Sven Nahnsen and Boris Macek*

Received 19th August 2010, Accepted 20th September 2010
DOI: 10.1039/c0mb00168f

Downloaded by UNIVERSITAET TUEBINGEN on 07 January 2013
Published on 21 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MB00168F

With the onset of modern DNA sequencing technologies, genomics is experiencing a revolution in terms of quantity and quality of sequencing data. Rapidly growing numbers of sequenced genomes and metagenomes present a tremendous challenge for bioinformatics tools that predict protein-coding regions. Experimental evidence of expressed genomic regions, both at the RNA and protein level, is becoming invaluable for genome annotation and training of gene prediction algorithms. Evidence of gene expression at the protein level using mass spectrometry-based proteomics is increasingly used in refinement of raw genome sequencing data. In a typical "proteogenomics" experiment, the whole proteome of an organism is extracted, digested into peptides and measured by a mass spectrometer. The peptide fragmentation spectra are identified by searching against a six-frame translation of the raw genomic assembly, thus enabling the identification of hitherto unpredicted protein-coding genomic regions. Application of mass spectrometry to genome annotation presents a range of challenges to the standard workflows in proteomics, especially in terms of proteome coverage and database search strategies. Here we provide an overview of the field and argue that the latest mass spectrometry technologies that enable high mass accuracy at high acquisition rates will prove to be especially well suited for proteogenomics applications.

Introduction

Major efforts in genome sequencing at the turn of the 21st century made a profound impact on biology and laid the framework for analysis of biological systems at a global level. Modern generations of DNA sequencing technologies are capable of identifying and quantifying vast amounts of nucleic acid sequence, giving rise to an unprecedented number of sequenced genomes and fueling development of other global approaches, such as transcriptomics and proteomics. The raw genomic sequence needs to be correctly annotated and this is mostly achieved using *ab initio* gene prediction algorithms trained to recognize specific features of the open reading frames (ORFs).[1] However, the prediction algorithms often suffer from low accuracy leading to a high number of false positive gene predictions and wrongly predicted protein termini or exon–intron structure in eukaryotic genes.[2–7] Moreover, different algorithms may produce contradictory results making it difficult to get to an unambiguous annotation of a certain genome. Therefore, the ultimate validation of genome annotation lies in expression analysis and detection of gene products at the transcript (RNA) or the protein level.

Transcript analysis has been predominantly used for genome annotation since the relatively simple extraction protocols and application of hybridization techniques enabled a routine high coverage of the transcriptome. Conversely, due to a much higher chemical diversity of proteins, proteomics was long plagued by elaborate sample preparation protocols and immature protein separation, sequencing and data processing workflows—which led to a comparably poorer proteome coverage and therefore to a more limited application of proteomics to genome annotation. However, this situation is rapidly changing. With development of new methodologies for protein extraction, separation and especially detection using high precision mass spectrometry,[8] proteomics becomes increasingly capable of comprehensively and reliably identifying gene products. Notably, smaller proteomes of *Protozoa* can be almost completely detected and quantified by mass spectrometry, albeit still at a considerable effort. Detection of the complete proteome of the yeast was recently reported[9] and other smaller proteomes, especially those of bacteria, are within reach.[10–12]

These improvements in proteome coverage resulted in an increased application of MS-based proteomics to genome annotation and refinement. In a modern "proteogenomics" experiment, the complete protein extract of an organism is digested into peptides, which are then mass-measured and fragmented in a mass spectrometer (Fig. 1a). Mass spectra are typically searched against a database containing six-frame translation of the raw genome assembly and can therefore identify new, unpredicted open reading frames and refine existing gene models in terms of protein start and stop positions, exon–intron structure as well as their exact boundaries (Fig. 1b). Although conceptually relatively simple, application of mass spectrometry to genome reannotation still presents a range of challenges, especially in terms of data analysis; six-frame translation databases significantly increase the search space, often requiring application of special strategies in database search and data processing. Here we will give a brief historical overview of the field of proteogenomics and outline current workflows in sample preparation, MS measurement and data processing strategies used in genome annotation by MS data.

Proteome Center Tuebingen, Interdepartmental Institute for Cell Biology, University of Tuebingen, Auf der Morgenstelle 15, 72076 Tuebingen, Germany. E-mail: boris.macek@uni-tuebingen.de; Fax: +49 (0)7071-29-5779; Tel: +49 (0)7071-29-70558

† This article is part of the 2011 Molecular BioSystems 'Emerging Investigators' issue: highlighting the work of outstanding young scientists at the chemical- and systems-biology interfaces.

This journal is © The Royal Society of Chemistry 2011

Downloaded by UNIVERSITAET TUEBINGEN on 07 January 2013
Published on 21 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MB00168F

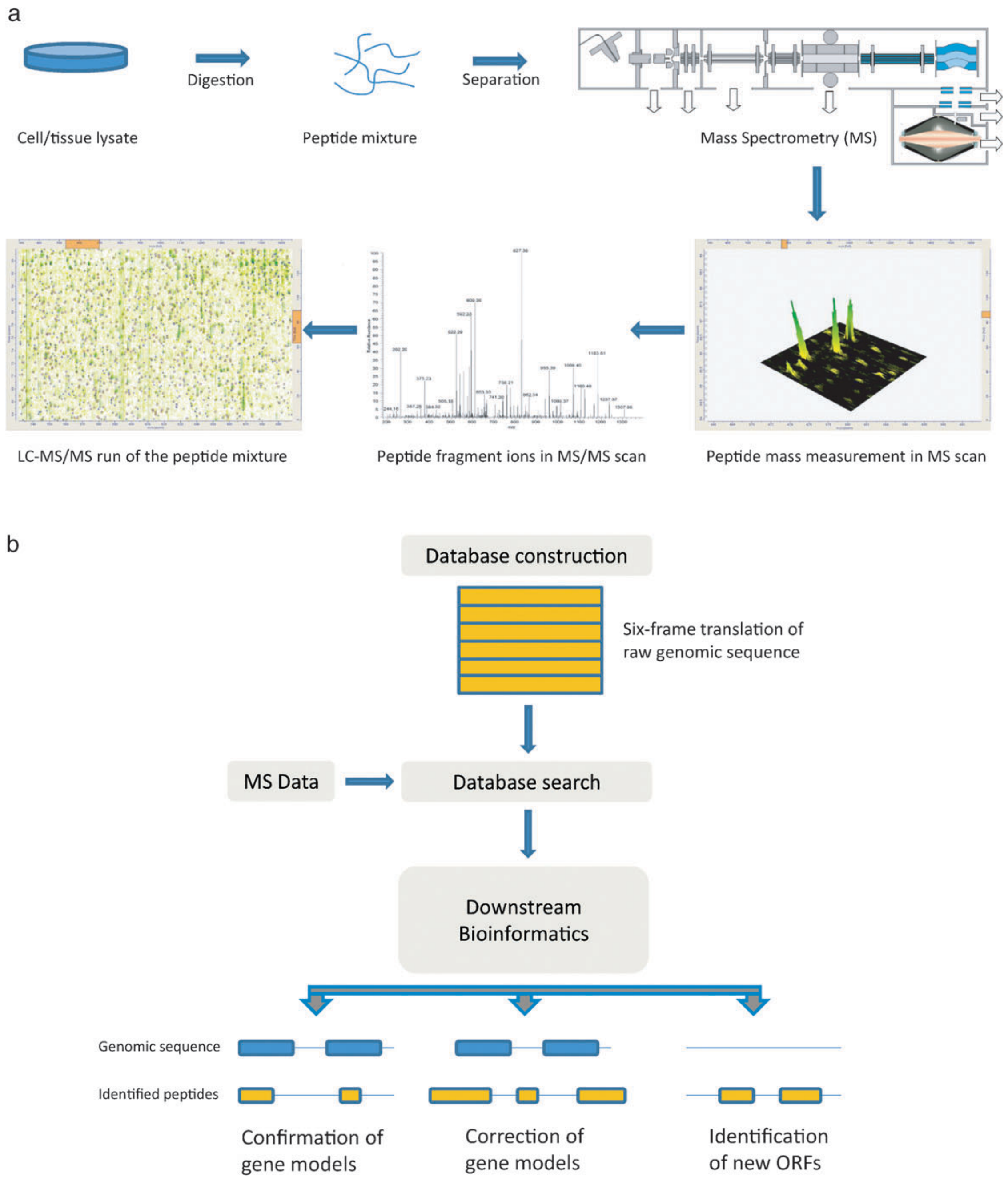

a
Digestion
Separation
Cell/tissue lysate
Peptide mixture
Mass Spectrometry (MS)
LC-MS/MS run of the peptide mixture
Peptide fragment ions in MS/MS scan
Peptide mass measurement in MS scan
b
Database construction
Six-frame translation of raw genomic sequence
MS Data
Database search
Downstream Bioinformatics
Genomic sequence
Identified peptides
Confirmation of gene models
Correction of gene models
Identification of new ORFs

Fig. 1 Workflows in a typical proteogenomics experiment. (a) Biochemical workflow. Proteins are extracted from a tissue or a cell line and digested by a protease into peptides. The resulting peptide mixtures are separated and analyzed by mass spectrometry; peptide masses are recorded in a "full scan" or "MS" spectrum; peptide ions are fragmented and the fragment ions are recorded in an "MS/MS" spectrum. Both levels of information are used in protein database search and peptide identification. (b) Data processing workflow. MS and MS/MS spectra are searched against a special database containing a six-frame translation of the whole genome assembly. Identified peptides are mapped onto the genome and provide three levels of information: (i) confirmation of the existing gene models; (ii) refinement of the existing gene models (*e.g.* repositioning of gene termini and exon/intron boundaries); (iii) identification of new expressed genomic regions (genes).

Historical overview

Combining proteomics and genomics is not a new concept. Already in 1995, the Yates group demonstrated the correlation of MS/MS spectra with nucleotide sequences[13] and since then several groups applied this concept to find new genes and to improve genome annotations in various organisms using different types of mass spectrometers and data analysis workflows. Early applications of MS to genome reannotation predominantly involved the use of low accuracy MS, often in combination with 2D gel separation of proteins. For example, Link *et al.* used triple-quadrupole and ion trap MS in combination with 2D PAGE to analyze abundant proteins in *Haemophilus influenzae*;[14] by searching the acquired spectra against the genomic sequence translated into six reading frames they identified several proteins that were not previously

Published on 21 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MB00168F
Downloaded by UNIVERSITAET TUEBINGEN on 07 January 2013

annotated. Neubauer *et al.* used triple quadrupole MS and searched expressed sequence tag (EST) databases with MS data to detect components of the mammalian spliceosome that were not contained in a comprehensive protein database[15] and Jungblut *et al.* demonstrated the expression of six genes in *Mycobacterium tuberculosis* not predicted by genomic approaches using 2D PAGE and MALDI-QTOF MS.[16] Soon after the first draft of the human genome became available, Choudhary *et al.* investigated the feasibility of searching the MS/MS spectra against a six-frame translation of all 23 human chromosomes and compared it to the search against protein and EST database.[17] More recent studies mainly employed ion trap mass spectrometers but with application of advanced, peptide-based "shot gun" proteomics[18] approaches instead of 2D PAGE protein separation. Oshiro *et al.* used a combination of expression profiling and ion trap MS to verify independent transcription and translation of genes in *Saccharomyces cerevisiae*,[19] whereas Jaffe *et al.* used ion trap MS and introduced the concept of a proteogenomic map to predict ORFs in *Mycoplasma pneumoniae* based on expressed protein-based evidence.[20] As the field of MS-based proteomics was developing, several public repositories of MS/MS data became available and were used in proteogenomics projects. Fermin *et al.* used MS data from the HUPO Plasma Project[21] (mostly ion trap MS) to search against a six-frame translation of the human genome in order to find novel blood proteins[22] and Tanner *et al.* used ion trap data as well as MS data obtained from the PeptideAtlas[23] to identify novel and extended genes, alternative splicing events and variant alleles of coding SNPs in human genome.[24] Recent applications of mass spectrometry to genomics included large-scale, mostly ion trap-based proteogenomics studies of model organisms such as *Shewanella oneidensis*,[25] several species of the *Mycobacterium* genus,[26–28] *Toxoplasma gondii*,[29] *Arabidopsis thaliana*,[30,31] *Caenorhabditis elegans*[32] and *Pristionchus pacificus*[33] (the latter study employed orbitrap MS and transcriptome analysis for genome annotation refinement). Importantly, all of the studies reported hundreds of completely new or reannotated genes at the protein level, demonstrating the potential and need of using mass spectrometry-based proteomics in genome reannotation.

Current workflows in proteogenomics

Application of mass spectrometry to genome annotation presents a range of challenges to all aspects of a proteomics workflow. Sample preparation is crucial for extraction of the complete proteome; mass spectrometry and data processing are essential for reliable peptide/protein detection; and downstream bioinformatics is important for interpretation of MS data and identification/validation of gene models. Here we will discuss current developments in all segments of a typical proteogenomics experiment.

Sample preparation

The aim of every proteogenomics experiment is to achieve as comprehensive proteome coverage as possible; therefore, the cell culture or tissue of interest needs to be homogenized and lysed in a way that enables the highest possible retrieval of its protein content. Ideally, proteins should be extracted and solubilized in a buffer containing a potent denaturing agent and a detergent, digested in solution, and the resulting peptide mixtures should be separated by at least two orthogonal separation methods prior to MS analysis. However, efficient extraction of membrane proteins requires potent ionic detergents, such as SDS, which complicates downstream processing of the sample as it inhibits commonly used proteases and hampers ionization in the MS. This problem is usually addressed either by separation of the protein extract by 1D SDS-PAGE and subsequent in-gel protein digestion,[34] or by application of less potent non-ionic detergents, such as *N*-octylglucoside, during sample homogenization/lysis. Both approaches, however, yield lower amounts of proteins—especially membrane proteins—compared to SDS-based lysis protocols. Recently, a sample preparation method that combines the advantages of the 1D gel-based workflow and in-solution digestion was introduced.[35,36] The method, termed filter-aided sample preparation (FASP) by the Mann lab, involves tissue homogenization/lysis in an SDS-containing denaturing buffer, which is followed by extensive washing, buffer exchange and protein digestion on a microcentricon filter. After digestion, peptides are spun through the filter in an MS-compatible buffer, collected and, if desired, further separated. Due to the straightforward sample processing, this method is applicable to low amounts of material, making it especially attractive for a comprehensive analysis of proteome in proteogenomics applications.

Peptide separation

After successful extraction, proteins are digested by a protease and the resulting peptide mixtures are mass-measured and fragmented in a mass spectrometer. Even if previously separated, tissue-derived protein mixtures usually give rise to thousands of peptides upon digestion. Since simultaneous ionization of such complex peptide mixtures would lead to a significant decrease of the measurement dynamic range (and therefore to the lower number of identifications), the peptides are further separated by liquid chromatography (LC) prior to MS measurement. Current LC-MS setups involve the use of nano-HPLC columns (inner diameter 25–75 μm and flow rates 100–500 nl min^{-1}), usually packed with reverse-phase C_{18} material and/or SCX material in the case of multi-dimensional chromatography.[18] These columns are directly coupled to an electrospray ionization source to minimize dead volume that may result in peak broadening after separation.[37] Such chromatographic setup results in a high peak capacity and resolving power, which directly influences the sensitivity and dynamic range of LC-MS measurement. Alternatively, HPLC fractions may be spotted on a target plate, mixed with a matrix and ionized using MALDI ionization.[38] An interesting development in peptide separation prior to MS is the introduction of the REPLAY chromatography,[39] which is suitable for analysis of extremely low amounts of material and enables performance of two consecutive and almost identical LC-MS analyses upon one sample injection, thus enabling targeted proteomics analysis and higher proteome coverage.

This journal is © The Royal Society of Chemistry 2011

Published on 21 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MB00168F
Downloaded by UNIVERSITAET TUEBINGEN on 07 January 2013

MS measurement

Despite the high importance of all upstream sample processing and separation workflows, mass spectrometry remains the most important component of every proteomics experiment.

Several different MS platforms and instruments are in routine use today and have been extensively reviewed elsewhere.[40] For the purpose of this review we will divide them into low accuracy (ion traps, triple quadrupoles) and high accuracy (Q-TOF, FT ICR, Orbitrap) mass spectrometers. Interestingly, ion traps are especially popular in proteogenomics applications, mainly due to their high sensitivity and fast scanning times that enable extensive proteome coverage. However, the ion trap data suffer from poor resolution and mass accuracy, which is typically 0.2–0.5 Da (200–500 ppm at 1000 m/z), on modern ion trap instruments. As discussed below, the low accuracy has wide-ranging implications for database search, as it requires high mass tolerances and therefore increases search space and decreases the search sensitivity. In addition, inability of most ion traps to simultaneously store all fragment ions[41] leads to a poorer coverage of the low mass fragment ions, often having as a consequence ambiguous sequence assignment of the peptide termini. Although this effect may be avoided in the newer ion traps (*e.g.* by "pulsed-Q" dissociation[42]), it can be problematic in combination with low mass accuracy of the precursor ion and may lead to increased false discovery rates when searching large proteogenomics databases, in which most of the entries are false (see below).

High accuracy instruments, such as FT ICR and Q-TOF typically have a very good resolution and can achieve sub-ppm mass accuracy, which decreases the search space and increases the sensitivity of the protein database search. However, a drawback of the high accuracy mass analyzers is a relatively low speed of acquisition, which leads to undersampling of the analyzed complex peptide mixtures and therefore to lower proteome coverage. Newer generations of hybrid mass spectrometers circumvent this problem by combining a low- and a high resolution mass analyzer and enabling their almost simultaneous action during peptide sequencing. The best example is the LTQ-Orbitrap, where the precursor (peptide) ion mass is typically measured at high resolution and accuracy in the Orbitrap mass analyzer, whereas the peptides are fragmented at high speed and sensitivity in the linear ion trap mass analyzer.[43] The resulting data provide a good trade-off between the sequencing speed and mass accuracy and therefore between proteome coverage and database search space. Especially the recently developed LTQ Orbitrap Velos shows a great promise for in-depth and accurate proteome analysis by combining a faster and more sensitive dual-pressure linear ion trap with a high accuracy Orbitrap mass analyzer. In addition, an improved higher energy collision-induced dissociation (HCD) collision cell enables acquisition of both MS and MS/MS spectra at a high speed and mass accuracy,[44] with a good fragment ion coverage across the mass range. The instrument can also be fitted with a chemical ionization source to enable electron transfer dissociation (ETD), another fragmentation method that achieves comprehensive fragmentation of multiply charged ions[45,46] and therefore provides data suitable for proteogenomics applications.

Data processing and database search

The final segment of a proteomics experiment is MS data processing. Specialized software converts the MS and MS/MS spectra into a format suitable for database search. At this stage, the latest generation of processing software uses high measurement mass accuracy of modern mass spectrometers to identify spectral features important for recalibration and quantification of stable isotope pairs, if these were used in the experiment; mass recalibration performed in this way can improve the mass accuracy by 5–10 fold.[47] Recalibrated and quantified peak lists are then submitted to a database search engine (such as Mascot,[48] Sequest[49] or OMSSA[50]), which identifies peptides by comparing measured fragmentation spectra with theoretical mass spectra of all peptides in a protein database, in this case a custom-made database containing six-frame translation of the raw genome assembly. The choice of the appropriate database is of great importance: specialized databases, consisting of forward and reversed (or randomized) protein sequences are nowadays commonly used for estimation of false discovery rates of database searches using a "target-decoy" approach,[51] but this approach is often impractical in proteogenomics due to the initial large size of six-frame translation databases. In a standard proteomics experiment, identified peptides are assembled into proteins or protein groups,[52] which represents a non-trivial and challenging task especially in eukaryotic proteomes where alternative splicing plays a key role. However, the protein interference problem does not explicitly occur in proteogenomics experiments, as only the identified peptide sequences—rather than assembled protein sequences—are used for the purpose of genome reannotation. These peptide sequences are used as extrinsic evidence for retraining of gene finding algorithms and, ultimately, refinement of gene models. Popular gene finding tools use generalized hidden Markov models (GeneZilla,[53] GlimmerHMM,[54] GeneScan,[55] SNAP[56]) and/or support vector machines (mGene[57,58]), which were shown to significantly improve the number of gene annotations. Retraining of these tools with experimentally determined gene expression data typically leads to their higher sensitivity and accuracy, as demonstrated in all proteogenomics studies so far.

Challenges in proteogenomics

Despite the recent breakthroughs in almost all segments of the proteogenomics workflow, the application of mass spectrometry to genome re-annotation is still a challenging task. The challenges mainly include the proteome coverage; dynamic range and sequencing speed of mass spectrometers; and construction of proteogenomics databases. As discussed, biochemical protocols for protein extraction and mass spectrometers are developing rapidly, leading to ever-increasing numbers of sequenced peptides and better proteome coverage; however, developments in DNA sequencing technologies appear to be even faster, leading to a more comprehensive coverage of gene expression at the transcript level. We recently reannotated the genome of the nematode model *P. pacificus* by sequencing the RNA libraries using 454 Roche pyrosequencing platform and by sampling the proteome using LTQ Orbitrap

Downloaded by UNIVERSITAET TUEBINGEN on 07 January 2013
Published on 21 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MB00168F

mass spectrometry.[30] The transcriptome analysis led to identification of >700 000 expressed sequence tags, whereas the proteome analysis led to identification of >30 000 non-redundant peptide sequences. Despite this apparent discrepancy in coverage, the proteome data are still of high qualitative importance as they provide evidence on the *translation* of the genetic message.

Another problem of proteogenomics concerns the construction of appropriate protein databases, search of MS spectra against these databases and the statistical validation of identified peptides. These computational challenges in proteogenomics were recently reviewed by Castellana and Bafna.[59] To enable identification of all possible gene products from MS data, the "raw" genome assembly needs to be translated *in silico* into all six reading frames, leading to at least six-fold increase in entries compared to the standard protein database and therefore to much larger database search space. While the construction of an appropriate proteogenomics database is relatively straightforward for microbial genomes due to their relatively small size and lack of alternative splicing, proteogenomics databases of eukaryotes are far more complex due to the exon–intron gene structure, which leads to a disproportional increase in database size. Since many more theoretical peptides from a typical proteogenomics database are considered for every acquired MS/MS spectrum, the overall peptide scores tend to be lower, leading to a decrease in the sensitivity (number of identifications) and specificity (number of correct identifications) during database search. Importantly, of all sequences in such a database, less than 20% correspond to true protein sequences (approximately one out of six reading frames), whereas the vast majority present non-sense protein entries. This makes the database searches in proteogenomics prone to high false discovery rates, especially when combined with low mass accuracy MS data.

High accuracy mass spectrometry has an intrinsic potential to decrease the search space and therefore increase the database search sensitivity and specificity. Fig. 2 depicts the number of theoretical peptide sequences (search space) as a function of precursor mass error exemplified on a relatively simple genome of *Bacillus subtilis*. For each theoretical tryptic peptide from the *B. subtilis* decoy protein database (13 000 entries) all peptides falling into a certain mass difference bin were counted—for example for a 1000 Da peptide in the 1 ppm mass tolerance bin all peptides in the window of 1000 ± 0.001 Da were counted. This procedure was repeated for all theoretical peptides and the distribution of all theoretical peptides observed in one mass tolerance bin is defined as the search space. Even in this simple case it is visible that high measurement mass accuracy enables a narrow mass tolerance during database search and limits the number of candidate peptide sequences that "compete" for an MS spectrum acquired at a given mass accuracy. Lower number of candidate peptide sequences leads to a smaller search space and this in turn leads to an increased sensitivity and specificity of the database search. However, the actual mass tolerance during database search usually exceeds the achieved measurement mass accuracy by several folds to provide ample number of candidate peptide sequences and prevent "forcing" the search engine to report a particular peptide. The basic considerations of database search space and strategies are covered elsewhere.[60]

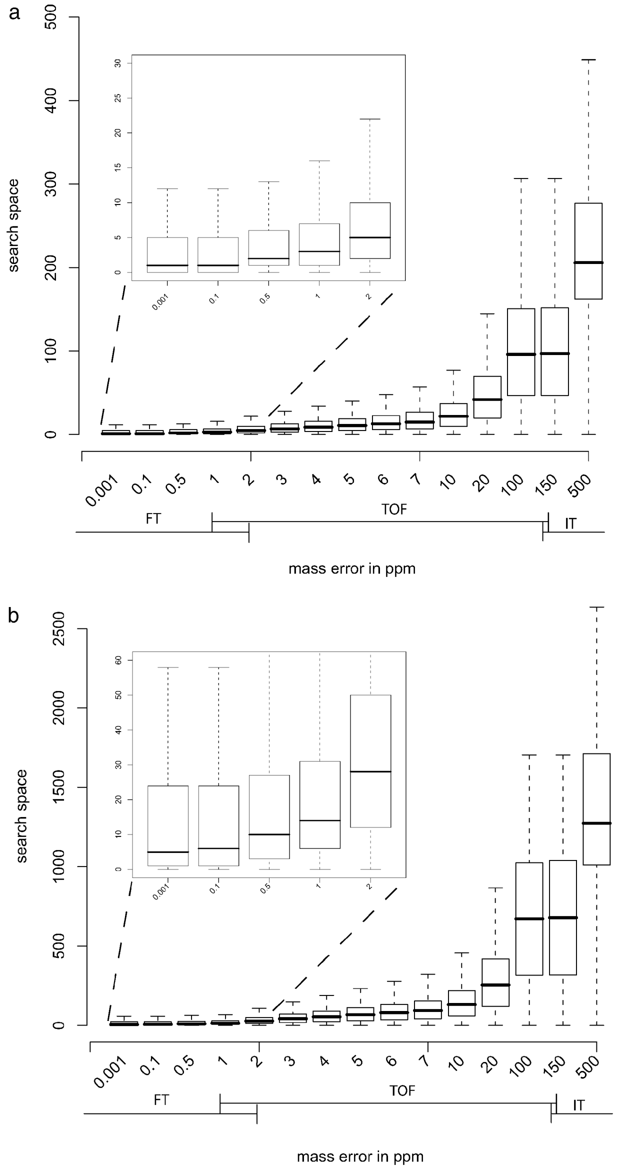


Fig. 2 Relationship of the search space and mass tolerance in database search. The data are presented as a box plot where the full horizontal line is the median size of the search space. This is only a simplified presentation using a small protein database and the simplest search parameters (no missed cleavages, one charge state, no modification, *etc.*). The actual search space depends on many parameters and is in practice much larger. (a) Database search space as a function of mass tolerance in a standard *B. subtilis* decoy protein database. Note that the number of theoretical peptides (search space) is >30 times higher at a mass accuracy of 200–500 ppm (commonly achieved by ion trap MS), compared to 1 ppm (commonly achieved by orbitrap MS). (b) Database search space as a function of mass tolerance in a six frame translation database of *B. subtilis*. The search space in this relatively simple database used in proteogenomics applications is about six times higher than in a standard proteome database. Note that the more complex eukaryotic databases lead to even higher increase of search space due to alternative splicing.

Various bioinformatics strategies have been applied to reduce the search space in proteogenomics experiments. Küster *et al.* used peptide sequence tags, short sequences of only a few

This journal is © The Royal Society of Chemistry 2011

Published on 21 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MB00168F
Downloaded by UNIVERSITAET TUEBINGEN on 07 January 2013

amino acids, to query the raw genome sequence of *A. thaliana*.[61] These short tags are determined *de novo* and define a sequence of few amino acids from the masses of adjacent fragment ions in the MS/MS spectrum. Using this information and the precursor ion mass, a search "template" is created for each peptide and used in database search, thereby significantly reducing the number of theoretical peptide sequences that "compete" for a positive identification of an MS/MS spectrum. Sevinsky *et al.* developed a method for reduction of the peptide search space by considering isoelectric point and accurately determined precursor ion mass in database search.[62] In this approach, peptides are fractionated using isoelectric focusing and the pI range of each fraction is determined; only peptides with theoretical pI matching the observed pI range are considered, greatly increasing the sensitivity of database search. Waridel *et al.* used a sequence similarity-driven approach by combining conventional database searches, *de novo* sequencing and MS BLAST searches to characterize the proteome of the unsequenced organism *Dunaliella salina*.[63] Spectra that could not be confidently assigned to the database were filtered and further processed by *de novo* sequencing algorithm; confidently identified peptide sequences were submitted to MS BLAST to identify unknown proteins. Some approaches make use of the evolutionary information encoded in multiple genomes of closely related organisms. Gupta *et al.* used such a comparative approach on three bacterial genomes of *Shewanella* under assumption that the proteins and their start sites are highly conserved between the analyzed species.[64] Based on this assumption there is a higher probability that a protein is expressed when it was seen in different species although it was not confidently identified. Furthermore, peptides that did not fall into previously annotated regions can be used to identify programmed frame shifts and sequencing errors. A similar approach, termed orthoproteogenomics was introduced by Gallien *et al.* to perform a simultaneous refinement and annotation of multiple genomes at once.[26] Recently, Merrihew *et al.* used shotgun proteomics to refine the genome annotation of *C. elegans*.[32] By using conserved sequences in related nematodes as putative ORFs, *in silico* gene predictions as well as the protein database from the Wormbase, they refined existing and identified new gene models in this well-studied model nematode. They avoided to explicitly search against a six-frame translation with MS/MS spectra but used translated intergenic regions to identify homologous sequences in other nematodes which were included as putative ORFs in the database used for MS identification.

Another challenge in proteogenomics studies of eukaryotes is the identification of peptides spanning a splice junction. While these peptides are crucial for determination of the exact exon–intron borders they cannot be detected by searching a raw six-frame translation with a standard search engine as mentioned above. Allmer *et al.*[65,66] described an approach to identify intron-split peptides by combining *de novo* MS/MS sequencing and classical search engines like Sequest or Mascot. *De novo* deduced peptide sequences are aligned to the genomic sequence to assemble a database of possible peptides that match a particular mass spectrum. This database can be queried by an MS/MS search engine to identify and validate peptides that span an intron–exon boundary. A different approach, proposed by Chen[67] and Colinge *et al.*[68] is to directly account for splice donor and acceptor sites within the genomic database. These groups used predicted donor/acceptor sites to generate putative spliced peptides that can be queried by MS/MS spectra. Finally, the search engine described in Roos *et al.*[69] can also be used to query genomic databases directly with MS/MS spectra taking potential GT-AT introns with a given gap size into account.

Conclusions and outlook

Recent developments in all segments of the proteomics workflow have enabled a wider application of mass spectrometry to genome annotation, which is documented in an increasing number of large-scale proteogenomics studies. Although this approach still suffers from the lower coverage compared to the transcriptome analysis, analysis of gene expression at the protein level is invaluable for determination and prediction of protein-coding (translated) genes. Interestingly, even the most recent proteogenomics studies predominantly employ low accuracy mass spectrometers that enable high proteome coverage but at the cost of database search sensitivity and specificity. Together with the potent bioinformatics strategies that were developed to circumvent this problem, the future of the field lies in the use of the newest generation of fast scanning *and* high accuracy hybrid mass spectrometers that have an intrinsic capability to provide both high proteome coverage and database search specificity. Combined with the necessary standardization of the data processing workflows, proteogenomics based on these new technologies will prove to be an invaluable tool in the future efforts in interpretation of genome sequencing data.

Abbreviations

SDS	sodium dodecyl sulfate
PAGE	polyacrylamid gel electrophoresis
HUPO	Human Proteome Organization
EST	expressed sequence tags
ppm	parts-per-million
CID	collision-induced dissociation
ETD	electron transfer dissociation
ECD	electron capture dissociation

Acknowledgements

The Proteome Center Tuebingen is funded by the Landesstiftung Baden Wuerttemberg and Ministerium fuer Wissenschaft und Kunst Baden Wuerttemberg (Juniorprofessorenprogramm to BM).

References

1 *Modern genome annotation. The biosapiens network*, Ed. D. Frishman and A. Valencia, Springer, New York, 1st edn, 2008.
2 J. Starmer, A. Stomp, M. Vouk and D. Bitzer, Predicting Shine-Dalgarno sequence locations exposes genome annotation errors, *PLoS Comput. Biol.*, 2006, **2**, e57.
3 D. Devos and A. Valencia, Intrinsic errors in genome annotation, *Trends Genet.*, 2001, **17**, 429–431.

Published on 21 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MB00168F
Downloaded by UNIVERSITAET TUEBINGEN on 07 January 2013

4 J. Lamontagne, M. Beland, A. Forest, A. Cote-Martin, N. Nassif, F. Tomaki, I. Moriyon, E. Moreno and E. Paramithiotis, Proteomics-based confirmation of protein expression and correction of annotation errors in the *Brucella abortus* genome, *BMC Genomics*, 2010, **11**, 300.

5 G. Parra, P. Agarwal, J. F. Abril, T. Wiehe, J. W. Fickett and R. Guigo, Comparative gene prediction in human and mouse, *Genome Res.*, 2003, **13**, 108–117.

6 J. Reboul, P. Vaglio, J. F. Rual, P. Lamesch, M. Martinez, C. M. Armstrong, S. Li, L. Jacotot, N. Bertin, R. Janky, T. Moore, J. R. Hudson, Jr., J. L. Hartley, M. A. Brasch, J. Vandenhaute, S. Boulton, G. A. Endress, S. Jenna, E. Chevet, V. Papasotiropoulos, P. P. Tolias, J. Ptacek, M. Snyder, R. Huang, M. R. Chance, H. Lee, L. Doucette-Stamm, D. E. Hill and M. Vidal, *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression, *Nat. Genet.*, 2003, **34**, 35–41.

7 P. Flicek, E. Keibler, P. Hu, I. Korf and M. R. Brent, Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map, *Genome Res.*, 2003, **13**, 46–54.

8 M. Mann and N. L. Kelleher, Precision proteomics: the case for high resolution and high mass accuracy, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 18132–18138.

9 L. M. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frohlich, T. C. Walther and M. Mann, Comprehensive mass-spectrometry-based proteome quantification of haploid *versus* diploid yeast, *Nature*, 2008, **455**, 1251–1254.

10 M. Iwasaki, S. Miwa, T. Ikegami, M. Tomita, N. Tanaka and Y. Ishihama, One-dimensional capillary liquid chromatographic separation coupled with tandem mass spectrometry unveils the *Escherichia coli* proteome on a microarray scale, *Anal. Chem.*, 2010, **82**, 2616–2620.

11 D. Becher, K. Hempel, S. Sievers, D. Zuhlke, J. Pane-Farre, A. Otto, S. Fuchs, D. Albrecht, J. Bernhardt, S. Engelmann, U. Volker, J. M. van Dijl and M. Hecker, A proteomic view of an important human pathogen – towards the quantification of the entire *Staphylococcus aureus* proteome, *PLoS One*, 2009, **4**, e8176.

12 B. Soufi, C. Kumar, F. Gnad, M. Mann, I. Mijakovic and B. Macek, Stable isotope labeling by amino acids in cell culture (SILAC) applied to quantitative proteomics of *Bacillus subtilis*, *J. Proteome Res.*, 2010, **9**, 3638–3646.

13 J. R. Yates, 3rd, J. K. Eng and A. L. McCormack, Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases, *Anal. Chem.*, 1995, **67**, 3202–3210.

14 A. J. Link, L. G. Hays, E. B. Carmack and J. R. Yates, Identifying the major proteome components of *Haemophilus influenzae* type-strain NCTC 8143, *Electrophoresis*, 1997, **18**, 1314–1334.

15 G. Neubauer, A. King, J. Rappsilber, C. Calvio, M. Watson, P. Ajuh, J. Sleeman, A. Lamond and M. Mann, Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex, *Nat. Genet.*, 1998, **20**, 46–50.

16 P. R. Jungblut, E. C. Muller, J. Mattow and S. H. E. Kaufmann, Proteomics reveals open reading frames in *Mycobacterium tuberculosis* H37Rv not predicted by genomics, *Infect. Immun.*, 2001, **69**, 5905–5907.

17 J. S. Choudhary, W. P. Blackstock, D. M. Creasy and J. S. Cottrell, Interrogating the human genome using uninterpreted mass spectrometry data, *Proteomics*, 2001, **1**, 651–667.

18 D. A. Wolters, M. P. Washburn and J. R. Yates, 3rd, An automated multidimensional protein identification technology for shotgun proteomics, *Anal. Chem.*, 2001, **73**, 5683–5690.

19 G. Oshiro, L. M. Wodicka, M. P. Washburn, J. R. Yates, 3rd, D. J. Lockhart and E. A. Winzeler, Parallel identification of new genes in *Saccharomyces cerevisiae*, *Genome Res.*, 2002, **12**, 1210–1220.

20 J. D. Jaffe, H. C. Berg and G. M. Church, Proteogenomic mapping as a complementary method to perform genome annotation, *Proteomics*, 2004, **4**, 59–77.

21 K. Cottingham, HUPO Plasma Proteome Project: challenges and future directions, *J. Proteome Res.*, 2006, **5**, 1298.

22 D. Fermin, B. B. Allen, T. W. Blackwell, R. Menon, M. Adamski, Y. Xu, P. Ulintz, G. S. Omenn and D. J. States, Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics, *GenomeBiology*, 2006, **7**, R35.

23 F. Desiere, E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich and R. Aebersold, The PeptideAtlas project, *Nucleic Acids Res.*, 2006, **34**, D655–D658.

24 S. Tanner, Z. X. Shen, J. Ng, L. Florea, R. Guigo, S. P. Briggs and V. Bafna, Improving gene annotation using peptide mass spectrometry, *Genome Res.*, 2007, **17**, 231–239.

25 N. Gupta, S. Tanner, N. Jaitly, J. N. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R. D. Smith and P. A. Pevzner, Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation, *Genome Res.*, 2007, **17**, 1362–1377.

26 S. Gallien, E. Perrodou, C. Carapito, C. Deshayes, J. M. Reyrat, A. Van Dorsselaer, O. Poch, C. Schaeffer and O. Lecompte, Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol, *Genome Res.*, 2009, **19**, 128–135.

27 G. A. de Souza, T. Softeland, C. J. Koehler, B. Thiede and H. G. Wiker, Validating divergent ORF annotation of the *Mycobacterium leprae* genome through a full translation data set and peptide identification by tandem mass spectrometry, *Proteomics*, 2009, **9**, 3233–3243.

28 G. A. de Souza, H. Malen, T. Softeland, G. Saelensminde, S. Prasad, I. Jonassen and H. G. Wiker, High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using *Mycobacterium tuberculosis* as an example, *BMC Genomics*, 2008, **9**, 316.

29 D. Xia, S. J. Sanderson, A. R. Jones, J. H. Prieto, J. R. Yates, E. Bromley, F. M. Tomley, K. Lal, R. E. Sinden, B. P. Brunk, D. S. Roos and J. M. Wastling, The proteome of *Toxoplasma gondii*: integration with the genome provides novel insights into gene expression and annotation, *GenomeBiology*, 2008, **9**, R116.

30 N. E. Castellana, S. H. Payne, Z. Shen, M. Stanke, V. Bafna and S. P. Briggs, Discovery and revision of Arabidopsis genes by proteogenomics, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 21034–21038.

31 K. Baerenfaller, J. Grossmann, M. A. Grobei, R. Hull, M. Hirsch-Hoffmann, S. Yalovsky, P. Zimmermann, U. Grossniklaus, W. Gruissem and S. Baginsky, Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics, *Science*, 2008, **320**, 938–941.

32 G. E. Merrihew, C. Davis, B. Ewing, G. Williams, L. Kall, B. E. Frewen, W. S. Noble, P. Green, J. H. Thomas and M. J. MacCoss, Use of shotgun proteomics for the identification, confirmation, and correction of *C. elegans* gene annotations, *Genome Res.*, 2008, **18**, 1660–1669.

33 N. Borchert, C. Dieterich, K. Krug, W. Schutz, S. Jung, A. Nordheim, R. J. Sommer and B. Macek, Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models, *Genome Res.*, 2010, **20**, 837–846.

34 M. Schirle, M. A. Heurtier and B. Kuster, Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography–tandem mass spectrometry, *Mol. Cell. Proteomics*, 2003, **2**, 1297–1305.

35 L. L. Manza, S. L. Stamer, A. J. Ham, S. G. Codreanu and D. C. Liebler, Sample preparation and digestion for proteomic analyses using spin filters, *Proteomics*, 2005, **5**, 1742–1745.

36 J. R. Wisniewski, A. Zougman, N. Nagaraj and M. Mann, Universal sample preparation method for proteome analysis, *Nat. Methods*, 2009, **6**, 359–362.

37 Y. Ishihama, J. Rappsilber, J. S. Andersen and M. Mann, Microcolumns with self-assembled particle frits for proteomics, *J. Chromatogr., A*, 2002, **979**, 233–239.

38 F. Hillenkamp, M. Karas, R. C. Beavis and B. T. Chait, Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers, *Anal. Chem.*, 1991, **63**, 1193A–1203A.

39 L. F. Waanders, R. Almeida, S. Prosser, J. Cox, D. Eikel, M. H. Allen, G. A. Schultz and M. Mann, A novel chromatographic method allows on-line reanalysis of the proteome, *Mol. Cell. Proteomics*, 2008, **7**, 1452–1459.

40 F. E. Ahmed, Utility of mass spectrometry for proteome analysis: part I. Conceptual and experimental approaches, *Expert Rev. Proteomics*, 2008, **5**, 841–864.

This journal is © The Royal Society of Chemistry 2011

Published on 21 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MB00168F
Downloaded by UNIVERSITAET TUEBINGEN on 07 January 2013

41 E. d. Hoffmann and V. Stroobant, *Mass spectrometry: principles and applications*, J. Wiley, Chichester, West Sussex, England, Hoboken, NJ, 3rd edn, 2007.

42 J. C. Schwartz, J. P. Syka and S. T. Quarmby, *Improving the Fundamentals of Msn on 2D ion traps: new ion activation and isolation techniques*, *Proceedings of the 53rd ASMS Conference on Mass Spectrometry (June 5–9)*, San Antonio, Texas, 2005.

43 J. V. Olsen, L. M. de Godoy, G. Li, B. Macek, P. Mortensen, R. Pesch, A. Makarov, O. Lange, S. Horning and M. Mann, Parts per million mass accuracy on an Orbitrap mass spectrometer *via* lock mass injection into a C-trap, *Mol. Cell. Proteomics*, 2005, **4**, 2010–2021.

44 J. V. Olsen, J. C. Schwartz, J. Griep-Raming, M. L. Nielsen, E. Damoc, E. Denisov, O. Lange, P. Remes, D. Taylor, M. Splendore, E. R. Wouters, M. Senko, A. Makarov, M. Mann and S. Horning, A dual pressure linear ion trap–Orbitrap instrument with very high sequencing speed, *Mol. Cell. Proteomics*, 2009, **8**, 2759–2769.

45 J. E. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz and D. F. Hunt, Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 9528–9533.

46 C. D. Wenger, G. C. McAlister, Q. Xia and J. J. Coon, Sub-part-per-million precursor and product mass accuracy for high-throughput proteomics on an electron transfer dissociation-enabled orbitrap mass spectrometer, *Mol. Cell. Proteomics*, 2010, **9**, 754–763.

47 J. Cox and M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, *Nat. Biotechnol.*, 2008, **26**, 1367–1372.

48 D. N. Perkins, D. J. Pappin, D. M. Creasy and J. S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, 1999, **20**, 3551–3567.

49 J. R. Yates, 3rd, J. K. Eng, A. L. McCormack and D. Schieltz, Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database, *Anal. Chem.*, 1995, **67**, 1426–1436.

50 L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi and S. H. Bryant, Open mass spectrometry search algorithm, *J. Proteome Res.*, 2004, **3**, 958–964.

51 J. E. Elias and S. P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, *Nat. Methods*, 2007, **4**, 207–214.

52 A. I. Nesvizhskii and R. Aebersold, Interpretation of shotgun proteomic data: the protein inference problem, *Mol. Cell. Proteomics*, 2005, **4**, 1419–1440.

53 W. H. Majoros, M. Pertea, A. L. Delcher and S. L. Salzberg, Efficient decoding algorithms for generalized hidden Markov model gene finders, *BMC Bioinf.*, 2005, **6**, DOI: 10.1186/1471-2105-6-16.

54 S. L. Salzberg, A. L. Delcher, S. Kasif and O. White, Microbial gene identification using interpolated Markov models, *Nucleic Acids Res.*, 1998, **26**, 544–548.

55 C. Burge and S. Karlin, Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, 1997, **268**, 78–94.

56 I. Korf, Gene finding in novel genomes, *BMC Bioinformatics*, 2004, **5**, 59.

57 G. Schweikert, A. Zien, G. Zeller, J. Behr, C. Dieterich, C. S. Ong, P. Philips, F. De Bona, L. Hartmann, A. Bohlen, N. Kruger, S. Sonnenburg and G. Ratsch, mGene: accurate SVM-based gene finding with an application to nematode genomes, *Genome Res.*, 2009, **19**, 2133–2143.

58 G. Schweikert, J. Behr, A. Zien, G. Zeller, C. S. Ong, S. Sonnenburg and G. Ratsch, mGene.web: a web service for accurate computational gene finding, *Nucleic Acids Res.*, 2009, **37**, W312–W316.

59 N. Castellana and V. Bafna, Proteogenomics to discover the full coding content of genomes: a computational perspective, *J. Proteomics*, 2010, **73**(11), 2124–2135.

60 I. Eidhammer, *Computational methods for mass spectrometry proteomics*, John Wiley & Sons, Chichester, England, Hoboken, NJ, 2007.

61 B. Küster, P. Mortensen, J. S. Andersen and M. Mann, Mass spectrometry allows direct identificationof proteins in large genomes, *Proteomics*, 2001, **1**, 641–650.

62 J. R. Sevinsky, B. J. Cargile, M. K. Bunger, F. Meng, N. A. Yates, R. C. Hendrickson and J. L. Stephenson, Jr., Whole genome searching with shotgun proteomic data: applications for genome annotation, *J. Proteome Res.*, 2008, **7**, 80–88.

63 P. Waridel, A. Frank, H. Thomas, V. Surendranath, S. Sunyaev, P. Pevzner and A. Shevchenko, Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automated *de novo* sequencing, *Proteomics*, 2007, **7**, 2318–2329.

64 N. Gupta, J. Benhamida, V. Bhargava, D. Goodman, E. Kain, I. Kerman, N. Nguyen, N. Ollikainen, J. Rodriguez, J. Wang, M. S. Lipton, M. Romine, V. Bafna, R. D. Smith and P. A. Pevzner, Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes, *Genome Res.*, 2008, **18**, 1133–1142.

65 J. Allmer, B. Naumann, C. Markert, M. Zhang and M. Hippler, Mass spectrometric genomic data mining: novel insights into bioenergetic pathways in *Chlamydomonas reinhardtii*, *Proteomics*, 2006, **6**, 6207–6220.

66 J. Allmer, C. Markert, E. J. Stauber and M. Hippler, A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases, *FEBS Lett.*, 2004, **562**, 202–206.

67 T. Chen, Gene-finding *via* tandem mass spectrometry, *RECOMB 2001 Proceedings of the Fifth Annual International Conference on Computational Biology*, 2001, 87–94.

68 J. Colinge, I. Cusin, S. Reffas, E. Mahe, A. Niknejad, P. A. Rey, H. Mattou, M. Moniatte and L. Bougueleret, Experiments in searching small proteins in unannotated large eukaryotic genomes, *J. Proteome Res.*, 2005, **4**, 167–174.

69 F. F. Roos, R. Jacob, J. Grossmann, B. Fischer, J. M. Buhmann, W. Gruissem, S. Baginsky and P. Widmayer, PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra, *Bioinformatics*, 2007, **23**, 3016–3023.

Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments

Karsten Krug, Alejandro Carpy, Gesa Behrends, Katarina Matic, Nelson C. Soares, Boris Macek[¶]

Proteome Center Tuebingen, University of Tuebingen, Germany

[¶] To whom correspondence should be addressed:

Prof. Dr. Boris Macek
Proteome Center Tuebingen
Interfaculty Institute for Cell Biology
University of Tuebingen
Auf der Morgenstelle 15
72076 Tuebingen
Germany
Phone: +49/(0)7071/29-70558
Fax: +49/(0)7071/29-5779
E-Mail: boris.macek@uni-tuebingen.de

Running Title: Proteogenomics of E. coli

Summary

Recent advances in mass spectrometry (MS) have led to increased applications of shotgun proteomics to the refinement of genome annotation. The typical “proteogenomic” workflows rely on mapping of peptide MS/MS spectra onto databases derived by six-frame translation of the genome sequence. These databases contain a large proportion of spurious protein sequences which make the statistical confidence of the resulting peptide spectrum matches (PSMs) difficult to assess. Here we perform a comprehensive analysis of the *Escherichia coli* proteome using LTQ-Orbitrap MS and map the corresponding MS/MS spectra onto a six-frame translation of the *E. coli* genome. We assume complete annotation of the *E. coli* genome and regard all six frame-specific (novel) PSMs as false positive identifications. This enables us to assess the sensitivity, specificity, accuracy and actual false discovery rate in a typical bacterial proteogenomic dataset. To increase reliability of our results we use two complementary computational frameworks for processing and statistical assessment of MS/MS data: MaxQuant and Trans-Proteomic Pipeline. We show that the posterior error probability distribution of novel hits is almost identical to that of reversed (decoy) hits, pointing to substantial underestimation of FDR even in “simple” proteogenomic experiments obtained by high accuracy MS. The use of a small and well annotated bacterial genome enables us to address genome coverage achieved in state-of-the-art bacterial proteomics: identified peptide sequences mapped to all estimated expressed *E. coli* proteins, but covered 27.5% of the total genome sequence. Our results point to the necessity for further technological and bioinformatic improvements in proteogenomic strategies.

Introduction

MS-based proteomics has become an indispensable tool to study *in vivo* protein expression on a global scale (1). Briefly, in a typical "shotgun" proteomic experiment the whole proteome of an organism is extracted and digested by a protease (e.g. trypsin). The resulting complex peptide mixtures are usually further fractionated and separated by liquid chromatography (LC) before ionization and analysis in the mass spectrometer. Recent innovations in MS technology (2-4) enable high peptide sequencing rates at high mass accuracy and sensitivity, making the routine analysis of entire proteomes within reach (5, 6).

Modern genome annotation uses computational *ab initio* approaches to predict coding regions and gene models from raw sequencing data (7, 8). Since the ultimate evidence of gene expression is the detection of its product, transcriptomic data are commonly used to train gene prediction algorithms (9). Similarly, MS-based proteomics is increasingly used in genome annotation. In a typical proteogenomics experiment MS/MS spectra of peptides are searched against databases derived by *in-silico* six frame translation of the whole genome sequence (10-14). This approach has been applied, alone or in combination with transcriptomic data, in order to refine genome annotation in several organisms, including *C. elegans* (15), *P. pacificus* (16), *S. cerevisiae* (17), *S. pombe* (18), *A. thaliana* (19) *, S. nodorum* (20), *T. gondii* (21), *A. gambiae* (22), mouse (23) and human (24, 25). Bacteria are especially well suited to MS-assisted genome annotation, due to their relatively simple genome structures and small genome sizes that lead to overall better sequence coverage in a typical proteomics experiment (26-33).

Use of six-frame databases in proteogenomics experiments is challenging due to their large sizes which increase the search space as well as affect the sensitivity of database searches (34). Additionally, these databases contain a high proportion of artificial sequences resulting from frames that are not transcribed (13, 35). These spurious protein sequences are difficult to discriminate from the true protein sequences, which make the statistical confidence of the resulting peptide spectrum matches (PSMs) difficult to calculate.

Here we take the advantage of the small size (4.6Mb), simple architecture and a high annotation level of the *Escherichia coli* genome, and use it as benchmark model for proteogenomic data interpretation. We derive a comprehensive dataset of proteins expressed in the exponential growth of *Escherichia coli* and map the corresponding MS/MS spectra onto a six-frame translation of the *E. coli* genome. We assume complete annotation of the *E. coli* genome, which enables us to regard all six frame-specific (novel) PSMs as wrongly identified and to assess the actual false discovery rate in a simple proteogenomic experiment. We show that the posterior error probability distribution of novel peptides is almost identical to that of decoy (reversed) hits, which validates our assumption and points to the accumulation of false positive PSMs within novel peptide identifications. Our dataset comprises 2,600 *E. coli* proteins, approaching the identification of the complete proteome expressed during the exponential growth (36). The combined peptide sequences cover 48.5% of the expressed genome sequence but only about 27.5% of the total genome sequence.

Experimental Procedures

Bacterial Cell Culture

Wild-type *E. coli* strain K12 (isolate BW25113) (37) was inoculated in 5 mL Lysogeny Broth Luria/Miller (LB) medium at 37 °C under vigorous shaking for 24 h (OD_{600}=1.9), then 1 mL of the stationary culture was spun down at 260 x g for 10 min, in order to remove any remaining from the LB medium. The bacterial cells were washed twice with M9 minimal medium consisting of M9 salts (6.78 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 0.5 g/L NaCl, 1 g/L NH_4Cl, Sigma-Aldrich) supplemented with additional 0.5% (w/v) glucose, 33 µM thiamine, 1mM $MgSO_4$, 0.1 mM $CaCl_2$. Next, the resultant pellet was resuspended in a final volume of 1 mL M9. Immediately after, 5 µL of this culture were used to inoculate 5 mL of fresh M9 medium, containing 0.25 mg/mL of lysine (Sigma-Aldrich,). Overnight minimal medium cell cultures were grown at 37 °C under vigorous shaking to an OD_{600}=0.5 and used to inoculate (1:100 dilution) 125 mL of fresh minimal medium containing 0.25 mg/mL lysine. The cell cultures were grown to OD_{600}=0.5, harvested by centrifugation at 3345 *x g* for 10 min, washed with phosphate buffered saline (PBS) and snap-frozen in liquid nitrogen.

Protein Extraction

The frozen cell pellets were resuspended in 3-5 mL lysis buffer (pH 7.5) containing 2 mg/mL lysozyme (Sigma-Aldrich) in 50 mM Tris/HCl buffer, 1 mM EDTA and 5 mM of each of the following phosphatase inhibitors: glycerol-2-phosphate; sodium fluoride (Sigma-Aldrich) and sodium orthovanadate (Alfa Aesar). Cell wall lysis was performed at 37 °C for 15 min and DNA was comminuted by benzonase (1875 U) (Merck) for additional 10 min. For solubilization of membrane proteins, lithiumdodecylsulfate (LDS) (Sigma-Aldrich) was added to a final concentration of 1% (w/v) and was incubated at 37 °C under vigorous shaking for 15 min. Cell debris was removed by centrifugation at 3345 x *g* for 5 min and repeated centrifugation of the supernatant at 11,300 x *g* for 10 min. The crude protein extract was methanol/chloroform precipitated and the protein precipitates were redissolved in denaturation buffer containing 6 M urea/2 M thiourea in 10 mM Tris buffer. For estimation of the protein concentration, each extract was measured by Bradford assay (Bio-Rad).

SDS-PAGE and In-Gel Digestion

In-Gel digestion was performed as previously described (16). Briefly, extracted proteins were separated on a NuPage Bis-Tris 4-12% gradient gel (Invitrogen). The gel was stained with Coomassie Blue and subsequently cut into 15 slices. Resulting gel pieces were destained by washing three times with 10 mM ammonium bicarbonate (ABC) and acetonitrile (ACN) (1:1, v/v). Proteins were then reduced with 10 mM dithiothreitol (DTT) in 20 mM (ABC) for 45 min at 56°C and alkylated with 55 mM iodoacetamide IAA in 20 mM ABC for 30 min at room temperature in the dark. After washing two times with 5 mM ABC and one time with ACN, the gel pieces were dehydrated in a vacuum centrifuge. Proteins were either digested with trypsin (Promega) or Lys-C (Wako) (12.5 ng/µL in 20 mM ABC) at 37°C over night. Resulting peptides were extracted in three subsequent steps with the following solutions: I) 3% TFA in 30% ACN II) 0.5% acetic acid in 80% ACN III) 100% ACN. After evaporation of the ACN in a vacuum centrifuge peptide fractions were desalted using StageTips (38).

In-Solution Digestion

Protein extracts were reduced for 1 hour at room temperature with 1mM dithiothreitol (DTT) and subsequently alkylated with 1 mM iodoacetamide (IAA) for 1 hour at room temperature in the dark. Proteins were pre-digested with Lys-C (1:100 w/w) for 3 hours at room temperature. After dilution with 4 volumes of 20 mM ammonium bicarbonate (ABC), proteins were digested overnight at room temperature with either trypsin (1:100 w/w) or Lys-C (1:100 w/w).

Off-Gel Isoelectric Focusing

Peptides derived from the in-solution digestion were separated according to their isoelectric point using the 3100 OffGel fractionator (Agilent) following the manufacturer´s instructions. Peptides mixtures were separated into 12 fractions using 13 cm Immobiline DryStrips with a pH 3-10 gradient (GE Healthcare). Separation was performed at a maximum current of 50 µA until 50 kVH were reached. Peptide fractions were acidified with acidic solution (30% CAN, 5% Acetic Acid and 10% trifluoracetic acid in water) and desalted using Stage-Tips.

Strong Anion Exchange Chromatography

Peptides from the in-solution digestion were desalted using solid phase extraction. Strong Anion Exchange Chromatography (SAX) was performed as described before (39). Briefly, desalted peptides were loaded at pH 11 onto an anion exchange column containing 6 layers of Empore/Disk Anion Exchange (Varian) in a 200 µL pipette tip. For conditioning and elution the Britton & Robinson Universal Buffer (0.02 M Ch_3COOH, 0.02 M H_3PO_4 and 0.02 M H_3BO_3) at pH 3, 4, 5, 6, 8 and 11 was prepared. The column was activated with Methanol and conditioned with 1M NaOH followed by buffer pH 11. The flow-through was acidified with acidic solution and loaded on a Stage-Tip. Peptides were eluted at pH 8, 6, 5, 4 and 3 also acidified with acidic solution and desalted using Stage-Tips.

Nano-LC-MS/MS analysis

All peptide fractions were measured on an EASY-nLC II nano-LC (Proxeon Biosystems) coupled to an Orbitrap Velos mass spectrometer (Thermo Fisher Scientific). Chromatographic separation was done on a 15 cm PicoTip fused silica emitter with an inner diameter (ID) of 75 µm and an 8 µm Tip ID (New Objective) packed in-house with reversed-phase ReproSil-Pur C18-AQ 3 µm resin (Dr. Maisch GmbH). Peptides were injected into the column with solvent A (0.5% acetic acid) at 700 nL/min using a maximum pressure of 280 Bar. Peptides were then eluted using an 81 min or a 221 min segmented gradient of 5-50% solvent B (80% ACN in 0.5% acetic acid) at a flow rate of 200 nL/min. The mass spectrometer was operated on a data-dependent mode. Survey full-scans for the MS spectra were recorded between 300 – 2000 Thompson at a resolution of 60,000 with a target value of 1E6 charges. The 15 most intense peaks from the survey scans were selected for fragmentation with collision induced dissociation (CID) at a target value of 5000 charges. The fragment spectra were recorded in the linear ion trap. Selected masses were included in a dynamic exclusion list for 90 seconds.

MS data processing

Acquired MS data were preprocessed by MaxQuant (v.1.2.2.9) (40) in order to generate peak lists that can be submitted to database search. Derived peak lists were submitted to

Andromeda (41) and Mascot v2.2.0 (Matrix Science, UK) search engines to query the genome database translated into all six reading frames. The genome sequence of *E. coli* (42, 43) was downloaded from the NCBI homepage (accession number NC_000913.2). The translation into all six reading frames was done from stop codon to stop codon by applying the bacterial and plant plasmid code (translation table 11) using the transeq tool that is part of the Emboss software package (44). We required a minimal length of six amino acids for each resulting putative open reading frame (ORF) which corresponds to the minimal peptide length that we required in the database search. To that database we added decoy sequences using the SequenceReverse.exe tool shipped with MaxQuant software. The resulting database consisted of 263,159 putative open reading frames, 248 commonly observed lab contaminants and 263,407 reversed sequences.

Database search was performed using the following parameters: precursor mass tolerance was set to 6 and 7ppm for Andromeda and Mascot database search, respectively. The fragment ion mass tolerance was set to 0.5 Da for both search engines. Full enzyme specificity for trypsin and Lys-C was required and up to two missed cleavages were allowed. Oxidation of methionine and protein N-terminal acetylation were defined as variable modifications; carbamidomethylation of cysteine was defined as fixed modification.

The resulting lists of peptide spectrum matches (PSMs) were further processed by MaxQuant and Trans Proteomic Pipeline (v4.5 RAPTURE rev 0) (45). Andromeda database scores calculated by MaxQuant were converted to posterior error probabilities (PEP) as described in (41). We calculated q-values based on PEPs to estimate false discovery rates. Mascot result files (.dat) files were converted to mzML format and further processed by PeptideProphet (45) module as part of the Trans-Proteomic Pipeline. We used the accurate mass binning option, excluded singly charged peptides, and used decoy hits to model the score distribution of false positives for semi-supervised mixture modeling. The FDR was controlled by filtering PSMs according to the probability assigned by PeptideProphet. The corresponding probability threshold was calculated by the 'calctppstat.pl' perl script as part of the TPP and the 'Approx. P threshold for FDR' was used to filter the list of PSMs.

Acquired MS data were additionally searched against a recent annotation of the *E. coli* genome (UniProt reference proteome set; downloaded on 18 January 2012; 4309 protein entries) using MaxQuant v1.2.2.9 operating the same database search parameters as described above. False discovery rates on peptide and protein group level were set to 1%, respectively.

Proteogenomic workflow

Detected peptide sequences that resulted from searching the six-frame database were matched to the proteome and the genome database using BLASTP and TBLASTN, respectively (Blast 2.2.25+) (46, 47). For BLAST searches of typically short peptide sequences against the genome and proteome database, we set the maximal E-value to 10000 and the number of alignments to 20 in order to ensure that the peptides can be found in the genome and proteome databases. To map these peptides unambiguously, we required a full length alignment and 100% similarity. Multiple occurrences of the same peptide in the genome or proteome were considered separately. All peptides that did not meet these criteria were defined as initial candidate list of novel peptides. To address the ambiguity of leucine and isoleucine, the initial set of novel peptides was checked once again using regular expression matching. Peptide sequences that could not be found in the proteome database because of any isobaric amino acids were removed from the initial set of novel peptides. In a second BLAST iteration all six-frame ORFs that were detected by one or more novel peptides were matched to the proteome database as well as the non-redundant protein database (NCBI nr) database. In addition we re-submitted the spectra of novel peptides to query NCBI nr database using Mascot search engine, to check consistency between PSMs derived from searching the six-frame translation and NCBI nr database. Together with the genome coordinates of the peptides and the annotated proteins, we used this information to classify the novel peptides into different types of annotation conflicts.

The proteogenomic pipeline and further down-stream data analysis was implemented in R v2.13 (48).

Results

We derived a comprehensive dataset of *E. coli* proteins by harvesting the cells in the exponential phase of growth, extracting the proteome and applying three separation methods (strong anion exchange chromatography, OffGel isoelectric focusing and GeLC-MS) in combination with protein digestion using two proteases, trypsin and Lys-C. We analyzed the resulting peptide mixtures by nano-LC-MS on an LTQ Orbitrap Velos mass spectrometer. We measured the precursor (peptide) ion masses at high resolution and mass accuracy in the Orbitrap analyzer, while performing peptide fragmentation and fragment ions measurement at low resolution in the linear ion trap analyzer. In total, we acquired 1,941,724 mass spectra in about 6 days of measurement time. The average absolute mass accuracy of the identified PSMs was 0.34 ppm and 99% of the PSMs were measured within 1.8 ppm, which enabled us to use narrow (up to 7 ppm) precursor mass tolerance windows during database search. We mapped these spectra onto the six-frame translation of the raw genome sequence in order to assess sensitivity, specificity, accuracy, and the actual false discovery rate in a typical bacterial proteogenomic experiment (**Supplemental Figure 1**). Separately, we mapped the spectra to the annotated genome sequence (UniProt reference *E. coli* proteome database) to assess genome coverage by detected peptide sequences.

Assigning Statistical Confidence to Six Frame Database Search Results

The translation of the *E. coli* genome sequence from stop codon to stop codon resulted in 263,159 putative ORFs, which were generally short database entries with a median length of 20 amino acids. Most of these ORFs represent spurious sequences since usually only one reading frame at a given locus is transcribed; this means that on average five out of six sequences are artificial database entries. To increase confidence in the interpretation of proteogenomic data analysis we used two common workflows for processing and statistical assessment of MS/MS data: MaxQuant (40), based on Andromeda search engine (41) and target-decoy approach (TDA) for FDR estimation (49, 50); and Trans-Proteomic Pipeline (51), used with Mascot search engine (Matrix Science, UK) and mixture model approach (MMA) for FDR estimation (45). Searching the acquired MS data against the six-frame database using Mascot and Andromeda search engines and controlling the false discovery rate at 1 percent yielded markedly different numbers of identified MS/MS spectra and peptide sequences (**Table 1**). The application of Mascot search engine in combination with the MMA identified almost 24% fewer peptide sequences and 48% fewer MS/MS spectra at the same FDR compared to Andromeda search engine in combination with the TDA. This was not surprising, as the more conservative character of the MMA to control FDR was reported previously (35, 52).

False Positive Identifications Accumulate among Novel Peptide Hits

We investigated whether identified peptide sequences are present in the annotated, protein coding portion of the genome. Peptides that could not be assigned to any annotated protein in the UniProt *E. coli* database, we refer to as six frame-specific or novel peptides. Assuming a complete and correct annotation of the *E. coli* genome, these peptide hits are false positives and can be used to assess the performance of the applied proteogenomic search strategies. In order to validate this hypothesis we processed the MS data without any control of the FDR and classified the resulting peptide sequences according to whether they are annotated ("target", 44,872 hits), reversed hits ("decoy", 35,370 hits) or novel peptides ("novel", 31,075 hits) (**Figure 1A**). Interestingly, the absolute number of decoy and novel hits was very similar and the corresponding PEP values followed almost the same distribution with median PEP values of 0.79 (decoy) and 0.787 (novel), respectively **(Figure 1B)**. Conversely, the PEP values of peptides that could be assigned to annotated proteins followed a very tight distribution around a median PEP of 3.26e-6 indicating that a high percentage was true positive identifications. Application of the TPP workflow confirmed these results (**Supplemental Figure 2**). Taken together, these findings support the initial assumption of a complete genome annotation of *E. coli* and point to the fact that majority of novel peptides are in fact false positive hits. Therefore, these peptides can be used as decoy hits to estimate a false discovery rate according to the target-decoy approach. We calculated q-values (53) for all peptides that could be assigned to the existing protein database using the novel peptides as decoy hits and correlated the calculated values to 'standard' q-values (**Figure 1C**). Overall there was a very high correlation of q-values calculated based on decoy peptides (x-axis) and novel peptides (y-axis) pointing to a substantial bias of FDR assessment using the target-decoy approach even in simple proteogenomic experiments.

Assessment of Proteogenomic Workflows

The assumption of complete annotation of the *E. coli* genome enabled the calculation of various features of the applied proteogenomic pipeline, such as sensitivity (SENS), specificity (SPC), accuracy (ACC) and actual false discovery rate (FDR_{act}). The general strategy to calculate these values is depicted in **Supplemental figure 3A**. An experimental outcome, in our case the result of the proteogenomic workflow, is compared to a 'golden standard', which in our case is the annotated, protein-coding part of the *E. coli* genome. We classified all peptide sequences returned by MaxQuant and Trans-Proteomic Pipeline into the four possible contingencies (true positive, false positive, false negative, true negative) of this comparison (**Supplemental figure 3B**). Based on the derived contingency tables we assessed sensitivity, specificity, accuracy and the actual false discovery rate as a function of the FDR utilized by both approaches (**Figure 2**). To assess FDR_{act} we used the number of false positive (FP) identifications to estimate the expected number of FPs among the list of all detected peptide sequences, equivalent to the target-decoy approach. Both workflows demonstrated high specificity and accuracy which are essential to discriminate false positive from true positive identifications. The sensitivity of the TPP based workflow was consistently lower, on average 42.3%, compared to the MQ workflow across different FDR thresholds. Strikingly, the actual false discovery rate as a function of the decoy FDR utilized by MaxQuant increased linearly, with a constant ratio FDR_{act}/FDR_{decoy} of about 3.5 in our particular study, whereas the FDR_{act} did not approach the probability based FDR used in the TPP workflow confirming the conservative character of MMA. We expect that these features of MMA and TDA will be applicable to other proteogenomics datasets of similar sizes and complexities.

Novel Peptides and Potential Annotation Conflicts

In total, 313 peptide sequences passing the default constraint of 1% FDR were specifically found in the genomic six-frame translation and are not annotated according to UniProt *E. coli* database. Of all peptide sequences 68.1% were identified by both workflows whereas only nine peptides (2.8% of the total novel peptides) were identified as novel in both datasets (**Supplemental figure 4A,B**). The poor overlap of detected novel peptides further pointed to their stochastic distribution in the two datasets, and thus to an increased likelihood of false positive identifications among them.

We next focused on the nine novel peptides identified by both data processing workflows. The corresponding PEPs of these peptides were noticeably better compared to other novel peptides (**Supplemental figure 5A**) and therefore had the highest likelihood of being correctly identified. Manual inspection of the corresponding MS/MS spectra validated eight of the nine novel peptides which we classified into potential annotation conflicts (**Table 2**). The fact that most of the best-scoring novel peptides were known annotation conflicts and therefore true positive hits pointed to the fact that our calculations are conservative in nature (represent the “worst scenario”). Their presence also points to substantial number of annotation conflicts even in the simplest genomes. The presence of at least one obvious false positive even in this “golden” set of novel peptides indicated increased FDR in this part of the dataset. Examples of a novel peptide resulting from known annotation conflict, as well as a novel peptide resulting from false identification are presented in the **Figure 3**. All nine novel peptide sequences, together with their annotation details are presented in the **Supplemental Figure 6 and Supplemental table 3.**

The Expressed Proteome of *E. coli* in Exponential Growth Phase

The dataset derived in this study represents one of the most comprehensive proteomics datasets of *E. coli*. In order to assess proteome coverage, we searched the acquired MS spectra against the UniProt proteome database using MaxQuant operating with default parameters. Resubmission of the 1.9M spectra to the Andromeda search engine identified 42,780 non-redundant peptide sequences (**Supplemental table 4**) corresponding to 2,626 distinct *E. coli* proteins (**Supplemental table 5/6**) with an FDR of 1% at protein level. A detailed summary of all sub-datasets concerning the different fractionation methods as well as the two enzymes used can be found in **Supplemental table 7**. Although 2,626 proteins represent about 61% of the annotated proteome, a dataset of similar size was reported before and comparison to transcriptome showed that about 2,600 *E. coli* genes are expected to be expressed during exponential growth in culture (36). In that study, combined proteome and transcriptome (microarray) analysis detected 2,602 and 2,543 *E. coli* gene products, respectively. We therefore conclude that our dataset approached full coverage of the *E. coli* proteome expressed at the point of culture harvesting.

This comprehensive dataset enabled us to address general features of bacterial proteomics experiments, especially in the context of the coverage of the genome sequence by detected peptides. We first defined the protein-coding part of the genome by mapping the 4,309 proteins present in the UniProt *E. coli* database onto the chromosome (**Figure 4A**). This analysis revealed that 86.8% of the genome is annotated in the protein database (4.0 Mb). We next used sequences of all proteins identified in our dataset to define the expressed part of the genome (3.0 Mb), which corresponded to 65.4% of protein-coding genome regions. Finally, mapping of the detected peptide sequences onto the chromosome captured 1.27 Mb of the raw genome sequence, matching 48.5% of the expressed part of the genome (**Figure 4B**). The number of MS/MS events with which each nucleotide is represented, ranged from 1 to 1344 with an average coverage of 20 MS/MS and median number of 7 MS/MS events per nucleotide (**Figure 4C**). However, despite this relatively high average coverage of the expressed genome, the coverage of the total genome sequence was only about 27.5%. This limited sequence coverage is a major limitation in proteogenomics experiments and points to the need for highly sensitive MS and proteogenomic workflows in experiments that aim at genome reannotation.

Discussion

Performance of different search strategies in proteogenomic applications was subject to a number of previous studies. For example, the application of TDA to searches of protein databases derived by six-frame translation has been assessed recently (35) and previous reports pointed to important general considerations for application of this approach in database search (54, 55). There is a global consensus that the increased size of the databases obtained by six-frame translation decreases the sensitivity and specificity of database search and that the spurious protein sequences present in such databases make the statistical confidence of the resulting peptide spectrum matches (PSMs) difficult to assess (13, 56). To circumvent this problem, in this study we assumed complete annotation of the *E. coli* genome which enabled us to call all six frame-specific (novel) PSMs as false positive identifications. This simple assumption was confirmed by almost identical distribution of the PEP values of the novel and decoy hits, as well as the low number of detected novel peptides. We note that the low number of true positive novel hits influences the reported values, but we expect their effect to be minimal. The assumption of full genome annotation is only valid for this system (*E. coli*) and is not applicable to organisms with large and/or partially annotated genomes. However, it proved to be useful to assess general features of a typical bacterial proteogenomic dataset, such as sensitivity, specificity, accuracy and actual false discovery rate.

We used two complementary MS/MS data processing frameworks, MaxQuant implementing the target-decoy approach (TDA) and Trans-Proteomic Pipeline/Peptide Prophet using mixture-model approach for FDR assessment. While both achieved deep proteome coverage, the MaxQuant-based workflow identified significantly higher number of peptides, whereas TPP-based workflow had significantly lower actual FDR. However, TPP also led to a decreased sensitivity, resulting in lower number of identified spectra, which can significantly affect the coverage of the genome sequence by detected peptides (see below). In our view, the

MaxQuant workflow led to a better tradeoff between maximal peptide identification rates (sensitivity) desired in proteogenomic studies at acceptable false positive rate (FP).

Somewhat surprisingly, our data point to substantial underestimation of FDR even in “simple” proteogenomic experiments obtained by high accuracy mass spectrometry. Although several strategies to decrease the search space in proteogenomic databases have been proposed (57) we argue that the use of high accuracy is one of the most effective ways to achieve this. In this context, the use of “high-high” acquisition methods (ones in which the survey and MS/MS scans are acquired at high (ppm to sub-ppm) accuracy), will further improve the confidence of detected PSMs and become indispensable in future proteogenomic experiments. However, novel peptides detected in such experiments should still undergo thorough investigation before treated as true positive identifications, regardless of the acquisition method or proteogenomic workflow used.

The comprehensive proteome dataset derived for the purpose of this study enabled us to assess another important aspect of a proteogenomics experiment: coverage of the genome sequence by identified peptide sequences. The field of proteomics is getting to the remarkable stage of identification and quantification of all gene products expressed under specific conditions, and this especially applies to organisms with small and relatively simple genomes, such as bacteria and yeast (5). In addition to the detection of a gene product, genome reannotation also requires high coverage of the genome sequence. In our study, we achieved a comprehensive detection of the expressed *E. coli* proteome, which was in agreement with previous studies (36); however, the identified peptide sequences covered 48.5% of the estimated expressed, 65.4% of the protein-coding but only 27.5% of the total genome sequence. Since NGS studies routinely achieve up to 50-fold base coverage of 99.9% of genome sequence (58, 59), our results demonstrate the limitation of using mass spectrometry-based proteomics for the sole purpose of genome annotation. Despite of the constant improvements in mass spectrometry technology it is hard to see how the genome sequence coverage by detected peptides will be improved to the level achieved by the NGS technology. Therefore, we believe that the major impact of proteogenomics will not be in genome reannotation, but in analysis of features that are beyond the reach of genomics, such as posttranslational modifications of proteins in the context of individualized protein databases derived by NGS. However, the routine application of proteomics in these areas will require further substantial improvements aimed at increasing the sequencing speed/coverage (MS level) and specificity/sensitivity (bioinformatic workflows).

References

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198-207
2. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol Cell Proteomics* 10, M111 011015
3. Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Muller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., Dunyach, J. J., Cox, J., Horning, S., Mann, M., and Makarov, A. (2012) Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol Cell Proteomics* 11, O111 013698
4. Olsen, J. V., Schwartz, J. C., Griep-Raming, J., Nielsen, M. L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., Wouters, E. R., Senko, M., Makarov, A., Mann, M., and Horning, S. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics* 8, 2759-2769
5. de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Frohlich, F., Walther, T. C., and Mann, M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 455, 1251-1254
6. Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B., and Aebersold, R. (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* 138, 795-806

7. Frishman, D., and Valencia, A. (2009) *Modern genome annotation : the BioSapiens Network*, Springer, New York

8. Brent, M. R. (2005) Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Research* 15, 1777-1786

9. Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24, 637-644

10. Kuster, B., Mortensen, P., Andersen, J. S., and Mann, M. (2001) Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* 1, 641-650

11. Armengaud, J. (2010) Proteogenomics and systems biology: quest for the ultimate missing parts. *Expert Rev Proteomic* 7, 65-77

12. Yates, J. R., 3rd, Eng, J. K., and McCormack, A. L. (1995) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem* 67, 3202-3210

13. Castellana, N., and Bafna, V. (2010) Proteogenomics to discover the full coding content of genomes: a computational perspective. *J Proteomics* 73, 2124-2135

14. Tanner, S., Shen, Z. X., Ng, J., Florea, L., Guigo, R., Briggs, S. P., and Bafna, V. (2007) Improving gene annotation using peptide mass spectrometry. *Genome Research* 17, 231-239

15. Merrihew, G. E., Davis, C., Ewing, B., Williams, G., Kall, L., Frewen, B. E., Noble, W. S., Green, P., Thomas, J. H., and MacCoss, M. J. (2008) Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. *Genome Res* 18, 1660-1669

16. Borchert, N., Dieterich, C., Krug, K., Schutz, W., Jung, S., Nordheim, A., Sommer, R. J., and Macek, B. (2010) Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. *Genome Res* 20, 837-846

17. Oshiro, G., Wodicka, L. M., Washburn, M. P., Yates, J. R., Lockhart, D. J., and Winzeler, E. A. (2002) Parallel identification of new genes in Saccharomyces cerevisiae. *Genome Research* 12, 1210-1220

18. Bitton, D. A., Wood, V., Scutt, P. J., Grallert, A., Yates, T., Smith, D. L., Hagan, I. M., and Miller, C. J. (2011) Augmented annotation of the Schizosaccharomyces pombe genome reveals additional genes required for growth and viability. *Genetics* 187, 1207-1217

19. Castellana, N. E., Payne, S. H., Shen, Z. X., Stanke, M., Bafna, V., and Briggs, S. P. (2008) Discovery and revision of Arabidopsis genes by proteogenomics. *P Natl Acad Sci USA* 105, 21034-21038

20. Bringans, S., Hane, J. K., Casey, T., Tan, K. C., Lipscombe, R., Solomon, P. S., and Oliver, R. P. (2009) Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen Stagonospora nodorum. *BMC Bioinformatics* 10, 301

21. Xia, D., Sanderson, S. J., Jones, A. R., Prieto, J. H., Yates, J. R., Bromley, E., Tomley, F. M., Lal, K., Sinden, R. E., Brunk, B. P., Roos, D. S., and Wastling, J. M. (2008) The proteome of Toxoplasma gondii: integration with the genome provides novel insights into gene expression and annotation. *Genome Biol* 9, R116

22. Kalume, D. E., Peri, S., Reddy, R., Zhong, J., Okulate, M., Kumar, N., and Pandey, A. (2005) Genome annotation of Anopheles gambiae using mass spectrometry-derived data. *BMC Genomics* 6, 128

23. Brosch, M., Saunders, G. I., Frankish, A., Collins, M. O., Yu, L., Wright, J., Verstraten, R., Adams, D. J., Harrow, J., Choudhary, J. S., and Hubbard, T. (2011) Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res* 21, 756-767

24. Bitton, D. A., Smith, D. L., Connolly, Y., Scutt, P. J., and Miller, C. J. (2010) An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. *PLoS One* 5, e8949

25. Fermin, D., Allen, B. B., Blackwell, T. W., Menon, R., Adamski, M., Xu, Y., Ulintz, P., Omenn, G. S., and States, D. J. (2006) Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol* 7, R35
26. Armengaud, J. (2012) Microbiology and proteomics, getting the best of both worlds! *Environ Microbiol*
27. Armengaud, J. (2009) A perfect genome annotation is within reach with the proteomics and genomics alliance. *Curr Opin Microbiol* 12, 292-300
28. Chen, W. B., Laidig, K. E., Park, Y., Park, K., Yates, J. R., Lamont, R. J., and Hackett, M. (2001) Searching the Porphyromonas gingivalis genome with peptide fragmentation mass spectra. *Analyst* 126, 52-57
29. Wang, R., Prince, J. T., and Marcotte, E. M. (2005) Mass spectrometry of the M. smegmatis proteome: protein expression levels correlate with function, operons, and codon bias. *Genome Res* 15, 1118-1126
30. de Souza, G. A., Malen, H., Softeland, T., Saelensminde, G., Prasad, S., Jonassen, I., and Wiker, H. G. (2008) High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using Mycobacterium tuberculosis as an example. *BMC Genomics* 9, 316
31. de Souza, G. A., Softeland, T., Koehler, C. J., Thiede, B., and Wiker, H. G. (2009) Validating divergent ORF annotation of the Mycobacterium leprae genome through a full translation data set and peptide identification by tandem mass spectrometry. *Proteomics* 9, 3233-3243
32. Kelkar, D. S., Kumar, D., Kumar, P., Balakrishnan, L., Muthusamy, B., Yadav, A. K., Shrivastava, P., Marimuthu, A., Anand, S., Sundaram, H., Kingsbury, R., Harsha, H. C., Nair, B., Prasad, T. S., Chauhan, D. S., Katoch, K., Katoch, V. M., Chaerkady, R., Ramachandran, S., Dash, D., and Pandey, A. (2011) Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. *Mol Cell Proteomics* 10, M111 011627
33. Venter, E., Smith, R. D., and Payne, S. H. (2011) Proteogenomic Analysis of Bacteria and Archaea: A 46 Organism Case Study. *PLoS One* 6
34. Krug, K., Nahnsen, S., and Macek, B. (2011) Mass spectrometry at the interface of proteomics and genomics. *Mol Biosyst* 7, 284-291
35. Blakeley, P., Overton, I. M., and Hubbard, S. J. (2012) Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J Proteome Res* 11, 5221-5234
36. Iwasaki, M., Miwa, S., Ikegami, T., Tomita, M., Tanaka, N., and 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
37. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., and Mori, H. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2, 2006 0008
38. Ishihama, Y., Rappsilber, J., and Mann, M. (2006) Modular stop and go extraction tips with stacked disks for parallel and multidimensional Peptide fractionation in proteomics. *J Proteome Res* 5, 988-994
39. Wisniewski, J. R., Zougman, A., and Mann, M. (2009) Combination of FASP and StageTip-Based Fractionation Allows In-Depth Analysis of the Hippocampal Membrane Proteome. *Journal of Proteome Research* 8, 5674-5678
40. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26, 1367-1372
41. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 10, 1794-1805

42. Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K., Blattner, F. R., Chaudhuri, R. R., Glasner, J. D., Horiuchi, T., Keseler, I. M., Kosuge, T., Mori, H., Perna, N. T., Plunkett, G., 3rd, Rudd, K. E., Serres, M. H., Thomas, G. H., Thomson, N. R., Wishart, D., and Wanner, B. L. (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res* 34, 1-9

43. Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., Ohtsubo, E., Baba, T., Wanner, B. L., Mori, H., and Horiuchi, T. (2006) Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Mol Syst Biol* 2, 2006 0007

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

45. Keller, A., Nesvizhskii, A. I., Kolker, E., and 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

46. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* 215, 403-410

47. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421

48. Team, R. C. (2012) R: A Language and Environment for Statistical Computing.

49. Elias, J. E., and Gygi, S. P. (2010) Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol* 604, 55-71

50. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 4, 207-214

51. Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. I., and Aebersold, R. (2010) A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 10, 1150-1159

52. Kall, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008) Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res* 7, 40-44

53. Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genomewide studies. *P Natl Acad Sci USA* 100, 9440-9445

54. Gupta, N., Bandeira, N., Keich, U., and Pevzner, P. A. (2011) Target-Decoy Approach and False Discovery Rate: When Things May Go Wrong. *J Am Soc Mass Spectr* 22, 1111-1120

55. Cooper, B. (2012) The Problem with Peptide Presumption and the Downfall of Target-Decoy False Discovery Rates. *Anal Chem*

56. Nesvizhskii, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* 73, 2092-2123

57. Helmy, M., Tomita, M., and Ishihama, Y. (2012) Peptide Identification by Searching Large-Scale Tandem Mass Spectra against Large Databases: Bioinformatics Methods in Proteogenomics. *Genes, Genomes and Genomics* 6, 76-85

58. Srivatsan, A., Han, Y., Peng, J. L., Tehranchi, A. K., Gibbs, R., Wang, J. D., and Chen, R. (2008) High-Precision, Whole-Genome Sequencing of Laboratory Strains Facilitates Genetic Studies. *Plos Genet* 4

59. Metzker, M. L. (2010) Next Generation Technologies: Basics and Applications. *Environ Mol Mutagen* 51, 691-691

Tables

Table 1: Identified MS/MS spectra and peptide sequences after searching a six frame translation of *E. coli* chromosome using two MS data processing workflows.

	MS/MS spectra identified	**MS/MS spectra identified (%)**	**Peptide sequences**	**Novel peptides**	**Decoy peptides**	**Lab contaminant peptides**	**Annotated *E. coli* proteins**
MQ workflow	370,231	19,1	33,964	263	336	306	2,653
TPP workflow	162,028	8.3	25,724	59	0	209	2,524

Table 2: Novel peptide sequences identified by both data processing workflows. Accession numbers shown in brackets correspond to proteins of another organism or *E. coli* strains.

No	**Peptide sequence**	**Annotation conflict in *E. coli* K12**	**Remarks**	**UniProt Accession(s)**
A	VGSESWWQSK	Erroneous initiation (upstream peptide)	known conflict	P13039
B	INQTSAMPEK	Erroneous initiation (spanning peptide)	known conflict	P32695
C	LAMPSGNQEPR	Erroneous initiation (spanning peptide)	Correct in *E. coli* O157:H7	P0CB62 (Q8X3T3)
D	MMQTVLAK	Erroneous initiation (spanning peptide)	known conflict (in *E. coli* O157:H7**)**	P00909 (Q8X7B7)
E	CSEFGEAIIENM	Point mutation/sequencing error or deamidated version	Present in other E. coli strains, e.g. MS 116-1, DH1	P13039 (D8AHK0) (E6P3Y4)
F	GVALHAVK	Not present	Present in *E .coli* strain MS 117-3	(E9THR2)
G	SLYSIALIR	Not present	78% similarity to a sequence in *Selaginella moellendorffi*	(D8RY37)
H	GLSGPASQATVAAP	Not present	Unclear	
I	LSIRIQPPK	Not present	Unclear/FP	

Data Access

All raw files, peak lists, fasta databases and result tables returned by MQ and TPP were submitted to PeptideAtlas data repository. The data can be accessed by following ID and password:

ID: PASS00147

Password: GW584mr

Acknowledgements

We thank Dr. Boumediene Soufi for comments regarding the experimental setup and manuscript. This work was supported by the Juniorprofessoren-Programm of the BW Stiftung, SFB766 and PRIME-XS.

Author Contributions

KK and BM designed the experiments; KK performed the bioinformatic analysis; AC, GB, KM and NCS performed sample preparation and MS measurements; KK and BM wrote the manuscript.

Disclosure Declaration

Authors declare no conflict of interest.

Figure 1

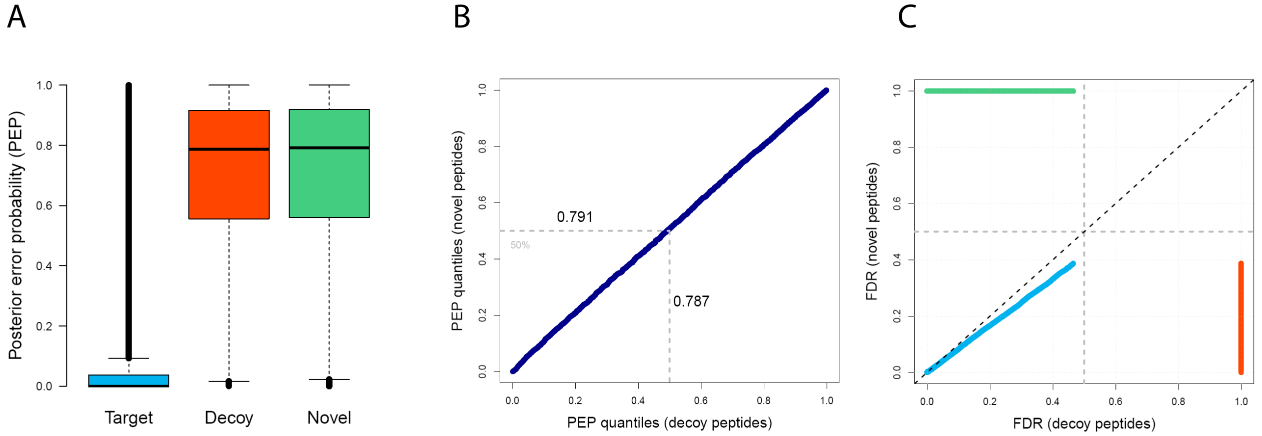


Figure 1: Distribution of PSM confidence scores
A) Distributions of posterior error probabilities (PEP) of different PSM populations that result from searching a genomic six-frame translation in a target-decoy database design. The PEPs of novel and decoy peptides are distinctly different distributed than the PEPs of all target peptides, that are contained in the UniProt proteome database of *E. coli*. **B)** Quantile-Quantile (QQ)-plot of PEPs resulting from decoy (x-axis) and novel peptides (y-axis) demonstrating an almost identical PEP distribution of novel and decoy peptides. **C)** Correlation of estimated false discovery rates (q-values) derived from decoy (x-axis) and novel peptide sequences (y-axis). Colors correspond to PSM populations shown in A).

Figure 2

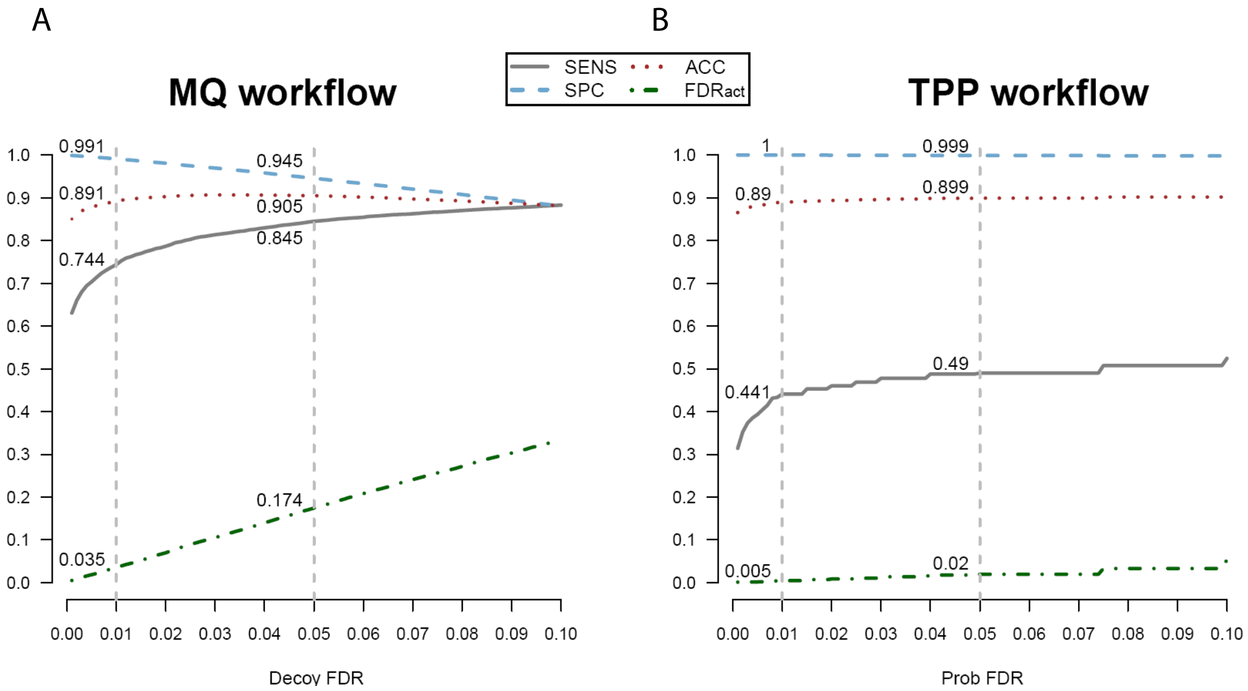


Figure 2: Assessing proteogenomic data processing workflows
The assumption of a complete annotation of the *E. coli* genome enabled the assessment of sensitivity (SENS), specificity (SPC), accuracy (ACC), and actual false discovery rate (FDR_{act}) as a function of the false discovery rate (FDR). Emphasized values correspond to commonly used FDR thresholds of 1% and 5%. **A)** MaxQuant (MQ) workflow utilizing decoy FDR approach demonstrated the best tradeof of high sensitivity, specificity and accuracy at a decoy FDR of 1%. The actual FDR increased linearly to the decoy FDR by a constant factor of 3.5.**B)** Trans-Proteomic Pipeline (TPP) workflow demonstrated high and constant specificity and accuracy across the probability based FDR, at much lower level of sensitivity and actual FDR compared to MQ based workflow.

Figure 3

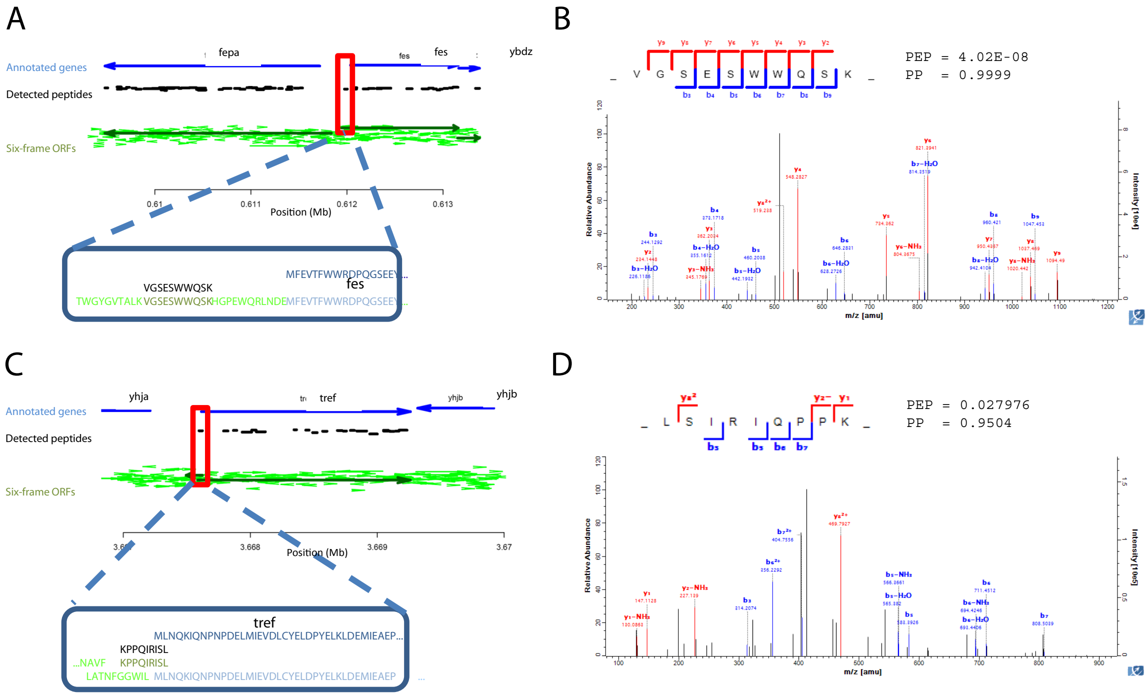


Figure 3: True positive and likely false positive example among highest scoring novel peptides
A) Schematic representation of the erroneous initiation of the *fes* gene. Annotated proteins are shown in blue, detected peptides are depicted in black, and six frame ORFs are shown in green. ORFs that were hit by peptides are shown in dark green. The novel peptide (VGSESWWQSK) was located upstream of the predicted protein N-terminus. The corresponding six frame ORF encompassed the complete sequence and employed the same reading frame as the *fes* gene. **B)** Corresponding MS/MS spectrum of the novel peptide depicted in A) annotated with a comprehensive series of b and y fragment ions. **C)** Schematic representation of a dubious novel peptide identified at 1% FDR by both data processing workflows used in this study. Although an adjacent cluster of peptides was detected that mapped to the *tref* gene, the novel peptide (LSIRIQPPK) utilized a different reading frame. **D)** MS/MS spectrum of the corresponding novel peptide shown in C) poorly annotated with b and y fragment ions.

Figure 4

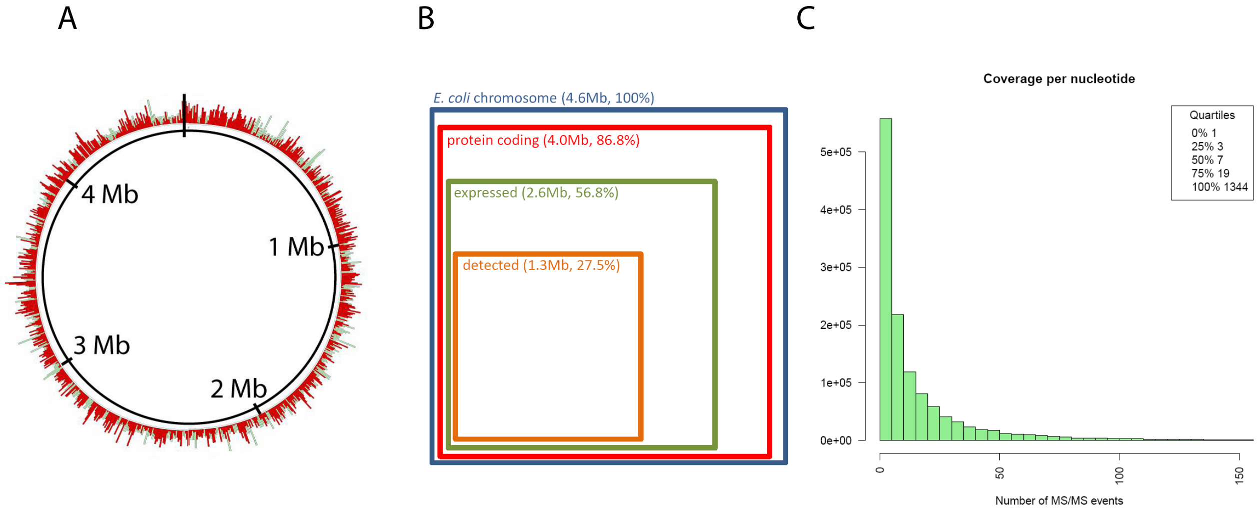


Figure 4: Achieved coverage of the *E. coli* genome by MS data
A) Schematic representation of achieved coverage of the *E. coli* chromosome by MS data derived in this study. Detected peptides are shown in red; protein sequences according to UniProt annotation are depicted in green. **B)** Venn diagram illustrating the coverage of detected peptide sequences to several layers of the *E. coli* genome. **C)** Histogram depicting the number of MS/MS scans per nucleotide that was detected in this study.

Downloaded from genome.cshlp.org on January 7, 2013 - Published by Cold Spring Harbor Laboratory Press

Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models

Nadine Borchert,[1,5] Christoph Dieterich,[1,2,5] Karsten Krug,[3,5] Wolfgang Schütz,[3]
Stephan Jung,[3] Alfred Nordheim,[4] Ralf J. Sommer,[1,6] and Boris Macek[3,6]

[1]Department for Evolutionary Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany; [2]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, 13125 Berlin, Germany; [3]Proteome Center Tübingen, University of Tübingen, 72076 Tübingen, Germany; [4]Department of Molecular Biology, University of Tübingen, 72076 Tübingen, Germany

Pristionchus pacificus is a nematode model organism whose genome has recently been sequenced. To refine the genome annotation we performed transcriptome and proteome analysis and gathered comprehensive experimental information on gene expression. Transcriptome analysis on a 454 Life Sciences (Roche) FLX platform generated >700,000 expressed sequence tags (ESTs) from two normalized EST libraries, whereas proteome analysis on an LTQ-Orbitrap mass spectrometer detected >27,000 nonredundant peptide sequences from more than 4000 proteins at sub-parts-per-million (ppm) mass accuracy and a false discovery rate of <1%. Retraining of the SNAP gene prediction algorithm using the gene expression data led to a decrease in the number of previously predicted protein-coding genes from 29,000 to 24,000 and refinement of numerous gene models. The *P. pacificus* proteome contains a high proportion of small proteins with no known homologs in other species ("pioneer" proteins). Some of these proteins appear to be products of highly homologous genes, pointing to their common origin. We show that >50% of all pioneer genes are transcribed under standard culture conditions and that pioneer proteins significantly contribute to a unimodal distribution of predicted protein sizes in *P. pacificus*, which has an unusually low median size of 240 amino acids (26.8 kDa). In contrast, the predicted proteome of *Caenorhabditis elegans* follows a distinct bimodal protein size distribution, with significant functional differences between small and large protein populations. Combined, these results provide the first catalog of the expressed genome of *P. pacificus*, refinement of its genome annotation, and the first comparison of related nematode models at the proteome level.

[Supplemental material is available online at http://www.genome.org. The 454 Life Sciences (Roche) sequencing data from this study have been submitted to the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession no. SRA010772. Sequences from targeted RT-PCR reactions have been submitted to GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) (accession numbers provided in Supplemental Tables 4 and 6). Mass spectrometry data have been uploaded to the Proteome Commons Tranche repository (https://proteomecommons.org/tranche/).]

Genome sequence data are useful only when genes are correctly annotated and information on their genomic localization, expression, and function is available. Comprehensive annotation of protein-coding genes is largely done in silico and is error-prone, especially when performed without experimental information on gene expression. Recent development of rapid techniques for nucleic acid sequencing has enabled comprehensive detection of transcribed genomic regions and use of this information in genome annotation. Especially, the platforms that implement high-throughput pyrosequencing, such as 454 Life Sciences (Roche) FLX, are powerful tools for genome annotation. This platform produces fewer reads (400 K–500 K) than other next-generation sequencers. However, these reads are on average longer (>200 bp vs. ~50 bp) and are essential for an accurate reconstruction of any metazoan transcriptome. This platform has recently been used in the annotation of eukaryotic and prokaryotic genomes (Shin et al. 2008; Vera et al. 2008).

In addition to the evidence of gene expression at transcription level, mass spectrometry (MS)-based proteomics is increasingly used for experimental identification of translated genomic sequence. In a "proteogenomics" approach (Ansong et al. 2008; Gupta et al. 2008), the complete protein extract of an organism is digested into peptides, which are then mass-measured and fragmented in a mass spectrometer. Mass spectra are typically searched against a database containing a six-frame translation of the raw genome assembly and can therefore identify new, unpredicted open reading frames and refine existing gene models. Pioneered already in 1995 (Yates et al. 1995), proteogenomics has since been used to provide experimental evidence for gene expression in various model organisms, such as *Arabidopsis thaliana* (Baerenfaller et al. 2008), *Plasmodium yoelii yoelii* (Carlton et al. 2002), *Toxoplasma gondii* (Xia et al. 2008), and *Homo sapiens* (Fermin et al. 2006). A recent study of *Caenorhabditis elegans* identified more than 6000 gene products by mass spectrometry and refined many gene models even in this well-studied organism (Merrihew et al. 2008).

Pristionchus pacificus is a nematode that has been established as a model organism in evolutionary developmental biology (Sommer et al. 1996; Hong and Sommer 2006). It shares many advantageous features with *C. elegans*, in that it can be grown easily under laboratory conditions by feeding on *Escherichia coli* OP 50, it has a short generation time (4 d at 20°C), and it is a self-fertilizing

---

[5]These authors contributed equally to this work.
[6]Corresponding authors.

E-mail boris.macek@uni-tuebingen.de; fax 49-7071-295779.
E-mail ralf.sommer@tuebingen.mpg.de; fax 49-7071-601498.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.103119.109. Freely available online through the *Genome Research* Open Access option.

Downloaded from genome.cshlp.org on January 7, 2013 - Published by Cold Spring Harbor Laboratory Press

hermaphrodite, which makes it amenable to forward and reverse genetics. The genome of *P. pacificus* was recently sequenced in a whole-genome shotgun approach with 10-fold coverage. The calculated genome size is 169 megabases (Mb) with a total number of 29,000 predicted protein-coding genes and a minimal gene content of 23,500 genes, as inferred from RT-PCR analyses (Dieterich et al. 2008). Many of these genes share no sequence similarity with already known genes in other nematodes and different phyla ("pioneer" genes). In comparison, the genome of *C. elegans* is completely assembled, consisting of a 100-Mb genome with 20,060 encoding genes (The *C. elegans* Sequencing Consortium 1998; Dieterich and Sommer 2009). *P. pacificus* and *C. elegans* belong to the same phylogenetic clade (Fig. 1A), which provides an ideal evolutionary distance for comparison of their proteome structures.

Here, we perform a comprehensive analysis of the *P. pacificus* transcriptome and proteome using 454 FLX sequencing and LTQ-Orbitrap mass spectrometry, respectively. We search high-accuracy MS data against the predicted proteome and six-frame translation of the raw genomic assembly. We identify more than 700,000 expressed sequence tags (ESTs) and 27,000 nonredundant peptide sequences and use these data to refine the genome annotation and compare the predicted and detected proteome of *P. pacificus* with that of the other nematode models. We show that >50% of all pioneer genes are transcribed and that pioneer proteins significantly contribute to the unimodal distribution of predicted protein sizes in *P. pacificus*. Finally, we observe that the predicted proteome of *C. elegans* follows a distinct bimodal distribution, with significant functional differences between small and large protein populations.

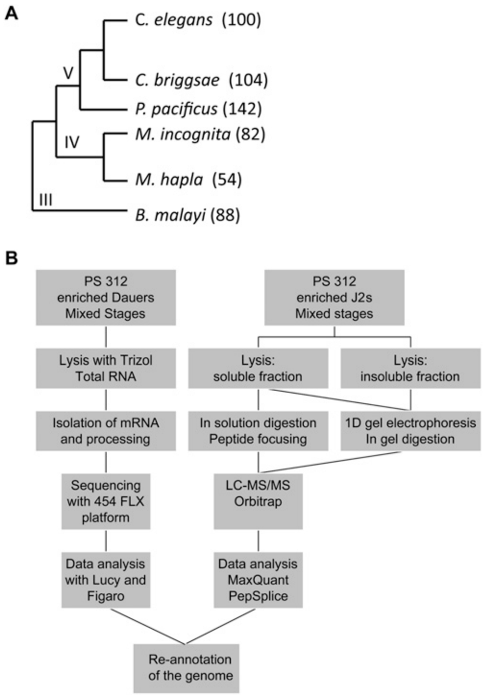

A
C. elegans (100)
C. briggsae (104)
P. pacificus (142)
M. incognita (82)
M. hapla (54)
B. malayi (88)
V
IV
III
B
PS 312 enriched Dauers Mixed Stages
PS 312 enriched J2s Mixed stages
Lysis with Trizol Total RNA
Lysis: soluble fraction
Lysis: insoluble fraction
Isolation of mRNA and processing
In solution digestion Peptide focusing
1D gel electrophoresis In gel digestion
Sequencing with 454 FLX platform
LC-MS/MS Orbitrap
Data analysis with Lucy and Figaro
Data analysis MaxQuant PepSplice
Re-annotation of the genome

Figure 1. Phylogeny of *Pristionchus pacificus* and proteogenomics workflow employed in this study. (*A*) Phylogenetic relationship of nematodes with sequenced genomes. The genome sizes (megabases) are written in brackets. The clades are depicted in the tree. (*B*) *P. pacificus* gene expression was assessed at the levels of transcription and translation and in three different developmental stages: dauer, J2, and "mixed stage" (containing all developmental stages, including eggs). In proteomics approach, several workflows for protein extraction and separation were used. ESTs detected with 454 pyrosequencing and peptide sequences detected with LTQ-Orbitrap mass spectrometry were used for genome reannotation.

Results

The aim of this study was to provide the first experimental catalog of the expressed genome of *P. pacificus* and to use this information for further refinement of the genome annotation. To obtain enhanced coverage of the expressed genome, we used two complementary approaches: transcriptome sequencing and high-accuracy MS proteomics. For the transcriptome analysis, total RNA was isolated from a mixed culture (containing all developmental stages, including eggs) and dauer stage culture of *P. pacificus* and sequenced on the 454 Life Sciences (Roche) FLX pyrosequencing platform.

For the proteome analysis, protein extracts were isolated from a mixed culture and second juvenile (J2) stage culture of *P. pacificus*. In all proteomics experiments, the protein extracts were divided into soluble and insoluble fractions, separated by 1D SDS-PAGE, and in-gel digested by trypsin. To achieve better analytical depth, soluble protein fractions were additionally digested in-solution by trypsin, and the resulting peptide mixtures were separated by isoelectric focusing. All peptide mixtures were subjected to nano-LC-MS/MS analysis on an LTQ-Orbitrap mass spectrometer. The MS data were processed and prepared for database search using the MaxQuant software suite. All MS/MS spectra were searched using the Mascot search engine against a decoy database consisting of the predicted *P. pacificus* proteome (based on old assembly; Dieterich et al. 2008), *E. coli* proteome, common laboratory contaminant proteins, and a six-frame translation of the *P. pacificus* raw genomic assembly. The complete workflow is summarized in Figure 1B.

Gene expression analysis of *P. pacificus*

In the 454 FLX transcriptome measurement, a sequencing run of the normalized mixed stage cDNA library yielded 334,441 ESTs that mapped uniquely to the genome top 965 contigs and had a median read length of 240 bp. In the normalized dauer stage library, 376,796 ESTs mapped to the top 965 contigs and had a median read length of 250 bp. In total, 711,237 ESTs were detected in both analyzed developmental stages.

In the proteome measurement, MS data pre-processing using MaxQuant software resulted in 1,190,811 spectra that were submitted to the Mascot search engine. Database searching led to the identification of 27,561 nonredundant peptide sequences at an estimated false discovery rate (FDR) of 0.2% at the peptide level. Of these, 22,208 were detected in the mixed culture and 17,412 in the J2 stage (Supplemental Table 1). The applied biochemical workflow enabled enhanced proteome coverage, as only 7989 (29%) peptides were identified in all three approaches (Fig. 2A). Robust recalibration algorithms integrated in the MaxQuant software led to the overall average absolute peptide mass deviation of 0.345 parts per million (ppm) with a standard deviation of 0.434 ppm and enabled the use of narrow individualized precursor ion mass tolerances in the database search (Fig. 2B; Cox and Mann 2008) .

Detected peptides were assembled into proteins and protein groups by MaxQuant software (see Methods). Of the detected protein groups, 3451 mapped to the *P. pacificus* predicted proteome (old assembly), 266 to the *E. coli* proteome, 50 to reversed sequences, 30 to contaminants included in the database, and the remainder to the raw genomic translation. The FDR at the protein

Downloaded from genome.cshlp.org on January 7, 2013 - Published by Cold Spring Harbor Laboratory Press

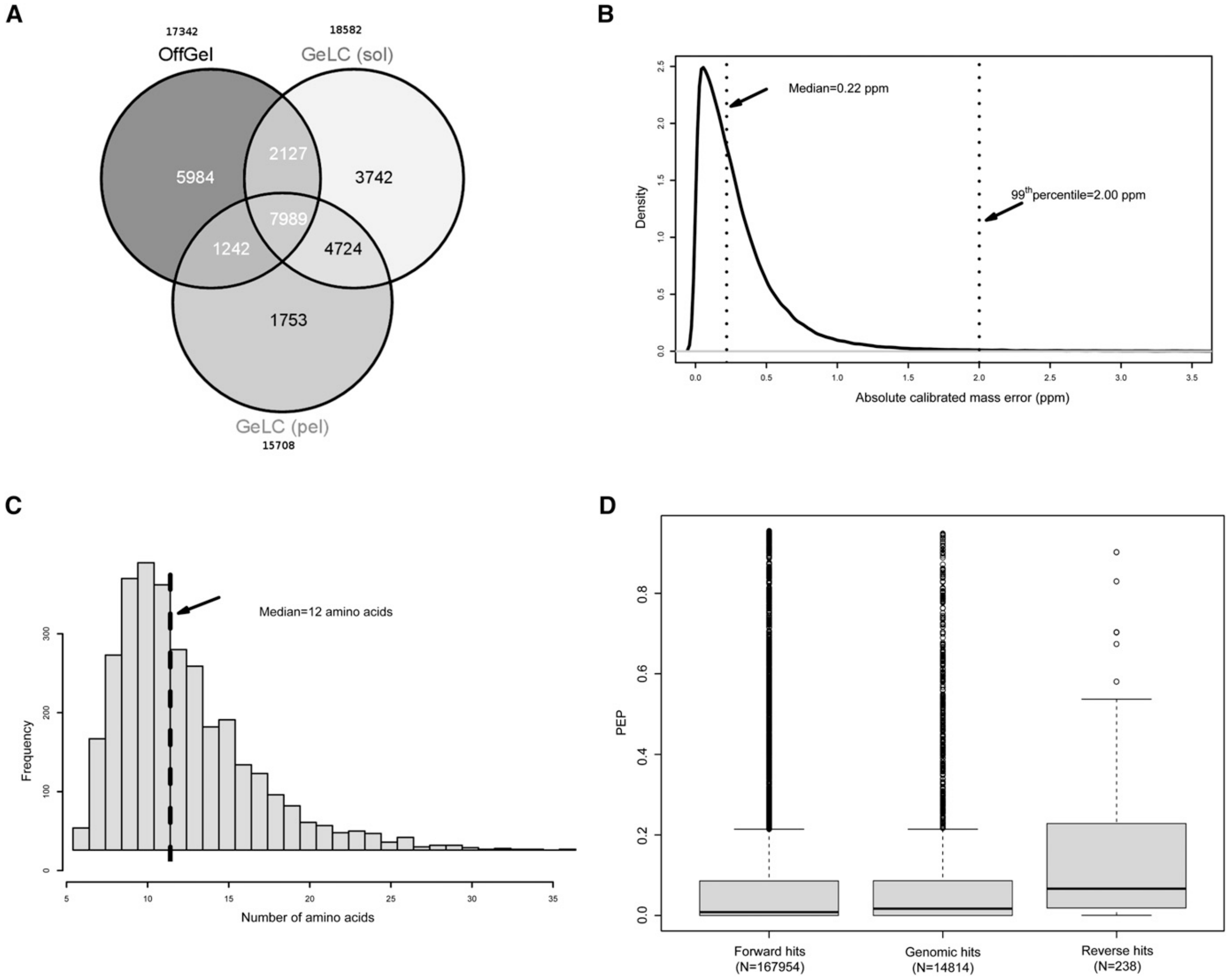


Figure 2. Overview of the proteomics results. (*A*) Application of complementary biochemical workflows for protein extraction and peptide separation led to enhanced proteome coverage. (sol) Soluble fraction; (pel) pellet. (*B*) Peptides were detected with a mean absolute mass deviation of 0.345 ppm. (*C*) Peptide sequences that mapped to the genome translation ("genomic peptides") had a median size of 12 amino acids. (*D*) Distribution of posterior error probabilities (PEP) was markedly different in the genomic and reversed peptide sequences.

group level was 1%. A list of all proteins detected by searching the six-frame translation database is available in Supplemental Table 2.

Refinement of *P. pacificus* gene predictions

We used the transcriptomics and proteomics data to refine the gene predictions in the *P. pacificus* genome sequence. In the transcriptome measurement, of a total of 711,237 detected ESTs, 223,849 ESTs corresponded to genomic regions that were not predicted by the old gene model (Dieterich et al. 2008): 96,754 ESTs in the mixed culture and 127,095 ESTs in the dauer stage. In the proteomics measurement, of 27,561 detected nonredundant peptide sequences, 2783 nonredundant peptides exclusively mapped to the translated genomic sequence, providing direct expression evidence for 1537 genomic regions (contigs and their corresponding reading frames) that were previously not predicted as protein-coding. The median length of genomic peptide hits was 12 amino acids (Fig. 2C), and their posterior error probability (PEP) distribution was distinctly different than that of highest-scoring reverse database hits (Fig. 2D), confirming the high reliability of the data set. To gain additional information on the splice junctions, MS/MS peak lists derived by MaxQuant software were submitted to the PepSplice search engine, which uses raw DNA sequence information to calculate peptides with gaps corresponding to potential GT–AG introns (Roos et al. 2007). The PepSplice database search resulted in identification of 541 spliced peptide sequences (Supplemental Table 3) that enabled identification of exact exon/intron boundaries in corresponding genes.

We used the information on the newly detected loci from both approaches to retrain the SNAP prediction algorithm (Dieterich and Sommer 2009), previously used for genome annotation of *P. pacificus* (Dieterich et al. 2008), and employed it to reannotation of gene predictions on the raw genome assembly. The genome reannotation led to a decrease in the number of predicted protein-coding genes from 29,424 (as reported at http://www.pristionchus.org), to 24,231, mainly through connection of

Downloaded from genome.cshlp.org on January 7, 2013 - Published by Cold Spring Harbor Laboratory Press

neighboring coding regions. Consequently, 3263 old gene models were not contained in the genome reannotation, 1848 new gene models (7736 exons) have been identified, and 11,313 existing gene models were extended (Fig. 3). The new gene prediction is available at http://www.pristionchus.org.

Although our experiments were not designed to perform a direct comparison of the transcriptomics and proteomics, our study provides insights into the main contributions of the two platforms to genome reannotation. Of 1848 new gene models, only 73 were covered by peptides, demonstrating the superior genome coverage of the transcriptomics platform and pointing to the gene model refinement—rather than gene discovery—as the main contribution of the proteomics platform. Indeed, <25% of peptides that mapped to translated genomic sequence were in the intergenic regions of the old gene model, whereas the majority were in the intragenic regions (i.e., in the intron sequences), therefore

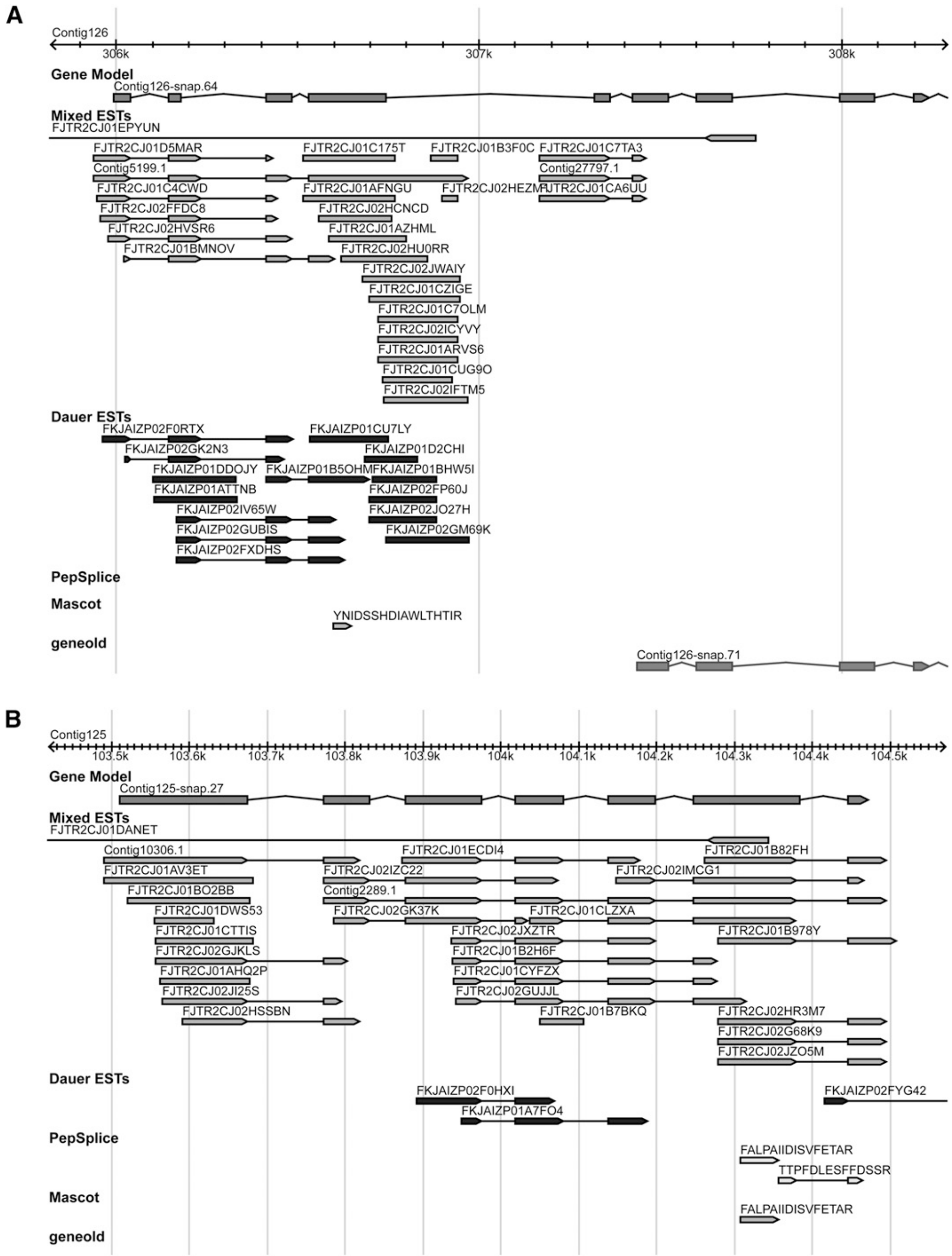


Figure 3. Genome reannotation resulted in new gene predictions and new gene models. (*A*) Example of a new gene model. New gene model "Contig126-snap.64" contains the old model "Contig126-snap.71". (*B*) Example of a new gene prediction. The gene model "Contig125-snap.27" appeared only after retraining of the SNAP prediction algorithm with gene expression data.

Downloaded from genome.cshlp.org on January 7, 2013 - Published by Cold Spring Harbor Laboratory Press

exclusively affecting the existing gene models. In addition, proteomics significantly contributed to determination of exon–exon splice junctions.

For independent confirmation of expression of newly predicted genes, we chose 99 candidates and amplified them with RT-PCR on cDNA from mixed-stage animals. Of analyzed transcripts, 60 could be amplified and sequenced, confirming their expression (Supplemental Table 4).

Catalog of detected *P. pacificus* proteins

The refinement of the *P. pacificus* genome using transcriptomics and proteomics data led to its most comprehensive and accurate annotation to date. To derive a comprehensive catalog of detected *P. pacificus* proteins, we used the refined genome database to create the corresponding decoy protein database and search our mass spectrometry data against it. Resubmission of 1,190,811 spectra to the Mascot search engine resulted in identification of 32,126 nonredundant peptide sequences that mapped to 4029 *P. pacificus* protein groups at FDR 1% (Supplemental Table 5). To gain insight into the distribution of functional protein classes among detected proteins, we used the Blast2GO software to perform BLAST searches of detected protein sequences against the complete nrNCBI database and to extract the Gene Ontology (GO) terms. The GO analysis of the detected *P. pacificus* proteins revealed overrepresentation of cytosolic and developmental proteins, and underrepresentation of membrane proteins (Supplemental Fig. 1). The distribution of GO terms compared favorably with the recent proteomics analysis of *C. elegans* (Schrimpf et al. 2009) and demonstrated a sampling of similar protein classes in *P. pacificus* despite the more comprehensive proteome coverage in the *C. elegans* study.

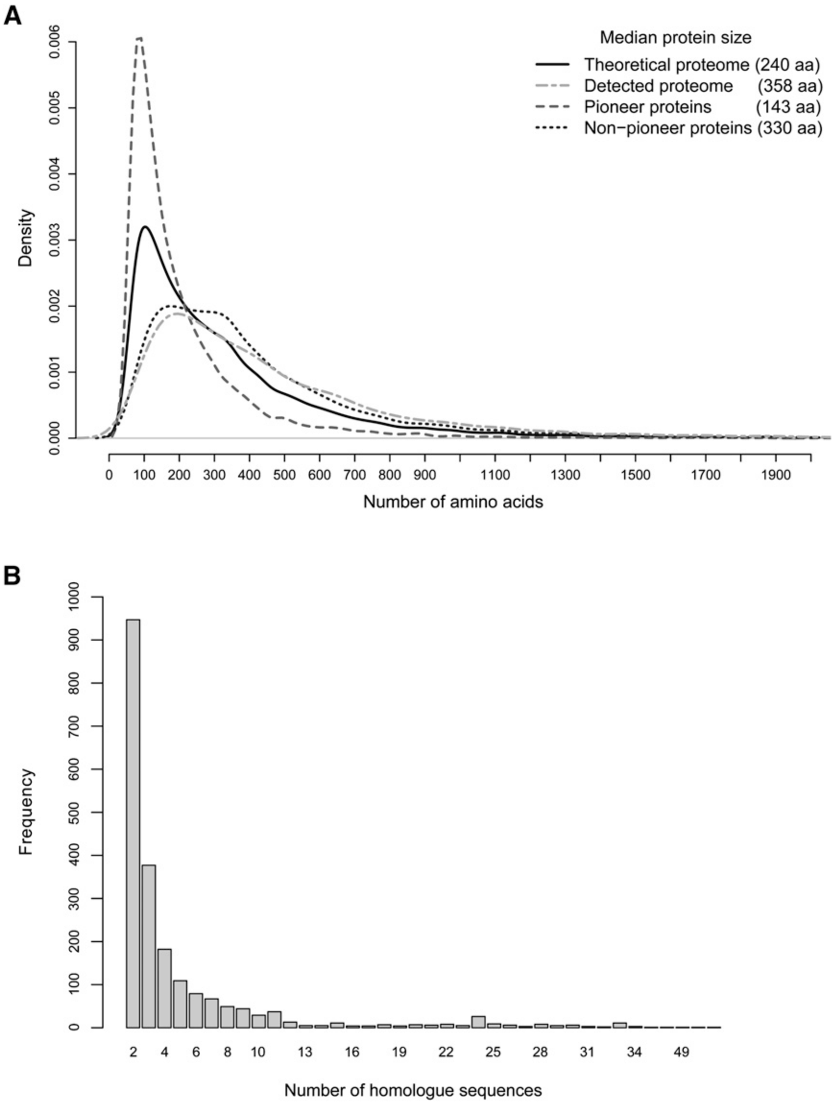


Figure 4. Features of the *P. pacificus* predicted proteome. (*A*) Protein size distribution shows that the pioneer proteins are mainly responsible for the unusually low median protein size in *P. pacificus*. (*B*) BLAST results of the pioneer proteins against themselves show presence of highly homologous proteins that may have a common origin.

Features of the predicted *P. pacificus* proteome

In silico translation of the predicted *P. pacificus* protein-coding genes showed an unusually low median predicted protein size of 240 amino acids (26.8 kDa) (Fig. 4A). BLAST analysis of the predicted proteome against the whole nrNCBI database did not retrieve any significant hits ($E < 1 \times 10^{-3}$) for 10,258 (42.3%) predicted proteins. We refer to them as “pioneer” proteins.

To gain insights into this group of proteins, we created a database consisting only of pioneer proteins and analyzed their features separately from the complete predicted proteome. Interestingly, the pioneer proteins are very short, with a median protein size of 143 amino acids (16 kDa). Their removal from the predicted proteome resulted in a significant increase in median size of the remaining proteins, from 240 amino acids (26.8 kDa) to 330 amino acids (36.9 kDa) (Fig. 4A), a value very close to the median size of proteins detected by MS (358 amino acids or 40 kDa). This leads to the conclusion that a majority of pioneer genes are not translated under tested conditions (environmental and/or developmental). Indeed, out of 4029 *P. pacificus* protein groups detected by MS, only 435 (10.8%) were products of pioneer genes. The search of unidentified MS spectra against a decoy database consisting only of pioneer proteins did not lead to significantly better coverage (data not shown). However, the coverage of pioneer genes was greater in the transcriptome analysis, where 5224 (51%) were detected to be expressed. To verify expression of pioneer genes we performed developmental stage-specific RT-PCR experiments. Out of 86 randomly chosen pioneer genes detected by MS, 56 could be amplified from mixed-stage cDNA and sequenced. Of this 56 transcripts, 31 showed different levels of expression in the tested second to fourth juvenile (J2–J4) stages, showing that at least some of the pioneer proteins may be

Downloaded from genome.cshlp.org on January 7, 2013 - Published by Cold Spring Harbor Laboratory Press

functionally relevant in different developmental stages (Supplemental Table 6).

To gain further insights into primary sequence characteristics and origin of pioneer proteins, we performed a stringent BLAST analysis ($E < 1 \times 10^{-20}$, bit score > 90) of every entry in the pioneer protein database against the whole database. Despite the stringent criteria, 2086 entries (20%) returned multiple (1–85) BLAST hits, pointing to the existence of close structural homologs among pioneer proteins (Fig. 4B). Indeed, the first genome draft of *P. pacificus* has already revealed that ~30% of the pioneer genes could be grouped into distinct protein families (Dieterich et al. 2008). We observe that some of the structurally related pioneer proteins reside on the same translated contigs, reflecting the proximity of their corresponding genes in the genome. Together, these data point to a likely common origin of part of the pioneer genes.

Comparison of the predicted *P. pacificus* proteome with proteomes of nematode models

We used the reannotated genome as a starting point for comparison of the predicted proteome of *P. pacificus* with three published nematode proteomes: *C. elegans* (The *C. elegans* Sequencing Consortium 1998), *C. briggsae* (Stein et al. 2003), and *Brugia malayi* (Ghedin et al. 2007). Surprisingly, the predicted protein sizes followed a unimodal distribution in *P. pacificus* and *B. malayi* and a distinct bimodal distribution in *C. elegans* and *C. briggsae* (Fig. 5A). To test whether this was a consequence of different qualities of gene annotations, we extended this comparison to three additional members of the *Caenorhabditis* genus: *C. remanei*, *C. brenneri*, and *C. japonica*, whose genome assemblies are publicly available (http://www.wormbase.org/), but are not yet peer-reviewed. These three organisms also showed distinct distributions of protein sizes, with *C. japonica* matching the unimodal protein size distribution and *C. remanei* and *C. brenneri* matching the bimodal distribution (Supplemental Fig. 2A,B). To assess the functional relevance of this observation, we performed GO analysis of the predicted *P. pacificus* and *C. elegans* proteomes. Whereas the GO analysis of the two proteomes showed a very similar overall distribution of GO classes (Supplemental Fig. 3), the GO analysis applied separately to the small and large protein populations in *C. elegans* pointed to a significant functional relevance (Fig. 5B). The short protein population was enriched in functions related to nucleosome assembly, translation, and development, while the long protein population was enriched in functions related to protein phosphorylation, signal transduction, and ion transport. Notably, protein functions (GO terms) enriched in one tested data set were depleted in the other, pointing to the functional complementarity of the two protein populations.

Discussion

In this study, we performed a comprehensive analysis of gene expression in *P. pacificus* with the goals of (1) genome refinement, (2) in-depth analysis of the detectable proteome, and (3) comparative predicted proteome analysis of the nematode model organisms. To achieve optimal gene expression coverage, we performed transcriptome and proteome analyses of *P. pacificus* cultures from different developmental stages, covering the mixed population, dauer, and J2 stages. By sequencing ESTs we achieved comprehensive coverage of the transcribed genomic regions and complemented it with information on the translated genomic regions from the proteomics experiment. Both technological platforms used in this study—454 FLX nucleic acid pyrosequencing and LTQ-Orbitrap mass spectrometry—represent the state of the art in the fields of transcriptomics and proteomics, respectively. The use of the 454 platform in this study enabled acquisition of one of the most comprehensive collection of ESTs so far used for genome refinement. In addition, the use of LTQ-Orbitrap MS resulted in one of the most accurate proteogenomics data sets to date, both in terms of mass accuracy and FDR (0.2%). We note that in this study the MS/MS spectra were recorded in the low-resolution linear ion trap analyzer, whereas the MS spectra were recorded in the high-resolution Orbitrap mass analyzer. This was needed to achieve a high speed of MS/MS acquisition at high precursor ion mass accuracy, both of which are crucial in proteogenomics. One of the challenges in the use of mass spectrometry in proteogenomics applications is the use of six-frame translation protein databases that result in the increase of search space and decrease in search specificity. Although high mass accuracy has an obvious potential to resolve this problem, the proteome complexity and high dynamic range of gene expression have so far made fast-scanning and low-accuracy mass spectrometers almost exclusively used in proteogenomics applications. While such instrumentation ensures in-depth proteome coverage, it requires the use of wide mass tolerance windows (up to 4 Da) during database search. In this study, the sub-ppm measurement mass accuracy of precursor ions enabled the use of narrow initial mass tolerance window during database search (7 ppm or 0.007 Da at $m/z = 1000$), and even higher tolerances for peptide acceptance (Cox and Mann 2008), thereby significantly increasing search specificity and reducing the FDR. This is especially important when searching the peptide mass spectra against the translation of the complete genome assembly as >80% of entries present nonsense protein sequences (e.g., wrong reading frames).

The refined *P. pacificus* gene predictions provided a unique opportunity to study its proteome features and compare them with *C. elegans*, a related and well-studied nematode model. *C. elegans* was recently a subject of a comprehensive proteogenomics study, in which 245 novel genes were identified and 151 existing gene models were modified (Merrihew et al. 2008). The drastically higher number of newly predicted and modified *P. pacificus* gene models in our study is caused in part by using a transcriptomics platform for genome reannotation, but also reflects more comprehensive existing genome annotation of *C. elegans*.

A distinct feature of the *P. pacificus* proteome is the presence of short proteins with no apparent homologs in current protein databases such as nrNCBI ("pioneer" proteins). A high proportion of genes without any known homologs was previously reported in the *P. pacificus* genome (Dieterich et al. 2008), but their expression was never demonstrated. Here, we show that at least a portion of these genes is expressed under tested conditions. Although only ~10% of predicted pioneer proteins were detected by MS, their coverage was higher in the transcriptome data set, where >50% were detected. At present, it is not clear whether this discrepancy is due to better transcriptome coverage, impaired translation, or low abundance (and therefore undersampling) of this class of proteins. The RT-PCR data showed that at least part of the pioneer genes show stage-specific expression, pointing to their potential roles in development. Interestingly, >20% of the pioneer proteins show similarity in primary structure and may therefore have a common origin. Although these data show that a fraction of pioneer proteins are synthesized and may be functional, their exact function and origin remain to be elucidated.

Downloaded from genome.cshlp.org on January 7, 2013 - Published by Cold Spring Harbor Laboratory Press

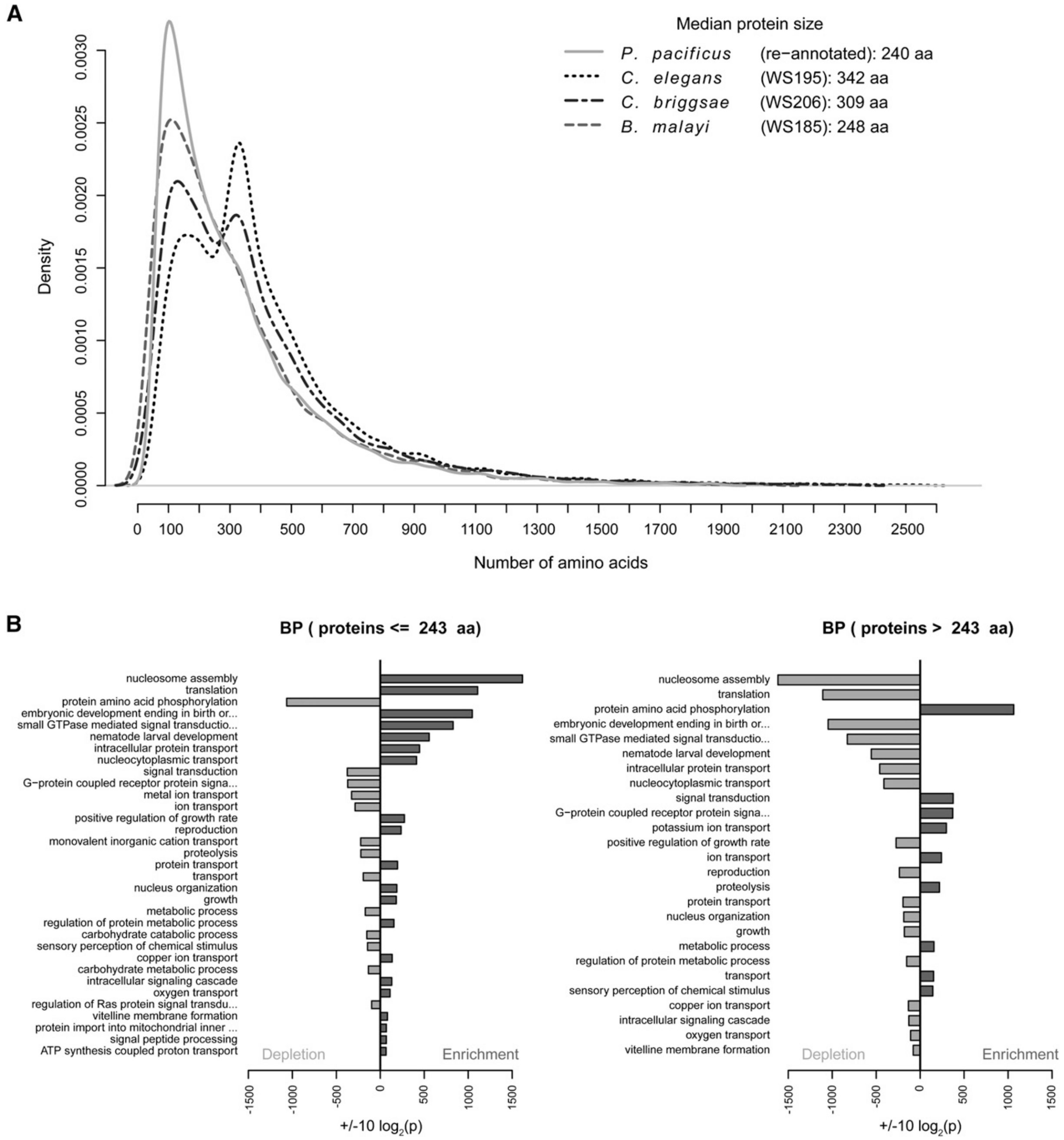


Figure 5. Comparison and functional analysis of protein size distributions in nematode models. (*A*) Predicted protein sizes in *P. pacificus* and *B. malayi* have a unimodal distribution, whereas *C. elegans* and *C. briggsae* have distinct bimodal distributions. (*B*) Gene Ontology enrichment analysis for short and long proteins in *C. elegans* shows distinct functional differences between the two classes of proteins.

An interesting aspect that arose from the comparison of *P. pacificus* and *C. elegans* proteomes is the bimodality of protein size distribution in the *C. elegans* proteome. To our knowledge, this is the first reported example of a bimodal distribution of protein sizes in any proteome, with pronounced functional differences between the two protein populations. At present it is not clear whether this distinct proteome feature is of biological relevance; however, it seems to represent a phylogenetic trait, as only species of the *Caenorhabditis* crown group show the bimodal distribution, whereas *C. japonica* follows a unimodal distribution. Also, we note that the enrichment of GO terms related to protein phosphorylation among the larger protein population may reflect the unusually high number of protein kinases reported in *C. elegans* (Manning et al. 2002). Since the recently published phosphoproteome of *C. elegans* showed an unusual functional distribution of phosphoproteins (Zielinska et al. 2009), a quantitative comparison of *P. pacificus* and *C. elegans* organisms at the phosphoproteome level is likely to give valuable insights into evolution of phosphorylation networks in Metazoa.

Downloaded from genome.cshlp.org on January 7, 2013 - Published by Cold Spring Harbor Laboratory Press

Methods

Culturing of worms and preparation of protein extracts

P. pacificus strain PS312 was grown on 10-cm NGM agar plates spotted with 2 mL of *E. coli* OP50 solution. Plates were inoculated with 50–100 worms and incubated at 25°C. The mixed-stage population was harvested shortly after the bacterial lawn was consumed, avoiding starvation of the animals. After thoroughly washing with distilled water and 0.9% sodium chloride, the animals were incubated with ampicillin (100 μL/mL) and chloramphenicol (34 μL/mL) in 0.9% sodium chloride for 48 h to remove residual bacteria. The worms were then pelleted and prepared for proteomics measurements. The animals in the J2 developmental stage were harvested as follows: Plates full of eggs were washed with distilled water and the animals were bleached with hydrogen peroxide and 5 M sodium hydroxide, leaving just the eggs alive. Animals were then spotted on 10-cm NGM agar plates for hatching for 24 h, and collected by washing after removing debris and corpses. Animals were pelleted and stored frozen until further analysis. For protein isolation, 100 μL of animals was solubilized in 300 μL of denaturation buffer (6 M urea, 2 M thiourea, 10 mM Tris at pH 8.0). After three cycles of freeze (liquid nitrogen) and thaw (37°C), 100 μL of glass beads were added and the solution was vortexed for 20 min. After centrifugation (20 min, 20.800*g*, 4°C), the protein concentration of the supernatant was determined using the Bradford assay. The pellet was solubilized in sample buffer for gel electrophoresis and further analysis.

OffGel electrophoresis and in-solution digestion

For OffGel fractionation the proteins were reduced by incubation in 1 mM dithiotreitol (DTT) for 1 h at room temperature. Alkylation was performed in 5.5 mM iodoacetamide (IAA) in 50 mM ABC for 1 h at room temperature in the dark. Proteins were digested using LysC (1:100 w/w) for 3 h at room temperature and trypsin (1:100 w/w) overnight at room temperature after diluting the sample with four volumes of 20 mM ammonium bicarbonate (ABC). The resulting peptides were separated using the 3100 OffGel fractionator (Agilent) according to manufacturer's instructions with a 12- or 24-well setup. Focusing was done with 13-cm (12-well) or 24-cm (24-well) Immobiline DryStrips pH 3–10 (GE Healthcare) at a maximum current of 50 μA for 50 kVh. Peptide fractions were harvested and desalted using C18 StageTips as previously described (Ishihama et al. 2006).

GeLC-MS and in-gel digestion

For GeLC-MS analysis 100 μg of the supernatant and the solubilized pellet were loaded on a NuPAGE Bis-Tris 4%–12% gradient gel (Invitrogen). After brief Coomassie staining, each lane was cut into 10 slices that were further cut into small pieces. Destaining was performed by washing three times with 10 mM ABC and acetonitrile (ACN) (1:1, v/v) and was followed by protein reduction with 10 mM DTT in 20 mM ABC for 45 min at 56°C, and alkylation with 55 mM iodoacetamide in 20 mM ABC for 30 min at room temperature in the dark. The gel pieces were then washed twice for 20 min in destaining solution followed by dehydration with ACN. The liquid was removed and gel pieces were swollen at room temperature by adding 13 ng/μL sequencing-grade trypsin (Promega) in 20 mM ABC. Digestion of proteins was performed at 37°C overnight. The resulting peptides were extracted in three subsequent incubation steps with 30% ACN/3% TFA; with 80% ACN/0.5% acetic acid; and with 100% ACN. Supernatants were combined, ACN was evaporated in a vacuum centrifuge, and peptides were desalted using C18 StageTips.

NanoLC-MS/MS analysis

All digested peptide mixtures were separated on a nanoLC-2D HPLC (Eksigent) coupled to a LTQ-Orbitrap-XL (Thermo Fisher Scientific) through a nano-LC-MS interface (Proxeon Biosystems). Binding and chromatographic separation of the peptides was performed on a 15-cm fused silica emitter of 75-μm inner diameter (Proxeon Biosystems), in-house packed with reversed-phase ReproSil-Pur C18-AQ 3-μm resin (Dr. Maisch GmbH). The peptide mixtures were injected onto the column in HPLC solvent A (0.5% acetic acid) at a flow rate of 500 nL/min and subsequently eluted with a 107-min segmented gradient of 2%–80% HPLC solvent B (80% ACN in 0.5% acetic acid) at a flow rate of 200 nL/min.

The mass spectrometer was operated in the data-dependent mode to automatically switch between MS and MS/MS acquisition. Survey full-scan MS spectra were acquired in the mass range of *m*/*z* 300–2000 in the orbitrap mass analyzer at a resolution of 60,000. An accumulation target value of 10^6 charges was set and the lock mass option was used for internal calibration (Olsen et al. 2005). The 10 most intense ions were sequentially isolated and fragmented in the linear ion trap using collision-induced dissociation (CID) at the ion accumulation target value of 5000 and default CID settings. The ions already selected for MS/MS were dynamically excluded for 90 sec. The resulting peptide fragment ions were recorded in the linear ion trap. In total, 101 LC-MS measurements were performed, corresponding to 10 d of measurement time.

The mass spectrometry data associated with this manuscript may be downloaded from the Proteome Commons Tranche repository (https://proteomecommons.org/tranche/) using the following hash: QgF9ukyrC8Y74IIE8L/y2ccmTd02ElO8UnFcVLVF wvy1C+/41QDGWVzZIR96f33MKIui57iuS6x8 8KNT2v4RiIuHRN4 AAAAAAAALhA==.

Data processing and analysis

MaxQuant data processing and Mascot database search

All raw files were processed together using the MaxQuant software suite (v. 1.12.35) (Cox and Mann 2008; Cox et al. 2009). Raw MS spectra were first processed by the Quant module to generate peak lists. This module performs a nonlinear mass recalibration for each individual precursor ion and calculates precise masses as well as individual mass errors. To retrieve peptide sequences from the processed spectra, we used the Mascot search engine v.2.2 (Matrix Science). The processed MS spectra were searched against an in-house assembled target-decoy database that consisted of the in silico–predicted proteome of *P. pacificus* (SNAPNG2.aa.annot, 27,103 sequences); a complete six-frame translation of its genome (sctg_plus_2000.fas, 14,654 contigs; 87,924 sequences after six-frame translation); *E. coli* proteome (4256 sequences); and 262 commonly observed contaminants. All protein sequences in the database were reversed and appended to the database. This enabled the estimation of false discovery rate (FDR) in the data set by a target-decoy search strategy (Elias and Gygi 2007).

In the database search, carbamidomethylation (Cys) was set as fixed modification, whereas oxidation (Met) and acetylation (protein N termini) were set as variable modifications. The mass tolerances for precursor and fragment ions were set to 7 ppm and 0.5 Da, respectively.

The retrieved peptide sequences were further processed with the Identify module of the MaxQuant software. This module considers all 10 peptide candidates suggested by Mascot for each fragmentation spectrum and filters them according to consistency with a priori information, e.g., the individual precursor mass errors. Furthermore, the probability that an individual peptide is a false hit given its score and length is estimated by a Bayesian

Downloaded from genome.cshlp.org on January 7, 2013 - Published by Cold Spring Harbor Laboratory Press

probability (posterior error probability [PEP]). All filtered peptide sequences are sorted according to their PEP values, starting with the best PEP. To control the FDR the peptides are accepted until 1% of reversed peptides have accumulated within the list. The identified peptides are then assembled back into proteins. If a set or subset of identified peptides can be assigned to more than one protein, these proteins are joined into a protein group (Nesvizhskii and Aebersold 2005; Cox and Mann 2008). Finally, the FDR on protein group level was also controlled to be at 1%.

PepSplice database search

The PepSplice search engine (Roos et al. 2007) was used to complement Mascot-based searches. PepSplice uses a cache-optimized peptide database search algorithm for aligning spectra to genome-wide spliced six-frame translations. MaxQuant-processed MS/MS spectra (J2 + Mixed Stage) were searched against a target database, which contained all spliced six-frame translations of the 965 largest supercontigs (≥2 kb). All possible splicing events up to an intron size of 2 kb were considered and the maximal FDR was set to 1% on the peptide level. PepSplice also employs a target-decoy search strategy to estimate the FDR.

Downstream bioninformatics analysis

All downstream bionformatics was done in R (v. 2.9.0; http://www.r-project.org).

Protein size distributions were determined from the most recent versions of publicly available protein databases (http://www.wormbase.org). Distributions were determined by the "density" function from the base R package using default bandwidth. For robust estimation of protein size distribution, 99% of all proteins within the particular databases were considered. All BLAST searches in this study were performed by BLASTP v.2.2.21.

Gene Ontology analysis

GO annotation for the predicted *P. pacificus* and *C. elegans* (WS140, WS195) proteomes was derived using Blast2GO software (Conesa et al. 2005). For each query sequence the software first detects up to 20 homolog sequences in the nrNCBI database (nrNCBI version was from August, 2009; E-value $<1 \times 10^{-3}$) by a BLAST search. Based on the GO terms associated with these candidate sequences the software applies an annotation rule that filters and reports the most specific annotations.

To test for enrichment or depletion of specific GO terms among the identified proteome, the topGO R package was used (Alexa et al. 2006). This package implements two scoring methods that take care of the underlying GO graph topology. We used the "elim" algorithm that starts at the leaves of the induced GO graph and subsequently removes all proteins from the corresponding parent nodes that have been already used for testing the children nodes. Therefore, only the most specific GO terms for each protein were considered.

Fisher's exact test served as test statistic assuming the hypergeometric distribution as null-distribution. The derived P-values were further adjusted for multiple hypothesis testing by the method of Benjamini and Hochberg (Benjamini and Hochberg 1995) to control the FDR.

PCR analysis

To validate expression of proteins that were identified by MS, we chose 184 genes for RT-PCR. The primers were designed with the online tool Primer3 (Rozen and Skaletsky 2000) with an average amplicon size of 100 bp and were purchased from Eurofins MWG. For stage-specific cDNA, J2 stages were collected as described above and grown to J3 and J4, respectively. Total RNA was isolated using TRIzol (Invitrogen) according to the manufacturer's instructions. cDNA was produced using the Superscript III cDNA synthesis kit (Invitrogen) for 2 h at 42°C for the reverse transcription. PCR reactions were performed for 35 cycles of 20 sec at 95°C, 30 sec at 55°C, and 30 sec at 72°C. The reactions were subsequently separated on a 2% TBE agarose gel, stained with ethidium bromide, and visualized under UV light.

Transcriptome sequencing on the 454 Life Sciences (Roche) FLX platform

Total RNA was isolated from a mixed and dauer stage culture of *P. pacificus* (Ppa 312, California) using TRIzol (Invitrogen) according to the manufacturer's instructions. The RNA was sequenced at the Genome Sequencing Center at Washington University, St. Louis, MO using the 454 FLX for 454 sequencing.

Transcriptome assembly

The 454 reads were processed prior to assembly. Low-quality base calls were removed from read ends by Lucy (Chou and Holmes 2001) using default settings. Highly repetitive sequence segments were removed by Figaro (White et al. 2008) using default settings. We assembled 28,599/26,092 contigs from the 350,839/394,453 remaining sequences with the EST version of PCAP.REP. These contigs encompass >10 Mb of transcribed sequence.

Transcriptome mapping

The assembled contigs and all trimmed reads were aligned to the genome using Exonerate (Slater and Birney 2005) with a maximal intron size of 20 kb. In summary, we could identify 98,254 unique acceptor and 95,210 unique donor sites. This data set was subsequently used to improve the Pristionchus genome annotation.

Gene prediction

We took the 11 largest supercontigs from the Hybrid Genome Assembly (Sanger + 454). We predicted a new set of genes with the current hidden Markov model (HMM) gene model plus external evidence as given by the 454 transcriptome data (98254 Acceptor and 95210 Donor sites). This new gene set was subsequently used to retrain our HMM gene model (SNAPNG2).

All protein database searches for peptide identification were carried out on this reference data set. We used the genomic hits from Mascot and PepSplice as additional external evidence in the next gene model training step (4431 data points/coding segments). We updated our gene model to its final version and reran the gene predictions including all available external evidence (MS/MS + 454).

Acknowledgments

This work is supported by grants of Ministerium fuer Wissenschaft und Kunst and Landesstiftung Baden-Wuerttemberg to the Proteome Center Tübingen and by the Innovation Fund of the Max-Planck Society grant to R.J.S.

References

Alexa A, Rahnenfuhrer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22:** 1600–1607.

Downloaded from genome.cshlp.org on January 7, 2013 - Published by Cold Spring Harbor Laboratory Press

Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD. 2008. Proteogenomics: Needs and roles to be filled by proteomics in genome annotation. *Brief Funct Genomics Proteomics* **7:** 50–62.

Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S. 2008. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* **320:** 938–941.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57:** 289–300.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, et al. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419:** 512–519.

Chou HH, Holmes MH. 2001. DNA sequence quality trimming and vector removal. *Bioinformatics* **17:** 1093–1104.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21:** 3674–3676.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26:** 1367–1372.

Cox J, Matic I, Hilger M, Nagaraj N, Selbach M, Olsen JV, Mann M. 2009. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat Protoc* **4:** 698–705.

Dieterich C, Sommer RJ. 2009. How to become a parasite—lessons from the genomes of nematodes. *Trends Genet* **25:** 203–209.

Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, Fulton L, Fulton R, Godfrey J, Minx P, et al. 2008. The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat Genet* **40:** 1193–1198.

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4:** 207–214.

Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ. 2006. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol* **7:** R35. doi: 10.1186/gb-2006-7-4-r35.

Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, Allen JE, Delcher AL, Guiliano DB, Miranda-Saavedra D, et al. 2007. Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317:** 1756–1760.

Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, Kerman I, Nguyen N, Ollikainen N, Rodriguez J, Wang J, et al. 2008. Comparative proteogenomics: Combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res* **18:** 1133–1142.

Hong RL, Sommer RJ. 2006. *Pristionchus pacificus*: A well-rounded nematode. *Bioessays* **28:** 651–659.

Ishihama Y, Rappsilber J, Mann M. 2006. Modular stop and go extraction tips with stacked disks for parallel and multidimensional peptide fractionation in proteomics. *J Proteome Res* **5:** 988–994.

Manning G, Plowman GD, Hunter T, Sudarsanam S. 2002. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* **27:** 514–520.

Merrihew GE, Davis C, Ewing B, Williams G, Kall L, Frewen BE, Noble WS, Green P, Thomas JH, MacCoss MJ. 2008. Use of shotgun proteomics for the identification, confirmation, and correction of *C. elegans* gene annotations. *Genome Res* **18:** 1660–1669.

Nesvizhskii AI, Aebersold R. 2005. Interpretation of shotgun proteomic data: The protein inference problem. *Mol Cell Proteomics* **4:** 1419–1440.

Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, Pesch R, Makarov A, Lange O, Horning S, Mann M. 2005. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics* **4:** 2010–2021.

Roos FF, Jacob R, Grossmann J, Fischer B, Buhmann JM, Gruissem W, Baginsky S, Widmayer P. 2007. PepSplice: Cache-efficient search algorithms for comprehensive identification of tandem mass spectra. *Bioinformatics* **23:** 3016–3023.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132:** 365–386.

Schrimpf SP, Weiss M, Reiter L, Ahrens CH, Jovanovic M, Malmstrom J, Brunner E, Mohanty S, Lercher MJ, Hunziker PE, et al. 2009. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol* **7:** e48. doi: 10.1371/journal.pbio.1000048.

Shin H, Hirst M, Bainbridge MN, Magrini V, Mardis E, Moerman DG, Marra MA, Baillie DL, Jones SJ. 2008. Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags. *BMC Biol* **6:** 30. doi: 10.1186/1741-7007-6-30.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31. doi: 10.1186/1471-2105-6-31.

Sommer RJ, Carta LK, Kim SY, Sternberg PW. 1996. Morphological, genetic and molecular description of *Pristionchus pacificus* sp n (Nematoda: Neodiplogastridae). *Fundam Appl Nematol* **19:** 511–521.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol* **1:** e45. doi: 10.1371/journal.pbio.0000045.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol* **17:** 1636–1647.

White JR, Roberts M, Yorke JA, Pop M. 2008. Figaro: A novel statistical method for vector sequence removal. *Bioinformatics* **24:** 462–467.

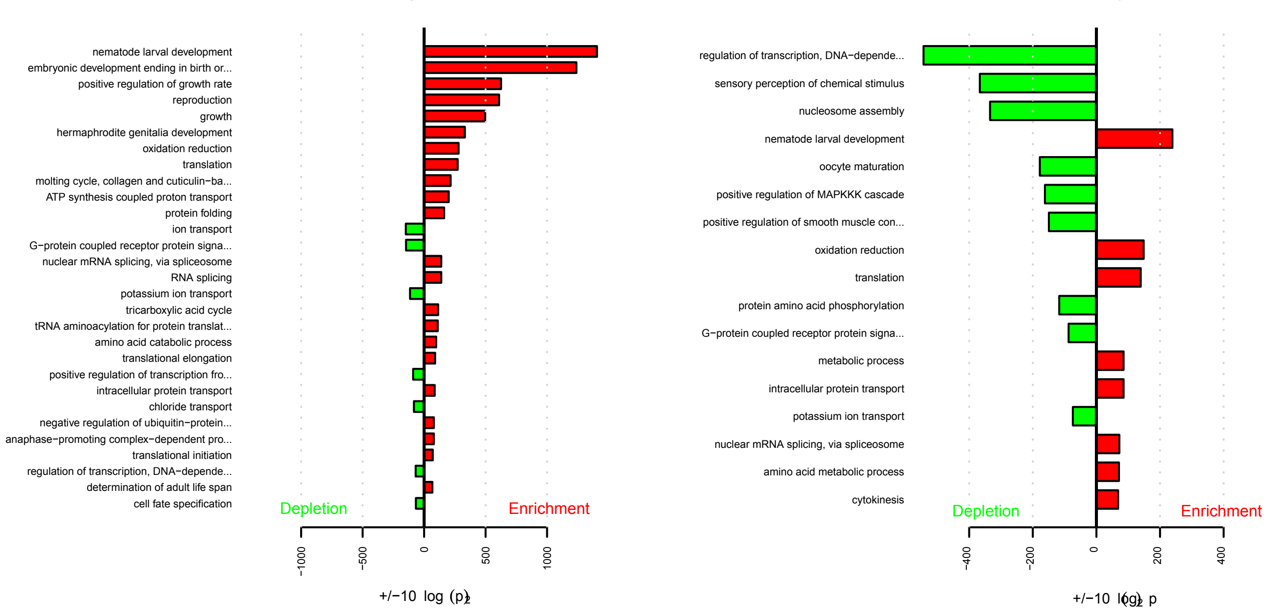
Xia D, Sanderson SJ, Jones AR, Prieto JH, Yates JR, Bromley E, Tomley FM, Lal K, Sinden RE, Brunk BP, et al. 2008. The proteome of *Toxoplasma gondii*: Integration with the genome provides novel insights into gene expression and annotation. *Genome Biol* **9:** R116. doi: 10.1186/gb-2008-9-7-r116.

Yates JR III. Eng JK, McCormack AL. 1995. Mining genomes: Correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem* **67:** 3202–3210.

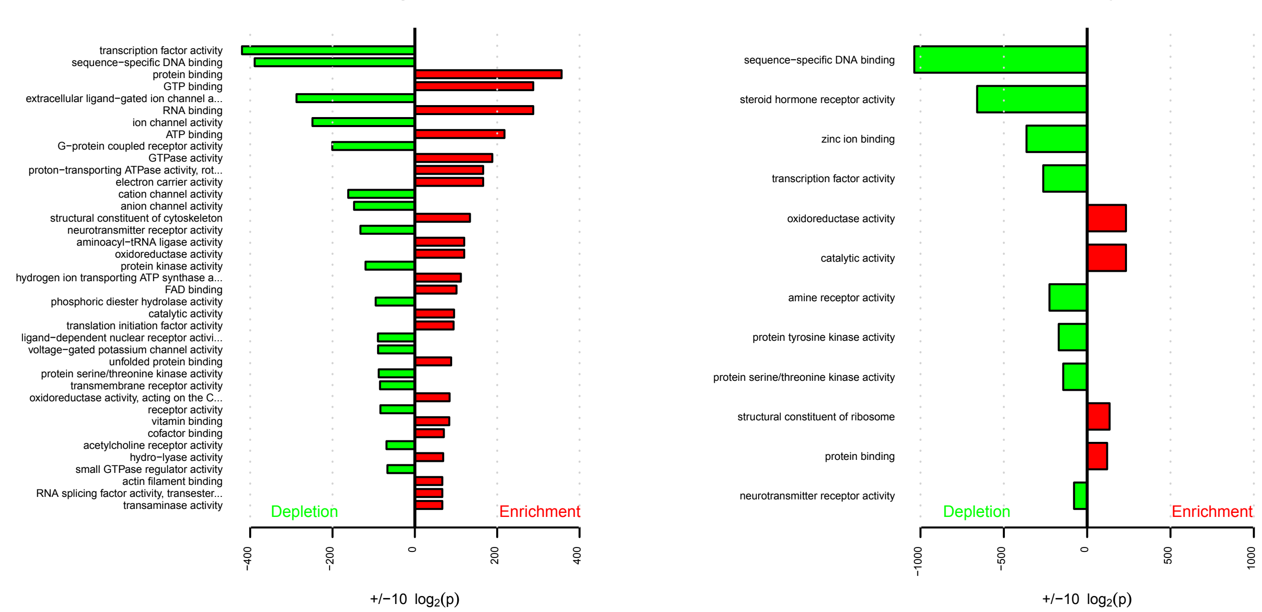
Zielinska DF, Gnad F, Jedrusik-Bode M, Wisniewski JR, Mann M. 2009. *Caenorhabditis elegans* has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins. *J Proteome Res* **8:** 4039–4049.

Received November 11, 2009; accepted in revised form March 10, 2010.

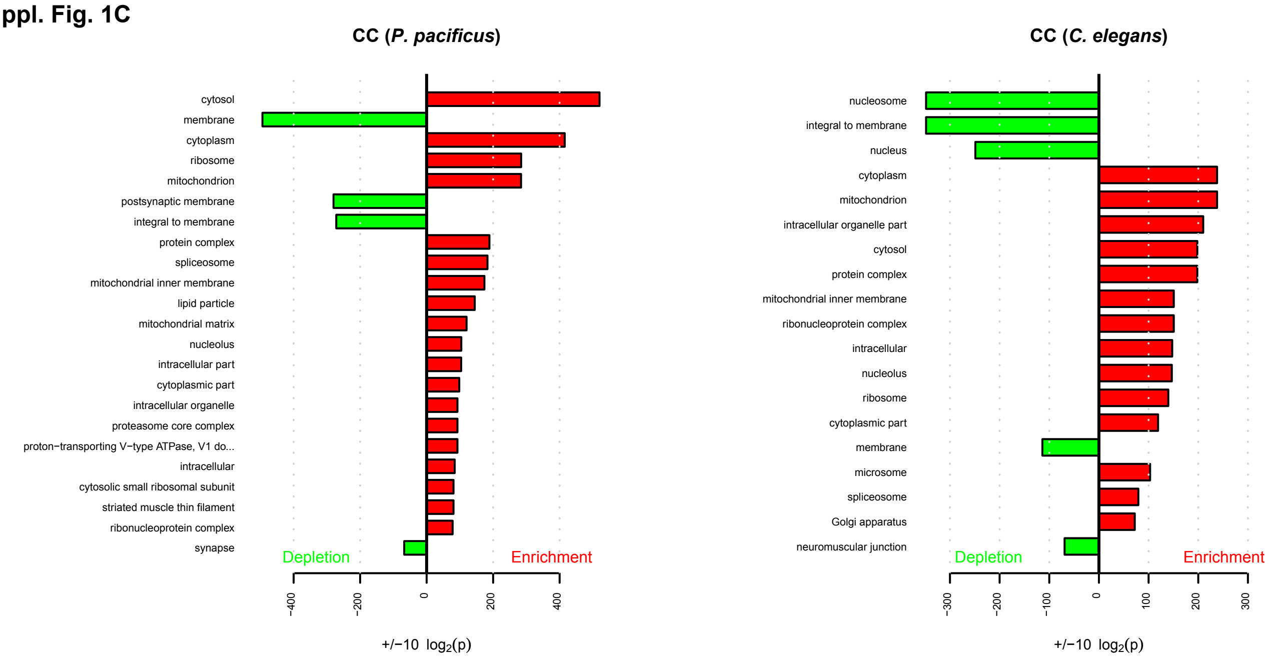
Suppl. Fig. 1A



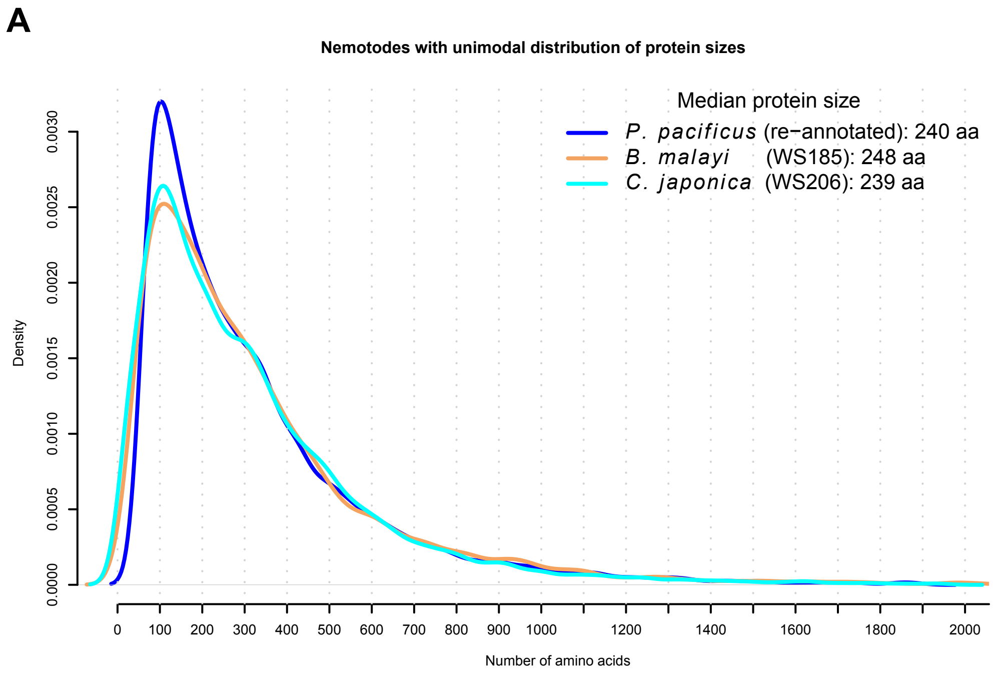
Suppl. Fig. 1B

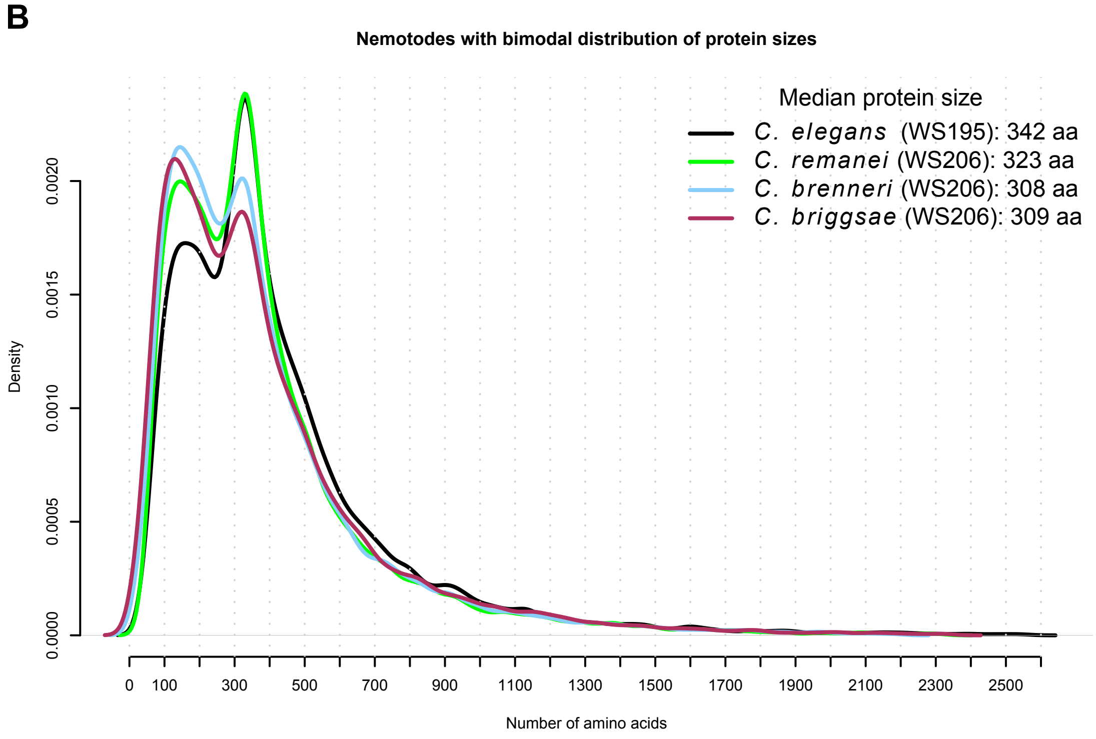


Suppl. Fig. 1C



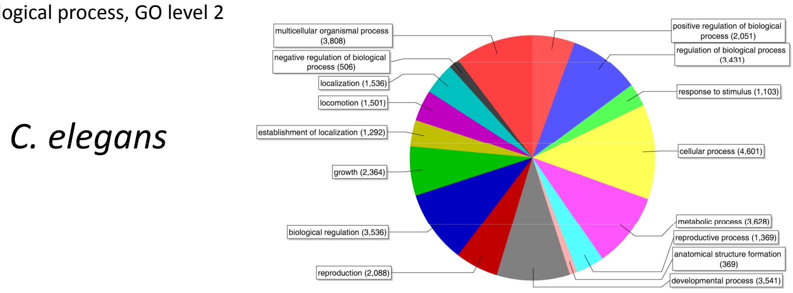
Suppl. Fig. 2

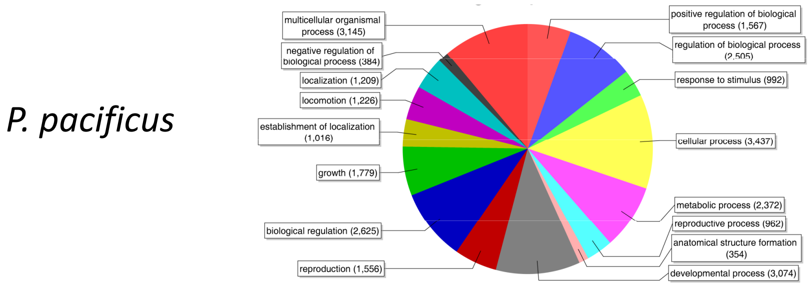

A
Nemotodes with unimodal distribution of protein sizes
Median protein size
P. pacificus (re−annotated): 240 aa
B. malayi (WS185): 248 aa
C. japonica (WS206): 239 aa
Density
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0
100
200
300
400
500
600
700
800
900
1000
1200
1400
1600
1800
2000
Number of amino acids


B
Nemotodes with bimodal distribution of protein sizes
Median protein size
C. elegans (WS195): 342 aa
C. remanei (WS206): 323 aa
C. brenneri (WS206): 308 aa
C. briggsae (WS206): 309 aa
Density
0.0000
0.0005
0.0010
0.0015
0.0020
0
100
300
500
700
900
1100
1300
1500
1700
1900
2100
2300
2500
Number of amino acids

Suppl. Fig. 3A

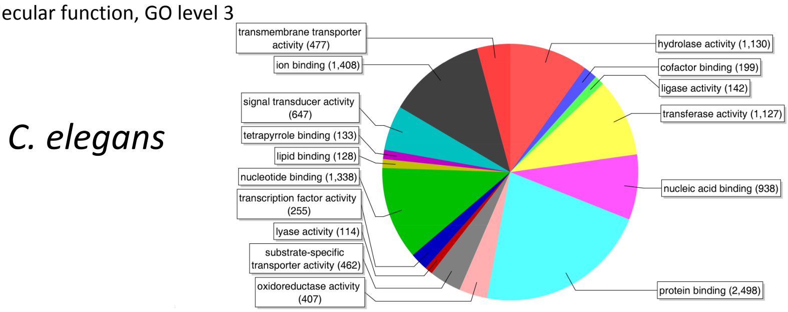
Biological process, GO level 2

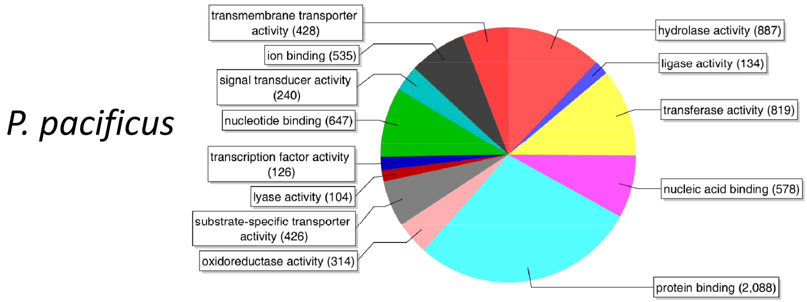




Suppl. Fig. 3B

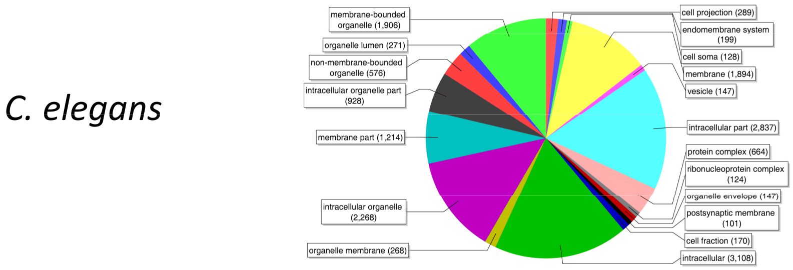
Molecular function, GO level 3

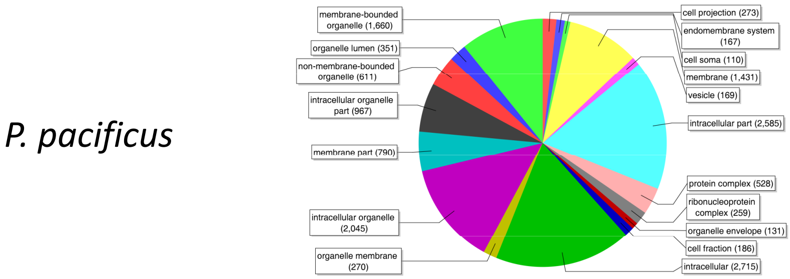




Suppl. Fig. 3C

Cellular component, GO level 3





© 2012 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org

Phosphoproteome of *Pristionchus pacificus* Provides Insights into Architecture of Signaling Networks in Nematode Models[S]

Nadine Borchert‡§, Karsten Krug§¶, Florian Gnad‖, Amit Sinha‡, Ralf J. Sommer‡‡‡, and Boris Macek¶‡‡**

Pristionchus pacificus **is a nematode that is increasingly used as a model organism in evolutionary biology. The genome of *P. pacificus* differs markedly from that of *C. elegans*, with a high number of orphan genes that are restricted to *P. pacificus* and have no homologs in other species. To gain insight into the architecture of signal transduction networks in model nematodes, we performed a large-scale qualitative phosphoproteome analysis of *P. pacificus*. Using two-stage enrichment of phosphopeptides from a digest of *P. pacificus* proteins and their subsequent analysis via high accuracy MS, we detected and localized 6,809 phosphorylation events on 2,508 proteins. We compared the detected *P. pacificus* phosphoproteome to the recently published phosphoproteome of *C. elegans.* The overall numbers and functional classes of phosphoproteins were similar between the two organisms. Interestingly, the products of orphan genes were significantly underrepresented among the detected *P. pacificus* phosphoproteins. We defined the theoretical kinome of *P. pacificus* and compared it to that of *C. elegans*. While tyrosine kinases were slightly underrepresented in the kinome of *P. pacificus*, all major classes of kinases were present in both organisms. Application of our kinome annotation to a recent transcriptomic study of dauer and mixed stage populations showed that Ser/Thr and Tyr kinases show similar expression levels in *P. pacificus* but not in *C. elegans*. This study presents the first systematic comparison of phosphoproteomes and kinomes of two model nematodes and, as such, will be a useful resource for comparative studies of their signal transduction networks.** ***Molecular & Cellular Proteomics 11: 10.1074/mcp.M112.022103, 1631–1639, 2012.***

Pristionchus *pacificus* is a nematode that is established as a model in evolutionary developmental biology (2). Like *Caenorhabditis elegans*, which was the first multicellular organism to have its genome completely sequenced (3), it has several advantageous features: it is easy to cultivate in the laboratory, it feeds on *E. coli*, it has a short generation time of 4 days (at 20 °C), and, because it is a self-fertilizing hermaphrodite, it is amenable to forward and reverse genetic techniques. Its genome has recently been sequenced, revealing a high number of predicted genes that share no sequence similarity to genes from any other organisms ("orphan" or "pioneer" genes) (4). Like many other nematodes, *P. pacificus* exhibits phenotypic plasticity of its life cycle and is able to quickly adapt to different environmental conditions. Under favorable conditions, *P. pacificus* undergoes direct development, but it can arrest development to form a stress resistant dauer stage when the environmental conditions turn unfavorable. These examples of phenotypic plasticity have allowed nematodes to invade many different habitats (5). *P. pacificus* occupies a completely different ecological niche than *C. elegans*. It has a necromenic lifestyle in which the developmentally arrested dauer larva infests a scarab beetle and resumes development upon the beetle's death, feeding on the microorganisms that decompose the beetle's carcass (6). The estimated evolutionary distance between *C. elegans* and *P. pacificus* is 250 to 420 million years, which makes them very attractive models in evolutionary developmental biology (4).

We recently performed a comprehensive analysis of the proteome and transcriptome of *P. pacificus*, with the aim of refining its genome annotation. Retraining the gene prediction algorithm with gene expression data estimated the number of predicted open reading frames to 24,000. Comparison of our data to the predicted proteome of *C. elegans* revealed differences in the proteome structures of the two nematodes. Whereas the predicted proteome of *P. pacificus* showed a unimodal distribution of protein sizes, the proteome of *C. elegans* followed a clearly bimodal distribution. Interestingly, this bimodal distribution seemed to be connected to functions related to protein phosphorylation, suggesting a potential difference in protein phosphorylation between the two organisms (7).

To gain further insights into the proteome of *P. pacificus*, we performed a large-scale analysis of *P. pacificus* phosphopro-

---

From the ‡Department for Evolutionary Biology, Max Planck Institute for Developmental Biology, Tuebingen, Germany; ¶Proteome Center Tuebingen, University of Tuebingen, Germany; ‖Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Martinsried, Germany

Received July 6, 2012, and in revised form, August 21, 2012
Published, MCP Papers in Press, August 25, 2012, DOI 10.1074/mcp.M112.022103

teome using phosphopeptide enrichment and high accuracy mass spectrometry. Here we report the first comprehensive phosphoproteome map of *P. pacificus,* measured to a depth of almost 7,000 localized phosphorylation sites, and compare it to the recently reported phosphoproteome of *C. elegans* (8). We show that the two phosphoproteomes are of similar sizes but differ significantly in the frequencies of phosphorylated serine, threonine, and tyrosine residues. We define direct orthologs between the two organisms and show that this discrepancy is also pronounced at the ortholog level. We show that the products of orphan genes are significantly underrepresented among the detected *P. pacificus* phosphoproteins. Finally, we define the predicted kinome of *P. pacificus* and show that it is slightly smaller than that of *C. elegans* but contains all major classes of kinases.

MATERIALS AND METHODS

Culturing of Worms and Preparation of Protein Extracts—P. pacificus strain PS312 was grown on 10 cm NGM agar plates spotted with 2 ml *E. coli* OP50 solution. Plates were inoculated with between 50 and 100 worms and incubated at 25 °C. The mixed stage population was harvested shortly after the bacterial lawn was consumed, avoiding the starvation of the animals. After thorough washing with distilled water and 0.9% sodium chloride, worms were pelleted and prepared for proteomics measurements.[1]

For protein isolation, 100 μl of animals were solubilized in 300 μl denaturation buffer (6 M urea, 2 M thiourea, 10 mM Tris pH 8.0). After three cycles of freezing (liquid nitrogen) and thawing (37 °C), 100 μl of glass beads were added, and the solution was vortexed for 20 min. After centrifugation (20 min, 20.800 $\times$ *g*, 4 °C), the protein concentration of the supernatant was determined using the Bradford assay and further processed using the filter-aided sample preparation (FASP)[2] method (9) (see below).

Protein Digestion—The soluble protein fraction was digested as described previously (7). Briefly, 5 mg of protein was reduced with a final concentration of 1 mM DTT and alkylated with a final concentration of 5.5 mM iodoacetamide. After the pH was adjusted to 8.0, 1 μg of trypsin was added per 100 μg of protein, and the mixture was incubated overnight at 37 °C.

The insoluble protein fraction was processed with a modified FASP protocol (9). The protein pellet was solubilized in 4% SDS, 100 mM DTT, and 100 mM Tris pH 7.6. An aliquot of the pellet was precipitated with chloroform/methanol and solubilized in denaturation buffer for Bradford analysis. Based on the Bradford measurement, a protein-SDS solution containing 5 mg of protein was diluted with urea buffer A (8 M urea in 100 mM Tris, pH 8.5) to a final volume of 6 ml and pipetted into the 15-ml Centriprep column YM-30 (Millipore, Billerica, MA). After the sample had been spun for 15 min at 6,000 g, 600 μl of iodoacetamide solution (550 mM) was added, and the sample was incubated for 1 h in the dark and then centrifuged for 15 min at 3,000 g. The protein was washed with UA three times, and the last centrifugation step was increased to 20 min. Six ml of ammonium bicarbonate was added, and the sample was centrifuged for another 15 min at 3,000 g. After that, Trypsin was added at a final concentration of 1 μg per 100 μg total protein and incubated overnight at 37 °C. After the next centrifugation step (15 min, 3,000 g), the peptides were collected in the flowthrough. Centrifugation was repeated with 3 ml water, and the flowthrough was collected for strong cation exchange (SCX).

Phosphopeptide Enrichment—After 5 mg of digested total protein lysate had been acidified to pH 2.7 with trifluoroacetic acid, the sample was loaded onto an ÄKTApurifier (GE Healthcare, Little Chalfont, UK) HPLC for SCX. The 16 resulting fractions were pooled according to the elution profile to 10 fractions for titanium dioxide enrichment. Five mg of TiO_2 beads were resuspended in 50 μl of a 30 mg/ml 2.5 dihydrobenzoic acid, 80% acetonitrile in water solution. After 10 min of incubation at room temperature, the TiO_2 loading solution was added to the sample and mixed for 30 min at room temperature using an orbital shaker. The beads were precipitated with centrifugation at 13,000 rpm for 2 min and washed with 1 ml Wash Solution I (30% acetonitrile (ACN), 3% TFA) for 10 min in a shaker and Wash Solution II (10) (80% ACN, 0.1% TFA) for 10 min in a shaker. The beads were then resuspended in 50 μl Wash Solution II and transferred to a 200 μl pipette tip plugged with one layer of Empore C8 tip. After the beads had been washed three times with 100 μl 40% ammonia solution (25% in water) in ACN pH 10.5, the eluate was reduced to 5 μl in a SpeedVac.

NanoLC-MS/MS Analysis—Enriched phosphopeptide mixtures were separated via Easy-LC nano-HPLC (Proxeon Biosystems, Odense, DK) coupled to an LTQ-Orbitrap-XL (Thermo Fisher Scientific) through a nano-LC-MS interface (Proxeon Biosystems). Chromatographic separation of the peptides was performed on a 15 cm fused silica emitter with a 75 μm inner diameter (Proxeon Biosystems), in-house packed with reversed-phase ReproSil-Pur C18-AQ 3 μm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, DE). The peptide mixtures were injected onto the column in HPLC solvent A (0.5% acetic acid) at a flow rate of 500 nl/min and subsequently eluted with a 107 min segmented gradient of 2% to 80% of HPLC solvent B (80% acetonitrile in 0.5% acetic acid) at a flow rate of 200 nl/min.

The MS was operated in the data-dependent mode so as to automatically switch between MS and MS/MS acquisition. Survey full scan MS spectra were acquired in the mass range from *m/z* 300 to 2,000 in the orbitrap mass analyzer at a resolution of 60,000. An accumulation target value of 10^6 charges was set, and the lock mass option was used for internal calibration (11). The five most intense ions were sequentially isolated and fragmented in the linear ion trap using collision-induced dissociation (CID) at an ion accumulation target value of 5,000 and default CID settings. Multistage activation (at -98, -49, and -32.66 Th relative to the precursor ion) was used to optimize fragmentation of Ser/Thr phosphopeptides. The ions already selected for MS/MS were dynamically excluded for 90 s. The resulting peptide fragment ions were recorded in the linear ion trap. In total, 41 LC-MS measurements were performed, corresponding to 4 days of measurement time.

Data Processing and Analysis—MS data were processed with MaxQuant (12), version 1.0.14.3. Peak lists were generated and subsequently submitted to the Mascot search engine (Matrix Science, London, UK) to query a database consisting of the latest annotation of *P. pacificus* (dataset "HYBRID1 proteomics gene models"; 24,231

---

[1] The data associated with this manuscript may be downloaded from ProteomeCommons.org Tranche using the following hash: pxGey/Jh9q186pz5hyUKK13Idzf8sjFVLW+ZZbNgv0IkOAH71q31oIfK2vNyvp8wb7ItfBczkQ8O5W/llVxLtpPhpEoAAAAAAAA1Ow==.

The hash may be used to prove exactly what files were published as part of this manuscript's data set, and the hash may also be used to check that the data have not changed since publication. The data can also be viewed through the PHOSIDA database www.phosida.com (1).

[2] The abbreviations used are: CID, collision-induced dissociation; EC, enzyme commission number; EGF, epidermal-growth factor; FASP, filter-aided sample preparation; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; Pfam, protein families and domains; SCX, strong cation exchange; VPC, vulva precursor cell.

protein entries) (7), 4,256 *E. coli* proteins, 262 commonly observed protein contaminants, and 28,749 reversed sequences. The initial precursor mass tolerance was set to 7 ppm for Orbitrap data (full scans); fragment ion mass tolerance was set to 0.5 Da for ion trap data (MS/MS scans). Full trypsin specificity was required, and up to two missed cleavages were allowed. Carbamidomethylation on cysteine was defined as fixed modification; methionine oxidation, protein N-terminal acetylation, and phosphorylation on serine, threonine, and tyrosine were defined as variable modifications. The database search results were parsed by MaxQuant to assemble protein groups, peptides, and phosphorylation sites at a false discovery rate of 1%. All phosphorylation events having a reported localization probability of at least 0.75 were considered as localized (assigned to a specific amino acid). Subsequent downstream analysis of the result tables was done in R v2.11.1 (13).

Determination of Orthologous and Homologous Relationships—Pairwise orthologs and homologs between *P. pacificus* and *C. elegans* were inferred using bidirectional and unidirectional BLASTP, respectively (14, 15). We used Wormbase WS200 for *C. elegans* and the latest genome annotation for *P. pacificus* (7) as input. Global alignments between orthologous proteins were derived using Needle (16, 17).

Determination of Orphan/Pioneer Proteins—Orphan proteins were defined by two BLAST analyses. First we regarded every *P. pacificus* protein having no homologue in the NCBInr database (BLASTP E-value $< 1 \times 10^{-3}$) as a potential orphan. Second, we used the information derived from the pairwise BLAST analysis of the theoretical proteomes of *P. pacificus* and *C. elegans* as described above. Orphan proteins were required to have no homologues in the NCBInr database or in the Wormbase WS200.

Functional Annotation of the P. pacificus Proteome—Blast2GO software was used to derive Gene Ontology (GO) (18) terms via a BLAST search of the theoretical proteome of *P. pacificus* against the nonredundant NCBI protein database (downloaded on April 29, 2010) using default parameters. Information on specific pathways on the basis of Kyoto Encyclopedia of Genes and Genomes (KEGG) terms (19) was obtained from the KEGG Automatic Annotation Server (20) using default parameters. The classification of proteins into protein families was performed using Pfam (21). The significance E-value threshold was gathered by the software automatically. All types of annotation were merged and exported to an Excel sheet using R.

Functional Enrichment Analysis of the Detected Phosphoproteome—The frequencies of functional annotation terms assigned to the detected phosphoproteome were tested against the corresponding frequencies in the entire proteome using Fisher's exact test (one-sided). A minimum of five occurrences of each term was required in order for the term to be taken into account for analysis. Derived *p* values were further adjusted for multiple hypothesis testing using the method proposed by Benjaminii and Hochberg (22).

Draft Kinome Annotation—We considered all proteins having predicted Pfam domains "Pkinase," "Pkinase_C," or "Pkinase_Tyr" as potential kinases. In order to classify these kinases into kinase groups, families, and subfamilies, we performed a BLAST search of predicted kinase domains against all nematode-specific kinase domains contained in Kinbase. BLAST hits were considered significant if the reported E-value was below 1×10^{-20}, resulting in a minimal bit score of 90.9. For further validation, we did a second BLAST search by querying the kinase domains contained in Kinbase against the predicted Pfam domains in the *P. pacificus* proteome and checking whether the results were consistent. All predicted Pfam domains that met these criteria were classified according to Kinbase annotation.

Phylogenetic distances between the domains were estimated by ClustalW and exported to Nexus format. Distances were logarithmized and imported into the Interactive Tree of Life online tool (23) to produce the phylogenetic trees. The trees were annotated with kinase groups using the classification obtained by the BLAST analysis described above.

Secondary Protein Structure Prediction—The secondary structures of all phosphorylated proteins detected in *P. pacificus* and *C. elegans* were calculated using PsiPred v3.3 (24) and PSIBLAST v2.2.23. Initial PSIBLAST searches were done against the Unrirefr90 database. Prior to the search, low complexity regions were removed from that database as described in the README file of the Psipred software.

Comparison with Transcriptome Data—The gene expression data from a dauer versus mix-stage comparison in both *C. elegans* and *P. pacificus* were obtained from Sinha *et al.* (25). For *P. pacificus*, the gene predictions and, hence, the gene identifiers used in Ref. 25 are different from those used in Ref. 7, although the underlying genome assembly is the same. Thus, the mapping from a microarray probe to a gene prediction corresponding to Ref. 7 was calculated using stringent BLAST criteria (E-value $< 1 \times 10^{-10}$, 100% identity between the 60 bp microarray probe and the target gene). Probes that matched multiple genes were removed from the analysis, and fold-changes were calculated using the same parameters and methods as in Ref. 25. Pfam domain annotations of *C. elegans* were based on wormpep-210. We used kinase domain annotation for *P. pacificus* from supplemental Table 2. The average expression values of all the kinase genes (expression ratio "Dauer/Mix-stage") for all genes annotated with a particular kinase domain ("Pkinase" or "Pkinase_Tyr") were compared within species, and the significance of the difference was assessed based on two-sample Wilcoxon tests. The number *n* in Fig. 5 is the total number of genes belonging to a particular gene family. The "average expression" is defined as log2(RedSignal * GreenSignal) on an arbitrary scale. Hence the values can be compared only within a nematode species and should not be compared across nematodes. The *P. pacificus* and *C. elegans* fold-change and average expression data on kinases are included in supplemental Table 5.

RESULTS

In this study we aimed to provide the reference phosphoproteome of the nematode model *P. pacificus* and compare it to the recently published phosphoproteome of *C. elegans* (8). To minimize experimental bias and enable direct comparison between the datasets, we employed similar sample preparation, measurement, and data processing workflows as in the phosphoproteomic study of *C. elegans*. Briefly, we lysed a well-fed mixed stage *P. pacificus* culture by rupturing the cuticle with freeze-thaw cycles and glass bead treatment. We extracted the proteins from the insoluble fraction in 4% SDS and processed them via the FASP protocol as described by Wisniewski *et al.* (9). We digested the soluble protein fraction in solution with trypsin and separately subjected both fractions to two stages of phosphopeptide enrichment, consisting of strong cation exchange and TiO_2 chromatographies (26, 27). We performed LC-MS analysis on an Easy-LC (Proxeon Biosystems) coupled to an LTQ-Orbitrap XL MS (Thermo Fisher Scientific) and processed the data using the MaxQuant software suite (12). The workflow employed in this study is depicted in Fig. 1.

Detected Phosphoproteome of P. pacificus—The analysis of the *P. pacificus* phosphoproteome resulted in 60,358 identified MS/MS spectra that detected 9,872 nonredundant pep-

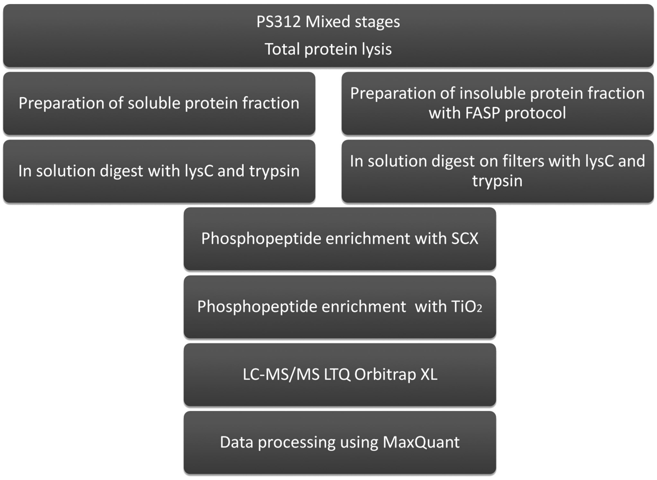

PS312 Mixed stages
Total protein lysis
Preparation of soluble protein fraction
Preparation of insoluble protein fraction with FASP protocol
In solution digest with lysC and trypsin
In solution digest on filters with lysC and trypsin
Phosphopeptide enrichment with SCX
Phosphopeptide enrichment with TiO_2
LC-MS/MS LTQ Orbitrap XL
Data processing using MaxQuant

FIG. 1. **Biochemical workflow used in this study.** A mixed population of *P. pacificus* worms was harvested, and the protein extract was split into soluble and insoluble fractions, which were processed further using the FASP protocol. After LysC/trypsin digestion, phosphopeptides were enriched by SCX and TiO2 chromatographies and measured on an LTQ Orbitrap XL mass spectrometer.

TABLE I
Number of (phospho)proteins detected in this study (at 1% false discovery rate (FDR))

MS data were searched against a decoy database containing *P. pacificus* and *E. coli* protein entries.

	Phosphoproteins	All proteins
P. pacificus	2,508	3,158
E. coli	11	23

tide sequences with a median absolute mass deviation of 255 ppb (supplemental Fig. 1). We detected 3,158 *P. pacificus* protein groups at a false discovery rate of 1%; of these, 2,508 were phosphorylated (Table I) and 1,518 were not detected in our previous large-scale proteomics study (7). This resulted in extension of the catalogue of *P. pacificus* proteins detected by MS to 5,547 (supplemental Fig. 2). In total, we localized 6,809 phosphorylation events to a specific amino acid residue with a median confidence level of 99.8%. The frequencies of phosphorylated serines, threonines, and tyrosines were found to be 87.8% (5,981 events), 11.1% (756 events), and 1.06% (72 events), respectively. All detected phosphorylation sites are presented in supplemental Table 1.

Functional Classes and Kinase Motifs of Detected P. pacificus Phosphoproteins—To gain insight into the functional distribution of proteins phosphorylated in *P. pacificus*, we first retrieved the latest functional annotation according to GO terms, KEGG pathways, enzyme commission numbers (ECs), and protein families and domains (Pfam) (supplemental Table 2). We then performed enrichment analyses of the GO, KEGG, Pfam, and EC terms of proteins detected as phosphorylated (supplemental Table 3). The GO term analysis showed an enrichment of functions related to protein and nucleoside binding, transcription repressor activities, and kinase regulator activities, terms commonly enriched in large phosphoproteome datasets. The Enzyme Class analysis showed significant enrichment of only two classes, protein tyrosine kinases (EC 2.7.10.0; 23 detected phosphoproteins) and protein serine kinases (EC 2.7.11.0; 35 detected phosphoproteins). This was expected because kinases and phosphatases themselves are commonly regulated by phosphorylation, and many kinases show autophosphorylation activity. In agreement with this, the Pfam analysis showed an enrichment of protein kinase domains, as well as phosphatase domains. Proteins with domains involved in protein–protein interactions and signaling were also overrepresented in comparison with the total gene predictions. Among the detected domains, WD40, VWD, Ankyrin, and PDZ domains were highly represented. Moreover, RNA binding domains such as rrm-1, helicase, and DEAD were also overrepresented. The results of the functional enrichment analysis are summarized in Fig. 2.

We next tested the representation of *P. pacificus* orphan gene products in the phosphoproteome. In total, we detected phosphorylation on 234 products of orphan genes (9.3% of the detected phosphoproteome). Compared with all orphan genes in the *P. pacificus* genome (9,957; 41.09% of the genome), this presented a significant underrepresentation ($p < 3.64 \times 10^{-303}$). However, it has to be noted that this class of gene products showed a similar underrepresentation at the proteome level (7), pointing to the fact that their underrepresentation in the phosphoproteome results from the lack of expression, not phosphorylation.

Next, we tested the enrichment of specific kinase target motifs on *P. pacificus* phosphoproteins detected in our dataset, as described by Zielinska *et al.* (8). On proteins phosphorylated on serine, three motifs were overrepresented—CAMK2 (RXX[pS]), CK2 ([pS]XXE), and PKA (RX[pS])—whereas on proteins phosphorylated on threonine, only the CAMK2 motif was overrepresented. For both phosphorylated residues there was also significant overrepresentation of proline adjacent to the phosphorylation site ([pS]P and [pT]P) (supplemental Fig. 3). No significant motifs were detected on proteins phosphorylated on tyrosine residues, most likely because of the small size of the dataset.

Comparison of P. pacificus and C. elegans Phosphoproteomes—We next compared the phosphoproteome of *P. pacificus* to the recently published phosphoproteome of *C. elegans* (8), in which 6,699 phosphorylation sites were localized on 2,365 proteins (Table II). The sizes of the two phosphoproteomes were very similar, and the enriched functional classes of detected phosphoproteins were almost identical, demonstrating that both nematodes likely use protein phosphorylation in similar biological processes. Interestingly, the two phosphoproteomes differed in frequencies of S/T/Y phosphorylation events. In *C. elegans*, the reported pSer, pThr, and pTyr frequencies were 80.2%, 18%, and 1.8%, whereas in *P. pacificus* they were 87.8%, 11.1%, and 1.1%, respectively.

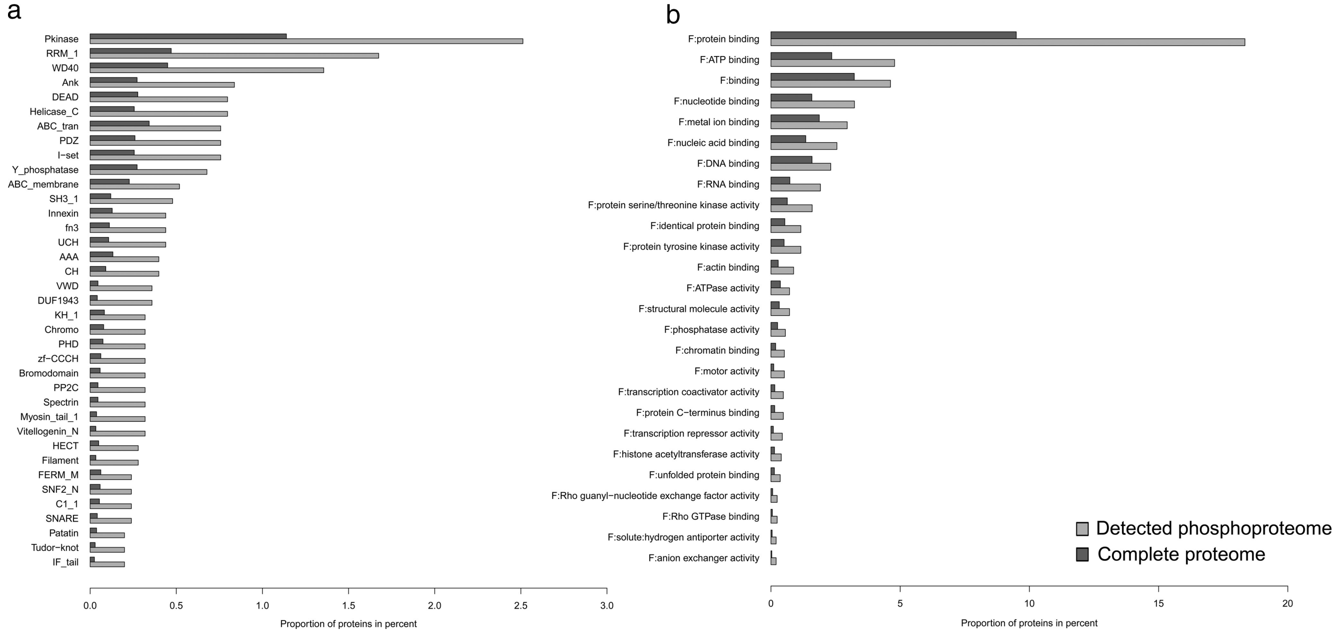


FIG. 2. **Functional enrichment analysis of the detected *P. pacificus* phosphoproteome.** *a*, enrichment of Pfam terms; *b*, enrichment of GO terms (molecular function).

TABLE II

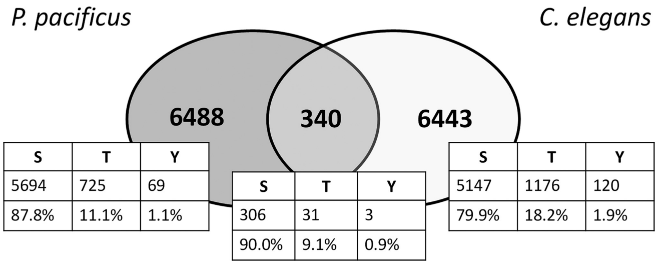
Numbers and frequencies of phosphorylation sites localized on serine, threonine, and tyrosine in P. pacificus and C. elegans

To test whether the frequencies of pS, pT, and pY were significantly different between the two nematodes, we calculated *p* values using a two-sided binomial test.

	Total	pS	pT	pY
Wormbase200	23,973 proteins	7.81%	5.85%	2.75%
Ppa database	24,231 proteins	8.12%	5.89%	3.13%
C. elegans (Zielinska et al. (8))	6,699 (2,365 proteins)	5,372 (80.19%)	1,207 (18.02%)	120 (1.79%)
P. pacificus (this study)	6,809 (2,401 proteins)	5,981 (87.84%)	756 (11.1%)	72 (1.06%)
Binomial *p* value		$p < 2.2 \times 10^{-16}$	$p < 2.2 \times 10^{-16}$	$p < 1.18 \times 10^{-6}$

The frequencies of all phosphorylated amino acids were significantly different despite very similar overall frequencies of these amino acids in the proteomes of *P. pacificus* and *C. elegans* (Table II).

To gain insight into the potential origin of this discrepancy, we investigated the frequencies of pSer, pThr, and pTyr in orthologs shared between *P. pacificus* and *C. elegans* and therefore likely present in their common ancestor. Based on the bidirectional BLASTP approach, 619 phosphoproteins from our dataset were defined as orthologs between *P. pacificus* and *C. elegans* and phosphorylated in both species. On these orthologs, 340 phosphorylation sites were determined as conserved (Fig. 3; supplemental Table 4). Interestingly, the frequencies of pSer, pThr, and pTyr at the ortholog level (90%, 9.1%, and 0.9%, respectively) resembled more closely the frequencies measured in the phosphoproteome of *P. pacificus* than those in the phosphoproteome of *C. elegans*. This means that the basal phosphoproteome of *P. pacificus* might resemble the phosphoproteome of the common ancestor of *P. pacificus* and *C. elegans*.



S	T	Y
5694	725	69
87.8%	11.1%	1.1%

S	T	Y
306	31	3
90.0%	9.1%	0.9%

S	T	Y
5147	1176	120
79.9%	18.2%	1.9%

FIG. 3. **Evolutionary conserved phosphorylated residues between *P. pacificus* and *C. elegans*.** Venn diagram depicting the overlap of conserved phosphorylation sites on direct orthologs found to be phosphorylated in the phosphoproteome datasets.

Whereas different frequencies of tyrosine phosphorylation may be explained by different usages of this modification in signal transduction (see Discussion), different frequencies of detected serine and threonine phosphorylation are more difficult to explain, mostly because of the dual specificity of Ser/Thr kinases. A potential reason could be different representation of these amino acids in the unstructured protein regions that are more accessible to protein kinases. To test

TABLE III

Number of predicted protein kinases in different nematodes according to Pfam annotations of protein kinase domains

	Domains	Proteins
C. elegans	441 (119 pTyr)	413 (117 pTyr)
P. Pacificus	408 (102 pTyr)	368 (94 pTyr)
B. malayi	406 (89 pTyr)	378 (83 pTyr)
M. incognita	392 (57 pTyr)	361 (53 pTyr)

this hypothesis, we calculated the frequencies of all serine, threonine, and tyrosine residues in coiled coils and in helical and strand regions of proteins from detected phosphoproteomes of *P. pacificus* and *C. elegans* and compared them to detected phosphorylation sites (supplemental Fig. 4). As expected, this analysis did not reveal any significant differences in the frequencies of serine, threonine, and tyrosine in the two organisms, demonstrating that different accessibility is not the reason for the observed differences in frequencies of phosphorylated amino acids.

Predicted Kinome of P. pacificus and Its Comparison with C. elegans—We next compared the predicted kinomes of several sequenced model nematodes. To define the predicted kinomes of *P. pacificus*, we used Pfam annotation and considered all proteins containing a "P-kinase" domain as potential kinases (see Methods). After collapsing all *C. elegans* kinase isoforms, we compared the predicted kinome to that of *P. pacificus*. The kinome of *P. pacificus* contained 368 kinases (supplemental Table 2) and was 11% smaller than that of *C. elegans*, which contained 413 kinases (Table III); interestingly, the number of predicted tyrosine kinases was 20% lower in *P. pacificus* (94 kinases) and therefore was underrepresented relative to *C. elegans* (117 kinases). Of the 368 predicted kinases in *P. pacificus*, 77 were detected as phosphorylated in our study. Of those, 61 had direct orthologs and 30 were detected as phosphorylated in *C. elegans* (8). Interestingly, two of the three (66.6%) conserved pTyr residues were located on kinases (cdk-1, mbk-1), one of the 31 (3.2%) conserved pThr residues was located on a kinase (sek-1), and 12 of 306 (3.9%) conserved pSer residues were located on kinases (unc-82, unc-22, pkc-1, grk-1, ZK524.4, gcy-28, ZC581.9, B0495.2).

To classify *P. pacificus* kinases into groups, families, and subfamilies, we performed a bidirectional BLAST analysis of predicted kinase domains against *C. elegans* kinase domains contained in Kinbase. The BLAST analysis resulted in 282 highly confident hits, indicating that the catalytic domains of predicted kinases appeared to be conserved between the two nematodes. All eight major protein kinase groups present in *C. elegans* were also present in *P. pacificus* (Fig. 4; supplemental Fig. 5).

Expression of Different Kinase Classes in C. elegans and P. pacificus—To assess the expression of different kinase classes in *C. elegans* and *P. pacificus*, we analyzed a recently published transcriptome dataset that addresses global changes in gene expression in the dauer and mixed populations of these two nematodes (25). Applying our Pfam-based kinome annotation, we extracted expression data for 404 kinases in *C. elegans* (288 Ser/Thr and 116 Tyr kinases) and 316 kinases in *P. pacificus* (239 Ser/Thr and 77 Tyr kinases). As expected, the transcriptome analysis showed good coverage of the kinome in both organisms, albeit slightly higher in *C. elegans* (404/413, 98%) than in *P. pacificus* (316/368, 86%). Interestingly, in *C. elegans*, the average expression of "Pkinase_Tyr" genes was significantly higher than the average expression of "Pkinase" genes in the dauer population. However, in *P. pacificus*, all kinase genes were expressed at a significantly higher level in the dauer population, and there was no difference in average expression between the two kinase categories, pointing to the fact that tyrosine kinases are expressed at levels similar to those of Ser/Thr kinases (Fig. 5). These data reveal that both nematodes express all classes of kinases and point to their potentially different usage in the dauer stage of the life cycle.

DISCUSSION

In this study, we have reported the first global phosphoproteomic dataset of the mixed stage population of the *P. pacificus* nematode. In order to increase the number of identified phosphorylation sites, we performed three biological replicates, two with the soluble and one with the insoluble protein fraction. In this way, we made all cellular compartments accessible to protein analysis. By using mixed stages, we aimed to get an in-depth catalog of phosphorylation sites of *P. pacificus* and compare it to the previously reported phosphoproteome of *C. elegans*, analyzed under similar conditions.

Although the two phosphoproteomes were very similar in terms of size, classes of phosphorylated proteins, and overrepresented kinase motifs, they were different in the extent of serine, threonine, and tyrosine phosphorylation. Interestingly, this difference might reflect the observed alterations in signal transduction during postembryonic development of these two species. Work over the past decade has compared signaling networks during vulva development and dauer formation between *P. pacificus* and *C. elegans* and identified substantial differences (28). In *C. elegans*, three vulva precursor cells (VPCs) are induced to form vulval tissue by a signal from the gonadal anchor cell. This signal is a secreted epidermal-growth factor (EGF)-type factor that is transmitted within the VPCs by EGFR-RAS-MAP kinase signaling and finally results in the initiation of cell division. A series of phosphorylation events by LIN-45/RAF, MEK-2/MAP kinase, and MPK-1/MAP kinase is at the center of *C. elegans* vulva induction (29). In *P. pacificus*, in contrast, vulva formation is regulated by a completely different regulatory mechanism (for a review, see (30)). While the same VPCs form vulval tissue, their induction requires regulatory input from Wnt signaling rather than EGF-MAP kinase signaling (31). This involves an unusual regulatory

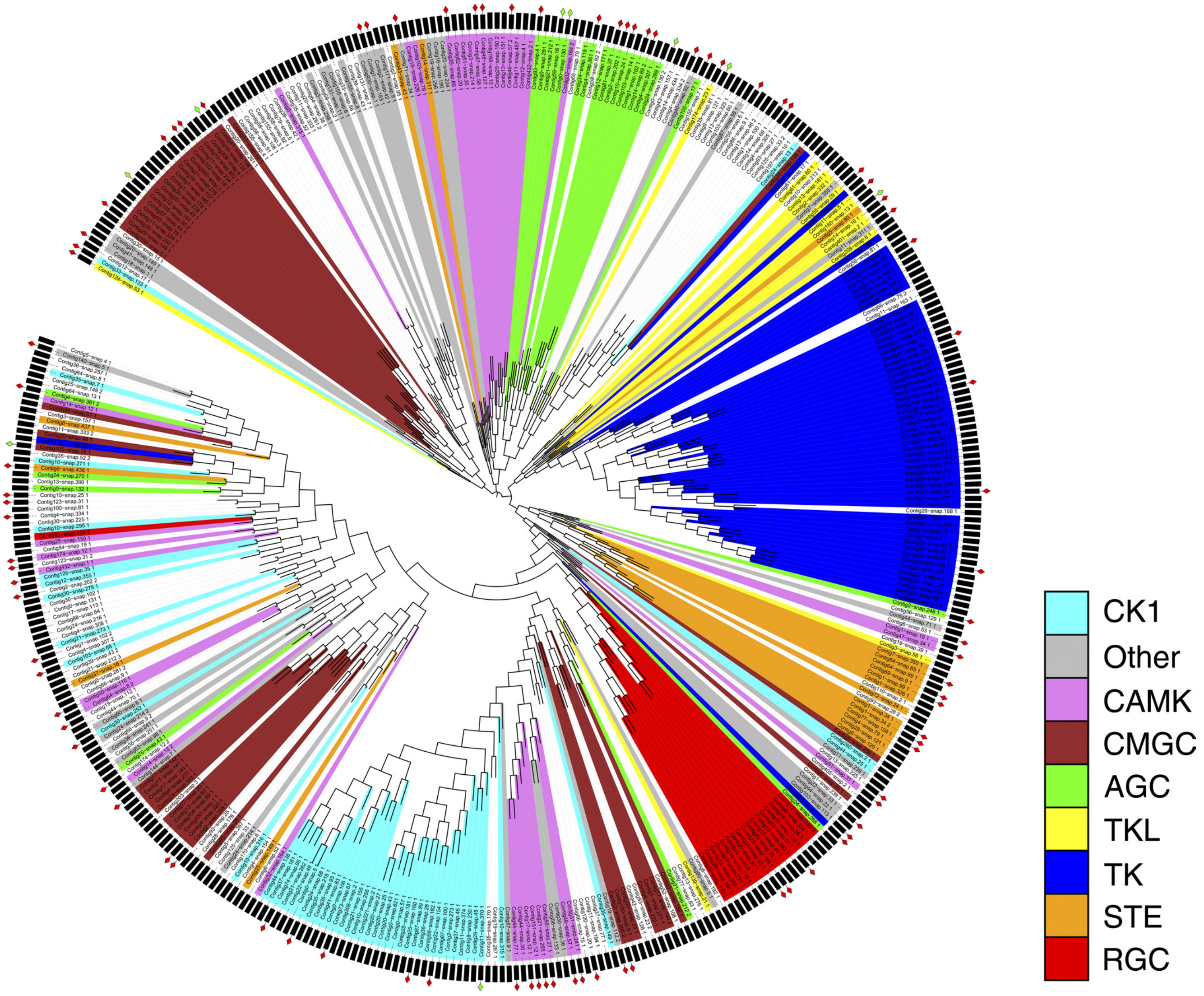


FIG. 4. **Phylogentic tree of the predicted *P. pacificus* kinome.** The tree shows the phylogentic relationships of predicted kinase domains and their classification into kinase groups according to the *C. elegans* kinome. Phylogenetic distances are based on multiple sequence alignments of predicted kinase domains. The classification of domains into kinase groups is shown by the different colors of the branches. Red rectangles at the outer edge of the circle indicate kinases that are detected as phosphorylated; green rectangles indicate kinases that are detected as nonphosphorylated in our study.

linkage of Wnt-type ligands and Frizzled-type receptors, as well as novel protein-interaction domains in LIN-18/Ryk/Derailed-type co-receptor (28). Thus, vulva induction in *C. elegans* is regulated by a kinase pathway involving a high extent of tyrosine phosphorylation, whereas the same process in *P. pacificus* depends much less on tyrosine phosphorylation. It has to be noted, however, that *P. pacificus* contains 1:1 orthologs for all of the EGF/Ras pathway genes/proteins known from *C. elegans*. Interestingly, *Ppa*-MPK-1 was the only kinase of the EGF/RAS pathway shown to be phosphorylated in our dataset. The functional significance of this finding, if any, has yet to be identified.

Similarly, work on dauer formation revealed potential differences in signaling activity during development. In *C. elegans*, the formation of dauer larvae, an arrested alternative life stage that facilitates the survival of harsh environmental conditions, involves insulin and TGF-β signaling activity that is coupled to transcriptional activity of the nuclear hormone receptor DAF-12 and the FOXO-transcription factor DAF-16 (5). In *P. pacificus*, both transcription factors have similar roles during dauer regulation, as indicated by the phenotype of mutations in the corresponding genes, whereas there is no report that would suggest similar roles of insulin and TGF-β signaling (6). However, as indicated above for vulva development, these differences in signaling activity in these two nematodes are not reflected in the copy number of genes encoding signaling components in the respective genomes. Thus, differences in phosphorylation patterns as revealed in our study can occur

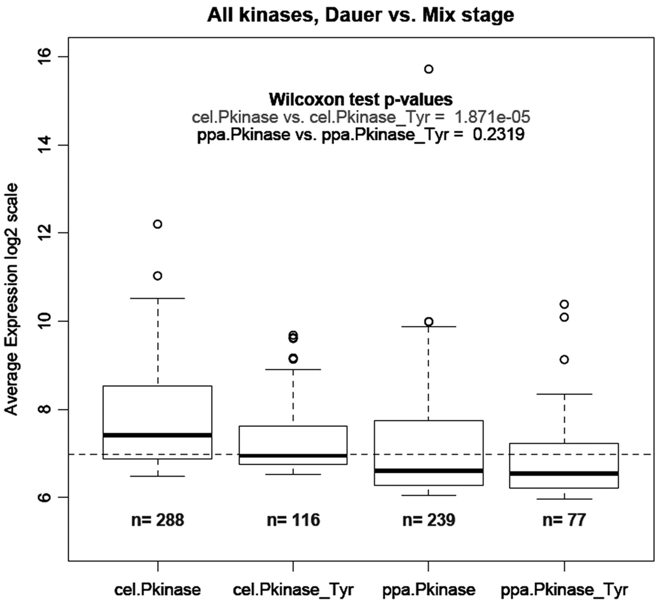


FIG. 5. **Average expression of different protein kinase classes in dauer versus mixed stages of *C. elegans* (cel) and *P. pacificus* (ppa).** Kinome annotation from supplemental Table 2 was applied to quantitative transcriptomics data derived from Sinha *et al.* (25).

in the absence of major changes in the signaling pathways that act during development.

The comparative analysis of the predicted kinomes of *P. pacificus* and *C. elegans* indicates that all major protein kinase groups are conserved between these two nematodes (supplemental Table 2), and recent transcriptome analysis suggests that all kinase classes are expressed (and presumably active) in both nematodes during the dauer stage of the life cycle (25). When compared with other protein classes, the kinome shows a relatively high level of conservation and low copy number variations. For example, many of the detoxification enzymes, such as cytochrome P450 proteins, show a more than 3-fold difference between the *P. pacificus* and *C. elegans* proteomes with 197 and 67 protein predictions, respectively (4). We speculate that the difference in cytochrome P450 enzymes reflects the adaptation to the different environments in which these nematodes are found. In contrast, the overall similarity of the two kinomes represents the conserved molecular and cellular processes, which evolved largely independent of ecological alterations. This evolutionary pattern becomes even stronger when data available for additional nematodes are considered: the numbers of predicted protein kinases of *P. pacificus*, *C. elegans,* the human parasite *B. malayi*, and the plant parasite *M. incognita* are surprisingly similar (Table III). Thus, the kinome represents a stable part of the nematode proteome, and most likely the analysis of a small number of selected model organisms will provide comprehensive insight into processes of phosphorylation.

Acknowledgments—We thank Matthias Mann for access to the Phosida database. We also thank Jonathan Goldberg for fruitful discussions on the kinome prediction.

* This work was supported by funding from Baden-Württemberg Stiftung (Juniorprofessoren-Programm), the German Research Foundation, and PRIME-XS (to B.M.).

S This article contains supplemental Figs. 1 to 5 and Tables 1 to 5.

§ These authors contributed equally to this work.

** Current address: Department of Bioinformatics and Computational Biology, Genentech Inc., San Francisco, CA.

‡‡ To whom correspondence may be addressed: Prof. Dr. Boris Macek, Proteome Center Tuebingen, Interfaculty Institute for Cell Biology, University of Tuebingen, Auf der Morgenstelle 15, 72076 Tuebingen, Germany, Tel.: +49/(0)7071/29-70558, Fax: +49/(0)7071/29-5779, E-mail: boris.macek@uni-tuebingen.de

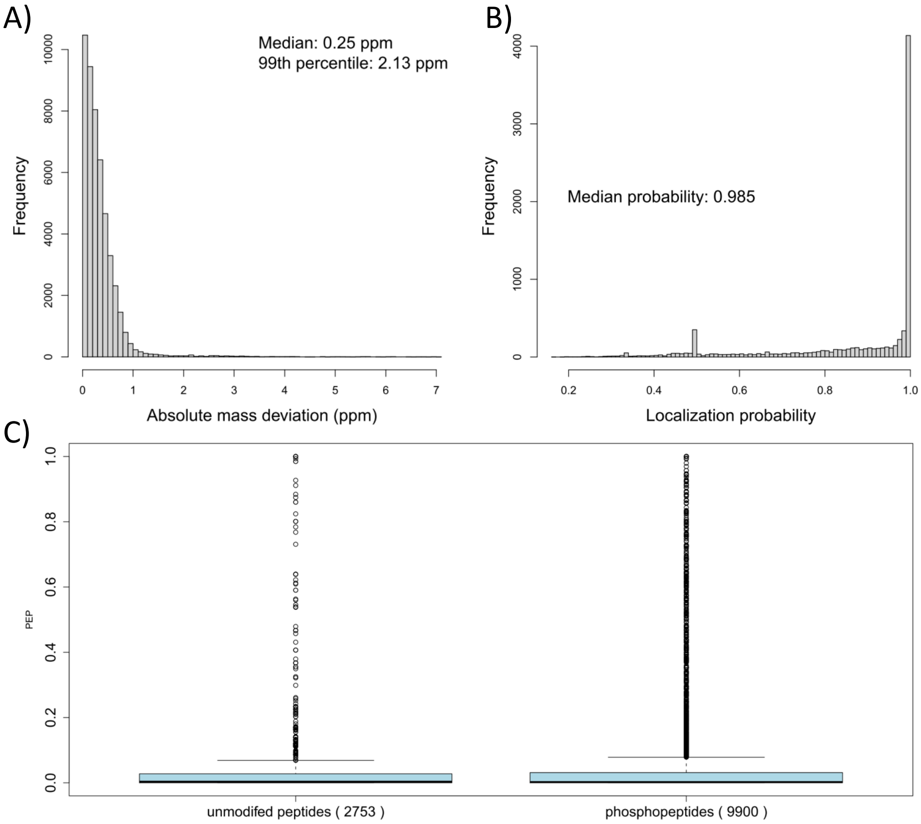
§§ To whom correspondence may be addressed: Prof. Dr. Ralf J. Sommer, Dept. for Evolutionary Biology, Max Planck Institute for Developmental Biology, Spemannstrasse 37, Tuebingen D-72076, Germany, Tel.: +49-7071-601371, Fax: +49-7071-601498, E-mail: ralf.sommer@tuebingen.mpg.de.

REFERENCES

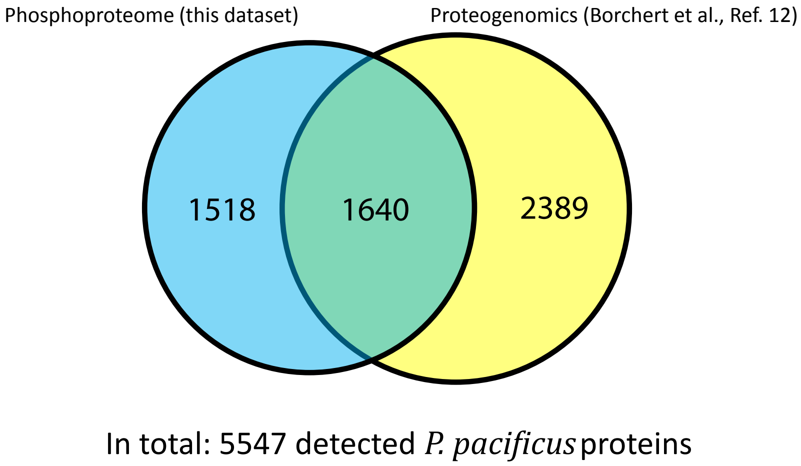
1. Gnad, F., Ren, S. B., Cox, J., Olsen, J. V., Macek, B., Oroshi, M., and Mann, M. (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* **8,** R250
2. Hong, R. L., and Sommer, R. J. (2006) Pristionchus pacificus: a well-rounded nematode. *Bioessays* **28,** 651–659
3. The C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282,** 2012–2018
4. Dieterich, C., Clifton, S. W., Schuster, L. N., Chinwalla, A., Delehaunty, K., Dinkelacker, I., Fulton, L., Fulton, R., Godfrey, J., Minx, P., Mitreva, M., Roeseler, W., Tian, H., Witte, H., Yang, S. P., Wilson, R. K., and Sommer, R. J. (2008) The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. *Nat. Genet.* **40,** 1193–1198
5. Coghlan, A. (2005) Nematode genome evolution. *WormBook* **2005,** 1–15
6. Sommer, R. J., and Ogawa, A. (2011) Hormone signaling and phenotypic plasticity in nematode development and evolution. *Curr. Biol.* **21,** R758–R766
7. Borchert, N., Dieterich, C., Krug, K., Schutz, W., Jung, S., Nordheim, A., Sommer, R. J., and Macek, B. (2010) Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. *Genome Res.* **20,** 837–846
8. Zielinska, D. F., Gnad, F., Jedrusik-Bode, M., Wisniewski, J. R., and Mann, M. (2009) Caenorhabditis elegans has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins. *J. Proteome Res.* **8,** 4039–4049
9. Wisniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat. Methods* **6,** 359–362
10. Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K., and Horning, S. (2006) Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **78,** 2113–2120
11. Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4,** 2010–2021
12. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
13. R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria
14. Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314,** 1041–1052
15. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410
16. Needleman, S. B., and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins.

J. Mol. Biol. **48,** 443–453

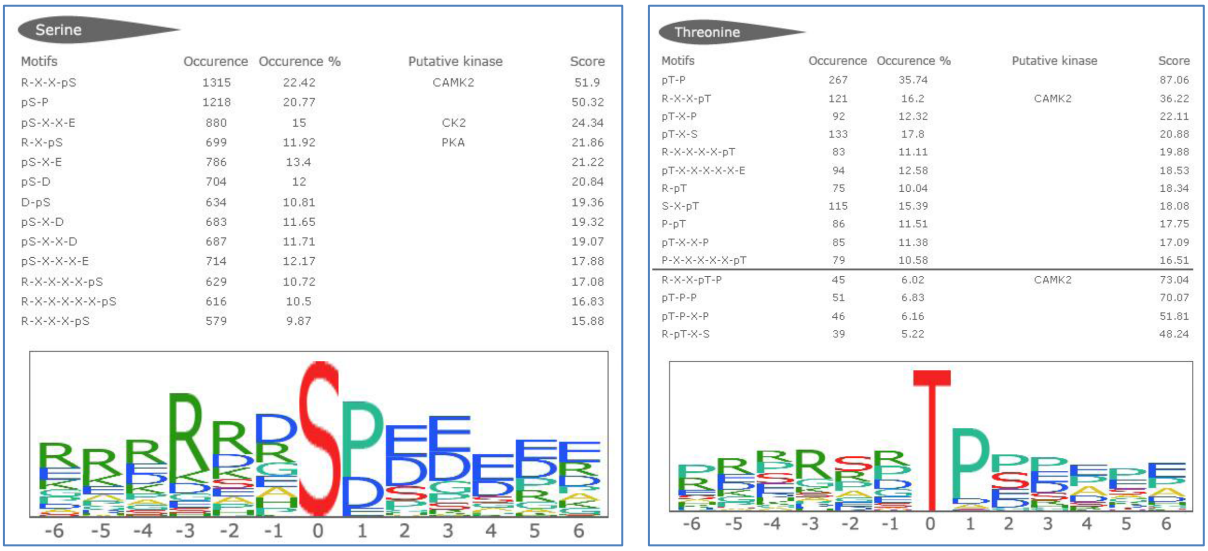
17. Rice, P., Longden, I., and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277
18. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29
19. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34,** D354–D357
20. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35,** W182–W185
21. Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., and Bateman, A. (2010) The Pfam protein families database. *Nucleic Acids Res.* **38,** D211–D222
22. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.* **57,** 289–300
23. Letunic, I., and Bork, P. (2007) Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23,** 127–128
24. Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292,** 195–202
25. Sinha, A., Sommer, R. J., and Dieterich, C. (2012) Divergent gene expression in the conserved dauer stage of the nematodes Pristionchus pacificus and Caenorhabditis elegans. *BMC Genomics* **13,** 254
26. Macek, B., Mann, M., and Olsen, J. V. (2009) Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu. Rev. Pharmacol. Toxicol.* **49,** 199–221
27. Olsen, J. V., and Macek, B. (2009) High accuracy mass spectrometry in large-scale analysis of protein phosphorylation. *Methods Mol. Biol.* **492,** 131–142
28. Wang, X., and Sommer, R. J. (2011) Antagonism of LIN-17/Frizzled and LIN-18/Ryk in nematode vulva induction reveals evolutionary alterations in core developmental pathways. *PLoS Biol.* **9,** e1001110
29. Sternberg, P. W. (2005) Vulval development. *WormBook* **2005,** 1–28
30. Sommer, R. J. (2008) Homology and the hierarchy of biological systems. *Bioessays* **30,** 653–658
31. Tian, H., Schlager, B., Xiao, H., and Sommer, R. J. (2008) Wnt signaling induces vulva development in the nematode Pristionchus pacificus. *Curr. Biol.* **18,** 142–146



Supplementary Figure 1. Technical details of the detected phosphorylation events. A) distribution of measured mass deviations of all identified phosphopeptides; B) distribution of localization probabilities of all phosphorylation events; C) distributions of peptide PEP values of all identified phosphopeptides



Supplementary Figure 2. Additional *P. pacificus* protein identifications derived from the phosphoproteome dataset (compared to Borchert et al.).



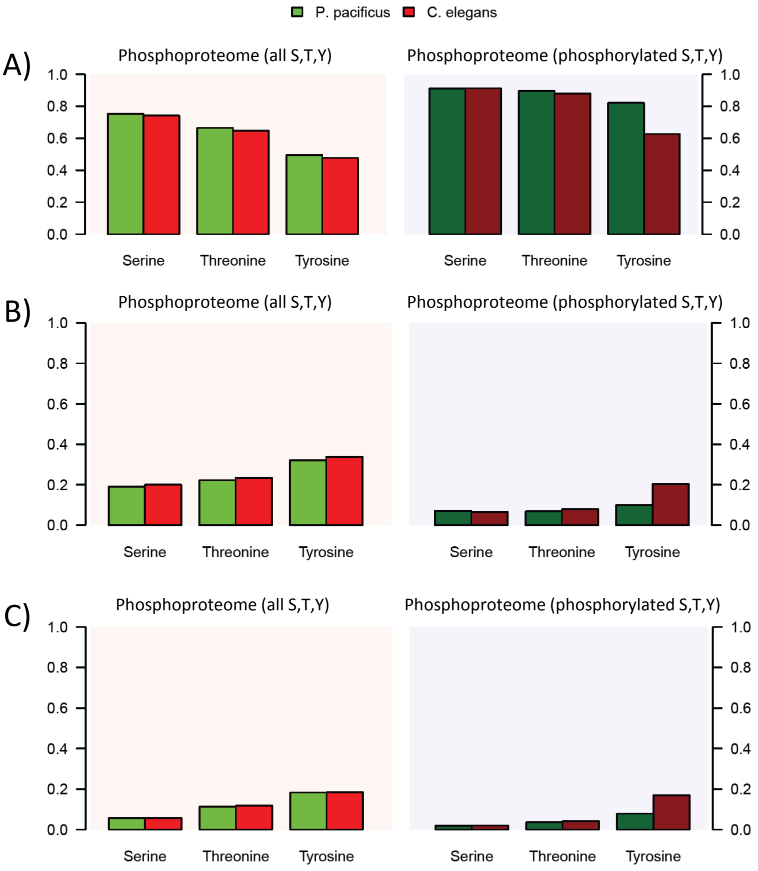
Serine

Motifs	Occurence	Occurence %	Putative kinase	Score
R-X-X-pS	1315	22.42	CAMK2	51.9
pS-P	1218	20.77		50.32
pS-X-X-E	880	15	CK2	24.34
R-X-pS	699	11.92	PKA	21.86
pS-X-E	786	13.4		21.22
pS-D	704	12		20.84
D-pS	634	10.81		19.36
pS-X-D	683	11.65		19.32
pS-X-X-D	687	11.71		19.07
pS-X-X-X-E	714	12.17		17.88
R-X-X-X-X-pS	629	10.72		17.08
R-X-X-X-X-X-pS	616	10.5		16.83
R-X-X-X-pS	579	9.87		15.88

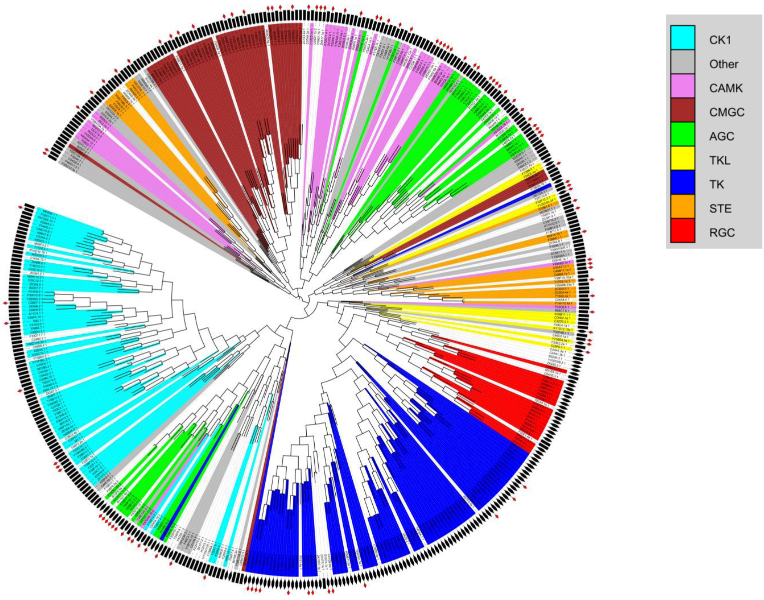
Threonine

Motifs	Occurence	Occurence %	Putative kinase	Score
pT-P	267	35.74		87.06
R-X-X-pT	121	16.2	CAMK2	36.22
pT-X-P	92	12.32		22.11
pT-X-S	133	17.8		20.88
R-X-X-X-X-pT	83	11.11		19.88
pT-X-X-X-X-X-E	94	12.58		18.53
R-pT	75	10.04		18.34
S-X-pT	115	15.39		18.08
P-pT	86	11.51		17.75
pT-X-X-P	85	11.38		17.09
P-X-X-X-X-X-pT	79	10.58		16.51
R-X-X-pT-P	45	6.02	CAMK2	73.04
pT-P-P	51	6.83		70.07
pT-P-X-P	46	6.16		51.81
R-pT-X-S	39	5.22		48.24

Supplementary Figure 3. Overrepresented kinase motifs in P. pacificus phosphoproteome.

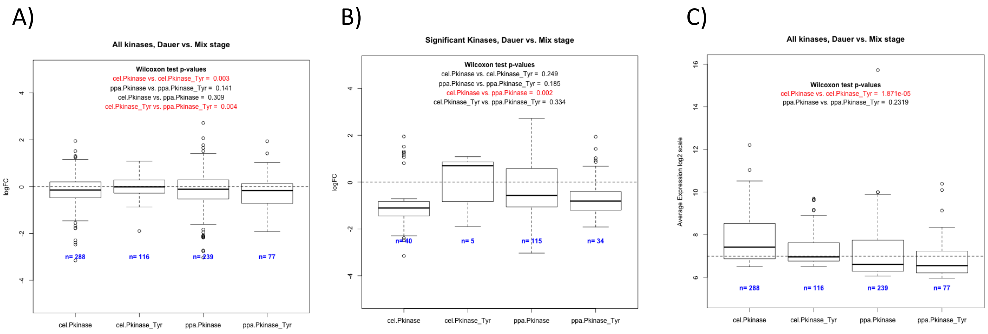


Supplementary Figure 4. Calculated frequencies of all serine, threonine and tyrosine residues in A) coiled coils; B) helical; and C) strand regions of proteins from detected phosphoproteomes of P. pacificus and C. elegans (left panel) and comparison with frequencies detected phosphorylation sites in the same protein regions (right panel)



C. elegans

Supplementary Figure 5. Annotated kinome of *C. elegans*. The kinome tree is based on kinase domains predicted using Pfam. The annotation of kinases is based on kinbase (http://kinase.com/).



Supplementary Figure 6. Expression of different protein kinase classes in dauer vs. mixed stage of *C. elegans* (cel) and *P. pacificus* (ppa). Data are derived from Sinha et.al. A) changes in expression of all annotated kinases from the dataset; B) Changes in expression of significantly regulated kinases from the dataset; C) Changes in average expression of kinases.

Curriculum vitae

Personal information	Name	Karsten Krug
	Date of birth	15 July 1980
	Place of birth	Staßfurt, Germany
Education/work experience	08/2008 - present Staff scientist/PhD student	Proteome Center Tübingen Eberhard-Karls Universität Tübingen Since 01/2009 working on PhD
	04/2008 – 07/2008 Graduate assistant, Genetic Statistics	Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig
	10/2006 – 12/2007 Diploma thesis	Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig
	09/2006 – 03/2008 Student assistant	Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig
	10/2001 - 03/2008 Diploma	Universität Leipzig Informatics/Bioinformatics
	09/1997 – 07/2000 Abitur	Staatliches Gymnasium Johann Gottfried Seume, Vacha
	09/1991 – 07/1997 Mittlere Reife	Realschule Dorndorf

Complete list of publications

1. **Krug K**, Carpy A, Behrends B, Matic K, Soares NC, Macek B
   Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments.
   under revision in *Mol Cell Proteomics*

2. Soares NC*, Spät P*, **Krug K**, Macek B.
   Global Dynamics of the Escherichia coli Proteome and Phosphoproteome During Growth in Minimal Medium.
   J Proteome Res. 2013 May 2. [Epub ahead of print]

3. Borchert N*, **Krug K***, Gnad F, Sinha A, Sommer RJ, Macek B. **Phosphoproteome of Pristionchus pacificus provides insights into architecture of signaling networks in nematode models.**
   Mol Cell Proteomics. 2012 Dec;11(12):1631-9

4. Sessler N, **Krug K**, Nordheim A, Mordmüller B, Macek B.
   Analysis of the Plasmodium falciparum proteasome using Blue Native PAGE and label-free quantitative mass spectrometry.
   Amino Acids. 2012 Sep;43(3):1119-29

5. Franz-Wachtel M*, Eisler SA*, **Krug K***, Wahl S, Carpy A, Nordheim A, Pfizenmaier K, Hausser A, Macek B.
   Global detection of protein kinase D-dependent phosphorylation events in nocodazole-treated human cells.
   Mol Cell Proteomics. 2012 May;11(5):160-70

6. Koch A, **Krug K**, Pengelley S, Macek B, Hauf S.
   Mitotic substrates of the kinase aurora with roles in chromatin regulation identified through quantitative phosphoproteomics of fission yeast.
   Sci Signal. 2011 Jun 28;4(179)

7. Schütz W*, Hausmann N*, **Krug K***, Hampp R, Macek B.
   Extending SILAC to proteomics of plant cell lines.
   Plant Cell. 2011 May;23(5):1701-5

8. Zelenak C, Föller M, Velic A, **Krug K**, Qadri SM, Viollet B, Lang F, Macek B.
Proteome analysis of erythrocytes lacking AMP-activated protein kinase reveals a role of PAK2 kinase in eryptosis.
J Proteome Res. 2011 Apr 1;10(4):1690-7

9. **Krug K**, Nahnsen S, Macek B.
Mass spectrometry at the interface of proteomics and genomics.
Mol Biosyst. 2011 Feb;7(2):284-91

10. Borchert N*, Dieterich C*, **Krug K***, Schütz W, Jung S, Nordheim A, Sommer RJ, Macek B.
Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models.
Genome Res. 2010 Jun;20(6):837-46

11. Muthreich N, Schützenmeister A, Schütz W, Madlung J, **Krug K**, Nordheim A, Piepho HP, Hochholdinger F.
Regulation of the maize (Zea mays L.) embryo proteome by RTCS which controls seminal root initiation.
Eur J Cell Biol. 2010 Feb-Mar;89(2-3):242-9

* = equal contribution