GAPP: A Proteogenomic Software for Genome Annotation and Global Profiling of Post-translational Modifications in Prokaryotes*

Jia Zhang‡¶, Ming-kun Yang‡¶, Honghui Zeng§, and Feng Ge‡§¶

Although the number of sequenced prokaryotic genomes is growing rapidly, experimentally verified annotation of prokaryotic genome remains patchy and challenging. To facilitate genome annotation efforts for prokaryotes, we developed an open source software called GAPP for genome annotation and global profiling of post-translational modifications (PTMs) in prokaryotes. With a single command, it provides a standard workflow to validate and refine predicted genetic models and discover diverse PTM events. We demonstrated the utility of GAPP using proteomic data from Helicobacter pylori, one of the major human pathogens that is responsible for many gastric diseases. Our results confirmed 84.9% of the existing predicted H. pylori proteins, identified 20 novel protein coding genes, and corrected four existing gene models with regard to translation initiation sites. In particular, GAPP revealed a large repertoire of PTMs using the same proteomic data and provided a rich resource that can be used to examine the functions of reversible modifications in this human pathogen. This software is a powerful tool for genome annotation and global discovery of PTMs and is applicable to any sequenced prokaryotic organism; we expect that it will become an integral part of ongoing genome annotation efforts for prokaryotes. GAPP is freely available at https://sourceforge.net/projects/gappproteogenomic/. Molecular & Cellular Proteomics 15: 10.1074/mcp.M116.060046, 3529–3539, 2016.

Thousands of prokaryotic genomes have been sequenced with the rapid advance of massively parallel sequencing technologies. Protein-coding gene annotation is a prerequisite for understanding the functional elements of the genome. Previously, conventional genome annotation was mainly dependent on computational ab initio algorithms to predict coding regions and gene models from raw sequencing data (1, 2). Computational ab initio algorithms, which are hard to avoid errors such as missing genes and incorrect gene boundaries, are limited in their prediction accuracy. Recently, the use of information from RNA-seq or expressed sequence tag libraries has dramatically improved the genome annotation confidence (3). However, abundance of transcripts does not necessarily reflect cellular protein levels, and the correlation between protein and mRNA levels is generally slight (4–6). The existence and function of many gene products remain to be confirmed. Thus, MS has emerged as an invaluable high-throughput method for accurate genome annotation, a concept called "proteogenomics" (7–9).

Proteogenomics is defined as the use of proteomic data, often derived from MS, to improve and refine genome annotation (10, 11). Pioneering work by Yates et al. (12) and Jaffe et al. (13) opened new avenues to high-throughput gene annotation. In recent years, proteogenomics has emerged as a promising and indispensable approach to genome annotation (10, 11). It has been applied for genome annotation including identification of novel genes, correction and validation of predicted genes in various organisms (14–21). Proteogenomics has become particularly important for prokaryotes where genomes are sequenced on a daily basis and in silico gene annotations are prone to errors (22, 23). It has been successfully applied to re-annotate bacterial and archaeal genomes and has provided invaluable information for accurate genome annotation of these prokaryotes (18, 24–36). Besides improving genome annotations, proteogenomic studies can provide valuable information about the presence of post-translational modifications (PTMs) on the proteome-wide level (36–38). PTMs are a vital cellular control mechanism, which regulates every protein property such as folding, conformation, activity, and functions (39). Almost all the proteins undergo certain amount of PTMs, and this reversible process occurs in most cell compartments to determine the functions of modified protein (40, 41). More than 600 different types of PTMs have been reported by the RESID database, and many more are

From the ‡Key Laboratory of Algal Biology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China; §Wuhan Branch, Supercomputing Center, Chinese Academy of Sciences, China

Received April 7, 2016, and in revised form, September 8, 2016

Published, MCP Papers in Press September 14, 2016, DOI 10.1074/mcp.M116.060046

Author contributions: F.G. designed research; J.Z. and M.Y. performed research; H.Z. contributed new reagents or analytic tools; J.Z., M.Y., and F.G. analyzed data; J.Z., M.Y., and F.G. wrote the paper.

© 2016 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available online at http://www.mcponline.org
still being identified (42). Identification of PTMs in a protein has improved dramatically, mainly because of rapid advancement of MS technology (43, 44). However, proteome-wide identification of PTMs remains a highly challenging task and is a much needed analytical tool (45). This situation is exemplified by very few software tools being developed to comprehensively annotate PTMs events at the proteomic level (40).

Over the last few years, computational proteomics has become a dramatically growing field, and a handful of tools have been developed to execute complete proteogenomic analyses (11, 46, 47). Two excellent reviews have described a comprehensive overview of the various problems commonly encountered and their current solutions for this growing research area (8, 11). A number of automated software, including several visualization and database-building tools, have been developed for integration of MS-based proteomic data into genome databases (48–59). However, there is still a lack of automated software for proteogenomic analyses that incorporate both genome annotation and proteome-wide PTM analysis (11). To fill this gap, we developed a one-stop open source software termed GAPP, which provides a complete proteogenomic pipeline for carrying out both genome annotation and large-scale PTM analysis on a proteome-wide level against prokaryotes. It is an open source and publicly comprehensively annotate PTMs events at the proteomic level (40).

EXPERIMENTAL PROCEDURES

Data and Databases—All MS data for proteogenomic analysis on *H. pylori* were obtained from PRIDE proteomics data repository (63) with the data set identifier PXD000054. A total of six different data sets that included two biological replicates were used (14). The genome and all annotated protein sequences were obtained from the NCBI ftp site (NCBI reference sequences NC_009115.1; 11.07.2012). In total, 1,102,122 MS/MS spectra were selected for further analysis. Experimental Design and Statistical Rationale—Development of GAPP—We developed a software called GAPP, an automated standalone pipeline for in-depth proteogenomic analysis and PTM discovery in prokaryotes. Three pregenerated search algorithms XITandem, MSAmanda, and MSGF have been integrated into the GAPP tool for proteogenomic analysis. The unrestricted search algorithm MODification via alignment (MODa) is already integrated as a part of the standard distribution for PTM discovery. Moreover, integration with the results from other existing search algorithms is easily assembled into this tool. It is highly configurable, and any identification results from existing frameworks can be used for further process. GAPP tool uses common file formats (such as FASTA, mascot generic format (MGF), GFF, and TXT) for configuration, and most outputs are in the CSV format, making it easy for other programs to read and write the configuration files. All inputs for the GAPP tool are defined in a configuration file. The core software was developed in Java and Perl, and the executables are distributed for Windows platforms. The source code is free for noncommercial use and available at https://sourceforge.net/projects/gappproteogenomic/.

MS Data Analysis—Proteins were identified by database searching using the GAPP tool with several parameters: (1) two missed cleavages were allowed for trypsin and LysC, whereas three were set for AspN; (2) precursor ion mass tolerance was 5 ppm; (3) fragment ion mass tolerance was 0.5 Da; and (4) fixed PTM was carbamidomethylation of cysteine residues. For parameters not common across search algorithms, we used the search algorithm defaults. The unrestricted database search algorithms, MODa (MODification via alignment) (64) were integrated into GAPP to find all known and even possibly unknown types of PTMs at once, allowing for arbitrary modification with mass shift up to 250 Da per peptide. The parameters were set as follows: (1) PPTolerance was 5 ppm; (2) Frag-Tolerance was 0.5 Da; (3) BlindMode was 2, which allows arbitrary modifications per peptide; (4) MinModSize was −250 and maxModSize was +250; (5) two missed cleavages were allowed for Trypsin and LysC, whereas three were set for AspN; and (6) fixed post-translational modification was carbamidomethylation of cysteine residues. Peptide spectrum matches (PSMs) identified with a 1% FDR threshold were considered using anal_moda.jar for further PTMs analysis. After confirming modification type, the MaxQuant computational proteomics platform (version 1.3.0.5) was used to further search against the protein database for determining the localization of modifications to specific amino acids in the modified peptide sequence as previously described (65). The estimated FDR thresholds for modification site, peptide and protein were specified at maximum 1%.

RESULTS

Overview of GAPP—GAPP is designed as an automated pipeline that contains two modules: (1) Module I (genomicRef) allows MS/MS data to be analyzed against the genome and proteome and (2) Module II (annoMod) is designed to discover PTM events. GAPP is easy to configure and run in one command line. The main input data type accepted is MGF (mascot generic format) format, which can be easily converted by msconvert tool (66) in this pipeline, and the output format...
includes CSV. The workflow and modules are outlined in Fig. 1. MS raw data files as well as processed data files are publicly available through the repository PeptideAtlas (www.peptideatlas.org) (data ID: PASS00860).

Module I (genomicRef)—All automated MS data analyses including peptide and protein identification, and annotation are performed by Module I. Three publicly available search algorithms, namely X!Tandem (67), MS Amanda (68), and MSGF (69), have been deployed within this step to improve the sensitivity for peptide identifications. All search criteria can be specified by the user using a configuration file. In addition, the results from other existing search algorithms can be integrated into this pipeline. All identified peptides from different search algorithms were then mapped onto the six-frame translated genome database and the existing predicted protein database using BLASTP. Finally, a set of unique orphan peptides that were not represented among the existing predicted proteins, were identified. These novel peptides, mapped exclusively to unique locations on the genome, were designated as genome search-specific peptides (GSSPs) as previously reported (65, 70, 71). Because all the results from the existing search algorithms were combined to improve the sensitivity for peptide identifications, it is conceivable that a spectrum would be matched to multiple different peptide sequences that identified by different search algorithms. Therefore, these peptides were not selected for further analyses. Moreover, the purpose of this pipeline was to annotate genomic loci as completely as possible, a peptide mapped to different genomic loci, namely a shared peptide, was also discarded in analogy to the strategy used in a previous publication (72).

For estimation of identification error rates, GAPP used the target-decoy search strategy to calculate FDR for all employed algorithms’ results (73–75). Only peptides that pass the FDR threshold (1%) from multiple search algorithms can be merged for identifying credible annotated peptides. Because of nonuniform distribution of false positives between annotated and novel peptides, the actual FDR of novel peptides could be much higher than the combined FDR estimation (1%). To further improve the identification accuracy of novel peptides, GAPP calculated the FDR using a more stringent filtering strategy, namely separated FDR (76, 77). The FDR for novel peptides was estimated according to the following equation:

\[
FDR_n = \frac{D^+ \times D_o}{T_n^+}
\]

(Eq. 1)

where \(D^+\) is the number of identified decoy peptides with scores above the score threshold, \(T_n^+\) is the number of identified novel peptides in the target database above the score...
Proteogenomic Software for Prokaryotic Genome Annotation

threshold, \(D_n\) is the number of identified decoy novel peptides, and \(D\) is the total number of identified decoy peptides. \(D_n/D\) is an approximation for the fraction of novel sequences in the search space. On the basis of stringent FDR control (FDR\(_n\) < 1%), the list of highly reliable GSSPs was generated by GAPP.

Once peptide hits generated by multiple search algorithms are directly mapped back to their original unique genomic locations, open reading frame (ORF) predictions will be performed by investigating the identified peptide sequence, peptide score, peptide length, and possible presence of start codons. Moreover, the proteins identified with only a single unique peptide and shared peptide were then not considered for further analysis. Finally, the ORFs will be categorized: (1) mapping to an intergenic region, (2) mapping to a different reading frame region with an annotated gene region, (3) mapping to an opposite strand to an existing annotated gene region, and (4) partially overlapping an annotated gene’s N terminus. In addition, the GFF file (General Feature Format) containing all novel peptides will also be provided for further visualization in genome browsers (such as the UCSC Genome Browser and the Integrative Genomics Viewer). Additional functionality for novel protein sequence alignment and gene ontology (GO) functional annotations are added to the pipeline. The identified novel genes were first mapped onto the Uniref database (ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref50/uniref50.fasta.gz) via blastp and the E-value for each gene with a cutoff score (E-value < 0.00001) was selected. GAPP then assigned the most suitable functional labels to these novel genes by mapping to existent annotation associations according to the GO function database (http://archive.geneontology.org/latest-lite/go_20160326-seqdb.fasta).

Module II (annoMod)—To identify unexpected PTM types at once on the proteome-wide level, we used unrestricted or blind approaches to analyze mass spectra. MODa (MODification via alignment) algorithms were integrated into GAPP to search high-throughput MS data against proteome database, allowing for arbitrary modification with mass shift up to 250 Da per peptide. Moreover, to assign the locations of modifications to specific amino acids, the specific PTM searches are performed by MaxQuant to further validate the PTM events.

Identification of Novel Protein-Coding Regions in \(H.\) pylori—To demonstrate the utility and scalability of GAPP, we benchmarked GAPP on a proteogenomic spectral data set from a publicly available, high-throughput MS/MS data set on \(H.\) pylori. All 1,102,122 spectra from six different data sets were searched against the \(H.\) pylori genome database. After eliminating the 507 peptides mapped to multiple genomic locations, a total of 511,560 peptide–spectrum matches corresponding to 35,487 unique peptides were identified with 1% FDR threshold. The whole peptides identified from the proteogenomic analysis of \(H.\) pylori are listed in supplemental Table S1.

Notably, about 84.9% (1248) of the existing predicted \(H.\) pylori proteins were identified with at least two unique peptides, indicating the high coverage of protein identification. More importantly, we identified a set of unique orphan peptides (302) that were designated as GSSPs and mapped to unique locations on \(H.\) pylori genome. Out of these peptides, 171 peptides either mapped to regions of the genome where no gene was annotated or did not match the gene model to which they were mapped. In total, we identified 20 novel protein-coding genes and modified four existing gene models in \(H.\) pylori, all of which were assigned by at least two unique GSSPs (supplemental Table S2 and supplemental Data S1). All annotated spectra for GSSPs were submitted to MS-Viewer (78) (http://msviewer.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msviewer) with the search key: lijxwvmvb3 (X!Tandem), 925qqmgeso (MSGF+) and e6bair7m6k (MS-Amanda). In contrast to the report by Stephan et al. (14), we observed more novel protein-coding genes using GAPP on the \(H.\) pylori genome. A graphical representation of all the GSSPs, novel protein-coding genes, and corrected proteins is presented in Fig. 2A. The length of 16 novel proteins was < 300 amino acids, and the average length of these proteins was 145, suggesting that most novel protein-coding genes are encoded by short ORFs (supplemental Fig. S1).

Among of these novel protein-coding genes, 17 were intergenic and 3 were on different frame. Because the presence of orthologous genes in other species will give a powerful evidence for the novel protein-coding gene models, 11 novel proteins had been previously annotated in other species. Additionally, further functional annotation was performed by assigning the identified novel protein-coding genes to biological process, molecular function, and cellular component categories. As shown in Fig. 2B, we found that 8 (40%) novel proteins were completely annotated according to their biological processes and molecular functions, implying the existence and importance of the identified novel proteins in \(H.\) pylori. Fig. 3A shows an example of a novel protein-coding protein identified with 15 peptides mapping to an intergenic region of the genome. Moreover, one of novel protein-coding genes mapped onto different reading-frame regions is present in supplemental Fig. S2. The GSSPs mapped upstream of annotated gene models on the same frame and strand indicate the changes in annotated gene models. Fig. 3B depicts an example of a correction to a gene model via extension of the N terminus. Two peptides map upstream to the currently annotated gene model for locus HP1302, which is also annotated in \(H.\) pylori HPAG1 as 30S ribosomal protein. Details of all novel genes and corrected gene models identified by GAPP are listed in supplemental Table S3.

Global PTM Discovery in \(H.\) pylori—The proteogenomic study can not only be employed to reannotated genomes but also be used to obtain PTM events (79, 80). Recently, we
reported a proteogenomic analysis and global profiling of PTMs in model cyanobacterium Synechococcus sp. PCC 7002, providing a holistic view of PTM events in this model cyanobacteria (65). In the present study, the MODa algorithm was first integrated into the proteogenomic pipeline for analyzing the mass spectra and identifying PTMs at once (81). As expected, we identified 20,057 unique peptides from 1308 proteins by MODa. Among these, 56,220 spectra of 7543 unique peptides and 1083 proteins contained PTMs with sizes accepted by the MODa search regardless of the modification classification in Unimod database (www.unimod.org). Details of all the PTMs are provided in supplemental Table S4. However, the unrestrictive database search algorithms are used to identify unexpected modifications and allocate post-translationally modified peptide to MS/MS spectra, but it is often difficult to assign the accurate localization of modifications to specific sites. To this end, the MaxQuant computational proteomics platform was used to search for the specific PTM and determine the localization probability of modifications in peptides (82). We chose 22 common modification types in prokaryotic organisms to further search against H. pylori protein database using MaxQuant. In all, 1854 unique peptides encompassing 22 different PTMs from 805 H. pylori proteins were identified with high confidence (FDR 1%). To improve the search accuracy and confidence, peptide identifications of which the score for localization is < 40 were eliminated from the search results. Finally, 3415 modification sites on 1854 unique peptides from 805 proteins were identified (Fig. 4A). A detailed list of the identified peptides with each PTM is presented in supplemental Data S2.

As shown in Fig. 4B, we further classified the entire set of predicted, MS-detected novel, and modified proteins into a wide range of functional classes by searching against the NCBI COG (Clusters of Orthologous Genes) database. Notably, we found that certain functional classes were statistically overrepresented (p < 0.05) in certain sets of modified proteins. For instance, the identified proteins with nitrosylation, geranylation, and diphthamide PTMs were overrepresented in coenzyme transport and metabolism, defense mechanisms, and inorganic ion transport, suggesting that these PTM events may be extensively involved in diverse functions of H. pylori.

Proteogenomic Analysis of Bradyrhizobium japonicum USDA110—To test GAPP’s generalizability and applicability
in annotating prokaryotic genomes, it was also applied on Bradyrhizobium japonicum USDA110 data (Pride Accession 10099–10104 and 10114–10116), a model organism to study agriculturally important rhizobium-legume symbiosis. We confirmed 37% (3110) of known genes, discovered 51 high reliable previously unannotated protein coding genes, and corrected annotation for 10 genes that show rapid proteogenomic analyses of another prokaryotic genome (supplemental Data S3). Among of these novel protein-coding genes, 28 were intergenic, 5 were on different frame, and 18 were on the strand opposite to an existing gene annotation. More importantly, GAPP was used to perform the global profiling of post-translational modifications (PTMs) and revealed a large repertoire of PTM events using the same proteomic data. Our results provided a rich resource that can be used to examine the functions of reversible modifications in B. japonicum. Details of all identified PTM events and modified proteins are provided in supplemental Data S3 and available at www.peptideatlas.org/PASS/PASS00860.

**Discussion**

GAPP is a one-stop proteogenomic software for prokaryotic genome refinement and global identification of PTM events from the same experimental proteomic data sets. This software allows concurrent querying of proteomic and genomic databases to refine genome and identify diverse PTM events. This includes MS data and database preprocessing, database searches, FDR calculations, statistical result integration, biological interpretation, and global PTM discovery. In addition, GAPP is a simple and efficient proteogenomic software, which requires minimal operator intervention.

GAPP integrated MS data and database preprocessing as well as multiple database search algorithms into the pipeline; this is different from other existing tools (72, 83). All database search algorithms in this software were designed to run in parallel on a single machine or on a computer cluster to increase the search efficiency. Two advantages of GAPP are the automation of steps and the simplified command line API.
FIG. 4. Overview of the identified proteins with large-scale PTMs in *Helicobacter pylori*. A, The histogram diagram of the number of sites and proteins from 22 types of PTMs in *H. pylori*. B, Distribution of predicted, identified, and modified proteins of *H. pylori* according to clusters of orthologous gene functional category on gene function information. Different color represents different PTMs. Red hexagram denotes overrepresentation. PTM, post-translational modification.
To simplify installation, GAPP has been packaged into the software with all functions installed and configured, enabling plug-and-play use. Therefore, GAPP is accessible to a broad class of users including laboratories with limited bioinformatic skills and could be applied as a standard part of the genome annotation projects.

In proteogenomic analysis, multiple algorithmic search approaches can increase sensitivity and specificity in peptide identification more than any single algorithm (84). The target-decoy strategy (TDS) is the most widely used strategy to estimate FDR (85) and many effective methods have been developed to improve the performance and reliability of TDS (86–88). However, the error may be inflated in novel PSM identifications. To accurately identify novel peptides, it is necessary to filter the search results to control the FDR. Recently, more stringent filtering strategies, such as posterior probability (89) or separate FDRs for annotated and novel peptides (46, 90) were employed in proteomic analysis. In this work, we used a more stringent strategy (novel FDR) to increase the accuracy of identified novel peptides (76, 77). However, because of the lower expression levels or worse scores of the novel peptides in proteogenomic study, the FDR calculation is difficult to explicitly estimate to exclude false peptides (90). To further confirm the identified novel genes, GAPP applied a postprocessing of identification strategy, including the novel protein sequence alignment and GO functional annotation. This was done to validate identified novel genes, which is independent of the discovery of novel peptides and indirectly verifies the existence of the identified novel genes. Moreover, if the strand-specific transcriptome information is available, it will provide additional evidence for the identification of novel genes (65, 91).

At present, identification of large numbers of PTMs from complex proteome mixtures remains a critical challenge because the database search space expands exponentially as the number of PTMs increases (11, 36, 92). The search space inflates dramatically with each additional variable modification added in the “restricted search” mode, leading to low interpretation rates of MS/MS spectra and high false-positive rates (93, 94). Recently, the unrestricted database search algorithms, which do not require prior knowledge of the modifications present in the sample, are used in high-throughput proteome analysis (90). GAPP integrated the unrestricted PTMs search strategy (MODa) into the pipeline (64) and identified a large number of novel modifications in H. pylori. However, MODa could not provide the accurate localization of modifications to specific sites in the identified peptide sequences. To address this problem, the restricted search mode was used to search for the specific PTM and determine the localization probability of modifications in peptides (82). As each strategy has its own strengths and weaknesses, the combination of both unrestricted and restricted algorithms could improve the overall accuracy of PTM identification.

In this study, the identification of many novel proteins and PTM events have expanded the proteomic landscape of H. pylori. These findings also reveal the capability and versatility of GAPP for in-depth annotation of prokaryotic genomes. We anticipate that GAPP could be applied to any sequenced prokaryotic organism to yield similar sets of novel discoveries and believe that it represents a useful contribution to the software toolset for proteogenomics.

* This work was supported by the National Key Research and Development Program (2016YFA0501304), National Basic Research Program of China (973 Program, 2012CB518700), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB14030202) and the National Natural Science Foundation of China (Grant No. 31570829).

\[ \text{S} \] This article contains supplemental material.

** To whom correspondence should be addressed: Chinese Academy of Sciences, Institute of Hydrobiology, Wuhan 430072 China. Tel.: 0086-27-68780500; Fax: 27-68780500; E-mail: gefeng@ihb.ac.cn.

|| These authors contributed equally to this work.

REFERENCES

1. Engel, S. R., and Cherry, J. M. (2013) The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple S. cerevisiae strains at the Saccharomyces Genome Database. Database 2013, bat012.

2. Brent, M. R. (2005) Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res. 15, 1777–1786.

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

4. Schwanhauser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. Nature 473, 337–342.

5. de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., and Vogel, C. (2009) Global signatures of protein and mRNA expression levels. Mol. Biol. syst. 5, 1512–1526.

6. Wu, L., Candille, S. I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013) Variation and genetic control of protein abundance in humans. Nature 499, 79–82.

7. Dimitrakopoulos, L., Prassas, I., Diamandis, E. P., Nesvizhskii, A., Kisliger, T., Jaffe, J., and Drabovich, A. (2016) Proteogenomics: Opportunities and Caveats. Clin. Chem. 62, 551–557.

8. Menschaert, G., and Fenyo, D. (2015) Proteogenomics from a bioinformatics angle: A growing field. Mass Spectrom Rev 3999, 1–18.

9. Faulkner, S., Dun, M. D., and Hondermarck, H. (2015) Proteogenomics: emergence and promise. Cell Mol. Life Sci. 72, 953–957.

10. Pandey, A., and Pevzner, P. A. (2014) Proteogenomics. Proteomics 14, 2631–2632.

11. Nesvizhskii, A. I. (2014) Proteogenomics: concepts, applications and computational strategies. Nat. Methods 11, 1114–1125.

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

14. Muller, S. A., Findeiss, S., Pernitzsch, S. R., Wissenbach, D. K., Stadler, P. F., Hofacker, I. L., von Bergen, M., and Kalkhof, S. (2013) Identification of new protein coding sequences and signal peptidase cleavage sites of Helicobacter pylori strain 26695 by proteogenomics. J. Proteomics 86, 27–42.

15. Trapp, J., Geffard, O., Imbert, G., Gaillard, J. C., Davin, A. H., Chaumont, A., and Armengaud, J. (2014) Proteogenomics of Gammarus fossarum to document the reproductive system of amphipods. Mol. Cell. Proteomics 13, 3612–3625.
26. Luge, T., Kube, M., Freiwald, A., Meierhofer, D., Seemuller, E., and Sauer, J. A., Christie-Oleza, J. M., Bosch, R., Nogales, B., and Armengaud, J. (2015) Proteomics of the Roseobacter clade, a bacterial case study. Proc. Natl. Acad. Sci. U.S.A. 112, 265–270

27. Christie-Oleza, J. A., Miotello, G., and Armengaud, J. (2012) High-throughput proteogenomics of Ruegeria pomeroyi: seeding a better understanding of metabolic pathways in marine bacteria. Mol. Cell. Proteomics 11, 251–260

28. Christie-Oleza, J. A., Pina-Villalonga, J. M., Bosch, R., Nogales, B., and Armengaud, J. (2012) Comparative proteogenomics of twelve Roseobacter bacteriae. Proteomics 12, 2072–2081

29. Venter, E., Smith, R. D., and Payne, S. H. (2011) Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS ONE 6, e27587

30. Gardebrecht, A., Markert, S., Sievert, S. M., Felbeck, H., Thummer, A., Albrecht, D., Wollher, A., Kaschis, J., Le Brin, L., Lehmann, R., Daniel, R., Liesegang, H., Hecker, M., and Schweder, T. (2012) Physiological homogeneity among the endosymbionts of Riftia pachyptila and T Even jerahoniana revealed by proteogenomics. ISME J. 6, 766–776

31. Christie-Oleza, J. A., and Armengaud, J. (2010) In-depth analysis of exo-proteomes from marine bacteria by shotgun liquid chromatography-tandem mass spectrometry: the Ruegeria pomeroyi DSS-3 case-study. Mar. Drugs 8, 2223–2239

32. Baudet, M., Ortel, P., Gaillard, J. C., Fernandez, B., Guerin, P., Ensajbel, C., Soth, S., de Groot, A., Barakat, M., Dedieu, A., and Armengaud, J. (2010) Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwanted use of non-canonical translation initiation codons. Cell Mol. Proteomics 9, 415–426

33. Delmote, N., Knief, C., Chaffron, S., Innerrebinger, G., Roschitzki, B., Schlapbach, R., von Mering, C., and Vorholt, J. A. (2009) Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. ISME J. 3, 57–64

34. Kelkar, D. S., Kumar, D., 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, V. K., Katoch, M. V., Chaerkady, R., Ramachandran, S., Dash, D., and Pandey, A. (2011) Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. Mol. Cell. Proteomics 10, 21034–21038

35. Chaerkady, R., Kelkar, D. S., Muthusamy, B., Kendasamy, D., Dwivedi, S. B., Sahasrabudhe, N. A., Kim, M. S., Renuse, S., Pinto, S. M., Sharma, R., Pawar, H., Sekhar, N. R., Mohanty, A. K., Getnet, D., Yang, Y., Zhong, J., Dash, A. P., MacCallum, R. M., Delanghe, B., Milamgo, G., Kumar, A., Keshava Prasad, T. S., Okutake, M., Kumar, N., and Pandey, A. (2011) A proteogenomic analysis of Anopheles gambiae using high-resolution Fourier transform mass spectrometry. Genome Res. 21, 1792–1818

36. Yang, M. K., Yang, Y. H., Chen, Z., Zhang, J., Lin, Y., Wang, Y., Xiong, Q., Li, T., Ge, F., Bryant, D. A., and Zhao, J. D. (2014) Proteogenomic analysis and global discovery of post-translational modifications in prokaryotes. Proc. Natl. Acad. Sci. U.S.A. 111, E5633–5642

37. Hart, G. W., and Ball, L. E. (2013) Post-translational modifications: a major focus for the future of proteomics. Mol. Cell. Proteomics 12, 3443

38. Arvanitis, A., Carpenter, A., Gengenrude, C., Palmez, A., Chiappetta, G., Marino, G., and Pucci, P. (2009) Technical advances in proteomics mass spectrometry: identification of post-translational modifications. Clin. Chem. Lab. Med. 7, 57–65

39. Xiong, Q., Chen, Z., and Ge, F. (2016) Proteomic analysis of post-translational modifications in cyanobacteria. J. Proteomics 134, 57–64

40. Na, S., and Paek, E. (2015) Software eyes for protein post-translational modifications. Mass Spectrom. Rev. 34, 133–147

41. Chicooree, N., Unwin, R. D., and Griffiths, J. R. (2015) The application of targeted mass spectrometry-based strategies to the detection and localization of post-translational modifications. Mass Spectrom. Rev. 34, 595–626

42. Garavelly, J. S. (2004) The RESID Database of Protein Modifications as a resource and annotation tool. Proteomics 4, 1527–1533

43. Ost, J. A., Olson, N., and Cordwell, S. J. (2014) Beyond gene expression: the impact of protein post-translational modifications in bacteria. J. Proteomics 97, 265–286

44. Marx, H., Lemeer, S., Schliep, J. E., Matheron, L., Mohammed, S., Cox, J., Mann, M., Heck, A. J., and Kuster, B. (2013) A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. Nat. Biotechnol. 31, 557–564

45. Olsen, J. V., and Mann, M. (2013) Status of large-scale analysis of post-translational modifications in yeast using mass spectrometry. Mol. Cell. Proteomics 12, 3443–3452

46. Branca, R. M., Orte, L. M., Johansson, H. J., Granholm, V., Huss, M., Perez-Bercoff, A., Forshed, J., Kall, L., and Lehtio, J. (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat. Methods 11, 59–62
Proteogenomic Software for Prokaryotic Genome Annotation

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

48. Wang, X., and Zhang, B. (2013) customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics 29, 3235–3237

49. Giddings, M. C., Shah, A. A., Gesteland, R., and Moore, B. (2003) Genomed-based peptide fingerprint scanning. Proc. Natl. Acad. Sci. U.S.A. 100, 20–25

50. Kuhring, M., and Renard, B. Y. (2012) IPiG: integrating peptide spectrum matches into genome browser visualizations. PLoS ONE 7, e50246

51. Brouwer, R. W., van Hijum, S. A., and Kuipers, O. P. (2009) MINOMICS: visualizing prokaryote transcriptomics and proteomics data in a genomic context. Bioinformatics 25, 139–140

52. Pang, C. N., Tay, A. P., Aya, C., Twine, N. A., Harkness, L., Hart-Smith, G., Kassem, M., and Wilkins, M. R. (2014) Tools to covisualize and coanalyze proteomic data with genomes and transcripts: validation of genes and alternative mRNA splicing. J. Proteome Res. 13, 84–98

53. Ferro, M., Tardif, M., Reguer, E., Cahuzac, R., Bruley, C., Vermaat, T., Nuges, E., Viguouroux, M., Vandenbrouck, Y., Gari, J., and Viali, A. (2009) Peptide Miner: a software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genome sequences. J. Proteome Res. 8, 1873–1883

54. Risk, B. A., Spitzer, W. J., and Giddings, M. C. (2013) Peppy: proteogenomic search software. J. Proteome Res. 12, 3019–3025

55. Tovchigrechko, A., Veneponly, P., and Payne, S. H. (2014) PGP: parallel prokaryotic proteogenomics pipeline for MPI clusters, high-throughput batch clusters and multicore workstations. Bioinformatics 30, 1449–1450

56. Sanders, W. S., Wang, N., Bridges, S. M., Malone, B. M., Dandass, Y. S., McCarthy, F. M., Nanduri, B., Lawrence, M. L., and Burgess, S. C. (2011) The proteogenomic mapping tool. BMC Bioinformatics 12, 115

57. Omasits, U., Ahrens, C. H., Muller, S., and Wollschle, B. (2014) Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics 30, 884–886

58. Lange, P. F., Huygens, L. F., and Overveld, C. M. (2012) TopFIND 2.0—linking protein termini with proteolytic processing and modifications altering protein function. Nucleic Acids Res. 40, D351–361

59. Peterson, E. S., McCue, L. A., Schrimpe-Rutledge, A. C., Jensen, J. L., Risk, B. A., Spitzer, W. J., and Giddings, M. C. (2013) Peppy: proteogenomic search software. J. Proteome Res. 12, 3019–3025

60. Kim, M. S., Pinto, S. M., Getnet, D., Nirjogi, R. S., Manda, S. S., Charekady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Beal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D. N., Patil, A. H., Nanappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, N. T., Raju, R., Kumar, M., Sreerivasanmuthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Satta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S., Jayaraj, S., Rajagopal, S., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mugal, K., Chatterjee, A., Huang, T. C., Zhong, J., Wu, Y. X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sridevshukh, R., Maltra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S. K., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahe, C. A., Gowda, H., and Pandey, A. (2014) A draft map of the human proteome. Nature 508, 575–581

61. Kumar, D., Yadav, A. K., Kadi, P. K., Nagaraj, S. H., Girnmond, S. M., and Dash, D. (2013) Proteogenomic Analysis of Bradyrhizobiurn japonicum USDA110 Using Genosuite, an Automated Multi-algorithmic Pipeline. Mol. Cell. Proteomics 12, 3388–3397

62. Elias, J. E., Haas, W., Faherty, B. K., and Gygi, S. P. (2005) Comparative evaluation of mass spectrometry platforms used in large-scale proteomic investigations. Nat. Methods 2, 667–675

63. 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

64. Nesvizhskii, A. I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat. Methods 4, 787–797

65. Karpova, M. A., Karpov, D. S., Ivanov, M. V., Pyshtynski, M. A., Cherno- brovkin, A. L., Lobas, A. A., Liatsis, A. V., Archakov, A. I., Gorshkov, M. V., and Moshkovskii, S. A. (2014) Exome-Driven Characterization of the Cancer Cell Lines at the Proteome Level: The NCI-60 Case Study. J. Proteome Res. 13, 5551–5560

66. Wei, B., Xu, S., Zhou, R., Zhang, B., Wang, X., Liu, X., Xu, X., and Liu, S. (2016) PGA: an R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq. BMC Bioinformatics 17, 1

67. Baker, P. R., and Chakley, R. J. (2014) MS-viewer: a web-based spectral viewer for proteomics results. Mol. Cell. Proteomics 13, 1392–1396

68. Cao, X. J., Dai, J., Xu, H., Nie, S., Chang, X., Hu, B. Y., Sheng, Q. H., Wang, L. Z., Ning, Z. B., Guo, X. K., Zhao, G. P., and Zeng, R. (2010) High-coverage proteome analysis reveals the first insight of protein modification systems in the pathogenic spirochete Leptospira interrogans. Cell Res. 20, 197–210

69. Gupta, N., Tanner, S., Jaitly, N., Adkins, J. N., Lipton, M., Edwards, R., Baker, P. R., and Chakley, R. J. (2014) MS-GF + makes progress towards a universal database search tool for proteomics. Nat. Communications 5, 2277

70. Kim, S., and Pezvner, P. A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Communications 5, 2277
81. Liu, L., Liu, X. J., Sun, W., Li, M. X., and Gao, Y. H. (2013) Unrestrictive identification of post-translational modifications in the urine proteome without enrichment. *Proteome Sci.* 11, 1–9
82. 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
83. Nagaraj, S. H., Waddell, N., Madugundu, A. K., Wood, S., Jones, A., Mandyam, R. A., Nones, K., Pearson, J. V., and Grimmond, S. M. (2015) PGTools: a software suite for proteogenomic data analysis and visualization. *J. Proteome Res.* 14, 2255–2266
84. Yu, W., Taylor, J. A., Davis, M. T., Bonilla, L. E., Lee, K. A., Auger, P. L., Farnsworth, C. C., Welcher, A. A., and Patterson, S. D. (2010) Maximizing the sensitivity and reliability of peptide identification in large-scale proteomic experiments by harnessing multiple search engines. *Proteomics* 10, 1172–1189
85. 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
86. Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M., Schmidt, A., Buhrmann, J. M., Hengartner, M. O., and Aebersold, R. (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* 8, 2405–2417
87. Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R. L., Aebersold, R., and Nesvizhskii, A. I. (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* 10, M111 007690
88. Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B., and Bantscheff, M. (2015) A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Mol. Cell. Proteomics* 14, 2394–2404
89. 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
90. Zhang, K., Fu, Y., Zeng, W. F., He, K., Chi, H., Liu, C., Li, Y. C., Gao, Y., Xu, P., and He, S. M. (2015) A note on the false discovery rate of novel peptides in proteogenomics. *Bioinformatics* 31, 3249–3253
91. Levin, J. Z., Yassour, M., Adiconis, X. A., Nusbaum, C., Thompson, D. A., Friedman, N., Gnimke, A., and Regev, A. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* 7, 709–U767
92. Hernandez, C., Waridel, P., and Quadroni, M. (2014) Database construction and peptide identification strategies for proteogenomic studies on sequenced genomes. *Curr. Topics Med. Chem.* 14, 425–434
93. Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C., Allen, N. P., Rexach, M., and Burlingame, A. L. (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry data set acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large data sets. *Mol. Cell. Proteomics* 4, 1194–1204
94. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Murnby, M., Pevzner, P. A., and Bafna, V. (2005) InsPeCT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* 77, 4626–4639