PROTEOMICS 2014, 14, 2790–2794
DOI 10.1002/pmic.201400153

TECHNICAL BRIEF

Discovery of rare protein-coding genes in model methylotroph Methylobacterium extorquens AM1

Dhirendra Kumar1, Anupam Kumar Mondal1,2, Amit Kumar Yadav1* and Debasis Dash1,2

1 G.N. Ramachandran Knowledge Center for Genome Informatics, CSIR-Institute of Genomics and Integrative Biology, Delhi, India
2 Academy of Scientific and Innovative Research, New Delhi, India

Proteogenomics involves the use of MS to refine annotation of protein-coding genes and discover genes in a genome. We carried out comprehensive proteogenomic analysis of Methylobacterium extorquens AM1 (ME-AM1) from publicly available proteomics data with a motive to improve annotation for methylotrophs; organisms capable of surviving in reduced carbon compounds such as methanol. Besides identifying 2482 (50%) proteins, 29 new genes were discovered and 66 annotated gene models were revised in ME-AM1 genome. One such novel gene is identified with 75 peptides, lacks homolog in other methylobacteria but has glycosyl transferase and lipopolysaccharide biosynthesis protein domains, indicating its potential role in outer membrane synthesis. Many novel genes are present only in ME-AM1 among methylobacteria. Distant homologs of these genes in unrelated taxonomic classes and low GC-content of few genes suggest lateral gene transfer as a potential mode of their origin. Annotations of methylotrophy related genes were also improved by the discovery of a short gene in methylotrophy gene island and redefining a gene important for pyrroquinoline quinone synthesis, essential for methylotrophy. The combined use of proteogenomics and rigorous bioinformatics analysis greatly enhanced the annotation of protein-coding genes in model methylotroph ME-AM1 genome.

Keywords:
Gene prediction / Genome annotation / Proteogenomics

Additional supporting information may be found in the online version of this article at the publisher’s web-site.

The availability of high-throughput omics data presents opportunities to make discoveries by implementing novel ways of analysis. Such analyses are largely based on genomic features such as protein-coding genes and their functions, making genome annotation a prerequisite for understanding biology of an organism at molecular as well as system levels. However, such genome annotations are largely computational and contain errors [1, 2]. The extent of misannotations might be higher in taxonomies with greater genome diversity [3]. Methylobacteriaceae is one such diverse taxonomic family that comprises Gram-negative soil bacteria with a distinct ability to grow on reduced carbon compounds such as methanol or methylamine and have a significant role in producing phytohormones to support vegetation. These bacteria are also of interest to biotechnology industry for the production of important biological molecules, proteins, and chemicals using methanol as feedstock [4]. Among methylobacteria, Methylobacterium extorquens AM1 (ME-AM1) is the most widely studied and is considered an experimental model to study methylotrophy. Besides, a 5.7 Mb long GC-rich circular chromosome with 4983 annotated protein-coding genes, the

Correspondence: Dr. Debasis Dash, G.N. Ramachandran Knowledge Center for Genome Informatics, CSIR-Institute of Genomics and Integrative Biology, South Campus, Sukhdev Vihar, Mathura Road, Delhi 110025, India
E-mail: dodash@igib.res.in, dodash@igib.in
Fax: +91-11-29879301

Abbreviations: FDR, false discovery rate; NSAF, normalized spectral abundance factor; PSM, peptide spectrum match

© 2014 The Authors. Proteomics Published by Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.
www.proteomics-journal.com
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

*Current address: Drug Discovery Research Center, Translational Health Science and Technology Institute, Plot no. 496, Phase-III, Udyog Vihar, Gurgaon-122016, Haryana, India
Colour Online: See the article online to view Figs. 1–3 in colour.
bacterium also has four additional replicons—one megaplasmid and three plasmids, which contribute additional 1258 proteins to the total proteome [5]. Despite being the most studied methylbacterium, its genome is poorly annotated. Using large-scale public proteomics data from this organism, we reannotated the genome using proteogenomics where experimentally identified peptides are utilized to revise and discover gene models. Proteogenomics has been very effective in improving gene annotations at a genome-wide scale for various prokaryotes of medical [2], environmental [6], and industrial importance [7]. Although Delmotte et al. [8] carried out an exhaustive metaproteogenomics of microbiota associated with plant leaves where methylbacteria are predominant genus; the study was not targeted to improve genome annotation. Additional in silico approaches like comparative genomics can also be utilized to improve genome annotation. Complete genome sequences of four closely related strains of *M. extorquens* and three genomes of other *Methylobacterium* species are available. For better ME-AM1 genome annotation, we performed comparative genomics of methylbacterial genomes.

Using genome-wide sequence similarity analysis (Supporting Information Materials and Methods), we observed that the genome identity between ME-AM1 and other ME strains ranged from 85.18 to 86.66%, which suggests that ≈15% of ME-AM1 genome differs from other related strains. The difference was much greater between ME-AM1 and other methylbacterial species where genome identity ranged from 16.37 to 72.74% (Supporting Information Table 1). To calculate the impact of genomic differences on the coded proteins, we estimated pan and core genome for the complete *Methylobacteriaceae* family. These eight genomes range from 5.47 to 7.77 Mb in size and contain 4829–7355 protein-coding genes. The pan genome contains 19,727 protein-coding genes indicating huge diversity among methylbacteria. The core genome, a subset of genes present in all genomes considered, is composed of 1028 genes, which is enriched for housekeeping functions. Fifty-two of 121 annotated methylotrophy genes of ME-AM1 fall in the core genome suggesting their essentiality for methylotrophy. Genomic regions, specific to one organism, highlight that 583 (11.69%) proteins of ME-AM1 genome do not have any

**Figure 1.** A genome-wide view of protein identifications, their expression (NSAF), methylotrophy genes, and novel genes.
ortholog in other methyllobacteria (Supporting Information File 1). Such high genomic differences and nonavailability of many related genomes pose limitations for comparative genomics to annotate genes. However, such analysis allows us to infer proteogenomic findings in the perspective of genomic conservation.

To carry out comprehensive proteogenomic analysis, tandem mass spectral dataset of ME-AM1 grown on acetate and methanol from the study by Schneider et al. [9] was downloaded from PRIDE repository [10]. Spectral data were searched against ME-AM1 genome using GenoSuite [3] with all four peptide identification algorithm options including MassWiz [11]. A concatenated target decoy search was implemented and false discovery rate (FDR) was calculated using formula implemented in ProteoStats [12] (see Supporting Information Materials and Methods for details).

From 538 820 spectra, 237 566 were assigned to peptides at a peptide spectrum match (PSM) FDR of <1% (Supporting Information File 2). In total 32 563 peptides were identified, of which 30 359 mapped to the main chromosome and rest mapped either to a contaminant or other replicons. A total of 2678 proteins were identified at a protein level FDR of 0.03%. A total of 2482 of these mapped to the main chromosome, 146 to the megaplasmid, 31 to the three plasmids, and 19 to the common contaminants (Supporting Information File 3). Our protein identifications show high overlap with the original study [9] as 97% (2327 of 2408) identified proteins originating from main chromosome in previous study overlap with 94% identified in this study.

From a set of 2678 protein identifications, 31 mapped to intergenic genomic regions and are spread throughout the genome (Fig. 1). Two of these were removed after manual inspection. We also utilized gene predictions by four ab initio predictors (Supporting Information Materials and Methods). Among remaining 29 novel proteins, 26 were supported by at least one gene prediction on the same frame and strand. Novel
proteins with their identified peptides and genomic context are tabulated in Supporting Information File 4 and annotated PSMs are provided in Supporting Information File 5. Among the novel proteins, is a protein identified with 75 unique peptides and is among the top 500 high expressed genes in both growth conditions based on normalized spectral abundance factor (NSAF) [13] estimates (Supporting Information Materials and Methods). Figure 2A depicts the identified peptides and genomic context of this novel protein. Notably, it does not have orthologs in any of the methylbacterial genomes and distant homologs can only be observed in far taxonomies. The protein is predicted to contain putative glycosyl transferase and lipopolysaccharide biosynthesis domains, which along with sequence similarity suggest its potential role in outer membrane synthesis and ligand binding. Among the novel proteins are ribonuclease, methyltransferase, kinase, and invertase domain containing proteins. These functional features indicate their potential role in important biological functions and thus should be pursued in future studies. Average length of novel proteins (238 aa) is comparable with that of all annotated proteins (304 aa) in the genome indicating no specific length bias.

Interestingly, 13 novel proteins did not have ortholog in any other seven related genomes (Fig. 3). Orthologs of the selected 13 proteins were only observed in relatively distant taxonomies ranging from rhizobia to firmicutes. Four of these had GC in the range 40–45% in comparison to 68.71% of the whole genome. Laterally transferred genes are known to have lower GC content compared to native genes [14]. Exclusivity to ME-AM1 and substantial relative low GC content suggest lateral gene transfer as a potential mode of origin of these genes in this bacterium. In related rhizobial groups, lateral gene transfer is a key aspect in evolution and might provide additional growth advantage in different niche [15].

Another important contribution of proteogenomics is in the correction of the translation initiation sites in a high-throughput manner. We revised 70 gene models based on the peptides mapping upstream to the currently annotated translation initiation site. Four of these 70 were not considered further for their poor quality of PSMs. Fifty-seven gene model changes were also supported by gene predictions (Supporting Information File 6). Figure 2B shows an example of the locus MexAM1 META1p1840, which is currently annotated for a short hypothetical protein. Three peptides on the same translation frame map upstream to this locus and are supported by gene prediction. This extension toward N-terminal to the annotated protein from the MexAM1 META1p1840 locus would add predicted Cytochrome_C3 heme-binding domain and zinc finger domain of HSP40 suggesting its potential role in electron transfer and protein folding. The novel peptides not only suggest a different regulatory region for the
expression but also suggest functional aspects, to a protein previously annotated as a hypothetical one.

We also checked annotation status of methylotrphy associated genes. First, we used a list of known methylotrphy genes from [5] to check for their quantitative expression using NSAF. Supporting Information File 7 lists NSAF estimates for all proteins identified in this study. As expected, 104 of 121 annotated methylotrphy genes are identified in this dataset suggesting extensive methylotrphic activity during the growth on methanol and acetate. Interestingly, we observed a short novel gene in one of the well-known methylotrphic island (Supporting Information Fig. 1), identified with three unique peptides. Although no probable function could be predicted, its genomic context in methylotrphic genomic island and expression during growth on C1 compound suggest its potential role in methylotrphic ability of ME-AM1. We also identified gene model change for a methylotrphic gene locus MexAM1 META1p1750. This protein is annotated to be involved in biosynthesis of pyrroloquinoline quinone, a compound essential for methylotrphic ability during growth on methanol [16].

Our analyses of methylotrphoria suggest that proteogenomic annotation of less explored taxonomies is promising to yield new and rare versions of protein-coding genes, which when included in public databases will enable better annotations both of gene models and protein functions for a wide range of organisms. Additionally, public availability of tandem mass spectral data and automated proteogenomic analysis software like GenoSuite should allow researchers to carryout similar analyses for less explored taxonomies like methylotrphoria.

We thank Drs. Abhay Sharma and Sheetal Gandotra for helpful discussions and insightful comments while proof-reading the manuscript. D. K. is supported by Department of Science and Technology-INSPIRE Senior Research Fellowship. A. K. M is supported by Department of Biotechnology-BINC Senior Research Fellowship. D. D. acknowledges CSIR Genesis project (BSC-0121) for compute infrastructure and publication charges.

The authors have declared no conflict of interest.

References

[1] Aivaliotis, M., Gevaert, K., Falb, M., Tekke, A. et al., Large-scale identification of N-terminal peptides in the halophilic archaea Halobacterium salinarum and Natronomonas pharaonis. J. Proteome Res. 2007, 6, 2195–2204.

[2] Kelkar, D. S., Kumar, D., Kumar, P., Balakrishnan, L. et al., Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. Mol. Cell. Proteomics 2011, 10, M111.011627.

[3] Kumar, D., Yadav, A. K., Kadimi, P. K., Nagaraj, S. H. et al., Proteogenomic analysis of Bradyrhizobium japonicum USDA110 using GenoSuite, an automated multi-algorithmic pipeline. Mol. Cell. Proteomics 2013, 12, 3388–3397.

[4] Smolukalova, H., Erb, T. J., Fuchs, G., Methanol assimilation in Methylobacterium extorquens AM1: demonstration of all enzymes and their regulation. PLoS ONE 2010, 5, e13001.

[5] Vuilleumier, S., Chistoserdova, L., Lee, M. C., Bringel, F. et al., Methylobacterium genome sequences: a reference blueprint to investigate microbial metabolism of C1 compounds from natural and industrial sources. PLoS ONE 2009, 4, e5584.

[6] Christie-Oleza, J. A., Miotello, G., Armengaud, J., High-throughput proteogenomics of Ruegeria pomeroyi: seeding a better genomic annotation for the whole marine Roseobacter clade. BMC Genomics 2012, 13, 2731–2742.

[7] D’haseeleer, P., Gladden, J. M., Allgaier, M., Chain, P. S. et al., Proteogenomic analysis of a thermophilic bacterial consortiun adapted to deconstruct switchgrass. PLoS ONE 2013, 8, e68465.

[8] Delmote, N., Knief, C., Chaffron, S., Innebben, G. et al., Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proc. Natl. Acad. Sci. USA 2009, 106, 16428–16433.

[9] Schneider, K., Peyraud, R., Kiefer, P., Christen, P. et al., The ethylmalonyl-CoA pathway is used in place of the glyoxylate cycle by Methylobacterium extorquens AM1 during growth on acetate. J. Biol. Chem. 2012, 287, 757–766.

[10] Martens, L., Hermjakob, H., Jones, P., Adamski, M. et al., PRIDE: the proteomics identifications database. Proteomics 2005, 5, 3537–3545.

[11] Yadav, A. K., Kumar, D., Dash, D., MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. J. Proteome Res. 2011, 10, 2154–2160.

[12] Yadav, A. K., Kadimi, P. K., Kumar, D., Dash, D., ProteoStats—a library for estimating false discovery rates in proteomics pipelines. Bioinformatics 2013, 29, 2799–2800.

[13] Zybairov, B., Mosley, A. L., Sardi, M. E., Coleman, M. K. et al., Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J. Proteome Res. 2006, 5, 2339–2347.

[14] Daubin, V., Lerat, E., Perriere, G., The source of laterally transferred genes in bacterial genomes. Genome Biol. 2003, 4, R57.

[15] Tian, C. F., Zhou, Y. J., Zhang, Y. M., Li, Q. Q. et al., Comparative genomics of rhizobia nodulating soybean suggests extensive recruitment of lineage-specific genes in adaptations. Proc. Natl. Acad. Sci. U S A 2012, 109, 8629–8634.

[16] Morris, C. J., Biville, F., Turlin, E., Lee, E. et al., Isolation, phenotypic characterization, and complementation analysis of mutants of Methylobacterium extorquens AM1 unable to synthesize pyrroloquinoline quinone and sequences of pqqD, pqqG, and pqqC. J. Bacteriol. 1994, 176, 1746–1755.