An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics

ID 4062208
Author(s) Omasits, Ulrich; Varadarajan, Adithi R.; Schmid, Michael; Goetze, Sandra; Melidis, Damianos; Bourqui, Marc; Nikolayeva, Olga; Québatte, Maxime; Patrignani, Andrea; Dehio, Christoph; Frey, Juerg E.; Robinson, Mark D.; Wollscheid, Bernd; Ahrens, Christian H.
Author(s) at UniBasel Dehio, Christoph ;
Year 2017
Title An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics
Journal Genome Research
Volume 27
Number 12
Pages / **Article-Number** 2083-2095
Accurate annotation of all protein-coding sequences (CDSs) is an essential prerequisite to fully exploit the rapidly growing repertoire of completely sequenced prokaryotic genomes. However, large discrepancies among the number of CDSs annotated by different resources, missed functional short open reading frames (sORFs), and overprediction of spurious ORFs represent serious limitations. Our strategy toward accurate and complete genome annotation consolidates CDSs from multiple reference annotation resources, ab initio gene prediction algorithms and in silico ORFs (a modified six-frame translation considering alternative start codons) in an integrated proteogenomics database (iPtgxDB) that covers the entire protein-coding potential of a prokaryotic genome. By extending the PeptideClassifier concept of unambiguous peptides for prokaryotes, close to 95% of the identifiable peptides imply one distinct protein, largely simplifying downstream analysis. Searching a comprehensive Bartonella henselae proteomics data set against such an iPtgxDB allowed us to unambiguously identify novel ORFs uniquely predicted by each resource, including lipoproteins, differentially expressed and membrane-localized proteins, novel start sites and wrongly annotated pseudogenes. Most novelties were confirmed by targeted, parallel reaction monitoring mass spectrometry, including unique ORFs and single amino acid variations (SAAVs) identified in a re-sequenced laboratory strain that are not present in its reference genome. We demonstrate the general applicability of our strategy for genomes with varying GC content and distinct taxonomic origin. We release iPtgxDBs for B. henselae, Bradyrhizobium diazoefficiens and Escherichia coli and the software to generate both proteogenomics search databases and integrated annotation files that can be viewed in a genome browser for any prokaryote.
Publisher Cold Spring Harbor Laboratory Press
ISSN/ISBN 1088-9051 ; 1549-5469
edoc-URL https://edoc.unibas.ch/59170/
Full Text on edoc Available;
Digital Object Identifier DOI 10.1101/gr.218255.116
PubMed ID http://www.ncbi.nlm.nih.gov/pubmed/29141959
ISI-Number WOS:000417047600011
Document type (ISI) Article