

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

Ulrich Omasits,<sup>1,6</sup> Adithi R. Varadarajan,<sup>1,2,6</sup> Michael Schmid,<sup>1</sup> Sandra Goetze,<sup>2</sup> Damianos Melidis,<sup>1</sup> Marc Bourqui,<sup>1</sup> Olga Nikolayeva,<sup>3</sup> Maxime Québatte,<sup>4</sup> Andrea Patrignani,<sup>5</sup> Christoph Dehio,<sup>4</sup> Juerg E. Frey,<sup>1</sup> Mark D. Robinson,<sup>3</sup> Bernd Wollscheid,<sup>2</sup> and Christian H. Ahrens<sup>1</sup>

<sup>1</sup>Agroscope, Research Group Molecular Diagnostics, Genomics and Bioinformatics & SIB Swiss Institute of Bioinformatics, CH-8820 Wädenswil, Switzerland; <sup>2</sup>Department of Health Sciences and Technology, Institute of Molecular Systems Biology, Swiss Federal Institute of Technology Zurich, CH-8093 Zurich, Switzerland; <sup>3</sup>Institute for Molecular Life Sciences & SIB Swiss Institute of Bioinformatics, University of Zurich, CH-8057 Zurich, Switzerland; <sup>4</sup>Biozentrum, University of Basel, CH-4056 Basel, Switzerland; <sup>5</sup>Functional Genomics Center Zurich, ETH & UZH Zurich, CH-8057 Zurich, Switzerland

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.

[Supplemental material is available for this article.]

Advances in next-generation sequencing technology and genome assembly algorithms have fueled an exponential growth of completely sequenced genomes, the large majority of which (>90%) originate from prokaryotes (Reddy et al. 2015). The accurate annotation of all protein-coding genes (interchangeably used with CDSs from here on) is essential to exploit this genomic information at multiple levels: from small, focused experiments, up to systems biology studies, functional screens, and accurate prediction of regulatory networks.

Yet, obtaining a high quality genome annotation is a challenging objective. Pipelines for automated de novo annotation of prokaryotic genomes have been developed (Aziz et al. 2008; Markowitz et al. 2009; Davidsen et al. 2010; Vallenet et al. 2013). Such annotations greatly benefit from a manual curation step to catch obvious errors (Richardson and Watson 2012), which is car-

ried out for selected reference genomes by resources like NCBI's RefSeq (Pruitt et al. 2012) or MicroScope (Vallenet et al. 2013). Major re-annotation efforts can affect hundreds of CDSs (Luo et al. 2009), highlighting the relevance of accurate genome annotations (Petty 2010).

Despite improvements in functional genome annotation, three major issues remain: the discrepancies of the number of CDSs annotated by different reference annotation resources (Poole et al. 2005; Bakke et al. 2009; Cuklina et al. 2016), the overprediction of spurious ORFs that do not encode a functional gene product (Dinger et al. 2008; Marcellin et al. 2013), and the underrepresentation of short ORFs (sORFs) (Hemm et al. 2008; Warren et al. 2010; Storz et al. 2014). True sORFs, which often belong to important functional classes like chaperonins, ribosomal proteins, proteolipids, stress proteins, and transcriptional regulators (Basrai et al. 1997; Zuber 2001; Hemm et al. 2008), are inherently difficult

<sup>6</sup>These authors contributed equally to this work.

Corresponding author: christian.ahrens@agroscope.admin.ch

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.218255.116>. Freely available online through the *Genome Research* Open Access option.

© 2017 Omasits et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

to differentiate from the large amount of spurious sORFs (Dinger et al. 2008; Marcellin et al. 2013).

Proteogenomics, a research field at the interface of proteomics and genomics (Ahrens et al. 2010; Nesvizhskii 2014), is one attractive approach to address these problems. The direct protein expression evidence provided by tandem mass spectrometry (MS) for CDSs missed in genome annotations differs from ribosome profiling data: While the latter can capture translational activity on a genome-wide scale (Ingolia 2014), proteogenomics allows detection of stable proteins. First used in the genome annotation effort for *Mycoplasma mobile* (Jaffe et al. 2004), proteogenomics has since been applied to both prokaryotes (Gupta et al. 2007; de Groot et al. 2009; Payne et al. 2010; Venter et al. 2011; Kumar et al. 2013; Marcellin et al. 2013; Kucharova and Wiker 2014; Cuklina et al. 2016) and eukaryotes (Nesvizhskii 2014; Menschaert and Fenyo 2015). Yet, the need for computational solutions to apply proteogenomics more broadly has been noted (Castellana and Bafna 2010; Renuse et al. 2011; Armengaud et al. 2014; Nesvizhskii 2014). Of particular interest are tools that create customized databases (DBs) to identify evidence for unannotated ORFs. RNA-seq data have been used to limit the protein search DB size to achieve better statistical power (Wang et al. 2012; Woo et al. 2013; Zickmann and Renard 2015). Other MS-friendly DB solutions that integrate data from different species or strains include MScDB (Marx et al. 2013), MSMSpddb (de Souza et al. 2010), and PG Nexus (Pang et al. 2014). Even pipeline solutions were developed that allow the search of proteomics data against a six-frame translation-based DB, including Peppy (Risk et al. 2013), GenoSuite (Kumar et al. 2013), and PGP (Tovchigrechko et al. 2014). However, an integration that leverages benefits of manually curated reference annotations and a six-frame translation into one highly informative, nonredundant, and transparent resource has not been accomplished so far.

Here, we address this unmet need of the microbiology and proteomics community and present an integrative strategy that takes the MS-friendly DB concept one important step further. We aimed to develop a solution that integrates and consolidates annotations from different sources and, at the same time, captures information about their overlap and differences in (1) informative identifiers, (2) an integrated proteogenomics search database (iPtgxDB), and (3) a GFF (generic feature format) file that can be viewed in a genome browser and overlaid with experimental evidence. One key objective was high information content, i.e., ensuring that the vast majority of peptides in the search DB unambiguously identify one distinct protein (Qeli and Ahrens 2010), thereby overcoming the need to dis-entangle protein groups implied by shared peptides and speeding up downstream data analysis. Stringency, i.e., allowing tight control of the false discovery rate (FDR), and flexibility, i.e., being applicable to both reference model organisms and newly sequenced strains, represented other key requirements. Here, we release the software as a public web server (<https://iptgxdb.expasy.org>), so that iPtgxDBs can enable research groups to take full advantage of completely sequenced genomes by improving genome annotations with proteogenomics.

## Results

### Experimental evidence underscores the need for a general proteogenomics approach

We used the  $\alpha$ -proteobacterium *Bartonella henselae* strain Houston-1 (Bhen) to explore how genome annotation differences could best

be integrated for a proteogenomics approach. A comparison of four Bhen reference genome annotations and results from two ab initio gene prediction tools (see Methods) confirmed reports for other organisms (Poole et al. 2005; Bakke et al. 2009; Cuklina et al. 2016) that both the number of predicted ORFs and their precise start sites largely differ (see Supplemental Fig. S1). Only 50% of the Bhen CDSs were annotated or predicted completely identical by all six resources, and 37% were unique to one resource (Supplemental Fig. S1A). Of note, 23% of the CDSs of the recent NCBI RefSeq2015 re-annotation differed from RefSeq2013 (Supplemental Table S1): 55 CDSs were removed (99 added), 74 CDSs were shortened (54 extended), and 64 pseudogenes were removed (15 added).

To assess the validity of the RefSeq2015 re-annotation, we relied on an ideal data set: a complete prokaryotic proteome (including many low abundant proteins) expressed under two conditions that mimic those encountered by Bhen in the arthropod vector midgut (uninduced condition) and the bloodstream of its mammalian host (induced condition), which had been searched against RefSeq2013 (Omasits et al. 2013). A search against a RefSeq2015 protein DB provided experimental evidence for many of the re-annotations, including six sORFs that we had previously identified with a prototype of our proteogenomics approach as novel (Supplemental Table S1) and which have since been added to RefSeq2015. Also, among the 55 removed CDSs, we found 32 of 52 proteins we had earlier singled out as potential overpredictions (Omasits et al. 2013). However, we also found several cases that supported the earlier RefSeq2013 annotation, including expression evidence for CDSs that were relabeled as pseudogenes and for removed CDSs. This highlights the need for an integrated, yet general approach to address this fundamental problem of gene annotation inconsistency.

### A general, integrative proteogenomics approach

An ideal solution to capture the full protein-coding potential of genome sequences should therefore (1) consider results from different reference genome annotations (Nesvizhskii 2014), which often include substantial manual curation efforts from experts, and from ab initio gene prediction tools, (2) allow the identification of the small fraction of true functional sORFs often missed by the above annotations or predictions, (3) aid in the annotation of newly sequenced genomes, and (4) enable scientists to visualize their experimental proteomics results in the context of both the genome and all available annotations.

To our knowledge, existing tools only address a subset of these requirements. These include pipeline solutions that rely on a six-frame translated genome like Peppy (Risk et al. 2013), which aims to improve the scoring function for peptide spectrum matches (PSMs), GenoSuite (Kumar et al. 2013), which uses four distinct search algorithms before integrating and visualizing the results, and PGP (Tovchigrechko et al. 2014), which draws on the experience of many proteogenomics studies (Venter et al. 2011) and highlighted the need for stringent criteria to accept novel ORFs. However, these tools do not integrate different annotation sources. Some MS-friendly integrated DBs accomplish this, such as MScDB (Marx et al. 2013), which uses a peptide-centric clustering algorithm to combine, e.g., cross-species DBs, or MSMSpddb, which allows the creation of a nonredundant protein DB for multiple closely related bacterial strains (de Souza et al. 2010). However, they do not integrate different annotations of the same genome. PG Nexus (Pang et al. 2014) uses the NCBI RefSeq annotation, a Glimmer ab

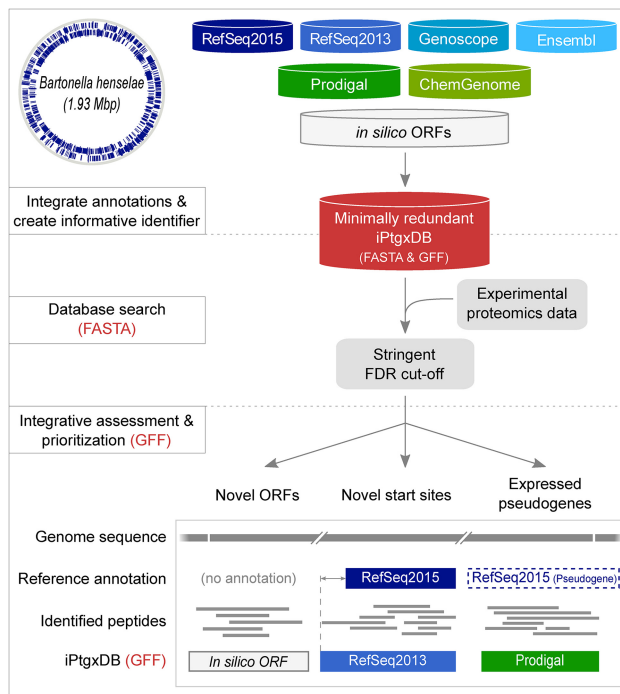
initio prediction (Delcher et al. 2007), and a six-frame translation against which peptides are searched with Mascot (Perkins et al. 1999) and later visualized onto the genome. However, the annotations are not integrated and consolidated; the boundaries of novel ORFs still have to be discovered based on peptide evidence, which requires substantial manual effort. In addition, Mascot is not ideal for the task of identifying novel ORFs in proteogenomics approaches (Omasits et al. 2013; Risk et al. 2013). To address all of the above objectives in one integrated solution, we devised a proteogenomics workflow that relies on three steps (Fig. 1).

First, popular reference genome annotations, the results of ab initio prediction algorithms, and in silico ORFs from a six-frame genome translation were combined into an integrated proteogenomics search database (iPtgxDB) (Fig. 1, upper panel) with the aim to capture the entire genomic protein-coding potential (see Methods). On top of genome annotations from NCBI RefSeq, Ensembl, and Genoscope, we included results of the ab initio gene prediction algorithms Prodigal (Hyatt et al. 2010), which performs well even for genomes with high GC content where gene calling is more difficult (Marcellin et al. 2013), and ChemGenome (Singhal et al. 2008). The latter relies on physicochemical characteristics of codons calculated by molecular dynamics simulations (Singhal et al. 2008) and is thus quite different from Prodigal and similar tools (Pati et al. 2010). Finally, to be able to

identify functional sORFs, which are often missed due to rather conservative length thresholds for ab initio predicted ORFs, all potential in silico ORFs (a modified six-frame translation considering alternative start codons) (see Supplemental Methods) above a selectable length threshold were added. A literature search for experimentally validated prokaryotic sORFs (Zuber 2001; Rowland et al. 2004; Venter et al. 2011) revealed that novel sORFs were longer than 20 amino acids (aa), with very few exceptions (Hemm et al. 2008). To balance comprehensiveness and avoid loss of statistical power when searching large DBs (Blakeley et al. 2012; Noble et al. 2015), we selected a length threshold of 18 aa.

In a second step, proteomics data—ideally comprehensive expression data obtained under multiple conditions (Ahrens et al. 2010)—is searched against the iPtgxDB and stringently filtered (Fig. 1, middle panel). We used the search engine MS-GF+, which rigorously computes *E*-values of PSMs based on the score distribution of all peptides (Kim and Pevzner 2014) and which had performed favorably in our hands for large shotgun proteomics data sets (Omasits et al. 2013) as well as in proteogenomics studies (Risk et al. 2013; Zickmann and Renard 2015; Cuklina et al. 2016).

In a third step, peptide evidence is visualized in the context of the genome and all annotations (contained in a GFF file) using a genome browser such as IGV (Fig. 1, lower panel; Robinson et al. 2011). Candidates in major classes of novelty (novel ORFs, different or additional start sites, expressed pseudogenes) can be inspected in the context of experimental data (e.g., proteomics and transcriptomics data), functional annotations, and other features to enable a comprehensive assessment and prioritization.

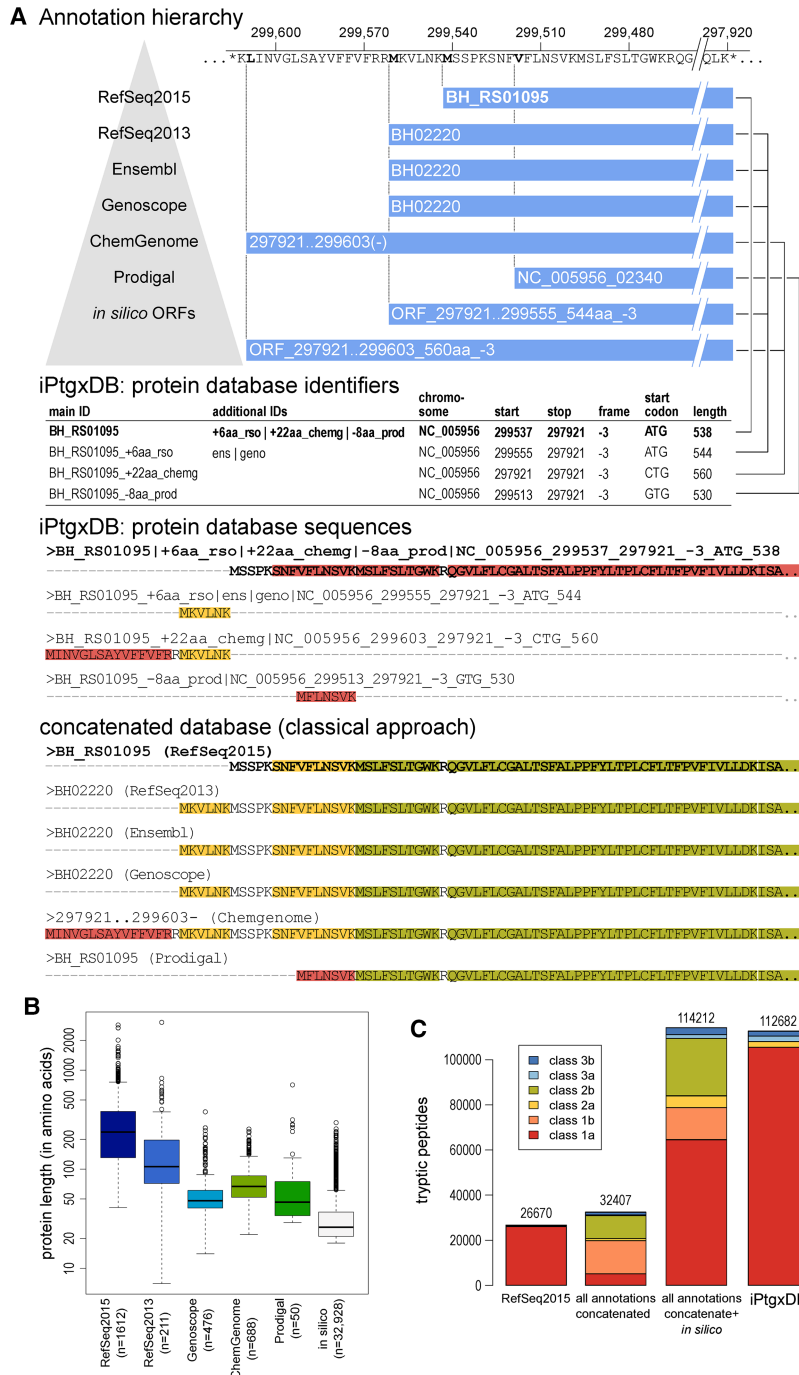


**Figure 1.** Integrative proteogenomics workflow. For a completely sequenced prokaryotic genome (Bhen is shown as an example with annotated CDSs), reference genome annotations (blue containers), results from ab initio gene prediction algorithms (green containers), and in silico ORFs (white container) are downloaded or computed and integrated in a first preprocessing step (upper panel). All CDS and pseudogene annotations are matched, and informative gene identifiers are created and stored in a minimally redundant iPtgxDB (red container; searchable protein sequences in FASTA format, integrated annotations in GFF format). Experimental proteomics data are matched to the DB using a target-decoy approach relying on stringent FDR cut-offs (middle panel). Identified PSMs and peptides are postprocessed to visualize novel candidates (lower panel) in the context of experimental data integrated with the GFF file.

### Creating minimally redundant but maximally informative protein search databases

A unique aspect of our proteogenomics approach is that almost all MS-identifiable peptides of the iPtgxDB unambiguously identify one specific protein (Fig. 2). To achieve this, we first extended our PeptideClassifier concept (Qeli and Ahrens 2010). PeptideClassifier was developed to classify the information content of peptides with respect to their originating gene model(s) into six classes for eukaryotes and three classes for prokaryotes: Class 1a peptides are most informative and allow unambiguous identification at the protein sequence, protein isoform, and gene model level (Supplemental Fig. S2). Our extension for prokaryotes now also considers six peptide evidence classes and treats protein sequences with a common stop codon and varying start positions (N termini) as a protein annotation cluster, i.e., an equivalent of a prokaryotic gene model (similar to isoforms of a eukaryotic gene model). Class 1a peptides remain most informative as they are unique to one entry in a DB, while class 1b peptides map uniquely to one annotation cluster with all identical sequences. Class 2a peptides identify a subset of sequences from an annotation cluster and class 2b peptides map to all sequences of an annotation cluster. Class 3a peptides map to identical sequences from different annotation clusters (typically duplicated genes). Class 3b peptides map to different sequences from different annotation clusters and are least informative.

The stepwise integration of resources, carried out in Figure 2A for Bhen as a model, follows a hierarchy: To leverage the quality of manual curation efforts, we start with reference annotations, then ab initio predictions, then in silico ORFs. The anchor sequence is selected from the annotation highest up in the hierarchy, i.e., here, RefSeq2015, unless no CDS is predicted in a given genomic region. Each subsequent resource added new protein clusters and



**Figure 2.** Generating an iPtgxDB with informative identifiers and a minimally redundant protein search DB in FASTA format. (A) CDSs and pseudogenes of seven resources are integrated in a stepwise fashion. Informative protein identifiers are created and illustrated for the annotation cluster, with the RefSeq2015 anchor sequence BH\_RS01095 shown in bold, where three additional start sites exist. The four different proteoforms are added to the protein search DB: the anchor sequence (bold) with the full protein sequence, the extensions (RefSeq2013 and ChemGenome) add the upstream sequence up to the first tryptic cleavage site within the anchor sequence. The shorter Prodigal prediction uses an alternative start codon resulting in a distinguishable N-terminal peptide and therefore also gets added. The two in silico ORFs are identical to annotations higher up in the annotation hierarchy and therefore are not added. Peptide classes are shown for the N-terminal sequences of the CDS annotation cluster (see also Fig. 2C). (B) Box plots of protein length for RefSeq2015 and those proteins that get added in each successive step to the protein search DB illustrate that we include many sORFs potentially missed in the reference annotations. (C) Bar chart showing the DB complexity and the peptide classes for RefSeq2015, all six integrated annotations without and with in silico ORFs, and the final iPtgxDB. The legend (*inset*) shows colors for the six peptide classes.

new extensions or reductions (i.e., alternative start sites) to an existing cluster, while identical annotations are collapsed (Table 1). Alternative start codons are captured with our approach, even for in silico ORFs (Fig. 2A).

The four distinct sequences (collapsed from eight annotations) of the protein cluster for apolipoprotein N-acyltransferase with the anchor sequence BH\_RS01095 (Fig. 2A) illustrate this key principle: By only adding the full sequence of the anchor protein plus class 1a peptides (red in Fig. 2A) that can unambiguously identify extensions or internal start sites and few class 2a peptides (yellow in Fig. 2A) to the iPtgxDB, we minimize redundancy and maximize information content of the peptides, compared to adding all protein sequences (Fig. 2A, lower panels). In the latter case, many peptides classified as 2a or 2b, which imply a subset or all annotations of a CDS cluster, would get added to the iPtgxDB. Identification of such shared peptides greatly impedes downstream analysis.

The protein identifiers of the four distinct sequences transparently capture overlap and differences of the annotations (see Methods); they show in which resource(s) the identified CDS is annotated, if and how the annotations differ, and whether it is a novel ORF or an alternative start site, again largely improving downstream data analysis (Fig. 2A).

The identifiers also contain genomic coordinates, allowing the visualization of all experimental peptide evidence for a novel ORF in its genomic context along with all integrated annotations provided in the iPtgxDB GFF file. Peptides implying any other sequence (e.g., one of the three identifiers below the anchor sequence identifier in Fig. 2A) would inform the experimentalist at a glance that novel information compared to RefSeq2015 was uncovered (see Supplemental Methods for examples of how to “interpret” the identifiers). A box plot of the lengths of the proteins added to our DB in the stepwise processing illustrates that we capture increasingly smaller proteins. Adding in silico ORFs down to a selectable length threshold allows us to query the entire protein-coding potential of the genome (Fig. 2B).

The final Bhen iPtgxDb contains 51,541 entries (Table 1), ~30 times as many as RefSeq2015 (1612). Importantly, 94% of all theoretically MS-identifiable tryptic peptides (6–40 aa) are class

**Table 1.** Result of the stepwise, hierarchical integration of resources for Bhen

| Annotation source | No. of annotations | New clusters | New extensions | New reductions | Cumulative clusters | Cumulative annotations |
|-------------------|--------------------|--------------|----------------|----------------|---------------------|------------------------|
| 1. RefSeq2015     | 1612               | 1612         | –              | –              | 1612                | 1612                   |
| 2. RefSeq2013     | 1612               | 211          | 79             | 66             | 1823                | 1968                   |
| 3. Ensembl        | 1612               | 0            | 1              | 0              | 1823                | 1969                   |
| 4. Genoscope      | 2114               | 476          | 9              | 64             | 2299                | 2518                   |
| 5. ChemGenome     | 2211               | 688          | 515            | 35             | 2987                | 3756                   |
| 6. Prodigal       | 1643               | 50           | 41             | 63             | 3037                | 3910                   |
| 7. In silico      | 54,099             | 32,928       | 16,514         | 129            | 35,965              | 53,481                 |

The number of annotations per source, new protein clusters, and new extensions and reductions are shown for each step and summarized under cumulative clusters and cumulative annotations. Overall, protein sequences for 53,481 annotations mapped to 35,965 annotation clusters. The final iPtgxDB had 51,541 entries, as sequences <6 aa (1750), i.e., not identifiable with shotgun proteomics, and indistinguishable internal start sites (190) were not considered.

1a peptides, which allow unambiguous identification of one protein (red color in the legend of Fig. 2C). Figure 2C illustrates how this is achieved: For RefSeq2015, almost all peptides mapped uniquely to one protein, which is common for a prokaryote. Combining all six annotations resulted in a modest increase of tryptic peptides (23%). However, most peptides (85%) now matched at least two annotations: either annotations for an identical sequence (class 1b) or annotations of proteins with different length but of the same annotation cluster (class 2a or 2b), which would greatly complicate the interpretation of proteomics search results. Adding in silico ORFs significantly increased the number of peptides, adding mainly new unique peptides (class 1a) for ORFs in regions without annotation (Fig. 2C). Our careful integration collapses identical sequences and removes 1b and 2b peptides; the other shared peptides (classes 2a, 3a, and 3b) account for ~6% of the iPtgxDB.

Defining DB complexity as the number of distinct tryptic peptides of 6–40aa in length, the complexity of the resulting iPtgxDB was ~50% of that of a full six-frame translated genome that Mascot (Perkins et al. 1999) or PG Nexus (Pang et al. 2014) would rely on to identify proteogenomic evidence for novel peptides. Despite the relatively large number of entries, the DB complexity is only 70% of that of baker's yeast and below 20% of a human protein DB (Supplemental Table S2).

### Searching Bhen proteomics against our iPtgxDB identifies novel ORFs

We next searched existing data from a comprehensive expressed proteome, an in vitro model mimicking interaction of Bhen with

**Table 2.** Summary of novel information uncovered by the integrated proteogenomics approach

| Annotation source      | Novel protein-coding ORF |                          | Evidence for alternative NH2-terminus |
|------------------------|--------------------------|--------------------------|---------------------------------------|
|                        | Not in RefSeq2015        | Pseudogene in RefSeq2015 |                                       |
| RefSeq2013/<br>Ensembl | 4                        | 10                       | 5                                     |
| Genoscope              | 4                        | 1                        | 0                                     |
| ChemGenome             | 1                        | 2                        | 5                                     |
| Prodigal               | 3                        | 1                        | 2                                     |
| In silico ORFs         | 10                       | 1                        | 5                                     |
| Total                  | 22                       | 15                       | 17                                    |

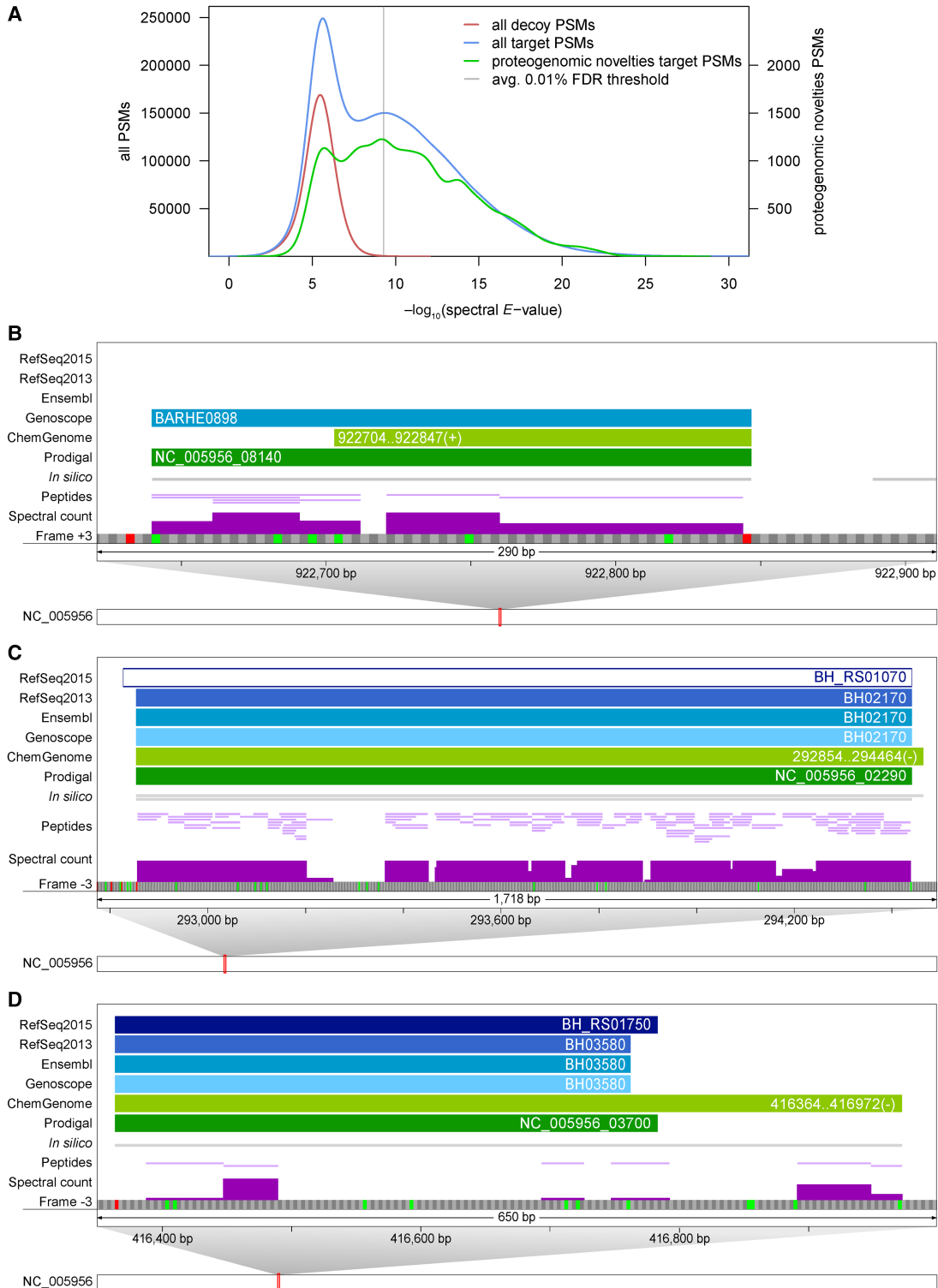
Compared to RefSeq2015, each resource added some novelty with respect to Bhen's overall protein-coding potential. Notably, close to 80% of the identified novelties could be independently confirmed by parallel reaction monitoring (PRM) (Supplemental Table S3).

the arthropod vector (uninduced condition) or its mammalian host (induced condition) (Omasits et al. 2013), against the Bhen iPtgxDB using MS-GF+. Relying exclusively on unambiguous class 1a peptides, this allowed the systematic identification of expression evidence for novel ORFs, novel start sites, and CDSs wrongly annotated as pseudogenes (Table 2). Importantly, each of the reference genome annotations, ab initio gene prediction tools, and in silico predicted ORFs provided unique novel hits, underlining the value of our integrated approach (Table 2; Supplemental Fig. S3). These hits are novel compared to the most common approach of using the latest reference annotation as the search DB, i.e., RefSeq2015, in this case.

When searching large data sets, it is imperative to use stringent cut-offs. This is particularly relevant for proteogenomics, where correctly identified novel information would require a genome annotation change. We relied on an estimated PSM-level false discovery rate cut-off of 0.01%, which resulted in a peptide-level FDR of 0.12%. This cut-off is about 10-fold more stringent than in other proteogenomics studies (Krug et al. 2013; Kumar et al. 2013; Chapman and Bellgard 2014; Zickmann and Renard 2015) and closer to the cut-offs used by Payne and colleagues (peptide level FDR cut-off 0.3%) (Venter et al. 2011). Of particular note, the *E*-value score distribution of PSMs that identify novel features is also bi-modal, similar to that of PSMs identifying annotated proteins in the target DB (Fig. 3A). To claim a potential novel ORF, we required at least three PSMs to class 1a peptides if predicted by a reference genome annotation/ab initio prediction tool, and four PSMs to class 1a peptides for in silico ORFs, in line with earlier recommendations (Nesvizhskii 2014). Furthermore, all genomic regions encoding novel candidates/start sites were expressed (Supplemental Table S3).

Overall, 37 novel Bhen ORFs (with respect to RefSeq2015) were identified (Table 2; Supplemental Fig. S3): 12 annotated by another resource or ab initio prediction tool, 10 in silico only predicted ORFs, and 15 with a pseudogene annotation. In addition, 17 alternative start sites were identified. The median length of 22 novel ORFs (excluding pseudogenes) was 48 aa and that of six novel ORFs previously identified with a prototype versus RefSeq2013 (i.e., prior to the re-annotation) (Supplemental Table S1) was 80 aa (Supplemental Fig. S4). This confirms that the novel ORFs represent sORFs commonly underrepresented in genome annotations. Analysis of the estimated expression levels of the novel ORF candidates including pseudogenes (see Methods) furthermore indicated that several of the sORFs are well-expressed proteins (Supplemental Fig. S4) that may carry out important functions.

Examples of novel ORFs included differentially expressed sORFs such as BARHE0898 (68 aa), that was expressed roughly



**Figure 3.** Examples of novel information uncovered by integrative proteogenomics. (A) *E*-value distribution of PSMs against proteins of decoy and target DB (red and blue lines, left scale) plus the distribution of those PSMs that uncovered novelties (green line, right scale). A PSM level FDR cut-off of 0.01% was selected per sample. (B–D) Zoomed-in views of genomic regions that harbor novelties. For illustration, a single frame of the forward/reverse strand with possible start (green) and stop codons (red) is shown, along with annotations and experimental evidence (spectral counts scaled from 0 to 20). (B) Example of a novel sORF of 68 amino acids (BARHE0898, frame +3). (C) Example of a highly expressed pseudogene (RefSeq2015: BH\_RS01070, frame –3); 2244 spectra are mapped to 117 peptides of NusA, which is annotated as a pseudogene in RefSeq2015 for unknown reasons. There is no experimental evidence for the +8-aa N-terminal extension predicted by ChemGenome. (D) Proteomic expression evidence supports a 63-aa-longer proteoform of BH\_RS01750 (frame –3) uniquely predicted by ChemGenome.

6.5-fold higher in the induced condition (Supplemental Table S3). Peptide evidence supported the longer form of this ORF annotated by Genoscope and Prodigal (Fig. 3B). Another novel sORF of 67 aa, a lipoprotein uniquely predicted by ChemGenome, was only identified in the uninduced condition (Supplemental Fig. S5A). Even smaller ORFs were identified, including a well-expressed (nine peptides, 143 PSMs) sORF of 49 aa that was uniquely predicted by Prodigal (Supplemental Fig. S5B; Supplemental Table S3) and an in silico only predicted sORF of 34 aa (Supplemental Fig. S5D). We also identified a highly expressed RefSeq2015 pseudogene (BH\_RS01070, frame -3) (Fig. 3C) annotated as normal CDS (transcription elongation factor NusA) by RefSeq2013, Ensembl, and Genoscope. Other misannotated pseudogenes included a potassium-efflux transporter (BH10840) and the *Bartonella* effector protein BepD (BH13410) (Supplemental Table S3). Finally, for BH\_RS01750, a hypothetical protein encoded in a prophage region, only ChemGenome correctly predicted a 63-aa-longer proteoform than annotated by other resources; its expression was supported by several peptides (Fig. 3D). Proteomic data thus can support novel start sites (even multiple start sites) (Supplemental Fig. S5C) and distinguish between those predicted by different reference annotations.

#### Confirmation of novelties by independent targeted proteomics

In order to confirm novel ORFs by independent methods, the expression of novel candidates at the protein level was assessed by targeted proteomics experiments (Fig. 3B–D; Supplemental Fig. S5A–D) with parallel reaction monitoring (PRM) assays (Peterson et al. 2014). For this highly sensitive method, cytoplasmic (cyt) and total membrane (TM) extracts were prepared from new biological samples as described (Omasits et al. 2013; see Methods). Overall, we were able to validate 107 of 138 targeted peptides (78%) (Supplemental Tables S3, S4), including low expressed novel proteins implied by one peptide and three PSMs. We had previously derived predominant subcellular localizations (SCL) for all proteins (Stekhoven et al. 2014), which we computed here also for the novel candidates (see Supplemental Methods; Supplemental Table S3). Importantly, the SCL data agreed with the PRM evidence in either cyt or TM fractions, thereby adding yet another layer of support to the confirmed novel ORFs. The validation success was 100% for the six novel ORFs identified previously by our prototype (novel with respect to the RefSeq2013 annotation) (Supplemental Fig. S4), ~80% for 15 expressed pseudogenes and 12 novel ORFs from another genome annotation/prediction, 60% for novel in silico ORFs, and ~55% for novel start sites (Supplemental Table S3).

Of note, we identified 38 of 51 lipoproteins predicted to have a SpII cleavage site (LipoP, version 1.0) (Supplemental Table S5). Two of these were identified among the 12 novel ORFs (one predicted by Genoscope, one by ChemGenome) and two others among the six novel ORFs identified previously, which have since been incorporated in RefSeq2015. All four candidates were validated by PRM, and their predominant SCL indicated that they were found exclusively in the total membrane or outer membrane fractions (Supplemental Table S3). Lipoproteins could thus represent a class of proteins for which a substantial percentage is missed in reference genome annotations, which is relevant given their important roles in signaling, protein folding and export, virulence, immunity, and antibiotic resistance (Kovacs-Simon et al. 2011).

#### De novo assembly and genome comparison of Bhen strains underlines the importance of a correct genome sequence

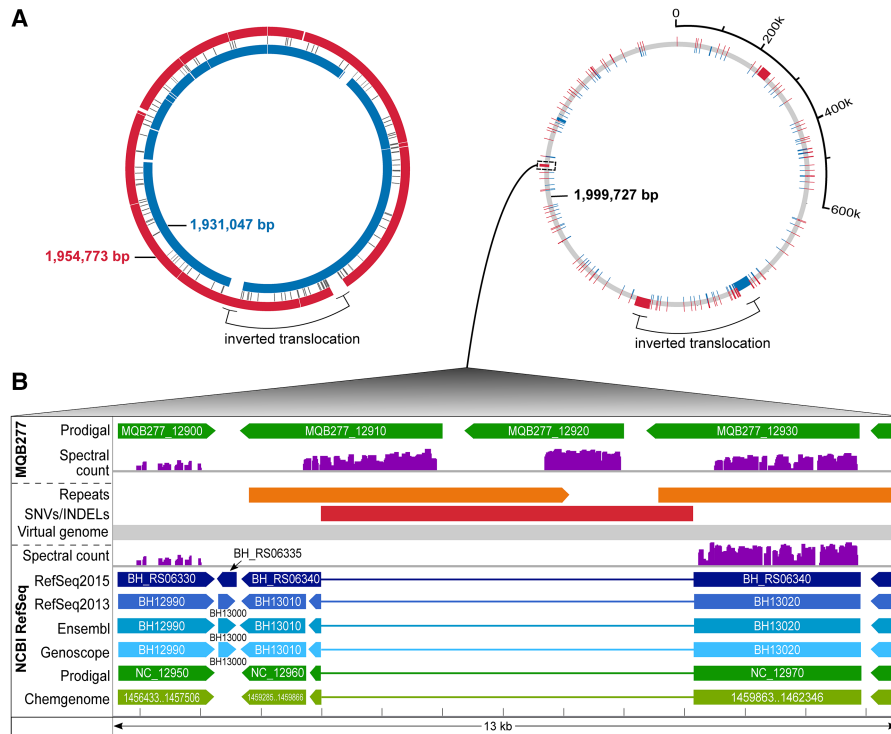
Massively reduced sequencing costs and improved assembly algorithms make it possible to determine the actual genome sequence of key bacterial strains used in a laboratory, which provides an optimal basis to integrate functional genomics data and to correctly identify novel sORFs. We thus explored to what extent our lab strain (MQB277) differed from the Bhen reference strain (Alsmark et al. 2004) and whether we could detect protein expression evidence for novel ORFs in unique genomic regions and for single amino acid variations (SAAVs). Conceptually, this allowed testing of our approach on a newly sequenced genome, now relying on an iPtgxDB integrating Prodigal predictions plus in silico ORFs, but without curated reference genome annotations.

We de novo-assembled (Chin et al. 2013) the PacBio-sequenced genome of the MQB277 lab strain, derived from Bhen CHDE101, a Bhen variant-1 strain (Lu et al. 2013), into one 1,954,773-bp high-quality contig (see Supplemental Methods), i.e., ~23.7 kbp longer than the NCBI reference genome (Fig. 4A). To compare closely related genomes in the context of experimental evidence, we devised a “virtual genome” concept, i.e., a coordinate system that integrates sequences from reference genome and de novo assembly (Fig. 4). This allowed us to integrate annotation and experimental data tracks, to efficiently zoom down to the single nucleotide level, and to inspect all lines of evidence for observed differences.

Overall, we noted a large inversion translocation (34.4 kbp) close to the terminus of replication, previously reported for some *Bartonella* isolates (Lindroos et al. 2006), and three insertions of 22.1, 6.1, and 1.4 kbp in the MQB277 assembly (Supplemental Table S6). The 22.1- and 1.4-kb insertions affected a genomic region encoding the surface protein BH01510, a BadA1 adhesin (Supplemental Fig. S6) and major pathogenicity factor that mediates binding of *B. henselae* to extracellular matrix proteins and endothelial cells (Riess et al. 2004). A complex repeat structure in this region of the assembly harbored additional ORFs, whose expression was supported by unambiguous peptide evidence. In line with the CDS MQB277\_01630 lacking a C-terminal membrane anchor, its predominant SCL was cytoplasmic (for more detail, see Supplemental Fig. S6). These data help to explain earlier experimental data that had demonstrated lack of or much lower BadA surface expression in the Bhen CHDE101 variant strain (Lu et al. 2013).

Of note, the 6.1-kbp insertion harbored two CDSs predicted to encode autotransporter proteins (Fig. 4B). Both MQB277\_12910 and MQB277\_12920 were unique to the high-quality PacBio assembly, which contains a direct repeat in this region (missing in the reference), indicative of a duplication event. Searching against the MQB277-based iPtgxDB, all three proteins were highly expressed (Fig. 4B, MQB277 track), compared to only BH\_RS06340 based on the RefSeq2015 protein DB (NCBI RefSeq track). Importantly, the novel CDS MQB277\_12910 was among the most up-regulated (42-fold) proteins in the induced condition (Supplemental Table S3). This data is in line with the dramatic re-organization of the membrane proteome reported earlier (Omasits et al. 2013). Our SCL data indicated that all three proteins were localized in the outer membrane (Supplemental Table S3).

The assembly also comprises a genomic region harboring a 1-bp insertion and a 81-bp deletion with respect to the reference. Because of the frameshift caused by the insertion, protein



**Figure 4.** Comparative analysis in the context of experimental data. Integrated visualization of two closely related genomes through a virtual genome concept. (A) On the left, the Bhen NCBI RefSeq genome (inner blue circle) is aligned to our de novo assembly (outer red circle). A large inversion-translocation (black bracket) is marked; several insertions or deletions in either genome are shown (white spaces) and a center track for single nucleotide variations (SNVs). To the right, the virtual genome (gray) is shown which incorporates both genome sequences, including all differences, into a common coordinate system. Unique sequences are shown in blue or red (the inversion-translocation present in both genomes is left as is). (B) Zoom into the region harboring a 6088-bp insertion in MQB277 (red bar), showing annotations for the RefSeq genome (below the virtual genome track) and assembly (above the virtual genome track), plus experimental proteomics evidence mapped against both genomes (spectral count scaled from 0–800). This region harbors a direct repeat only in the assembly (orange bars). Three CDSs (MQB277\_12910, MQB277\_12920, MQB277\_12930) annotated as autotransporters are highly expressed; the first two (novel CDSs) are only detected (unambiguous 1a peptides) with the correct genome sequence available.

expression of a CDS annotated as an ABC transporter (downstream from the insertion) was only observed in the lab strain assembly (Fig. 5A; Supplemental Table S6), which is also supported by transcriptomics data (see Methods). Due to the frameshift, CDSs in this region were either annotated as pseudogenes or as split CDSs (NCBI RefSeq track). This example is furthermore noteworthy, as our transcriptomics data set (Omasits et al. 2013) had been re-analyzed with an RNA-seq based proteogenomics approach (Zickmann and Renard 2015), and a novel ORF was reported in this region (Fig. 5A, orange arrow). However, only with the correct assembly at hand can novel ORFs be identified completely accurately. Our integrated analysis demonstrates that the novel ORF in question is, in fact, longer, and its expression is supported by class 1a peptides (MQB277 track, Supplemental Fig. S7).

Of note, we found several examples where single nucleotide variations (SNVs) led to nonsynonymous protein sequence differences, as, e.g., for transcription elongation factor GreA (Fig. 5B). The expression of this and 11 additional SAAVs (Supplemental Table S3) was again independently validated by PRM assays. Analysis of the 274 SNVs observed between the two genomes indicated that they were significantly enriched in a limited number of regions encoding surface proteins (114/274), including four of

eight hemagglutinins, and four hemolysin activator proteins (HECs) (Supplemental Table S6). Together, our data provide multiple lines of evidence for the earlier postulation that genome rearrangements observed in natural Bhen populations affect variation of surface proteins (Lindroos et al. 2006).

Finally, the search against the assembly-based protein DB led to overall 10,410 more assigned PSMs and 441 peptides (same 0.01% PSM level FDR threshold) (Supplemental Table S7). Together, these results emphasize the value that research groups can gain by sequencing and de novo-assembling their most important strains.

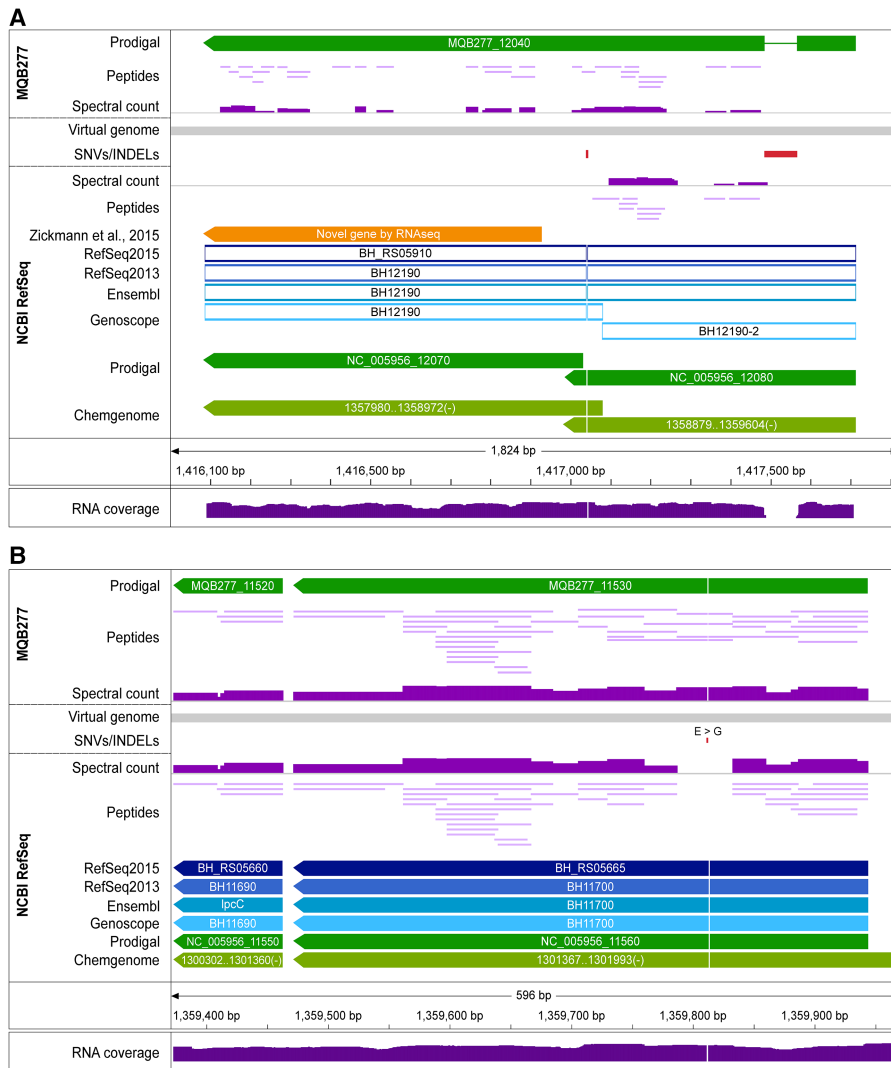
### Integrated proteogenomics approach is generically applicable

Genome annotation is more difficult and error-prone for genomes with high GC content; they contain more spurious ORFs and fewer stop codons, which leads to a reduced accuracy of translation start site prediction (Hyatt et al. 2010; Marcelin et al. 2013). To demonstrate that our approach can work beyond Bhen (38.2% GC), we have applied an earlier prototype on genomes with higher GC content: We could identify novel ORFs in the genome of *Burkholderia kirkii* (62.9% GC) including metabolic enzymes missed in a RAST annotation (Aziz et al. 2008), which carry out critical functions in the obligate symbiosis with plants of the genus *Psychotria* (Carlier et al. 2013). For *Bradyrhizobium diazoefficiens* (64.1% GC) and in combination

with dRNA-seq data, we uncovered many novel short ORFs and internal start sites expressed under free-living conditions and in symbiosis with soybean (Cuklina et al. 2016). Finally, we applied our approach on shotgun proteomics data sets from *Escherichia coli* K-12 BW25113 (50.3% GC) during exponential growth (Krug et al. 2013) or grown under multiple conditions (Schmidt et al. 2015; see Supplemental Methods). Even for the well-annotated genome of the parental strain of the Keio knockout strain collection (Baba et al. 2006), we could identify evidence for novel ORFs. These included six pseudogenes with solid expression evidence but also short in silico ORFs, including a highly conserved sORF of 57 aa (three peptides, nine PSMs) plus several novel start sites (Supplemental Tables S8, S9).

To enable proteogenomics for a larger microbiology research community with access to proteomics core facilities, we provide both a set of precomputed iPtgxDBs for several key prokaryotic model organisms, including founder strains of gene knockout collections, and the software to create them in a public web server (<https://iptgxdb.expasy.org>) for any prokaryote (Fig. 6). iPtgxDBs can thus be created for newly sequenced organisms, e.g., type strains for the 11,000 named species targeted by the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project





**Figure 5.** Protein evidence for single amino acid variations (SAAVs). (A) Genomic region encoding an ABC transporter (BH\_RS05910). RefSeq and Ensembl annotate it as a pseudogene, Genoscope as a fragmented pseudogene, while Prodigal and ChemGenome predict two CDSs. The reference genome (below gray virtual genome bar; NCBI RefSeq track) differs from the MQB277 assembly (MQB277 track above the virtual genome) by an insertion of 81 bp and a 1-bp deletion (red boxes); the 1-bp deletion causes a frameshift, evidenced by the lack of protein expression downstream from it (spectral count below the virtual genome; scaled from 0 to 800) and by transcriptomic data (reads mapped to the reference genome all support the insertion; lower panel). In contrast, the protein encoded by MQB277\_12040 in the assembly is expressed over almost its entire length (class 1a peptides; one peptide identified by seven PSMs spans the frameshift region), also supported by transcriptomic reads mapping without any mismatch (Supplemental Fig. S7). (B) Evidence for a SNV causing a nonsynonymous SAAV in the CDS of transcription elongation factor GreA. Four peptides (two, four, eight, 39 PSMs) confirming this SAAV (glycine in reference to glutamic acid in our assembly) are mapped to this position in MQB277.

(Kyrpides et al. 2014) or environmental isolates. These organisms offer unique opportunities to study various fundamental aspects such as the development and spread of antibiotics resistance (ABR), the presence of novel biochemical reactions (Montes Vidal et al. 2017) or pathways relevant for biotechnological applications, or key functions encoded by important strains isolated from complex microbiomes, whose importance for, e.g., human health (Cho and Blaser 2012) and plant protection from pathogen attack in agricultural settings (Berendsen et al. 2012), has been recognized.

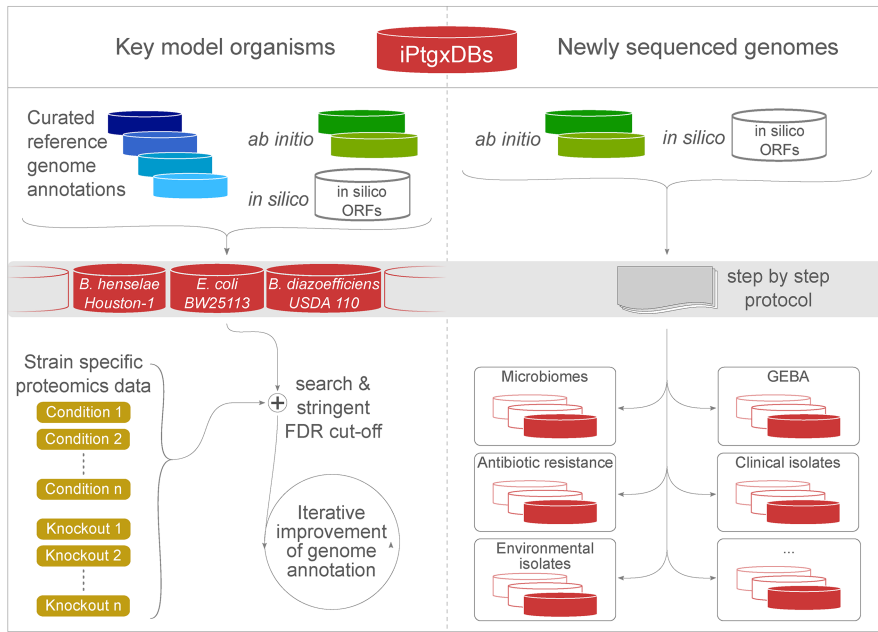
six-frame translation and later visualize identified peptides onto a genome sequence but lack integrated and consolidated annotations.

The proteogenomics community is still to agree upon the best practice for required FDR thresholds and confirmation of novel candidates. Using very stringent FDR thresholds, as also advocated by Venter et al. (2011), we show that the *E*-value distributions of PSMs for novel hits and target proteins are similar. Furthermore, we invested an extra effort to confirm the expression of novel ORFs with selective and sensitive PRM assays. The validation

## Discussion

We present a flexible, yet general proteogenomics strategy that allowed us to identify novelties in the genome of prokaryotes of different taxonomic origin ( $\alpha$ -,  $\beta$ -,  $\gamma$ -proteobacteria) and widely ranging GC content. Investing a major effort in a preprocessing step to hierarchically integrate reference genome annotations and predictions into an iPtgxDB that covers the entire protein-coding potential pays off: Close to 95% of the peptides unambiguously imply one protein (based on an extension of the PeptideClassifier concept [Qeli and Ahrens 2010] for prokaryotes), facilitating swift data analysis and mining. In addition, informative identifiers capture overlap and differences of all resources, start codon, and genomic coordinate information, such that novel ORFs, start sites, or expressed pseudogenes can readily be identified and visualized. These features are unique to our solution. Our iPtgxDBs come in the form of a protein search DB and a GFF file containing all annotations and identifiers.

For prokaryotes, the complexity of iPtgxDBs is lower than that of a regular protein search DB for, e.g., yeast or human (Supplemental Table S2). In our view, the benefit of generating a single iPtgxDB against which proteomics data from any condition (or knockout strain) can be searched to identify novel ORFs outweighs that of other elegant solutions that were developed for the more complex eukaryotes. Both splice graphs (Woo et al. 2013) and RNA-seq data (Wang et al. 2012) reduce the complexity and size of the search DB. However, in both cases, DBs specific for the conditions studied are generated, requiring bioinformatics expertise and limiting the general applicability of the resource. The GFF file we provide can be very valuable for other proteogenomics software solutions like GenoSuite (Kumar et al. 2013), PGP (Tovchigrechko et al. 2014), and PG Nexus (Pang et al. 2014), which allow users to search their data against a



**Figure 6.** Application of our integrated proteogenomics approach. We release open source iPtgxDBs for several model organisms (<https://iptgxdb.expasy.org>); here, for Bhen, *E. coli* BW25113, and *B. diazoefficiens* USDA 110 (left panel). Using proteomics data from any condition or knockout strain (light brown boxes; here, schematically shown for *E. coli*), researchers can identify novelties and iteratively improve the genome annotation, e.g., in a community-driven genome Wiki approach (Salzberg 2007). The release of the software to integrate ab initio predictor(s) and in silico predictions (Supplemental Fig. S8) can help to improve genome annotations of many newly sequenced genomes (right panel).

success (80% overall) ranged from 100% for SAAVs and highly expressed novel sORFs to around 55% for novel start sites. Reasons for the lower success with start sites can include N-terminal cleavage or modification, both of which can prevent detection of the single peptide from being confirmed (Goetze et al. 2009). Identification of internal start sites is even more difficult but greatly benefits from the availability of dRNA-seq data (Cuklina et al. 2016) and/or N-terminal enrichment steps.

Focusing on the description of our novel strategy, we did not further characterize or functionally validate novel sORFs beyond the PRM confirmation. More effort will be required to assess the functional relevance of sORFs, e.g., by individual gene deletion or genome-wide transposon mutagenesis screens (Christen et al. 2011). Recent work in yeast suggested that sORFs may represent a pool of proto-genes that are under evolutionary pressure and may lead to the birth of novel genes (Carvunis et al. 2012). Indeed, genes that emerged more recently tend to be shorter (Tautz and Domazet-Loso 2011).

Besides identifying missed sORFs, our data indicated that (1) the procedures to annotate pseudogenes differ between resources (and even releases) and have to be treated with caution, and (2) likely overpredicted ORFs can be uncovered when relying on complete, condition-specific expressed proteomes. Recent advances to comprehensively identify expressed proteomes within a few days (Nagaraj et al. 2012; Richards et al. 2015) suggest that proteomics data can, at least in part, address this issue of overprediction. Such extensive data sets can also uncover functionally relevant genomic changes down to the SAAV level, with implications for clinical proteomics and beyond. For example, by tracking clinically relevant pathogens either over time (Lee et al. 2017) or comparing different strains, genome changes that correlate with higher pathogenicity

(de Souza et al. 2011; Nasser et al. 2014; Malmstrom et al. 2015) can be identified, some of which ideally are supported by direct protein expression evidence for SAAVs.

Importantly, our data show that assembling the correct genome sequence of the strain under study is of critical importance: It is the optimal basis not only to comprehensively identify expression differences between the conditions studied but also to accurately identify novel sORFs by proteogenomics. An initial de novo assembly should thus be carried out routinely for the most important strains, in particular those that form the basis for long-term projects aiming to integrate functional genomics data.

We favor a conservative approach to genome re-annotation, ideally carried out by consortia that iteratively improve the annotation of their respective model or nonmodel organisms (Armengaud et al. 2014), e.g., relying on a genome Wiki concept (Fig. 6; Salzberg 2007). By releasing iPtgxDBs initially for three model organisms (<https://iptgxdb.expasy.org>; Supplemental File S10) and the software to create them (Supplemental File S11), we hope to enable a large user base to apply proteogenomics in the initial genome annotation step. This will provide an optimal basis for systems-wide functional studies and genome-scale regulatory or metabolic predictions and help to fully capitalize on the genome information and decode its function.

## Methods

### Source of reference genome annotations and ab initio predictors

Annotations of the Bhen reference genome (Alsmark et al. 2004) were obtained from NCBI's RefSeq (Pruitt et al. 2012) (NC\_005956.1; from 06/10/2013, called RefSeq2013, and 07/30/2015, called RefSeq2015) (for ftp links, see Supplemental Methods), from Ensembl's Genomes project (GCA\_000046705.1, Feb/2015), and from Genoscope's microbial genome annotation and analysis platform (v2.7.3, accessed 03/09/2016) (Vallenet et al. 2013). Ab initio gene predictions from Prodigal (v2.6) (Hyatt et al. 2010) and ChemGenome (v2.0, <http://www.scbio-iitd.res.in/chemgenome/chemgenomeweb.jsp>; with parameters: method, Swissprot space; length threshold, 70 nt; initiation codons, ATG, CTG, TTG, GTG) were used (Singhal et al. 2008). Files were parsed to extract the identifier, coordinates, and sequences of bona fide protein-coding sequences and pseudogene entries.

### Integrative proteogenomics approach

The annotations were collapsed into singletons (same sequence in all sources) or annotation clusters of two or more sequences with the same stop codon but different start sites. For clusters, we define an anchor sequence from the annotation highest up in the hierarchy, e.g., RefSeq2015. We construct an informative and transparent protein identifier that integrates all relevant information: A

code is added to the anchor sequence for each identical annotation (RefSeq2013=rso, Ensembl=ens, Genoscope=geno, Prodigal=prod, ChemGenome=chemg, in silico ORF=orf) separated by a pipe sign (e.g., BH\_RS00220|rso|ens|geno). Identical in silico ORFs are not considered. For alternative start sites, the length difference compared to the anchor annotation is added prior to the code (e.g., ...|-17aa\_prod|+6aa\_chemg). Finally, chromosome, start and stop position, reading frame, start codon, and CDS length complete the identifier. The anchor sequence identifier thus integrates relevant information of the genomic location and all annotation sources for this region, including possible reductions and extensions (see Fig. 2A). Identifiers for entries with alternative initiation sites contain a reference to the anchor annotation, the length difference, and the annotation source (e.g., BH\_RS00220\_+6aa\_chemg). To create an iPtgxDB, we only add the complete sequence of the anchor sequence of a cluster plus sequences of N-terminal regions that give rise to identifiable (i.e., different from the anchor sequence) tryptic peptides for the additional proteoforms of the cluster. For extensions, the N-terminal sequence up to the first tryptic cleavage site in the anchor sequence is added, for internal start sites the N-terminal tryptic peptide if it starts from an alternative initiation codon other than ATG (TTG, GTG, or CTG) giving rise to a N-terminal Met instead of a Leu, Val, or Leu, respectively. For pseudogenes, we added the suffix “\_p” (e.g., BH\_RS02905\_p) or “\_fCDS\_p” (e.g., BHGENO0333\_fCDS\_p; “fragmented CDS”, for Genoscope pseudogenes) to the identifier, and a sequence translated to the first stop codon to the protein DB. In silico ORFs above a selectable length threshold (18 aa) were added (Supplemental Methods). For the de novo-assembled Bhen MQB277 genome, Prodigal predictions and in silico ORFs were integrated (same length cut-off).

### PeptideClassifier analysis of protein search DBs

The complexity and redundancy of protein search DBs was assessed with the web-based PeptideClassifier tool (<http://peptideclassifier.expasy.org>) to derive an evidence class for every tryptic peptide of 6 to 40 aa (Supplemental Fig. S2). To deal with multiple different annotations for the same gene model, we generated the required gene-annotation mapping files using the stop codon coordinates as the common gene name across annotations, i.e., an extension of the original concept of gene-protein mapping (see text and Fig. 2; Qeli and Ahrens 2010). A web service to support peptide classification for proteogenomics in prokaryotes will soon be released.

### Stringent re-analysis of proteomics and transcriptomics data

Proteomics data (ProteomeXchange, acc.# PXD000153) was searched with MS-GF+ (v.10.0.72) (Kim and Pevzner 2014) and described parameters (Omasits et al. 2013) against the RefSeq2015-based DB, the iPtgxDB (51,541 proteins), and the iPtgxDB of our de novo assembly (52,687 proteins). A PSM FDR threshold of 0.01% was used; estimated peptide and protein level FDRs were 0.12% and 0.6%, respectively (Supplemental Table S7). Protein expression estimates and differential expression values were computed as described (Omasits et al. 2013). Reads from the matched Bhen transcriptomics data set (GEO, acc.# GSE44564) were stringently remapped both to the NCBI reference genome (Alsmark et al. 2004) and to our de novo assembly using NovoAlignCS v1.06.04 (Novocraft). For reads supporting coding SNVs, we only considered reads without a mismatch. This allowed us to provide transcriptomic support for a substantial amount of observed genomic differences, both coding and noncoding. For more details, see Supplemental Methods.

### Bacterial strains, genomic DNA, and protein extracts

High-quality gDNA was extracted from Bhen strains MQB277 and CHDE101 (Schmid et al. 2004), a close laboratory variant of the NCBI Bhen Houston-1 ATCC49882 reference strain (Alsmark et al. 2004) and parental strain of MQB277, using Sigma-Aldrich's GenElute kit. Both were sequenced (Pacific Biosciences [PacBio]) and assembled into one high-quality contig (Supplemental Methods). Protein extracts of cytoplasmic and total membrane fractions were prepared from bacterial cells grown under uninduced and induced conditions as described (Omasits et al. 2013).

### Protein features, functional annotation, conservation

Several protein features including signal peptides, transmembrane topology, lipoproteins, and protein domains were predicted, and protein sequences functionally annotated by eggNOG (Huerta-Cepas et al. 2016). Conservation of novel ORFs was assessed with tblastn. Predominant SCL information was computed for all proteins (including novel ORFs) similar to Stekhoven et al. (2014). For details, see Supplemental Methods.

### Independent validation by targeted proteomics

Peptides for novel ORFs, start sites, expressed pseudogenes, or assembly-specific changes were selected based on spectral count, number of tryptic sites, number of missed cleavage sites, and PeptideRank prediction (Qeli et al. 2014). Heavy-labeled reference peptides were purchased from JPT Peptide Technologies GmbH and used to set up PRM assays (Supplemental Table S4; Peterson et al. 2014). Specific transitions were measured in cyt and TM extracts of biological replicates of both conditions (new fractions). Only traces within a mass accuracy of 10 ppm were evaluated; we excluded transition interference by manually validating co-elution of peptide traces. For details of the sample preparation and MS set-up, see Supplemental Methods.

### Software availability

Software to create iPtgxDBs for any prokaryote is available in Supplemental File S11 and via the public web server (<https://iptgxdb.expasy.org>).

### Data access

The genome sequence of Bhen variant-1 strain CHDE101 from this study has been submitted to NCBI GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>) under accession number CP020742. iPtgxDBs for Bhen Houston-1, Bhen CHDE101, *E. coli* BW25113, and *B. diazoefficiens* USDA 110 are available in Supplemental File S10. PRM data from this study have been submitted to Panoramaweb ([https://panoramaweb.org/labkey/Bartonella\\_Proteogenomics.url](https://panoramaweb.org/labkey/Bartonella_Proteogenomics.url)).

### Acknowledgments

We thank A. Wicki for initial work on the genome assemblies; R. Schlapbach for access to the FGCZ and continued support; F. Freimoser, G. Pessi, and three anonymous reviewers for valuable input on the manuscript; and sciCORE for hosting the iPtgxDB website. The project was supported by B.W. through the D-HEST BioMedical Proteomics Platform (BMPP). C.H.A. acknowledges funding for U.O. and A.R.V. from the SNSF under grant 31003A-156320.

*Author contributions:* U.O. and C.H.A. devised the integrated proteogenomics strategy; U.O., A.R.V., and C.H.A. analyzed data; U.O., A.R.V., and M.S. made figures; A.R.V., M.S., and D.M.

assembled and annotated the genomes with input from A.P., J.E.F., O.N., and M.D.R.; M.Q. and C.D. provided Bhen gDNA and new protein extracts; S.G. and B.W. validated selected candidates by PRM; M.B. developed the iPTgxDB website with input from A.R.V. and C.H.A.; C.H.A. wrote the manuscript. All authors commented on the manuscript.

## References

- Ahrens CH, Brunner E, Qeli E, Basler K, Aebersold R. 2010. Generating and navigating proteome maps using mass spectrometry. *Nat Rev Mol Cell Biol* **11**: 789–801.
- Alsmark CM, Frank AC, Karlberg EO, Legault BA, Ardell DH, Canback B, Eriksson AS, Naslund AK, Handley SA, Huvet M, et al. 2004. The louse-borne human pathogen *Bartonella quintana* is a genomic derivative of the zoonotic agent *Bartonella henselae*. *Proc Natl Acad Sci* **101**: 9716–9721.
- Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM. 2014. Non-model organisms, a species endangered by proteogenomics. *J Proteomics* **105**: 5–18.
- Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, et al. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**: 75.
- Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**: 2006.0008.
- Bakke P, Carney N, Deloache W, Gearing M, Ingvorsen K, Lotz M, McNair J, Penumetcha P, Simpson S, Voss L, et al. 2009. Evaluation of three automated genome annotations for *Halorhabdus utahensis*. *PLoS One* **4**: e6291.
- Basrai MA, Hieter P, Boeke JD. 1997. Small open reading frames: beautiful needles in the haystack. *Genome Res* **7**: 768–771.
- Berendsen RL, Pieterse CM, Bakker PA. 2012. The rhizosphere microbiome and plant health. *Trends Plant Sci* **17**: 478–486.
- Blakeley P, Overton IM, Hubbard SJ. 2012. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J Proteome Res* **11**: 5221–5234.
- Carlier AL, Omasits U, Ahrens CH, Eberl L. 2013. Proteomics analysis of *Psychotria* leaf nodule symbiosis: improved genome annotation and metabolic predictions. *Mol Plant Microbe Interact* **26**: 1325–1333.
- Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and *de novo* gene birth. *Nature* **487**: 370–374.
- Castellana N, Bafna V. 2010. Proteogenomics to discover the full coding content of genomes: a computational perspective. *J Proteomics* **73**: 2124–2135.
- Chapman B, Bellgard M. 2014. High-throughput parallel proteogenomics: a bacterial case study. *Proteomics* **14**: 2780–2789.
- Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**: 563–569.
- Cho I, Blaser MJ. 2012. The human microbiome: at the interface of health and disease. *Nat Rev Genet* **13**: 260–270.
- Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, Collier JA, Fero MJ, McAdams HH, Shapiro L. 2011. The essential genome of a bacterium. *Mol Syst Biol* **7**: 528.
- Cuklina J, Hahn J, Imakaev M, Omasits U, Forstner KU, Ljubimov N, Goebel M, Pessi G, Fischer HM, Ahrens CH, et al. 2016. Genome-wide transcription start site mapping of *Bradyrhizobium japonicum* grown free-living or in symbiosis – a rich resource to identify new transcripts, proteins and to study gene regulation. *BMC Genomics* **17**: 302.
- Davidson T, Beck E, Ganapathy A, Montgomery R, Zafar N, Yang Q, Madupu R, Goetz P, Galinsky K, White O, et al. 2010. The comprehensive microbial resource. *Nucleic Acids Res* **38**: D340–D345.
- de Groot A, Dulerio R, Ortet P, Blanchard L, Guerin P, Fernandez B, Vacherie B, Dossat C, Jolivet E, Siguier P, et al. 2009. Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium *Deinococcus deserti*. *PLoS Genet* **5**: e1000434.
- de Souza GA, Arntzen MO, Wiker HG. 2010. MSMSpddb: providing protein databases of closely related organisms to improve proteomic characterization of prokaryotic microbes. *Bioinformatics* **26**: 698–699.
- de Souza GA, Arntzen MO, Fortuin S, Schurch AC, Malen H, McEvoy CR, van Soelingen D, Thiede B, Warren RM, Wiker HG. 2011. Proteogenomic analysis of polymorphisms and gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. *Mol Cell Proteomics* **10**: M110 002527.
- Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**: 673–679.
- Dinger ME, Pang KC, Mercer TR, Mattick JS. 2008. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol* **4**: e1000176.
- Goetze S, Qeli E, Mosimann C, Staes A, Gerrits B, Roschitzki B, Mohanty S, Niederer EM, Laczko E, Timmerman E, et al. 2009. Identification and functional characterization of N-terminally acetylated proteins in *Drosophila melanogaster*. *PLoS Biol* **7**: e1000236.
- 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**: 1362–1377.
- Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. 2008. Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol* **70**: 1487–1501.
- Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattai T, Mende DR, Sunagawa S, Kuhn M, et al. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**: D286–D293.
- Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.
- Ingolia NT. 2014. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* **15**: 205–213.
- Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T, FitzGerald MG, Hafez N, et al. 2004. The complete genome and proteome of *Mycoplasma mobile*. *Genome Res* **14**: 1447–1461.
- Kim S, Pevzner PA. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* **5**: 5277.
- Kovacs-Simon A, Titball RW, Michell SL. 2011. Lipoproteins of bacterial pathogens. *Infect Immun* **79**: 548–561.
- Krug K, Carpy A, Behrends G, Matic K, Soares NC, Macek B. 2013. Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol Cell Proteomics* **12**: 3420–3430.
- Kucharova V, Wiker HG. 2014. Proteogenomics in microbiology: taking the right turn at the junction of genomics and proteomics. *Proteomics* **14**: 2360–2675.
- Kumar D, Yadav AK, Kadimi PK, Nagaraj SH, Grimmond SM, Dash D. 2013. Proteogenomic analysis of *Bradyrhizobium japonicum* USDA110 using GenoSuite, an automated multi-algorithmic pipeline. *Mol Cell Proteomics* **12**: 3388–3397.
- Kyrpides NC, Hugenholtz P, Eisen JA, Woyke T, Goker M, Parker CT, Amann R, Beck BJ, Chain PS, Chun J, et al. 2014. Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol* **12**: e1001920.
- Lee AH, Flibotte S, Sinha S, Paiero A, Ehrlich RL, Balashov S, Ehrlich GD, Zlosnik JE, Mell JC, Nislow C. 2017. Phenotypic diversity and genotypic flexibility of *Burkholderia cenocepacia* during long-term chronic infection of cystic fibrosis lungs. *Genome Res* **27**: 650–662.
- Lindroos H, Vinnere O, Mira A, Repsilber D, Naslund K, Andersson SG. 2006. Genome rearrangements, deletions, and amplifications in the natural population of *Bartonella henselae*. *J Bacteriol* **188**: 7426–7439.
- Lu YY, Franz B, Truttmann MC, Riess T, Gay-Fraret J, Faustmann M, Kempf VA, Dehio C. 2013. *Bartonella henselae* trimeric autotransporter adhesin BadA expression interferes with effector translocation by the VirB/D4 type IV secretion system. *Cell Microbiol* **15**: 759–778.
- Luo C, Hu GQ, Zhu H. 2009. Genome reannotation of *Escherichia coli* CFT073 with new insights into virulence. *BMC Genomics* **10**: 552.
- Malmstrom L, Bakochi A, Svensson G, Kilsgard O, Lantz H, Petersson AC, Hauri S, Karlsson C, Malmstrom J. 2015. Quantitative proteogenomics of human pathogens using DIA-MS. *J Proteomics* **129**: 98–107.
- Marcellin E, Licona-Cassani C, Mercer TR, Palfreyman RW, Nielsen LK. 2013. Re-annotation of the *Saccharopolyspora erythraea* genome using a systems biology approach. *BMC Genomics* **14**: 699.
- Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC. 2009. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* **25**: 2271–2278.
- Marx H, Lemeer S, Klaeger S, Rattai T, Kuster B. 2013. MSDB: a mass spectrometry-centric protein sequence database for proteomics. *J Proteome Res* **12**: 2386–2398.
- Menschaert G, Fenyo D. 2015. Proteogenomics from a bioinformatics angle: a growing field. *Mass Spectrom Rev* **36**: 584–599.
- Montes Vidal D, von Rymon-Lipinski AL, Ravella S, Groenhagen U, Herrmann J, Zaburanyi N, Zarbin PH, Varadarajan AR, Ahrens CH, Weisskopf L, et al. 2017. Long-chain alkyl cyanides: unprecedented volatile compounds released by *Pseudomonas* and *Micromonospora* bacteria. *Angew Chem Int Ed Engl* **56**: 4342–4346.

- Nagaraj N, Kulak NA, Cox J, Neuhauser N, Mayr K, Hoerning O, Vorm O, Mann M. 2012. System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol Cell Proteomics* **11**: M111.013722.
- Nasser W, Beres SB, Olsen RJ, Dean MA, Rice KA, Long SW, Kristinsson KG, Gottfredsson M, Vuopio J, Raisanen K, et al. 2014. Evolutionary pathway to increased virulence and epidemic group A Streptococcus disease derived from 3,615 genome sequences. *Proc Natl Acad Sci* **111**: E1768–E1776.
- Nesvizhskii AI. 2014. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **11**: 1114–1125.
- Noble WS. 2015. Mass spectrometrists should search only for peptides they care about. *Nat Methods* **12**: 605–608.
- Omasits U, Quebatte M, Stekhoven DJ, Fortes C, Roschitzki B, Robinson MD, Dehio C, Ahrens CH. 2013. Directed shotgun proteomics guided by saturated RNA-Seq identifies a complete expressed prokaryotic proteome. *Genome Res* **23**: 1916–1927.
- Pang CN, Tay AP, Aya C, Twine NA, Harkness L, Hart-Smith G, Chia SZ, Chen Z, Deshpande NP, Kaakoush NO, et al. 2014. Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. *J Proteome Res* **13**: 84–98.
- Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, Kyrpides NC. 2010. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods* **7**: 455–457.
- Payne SH, Huang ST, Pieper R. 2010. A proteogenomic update to *Yersinia*: enhancing genome annotation. *BMC Genomics* **11**: 460.
- Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**: 3551–3567.
- Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ. 2014. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol Cell Proteomics* **11**: 1475–1488.
- Petty NK. 2010. Genome annotation: man versus machine. *Nat Rev Microbiol* **8**: 762.
- Poole FL II, Gerwe BA, Hopkins RC, Schut GJ, Weinberg MV, Jenney FE Jr, Adams MW. 2005. Defining genes in the genome of the hyperthermophilic archaeon *Pyrococcus furiosus*: implications for all microbial genomes. *J Bacteriol* **187**: 7325–7332.
- Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**: D130–D135.
- Qeli E, Ahrens CH. 2010. PeptideClassifier for protein inference and targeted quantitative proteomics. *Nat Biotechnol* **28**: 647–650.
- Qeli E, Omasits U, Goetze S, Stekhoven DJ, Frey JE, Basler K, Wollscheid B, Brunner E, Ahrens CH. 2014. Improved prediction of peptide detectability for targeted proteomics using a rank-based algorithm and organism-specific data. *J Proteomics* **108**: 269–283.
- Reddy TB, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, Mallajosyula J, Pagani I, Lobos EA, Kyrpides NC. 2015. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res* **43**: D1099–D1106.
- Renuse S, Chaerkady R, Pandey A. 2011. Proteogenomics. *Proteomics* **11**: 620–630.
- Richards AL, Hebert AS, Ulbrich A, Bailey DJ, Coughlin EE, Westphall MS, Coon JJ. 2015. One-hour proteome analysis in yeast. *Nat Protoc* **10**: 701–714.
- Richardson EJ, Watson M. 2012. The automatic annotation of bacterial genomes. *Brief Bioinform* **14**: 1–12.
- Riess T, Andersson SG, Lupas A, Schaller M, Schafer A, Kyme P, Martin J, Walzlein JH, Ehehalt U, Lindroos H, et al. 2004. *Bartonella* adhesin A mediates a proangiogenic host cell response. *J Exp Med* **200**: 1267–1278.
- Risk BA, Spitzer WJ, Giddings MC. 2013. Peppy: proteogenomic search software. *J Proteome Res* **12**: 3019–3025.
- Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative Genomics Viewer. *Nat Biotechnol* **29**: 24–26.
- Rowland SL, Burkholder WF, Cunningham KA, Maciejewski MW, Grossman AD, King GF. 2004. Structure and mechanism of action of Sda, an inhibitor of the histidine kinases that regulate initiation of sporulation in *Bacillus subtilis*. *Mol Cell* **13**: 689–701.
- Salzberg SL. 2007. Genome re-annotation: a wiki solution? *Genome Biol* **8**: 102.
- Schmid MC, Schulein R, Dehio M, Denecker G, Carena I, Dehio C. 2004. The VirB type IV secretion system of *Bartonella henselae* mediates invasion, proinflammatory activation and antiapoptotic protection of endothelial cells. *Mol Microbiol* **52**: 81–92.
- Schmidt A, Kochanowski K, Vedelaar S, Ahne E, Volkmer B, Callipo L, Knoop K, Bauer M, Aebersold R, Heinemann M. 2015. The quantitative and condition-dependent *Escherichia coli* proteome. *Nat Biotechnol* **34**: 104–110.
- Singhal P, Jayaram B, Dixit SB, Beveridge DL. 2008. Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. *Biophys J* **94**: 4173–4183.
- Stekhoven DJ, Omasits U, Quebatte M, Dehio C, Ahrens CH. 2014. Proteome-wide identification of predominant subcellular protein localizations in a bacterial model organism. *J Proteomics* **99**: 123–137.
- Storz G, Wolf YI, Ramamurthi KS. 2014. Small proteins can no longer be ignored. *Annu Rev Biochem* **83**: 753–777.
- Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet* **12**: 692–702.
- Tovchigrechko A, Venepally P, Payne SH. 2014. PGP: parallel prokaryotic proteogenomics pipeline for MPI clusters, high-throughput batch clusters and multicore workstations. *Bioinformatics* **30**: 1469–1470.
- Vallenet D, Belda E, Calteau A, Cruveiller S, Engelen S, Lajus A, Le Fevre F, Longin C, Mornico D, Roche D, et al. 2013. MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res* **41**: D636–D647.
- Venter E, Smith RD, Payne SH. 2011. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *PLoS One* **6**: e27587.
- Wang X, Slebos RJ, Wang D, Halvey PJ, Tabb DL, Liebler DC, Zhang B. 2012. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J Proteome Res* **11**: 1009–1017.
- Warren AS, Archuleta J, Feng WC, Setubal JC. 2010. Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics* **11**: 131.
- Woo S, Cha SW, Merrihew G, He Y, Castellana N, Guest C, Maccoss M, Bafna V. 2013. Proteogenomic database construction driven from large scale RNA-seq data. *J Proteome Res* **13**: 21–28.
- Zickmann F, Renard BY. 2015. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. *Bioinformatics* **31**: i106–i115.
- Zuber P. 2001. A peptide profile of the *Bacillus subtilis* genome. *Peptides* **22**: 1555–1577.

Received November 11, 2016; accepted in revised form October 25, 2017.