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


1 Agroscope, Research Group Molecular Diagnostics, Genomics and Bioinformatics & SIB Swiss Institute of Bioinformatics, CH-8820 Wädenswil, Switzerland; 2 Department of Health Sciences and Technology, Institute of Molecular Systems Biology, Swiss Federal Institute of Technology Zurich, CH-8093 Zurich, Switzerland; 3 Institute for Molecular Life Sciences & SIB Swiss Institute of Bioinformatics, University of Zurich, CH-8057 Zurich, Switzerland; 4 Biozentrum, University of Basel, CH-4056 Basel, Switzerland; 5 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 novelies 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.]
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), MSMSpdbb (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 disentangle 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 α-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 MSMSpdbb, 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
integratio 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 integrative 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 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; Cukлина 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.

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
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 include many sORFs potentially missed in the reference annotations (Fig. 2B).

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 include many sORFs potentially missed in the reference annotations (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

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 (RefSeq2015 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.
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.

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

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

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 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

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

<table>
<thead>
<tr>
<th>Annotation source</th>
<th>Novel protein-coding ORF</th>
<th>Evidence for alternative NH2-terminus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not in RefSeq2015</td>
<td>Pseudogene in RefSeq2015</td>
</tr>
<tr>
<td>RefSeq2013/Ensembl</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Genoscope</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>ChemGenome</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Prodigal</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>In silico ORFs</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>15</td>
</tr>
</tbody>
</table>

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).
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.
We also identified a highly expressed RefSeq2015 pseudogene um-efflux transporter (BH10840) and the transcription elongation factor NusA) by RefSeq2013, Ensembl, and protein BepD (BH13410) (Supplemental Table S3). Finally, for lipoproteins, they could represent a class of proteins for which a substantial percentage is missed in reference genome annotations, which is relevant given their importance among the six novel ORFs identified previously, which have since been incorporated in RefSeq2015. All four candidates were validated by independent targeted proteomics experiments (Supplemental Table S3). Lipoproteins are among the most up-regulated (42-fold) proteins 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. SSA). 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. SSB; Supplemental Table S3) and an in silico only predicted sORF of 34 aa (Supplemental Fig. SSD). 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 (BH110840) and the Bartonella effector protein BepD (BH13410) (Supplemental Table S3). 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 computationally 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).

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. SSA–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 computationally 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 SpI 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).
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 reanalyzed 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; Marcellin 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 Archea (GEBA) project.
Kyrpides et al. (2014) or environmental isolates. These organisms offer unique opportunities to study various fundamental aspects such as the development and spread of antibiotic 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.

**Discussion**

We present a flexible, yet general proteogenomics strategy that allowed us to identify novelties in the genome of prokaryotes of different taxonomic origin (α-, β-, γ-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 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 ORFs 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

**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.
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 IPTxDBs initially for three model organisms (https://iptxdb.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.scbioitd.res.in/chemgenome/chemgenomenew.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) 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 based on Stringent re-analysis of proteomics and transcriptomics data

Proteomics data (ProteomeXchange, acc.# PXD000153) was searched with MS-GF+ (v.10.0.7.2) (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 value computations were described 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 NovaAlignCS 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 unduced 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 Panorama (https://panoramaweb.org/labkey/Bartonella_Proteogenomics.url).

**Acknowledgments**

We thank A. Wicki for initial work on the genome assemblies; R. Schlabach 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 (BMP). C.H.A. acknowledges funding for U.O. and A.R.V. from the SNSF under grant 31003A-156320.

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 Bhgen DNA 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

Omas et al.


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

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


*Genome Res.* 2017 27: 2083-2095 originally published online November 15, 2017

Access the most recent version at doi:10.1101/gr.218255.116

**Supplemental Material**

http://genome.cshlp.org/content/suppl/2017/11/15/gr.218255.116.DC1

**References**

This article cites 84 articles, 16 of which can be accessed free at:

http://genome.cshlp.org/content/27/12/2083.full.html#ref-list-1

**Open Access**

Freely available online through the *Genome Research* Open Access option.

**Creative Commons License**

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/.

**Email Alerting Service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to *Genome Research* go to:

http://genome.cshlp.org/subscriptions

© 2017 Omasits et al.; Published by Cold Spring Harbor Laboratory Press