Supplemental Information

Table of Contents

Supplemental Methods

Genome sequencing and *de novo* genome assembly

The Bhen reference genome contains long, almost identical repeats and is classified as a difficult to assemble class III genome (Koren et al. 2013). We thus used size-selected long gDNA fragments (BluePippin) to *de novo* assemble the genomes of MQB277 and its parental strain Bhen CHDE101, a laboratory variant of Bhen ATCC49882 (Schmid et al. 2004). MQB277 gDNA was sheared, size-selected (BluePippin, >15 kbp insert size) and sequenced with two SMRT cells (PacBio RSII platform; P6-C4 chemistry). Quality filtering, genome assembly, and polishing steps were performed using PacBio SMRT Portal 2.3.0 (Chin et al. 2013). After the first quality filtering, 131,221 reads with a N50 of 10,995 bp were obtained and *de novo* assembled using protocol RS_HGAP_Assembly.3. Terminal repeats were removed, the genome was circularized using Circlator v1.1.1 (Hunt et al. 2015), prior to several rounds of sequence polishing with stringent filter criteria ("Minimum Polymerase Read Quality": 86), resulting in one 1,954,773 bp contig with a mean coverage of 104-fold. A quality assessment of this assembly with respect to the reference NC_005956.1 was performed using QUAST v4.0 (Gurevich et al. 2013) and the *de novo* assembly likelihood estimator, ALE v0.9 (Clark et al. 2013). To remove potential homopolymer errors that may occur in PacBio assemblies (Ross et al. 2013) we sequenced a mate-pair library (Nextera) with an Illumina MiSeq (2 × 300 bp) and mapped reads to the PacBio assembly using BWA-MEM v0.7.12 (Li 2013). Four single nucleotide (nt) insertions and four single nt deletions were corrected in the final assembly, which was start-aligned with the *Bhen* Houston-1 reference sequence to simplify downstream comparative analyses. A similar approach was followed to assemble the strain Bhen CHDE101 into one high quality contig (1,955,425 bp contig, 83-fold coverage), here without additional MiSeq data.

Virtual genome approach to compare closely related genomes

To compare and visualize both reference genome and actual assembly in the context of integrated experimental evidence, we manually processed the result of a pairwise global sequence alignment from EMBOSS Stretcher into a virtual genome sequence containing all nucleotides from both genomes (i.e. including insertions and deletions), and created a combined coordinate system. Feature tracks (GFF and WIG format) can then be projected to this virtual genome allowing us to analyze e.g. gene annotation, transcriptomics and proteomics evidence simultaneously in the context of both genomes (see Figure 4, main text).

In silico ORFs for iPtgxDBs

The *in silico* ORF annotation was generated by scanning all six frames of the genome sequence for the longest possible ATG-initiated ORFs. Extensions to such ORFs are considered for the alternative start codons TTG, GTG, and CTG. For regions between two stop codons without ATG codon, the longest alternative start codon initiated ORF is considered. Only ORFs above a selectable length threshold (here 18 aa) are considered. The length cut-off for *in silico* ORFs can be adapted; for Bhen, a search against an iPtgxDB based on a 10 aa cut-off did not identify additional sORFs. From the final integrated protein search DBs, all extensions or internal start peptides smaller than 6 aa were excluded (see Table 1). We encourage researchers to contact us with suggestions to create iPtgxDBs for additional model organisms.

Stringent re-analysis of proteomics and transcriptomics data

The NCBI does not use release numbers for RefSeq database versions. Annotations are available from these links, respectively for RefSeq2013 (first link: ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/Bartonella_henselae_Houston_1_uid57745/;ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/Bartonella_henselae/all_assembly_versions/GCA_000046705.1_ASM4670v1) and RefSeq2015 (second link). The protein search DBs also contained sequences from 349 sheep contaminants (identified in a prior search against sheep proteins present in RBA plates used to grow Bhen), 17 proteins encoded by genes of the complementation plasmid (pDT024) of strain MQB307 (MQB277 with pDT024; Omasits et al. 2013), a positive control (myc-*gfp*) and sequences of 256 common contaminants.

Reads from the matched Bhen transcriptomics dataset (Omasits et al. 2013) were pooled (two replicates each for uninduced and induced condition) and stringently re-mapped both to the NCBI reference genome (Alsmark et al. 2004) and to our *de novo* assembly using NovoalignCS v1.06.04 (Novocraft, Selangor, Malaysia). Color space based error correction (part of the NovoalignCS algorithm) and quality based filtering as described (Omasits et al. 2013) were applied before specifically extracting the coverage information for protein coding genomic positions in both reference and *de novo* genome sequence using the HTSeq package v0.6.1 (Anders et al. 2015). Count summary for the reads mapping to *in silico* ORFs encoded in the reference sequence were extracted separately due to overlaps with other genome annotations. 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 non-coding.

Confirmation by parallel reaction monitoring (PRM)

For tryptic digestion Bhen protein extracts from cytoplasmic (cyt) and total membrane (TM) fractions were precipitated with 80% acetone before reduction. Alkylation was carried out with 10 mM iodoacetamide to modify cysteine residues. After trypsin digestion samples were purified by reverse phase C-18 chromatography (Sep-PacK, Waters). For mass spectrometry analysis samples were resuspended in 3% acetonitrile, 0.1% formic acid. The synthetic peptide sequences ordered through JPT Peptide Technologies GmbH (Berlin, Germany) were validated via shotgun-MS2 analysis. Parallel Reaction Monitoring (PRM) (Peterson et al. 2014) was performed on an Orbitrap Fusion Mass Spectrometer (Thermo Fisher Scientific) coupled to a Proxeon EASY-nLC. Columns were packed in house with ReproSil-Pur C18-AQ, 1.9 µm (15 cm length × 75 µm (internal diameter). Peptides were loaded onto the column with 100% buffer A (99.9% H2O, 0.1% FA) and eluted at a constant flow rate of 300 nl/min with a 80 min linear gradient from 3–25% buffer B (99.9% ACN, 0.1% FA) and 10 min 25-50%B. 138 peptides were tested (107 confirmed, 78%), plus 11 peptides as retention time standards (149 overall).

Protein features, SCL and functional annotation

Various protein features were computed for 2932 integrated reference genome CDS (including selected *in silico* ORFs identified with 4 or more PSMs) and for 1697 CDS of the assembly (Prodigal predictions, plasmid proteins, and *in silico* ORFs/ORF extensions identified by 2 PSMs or more): transmembrane topology by Phobius (v1.01) (Kall et al. 2007), signal peptides by SignalP (v4.1) (Petersen et al. 2011), lipoproteins and their corresponding signal peptides by the LipoP server (v1.0) (Juncker et al. 2003), and various protein features including motifs, domains, etc. with InterProScan, v5.13-52.0 (Jones et al. 2014). Furthermore, protein sequences were functionally annotated with the EggNOG DB (v4.5) (Huerta-Cepas et al. 2016), considering the NOG, bactNOG, proNOG and aproNOG levels. Finally, experimental predominant SCL information was computed similar as before: spectral counts were normalized for the different amounts of protein present in the samples and P-scores were calculated using the same cut-offs and binomial tests as described (Stekhoven et al. 2014). In many cases, the experimental data from Cyt = cytoplasmic, TM = total membrane, IM = inner membrane, OM = outer membrane fractions provided SCL information where the software PSORTb (v3.0.2) (Yu et al. 2010) had no prediction (Unknown). The conservation of novel protein-coding ORFs was assessed by performing a tblastn search (blast v2.2.31+) of their protein sequence against NCBI's non-redundant 'nt' DB, restricted to the GI list of proteobacteria. Hits below an e-value cut-off of 1e-5 and with a query coverage per HSP of at least 30% (remaining parameters were set to default) were analyzed using in house scripts.

Source of *E. coli* BW25113 reference genome annotations and proteomics data

Annotations of the *E. coli* K-12 strain BW25113 reference genome (Grenier et al. 2014), the parent strain of the Keio knockout collection, were obtained from the following resources: a GenBank file from NCBI's RefSeq (CP009273.1; from 30/10/2014 called RefSeq), one from the Integrated Microbial Genomes (IMG) initiative of the Joint Genome Institute (JGI) (Markowitz et al. 2012) (Ga0058822; from 12/08/2014 called JGI). *Ab initio* gene predictions from Prodigal and ChemGenome were added using the same parameters as for Bhen, and *in silico* ORFs of 18 aa or longer. The iPtgxDB was created using the hierarchy RefSeq > JGI > Prodigal > ChemGenome > *in silico*. In contrast to Bhen and *B. diazofficiens*, the iPtgxDB for *E. coli* K-12 BW25113 was created without including annotated pseudogenes, as this makes the approach to create iPtgxDBs more robust. Expression evidence for pseudogenes in RefSeq, identified as novel protein coding ORF, was confirmed by visually checking the search results against a GFF file that included the RefSeq annotation (which contains the pseudogenes). Data from existing shotgun proteomics studies (Krug et al. 2013; Schmidt et al. 2015) were searched against the *E. coli* iPtgxDB and stringently processed as described (see Supplemental Table S8).

Annotation source	No. of annotations	New clusters	New extensions	New reductions	Cumulative clusters	Cumulative annotations
RefSeq	4132	4130	2	-	4130	4132
JGI (IMG)	4267	230	94	143	4360	4599
Prodigal	4301	20	2	3	4380	4624
ChemGenome	9007	4903	1727	58	9283	11,312
in silico	135,979	77,590	37,701	-	86,873	126,603

Overview of results of the stepwise, hierarchical integration of resources for *E. coli* BW25113.

Examples for iPtgxDB identifiers and how to "interpret" them

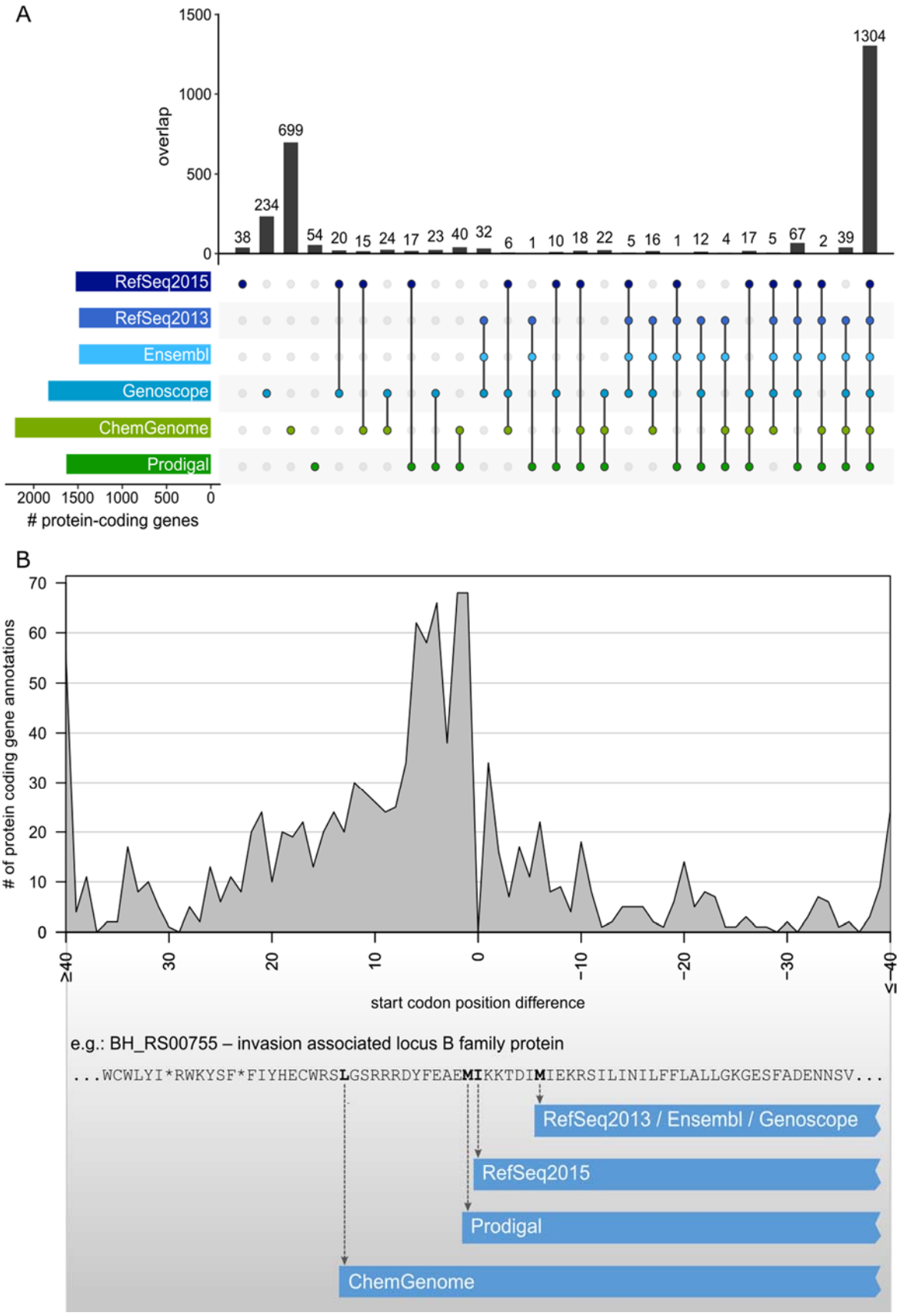
The protein BHGENO0898|-21aa_chemg|prod|NC_005956_922641_922847_+3_ATG_68 (Figure 3B) was identified by 6 unambiguous peptides: IIQYITDMSK, SPLLALLLDMAAK, IIQYITDMSKQMNQMAR, MFINHLKIIQYITDMSK, MFINHLKIIQYITDMSKQMNQMAR, EGQSIINSAEALGEDQNSTMTHNHQSVE. Two peptides (IIQYITDMSK, SPLLALLLDMAAK) were selected for validation because they do not have missed cleavage sites and they are less than 20 aa long. Both peptides could be validated in PRM assays.

This iPtgxDB identifier indicates it is a novel ORF in Bhen (BH) predicted by Genoscope (GENO). The additional annotation sources encoded in the identifier imply that the corresponding annotation cluster consists of a shorter ChemGenome prediction (-21aa... 21

aa shorter) and an identical Prodigal prediction, but no annotation in RefSeq. The encoding gene is located on the NC_005956 chromosome, starting at 922,641 bp and ending at 922,847 bp in the +3 reading frame. The initiation codon is ATG and the protein is 68 aa long.

The protein BHPROD|NC_005956_248423_248274_-3_ATG_49 (Supplemental Fig S5B) was identified by 9 unambiguous peptides: MNTYNNGKTTSSK, NTYNNGK, HNTLPSNNGK, HNTLPSNNGKK, HNTLPSNNGKKSIQSDQANTR, KSIQSDQANTR, KSIQSDQANTRQR, SIQSDQANTR, SIQSDQANTRQR. Two peptides without missed cleavage sites (SIQSDQANTR, HNTLPSNNGK) were selected and could be validated in PRM assays.
This iPtgxDB identifier indicates it is a novel ORF in Bhen (BH), in this case solely predicted by Prodigal (PROD), as no additional annotation sources are listed in the identifier. The encoding gene is located on the NC_005956 chromosome, starting at 248,423 bp and ending at 248,274 bp in the -3 reading frame. The initiation codon is ATG and the protein is 49 aa long.
BH_RS01750|-7aa_rso|+63aa_chemg|prod|NC_005956_416783_416364_-3_TTG_139 is the main ID (see Figure 3D) of a complex CDS annotation cluster. This iPtgxDB identifier indicates it is an annotated protein-coding ORF in Bhen (BH), here by RefSeq2015 (RS). The additional annotation sources encoded in the identifier imply that the corresponding annotation cluster consists of an alternative start site in RefSeq2013 annotation (-7aa_rso...7 aa shorter), an alternative start site in ChemGenome annotation (+63aa_chemg… 63 aa longer), and an identical Prodigal prediction. The encoding gene is located on the NC_005956 chromosome, starting at 416,783 bp and ending at 416,364 bp in the -3 reading frame. The initiation codon is TTG and the protein is 139 aa long.
The 7 aa shorter variant of the annotation cluster (BH_RS01750_-7aa_rso|ens|geno|NC_005956_416762_416364_-3_ATG_132) is annotated by RefSeq2013 (rso), Ensembl (ens) and Genoscope (geno). It starts with the standard start codon (ATG) and is 132 aa long. As the usage of the standard start codon does not result in a difference in the protein sequence it is not possible to unambiguously identify this proteoform. Therefore, this shorter variant will not be added to the iPtgxDB.
The 63 aa longer variant of the annotation cluster (BH_RS01750_+63aa_chemg|NC_005956_416972_416364_-3_ATG_202, Figure 3D) is predicted by ChemGenome. It is 202 aa (63 aa longer than the main ID with 139 aa above) and will be represented in the iPtgxDB with a protein sequence up to the first tryptic cleavage site within the anchor sequence of this annotation cluster (here: 75 aa). This sequence was identified by 3 unambiguous peptides: MEHKPIPK, QDNSAVVPIELHNTCPISK, and IISLGNEALQMLDEK. The latter peptide (IISLGNEALQMLDEK) could be independently confirmed in PRM assays.

Supplemental Fig S1. Large differences among Bhen reference genome annotations



(A) UpSet bar diagram (Lex et al. 2014) with a detailed overview how many of the CDSs predicted for Bhen by two NCBI RefSeq releases (RefSeq2013, RefSeq2015), Genoscope (Vallenet et al. 2013), Ensembl (Kersey et al. 2012), Prodigal (Hyatt et al. 2010) and ChemGenome (Singhal et al. 2008) (overall 2725 CDSs) range from being uniquely predicted (first 4 columns: 38 for RefSeq2015, 54 for Prodigal, 234 for Genoscope, and 699 for ChemGenome) to commonly predicted by all six (1304, last column). The overlap criterion is a matched stop codon; pseudogenes were not considered. Compared to RefSeq2013, most other annotation resources predicted more CDSs and/or different start sites including a substantial number (1025, 38%) of uniquely predicted CDSs: Only about 48% of CDSs were annotated/predicted completely identical by all six resources. RefSeq2013 and Ensembl are - at least for Bhen - almost identical and only contribute unique start sites.

(B) Histogram of total length differences for predicted protein start sites. The length difference was calculated from pairwise comparisons of annotated proteins relative to their respective cluster's anchor sequence (see main text). Exactly matching annotations (i.e. a difference of zero) are omitted for clarity.
Annotations for the same protein-coding gene that differ with respect to their predicted start codon/initiation site form annotation clusters, as illustrated for the cluster with the anchor sequence BH_RS00755 (invasion associated locus B family protein; RefSeq2015): RefSeq2013, Ensembl, and Genoscope all predicted a shorter protein (by 6 aa), while Prodigal and ChemGenome predicted a longer protein (by 1 and 13 aa, respectively). Alternative start sites (bold letters) are shown with part of the aa sequence in the correct reading frame (drawn to scale of the histogram), and N-terminal parts of the corresponding proteins (blue bars). Most start site differences affect initiation codons further upstream, leading to longer protein sequences. The RefSeq2015 and ChemGenome annotations used alternative start codons (here AUA and GUG, respectively, Supplemental Fig S1B); these should also be considered by a generic proteogenomics approach.

Supplemental Fig S2. Creating integrated search DBs with high peptide information content

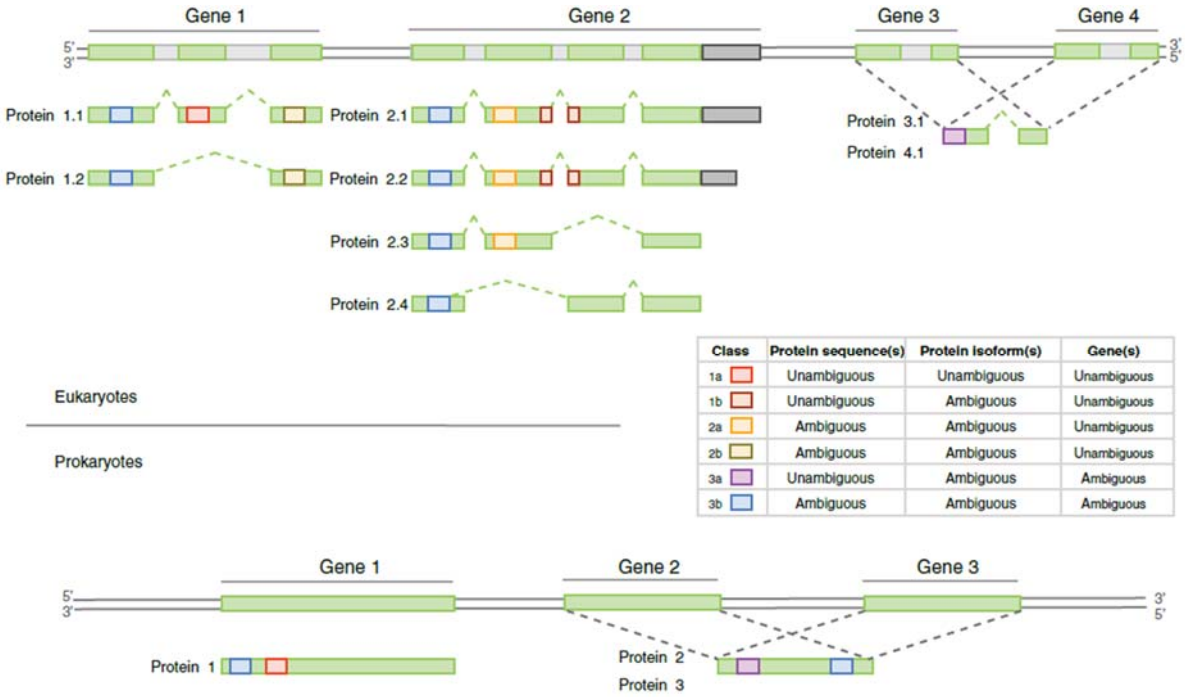
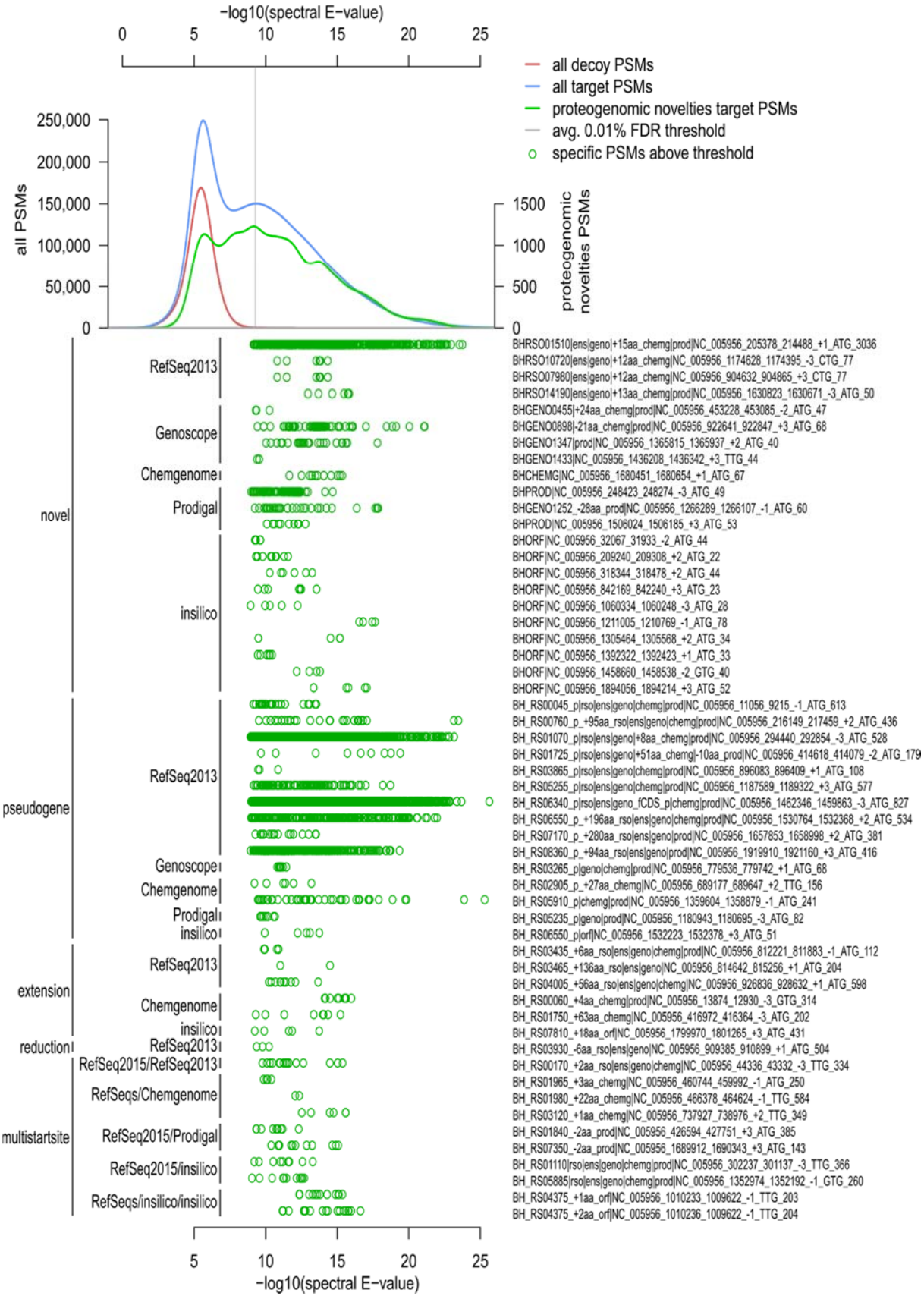


Figure reproduced with permission from (Qeli and Ahrens 2010).

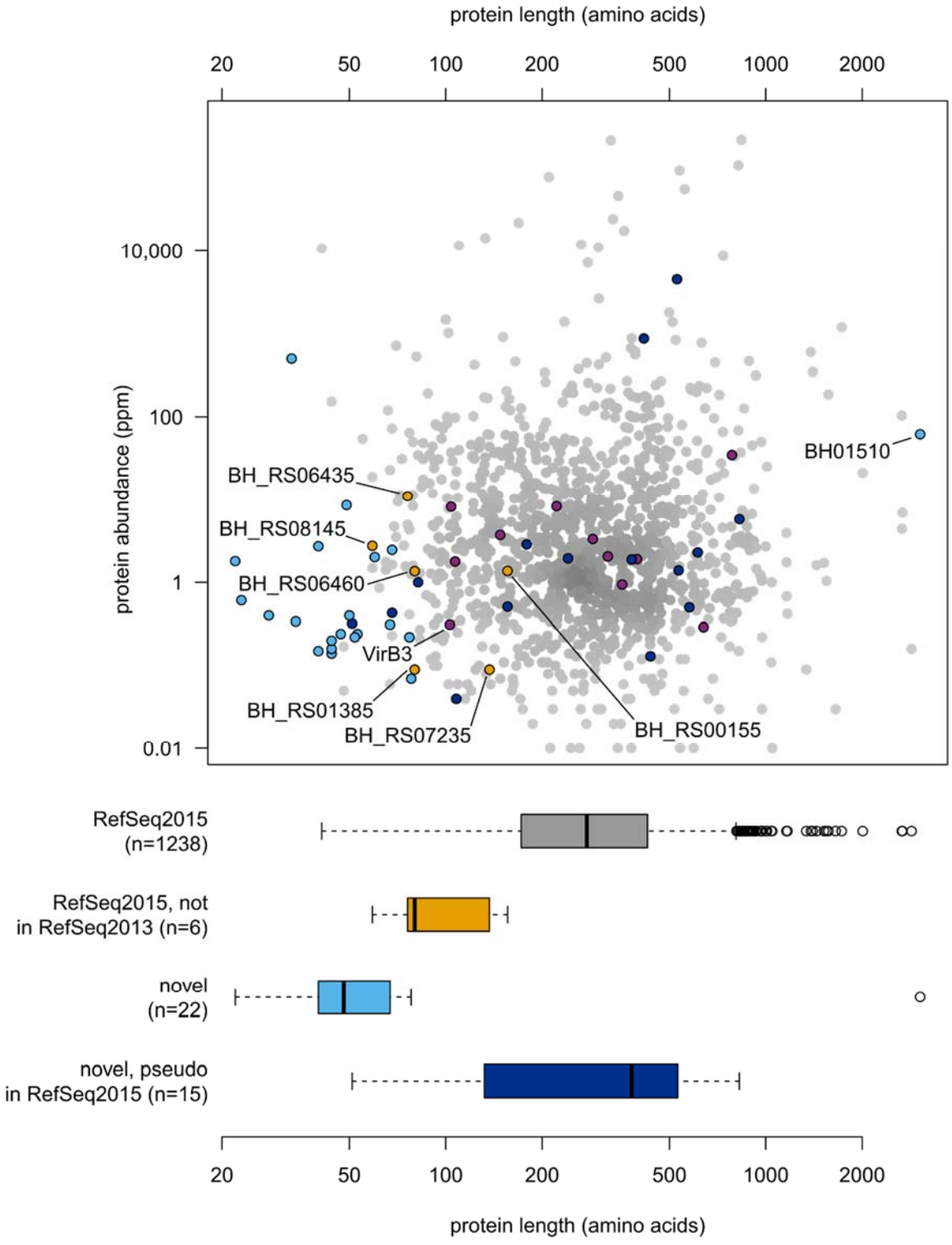
Original PeptideClassifier concept showing the peptide evidence classes and their information content at the protein sequence, protein isoform and gene model level for eukaryotes (six classes, upper panel) and prokaryotes (three classes, lower panel) (Qeli and Ahrens 2010).
The different start sites predicted by various annotation resources of prokaryotic genomes can be covered with an extension of this model. This extension for prokaryotes now also considers six peptide evidence classes as for eukaryotes. Such protein annotation clusters of protein sequences with a shared stop codon (see also Supplemental Fig S1B, Figure 2A) are then treated as variants of a gene model. As a consequence, most of their peptides get re-classified as either class 2a or 2b, i.e. they either imply either a subset or all protein isoforms encoded by one gene model, respectively (Figure 2A). For the cases where all annotations agree (Supplemental Fig S1A), many class 1b peptides are added (Figure 2C). Through the careful, hierarchical integration, we mainly include peptides that unambiguously identify distinct protein sequence(s) of an annotation cluster in the iPtgxDB. Therefore, the percentage of class 1a peptides reaches almost 95%. Through collapsing identical annotations, we remove class 1b and 2b peptides. We do keep shared peptides of classes 2a (i.e. peptides shared within one annotation cluster), as well as 3a and 3b (i.e. peptides shared between different annotation clusters; Figure 2C).

Supplemental Fig S3. PSM score distribution for novelties



The PSM score distribution indicates that the distribution of PSMs implying novelties (green line) is resembling that of hits against the target protein database (blue line), an important indicator that the proteogenomic novelties we identify are not based on false-positive identifications. The figure also shows that our overall FDR threshold is rather stringent (Venter et al. 2011) and that many PSMs with very low E-values are identified.

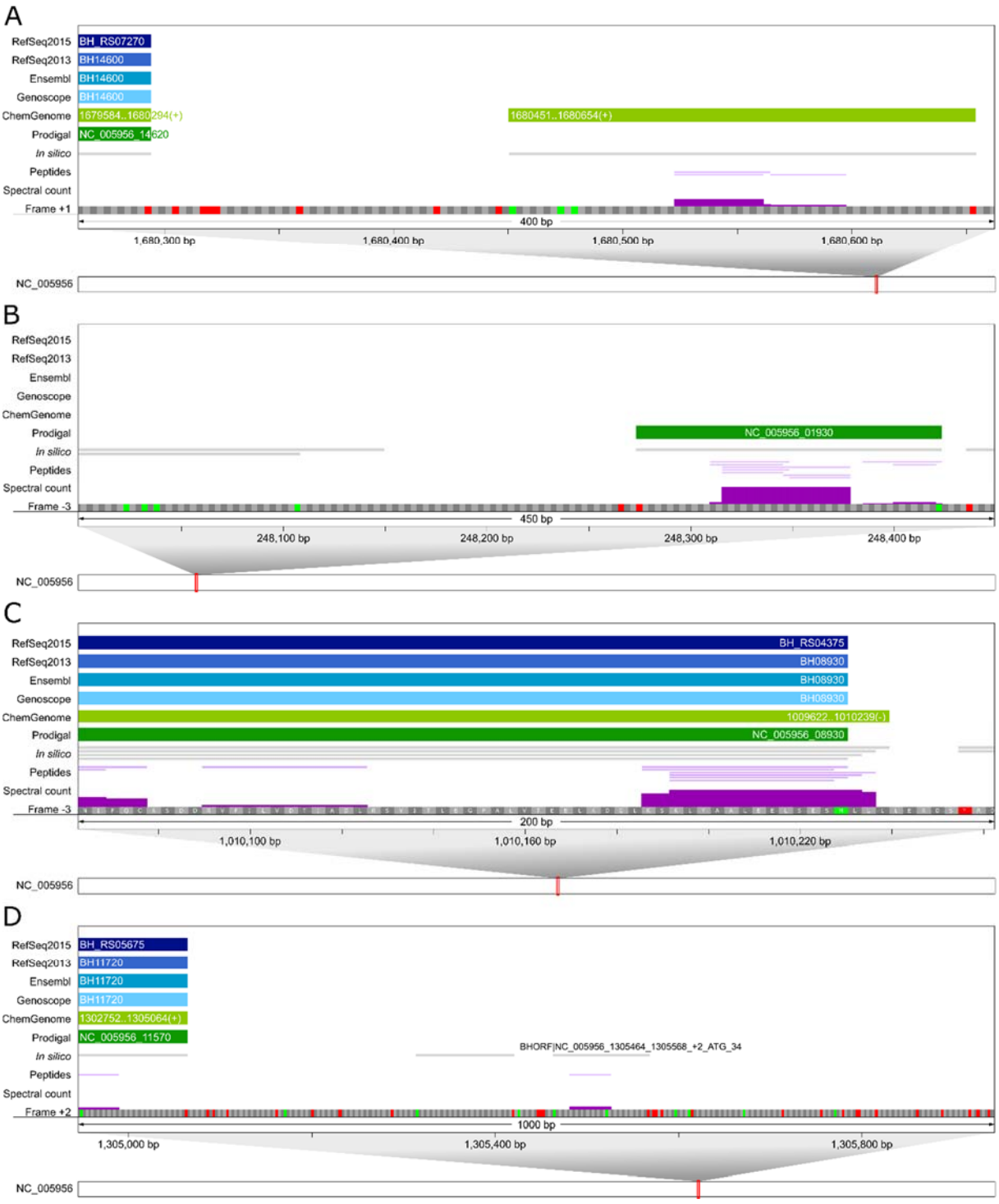
Supplemental Fig S4. Novel ORFs are predominantly short and less abundant



Two-dimensional plot of protein length versus protein abundance, estimated based on spectral counts as described (Schrimpf et al., 2009). Several of the sORFs are relatively well expressed. Gray dots represent RefSeq2015 proteins that we identified. The 6 orange dots represent proteins we had previously identified as novel with a prototype of our

proteogenomics solution (all were predicted by either Genoscope, Prodigal or ChemGenome), and which got incorporated in the RefSeq2015 annotation. Novel ORFs identified with respect to RefSeq2015 are shown in light blue. We labeled VirB3, a component of the VirB/D4 type IV secretion system (T4SS, violet dots) that gets up-regulated in our model system; it is an example of a protein that is highly relevant for the biological response in our model system but that is difficult to identify by mass spectrometry (short, transmembrane domain). The longest "novel" protein is BH01510, the BadA1 adhesin annotated in RefSeq2013 but not in RefSeq2015. For this region, the actual assembly of our lab strain provided additional insights (see main text and Supplemental Fig S6). Protein expression indicates that several of the RefSeq2015 pseudogenes (dark blue dots) were well expressed.

Supplemental Fig S5. Further examples of novelty uncovered by our integrative proteogenomics approach



(A) A differentially expressed, novel sORF (67 aa) uniquely predicted by ChemGenome. Peptide evidence (3 peptides, 11 PSMs) was only observed in the uninduced condition; a predicted SPaseII cleavage signal indicated it could be a lipoprotein. The predicted SCL

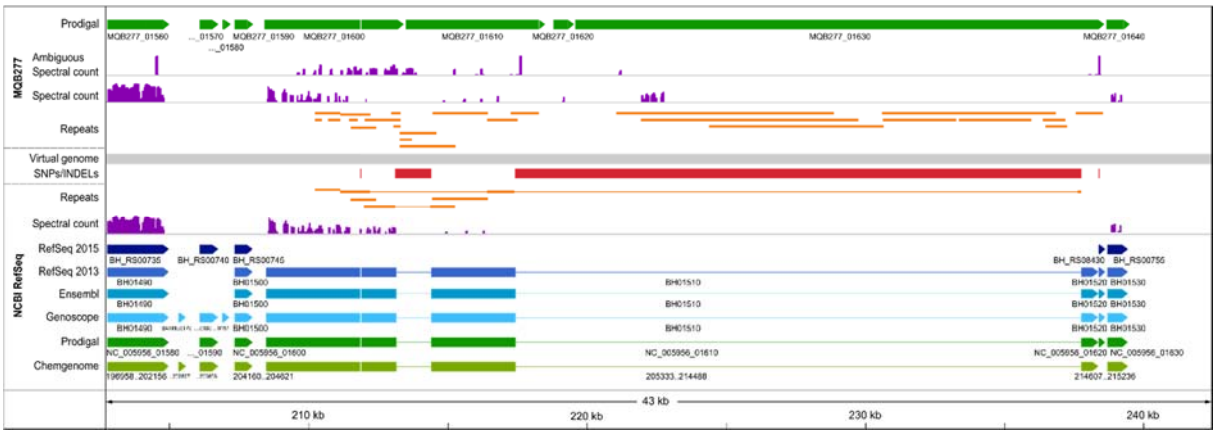
classified it as exclusively total membrane, which was confirmed in the PRM assays, where it was also detected only in the TM fractions (Supplemental Table S3).

(B) NC_005956_01930 is a novel sORF (49 aa) uniquely predicted by Prodigal. We observed protein expression evidence from 9 peptides and a total of 143 PSMs. It is predicted to be localized predominantly in the cytoplasm (Supplemental Table S3).

(C) For BH_RS04375 (NADH dehydrogenase subunit C) all annotations except ChemGenome agree on the translation start site; there is proteomics evidence for the N-terminal peptide with and without initiator methionine (Goetze et al. 2009). In addition, the extended N-termini (+1 and +2 aa) of two *in silico* predicted ORFs are supported by the peptides MMSESLEELAAYLK and MLMSESLEELAAYLK, respectively, while there is no evidence for the +3 aa prediction by ChemGenome (although it cannot be ruled out that the initiator Methionine gets cleaved; Goetze et al., 2009).

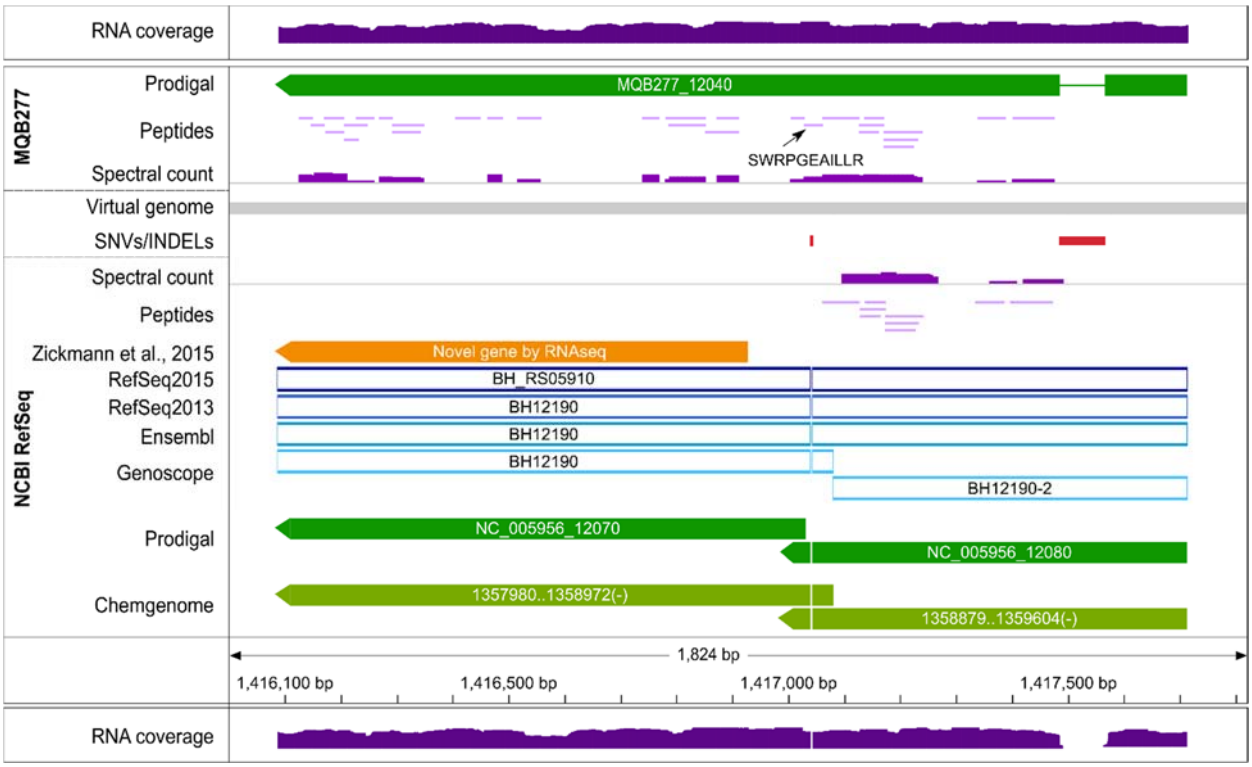
(D) A novel *in silico* ORF (BHORF|NC_005956_1305464_1305568_+2_ATG_34) was identified with one single peptide and 4 PSMs and does not overlap with any other annotation. Its expression could be validated by PRM; our predominant SCL computation (that we extended to all novel ORFs) indicated that it is localized to the IM (Supplemental Table S3).

Supplemental Fig S6. Genomic region of MQB277 that harbors insertions affecting the BadA1 surface antigen, a major pathogenicity factor



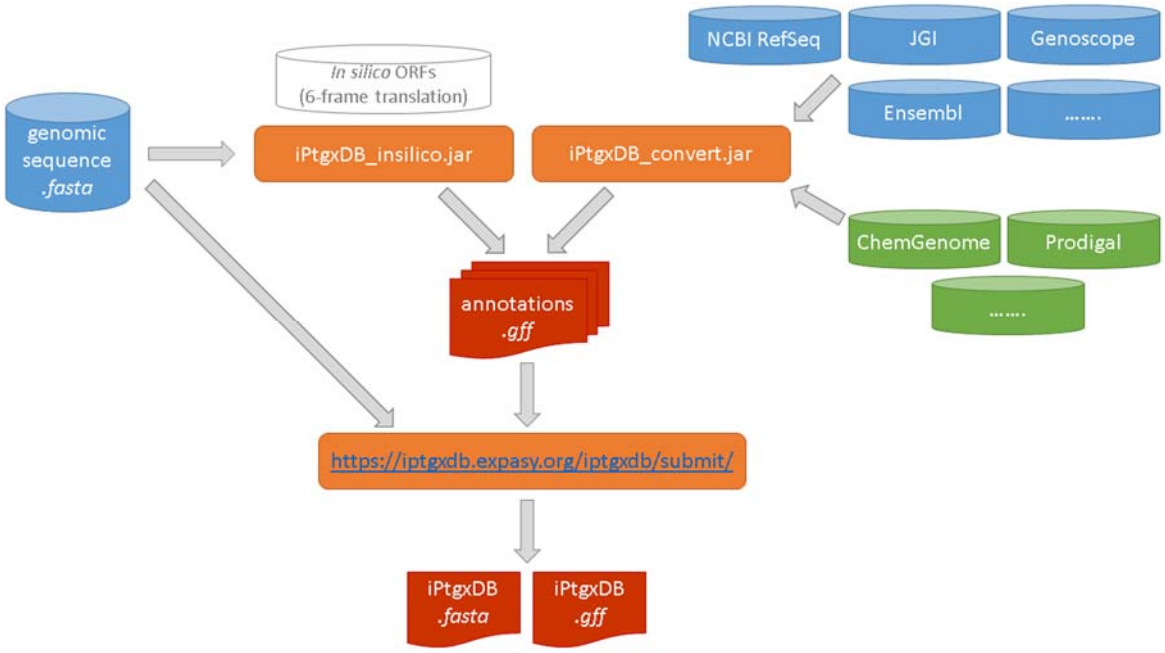
The 22.1 and 1.4 kbp insertions into MQB277 (red bars) compared to the NCBI reference genome (NC_005956) affect a region encoding the Bhen BadA1 adhesion. While this CDS of 3036 aa is annotated in RefSeq 2013 (BH01510), it is missing in the RefSeq2015 annotation. This protein is a major pathogenicity factor that mediates binding of *B. henselae* to extracellular matrix proteins and endothelial cells, and was shown to be immuno-dominant in *B. henselae*-infected patients (Riess et al. 2004). In our high quality PacBio assembly, two CDS are predicted (MQB277_01600, 1776 aa; MQB277_01610, 1731 aa) that have similar domains to MQB277_01590 (RefSeq2013: BH01490), a BadA1 homolog that displays 47% aa identity with BadA1 (Riess et al. 2004). In addition, a protein of 6897 aa was predicted (MQB277_01630). The highlighted repeat regions (orange bars) indicate that both the N-terminal part of BH01510 is rich in repeats, as well as the 22.1 kbp insert, where several repeat units (around 1 kbp) are forming tandem arrays of larger repeat regions. Therefore, we here further separated the spectral count data into those from unambiguous (class 1a) and unambiguous (class 3b; upper spectral count panel for the assembly) peptides (Qeli and Ahrens 2010). There is solid, unambiguous expression evidence for MQB277_01600, few PSMs for MQB277_01610, and the N-terminal portion of MQB277_01630, although due to internal repetitive peptides, these cannot be unambiguously localized. Importantly, the predominant SCL (Stekhoven et al. 2014) for MQB277_01630 is cytoplasmic (Supplemental Table S3), supporting earlier experimental data that had demonstrated lack of or much lower BadA surface expression in the Bhen CHDE101 variant strain (Lu et al. 2013).

Supplemental Fig S7. Proteomics and transcriptomics evidence support a longer ORF only found in the *de novo* assembly.



The same genomic region as in Figure 5A of the main manuscript is shown. Here, we add the sequence of the class 1a peptide (SWRPGEAILLR; see arrow) that spans the region where a 1 bp deletion was observed in the NCBI RefSeq sequence; it was identified by 7 PSMs. In addition, transcriptomics mapped against the assembly (top track “RNA coverage” in the MQB277 assembly) also supports this change (1 additional bp): all reads map perfectly to the assembly compared to a mapping against the NCBI RefSeq genome (bottom track “RNA coverage”).

Supplemental Fig S8. Overview of the software modules to create iPtgxDBs



We release the code to generate custom iPtgxDBs in the form of three java jar files and a public web server (https://iptgxdb.expasy.org/iptgxdb/submit/), which runs the script iPtgxDB_combiner.jar (with additional functionality to be integrated in future releases). Using a genome sequence file in FASTA format as input, the script iPtgxDB_insilico.jar can be executed to create a 6-frame translation (with additional options like alternative start codons). Alternatively, the 6-frame translated database can also be generated using the last module in the public web server. One or several existing reference genome annotations (blue containers) or results of *ab initio* gene prediction tools (green containers) can be processed with iPtgxDB_convert.jar to create several annotation.gff files (one per input file). All of these gff files are then combined with the script iPtgxDB_combiner.jar or on our public web server (https://iptgxdb.expasy.org/iptgxdb/submit/) to create both the searchable iPtgxDB.fasta file and the iPtgxDB.gff file that contains all integrated annotations.

Supplemental Table S1. Detailed analysis of annotation changes for RefSeq2013 vs. 2015

Mapping	RefSeq 2013	RefSeq 2015	# of map-pings	Proteomics evidence (R2013/R2015)*	Transcrip-tomics evidence**
identical protein-coding ORF	n	n	1256	1138	1225
shortened protein-coding ORF	n	n-	74	60 (4/15)	71
extended protein-coding ORF	n	n+	54	45 (1/17)	52
removed protein-coding ORF	n	--	55	4	37
removed pseudogene	p	--	64	0	53
added protein-coding ORF	--	n	99	5[#]	84
added pseudogene	--	p	15	0	15
modified pseudogene region(s)	p	p	28	4	27
	2*p	p	2	0	1
pseudogene(s) to protein-coding ORF(s)	p	n	15	0	15
	p	2*n	1	0	1
	p	3*n	4	1[#]	1
	2*p	6*n	1	0	1
protein-coding ORF(s) to pseudogene	n	p	28	8	23
	2*n	p	8	1	1
complex mappings	n	2*n & p	1	1	n.d.
	p	n & p	2	0	n.d.
	p	2*n & p	1	0	n.d.
	n & p	p	2	0	n.d.
	2*n & p	p	1	0	n.d.

n ('normal' CDS), p (pseudogene), n.d. (not determined). * Cases where either RefSeq2013 or RefSeq2015 annotations are supported by experimental evidence. ** We only considered unambiguously mapped reads (see Suppl. Methods), thus likely underestimating the number of expressed protein-coding ORFs. [#] **6 ORFs we previously identified with a prototype of our proteogenomics approach, i.e. novel ORFs with respect to RefSeq2013**.

A closer inspection of NCBI RefSeq2013 (1488 CDSs, 124 pseudogenes), the basis for the initial study (Omasits et al. 2013), and the recent re-annotation RefSeq2015 (1525 CDSs, 87

pseudogenes) revealed that 23% of the CDSs changed: 55 CDSs were removed (99 added), 74 CDSs were shortened (54 extended), and 64 pseudogenes were removed (15 added) compared to Refseq2013.

Several of the re-annotations were supported by experimental evidence: among 74 shorter RefSeq2015 ORFs, 60 were expressed under the two conditions, with peptide evidence supporting the shorter protein in 15 cases. In contrast, experimental data implied the longer RefSeq2013 protein for 4 cases. For 54 longer RefSeq2015 proteins, 45 were expressed with peptide support for 17 longer proteins, but also implying the shorter protein in one case. In addition, among 48 ORFs relabeled as pseudogenes, we found expression evidence for 10. Similarly, among 55 CDSs removed in RefSeq2015, our data provide expression evidence for 4 (Supplemental Table S1). Importantly, using a prototype of our integrated proteogenomics approach, we had previously found expression evidence for 6 ORFs that were added in RefSeq2015, i.e. novel ORFs with respect to RefSeq2013 (Supplemental Table S1; Supplemental Fig S4). Of further note, among the 55 removed ORFs, we found 32 of 52 proteins that we had previously singled out as potential over-predictions based on lacking orthologs, functional annotation and expression (Omasits et al. 2013).

The matching of CDSs that were relabeled as pseudogenes and pseudogenes that became *bona fide* CDSs was very complex. This may partly be attributed to Bhen being a facultative intracellular pathogen, which often undergo genome adaptation or erosion (Toft and Andersson 2010).

Supplemental Table S2. Comparison of DB complexity and information content

Database	# protein entries	DB complexity (tryptic peptides of 6-40aa)	relative complexity Bhen iPtgxDB versus other Databases	%age class 1a peptides
Bhen RefSeq2015	1,525	25,571	440.7%	97.7%
Bhen iPtgxDB	51,541	112,682	100.0%	93.7%
6-frame translated Bhen genome DB (Mascot)	191,509 (stop-to-stop sequences)	221,029	51.0%	96.2%
Yeast	6,721	159,600	70.6%	97.5%
Human	42,024	603,141	18.7%	45.7%

The Bhen iPtgxDB has many protein entries. However, its complexity (the number of distinct tryptic peptides identifiable by mass spectrometry) is lower than that of a 6-frame translated genome DB as it would be used by Mascot, and that of regular protein search DBs for yeast (*Saccharomyces cerevisiae,* strain ATCC 204508 S288c, UniProt release 2016_08) or human (NextProt release 2016_08_08).

Supplemental Table S3 (separate Excel file).

See separate Excel file with summary of novel ORFs and start sites, expressed pseudogenes plus selected additional information.

Supplemental Table S4 (separate Excel file).

See separate Excel file with the masses of peptides selected for PRM; heavy denotes the labeled peptides from JPT, light the endogenous peptides.

Supplemental Table S5 (separate Excel file).

See separate Excel file (Master table) describing, based on the integrated proteogenomics DB, the entire protein-coding potential of the RefSeq genome (NC_005956) and that of the assembly. Information provided in Supplemental Table S5 includes matching annotations and their identifiers from the different reference genome annotations, expression information, SCL info, and more. The table for the RefSeq genome contains 2932 CDSs, which correspond to the 3037 annotation clusters from the iPtgxDB (without *in silico* ORFs, see Table 1), minus 186 pseudogenes which either had no expression evidence or could manually be reassigned to another cluster, and adding 81 *in silico* ORFs that have at least one PSM in our proteomics data. The table for the assembled genome contains 1639 CDSs predicted by Prodigal and 58 *in silico* ORFs with at least one PSM in our proteomics data (1697 entries overall).

Supplemental Table S6. Summary of genomic differences between MQB277 and NC_005956.

Type	Total #	Selected examples (size, observations)
Inversion - translocation	1	34.4 kbp region; it harbors 89 of 274 SNVs. ➢ 82 SNVs affect 9 protein coding genes* in MQB277; 7 non-coding SNVs. The 2 hemagglutinins comprise 63 SNVs. ➢ 31 SNVs with proteomic support, 51 SNVs without ➢ 26 SNVs with ambiguous peptide evidence and ➢ 5 SNVs with unambiguous peptide evidence (< 4 PSMs) ➢ No transcriptomic support for these SNVs because of poor mapping of reads (duplicated regions)
Insertion/ deletion (>30 bp)	20	3 insertions (22.1, 6.1, 1.4 kbp) in MQB277, harboring ORFs identified with unambiguous pept. evidence (Figure 4, Supplemental Fig S6). 8 insertions in MQB277, 5 affect CDS, no expr. evidence; 7 deletions in NC_005956, 5 affect CDS, no expr. evidence; (2 engineered changes in MQB277: myc-gfp, batR-batS deletion; see Omasits et al. 2013)
Insertion/ deletion (<30 bp)	22	11 changes affected coding regions in MQB277 ** ➢ 4 with transcriptomics and proteomics support ➢ 3 with only proteomics support ➢ 3 with only transcriptomics support ➢ 1 without support
SNV	185 (89 descr. above; 274 in total)	134 SNVs affect protein-coding regions in MQB277 ➢ 42 with proteomics and transcriptomics support (Figure 5) ➢ 12 with proteomics support ➢ 71 with only transcriptomics support ➢ 9 without any support 51 non-coding SNVs (42 with transcriptomics support)

* 2 hemagglutinins, 1 hemin transporter, 1 exonuclease, 1 recombinase, 1 terminase, 1 pemI, 2 hypothetical proteins

** 2, 5 and 4 of 11 with unambig., ambiguous, and no peptide evidence, 7 with transcript support.

Supplemental Table S7. Overview of Bhen search results with MS-GF+.

Genome	**NC_005956**	**MQB277**
Annotation	iPtgxDB	iPtgxDB
Total number of spectra in dataset	2,207,109	2,207,109
Spectra passing the quality criteria by MS-GF+ (spectra searched by MS-GF+)	2,205,799	2,205,799
Spectra assigned to peptides	2,138,758	2,138,716
Spectra passing the PSM FDR threshold	841,411	851,821
Estimated PSM FDR	0.01%	0.01%
Peptides	47,032	47,473
Estimated peptide FDR	0.12%	0.12%
Identified Bhen proteins; RefSeq 2015	1244 (22 novelties, w/o *in silico*)	1275 (Prodigal, w/o *in silico*)
Estimated protein FDR	0.60%	0.60%

Supplemental Table S8. Novel ORFs identified in *E. coli* BW25113.

Annotation source	Novel protein-coding ORF		Evidence for alternative N-terminus
	not in RefSeq	pseudogene in RefSeq	
JGI/Prod (> 5 PSMs)	1	6	2
ChemGenome	1	0	3
in silico ORFs	5	0	1
total	7	6	6

Summary of novel information uncovered for *E. coli* K-12 strain BW25113. We here considered JGI's Integrated Microbial Genomes resource (Markowitz et al. 2009) as alternative reference genome annotation. We again stringently processed the data (PSM level FDR 0.01%); here we required at least 5 PSMs for *in silico* predicted ORFs and at least 3 PSMs for novel start sites for proteins whose expression was supported by additional peptides. The 19 cases of novelty are further detailed in Supplemental Table S9.

Supplemental Table S9 (separate Excel file).

See separate Excel file with summary of novel ORFs, start sites, expressed pseudogenes and additional information for *E. coli* BW25113.

Supplemental File S10. Separate file with iPtgxDbs for Bhen Houston-1, Bhen CHDE101, *E. coli* BW25113 and *B. diazoefficiens* USDA 110.

Supplemental File S11. Separate file with Java code (version 1.0) to create iPtgxDBs.

References

Alsmark, C.M., Frank, A.C., Karlberg, E.O., Legault, B.A., Ardell, D.H., Canback, B., Eriksson, A.S., Naslund, A.K., Handley, S.A., 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 U S A* 101: 9716-9721.

Anders, S., Pyl, P.T., and Huber, W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31: 166-169.

Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E. et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10: 563-569.

Clark, S.C., Egan, R., Frazier, P.I., and Wang, Z. 2013. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics* 29: 435-443.

Goetze, S., Qeli, E., Mosimann, C., Staes, A., Gerrits, B., Roschitzki, B., Mohanty, S., Niederer, E.M., Laczko, E., Timmerman, E. et al. 2009. Identification and functional characterization of N-terminally acetylated proteins in *Drosophila melanogaster*. *PLoS Biol* 7: e1000236.

Grenier, F., Matteau, D., Baby, V., and Rodrigue, S. 2014. Complete Genome Sequence of *Escherichia coli* BW25113. *Genome Announc* 2.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29: 1072-1075.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., 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: D286-293.

Hunt, M., Silva, N.D., Otto, T.D., Parkhill, J., Keane, J.A., and Harris, S.R. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol* 16: 294.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G. et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30: 1236-1240.

Juncker, A.S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H., and Krogh, A. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* 12: 1652-1662.

Kall, L., Krogh, A., and Sonnhammer, E.L. 2007. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res* 35: W429-432.

Kersey, P.J., Staines, D.M., Lawson, D., Kulesha, E., Derwent, P., Humphrey, J.C., Hughes, D.S., Keenan, S., Kerhornou, A., Koscielny, G. et al. 2012. Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res* 40: D91-97.

Koren, S., Harhay, G.P., Smith, T.P., Bono, J.L., Harhay, D.M., McVey, S.D., Radune, D., Bergman, N.H., and Phillippy, A.M. 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol* 14: R101.

Krug, K., Carpy, A., Behrends, G., Matic, K., Soares, N.C., and 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.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. 2014. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* 20: 1983-1992.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*: 1303.3997v1302.

Lu, Y.Y., Franz, B., Truttmann, M.C., Riess, T., Gay-Fraret, J., Faustmann, M., Kempf, V.A., and 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.

Markowitz, V.M., Chen, I.M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P. et al. 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* 40: D115-122.

Markowitz, V.M., Mavromatis, K., Ivanova, N.N., Chen, I.M., Chu, K., and Kyrpides, N.C. 2009. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 25: 2271-2278.

Omasits, U., Quebatte, M., Stekhoven, D.J., Fortes, C., Roschitzki, B., Robinson, M.D., Dehio, C., and Ahrens, C.H. 2013. Directed shotgun proteomics guided by saturated RNA-Seq identifies a complete expressed prokaryotic proteome. *Genome Research* 23: 1916-1927.

Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 8: 785-786.

Peterson, A.C., Russell, J.D., Bailey, D.J., Westphall, M.S., and Coon, J.J. 2014. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol Cell Proteomics* 11: 1475-1488.

Qeli, E. and Ahrens, C.H. 2010. PeptideClassifier for protein inference and targeted quantitative proteomics. *Nat Biotechnol* 28: 647-650.

Riess, T., Andersson, S.G., Lupas, A., Schaller, M., Schafer, A., Kyme, P., Martin, J., Walzlein, J.H., Ehehalt, U., Lindroos, H. et al. 2004. *Bartonella* adhesin a mediates a proangiogenic host cell response. *J Exp Med* 200: 1267-1278.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. 2013. Characterizing and measuring bias in sequence data. *Genome Biol* 14: R51.

Schmid, M.C., Schulein, R., Dehio, M., Denecker, G., Carena, I., and 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., Ahrne, E., Volkmer, B., Callipo, L., Knoops, K., Bauer, M., Aebersold, R., and Heinemann, M. 2015. The quantitative and condition-dependent *Escherichia coli* proteome. *Nat Biotechnol* 34: 104-110.

Schrimpf, S.P., Weiss M., Reiter L., Ahrens C.H., Jovanovic M., Malmstrom J., Brunner E., Mohanty S., Lercher M.J., Hunziker P.E., et al. 2009. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. PLoS Biol 7: e48.

Singhal, P., Jayaram, B., Dixit, S.B., and Beveridge, D.L. 2008. Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. *Biophys J* 94: 4173-4183.

Stekhoven, D.J., Omasits, U., Quebatte, M., Dehio, C., and Ahrens, C.H. 2014. Proteome-wide identification of predominant subcellular protein localizations in a bacterial model organism. *J Proteomics* 99: 123-137.

Toft, C. and Andersson, S.G. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat Rev Genet* 11: 465-475.

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: D636-647.

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

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J. et al. 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26: 1608-1615.