Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome

Omasits, Ulrich; Quebatte, Maxime; Stekhoven, Daniel J; Fortes, Claudia; Roschitzki, Bernd; Robinson, Mark D; Dehio, Christoph; Ahrens, Christian H

Abstract: Prokaryotes are, due to their moderate complexity, particularly amenable to the comprehensive identification of the protein repertoire expressed under different conditions. We applied a generic strategy to identify a complete expressed prokaryotic proteome, which is based on the analysis of RNA and proteins extracted from matched samples. Saturated transcriptome profiling by RNA-seq provided an endpoint estimate of the protein-coding genes expressed under two conditions which mimic the interaction of Bartonella henselae with its mammalian host. Directed shotgun proteomics experiments were carried out on four subcellular fractions. By specifically targeting proteins which are short, basic, low abundant and membrane localized, we could eliminate their initial under-representation compared to the estimated endpoint. A total of 1,250 proteins were identified with an estimated false discovery rate below 1%. This represents 85% of all distinct annotated proteins and about 90% of the expressed protein-coding genes. Genes that were detected at the transcript but not protein level, were found to be highly enriched in several genomic islands. Furthermore, genes that lacked an ortholog and a functional annotation were not detected at the protein level; these may represent examples of over-prediction in genome annotations. A dramatic membrane proteome re-organization was observed including differential regulation of autotransporters, adhesins and hemin binding proteins. Particularly noteworthy was the complete membrane proteome coverage, which included expression of all members of the VirB/D4 type IV secretion system, a key virulence factor.

DOI: https://doi.org/10.1101/gr.151035.112

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-80861
Akzeptierte Version

Originally published at:
Omasits, Ulrich; Quebatte, Maxime; Stekhoven, Daniel J; Fortes, Claudia; Roschitzki, Bernd; Robinson, Mark D; Dehio, Christoph; Ahrens, Christian H (2013). Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome. Genome Research, 23(11):1916-1927.
DOI: https://doi.org/10.1101/gr.151035.112
Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome

Ulrich Omasits, Maxime Quebatte, Daniel J. Stekhoven, et al.

*Genome Res.* published online July 22, 2013
Access the most recent version at doi:10.1101/gr.151035.112

Supplemental Material

http://genome.cshlp.org/content/suppl/2013/08/20/gr.151035.112.DC1.html

Published online July 22, 2013 in advance of the print journal.

Peer-reviewed and accepted for publication but not copyedited or typeset; preprint is likely to differ from the final, published version.

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported), as described at http://creativecommons.org/licenses/by-nc/3.0/.

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.

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

© 2013, Published by Cold Spring Harbor Laboratory Press
Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome

Ulrich Omasits\textsuperscript{1,2}, Maxime Quebatte\textsuperscript{3}, Daniel J. Stekhoven\textsuperscript{1}, Claudia Fortes\textsuperscript{4}, Bernd Roschitzki\textsuperscript{4}, Mark D. Robinson\textsuperscript{1,5}, Christoph Dehio\textsuperscript{3}, Christian H. Ahrens\textsuperscript{1}

\textsuperscript{1}Quantitative Model Organism Proteomics, Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
\textsuperscript{2}Zurich Life Sciences Graduate School Program in Systems Biology, Zurich, Switzerland
\textsuperscript{3}Biozentrum Basel, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland
\textsuperscript{4}Functional Genomics Center Zurich, ETH & University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
\textsuperscript{5}SIB Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland

Corresponding author:
Christian H. Ahrens
Quantitative Model Organism Proteomics, Institute of Molecular Life Sciences
University of Zurich
Winterthurerstr. 190
CH-8057 Zurich, Switzerland
Phone: +41 44 635 3960 E-Mail: christian.ahrens@imls.uzh.ch
ABSTRACT

Prokaryotes are, due to their moderate complexity, particularly amenable to the comprehensive identification of the protein repertoire expressed under different conditions. We applied a generic strategy to identify a complete expressed prokaryotic proteome, which is based on the analysis of RNA and proteins extracted from matched samples. Saturated transcriptome profiling by RNA-seq provided an endpoint estimate of the protein-coding genes expressed under two conditions which mimic the interaction of Bartonella henselae with its mammalian host. Directed shotgun proteomics experiments were carried out on four subcellular fractions. By specifically targeting proteins which are short, basic, low abundant and membrane localized, we could eliminate their initial under-representation compared to the estimated endpoint. A total of 1,250 proteins were identified with an estimated false discovery rate below 1%. This represents 85% of all distinct annotated proteins and about 90% of the expressed protein-coding genes. Genes that were detected at the transcript but not protein level, were found to be highly enriched in several genomic islands. Furthermore, genes that lacked an ortholog and a functional annotation were not detected at the protein level; these may represent examples of over-prediction in genome annotations. A dramatic membrane proteome re-organization was observed including differential regulation of autotransporters, adhesins and hemin binding proteins. Particularly noteworthy was the complete membrane proteome coverage, which included expression of all members of the VirB/D4 type IV secretion system, a key virulence factor.
A major goal of the post-genome era is to understand how expression of the functional elements encoded by a genome is orchestrated to allow an organism to develop and adapt to life under varying conditions. Transcriptomics and proteomics technologies both provide important and complementary insights: the former allow researchers to generate global quantitative gene expression profiles and to study gene regulatory aspects like the impact of short RNAs. However, due to the varying correlation of transcriptomics and proteomics data reported in the literature (de Godoy et al. 2008; de Sousa Abreu et al. 2009; Maier et al. 2011; Marguerat et al. 2012), the direct measurement of protein expression levels is often desirable. For certain aspects, proteomics data can provide more informative and accurate data, as it reflects the effects of other important regulatory processes like protein translation rates and protein stability (Schwanhausser et al. 2011). Furthermore, proteomics provides unique functional insights including post-translational modifications, subcellular localization information and identification of interaction partners of proteins.

Due to enormous advances in mass spectrometry instrumentation, biochemical fractionation methods and computational approaches, proteomics has matured into a state where the description of complete proteomes expressed in a specific condition is within reach. So far, only one study has claimed the identification of a complete proteome expressed in haploid and diploid baker’s yeast (de Godoy et al. 2008), while extensive proteome coverage has been reported for several prokaryotes (Becher et al. 2009; Jaffe et al. 2004; Malmstrom et al. 2009) and archaea (Giannone et al. 2011). Describing extensive proteome maps under different conditions with a discovery proteomics approach is an important first step in defining the protein expression landscape for an organism and facilitates a subsequent shift away from the discovery mode to a re-measurement or scoring mode (Ahrens et al. 2010; Kuster et al. 2005).
Due to the lower transcriptome and proteome complexity compared to eukaryotes, an exhaustive discovery proteomics approach is particularly amenable for prokaryotes. We describe here a generic strategy to achieve an essentially complete coverage of a prokaryotic proteome expressed under specific conditions. Key elements of the strategy are the parallel extraction of RNA and protein from matched samples, and a saturated transcriptome analysis by RNA-seq (Wang et al. 2009). This in turn allows to generate a condition-specific endpoint estimate of the number of actively transcribed protein-coding genes, which is a more appropriate estimate than considering all annotated protein-coding genes. A combination of experimental and computational strategies is then used to dig very deep into the proteome.

We apply the strategy to two conditions that mimic the changing environment encountered by Bartonella henselae upon transfer by its arthropod vector into its mammalian host. The Gram-negative α-proteobacterium B. henselae is a hemotopic, zoonotic pathogen that frequently causes cat scratch disease in immuno-competent humans, as well as bacteremia, endocarditis, and vasoproliferative lesions in immuno-compromised patients. Members of the genus Bartonella are considered re-emerging pathogens and are primarily being studied as models for host-pathogen interaction (Harms and Dehio 2012). A particular emphasis was put on achieving an extensive coverage of the important membrane proteome (Savas et al. 2011). Membrane proteins carry out essential functions as transporters, enzymes, receptors to sense and transmit signals, and adhesion molecules. In light of the resurgence of infectious diseases, membrane proteins are furthermore prime candidates for the development of urgently needed novel anti-infectives (Norrby et al. 2005).

Relying on a very stringent false discovery rate (FDR) cutoff, we were able to identify 1,250 of the 1,467 annotated distinct B. henselae proteins, i.e. a proteome coverage of 85%.
Several lines of evidence indicated that we have exhaustively measured the expressed proteome and can claim to have identified a complete membrane proteome. This included expression evidence – to our knowledge for the first time – for all protein components of a bacterial type IV secretion system (T4SS) which spans the inner and outer bacterial cell membranes.
RESULTS AND DISCUSSION

Model system to explore complete proteome coverage

We chose *B. henselae* as a model system due to several reasons: i) its relatively small genome (1.93 Mbp) comprises 1,488 predicted protein-coding genes (Alsmark et al. 2004); ii) it is a facultative intracellular pathogen that can be grown in pure culture; iii) protocols for subcellular fractionation have been described (Rhomberg et al. 2004); and iv) *in vitro* conditions that mimic the pH dependent induction of virulence genes required for the successful interaction with host endothelial cells, the likely primary niche for *B. henselae* (Harms and Dehio 2012), have been established (Quebatte et al. 2010). The availability of a model system that eliminates the need for co-culture with human endothelial cells is critical to achieve complete coverage of an expressed proteome.

Our *in vitro* model system relies on the induction of the transcription factor BatR (BH00620) that is essential for the pathogenicity of *B. henselae* (Quebatte et al. 2010) (for details, see Supplementary Methods, Tables S1, S2). In the absence of IPTG (uninduced condition), the BatR regulon is not induced, resembling the situation encountered in the arthropod midgut. In contrast, BatR expression is up-regulated in the induced condition, resulting in a marked induction of the BatR regulon, including the VirB/D4 type IV secretion system (T4SS), which is required for infection of endothelial cells (Schulein and Dehio 2002). This state mimics the environment encountered by bacteria in the mammalian host.

A generic strategy for complete proteome coverage by discovery proteomics

We rely on our previous definition of complete proteome coverage, i.e. having identified protein expression evidence for the annotated protein-coding genes actively transcribed in a given state (Ahrens et al. 2010). A recent proteogenomics study of 46 prokaryotes indicated that on average only 0.4% protein-coding genes were missed in the original genome.
Our strategy to achieve an as complete as possible coverage of the expressed proteome of a prokaryote consists of three stages. In a first stage, RNA and proteins are extracted from identical samples, and whole transcriptome libraries are sequenced to saturation by RNA-seq (Figure 1A). Thereby, the number of protein-coding genes actively transcribed in a given state can be estimated, shown here for the sum of protein-coding genes expressed in the uninduced and induced condition (orange dashed line, Figure 1B). Based on such an optimal endpoint estimate, in a second stage several pilot experiments are performed on cytoplasmic and total membrane fractions of the respective conditions. Following a statistical comparison of the pilot phase proteome (green line, Figure 1B) to the predicted endpoint, areas of under-representation can be targeted by the analysis-driven experimentation (ADE) feedback-loop strategy (Brunner et al. 2007), which can help to overcome the premature saturation of distinct protein identifications and sequence deeper into the expressed proteome (blue lines, Figure 1B). In a third stage, evidence is presented that virtually no biases remain when comparing protein parameters of all identified proteins to those called actively expressed, justifying the claim to have identified a complete proteome expressed in a specific condition. Analysis of such a dataset is expected to provide novel insights regarding the achievable membrane proteome coverage, differential protein expression, and evolutionary conservation and genome structure (Figure 1C).

Transcriptome exploration by RNA-seq

We relied on RNA-seq (Wang et al. 2009) primarily to generate an endpoint estimate for the number of expressed protein-coding genes. Whole transcriptome libraries of two biological replicates per condition were generated using a protocol that enriches for mRNA transcripts (see Methods). We sequenced very deep into the transcriptome and obtained 55-87 million
single end 50-mer reads per sample. Of these, 10.7-26.7 million reads mapped
unambiguously, while the vast majority of remaining reads originated from multiple-copy
rRNA genes (see Methods, Table S3). RPKM values (Mortazavi et al. 2008) showed very
high concordance of the biological replicates (r > 0.97, Figure S1).

To estimate how many protein-coding genes are actively expressed in the two conditions,
we plotted the number of distinct expressed protein-coding ORFs as a function of the sum of
uniquely mapping reads. We required at least 5 distinct reads within a 50 nt window of the 5'
end to deem a protein-coding gene actively expressed (Figure S2), a cutoff similar to that
used by (Wang et al. 2009). Saturation is characterized graphically through flattening of the
curves as the number of reads increases. Due to the asymptotic nature of saturation curves,
reaching complete coverage is theoretically only possible with infinite effort. Therefore, we
define saturation as the number of discoveries from where, based on nonlinear modeling
and extrapolation, a doubling of effort is expected to increase the number of discoveries only
marginally. Figure 2A indicates that doubling the number of reads would increase the
number of detected protein-coding genes by less than 3.5% for sample uninduced2 and by
roughly 1% for induced2. Therefore, our analysis indicated that the transcriptome was
sequenced to saturation (Figure 2A). We acknowledge that different library preparations
might potentially identify additional genes, and that very low abundance transcripts (and
proteins) expressed in only few cells of the population may not be identified with this
approach.

We also plotted the density of the RPKM values in order to assess the distribution of
transcription levels for all annotated protein-coding genes: the resulting bimodal graph
suggested that under the conditions studied not all protein-coding genes are actively
expressed; RPKM=10 might be considered a conservative lower cutoff (Figure 2B). The
average RPKM values for members of the virB/D4 operon in condition uninduced2 (30),
where the operon is expected to be expressed at low levels, versus induced (160) support
this observation.

Based on the combined thresholds 1,353 protein-coding genes were expressed in the two
conditions (uninduced 1,254 and induced 1,349). An inter-replicate analysis revealed more
than 95% overlap of the expressed protein-coding genes (Table S4). We include an error
envelope of ±2.5% to account for uncertainty in the thresholds (Figure 3A).

Extended proteome coverage strategy: experimental and computational approaches

Our experimental strategy to reach very deep into the proteome relied on four elements:
First, we used a combination of sub-cellular fractionation and additional biochemical
fractionation regimens to reduce the overall sample complexity, a measure that had been
key to describe the complete expressed proteome of baker’s yeast (de Godoy et al. 2008).
Second, an exclusion list approach (Kristensen et al. 2004) was applied, which helped to
identify a significant amount of low abundant proteins (Figure S5). Third, we relied on the
analysis-driven experimentation (ADE) feedback-loop strategy (Brunner et al. 2007) (Figure
1B) to target under-represented areas of the proteome and overcome premature saturation.
Finally, for all membrane-derived fractions we used chymotrypsin in addition to trypsin,
thereby maximizing the per-protein sequence coverage and the overall membrane proteome
coverage (Fischer et al. 2006).

In terms of computational approaches, we combined results from two database search
engines, Mascot-Percolator (Brosch et al. 2009) and MS-GF+, an updated version of MS-
GFDB (Kim et al. 2010) (see Methods), which employs the generating function approach
(Kim et al. 2008) to compute statistical significance of peptide identifications (spectral
probabilities). Based on these spectral probabilities or the target-decoy option, one can
estimate and stringently control the FDR rate, a critical step for a complete proteome
discovery project. Otherwise, lower quality PSMs will start to accumulate false-positive
peptide evidence for proteins in a random fashion (Reiter et al. 2009). In addition, the error propagates and increases from spectra to peptides and proteins (Nesvizhskii 2010); a PSM level FDR of 1% can correspond to a protein level FDR of 8-11% (Balgley et al. 2007). We therefore chose a very stringent PSM FDR cutoff of 0.01%, allowing us to report protein identifications with an FDR below 1% (see below).

**Identification of the complete expressed B. henselae proteome**

The induction of batR and virB/D4 T4SS expression was more pronounced for the sample pair uninduced2/induced2 than for its biological replicate based on the RNA-seq data. Subcellular fractions from this sample pair (i.e. cytoplasmic (Cyt), total membrane (TM), inner (IM) and outer membrane (OM) fractions) were thus analyzed in detail using different biochemical fractionations (see Methods, Figure 1A).

We first measured the Cyt and TM fractions of both conditions using OFFGEL electrophoresis at the protein level (OGEprot). When requiring at least two independent PSMs to identify a protein, 924 distinct proteins were identified in four experiments, i.e. 63% of all 1,467 distinct annotated proteins or 68%±2% compared to the RNA-seq endpoint estimate of 1,353±34 expressed proteins (Figure 3A). Analysis of the IM fractions from uninduced and induced condition (IMu/i) and the OMu/i fractions contributed 130,000 additional PSMs (72% more PSMs), but only added 22 previously not identified proteins (Figure 3A), indicating that we were already in the saturation phase. We fitted a saturation curve to the eight OGEprot experiments, which shows the anticipated trend of further protein identifications assuming no change in the experimental approach, and also calculated confidence intervals (see Methods, Figure 3A). Carrying out further OGEprot experiments is predicted to lead only to a handful of new protein identifications.

Instead, we relied on the ADE strategy to break the saturation trend. We computed several physicochemical parameters for all distinct B. henselae proteins (see Supplementary
The statistical comparison of the parameters of 946 proteins identified by OGEprot in the pilot phase versus the RNA-seq endpoint estimate of 1,353 expressed proteins in both conditions provided evidence for a significant under-representation of short, low abundant, basic and hydrophobic proteins. These areas of the proteome were subsequently targeted by specific experimental approaches (see Supplementary Methods). Under-representation with respect to length was targeted using size exclusion chromatography (gelfiltration) (Brunner et al. 2007). These experiments added 83 new protein identifications compared to the OGEprot pilot phase (Figure 3A, blue color). The enrichment for shorter proteins can be appreciated in the upper left panel of Figure 3B. Low abundant proteins were targeted using ProteoMiner (Fonslow et al. 2011; Guerrier et al. 2008). These experiments (Figure 3A, grey) helped to identify 42 additional proteins, which were preferentially lower-abundant proteins as evidenced from the density distribution of their Codon Adaptation Index (CAI) values (Sharp and Li 1987) (upper right panel in Figure 3B). Basic and membrane localized proteins were targeted using OFFGEL electrophoresis at the peptide level (OGEpep). The 285 proteins newly added by the OGEpep experiments (Figure 3A, red) were highly enriched for basic proteins (Figure 3B lower left panel) and membrane proteins (with a high grand-average hydropathicity (gravy) value (Figure 3B lower right panel).

Overall, we identified 1,250 distinct proteins requiring at least two PSMs per protein (Figure S3), and only considering peptides that unambiguously identify one bacterial protein (Qeli and Ahrens 2010) (Table 1), i.e. 85% of the 1,467 distinct protein sequences. The FDRs at the PSM, peptide and protein level are below 0.01%, 0.1% and 1%, respectively (Table 1). Only few among the 1,228 proteins identified in the uninduced and 1,231 in the induced condition were selectively expressed (Figure S4); these included several members of the VirB/D4 T4SS in the induced condition. Compared to the expressed transcriptome, the proteome coverage reaches 90% for both the uninduced and uninduced condition.
Although each experimental and computational approach contributed unique protein identifications to the final dataset (see Figure S5), for similar studies aiming to maximize coverage of an expressed proteome with a minimum number of experiments we recommend to use subcellular fractionation (Cyt and TM), perform OGEpep and measure each fraction twice using the exclusion list approach. This approach would identify 1,153 proteins, i.e. 92%, while requiring only 15% of the mass spectrometry runs needed to identify all 1,250 proteins.

**Evidence for having reached an expressed proteome endpoint**

Several lines of evidence indicated that the 1,250 distinct protein groups are very close to the complete proteome endpoint that is actively expressed under the investigated conditions. First, a comparison of the total number of PSM identifications showed that MS-GF+ added 67% more PSMs than Mascot-Percolator (Figure S3A). Yet, at the level of distinct peptides, this increase was smaller (+37%, Figure S3B), and amounted to a mere 3% or 33 additional proteins at the protein level (Figure S3C) despite having added several hundred thousand additional PSMs. Using a third search engine, Sequest, would have only added one additional protein for all experimental spectra. This indicates that, similar to the transcriptome, we have also measured the expressed proteome to saturation. The exponential model fitted to the eight OGEpep experiments (Figure 3A), supports this: doubling the number of PSMs on OGEpep samples (roughly 305,000 additional PSMs, i.e. approx. 36% more PSMs overall) would only identify 5 new proteins (red number on top of red dashed line, Figure 3A).

Second, our expressed proteome encompassed all proteins identified in three previous *B. henselae* proteomics studies (Eberhardt et al. 2009; Li et al. 2011; Rhomberg et al. 2004), while adding many more low abundant proteins (Figure S6A-C).
Third and most importantly, a comparison of the protein parameter distributions of the datasets expressed protein-coding genes (1,353) and final expressed proteome (1,250) showed that there is virtually no under-representation any more in those areas of the proteome that we had specifically targeted, i.e. ADE successfully eliminated these differences present in the OGEprot pilot study (Figure S7). Two examples to illustrate this point: 1) for the parameter isoelectric point (pI) basic proteins are under-represented in the OGEprot dataset. After carrying out the ADE approach, there is only a small difference between the densities of the datasets final and expressed (Figure S7, top panels). 2) For the parameter gravy, membrane proteins with one or more predicted transmembrane domains (gravy values above 0.5) are under-represented in the OGEprot dataset. Again, after the ADE approach, the densities for the datasets “expressed” and “final” are virtually identical (Figure S7, middle panels). This comparison also showed that ADE could add proteins encoded by genes that are expressed at lower levels under the conditions studied (Figure S7, last panels). Two-dimensional density plots of the gene expression level versus the parameters length, pI and gravy (Figure S8) for the dataset final expressed proteome (1,250) versus not seen proteins (217) showed that there is still a noticeable tendency for short and basic proteins to be enriched among genes with expression levels close to the threshold whose proteins were not identified (Figure S8 A,B). These are not expected to be detectable with the shotgun proteomics approach since short and basic proteins have fewer tryptic peptides in the detectable range of the mass spectrometer. In contrast, for the two-dimensional density plot with the protein parameter gravy (values above 0.5 are found in proteins with transmembrane domains), we observed no bias (Figure S8C), indicative of a complete membrane proteome coverage.

To correlate the gene expression level with the proteome coverage, we binned the protein-coding genes according to gene expression strength (RPKM values) and plotted for each bin the respective percentage of proteins identified (Figure 4). A clear correlation between
higher levels of gene expression with a higher success rate of protein identification can be observed. However, several proteins of highly expressed genes were not identified: among 26 such cases from the five top expression bins, 23 had no conserved ortholog in bartonellae, and 16 were located in a novel, plastic genome region (see next section).

**Integration of genome structure information and evolutionary conservation**

We projected transcriptomic and proteomic evidence, ortholog predictions and repeat regions onto the *B. henselae* genome sequence (Figure 5), which contains a large prophage region and three major genomic islands (Alsmark et al. 2004). Genes in such genomic regions are often subject to regulation and become actively expressed only under specific conditions (Juhas et al. 2009). Intriguingly, for 109 of the 198 genes that are located in these four genomic regions we could not detect any expressed proteins (Figure 5, fourth ring). This is a significant enrichment, given that only 227 annotated protein-coding genes did not express any protein (p-value <10^-9, see Figure 5).

We next investigated whether the products of evolutionary conserved protein-coding genes were enriched or selected against. In a comparison with *B. tribocorum*, *B. quintana* and *B. grahamii*, 1,093 of the 1,488 *B. henselae* protein-coding genes were predicted to have an ortholog (Engel et al. 2011), while 395 were not (Figure 5, third ring, turquoise bars). We detected significant over-representation of genes lacking an ortholog (187 of 395) among the 227 protein-coding genes whose proteins were not identified (p-value <10^-9, Figure 5).

To extend the evolutionary conservation analysis beyond members of the genus *Bartonella*, we relied on the eggNOG resource, which contains orthology information from 1,133 organisms including *B. henselae* (Powell et al. 2012). Among the 1,488 *B. henselae* proteins only 55 proteins lack any functional annotation; they are a subset of the 395 without ortholog (black bars, third ring, Figure 5). Strikingly, 52 of these 55 were not detected, again a significant enrichment (p-value <10^-9). A significant number of the genes (16) encoding these
55 proteins clustered in a region from 1,612-1,674kbp that harbors 59 predicted ORFs (p-value <10^-9) (yellow box, Figure 5). Location in this plastic, repeat-rich genome region (orange bars, fourth ring) may lead to strong transcription of genes that do not represent a bona fide protein-coding ORF.

The evolutionary conservation information provided by eggNOG together with high quality experimental proteomics data, represents a particular useful combination to identify candidates for over-predicted protein-coding genes in genome annotations: the densities of the protein length distribution of the proteins not identified (217) was clearly separated from that of the proteins seen (1,250) (Figure S9A). Among the proteins not seen, those that lack any functional annotation are considerably shorter than those with a functional annotation (Figure S9B). Since we can detect short proteins with our set-up (see density of the 150 shortest proteins detected compared to all, Figure S9C), the proteins that lack an ortholog and any functional annotation may either only be expressed under different conditions, or are potential over-predicted ORFs.

Coverage of the membrane proteome and the VirB/D4 T4SS

The membrane proteome serves many essential roles in cellular communication, transport, adhesion to host cells and evasion of the host immune system. While accounting for up to one third of the gene products, more than 50% of the druggable targets fall into this category (Hopkins and Groom 2002). However, due to the amphipathic nature and low abundance of membrane proteins, they are notoriously under-represented in proteomics studies (Helbig et al. 2010; Poetsch and Wolters 2008; Tan et al. 2008).

To reach a high protein sequence coverage for membrane proteins, we used a combination of trypsin and chymotrypsin in all membrane samples and furthermore applied proteolytic digestion in 60% (v/v) Methanol to improve cleavability of hydrophobic proteins (Fischer et al. 2006) (Supplementary Methods). Among 924 proteins identified in the first four pilot
phase experiments (63% of all distinct proteins), 182 contained predicted transmembrane
domains (54%, Figure 6A, upper panel). However, the ADE approach was able to eliminate
this under-representation of membrane proteins: among the final 1,250 identified proteins
(85% of all distinct annotated proteins) 289 of the 338 distinct proteins with one or more
predicted transmembrane regions were found, i.e. 86% (Figure 6A, lower panel; Figure
S10A). Notably, the OGEpep fractionation regimen was particularly successful to identify
membrane proteins. We also identified 54 of the 58 predicted secreted proteins (95%).
These include many proteins for which PSORTb (Yu et al. 2010) predicts localization in the
membrane space, and where other studies could confirm their localization in inner or outer
membrane, periplasm or the extra-cellular space (Figure S10B). Together with the striking
result that transmembrane proteins with high gravy values are not over-represented among
the 217 non-identified proteins compared to 1,250 seen proteins (see Figure S8C), the data
suggested that we have identified a complete membrane proteome expressed under two
specific conditions.

This includes all eleven protein members encoded by the virB/D4 operon in the induced
condition (Figure 6B). To our knowledge, this is the first complete coverage of this important
molecular machinery spanning both inner and outer membrane by a shotgun proteomics
approach. We also detected all seven Bartonella effector proteins (Beps), which are
secreted by the VirB/D4 T4SS into eukaryotic host cells (Figure 6B). In contrast, many
proteins of the Trw complex, a second B. henselae T4SS that is essential for the infection of
erthrocytes (Vayssier-Taussat et al. 2010) but dispensable under the conditions studied,
were not detected (9 of 24, 38%) (Figure 5, first and fifth ring), nor was their expression
regulated (Figure S11).

When we assessed the level of induction at the RNA and protein level, we observed that the
induction of virB/D4 and bep operons, which are direct targets of the transcriptional regulator
BatR, seemed to be more prominent at the protein level. They also included more cases with
statistical significance of the up-regulation (Figure 6B, log\textsubscript{2} fold changes, left panel). A comparison of the log\textsubscript{2} fold changes at the RNA level versus those at the protein level indicated that several of the virB/D4 and bep genes appear to be regulated preferentially at the post-transcriptional level, indicated in Figure 6C by their position close to the vertical axis.

The ability to identify complete membrane proteomes of prokaryotes has important implications for studying their expression under different conditions in a quantitative fashion. Ideally, such a task would be performed with the more sensitive targeted proteomics approach (Schmidt et al. 2011), which typically relies on predicted PTPs using tools like PeptideSieve (Mallick et al. 2007). Our data indicate that a comprehensive discovery proteomics approach adds clear value with respect to experimentally identified PTPs as we could identify peptides for 145 proteins for which PeptideSieve predicted no PTP (see Supplementary Methods). We provide the proteomics and transcriptomics data with results of several prediction algorithms (Table S5A), and all experimentally identified peptides (Table S5B), from which the best-suited PTPs can be selected using available guidelines (Picotti and Aebersold 2012).

Identification of differentially expressed proteins

Our in depth proteome analysis precluded the measurement of biological replicates. We thus relied on DESeq to identify the most significantly differentially regulated proteins between induced and uninduced states (see Methods). The top 10% differentially expressed proteins (Table S6), including 68 up-regulated (red dots), and 57 down-regulated proteins (green dots) in the induced condition, are highlighted in Figure 7.

Among these 125, 36 transmembrane and 12 secreted proteins were found, a significant enrichment (p-value <0.0018) compared to 343 membrane and secreted proteins among the
1,250 proteins. A striking feature was the strong regulation of different families of autotransporters, which rely on the type V secretion pathway for their delivery to the surface of Gram-negative bacteria (Leyton et al. 2012). These included two representatives of the trimeric autotransporter adhesins (BH01490, BH01510), a class of virulence factors essential for Bartonella pathogenicity (Franz and Kempf 2011). Furthermore, 7 of 10 proteins with an autotransporter beta domain (as predicted by SMART version 7, (Letunic et al. 2012)) were among the top 10% differentially regulated proteins (6 up-regulated, 1 down-regulated; yellow dots, Figure 7; Table S6), i.e. a significant enrichment (p-value <4x10^{-7}). BH13020, BH13180 and BH13010 were the top three up-regulated proteins, which ranked even higher than members of the VirB/D4 operon. While less is known about the role of this family of autotransporters in Bartonella, they were found to be up-regulated during infection of endothelial cells (Quebatte et al. 2010) and may be involved in adhesion to host cells (Litwin et al. 2007). Finally, two of the four outer membrane proteins of the hemin binding protein family (HbpC and HbpB) were found. HbpC was shown to protect B. henselae against hemin toxicity and to play a role during host infection (Roden et al. 2012).

The top 10% regulated proteins included 6 of the 7 Beps and all VirB/D4 T4SS proteins except VirB3. For this small protein (103 amino acids) with one predicted transmembrane domain we only found 4 spectra, all in the induced condition. This indicates that a large experimental effort is required to detect proteins that combine several parameters which complicate their mass spectrometric identification with shotgun proteomics, i.e. they are short, basic and hydrophobic. Another protein exclusively identified in the induced condition is BH13250, a hypothetical protein with a transmembrane domain (Table S6). Its location right upstream of the virB/D4 operon is conserved in other Bartonella, suggesting that it may potentially carry out a yet to be determined function as virulence factor. Finally, another interesting upregulated protein is RpoH1 (BH15210), an alternative RNA-polymerase sigma factor 32. It has recently been identified as an essential component for the expression of the
VirB/D4 T4SS by independent approaches (M. Quebatte et al., personal communication). A role in virulence has been documented for its gene in an in vivo mouse infection model for the closely related Brucella (Delory et al. 2006).
CONCLUSION

Using a discovery proteomics approach, the expressed proteome of *B. henselae* was exhaustively studied under two conditions that mimic those encountered in different hosts. The saturated transcriptome analysis of RNA extracted from matched samples provided the best-possible endpoint estimate for the number of actively transcribed protein-coding genes. ADE was able to virtually eliminate the biases of commonly under-represented short, basic, and particularly lower abundant and membrane protein classes, all of which are experimentally tractable. Based on a very stringent FDR at the PSM level, we identified 85% of all distinct, annotated proteins and about 90% compared to the expressed protein-coding genes in the two conditions. Several lines of evidence indicated that this is very close to all proteins that can be identified by a discovery proteomics approach with current technology. This is best illustrated by the complete membrane proteome coverage, including evidence for all members of the important VirB/D4 T4SS. The analysis of the genome organization revealed that genes whose transcripts were detected, but not their corresponding protein products, were highly enriched in genomic islands. Information regarding evolutionary conservation provided evidence for preferential expression of genes with a predicted ortholog. In contrast, genes that lacked an ortholog and functional annotation were mostly not observed at the protein level, suggesting possible over-prediction in genome annotations.

Our report is the second complete expressed proteome reported (de Godoy et al. 2008). Using a similarly extensive fractionation strategy, our matched transcriptomics and proteomics data correlated quite well (r=0.57) while identifying the VirB/D4 T4SS as a prominent target of post-transcriptional regulation. The rigorous approach to sequence transcriptome and proteome to saturation and to provide proof for having eliminated observed biases at the protein level is unique. It supports a recent perspective article showing that up to 90% of an expressed proteome ("nearly complete") can be measured
quite fast (Mann et al. 2013), but the remaining 10% require extensive effort. It also
underlines that the difference between “comprehensive” and complete can be quite large, in
particular with respect to coverage of the membrane proteome (Beck et al. 2011). The
higher coverage of distinct annotated proteins (85%) compared to the proteome expressed
by haploid and diploid yeast (67%) suggests that prokaryotes express a higher fraction of
the encoded proteins potentially reflecting their need to quickly adapt to changing conditions.
This fraction may be lower for more complex prokaryotes.

The data attest to the value of a discovery proteomics approach in providing experimentally
identified PTPs beyond those predicted in silico. The sensitive quantitative measurement of
such PTPs by SRM holds particular promise to be able to screen entire bacterial
surfaceomes and to identify targets for novel anti-infectives. Ideally, such studies would be
carried out using in vivo infection models. Enabled by the consideration of organism-specific
peptide information (Delmotte et al. 2010), they will bring the analysis of mixed in vivo
proteomes within reach and complement the power of dual RNA-seq (Westermann et al.
2012) for this task. We expect that the strategy described here will be useful for some of
these exciting applications.
METHODS

Bacterial growth and subcellular fractionation

*B. henselae* strain MQB307 harbors a deletion of the response regulator *batR* (BH00620) and its cognate sensor histidine kinase *batS* (BH00610), and carries a plasmid-encoded copy of *batR* under the control of an IPTG inducible promoter (for details, see Supplementary Methods, Tables S1, S2). MQB307 was grown on Columbia blood agar (CBA) plates supplemented with 30 mg/l kanamycin with (induced condition) or without (uninduced condition) 500 μM IPTG at 35°C and 5% CO₂ for 60 h. The subcellular fractionation was performed as previously described (Rhomberg et al. 2004) (Supplementary Methods). To maximize the recovery of membrane proteins, the total membrane fraction (TM) was further separated into inner membrane (IM) and outer membrane (OM) fraction.

RNA extraction and whole transcriptome sequencing

RNA was isolated from bacterial cells as described (Quebatte et al. 2010). Whole Transcriptome libraries were produced using the Ribominus kit, Bacteria Module (Invitrogen), and the SOLiD™ Total RNA-seq kit (Applied Biosystems). Briefly, cDNA libraries were size selected and amplified for 18 cycles of PCR. The whole transcriptome library was used for emulsion-PCR based on a concentration of 0.5 pM. Sequencing beads were pooled and loaded on a full SOLiD™-4 slide; between 55-87 million 50 base sequencing reads were generated per library (Table S3). For details, see Supplementary Methods.

RNA-seq data processing and transcriptome coverage analysis

The sequenced reads were mapped to the genome sequence of *B. henselae* Houston-1 strain using the BioScope 1.3.1 mapping pipeline. Among all uniquely mapping reads, those of lower quality were removed (for more detail, see Supplementary Methods, Figure S12). The count data summary for annotated *B. henselae* ORFs was generated using the HTSeq package. To create Figure 2A, the filtered reads were shuffled and sequentially mapped to the genome; a protein-coding ORF was classified as expressed when accumulating five or more distinct reads in the 5’ end of the ORF. Based on this data, nonlinear regression models were constructed to estimate the effect of doubling the number of reads. For details, see Supplementary Methods.

Protein and peptide fractionation & mass spectrometry
The subcellular fractions (Cyt_u/i, TM_u/i, IM_u/i, OM_u/i) were further fractionated biochemically, including OFFGEL electrophoresis at the protein (OGEprot) and peptide level (OGEpep), and size exclusion chromatography (SEC, “gelfiltration”). To enrich for low-abundant proteins, we used the ProteoMiner approach (Guerrier et al. 2008). More detail on the biochemical fractionations, digest conditions and the mass spectrometry set-up is given in the Supplementary Methods and Figure S13. Samples were injected into an Eksigent-nano-HPLC system (Eksigent Technologies) by an autosampler, separated on a self-made reverse-phase tip column packed with C18 material, and acquired on an LTQ-Orbitrap XL or LTQ-ICR-FT-Ultra mass spectrometer (both Thermo Scientific).

**Database searching and data processing**

To minimize the chance for false positive assignments spectra were searched against a combined database (1,488 *B. henselae* proteins, 3,336 sheep proteins, a positive control (myc-gfp), and sequences of 256 common contaminants (keratins, trypsin, etc.)) either with Mascot (version 2.3.0, Matrix Science) or with MS-GF+ (MS-GFDB v7747). For Mascot, data were further post-processed with Percolator (Brosch et al. 2009). Based on the target-decoy search approach, a Percolator/MS-GF+ score cutoff was determined that resulted in an estimated 0.01% FDR at the PSM level. All PSMs above this cutoff were classified with the PeptideClassifier software (Qeli and Ahrens 2010), and only peptides (tryptic or semi-tryptic) that unambiguously imply one bacterial protein sequence were considered (Table 1). For details, see Supplementary Methods.

**ADE analysis**

Exponential curves were fitted to each block of experiments with a shared biochemical fractionation regimen to find a saturation threshold (Figure 3A). We then used this fit to predict the saturation beyond the point of experimentally observed PSMs for each biochemical fractionation regimen (Figure 3A, dashed lines). For details on the exponential model, approximating confidence bands, density estimation of physicochemical parameters, and computation of physicochemical parameters and other protein sequence features, see Supplementary Methods.

**Statistical analysis**

Statistical tests were performed using the statistical software R 2.15.2 (www.R-project.org). All reported p-values are from hypergeometric tests and are adjusted for multiple testing.
controlling the corresponding FDR (Benjamini and Hochberg 1995). Significance is based on an alpha level of 5%.

Transcript and protein abundance estimation

Transcript abundance was estimated via RPKM values calculated similar to (Mortazavi et al. 2008). The sum of mapped and filtered reads per gene was divided by its length (in kilobases) and the sum of reads for all B. henselae protein-coding genes (in million reads).

Relative protein abundance (in ppm, see Figure S6C) was estimated based on spectral counts as described (Schrimpf et al. 2009).

Orthologs, sequence repeats, and functional protein classification

Orthologous genes conserved in B. henselae, B. tribocorum, and B. grahamii were taken from (Engel et al. 2011). To find duplicated regions of 50nt or longer in the B. henselae genome, we used RepSeek (version 6.5, (Achaz et al. 2007)). For functional protein classification, we relied on the eggNOG resource (http://eggnog.embl.de). For details, see Supplementary Methods.

Differential expression analysis

Differential transcript and protein expression analysis was carried out with the R package DESeq (version 1.6.1, (Anders and Huber 2010)). Our description of condition-specific complete expressed proteomes precluded the analysis of biological replicates. Since DESeq ranks proteins according to statistical significance, i.e. the top-ranked proteins are observed by many spectra, we minimized the potential to erroneously identify differentially expressed proteins by chance. On the other hand, without replicates we lack the power to detect lower expressed, truly differentially regulated proteins.

DATA ACCESS

RNA-seq data is accessible under the GEO Series accession number GSE44564. Proteomics data associated with this manuscript can be downloaded from ProteomeXchange under accession PXD000153.
ACKNOWLEDGEMENTS

We thank Dr. Ermir Qeli for initial work on the project, Dr. Sangtae Kim (PNNL, USA) for the MS-GF+ software, and Drs. Bernd Wollscheid (IMSB, ETH Zürich), Aurelien Carlier and Gabriella Pessi (Institute of Plant Biology, UZH) for critical reading of the manuscript. CHA acknowledges support from the Swiss National Science Foundation (SNF) under grant 31003A_130723. A part of the research was performed using EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.
FIGURE AND TABLE LEGENDS

FIGURES

Figure 1: Overview of the complete expressed proteome discovery workflow.

A. Extraction of RNA and proteins from matched samples, transcriptome analysis. Total RNA and proteins were extracted in parallel from bacteria grown either under uninduced or induced conditions (schematically shown by black knobs representing the VirB/D4 T4SS). Protein extracts were sub-fractionated into cytoplasmic (Cyt), total membrane (TM), inner (IM) and outer membrane (OM) fractions. To estimate an upper bound for the number of actively transcribed protein-coding genes, the transcriptome was sequenced to saturation using RNA-seq.

B. Analysis-driven experimentation (ADE). In a first pilot phase, samples are analyzed by LC-MS/MS. Under-represented proteome areas are identified based on a statistical analysis comparing experimentally identified proteins to all expressed proteins (the estimated RNA-seq endpoint indicated by the orange dashed line within an error envelope). All distinct annotated proteins are indicated by the black, dashed line. Subsequently, these areas are investigated by targeted experiments, aiming to overcome the saturation trend.

C. Integrative data analysis. Data from the expressed proteome are integrated with genomic, transcriptomics, orthology, and other information to enable further analyses.

Figure 2: Transcriptome coverage by RNA-seq.

A. Saturated coverage of protein-coding genes. An estimate of the number of actively expressed protein-coding genes based on the number of uniquely mapped RNA-seq reads is shown for both conditions and biological replicates.

B. Density distribution of RPKM values. In addition, boxplots representing the expression level of the eleven members of the virB/D4 operon are shown for the uninduced (black), and induced (red) condition. For clarity, we only show data for the sample pair uninduced2/induced2.

Figure 3: Overcoming the saturation of protein identifications using ADE-guided shotgun proteomics.

A. Increase of distinct identified proteins given the number of PSMs observed in different experiments. We fitted an exponential curve (see Methods) to all experiments for a given biochemical fractionation in order to find a saturation limit (see colored numbers on the right-
hand side). We also approximated confidence bands for the fitted points (thin lines, see Methods). The black dashed line at the top signifies the total number of distinct *B. henselae* proteins (1,467), the orange dashed line below represents the estimated RNA-seq endpoint of expressed distinct proteins (1,353) including a ±2.5% error envelope (orange shaded area).

B. Density estimates of four physicochemical protein parameters for different protein subsets. The parameter density for proteins newly identified by the ADE approach is contrasted to that of all expressed proteins (orange) and those identified in pilot experiments using OGEprot (green). The most important aspects of over- or under-representation can be seen on the abscissa; they indicate that the targeted experiments successfully add new protein identifications in areas of the proteome that were under-represented in the pilot experiments. For details on the density estimation and the bootstrap confidence bands (shaded areas) see Supplementary Methods.

**Figure 4:** Correlation of gene expression strength and successful protein identification rate. Protein-coding genes are binned according to strength of gene expression (the maximum RPKM value of both states). The success rate in identifying the encoded proteins in each bin is represented by the blue area of the bars; orange dots above the barplot indicate the respective percentage. The numbers above the bars show how many proteins were not identified within a given bin (for a total of 217 distinct proteins).

**Figure 5:** Integration of expression evidence with structural genome information and evolutionary conservation. Genes whose proteins were not identified cluster in specific regions of the *B. henselae* genome. Outer ring: genes whose proteins were identified (light blue), or not identified (red). Second ring: protein-coding genes classified by the RNA-seq analysis as expressed (grey), or not (dark green). Third ring: genes without a detectable ortholog among species of lineage 4 of the genus *Bartonella* (Engel et al. 2011) (turquoise), and genes without any functional annotation by the eggNOG classification (black). Fourth ring: repeat regions identified by RepSeek (orange) (Vallenet et al. 2006), and rRNA repeat regions (light orange). Fifth ring: location of a prophage region (ochre), three genomic islands (blue), the *virB/D4* and *trw* operons (skyblue), and a novel genomic region enriched in repeats as well as highly expressed genes whose encoded proteins were not identified (yellow). The results of hypergeometric tests for selected datasets are also shown (asterisks indicate statistically significant enrichment, see text). For the hypergeometric test, we used all possible protein-
coding genes for the identified “seen” proteins (1,250 distinct proteins encoded by 1,261
gene models) and “not seen” proteins (217 distinct proteins encoded by 227 gene models).
The circular plot was generated using DNAPlotter (Carver et al. 2009).

Figure 6: Membrane proteome coverage and dynamics.
A. Comparison of the membrane proteome coverage achieved in four pilot experiments
(upper panel) and the final dataset (lower panel). Membrane proteins are binned according
to the number of predicted transmembrane domains; the percentage of proteins identified
per bin is shown above each bar. The legends summarize the respective coverage achieved
comparing the respective dataset (pilot phase/final) against all distinct proteins and for the
subset of proteins with transmembrane domains. Membrane proteins are under-represented
in the pilot phase but not in the final dataset.

B. Transcript and protein expression changes of the virB/D4 T4SS and downstream bep
operon. Operon structures (upper panel) are drawn to scale. The lower left panel shows the
log₂ fold changes at the transcript and protein level for the induced versus uninduced state
(the ∞ indicates that the protein was only identified in the induced condition). Fold changes
and significance were calculated with DESeq. Regulation at the protein level appears to be
more pronounced compared to the transcript level. The lower right panel visualizes the
protein expression changes upon induction onto a schematic representation of the
assembled VirB/D4 T4SS using different shades of blue.

C. Comparison of expression changes at transcript and protein level. The respective log₂
fold changes based on the RPKM values and normalized spectral counts are shown.
Members of the VirB/D4 T4SS are shown in blue (BH13360 in light blue), Bartonella effector
proteins (Beps) in dark blue. Three proteins that exhibited the most significant differential
expression (Table S6, Figure 7) are also shown with their identifiers.

Figure 7: Differential protein expression analysis.
The log₂ fold change of the expression of all experimentally identified B. henselae proteins in
the induced versus uninduced condition is shown against the mean normalized spectral
count (MA plot). The 10% most significant differentially expressed proteins are highlighted,
including 68 up-regulated proteins (red dots) and 57 down-regulated proteins (green dots).
Selected regulated proteins are highlighted in different colors: members of the VirB/D4 T4SS
(blue dots), Beps (dark blue dots), and several proteins containing autotransporter beta-
domains (yellow dots). Proteins in these categories that rank below the 10% cutoff are
shown as open circles.
Table 1: Summary of identified PSMs, peptides and proteins and estimated FDR levels

<table>
<thead>
<tr>
<th></th>
<th>No. of PSMs</th>
<th>No. of distinct peptides</th>
<th>No. of distinct proteins*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1a</td>
<td>747,352</td>
<td>43,193</td>
<td>1,240</td>
</tr>
<tr>
<td>Class 3a</td>
<td>7,356</td>
<td>283</td>
<td>10</td>
</tr>
<tr>
<td>Class 3b</td>
<td>12,161</td>
<td>663</td>
<td>n.a.</td>
</tr>
<tr>
<td>Total B. henselae</td>
<td>766,869</td>
<td>44,139</td>
<td>1,250</td>
</tr>
<tr>
<td>Decoy hits</td>
<td>54</td>
<td>42</td>
<td>7</td>
</tr>
<tr>
<td>estimated FDR</td>
<td>&lt; 0.01%</td>
<td>&lt; 0.1%</td>
<td>&lt; 1.0%</td>
</tr>
</tbody>
</table>

The total number of PSMs, distinct peptides and distinct proteins is shown, further separated by peptide evidence class (Grohe et al. 2009). We only considered proteins implied by class 1a and 3a peptides, not those implied by ambiguous class 3b peptides (n.a.). *Protein groups identified by 3a peptides are unique protein sequences that can be encoded by two or more distinct gene models. The 1,250 experimentally identified proteins are encoded by 1,261 gene models; the 217 non-identified proteins are encoded by 227 gene models (in total: 1,467 distinct proteins are encoded by 1,488 protein-coding genes).
REFERENCES


under-represented proteome

induced RNA-seq saturates total RNA

uninduced RNA-seq

proteins

subcellular fractionation:
  - Cyt
  - TM
  - IM
  - OM

proteins

under-represented proteome

statistical analysis

identify proteins

targeted experiments

complete expressed proteome

orthology analysis

subset correlations

differential expression

membrane proteome

targeted experiments

A B C
A

B

Expressed OGEprot Gelfiltration ProteoMiner OGEpep

Cold Spring Harbor Laboratory Press on September 15, 2013 - Published by genome.cshlp.orgDownloaded from
1,488 protein-coding genes

198 genes in gen. islands or prophage region

59 genes in novel genomic region

395 genes without ortholog

55 genes without functional annotation
Figure 7

The figure shows a scatter plot of gene expression levels. The x-axis represents the mean normalized spectral count, while the y-axis represents the log2 ratio of induced to uninduced expression levels. The plot includes data points for induced genes only, uninduced genes only, 3-fold upregulated genes, 3-fold downregulated genes, and genes induced by VirB/D4, Beps, or Autotransporters.

Key markers and genes include:
- **VirB3**
- **BH13250**
- **RpoH1**
- **HbpC**
- **HbpB**
- **BH13010**
- **BH13020**
- **BH01510**
- **BH01490**

Legend:
- Blue: VirB/D4
- Dark blue: Beps
- Yellow: Autotransporters

The plot highlights the different expression patterns and fold changes for various genes and transcription factors.