Direct mapping of Peptide-to-Spectra-Matches to genome information facilitates qualifying proteomics information

John Anders (john@bioinf.uni-leipzig.de)
Leipzig University

Hannes Petruschke
Helmholtz Centre for Environmental Research

Nico Jehmlich
Helmholtz Centre for Environmental Research

Sven-Bastiaan Haange
Helmholtz Centre for Environmental Research

Martin von Bergen
Helmholtz Centre for Environmental Research

Peter F Stadler
Leipzig University

Research Article

Keywords: small Proteins, metaproteogenomics, Peptide-to-Spectra-Matches, microbial communitities

Posted Date: February 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-199254/v1

RESEARCH

Direct mapping of Peptide-to-Spectra-Matches to genome information facilitates qualifying proteomics information

John Anders[1*†], Hannes Petruschke[3], Nico Jehmlich[3], Sven-Bastiaan Haange[3], Martin von Bergen[3,2] and Peter F Stadler[1,4,5,6,7,8]

Abstract

Background: Small Proteins have received increasing attention in recent years. They have in particular been implicated as signals contributing to the coordination of bacterial communities. In genome annotations they are often missing or hidden among large numbers of hypothetical proteins because genome annotation pipelines often exclude short open reading frames or over-predict hypothetical proteins based on simple models. The validation of novel proteins, and in particular of small proteins (sProteins), therefore requires additional evidence. Proteogenomics is considered the gold standard for this purpose. It extends beyond established annotations and includes all possible open reading frames (ORFs) as potential sources of peptides, thus allowing the discovery of novel, unannotated proteins. Typically this results in large numbers of putative novel small proteins fraught with large fractions of false-positive predictions.

Results: We observe that number and quality of the Peptide-to-Spectra-Matches (PSMs) that map to a candidate ORF can be highly informative for the purpose of distinguishing proteins from spurious ORF annotations. We report here on a workflow that aggregates PSM quality information and local context into simple descriptors and reliably separates likely proteins from the large pool of false-positive, i.e., most likely untranslated ORFs. We investigated the artificial gut microbiome model SIHUMIx, comprising eight different species, for which we validate 5114 proteins that previously have been annotated only as hypothetical ORFs. In addition, we identified 37 non-annotated protein candidates for which we found evidence in proteomic and transcriptomic level. Half (19) of these candidates have close functional homologs in other species. Another 12 candidates have homologs designated as hypothetical proteins in other species. The remaining six candidates are short ($<$ 100 AA) and are most likely *bona fide* novel proteins.

Conclusions: The aggregation of PSM quality information for predicted ORFs provides a robust and efficient method to identify novel proteins in proteomics data. The workflow is in particular also capable of identifying small proteins and frameshift variants. Since PSMs are explicitly mapped to genomic locations, it furthermore facilitates the integration with transcriptomics data and other source of genome-level information.

Keywords: small Proteins, metaproteogenomics, Peptide-to-Spectra-Matches, microbial communitities

Background

Small proteins (sProteins) with a size below 100 amino acids have received increasing attention in particular in prokaryotes. Recent studies has revealed indispensable biological functions of some sProteins. CydX (37 AA), for instance, regulates the activity of cytochrome oxidase and thus ATP production in *E. coli* [1], and SgrT (43 AA) is an inhibitor of the EIICBGlc glucose transporter regulating glucose uptake [2]. Systematic surveys keep identifying large number of sProteins in prokaryotes, see e.g. [3, 4], hence it has become clear that sProteins are not rare peculiarities. The human gut microbiome, for instance, features thousands of sProteins, many of which are to predicted to function in in cell-cell communication [5]. Nevertheless, the available information has remained comparably sparse due the technical difficulties with their detec-

*Correspondence: john@bioinf.uni-leipzig.de
[1] Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16–18, D-04107 Leipzig, Germany
Full list of author information is available at the end of the article
†Corresponding author

tion and identification with both computational and experimental methods.

The annotation of newly sequenced genomes is primarily based on homology using already existing gene annotations from related species as a basis. By definition, this approach is limited to homologs of genes that have been described already in at least one species. The method is also susceptible to incorrect entries in protein data bases. Complementarily, putative coding sequences can be recognized with the help of Markov models that classify open reading frames (ORFs). To obtain a reliable signal, usually a minimum length of 100 codons is required in genome annotation [6]. These methods become unreliable for shorter ORFs, including those compiled in the `BactPepDB` [7], which surveys the complete prokaryotic genomes available for peptides with a length between 10 and 80 amino acids. Comparative approaches, in particular methods such as `RNAcode` [8] that evaluate sequence alignments rather than single sequences, can reliably recognize even very short coding sequences. They lose their power, however, if not enough genomes in a suitable genetic distance are available. To-date, the computational prediction of sProteins is thus by no means an easy routine task. Ribosome profiling [9] also provides information on translated regions and thus constitutes an alternative manner to identify putative novel proteins.

The gold standard for detecting sProteins is their direct identification in bottom-up proteomics. This technique relies on proteolytically cleaved proteins and subsequent analysis by LC-MS/MS [10]. Classic bottom-up proteomics protocols, however, tend to identify few sProteins since the small size implies that sProteins often yield only a single proteotypic peptide [11–13]. This issue is aggravated by the fact that peptide identification itself depends on underlying databases of predicted polypeptides and corresponding decoys. Tools such `Mascot` [14], `comet` [15], `MS-GB+` [16] and many others, therefore cannot identify peptides that are not in the set of protein annotations provided *a priori*. A peptide identified in this manner is referred to as Peptide-to-Spectra-Match (PSM).

Proteogenomics approaches typically make use of a conceptual translation of the genome into all six reading frames as the basis for peptide identification. This results in much larger `6frame` databases and thus a (moderate) reduction of sensitivity, but completely avoids all annotation-related biases [17–19]. With a focus on sProteins, it is also possible to extend annotations with additional predictions of (short) ORFs with high coding potential [20]. Already two decades ago EST data have been used to predict novel isoforms to allow the identification of proteins arising from splice variants [21]. More recently, the same idea has been used with hypothetical splice variants to identify missense SNPs, short indels, chimeric proteins, and intron retention [18, 19]. Metaproteomics [22], i.e., the application of proteogenomics to entire communities, incurs an additional layer of complexity for data analysis due to the need of disentangling different, but often closely related species [23, 24].

The focus of this study was the discovery of novel, unannotated proteins, in particular those that have not been flagged as likely candidates by homology-based genome annotation. This problem is more difficult than just verifying an annotated protein candidate because the overlooked cases are often short, have no or only poorly described homologs in other species, harbor unusual features such as frameshifts, or overlap incorrect annotations. As a consequence, the sensitivity needs to be increased, which necessarily leads to a rapidly growing number of false positive predictions. Here we describe a workflow to prioritize the candidates based on aggregated quality measures of the PSMs that map to candidate and translational status of overlapping annotation items.

Results

Accuracy of identifying candidate proteins

Our goal is to call protein candidates with high sensitivity and to tightly control the false positive rate at the same time. Since we are primarily interested in novel sProteins, we want to do this is in a manner that does not rely on any existing annotation. Thus we start by mapping all PSMs of sufficient quality (see Methods) to the genome and use the genomic map of PSMs to determine candidate proteins using a set of rules (see Methods). A candidate extends downstream to the closest stop codon, while the upstream end is determined by first start codon upstream of the upstream-most PSM mapped to the candidate.

In order to determine how well true proteins can be discriminated from false candidates on the basis of properties of mapped PSMs, we use an extensive data set [25] for *E. coli*. *E. coli* is nearly perfectly annotated, hence unannotated candidates most likely are false positive calls. In addition, we compare candidate calls using a `6frame` database with calls based on a database of annotated proteins (`proteome`). While we expect that the sensitivity of `6frame` is reduced compared to `proteome`, we can use candidates found only with the `6frame` but not with the `proteome` database to estimate the false positive rate.

If only a single PSM is required to identify a protein, we observe that the majority of the annotated *E. coli* proteins is called using both the `6frame` data base and proteome database. The consistency increases rapidly

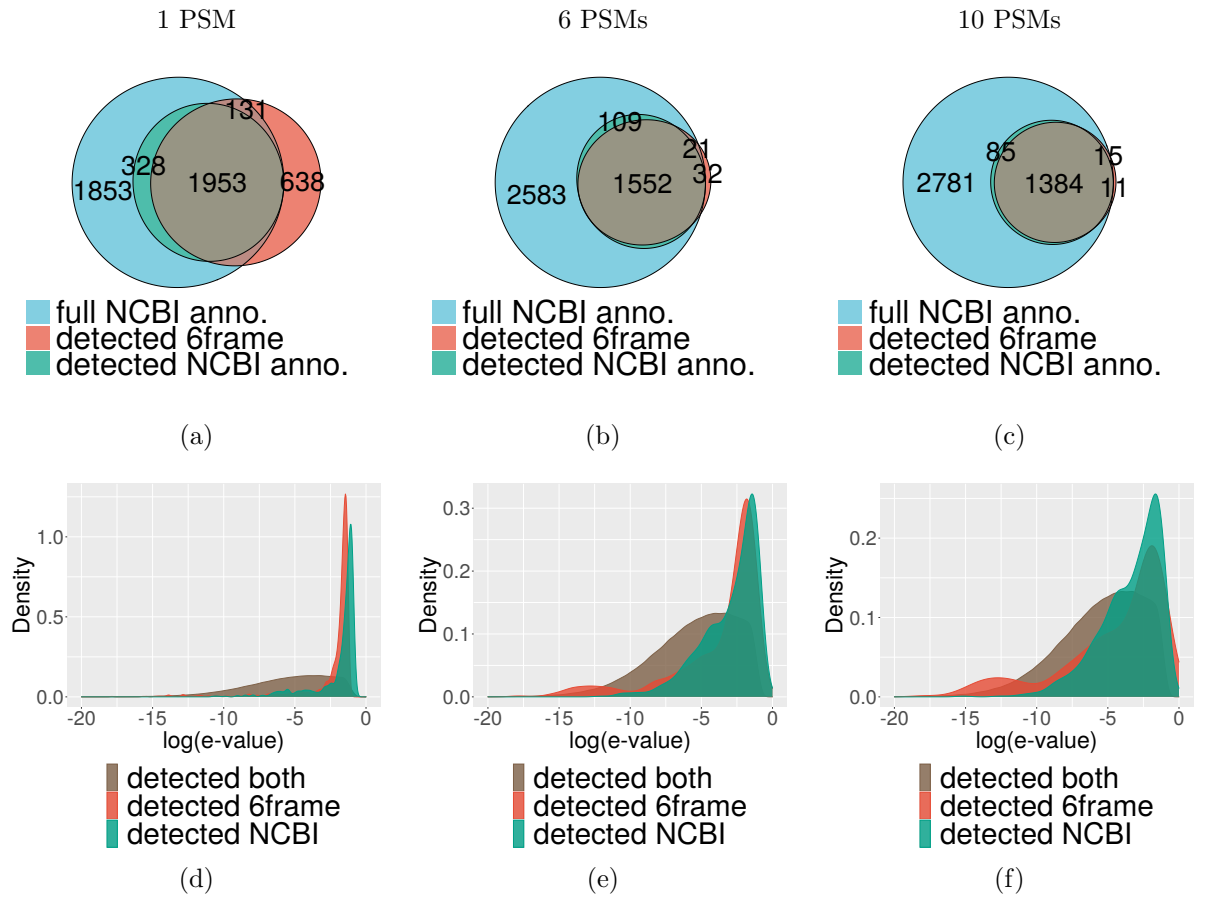


Figure 1: Top: Comparison of `6frame` and proteome data bases for peptide identification using comet search algorithm. Three sets of proteins are displayed. Blue is used for the set of all proteins in the NCBI annotation. The green set represents all proteins that are identified by at least n PSMs using the NCBI annotation as a search data base. The red set are proteins which are identified by at least n PSMs using the `6frame` data base.
Below: The score distributions of the PSMs for each intersection are shown. The score is given in a common logarithm. Red shows the set of proteins which are only detected by the `6frame` database. The green set are the proteins which are only detected by using the NCBI annotation as the search database. Grey is the intersection of proteins that are identified by both databases.
In (a) only a single PSM is required, there are noticeable differences between annotation based and `6frame` based searches. The differences quickly decrease when larger number of PSMs per candidate in required. In (b) and (c) we require six and ten not necessarily distinct PSMs. While recovery is reduced by about 15% and 20%, resp., the two methods yield nearly identical results. In the `6frame`-based data, very few unannotated candidates remain.
Discrepancies between the data bases are related almost exclusively to PSMs with poor scores. The number of such PSMs decreases drastically if only PSMs mapping to confidently identified proteins are considered (e,f). Candidate proteins that are only detected by one data base show an accumulation of poorly scoring PSMs, we thus have to interpret these as mostly false positives.

if more – not necessarily distinct – PSMs are required, Fig. 1 (a-c) and Additional File 1: Figure 1. Considering the score distribution of PSMs, i.e., the confidence with which they are identified from the MS/MS spectra, we observe that most of PSMs that are mapped by one but not the other database are of low confidence. Low confidence PSMs, furthermore, are strongly enriched in proteins to which only very few PSMs are

mapped, Fig. 1 (d-f). This matches the observation that false positive PSMs accumulate among unannotated ORFs [26]. These observations suggest that it is possible to devise an aggregate statistics of PSM quality scores that is capable of assessing candidate proteins even in the absence of multiple, distinct peptides.

As a second measure how well we can replicate the original annotation by using a 6frame database, we quantify the differences between start sites predicted with 6frame database and start sites reported in the original annotation. Their 3'-ends match perfectly as they are determined by the stop codons. For most of the annotated candidates, we recover the original length of the annotated protein (dominating peak at 0 in Fig. 2). For a fraction of the data we predict shorter candidates as compared to the annotation, presumably due to a lack of PSM coverage on the N-terminal part of the candidate. In a small number of cases our candidates begin upstream of the given annotation. This concerns 24 proteins with 6 PSMs. Fig. 2 shows one example, the fatty acyl-CoA synthetase *FadD*. Here, PSM evidence clearly shows that the true start codon is located upstream of the annotated coding sequence (CDS). Similar arguments can be made for 4 of the 24 cases with extended N-termini, the full list can be found in the result web page in the supplement material., indicating the despite outstanding quality of the annotation of the *E. coli* K12 reference genome, it is still not perfect and proteogenomics data are able to correct some of the remaining inaccuracies.

This observation prompted us to also inspect the 11 "false positives" that are supported by 10 or more PSMs. It turns out that two of them correspond to two parts of the formate dehydrogenase O subunit alpha, which our pipeline did not recognize due to a (presumably erroneous) stop codon in the genomic sequence. Two candidates are the two parts of the peptide chain release factor RF2, which has long been known to contain an obligatory frameshift [27]. Its peptides thus appear in two distinct predicted ORFs, neither of which completely matches the annotation. Several mRNAs in *E. coli* are known to produce minor variants that include a frameshift [28]. Two additional candidates are an IS5 transposase, for which frameshift has also been reported [29], and the transcriptional regulator GlpR, which, according to the UniProt annotation also harbors a frameshift.

This leaves only 5 ORFs as likely false positives. Surprisingly, these candidates are well distinguished by the distribution of PSM scores: while the frameshift proteins harbor mostly well-scoring PSMs, the remaining, likely false positives are matched only by PSMs with poor scores. This observation further supports the idea to aggregate PSM quality statistics into a quality measure for predicting a protein.

Table 1: Ambiguous mapping of PSMs in the SIHUMIx dataset with annotation based and `6frame` databases. More than 95% of the PSMs are unique, and thus can be unambiguously assigned to one of the species of the consortium, and the majority of the remaining PSMs matches only two positions on the metagenome (multiplicity= 2)

multiplicity	proteome	6frame
1	0.9582 %	0.9599 %
2	0.0288 %	0.0271 %
3	0.0056 %	0.0051 %
4	0.0052 %	0.0051 %
5	0.0009 %	0.0014 %
6	0.0006 %	0.0006 %

Increasing the sensitivity of requiring 6 PSMs per candidate moderately increases the number of candidate proteins to 29. Using the number of candidates predicted from with `6frame` proteogenomics approach that do not match the annotation (or are not called using a `proteome` database) shows that the FDR quickly drops with the number of PSMs that are required to call a candidate, Addition File 1: Figure 2. Our analysis of the *E. coli* data suggests that a coverage of 6-10 PSMs is sufficient to identify likely candidate proteins. Notably, these PSMs may correspond to the same peptide. It is unlikely that the *E. coli* genome harbors many undiscovered candidates. We therefore analyse a larger, much less well annotated data set next.

Metaproteogenomics of SIHUMIx

The proteomics data for SIHUMIx was analyzed using a combined `6frame` database for the eight species. In order to verify that this approach can properly separate the spectra from the different species we determined the number of PSMs mapped to more than one species, Table 1. For this model system we also analysed extensive RNA-seq data as a means of supporting proteogenomics-based predictions. It is not unexpected that there is only moderate agreement between protein and RNA abundances in Fig. 4, since RNA/protein ratios are known to vary considerably between organism [30].

The rate of detection of known and hypothetical proteins in the eight SIHUMIx species, as expected, correlates with the relative abundance in the mixture, see Table 2. There is near perfect congruence between `6frame` and `proteome` database, see Additional File 1: Table 1.

The distribution of known and hypothetical protein differs dramatically across the eight SIHUMIx species.

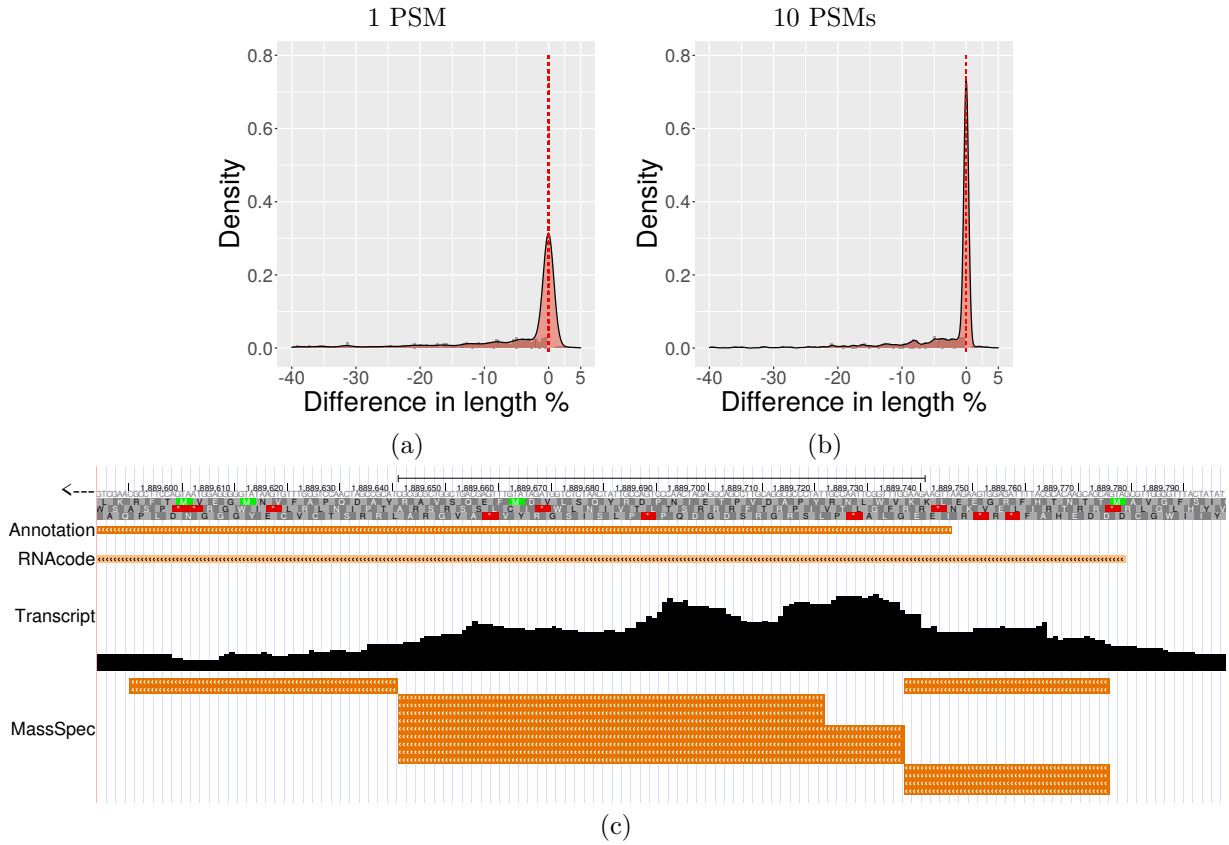


Figure 2: Comparison between the length difference of candidates from `6frame` search and the NCBI annotation, expressed as fraction of the total protein length. A value of 0% indicates that candidate and annotation are identical. Negative values mean that the candidate is shorter than the annotated protein due to lack of coverage by PSMs towards the N-terminus. Positive values imply that the candidate has a longer N-terminal sequence that extended the annotation. Below, the N-terminus of the fatty acyl-CoA synthetase *FadD* is shown. It is misannotated and extends to canonical start. The extension is also supported by evidence of conserved coding sequences determined by `RNAcode` [8].

In most species, the majority of the proteins is annotated as hypothetical based on the level of evidence available in the data sources. Since the confidence levels are unlikely to be truly consistent between species due to differences in the efforts that have been expended for their annotation, these number have to be interpreted with caution. They do, however, at least reflect qualitative trends.

Novel proteins in SIHUMIx

We discovered a total of 419 unannotated protein candidates supported by at least 6 PSMs in SIHUMIx. Since these initial candidates also include all those predictions that overlap annotated proteins in a different reading frame, we expect *a priori* that most of them are false positives. While it is manageable to manually evaluate a few hundred candidate proteins in a data set of particular interest, this is not feasible for routine applications and thus requires computational support. In order to better understand this candidate set we systematically gathered all information on them that is readily accessible by computational means. This leads to a natural workflow for prioritizing and validating the candidates.

Homology search against the NCBI protein database identifies 60 of 419 candidates with extensive similarity to proteins with a functional annotation in related species. These cases are clearly shortcomings in the available annotations and constitute a positive control for our approach and help to establish the criteria that can be applied to the remaining candidates. We exclude these 60 proteins from further analysis because we are interested here in those candidates proteins that cannot be found trivially by homology-based methods. In addition to these homologs of known proteins, we

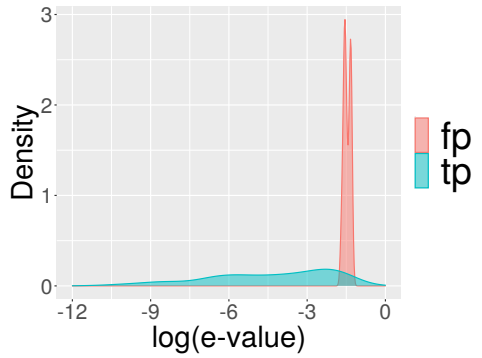


Figure 3: Score distributions of the PSMs mapping to the unannotated candidate proteins in *E. coli*. Those identified as likely true positives harbours PSMs with excellent scores, while those identified as likely false positives by manual inspection have only low-scoring PSMs.

Figure 4: Species composition of SIHUMIx from proteogenomic (red) and transcriptome data (blue). The relative frequency is given by number of PSMs per species divided by total number of PSMs for the proteogenomics composition respectively number of reads for transcriptome composition. Both the mapped reads and the PSMs were normalized to complete genomic size and proteogenomic (6frame DB) size respectively.

have another 47 of 419 candidates are homologs of hypothetical proteins.

We first considered the distribution of the e-values of the PSMs that contribute to each candidate protein. Fig. 3 already strongly suggests that this is a reliable predictor. We use the average $\hat{s}$ of the scores $s = -\log(\text{e-value})$ for the three best PSMs as an aggregate descriptor. Fig. 7 summarizes the data with at least 6 supporting PSMs. Almost all candidates with $\hat{s} > 3.5$ have homologous known proteins in other species. As an example, the *B. producta* candidate

Table 2: Summary of the number of proteins detected with at least 10 and 6 PSMs in SIHUMIx proteomics data using the `6frame` translation. Novel (nov) proteins are not contained in the annotation, hypothetical (hyp) proteins are annotated but tagged with low confidence (see Methods for details), known refer to all proteins for which higher levels of confidence are associated with the available annotation. The eight species are ordered by decreasing abundance. The last column gives the fraction of the annotated proteins that were detected.

At least 10 PSMs per candidate

Species	nov	hyp	known	%
B. theta.	37	1975	248	45.9
B. producta	52	1138	132	23.2
E. coli	26	150	988	26.8
E. ramosum	10	355	53	13.7
B. longum	16	128	0	7.4
A. caccae	17	549	100	19.3
L. plantarum	31	83	28	3.7
C. butyricum	14	135	32	4.1

A least 6 PSMs per candidate

Species	nov	hyp	known	%
B. theta.	72	2118	256	49.0
B. producta	103	1289	143	26.1
E. coli	65	182	1127	30.9
E. ramosum	30	431	65	16.7
B. longum	42	170	0	9.8
A. caccae	39	632	119	22.3
L. plantarum	48	116	36	5.1
C. butyricum	26	176	39	5.3

nov_57 is shown in Fig. 6 (top). It has a probable length of 72 amino acids and shows a recognizable homology with an deny late kinase of similar length from Listeria.

In total, we find 47 of 419 candidates with $\hat{s} > 3.5$. We first inspect all candidates with more than 10 high scoring PSMs. Interspersed among these known genes are three novel protein (*B. producta* nov_5, *B. theta.* nov_59 and nov_131). Nov_5 is clearly a complete protein, while nov_59 and nov_131 may be associated with frameshifts and constitute only parts of a protein. The most prominent candidate, *B. producta* nov_5 is shown in Fig. 5, lower panel. It has a likely length of 62 amino acids, judging from both the observed PSMs and the transcriptome data. Most but not all of these high-confidence candidates show evidence of transcription. Low RNA levels do not necessarily imply that the predicted protein is a false positive. In fact detection limits for RNA and protein may be vastly different. The typically much longer half-life of proteins may also contribute to explaining the presence of protein with low or undetectable RNA levels.

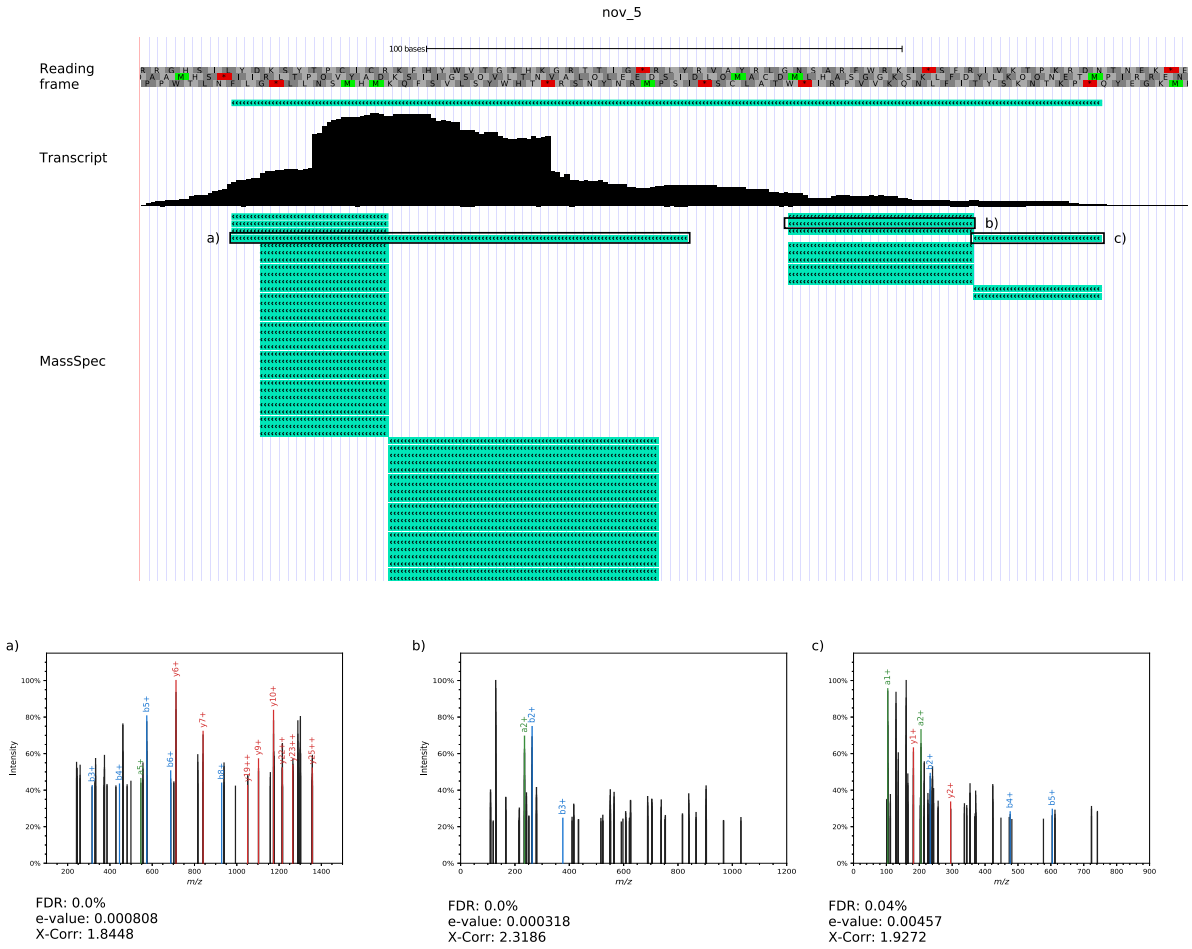


Figure 5: Protein candidate 5 in *Blautia producta*, for which no homologs can be found. The list of PSMs is cropped and not all 168 PSMs are shown. Below three mass spectra with top confidence PSMs are shown.

In five cases (*B. producta* nov_1, *E. coli* nov_8, *B. theta.* nov_34, *B. theta.* nov_131, *B. theta.* nov_61) there also an annotated protein in the same reading direction. Owing to our definition of the candidates, which extends to the nearest in-frame stop and the nearest in-frame start codon, this kind of overlap is indicative of an annotation error or a frameshift. Inspection shows that for *B. producta* nov_1 the available annotation of a TetR family transcriptional regulator extends across the stop codon. The remaining signals likely pertain to frameshifts. For *B. producta* nov_126 there is some weak evidence for translation of the annotated gene on the opposite strand, and convincing evidence for translation of a Cna B-type domain-containing protein corresponding to nov_126 that has been left unannotated.

Only three candidates with 6-9 PSMs have $\hat{s} \geq 3.5$: *B. producta* nov_216, an IS66 family transposase, nov_307 in hypothetical protein without functional annotation, and *E. coli* nov_302, the frameshift fragment of peptide chain release factor RF2 already discussed above.

The analysis of the remaining candidates with $\hat{s} < 3.5$ is much less straightforward. Although the overwhelming majority of them shows no homology to a known or hypothetical protein, this set contains at least a small number of proteins with known homologs with convincing proteomics evidence: *B. producta* nov_174 $\hat{s} = 3.3$, *B. producta* nov_215 $\hat{s} = 2.9$, *B. theta.* nov_180 $\hat{s} = 2.8$, and possibly *E. coli* nov_122 $\hat{s} = 2.3$. Some others, such as *B. producta* nov_84 $\hat{s} = 3.0$ and nov_28 $\hat{s} = 2.4$, however, are almost certainly false positives. A few curious cases, such as *E.*

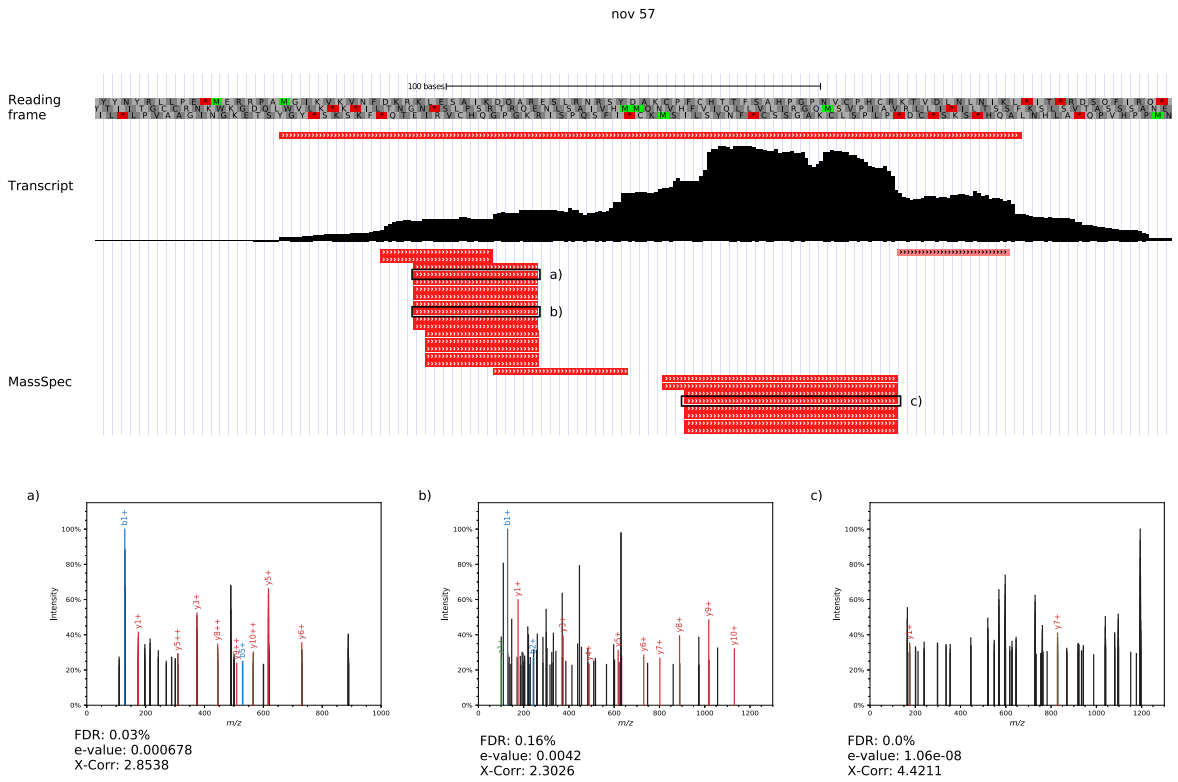


Figure 6: Protein candidate 57 in *B. producta*, for which only hypothetical homologs can be found. The rightmost PSM is ambiguous and provides little evidence for the candidate. Below three mass spectra for top confidence PSMs are shown.

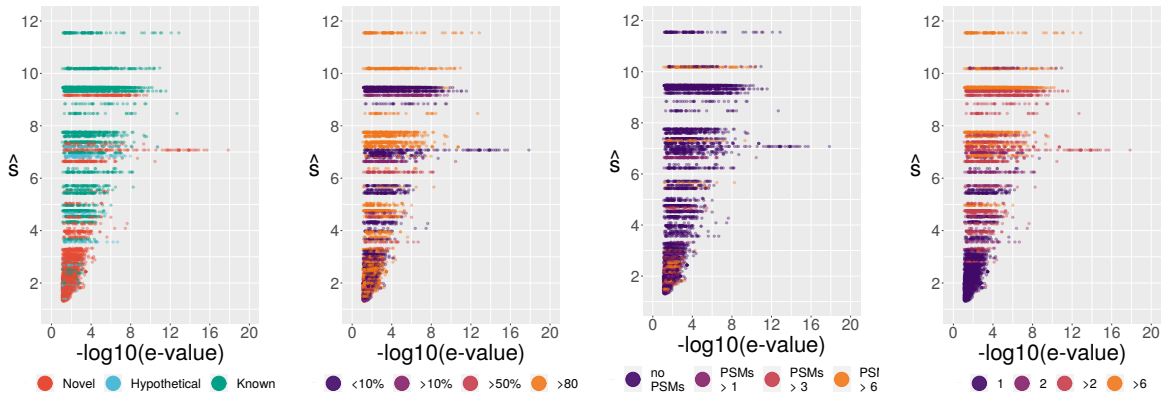
coli nov_123, $\hat{s} = 2.1$, are indicative of incorrect stop-codons or read-through; here the candidate sequence matches a GntR family protein from related species whose sequences extend beyond the stop codon of the annotated *E. coli* GntR gene immediately upstream of nov_123.

Protein expression of the opposite strand is a good indication that a candidate is a false positive: while overlapping ORFs are not uncommon in bacteria, long overlaps of coding regions are very rare [31, 32]. There are, however, a handful of exceptions: As already mentioned above, *B. producta* nov_126 is much more plausible than the potentially expressed ORF on the opposite strand. A few additional cases are supported by many good PSMs mapping to two or more distinct peptides. The best example in our data is *L. plantarum* nov_19, $\hat{s} = 3.28$. It deserves a more detailed follow-up.

For moderate values of $\hat{s} < 3.5$, therefore, we need additional criteria to distinguish between *bona fide* protein detections, likely parts of other proteins that should prompt an update of an known protein, and false positives. We therefore inspected additional descriptors. First, we consider number of distinct peptides corresponding to the PSMs belonging to a given candidate. Supporting the use of $\hat{s}$ as a valuable indicator, we find that with few exceptions, the candidates with large $\hat{s}$ values have multiple peptides, while for small $\hat{s}$, most candidates are supported only by a single peptide. The few notable exceptions (nov_174, nov_215, nov_180) with more than 3 distinct peptides have already been noted above as proteins with known homologs.

Workflow for Identifying and Prioritizing Candidate Proteins

The detailed evaluation of both the *E. coli* and the SIHUMIx metaproteomics date reported above informs the workflow for the identification of novel proteins shown in Fig. 8. It primarily relies of the number of PSMs mapped to an ORF and the distribution of their e-values, irrespective of whether or not there are multiple distinct peptides. The initial decision is based



(a) (b) (c) (d)

Figure 7: Distribution of $-\log$(e-value) for the PSMs contributing to each candidate. All candidates with at least 6 mapped PSMs are shown. Each dot represents a single PSM. On the Y-Axis the median of the top three PSMs $\hat{s}$ is plotted. Thus the observed horizontal lines of the dots, show all PSMs belonging to one candidate. In panel (a) candidates with functionally annotated homologs (green), homologous hypothetical proteins (blue), and no homologs in the NCBI protein database are distinguished. In (b), the same data are colored by RNA expression. Panel (c) shows whether translation was observed in a different ORF or the same or the opposite reading directions, and in panel (d) color coding refers to number of distinct peptide contributing to a candidate. Interactive versions of these plot are available as Online Supplemental Data (see *Availability of data and material*), allowing a detailed inspection.

on the number of PSMs, followed by a cut-off on the average score of the three best PSMs. Together the two values ensure reproducibility of good matches in the data set. For values of $\hat{s} \geq 3.5$, unlikely candidates are only those without distinct peptide matches and no evidence for transcription. For values $2.5 \leq \hat{s} < 3.5$ multiple distinct peptides may rescue an initial negative decision. Here, transcriptomics date are not helpful, since prokaryotic genomes produce diverse non-coding transcripts [33–35], so that transcription in itself cannot be used as a reliably predictor of translation.

Discussion

We have shown here that prokaryotic proteins can be identified with high reliability by considering the PSMs that map to the corresponding genomic location. Using SIHUMIx as an example we found that $\hat{s}$ (the average logarithm of the e-value of the best few PSMs that map to a candidate ORF) is an excellent discriminator between *bona fide* proteins and other false positive signals. In conjunction with the number of PSMs, it is sufficient to identify nearly all of the ORFs in the SIHUMIx data that have functionally annotated homologs in related species and thus are most likely true proteins. In a fine-grained analysis, the number of distinct peptides helps to distinguish likely candidates from background noise in the case of moderate values of $\hat{s}$. Manual inspection also revealed that translation products involving frameshifts can also be detected even if the frame-shifted part is contains only a single detectable peptide. Somewhat surprisingly, RNA expression data add very little to the task of identifying novel proteins.

Taken together, our observations leads us to the workflow summarized in Fig. 8. It is designed to efficiently identify previously unannotated candidates. It can also be employed to validate previously annotated proteins using the same decision criteria, since it accurately reproduces the annotation for known proteins from the PSM data and in some cases can help correcting annotation errors such as incorrect start codons.

Since PSMs are mapped directly to the genomic sequence in our workflow, standard genome browsers can be used to visualize the data. This also facilitates the integration with other data source, including transcriptome data and information on sequence conservation. The presentation of the data in a genome browser supports the manual evaluation of protein candidates in their genomic context, since information of overlapping features, including predicted proteins and PSM data mapping to other reading frames directly accessible.

Candidates identified as (likely) novel proteins can be followed up on by further computational steps. Most importantly, a homology search is likely to identify a large fraction of candidates as homolog of pro-

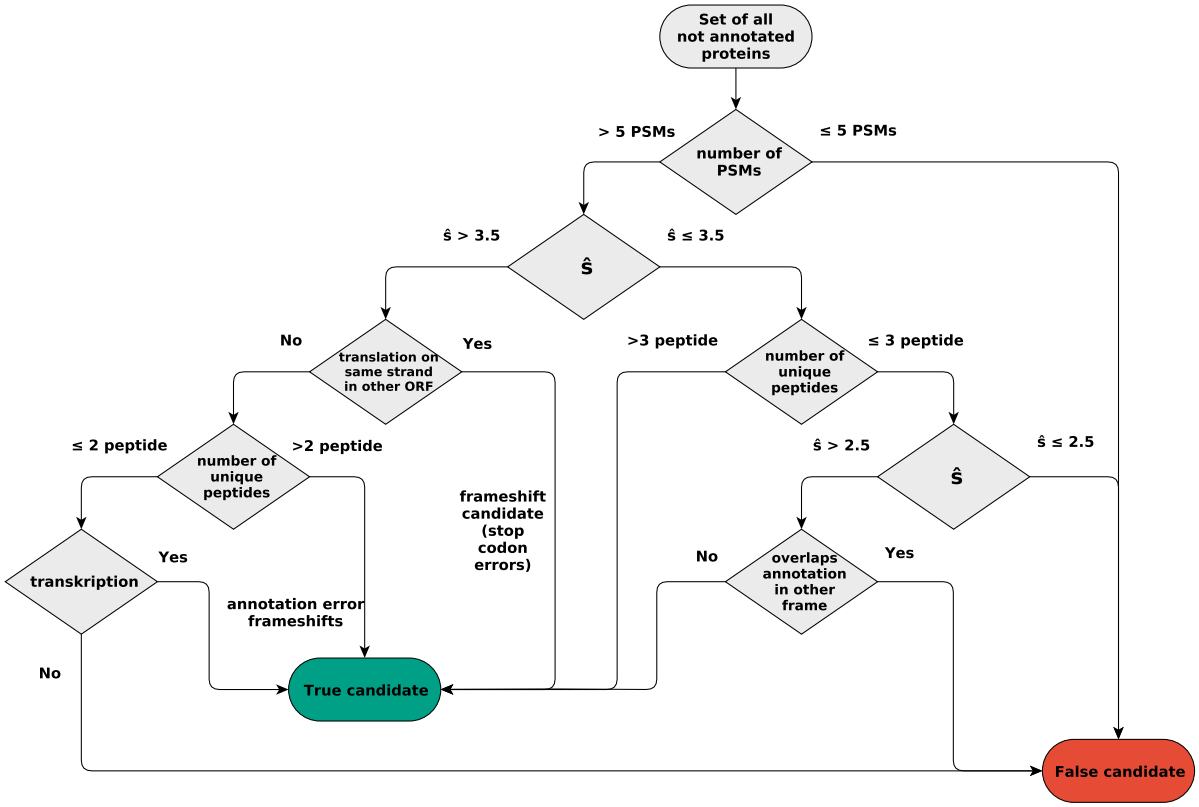


Figure 8: Rules to prioritize candidate proteins for further investigation. A candidate is classified as transcribed if more than 70% of its length is above the median RNA expression level of the organism. Annotated genes which overlap a candidate are classified as translated if they are identified by more than 6 unique PSMs.

teins that have been described already in other species. As in the case of the SIHUMIx example, we expect this to leave only a small fraction of novel proteins and homologs that so far have appeared only as "hypothetical proteins".

The workflow of Fig. 8 provides a robust way to identify novel proteins, including sProteins, from large mass spectrometry data. The method is applicable not only to a single species but also to metaproteomics data, provided the species composition of the sample is known. In the artificial gut community SIHUMIx we found 37 non-annotated novel proteins, among them six sProteins. Applications to microbial communities, however, are likely to be limited to the most abundant species, since the probability to identify a protein depends on its relative abundance in the sample.

Materials and Methods

Proteomics data sets For our analysis we used two different tandem mass spectrometry data set. One is a data set from a single strain *E. coli K-12*, grown under standard conditions. The data set consist of seven experimental replicas and part of the publication [25]. The SIHUMIx datasets are described in detail in [36, 37]. They comprise 166 independent measurements, of which 90 used a standard protein preparation protocol and the remaining 76 cover different enrichment protocols to elevate the level of small proteins in solution.

More than 5.7 million PSMs were analysed for the project. The search of the SIHUMIx data sets against the `6frame` data base (the main analysis to find new protein candidates) resulted in over 2,5 million PSMs. Beside different protein enrichment protocols, both Trypsin and Asp-N were used as different cleavage enzymes.

Peptide identification We used `getorf` [38] to retrieve all open reading frame between to stop codons from the genomic DNA sequence without any length constraints. For each ORF we store its amino acid sequence as well as its genomic start and end coordi-

nates. The reading frame is defined as that start coordinate $k \mod 3$ in forward direction and $(k \mod 3) - 3$ in negative direction. We then used `Comet` [15, 39] to search tandem mass spectra against protein sequence databases. Standard search parameters were used form both the `6frame` and the annotated protein databases, with the following exceptions: (i) we allowed semi-digestion at the N-terminus to accommodate fragmentation at the start codon, (ii) we conducted a concatenated search against a decoy database, and (iii) we used the full resolution of the MS/MS spectra.

Estimation of false discovery rates for PSMs. In addition to calculating the FDR a PSM by using a decoy data base within the `comet` software, two alternative approaches to estimate the FDR have been proposed [26, 40]. In the first approach, we assume that a false positive PSM is mapped with equal rate to a translated and non-translated locus. Ignoring the possibility of overlapping proteins in different frames we interpret all n PSMs mapping to one of the five incorrect reading frames of an annotated protein as false positives, resulting in an estimated number of $(6/5)n$ false positives. Of the N PSMs mapping in the correct reading frame, we expect $N - (1/5)n$ to be true positives. We can therefore estimate the false discovery rate as

$$FDR_{ann} = \frac{6}{5}\frac{n}{N+n} \tag{1}$$

where $n + N$ is the total number of PSMs mapped to an annotated locus irrespective of the frame.

Alternatively, we make the assumption that the protein annotation is complete and assume that a fraction α of the genome is covered by annotated proteins. All n' PSMs mapped outside this annotation are counted as false positives. Thus we have

$$FDR_{genom} = \frac{1}{1-\alpha}\frac{n'}{N'} \tag{2}$$

where $N' = N + n$ is the total number of mapped spectra. The prefactor extrapolates the same FDR to the annotated part of the genomes. In order to account for very short ORFs to which no ORFs can be mapped by construction, the factor α can be estimated more accurately by estimating the chance that a randomly drawn PSM from the `6-frame` annotation falls into an annotated region. For *E. coli* this yields $\alpha = 0.293$.

We note that FDR_{ann} is robust against incomplete annotation and also will not change substantially if many wrong genes are falsely annotated. In constrict, FDR_{genome} will only work well for genomes with reasonably complete annotations [26]. We checked consistency of the PSM estimates for the *E. coli* data. Among the $N = 180059$ mapped PSMs we observed $n = 829$ hits to an incorrect reading frames obtain $FDR_{ann} = 0.55\%$, i.e., a slight improvement over `comet`'s internal estimate of 1% from hits in the decoy database. Alternatively, at least in a well-annotated genome such as *E. coli* we may use PSMs mapped to unannotated regions as an estimator. This yields $FDR_{genome} = 0.52\%$. We also validated that, as expected [26] the genome-based FDR estimates are proportional to the FDRs estimated for the decoy database (Additional File 1: Figure 3).

Mapping PSMs to the genome To map PSMs to the genome, we determine its relative position in the ORF or ORFs of the protein or `6frame` database. This position is then directly translated to the genomic coordinates using the known genomic coordinates of the ORFs/proteins. Peptides may map to multiple ORFs/proteins. If this is the case, the multiplicity of the mapping is stored and can be accessed in the genome browser.

Construction and annotation of candidate proteins. We start from the collection of ORFs for a genome. For each ORF, we determine all PSMs that map inside it. The C-terminus of the candidate is determined by the stop codon of the ORF. The N-terminus is the closest canonical start codon before the first mapped PSM, or if no such start codon exists within the ORF, the first position of the ORF.

The candidate proteins are then compared to the protein annotation that is available for the genome in question. A candidate is considered annotated if it overlaps an annotation item in the correct reading frame and reading direction. In each case, we record the difference between the genomic start positions of annotation and candidate.

Protein contained in the available annotations are classified as *known* unless they are tagged with validation levels 1, protein uncertain or 2, protein predicted in UniProt (i.e., lacking evidence from experiment or homology), or carry the annotations frameshifted, internal stop, hypothetical, Putative, or pseudogene. All of these are interpreted as *hypothetical* in Tabel 2.

Transcriptome data The transcriptome data were taken from [25] and mapped with `segemehl` [41] to an index comprising the eight SIHUMIx species as separate chromosomes. Default parameters were used. Annotation files were generated with `samtools` (`http://www.htslib.org/`). Total expression per species was averaged over all replicates.

Visualization. We display the data using the UCSC genome browser [42], which make it easy to integrate them with other data, including transcriptome data, available annotations, as well as custom annotations.

Availability of data and material

The transcriptomics data is available under the bioproject PRJNA655119 https://www.ncbi.nlm.nih.gov/bioproject/655119. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [43] partner repository with the dataset identifier PXD023243. The genomes and corresponding annotations used for the project are all publicly available by The NCBI Assembly database [44] a full list can be found in Additional File 1: Table 2. The following material is available for download from www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/20-002/:

- SIHUMIx track hub (track hub for the UCSC genome browser)
- Result web page (full list of candidates, ecoli annotation errors and interactive plots)
- Validation hash map (Maps each annotated protein in SIHUMIx to a validation level as of the time of the publication)

All scripts which are used to generated the data for this publication are available under https://github.com/studla/PROTMAP.

List of Abbreviations

PSM peptide spectrum map
FDR false discovery rate
AA amino acid

Declarations

Acknowledgements

Funding

This work was supported by German Research Foundation (*Deutsche Forschungs Gemeinschaft*, DFG), grants BE 3184/9-1 (to MvB) and STA 850/36-1 (to PFS) as part of SPP 2002 "Small Proteins in Prokaryotes, an Unexplored World". Publication costs are supported by the DFG and Leipzig University within the program of Open Access Publishing.

Author's contributions

PFS and MvB designed the study, JA wrote the software and analyzed the data, HP, NJ, and SBH produced the MS data and contributed to the analysis. PFS and JA drafted the manuscript. All authors contributed to the interpretation and writing and approved of the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Author details

[1] Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16–18, D-04107 Leipzig, Germany. [2] Institute of Biochemistry, Faculty of Life Sciences, University of Leipzig, Talstraße 33, D-04103 Leipzig, Germany. [3] Department of Molecular Systems Biology, Helmholtz Centre for Environmental Research - UFZ, Permoserstrasse 15 , D-04318 Leipzig, Germany. [4] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig & Competence Center for Scalable Data Services and Solutions Dresden-Leipzig & Leipzig Research Center for Civilization Diseases University Leipzig D-04107 Leipzig, Germany. [5] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany. [6] Department of Theoretical Chemistry, University of Vienna Währinger Straße 17, A-1090 Vienna, Austria. [7] Faculdad de Ciencias, Universidad Nacional de Colombia, Sede Bogotá, Ciudad Universitaria, COL-111321 Bogotá, D.C., Colombia. [8] Santa Fe Institute, 1399 Hyde Park Rd., NM87501 Santa Fe, USA.

References

1. VanOrsdel, C.E., Bhat, S., Allen, R.J., Brenner, E.P., Hobson, J.J., Jamil, A., Haynes, B.M., Genson, A.M., Hemm, M.R.: The *Escherichia coli* CydX protein is a member of the CydAB cytochrome oxidase complex and is required for cytochrome oxidase actvity. J Bacteriology **195**, 3640–3650 (2013). doi:10.1128/JB.00324-13
2. Kosfeld, A., Jahreis, K.: Characterizaton of the interaction between the small regulatory peptide SgrT and the EIICBGlc of the glucose-phosphotransferase system of *E. coli* K-12 metabolites **2**, 756–774 (2012). doi:10.3390/metabo2040756
3. Makarewich, C.A., Olson, E.N.: Mining for micropeptides. Trends Cell Biol **27**, 685–696 (2017). doi:10.1016/j.tcb.2017.04.006
4. Miravet-Verde, S., Ferrar, T., Espadas-García, G., Mazzolini, R., Gharrab, A., Sabido, E., Serrano, L., Lluch-Senar, M.: Unraveling the hidden universe of small proteins in bacterial genomes. Mol Syst Biol **15**, 8290 (2019). doi:10.15252/msb.20188290
5. Sberro, H., Fremin, B.J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M.P., Pavlopoulos, G.A., Kyrpides, N.C., Bhat, A.S.: Large-scale analyses of human microbiomes reveal thousands of small, novel genes. Cell **178**, 1245–1259 (2019). doi:10.1016/j.cell.2019.07.016
6. Su, M., Ling, Y., Yu, J. J.and Wu, Xiao, J.: Small proteins: untapped area of potential biological importance. Frontiers Genetics **4**, 286 (2013). doi:10.1016/j.cell.2019.07.016
7. Rey, J., Deschavanne, P., Tuffery, P.: BactPepDB: a database of predicted peptides from a exhaustive survey of complete prokaryote genomes. Database **2014**, 106 (2014). doi:10.1093/database/bau106
8. Washietl, S., Findeiß, S., Müller, S., Kalkhof, S., von Bergen, M., Hofacker, I.L., Stadler, P.F., Goldman, N.: RNAcode: robust prediction of protein coding regions in comparative genomics data. RNA **17**, 578–594 (2011). doi:10.1261/rna.2536111
9. Olexiouk, V., Van Criekinge, W., Menschaert, G.: An update on sORFs.orgt a repository of small ORFs identified by ribosome profiling. Nucleic Acids Research **46**, 497–502 (2017). doi:10.1093/nar/gkx1130
10. Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.-C., Yates, J.R.: Protein analysis by shotgun/bottom-up proteomics. Chem. Rev. **113**, 2343–2394 (2013). doi:10.1021/cr3003533
11. Müller, S.A., Kohajda, T., Findeiß, S., Stadler, P.F., Washietl, S., Kellis, M., von Bergen, M., Kalkhof, S.: Optimization of parameters for coverage of low molecular weight proteins. Anal. Bioanal. Chem. **398**, 2867–2881 (2010). doi:10.1007/s00216-010-4093-x
12. Ma, J., Diedrich, J.K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., Yates, J.R., Saghatelian, A.: Improved identificaton and analysis of small open reading frame encoded polypeptides. Analytical Chem. **88**, 3967–3975 (2016). doi:10.1021/acs.analchem.6b00191
13. Shishkova, E., Hebert, A.S., Coon, J.J.: Now, more than ever, proteomics needs better chromatography. Cell Systems **3**, 321–324 (2016). doi:10.1016/j.cels.2016.10.007
14. Koenig, T., Menze, B.H., Kirchner, M., Monigatti, F., Parker, K.C., Patterson, T., Steen, J.J., Hamprecht, F.A., Steen, H.: Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. J. Proteome Res. **7**, 3708–3717 (2008). doi:10.1021/pr700859x
15. Eng, J.K., Jahan, T.A., Hoopmann, R. Micheal: Comet: an open source tandem mass spectrometry sequence database search tool. Proteomics **13**, 22–24 (2013). doi:10.1002/pmic.201200439
16. Kim, S., Pevzner, P.A.: MS-GF$^+$ makes progress towards a universal database search tool for proteomics. Nat. Commun. **5**, 5277 (2014). doi:10.1038/ncomms6277
17. Nesvizhskii, A.: Proteogenomics: concepts, applications, and computational strategies. Nat Methods **11**, 1114–1125 (2014). doi:10.1038/nmeth.3144
18. Walley, J.W., Briggs, S.P.: Dual use of peptide mass spectra: Protein atlas and genome annotation. Current Plant Biology **2**, 21–24 (2015). doi:10.1016/j.cpb.2015.02.001
19. Sheynkman, G.M., Shortreed, M.R., Cesnik, A., Smith, L.M.: Proteogenomics: Integrating next-generation sequencing and mass spectrometry to characterize human proteomic variation. Annu. Rev. Anal. Chem. **9**, 521–545 (2016). doi:10.1146/annurev-anchem-071015-041722
20. Fuchs, S., Kucklick, M., Lehmann, E., Beckmann, A., Wilkens, M., Kolte, B., Mustafayeva, A., Ludwig, T., Diwo, M., Wissing, J., Jänsch,

L., Ahrens, C.H., Ignatova, Z., Engelmann, S.: A proteogenomics workflow to uncover the world of small proteins in *Staphylococcus aureus*. Technical report, bioRxiv. doi:10.1101/2020.05.25.114132
21. Choudhary, J.S., Blackstock, W.P., Creasy, D.M.C., S., J.: Matching peptide mass spectra to EST and genomic DNA databases. Trends Biotech. **19**, 17–22 (2001). doi:10.1016/S0167-7799(01)00004-X
22. Maron, P.-A., Ranjard, L., Mougel, C., Lemanceau, P.: Metaproteomics: A new approach for studying functional microbial ecology. Microbial Ecology **53**, 486–493 (2007). doi:10.1007/s00248-006-9196-8
23. Seifert, J., Herbst, F., Nielsen, P.H., Planes, F.J., Jehmlich, N., Ferrer, M., von Bergen, M.: Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities **13**, 2786–2804 (2013). doi:10.1002/pmic.201200566
24. Schiebenhoefer, H., Van Den Bossche, T., Fuchs, S., Renard, B.Y., Muth, T., Martens, L.: Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. Expert Rev Proteomics **16**, 375–390 (2019). doi:10.1080/14789450.2019.1609944
25. Mendler, A., Geier, F., Haange, S.B., Pierzchalski, A., Krause, J.L., Nijenhuis, I., Froment, J., Jehmlich, N., Berger, U., Ackermann, G., Rolle-Kampczyk, U., von Bergen, M., Herberth, G.: Mucosal-associated invariant T-Cell (MAIT) activation is altered by chlorpyrifos- and glyphosate-treated commensal gut bacteria. J Immunotoxicol. **17**, 10–20 (2020). doi:10.1080/1547691X.2019.1706672
26. Krug, K., Carpy, A., Behrends, G., Matic, K., Soares, N.C., Macek, B.: Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments. Mol. Cell Proteomics **12**, 3420–3430 (2013). doi:10.1074/mcp.M113.029165
27. Craigen, W.J., Caskey, C.T.: Expression of peptide chain release factor 2 requires high-efficiency frameshift. Nature **322**, 273–275 (1986). doi:10.1038/322273a0
28. Gurvich, O.L., Baranov, P.V., Zhou, J., Hammer, A.W., Gesteland, R.F., Atkins, J.F.: Sequences that direct significant levels of frameshifting are frequent in coding regions of *Escherichia coli*. EMBO J. **33**, 5941–5950 (2003). doi:10.1093/emboj/cdg561
29. Siguier, P., Gourbeyre, E., Chandler, M.: Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol. Rev. **38**, 865–891 (2014). doi:10.1111/1574-6976.12067
30. Karpinets, T.V., Greenwood, D.J., Sams, C.E., Ammons, J.T.: RNA:protein ratio of the unicellular organism as a characteristic of phosphorous and nitrogen stoichiometry and of the cellular requirement of ribosomes for protein synthesis. BMC Biology **4**, 30 (2006). doi:10.1186/1741-7007-4-30
31. Johnson, Z.I., Chisholm, S.W.: Properties of overlapping genes are conserved across microbial genomes. Genome Res. **14**, 2268–2272 (2004). doi:10.1101/gr.2433104
32. Pallejà, A., Harrington, E.D., Bork, P.: Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? BMC Genomics **9**, 335 (2008). doi:10.1186/1471-2164-9-335
33. Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P.F., Vogel, J.: The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature **464**, 250–255 (2010). doi:10.1038/nature08756
34. Cohen, O., Doron, S., Wurtzel, O., Dar, D., Edelheit, S., Karunker, I., Mick, E., Sorek, R.: Comparative transcriptomics across the prokaryotic tree of life. Nucleic Acids Res. **44**, 46–53. doi:10.1093/nar/gkw394
35. Harris, K.A., Breaker, R.R.: Large noncoding RNAs in bacteria. In: Storz, G., Papenfort, K. (eds.) Regulating with RNA in Bacteria and Archaea, pp. 515–526. ASM Press, Washington, DC (2019). doi:10.1128/microbiolspec.RWR-0005-2017
36. Petruschke, H., Anders, J., Stadler, P.F., Jehmlich, N., von Bergen, M.: Enrichment and identification of small proteins in a simplified human gut microbiome. J. Proteomics **213**, 103604 (2020). doi:10.1016/j.jprot.2019.103604
37. Schäpe, S.S., Krause, J.L., Engelmann, B., Fritz-Wallace, K., Schattenberg, F., Liu, Z., Müller, S., Jehmlich, N., Rolle-Kampczyk, U., Herberth, G., von Bergen, M.: The simplified human intestinal microbiota (SIHUMIx) shows high structural and functional resistance against changing transit times in in vitro bioreactors. Microorganisms **7** (2019). doi:10.3390/microorganisms7120641
38. Rice, P., Longden, I., Bleasby, A.: EMBOSS: The European Molecular Biology Open Software Suite. Trends Genetics **16**, 276–277 (2000). doi:10.1016/s0168-9525(00)02024-2. The EMBOSS Website: http://emboss.open-bio.org/
39. Eng, J.K., Hoopmann, M.R., Jahan, T.A., Egertson, J.D., Noble, W., MacCoss, M.J.: A deeper look into Comet – implementation and features. J Am Soc Mass Spectrom **26**, 1865–1874 (2015). doi:10.1007/s13361-015-1179-x
40. Zhang, K., Fu, Y., Zeng, W.-F., He, K., Chi, H., Liu, C., Li, Y.-C., Gao, Y., Xu, P., He, S.-M.: A note on the false discovery rate of novel peptides in proteogenomics. Bioinformatics **31**, 3249–3253 (2015). doi:10.1093/bioinformatics/btv340
41. Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., Kunz, M., Holdt, L.M., Teupser, D., Hackermüller, J., Stadler, P.F.: A multi-split mapping algorithm for circular RNA, splicing, trans-splicing, and fusion detection. Genome Biology **15**, 34 (2014). doi:10.1186/gb-2014-15-2-r34
42. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringe, T.H., Zahler, A.M., Haussler, D.: The human genome browser at UCSC. Genome Res. **12**, 996–1006 (2002). doi:10.1101/gr.229102
43. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Pérez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yılmaz, Ş., Tiwary, S., Cox, J., Audain, E., Walzer, M., Jarnuczak, A.F., Ternent, T., Brazma, A., Vizcaíno, J.A.: The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res **47**(D1), 442–450 (2019). doi:10.1093/nar/gky1106
44. Kitts, P.A., Church, D.M., Thibaud-Nissen, F., Choi, J., Hem, V., Sapojnikov, V., Smith, R.G., Tatusova, T., Xiang, C., Zherikov, A., DiCuccio, M., Murphy, T.D., Pruitt, K.D., Kimchi, A.: Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Res **44**(D1), 73–80 (2016). doi:10.1093/nar/gkv1226

Additional Files

- **Additional File 1:** Supplemental Figures and Tables

Figures

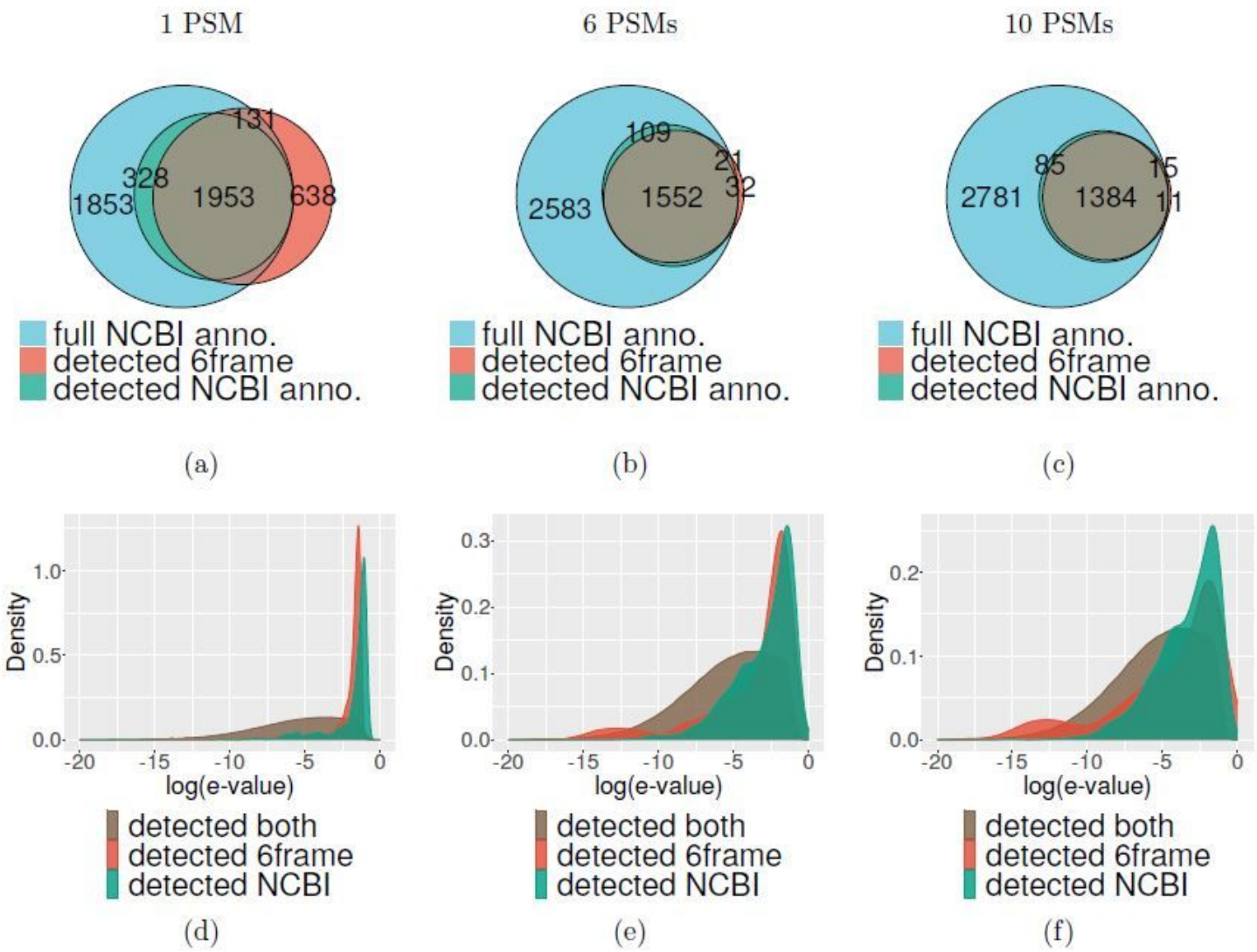


Figure 1

Due to technical limitations, the caption for figure 1 is only available in the manuscript.

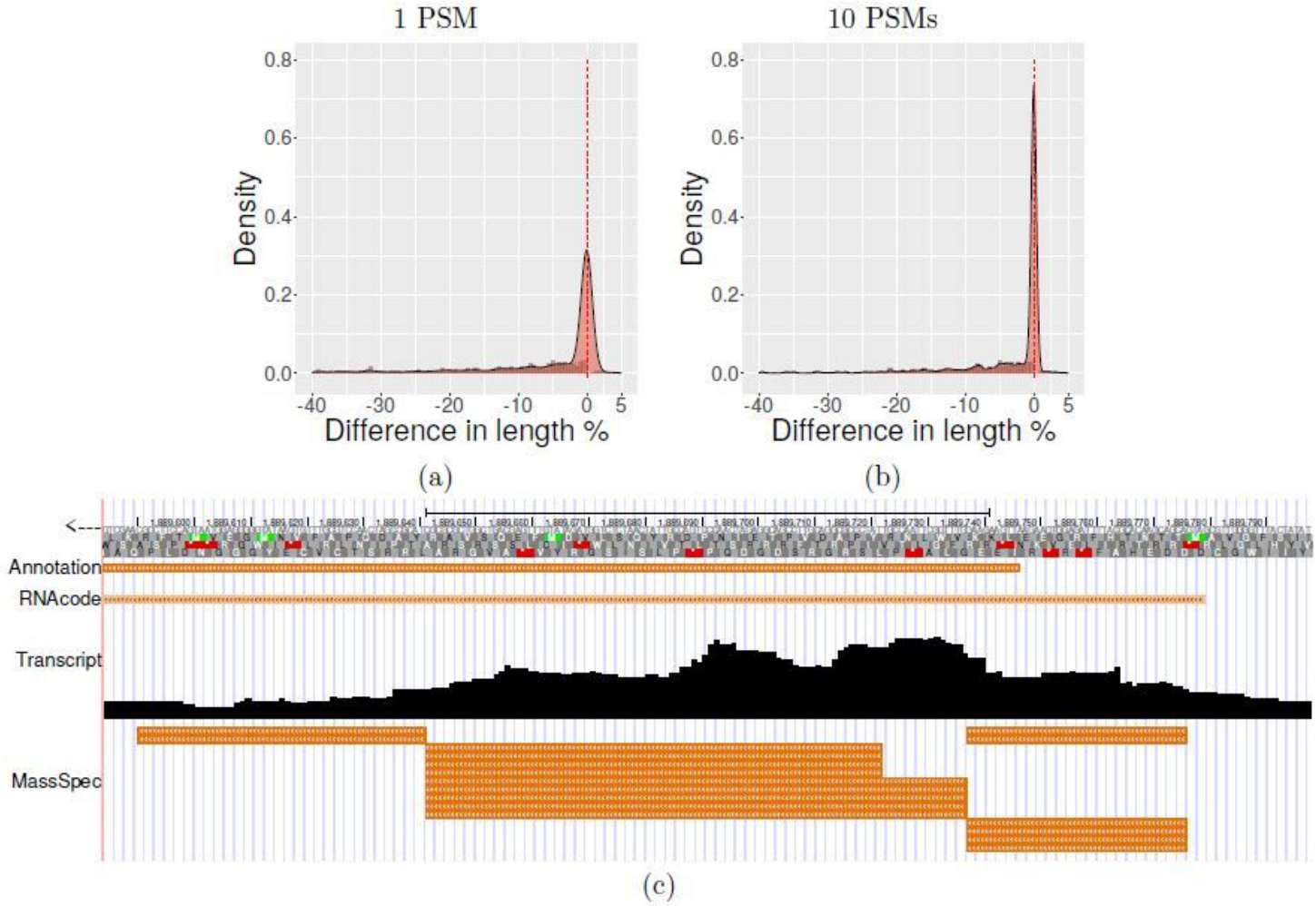


Figure 2

Due to technical limitations, the caption for figure 2 is only available in the manuscript.

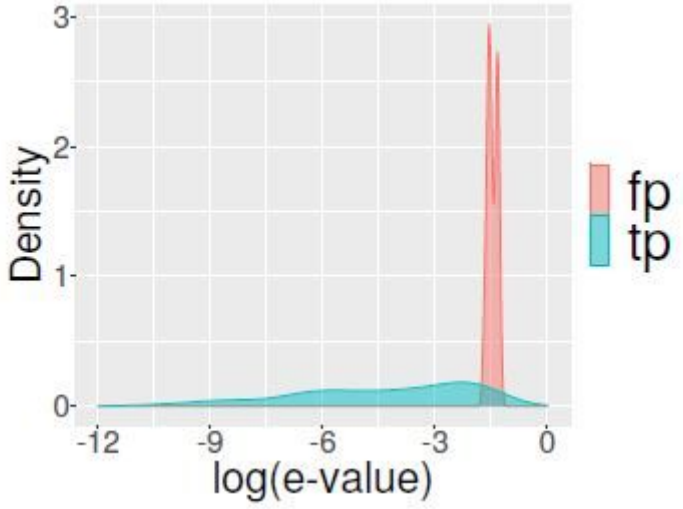


Figure 3

Due to technical limitations, the caption for figure 3 is only available in the manuscript.

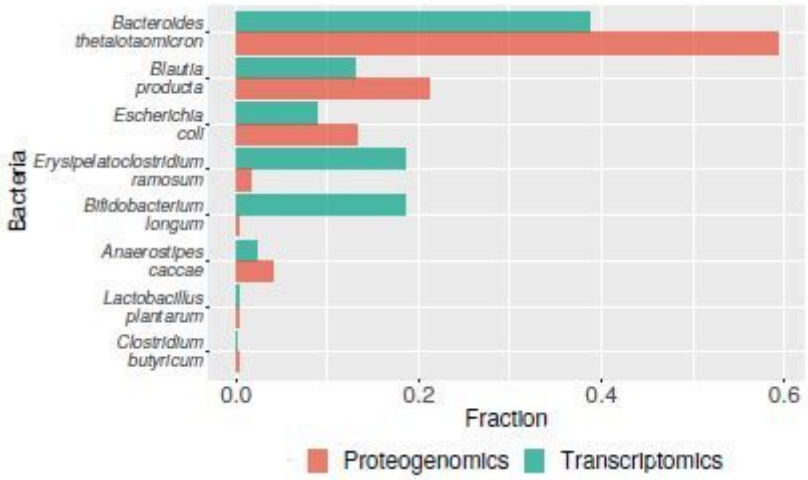


Figure 4

Due to technical limitations, the caption for figure 4 is only available in the manuscript.

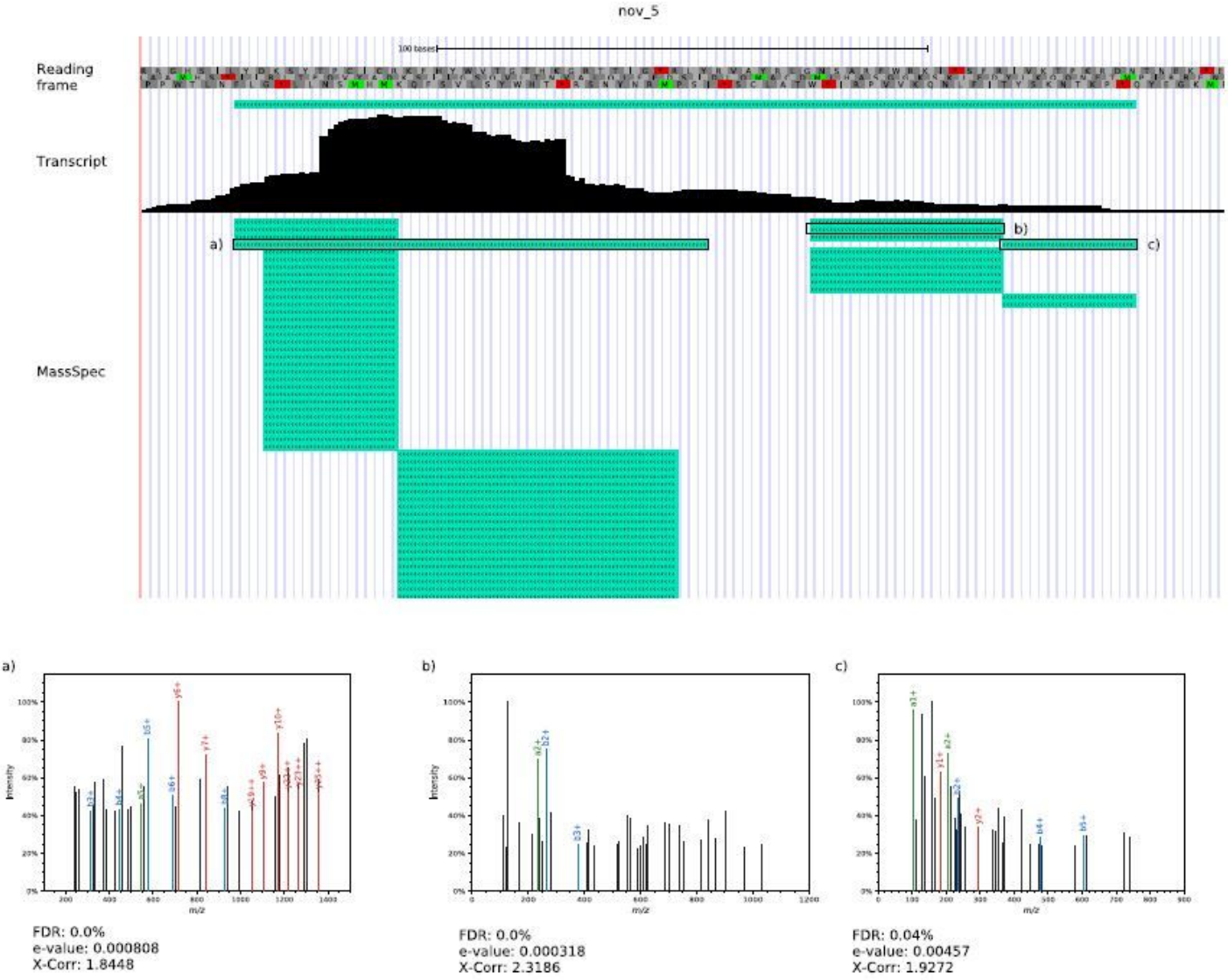


Figure 5

Due to technical limitations, the caption for figure 5 is only available in the manuscript.

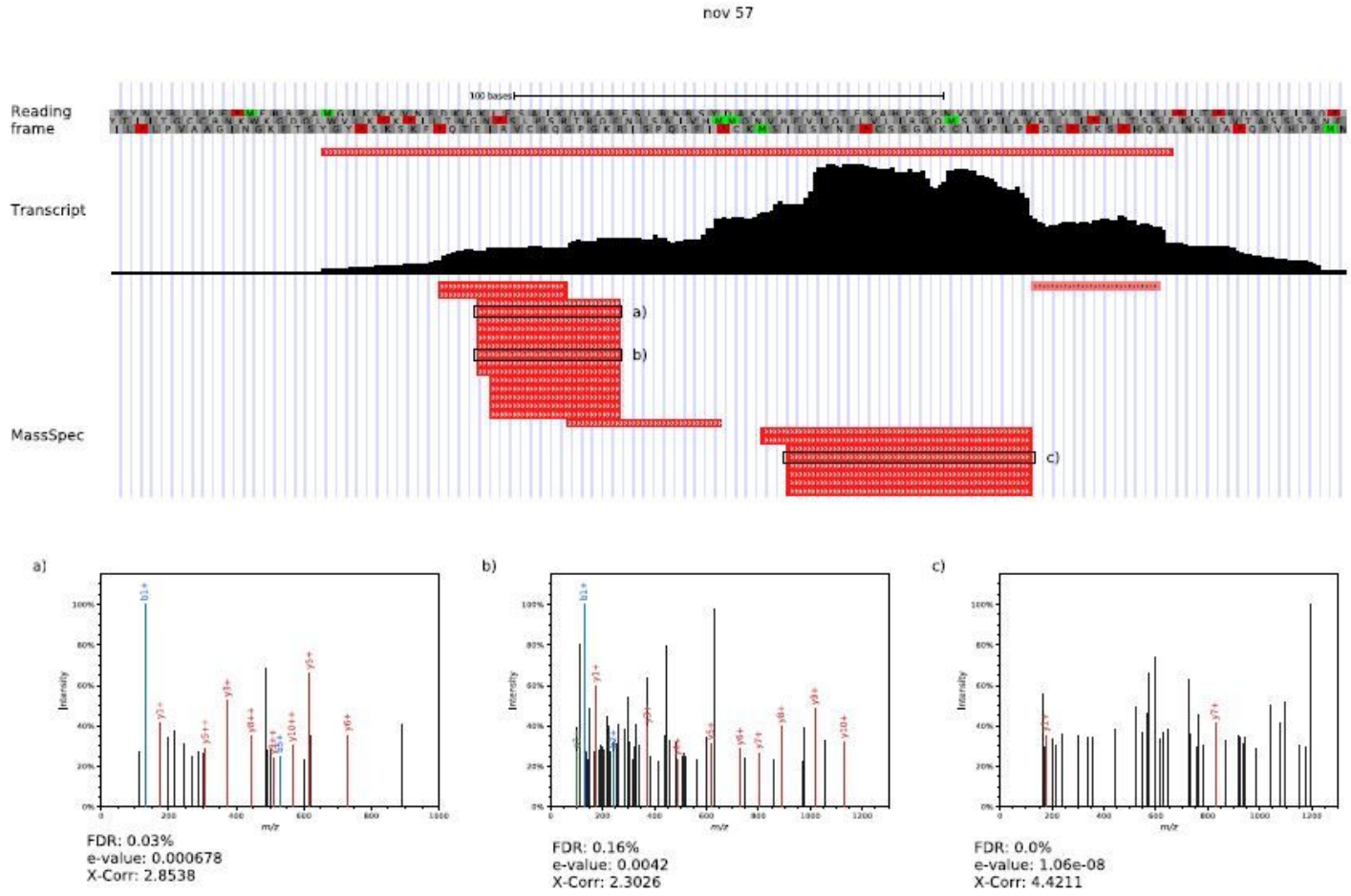


Figure 6

Due to technical limitations, the caption for figure 6 is only available in the manuscript.

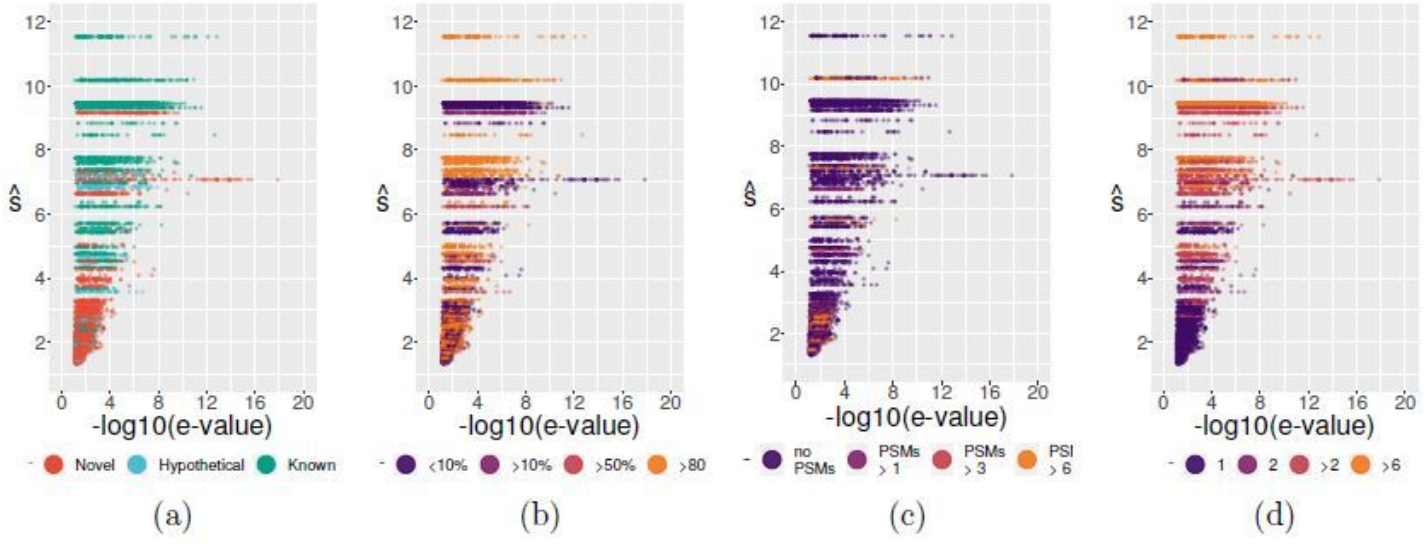


Figure 7

Due to technical limitations, the caption for figure 7 is only available in the manuscript.

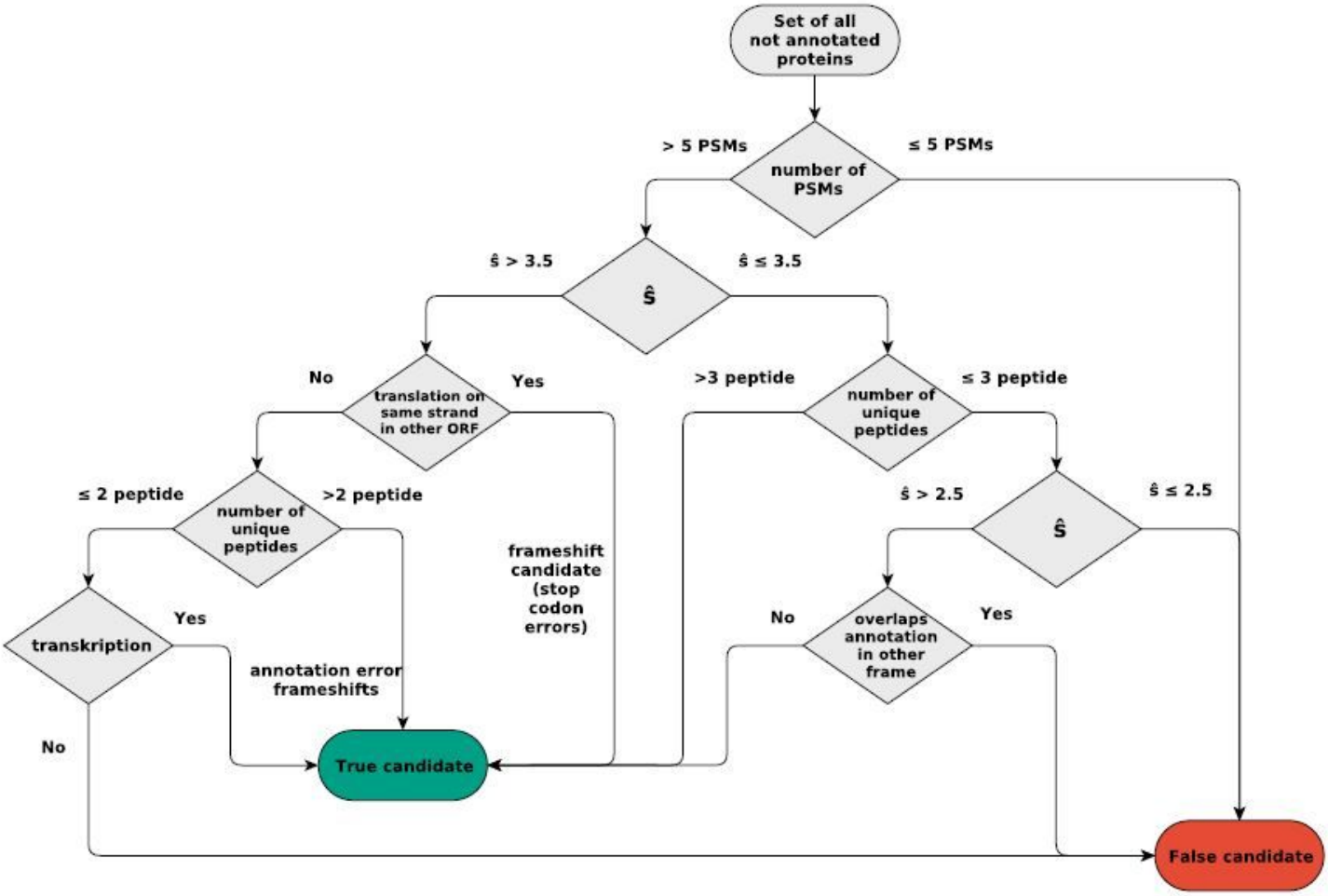


Figure 8

Due to technical limitations, the caption for figure 8 is only available in the manuscript.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- protmapsupplement.pdf