

# **Title: Rescuing discarded spectra: Full comprehensive analysis of a minimal proteome.**

**Running title:** Comprehensive analysis of a minimal proteome

**Authors:** Maria Lluch-Senar<sup>1,2,†,\*</sup>, Francesco M. Mancuso<sup>2,3,§,†</sup>, Héctor Climente-González<sup>1,2</sup>, Marcia I. Peña-Paz<sup>2,3</sup>, Eduard Sabido<sup>2,3</sup>, Luis Serrano<sup>1,2,4,\*</sup>

## **Affiliations:**

† These authors contributed equally to this work

<sup>1</sup> EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain

<sup>2</sup> Universitat Pompeu Fabra (UPF), Dr. Aiguader 88, 08003 Barcelona, Spain

<sup>3</sup> Proteomics Unit, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain

<sup>4</sup> Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain

§ Current address: Genomic Cancer Group, Vall d'Hebron Institute of Oncology (VHIO), Pg. Vall d'Hebron 119, 08035 Barcelona, Spain

\* to whom correspondence should be addressed: [maria.lluch@crg.es](mailto:maria.lluch@crg.es) and [luis.serrano@crg.es](mailto:luis.serrano@crg.es)

**Abbreviations:** ALC, average local confidence; PSM, peptide-spectrum match; WT, wild-type

**Keywords:** MS, PTM, Protein identification, Proteome coverage, Translational errors, Unassigned spectra

**Total number of words:** 5766

## ABSTRACT

A common problem encountered when performing large-scale MS proteome analysis is the loss of information due to the high percentage of unassigned spectra. To determine the causes behind this loss we have analyzed the proteome of one of the smallest living bacteria that can be grown axenically, *Mycoplasma pneumoniae* (729 ORFs). The proteome of *M. pneumoniae* cells, grown in defined media, was analyzed by MS. An initial search with both Mascot and a species-specific NCBIInr database with common contaminants (NCBImpn), resulted in around 79% of the acquired spectra not having an assignment. The percentage of non-assigned spectra was reduced to 27% after re-analysis of the data with the PEAKS software, thereby increasing the proteome coverage of *M. pneumoniae* from the initial 60% to over 76%. Nonetheless, 33 413 spectra with assigned amino acid sequences could not be mapped to any NCBIInr database protein sequence. Approximately, 1% of these unassigned peptides corresponded to PTMs and 4% to *M. pneumoniae* protein variants (deamidation and translation inaccuracies). The most abundant peptide sequence variants (Phe-Tyr and Ala-Ser) could be explained by alterations in the editing capacity of the corresponding tRNA synthases. About another 1% of the peptides not associated to any protein had repetitions of the same aromatic/hydrophobic amino acid at the N-terminus, or had Arg/Lys at the C-terminus. Thus, in a model system, we have maximized the number of assigned spectra to 73% (51 453 out of the 70 040 initial acquired spectra).

## **STATEMENT OF SIGNIFICANCE OF THE STUDY**

Proteomic analysis of *Mycoplasma pneumoniae*, one of the smallest bacteria with only 729 ORFs, enabled us to identify the causes behind the high percentage of unassigned spectra in MS. An initial database search left around 79% of the acquired spectra without an assignment. Re-analysis of the data with the PEAKS software and a large protein database, reduced this percentage to 27%, and thereby increased the *M. pneumoniae* coverage from 60% to over 76%. Approximately, 1% of the unassigned peptides corresponded to PTMs and 4% to protein variants. Thus, we have substantially increased the level of assigned spectra from 21% to 73%. Common alterations found in protein peptides could be explained by poor editing capacity of the corresponding tRNA synthases.

## 1. INTRODUCTION

During the last few years, shotgun proteomics has become the method of choice in biological studies that demand deep proteome coverage [1, 2]. Technological innovation has largely contributed to increasing the number of proteins identified by MS, with high-resolution instruments facilitating the correct detection of co-eluting peptides with similar masses, as well as the correct assignment of their charge states, thus resulting in a significant improvement of the number of assigned spectra. Despite the advances in the rate of protein and peptide identifications, the number of spectra assigned after MS/MS fragmentation remains low. For instance, for data-dependent LC-MS/MS, Michalski et al. [3] found that out of the over 100 000 likely peptide features eluted in single shotgun runs, only 16% were targeted for MS/MS fragmentation, of which less than 60% generated peptide identifications.

Several initiatives have emerged in order to increase the proteome coverage either by increasing the peptide fragmentation rate, i.e. the number of peptide features targeted for MS/MS analysis, or by reducing the number of unassigned spectra [4]. The availability of increasingly fast mass spectrometers, with current-generation instruments reaching scan rates of around 100 Hz, has significantly improved the fragmentation rate of peptide features. Similarly, several efforts have been made at improving sample preparation in order to enlarge the population of targeted peptide features. These efforts include refinements in protein isolation and extraction [5], new multidimensional separation and fractionation strategies for protein and peptide mixtures [6], as well as refined proteolytic digestion protocols [7], all of which have made it possible to achieve a comprehensive analysis of the human proteome [8]. However, despite the progress made in instrument acquisition methods and sample preparation techniques, these approaches often require high sample consumption and instrument time, thereby limiting the experimental throughput. As such, apart from increasing feature fragmentation rates, reducing the number of unidentified peptides and increasing the identification success rate within a mass spectrometric analysis is highly desirable.

Currently, only a relatively small fraction of spectra is eventually assigned to a peptide sequence in shotgun proteomic experiments, with an identification success rate laying around 60%. Different alternatives have been explored to reduce the number of unassigned spectra, including

attempts to exploit different fragmentation techniques [9], to improve the spectra quality [10, 11], and to combine different search engines [12], among others. Similarly, information related to protein splicing variants [13], amino acid variants (polymorphisms) [14], PTMs [15, 16] and chemical modifications due to secondary reactions that occur during sample preparation [17] have gradually been introduced into data analysis pipelines in order to increase the rate of identifiable peptides. Although these initiatives have improved the fraction of assigned spectra and identified peptides, the rate of assignment is still far from perfection.

To ascertain the importance of several of these factors, we analyzed the proteome of *Mycoplasma pneumoniae*, one of the smallest bacteria that can be grown in axenic culture and in defined medium. *M. pneumoniae*, with a 816 kbp genome, is the causative agent of walking pneumonia and other extra-pulmonary pathologies in humans [18]. The current annotation of its genome contains 662 ORFs larger than 100 amino acids, 67 smORFs (<100 amino acids), 311 ncRNAs, and 43 conventional RNAs (e.g. tRNAs, rRNAs, etc.) [19]. Its genome, transcriptome and proteome have been characterized in depth by different “omics” approaches [20-26], with several proteomic studies almost covering its entire proteome (75%) [27]. PTM analysis revealed phosphorylation of Ser, Thr, and Tyr residues in addition to Lys acetylation [26]. In our group, MS analysis (using the Mascot software) of trypsin-digested, non-fractionated *M. pneumoniae* cell extract, grown in rich medium [28], detected 413 different proteins. This represents 60% of all predicted ORFs [24].

Despite the proteomic simplicity of *M. pneumoniae*, only 21% of the 70 040 spectra could be assigned to proteins after Mascot software analysis and a species-specific NCBI non-redundant database search (NCBIInr). Upon re-analyzing the data using the PEAKS software [29], *de novo* assembly on the spectra and an entire NCBIInr database search, we were able to increase the spectra assignment up to 73.5% in the samples derived from cells grown in defined medium. In the latter analysis, half of the peptide-spectrum matches (PSMs) mapped to *Mycoplasma* and possible common contaminants, 1% to PTMs, and 4% to protein variants (deamidation and translation inaccuracies). Further analyses suggested that, aside from spontaneous Asn to Asp and Gln to Glu deamidations, some of the observed amino acid changes could be the result of translational errors, possibly due to mischarging the tRNA synthases. In fact the two most

common sequence variants found (Phe into Tyr, and Ala into Ser) could be due to the altered editing capacity of the corresponding tRNA synthases [30]. For the cases in which we could detect both the wild-type (WT) and altered peptides with good quality spectra (Phe into Tyr and Ala into Ser changes), we estimated a translational error rate of approximately 0.06 and 0.08, respectively (average ratio of the areas of peptide variants to WT). Despite these efforts, 26.5% of good quality spectra remained unassigned, which suggests the presence of other causes that impede spectra assignment, including peptide co-isolation, the presence of uncharacterized biomolecules, and unknown chemical reactions that still need to be identified.

## 2. MATERIAL AND METHODS

### *Mycoplasma pneumoniae* growth conditions

*M. pneumoniae* was grown in T-75 cm<sup>2</sup> tissue culture flasks with 50 mL of modified Hayflick medium for 96 h at 37°C as previously described [26]. Then they were washed with defined medium [26] and incubated in defined medium for 6 h at 37°C.

### *Sample preparation for LC-MS/MS*

Total protein lysates from *M. pneumoniae* were obtained by breaking the cells with 200 µl of lysis buffer (4% SDS, 0.1M DTT and 0.1M Hepes). Total protein extracts of two biological replicates were digested with trypsin and subsequently analyzed by MS. Briefly, samples were dissolved in 6 M urea, reduced with DTT (10 mM at 37°C, 60 min), and alkylated with iodoacetamide (20 mM at 25°C, 30 min). Samples were diluted 10-fold with 0.2 M NH<sub>4</sub>HCO<sub>3</sub> before being digested at 37°C overnight with trypsin (with a protein:enzyme ratio of 10:1). Peptides generated in the digestion were desalted, evaporated to dryness, and dissolved in 300 µl of 0.1% formic acid.

An aliquot of 2.5 µl of each fraction (amounts ranging from 0.17 to 4 µg) was run on an LTQ-Orbitrap Velos (ThermoFisher) fitted with a nanospray source (ThermoFisher) after a nanoLC separation in an EasyLC system (Proxeon). Peptides were separated in a reverse phase column, 75 µm x 150 mm (Nikkyo Technos Co., Ltd.) with a gradient of 5 to 35% acetonitrile in 0.1% formic acid for 60 min at a flow rate of 0.3 mL/min. The Orbitrap Velos was operated in positive ion mode with the nanospray voltage set at 2.2 kV and source temperature at 325°C. The instrument was externally calibrated using Ultramark 1621 for the FT mass analyzer and the background polysiloxane ion signal at m/z 445.120025 was used as lock mass. The instrument was operated in data-dependent acquisition (DDA) mode and full-MS scans were acquired in all experiments over a mass range of m/z 350-2000, with detection in the Orbitrap mass analyzer set at a resolution setting of 60 000. Fragment ion spectra produced via collision induced dissociation (CID) were acquired in the ion trap mass analyzer. In each cycle of data-dependent analysis, the top 20 most intense ions with multiple charges above a threshold ion count of 5000 were selected for fragmentation at a normalized collision energy of 35% following each survey scan. All data were acquired with Xcalibur 2.1 software. Total extract (20 µg) was also digested

and desalted, and 1  $\mu\text{g}$  of the resulting peptides were analyzed on an Orbitrap Velos Pro in the same conditions as the fractions but with a longer gradient (120 min).

The MS proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD002779 [31].

### *Proteomics data analysis*

Protein identification was performed by Proteome Discoverer software v.1.3 (ThermoFisher) using the Mascot search engine (v2.4.01, Matrix Science) [32]. MS/MS spectra were searched against a species-specific database containing all *M. pneumoniae* proteins of NCBIInr (7996 entries), and a list of the common contaminants (599 entries), NCBImpn. We set a precursor ion mass tolerance of 7 ppm (parts per million) at the MS1 level and a fragment ion mass tolerance of 0.5 Da. Up to three missed cleavages for trypsin digestion were allowed. Oxidation of methionine and protein acetylation at the N-terminus were defined as variable modifications, whereas carbamidomethylation on cysteines was set as a fixed modification. False discovery rates (FDR) in peptide identification were evaluated using a decoy database set to a maximum of 5%.

To increase the percentage of assigned spectra, peptide *de novo* sequencing coupled with a database search was used as implemented in the PEAKS software suite (v7) [29]. After importing the raw files, bad quality spectra were filtered out using a quality cutoff of 0.05. Search parameters were the same as for the previous Mascot search but with the entire NCBIInr database (January 2014 version). The decoy-fusion database method was used to calculate the FDR, and matched spectra were filtered for 5% FDRs, and only proteins with at least one unique peptide were considered. A *de novo* average local confidence (ALC) score of 50 was used as the cutoff for *de novo* peptides. ALC is the average of the total local confidence divided by the number of amino acids in the peptide sequence. Total local confidence is the sum of the local confidence scores from each amino acid in the peptide sequence. The PEAKS PTM [33] search module was used to identify peptides with unspecified modifications and translation inaccuracies; the modifications set are listed in Table 1. Finally, the SPIDER module [33] was used for finding novel peptides that are homologous to peptides in a given protein database. The



spectra associated to peptides after SPIDER analysis but not assigned to any protein in the NCBI database search were analyzed by BLAST [34, 35]. Only peptides showing no gaps and an identity percentage higher than 87% in the alignment were assigned to the identified protein. When one peptide was assigned with these parameters to more than one protein, it was discarded for the protein identification, so that all the studied peptides were unique.

#### *RNA extractions and sample preparations*

After a 6 h exponential growth at 37°C, *M. pneumoniae* strain M129 cells were washed twice with PBS and lysated with 700 µl of Qiazol buffer. RNA was extracted using the miRNeasy mini kit (Qiagen) following the manufacturer's instructions. Libraries for RNA-seq were prepared following directional RNA-seq library preparation and sequencing. Briefly, 1 µg of total RNA was fragmented to around 100-150 nt using NEB Next Magnesium RNA Fragmentation Module (ref. E6150S, NEB). RNA was treated with Antarctic phosphatase (ref. M0289S, NEB) and PNK (ref. M0201S, NEB) to make the 5 and 3 ends of the RNA available for adapter ligation. Samples were further processed using the TruSeq small RNA Sample Prep Kit (ref. RS-200-0012, Illumina) according to the manufacturer's protocol. In summary, 3' adapters and subsequently 5' adapters were ligated to the RNA. cDNA was synthesized using reverse transcriptase (SuperScript II, ref. 18064-014, Invitrogen) and a specific primer (RNA RT Primer) complementary to the 3' RNA adapter. cDNA was further amplified by PCR using indexed adapters supplied in the kit. Finally, size selection of the libraries was performed using 6% Novex® TBE Gels (ref. EC6265BOX, Life Technologies). Fragments with insert sizes of 100 to 130 bp were cut out from the gel, and cDNA precipitated and eluted in 10 µl of elution buffer. Double-stranded templates were cluster amplified and sequenced on an Illumina HiSeq 2000. The raw data of RNAseq was submitted to the ArrayExpress database (<http://www.ebi.ac.uk/arrayexpress>) and assigned the identifier E-MTAB-3076 [19].

#### *Read mapping to the genome and identification of variants*

Raw reads were mapped to the *M. pneumoniae* M129 complete genome (GenBank accession number NC\_000912.1), September 2013 version, using Bowtie 2. The parameters used corresponded to the default option E-MTAB-3076 E-MTAB-300 allowing a mismatch in the

seed (i.e. D 20, R 3 -N 1 -L 20 -i S,1,0.50). The identification of nucleotide variants was performed with SAMTools and BCFTools 0.1.19.

#### *Analysis of peptide and RNA mutations*

Mutations in both RNA and peptides were analyzed through an in-house script (publicly available at <https://switchlab.svn.cloudforge.com/sequencing>). All the protein variants identified by SPIDER and the RNA variants found by SAMTools were compared to the reference genome, as well as to a database with all the possible ORFs longer than 30 amino acids. The most abundant amino acid changes were determined after normalizing the change count by multiplying by the frequency of the WT amino acid in the proteome of *M. pneumoniae*.

#### *Analyzing codon enrichment in mutations*

We analyzed if any peptide alteration happened preferably on a particular codon. We searched this enrichment using Fisher's exact test. Accordingly, for each codon and alteration, every codon in the genome encoding the wild type amino acid were classified in four groups: (1) involved in an alteration, and the codon is the one studied; (2) involved in an alteration, and the any other codon of the wild type amino acid; (3) not involved in an alteration, and the codon is the studied one; and (4) not involved in an alteration, and the codon is not the studied one.

### 3. RESULTS

#### *Coverage in a minimal cell*

Two biological replicates of a non-fractionated extract of *M. pneumoniae*, grown in defined medium [22] to the exponential phase of growth, were analyzed by MS after trypsin digestion. Analysis of the MS/MS data by Mascot software [28] using the *M. pneumoniae* NCBImpn database with common MS contaminants [19] (see Methods) retrieved 4687 identified peptides (from 14 481 spectra), from which 4267 were unique peptides corresponding to 485 Mycoplasma proteins and 51 contaminant proteins (70 040 input spectra) (Figure 1A). In this initial data interpretation process, only ~21% of spectra were assigned to a peptide sequence (14 481 PSMs).

In order to increase the percentage of assigned spectra, we performed subsequent analyses on the samples. A *de novo* sequencing approach using the PEAKS 7 software suite [29] was considered as an alternative analytical process to assign spectra. By combining database searches, *de novo* sequencing as implemented in PEAKS, and using the NCBIInr database, 51 230 spectra out of 70 040 were assigned to peptide sequences (73%; Figures 1B and S1) with a FDR of 5%. However, only 35% of these (17 817) were matched to sequences present in the NCBIInr database (Supporting Information Table S1; Figures 1B and S1), and of those, 14 239 PSMs were associated to 496 *M. pneumoniae* proteins (72% of the proteome). 33 413 *de novo* good-quality spectra were assigned to peptide sequences but not to any known protein in the NCBIInr database. The identification of these peptide sequences could originate either from false assignments due to FDR underestimation [36], or from not considering certain peptide sequence variants during the database search (e.g. PTMs, genome mutations, transcriptional and translational errors, and new ORFs not yet been annotated in the *M. pneumoniae* genome; Table S2).

#### *Post-translational modifications represent 1% of the assigned spectra*

*M. pneumoniae* proteins are known to be phosphorylated at Ser, Thr, and Tyr, and acetylated at Lys residues [26]. Besides these modifications, the genome of *M. pneumoniae* encodes for protein methyltransferases (MPN362 and MPN576) [37, 38], for a tmRNA that can incorporate peptides at the C-termini of incomplete peptides in stalled ribosomes [39], and for a toxin that could ADP-rybosilate proteins (MPN348a) [40]. Additionally, the presence of glycosyltransferases (MPN483, MPN028 and MPN075) involved in the transfer of glycosyl

groups suggests that glycosylation is also possible. Thus, we decided to study the number of spectra that could be re-assigned to proteins when common PTMs (Table 1) were included in the search by using the PEAKS PTM module [33] (Supporting Information Table S3 and S4).

In this new search, a total of 51 314 spectra were associated to peptide sequences (Figure 1B and S1A) with 18 513 PSM matching to known proteins in the NCBI database (Supporting Information Table S3) and 32 801 identified by a *de novo-only* strategy (Supporting Information Table S4; Figure 1B and S1B). Around 79.5% (14 723 out of 18 513) of the spectra with peptides that matched the NCBI database were from *M. pneumoniae* and corresponded to 497 proteins (72% proteome coverage). Thus, upon considering PEAKS PTMs, 1% of total spectra were recovered and an additional protein from *M. pneumoniae* was identified. Out of the 14 716 spectra assigned to *M. pneumoniae* proteins, 25% (3736 of which 1066 unique peptides) corresponded to putative PTMs: 0.08% methylated, 0.2% di-methylated, 2% acetylated and 0.03% phosphorylated (Supporting Information Table S3; Figure 2A). Only 17 of the previously experimentally detected acetylations (after enrichment) [26], but none of the phosphorylated peptides, were found in this analysis (Supporting Information Table S5), reflecting that experimental enrichment is required to detect these types of PTMs.

#### *Genome and transcriptome mutations*

Spontaneous mutations in the genome sequence could generate variability in the amino acid sequence of a protein and explain some of the unassigned peptides. Genome sequencing of the strain used in this study revealed 93 genomic mutations in 70 different ORFs [20]. However, no peptide bearing these mutations was identified by MS.

Another possible source of amino acid alterations is transcriptional errors. To determine the possible impact transcriptional errors have on non-assigned spectra, we studied the frequency of base changes in *M. pneumoniae* RNA-seq data (See Methods and Supporting Information Table S6). After filtering for artifact chimeric RNAs [41] and mapping the reads against the reference genome, we found 252 mutations in the RNA, 249 of which were assigned to putative ORFs (Supporting Information Table S6). However, no mutation was matched to any aa alteration. This suggests that detected mutations are sequencing artifacts, or are too low in abundance to be detected by MS.

### *Translational errors*

Another source of unassigned spectra could be a consequence of errors in protein translation that alter the canonical peptide sequences. Using the SPIDER module [33], the input spectra (70 040 cases) were searched again and both amino acid substitutions at single positions as well as PTMs were allowed (see Table 1 for the ones considered; Table S8). 51 453 spectra were assigned to peptide sequences (73%; Figure 1B and S1A), of which 20 734 spectra corresponded to peptide sequences found in the NCBI nr database (Figure 1B and S1B). These 20 734 PSM corresponded to 5 837 unique peptides, and the majority of them (80.6 %) could be assigned to *M. pneumoniae* proteins (Supporting Information Table S7). Interestingly, these spectra correspond to 270 aa sequence alterations. Thus, taking into consideration single amino acid substitutions and PTMs, the proteome coverage increased from 72 to 72.5%.

We found a positive correlation between protein abundance in *M. pneumoniae* [24] and the number of peptides with single alterations identified [24]. In agreement with this, we find proportionally more mutant peptides from proteins with a higher peptide coverage (Figure S2). These results suggest that even though almost every protein likely has a similar proportion of peptides with single amino acid variants, their detection is highly dependent on the abundance of the corresponding protein (correlation coefficient=0.6; Figure 2B). Also, study of functionality of those proteins functional study by enrichment in COG categories (using Fisher's test, Table S9), revealed that proteins with more aa alterations are involved in metabolism, already described as most abundant [24].

We find mainly four amino acid changes (Figure 2C), of which two, Asn to Asp and Gln to Glu correspond to deamidations. To estimate the error rate we looked at the areas of those peptides where we could identify both the WT and the sequence altered peptides considering only those peptides that both the WT and aa variant have good quality criteria ( $\Delta RT < 0.5$ ; [42]; Tables 2 and S11). For the Asn to Asp and Gln to Glu deamidations the error rate was estimated in 0.95 and 1.4; respectively (Table S11). Also, we were able to obtain statistically significant data for the Phe to Tyr and Ala to Ser alterations, showing error rates of 0.06 and 0.08, respectively (Table 2). A simple calculation considering the total number of Phe and Ala in the proteome (13

377 and 15965), the average protein length (300 aa) and the detected error rates by MS (0.06 and 0.08), suggests that each protein has a 0.4 probability of having at least a Phe residue changed into a Tyr and a probability of 0.3 for Ala to Ser aa change.

There are three possible explanations for the observed amino acid changes:

i) Asn to Asp and Gln to Glu alterations can be explained by errors of the glutamyl-tRNA amidotransferase that amidates Asp and Glu loaded onto Asn and Gln-tRNA, respectively, as happens in Gram+ bacteria and as previously described for the metabolism of *M. pneumoniae* [43]. In addition to the above, spontaneous chemical deamidation reactions could happen *in vivo* in long-life proteins [44], and/or during sample preparation [45]. Deamidation rates depend on pH, temperature, ionic strength, buffer, and the amino acid residues of the peptide chain that are in close proximity to the Asn or Gln (at neutral pH and 37°C deamidation half-lives have been reported to be in the range of 1–500 days for Asn and 100–5000 days for Gln). This difference in Asn and Gln deamidation and the fact that Mycoplasma proteins have long protein half-lives (longer than 20 hours) [24, 45] could explain the large number of Asn to Asp cases yet lower number of Gln to Glu ones. This suggests that part of the deamidation of Asn and Gln could take place in the cell because of spontaneous chemical reactions. However, we cannot rule that errors introduced by the glutamyl-tRNA amidotransferase and deamidation during sample preparation could also be partly responsible.

ii) Incorrect recognition of the codon by the corresponding tRNA (wobble base theory) could come to mind as another possible cause. If this were the case we would expect to see mutations corresponding to amino acids whose recognition codon differs at the third base. We examined the frequency of codon usage at the position bearing the observed error to determine if the wobble base theory could be applicable in these other minority cases (Supporting Information Table S10). In principle, if codon usage displays equal probabilities for a particular alteration, we could conclude that incorrect tRNA loading is mainly responsible for the amino acid change. In general, we did not find a specific codon bias for any of the alterations (Table S10).

iii) Neither a simple chemical reaction or the wobble base theory is capable of explaining the other most frequently observed amino acid substitutions (Ser for Ala and Phe for Tyr). Aminoacyl-tRNA synthetases could be responsible for such translational errors, because of having both tRNA mischarging activity and poor editing rates. In this respect, it has been found that the Phe-tRNA synthase of *M. mobile* has poor editing incorrectly charging Tyr instead of

Phe [30]. A detailed sequence comparison of Mycoplasma tRNA synthases with those of other bacteria revealed in many cases deletions or mutations within their editing domains [46]. This is especially true for the Ala-tRNA synthetase that has a degenerated C-Ala domain at its C-terminus (Figure S3). Mutations at the C-Ala domain increase Ala-Ser miss-incorporation by 1000 times [47].

Therefore, overall our data supports the idea that these alterations could be associated with the poor editing of some tRNA-synthetases.

#### *Putative new proteins and non-tryptic digested peptides*

Finally, after considering PTMs and amino acid substitution by PEAKS there were still 30 719 spectra with peptide sequences assigned by the *de novo*-only approach that were not associated to any known protein in the NCBI database. Therefore, we did a BLAST alignment [34, 35] of the peptide sequences assigned to those 30 719 spectra against a *M. pneumoniae*-specific database (HomoConTrans19) [48]. This database comprises all putative proteins longer than 19 amino acids encoded by six open reading frames of the *M. pneumoniae* genome. By using this database, we were able to study the new putative translated products. About 34% of the considered spectra (10 502) showed >87.5% identity and 0 gaps in the BLAST results, and could be associated with proteolytic products, or non-tryptic digested peptides. Interestingly, the 10 502 spectra corresponded to 272 *M. pneumoniae* proteins, 224 of which were identified in the previous search and 48 that had not been previously identified [24] (with 27 corresponding to annotated ORFs by bioinformatics predictions, and 21 to putative new proteins). Nine out of the 21 putative new ORFs were supported by RNAseq reads, and thus the existence of the mRNA confirmed (Supporting Information Table S12).

Nonetheless 20 217 assigned spectra remained unassociated to any protein in our database, with 1% of them showing repetitions of the same amino acid at the beginning or at the end of the sequence. Repetitions at the N-terminus were comprised of aromatic or hydrophobic amino acids, whilst the ones located at C-terminus were either Arg or Lys (Supporting Information Figure S4). Since PEAKS only considered tryptic peptides, all peptides should end in an Arg or Lys. However, the amino acid repetitions found in these analyses suggested that some unknown reaction was happening at the termini of these peptides. Sequence analysis by BLAST, after

removing the repetitive sequences, revealed that 50 of these peptides with repetitions of Trp, Ala, or Met at the N-termini, could be associated to proteins (with an identity of >87,5% and 0 gaps; Supporting Information Table S13). This suggests that these peptides might indeed be present, but could be covalently bound to an uncharacterized molecule.

#### 4. DISCUSSION

Increasing the assignment of spectra in proteomic analysis is desirable for improving both protein identification and protein quantitation. Here, we assessed the impact of different factors in a minimal bacterium, *M. pneumoniae*. We selected this system because of its reduced number of proteins and scarce PTMs, enabling a better evaluation of the impact of different factors in spectra assignment. By using a *de novo* approach associated with database searches (PEAKS) and considering PTMs and amino acid substitutions (PEAKS PTM and SPIDER) we achieved an assignment of 73% of the total spectra to peptides. Despite our comprehensive approach, the simplicity of our model organism and the use of a defined media, 26% of good quality spectra remained unassigned and 40% of the assigned spectra did not map to any annotated protein.

Our analysis showed that the rate of detected peptide amino acid alterations is around 10 times higher than in *E. coli* [49] (0.06 for Phe to Try and 0.08 from Ala to Ser [50]). The observed amino acid deamidation is most likely a consequence of the long half-life of *M. pneumoniae* proteins [24], and not because of errors made by the glutamyl-tRNA amidotransferase, or sample preparation, since we see many more changes for Asn than for Gln. The fact that the Asn to Asp change mainly happens in proteins in the membrane could be the result of a higher exposure of these sequences to free radicals from the medium or by the peroxide secreted by this bacterium [36, 51, 52]. About the other alterations we observed, analysis of the codon usage of the mutations discriminates against tRNA codon recognition errors and favors errors at the level of the tRNA loading and editing by the tRNA synthases. In fact, it has been shown that the specificity for Phe over Tyr of the Phe-tRNA synthetase in *M. mobile* is 14-fold lower than that of between Phe and Tyr [30]. Also sequence comparison of *M. pneumoniae* tRNA synthases shows that for many of these enzymes the editing domain is compromised [53]. In fact, sequence comparison with the Ala-tRNA synthases of *E. coli* and *B. subtilis* shows a degeneration of the C-Ala domain which could increase Ala into Ser miss-incorporation a 1000 times [47]. Larger



translational errors in *M. pneumoniae* could be more easily tolerated than in *E. coli* due to its slow division rate (of ~8-20 h).

In conclusion, we showed that combining a *de novo* and database search approach together with consideration of PTMs and amino acid substitutions, increases the percentage of assigned spectra. However, despite the use of a small bacterium with a simple genome, which was grown in a defined medium with no contaminant proteins, there is still a considerable fraction of good quality spectra that could not be mapped to neither the bacterial proteome or to possible contaminants.

## **ACKNOWLEDGMENTS**

We thank Dr. Javier Delgado for fruitful discussions and the Genomics Core Facility at CRG. This work was supported by the European Research Council (ERC), the Fundación Marcelino Botin, the Spanish Ministerio de Economía y Competitividad BIO2007-61762 and the ISCIII (PI10/01702). The CRG/UPF Proteomics Unit is part of the “Plataforma de Recursos Biomoleculares y Bioinformáticos (ProteoRed)” supported by grant PT13/0001 of Instituto de Salud Carlos III (ISCIII). We acknowledge the support of the Spanish Ministry of Economy and Competitiveness, ‘Centro de Excelencia Severo Ochoa 2013-2017’, SEV-2012-0208. This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 634942.

## **CONFLICT OF INTEREST**

The authors declare no competing financial interests.

## 5. REFERENCES

- [1] Aebersold, R., Mann, M., Mass spectrometry-based proteomics. *Nature* 2003, 422, 198-207.
- [2] Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C., Yates, J. R., 3rd, Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews* 2013, 113, 2343-2394.
- [3] Michalski, A., Cox, J., Mann, M., More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *Journal of proteome research* 2011, 10, 1785-1793.
- [4] Mancuso FM, C. S., Sabido E., *Bioinformatic Approaches to Increase Proteome Coverage*, 2014.
- [5] Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N., Mann, M., Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nature methods* 2014, 11, 319-324.
- [6] Di Palma, S., Hennrich, M. L., Heck, A. J., Mohammed, S., Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis. *Journal of proteomics* 2012, 75, 3791-3813.
- [7] Glatter, T., Ludwig, C., Ahrne, E., Aebersold, R., *et al.*, Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion. *Journal of proteome research* 2012, 11, 5145-5156.
- [8] Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., *et al.*, Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* 2011, 7, 548.
- [9] Frese, C. K., Altelaar, A. F., Hennrich, M. L., Nolting, D., *et al.*, Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. *Journal of proteome research* 2011, 10, 2377-2388.
- [10] Gentzel, M., Kocher, T., Ponnusamy, S., Wilm, M., Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics* 2003, 3, 1597-1610.
- [11] Na, S., Paek, E., Quality assessment of tandem mass spectra based on cumulative intensity normalization. *Journal of proteome research* 2006, 5, 3241-3248.
- [12] Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L., Deutsch, E. W., Combining results of multiple search engines in proteomics. *Molecular & cellular proteomics : MCP* 2013, 12, 2383-2393.
- [13] Omenn, G. S., Menon, R., Zhang, Y., Innovations in proteomic profiling of cancers: alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology. *Journal of proteomics* 2013, 90, 28-37.
- [14] Su, Z. D., Sun, L., Yu, D. X., Li, R. X., *et al.*, Quantitative detection of single amino acid polymorphisms by targeted proteomics. *Journal of molecular cell biology* 2011, 3, 309-315.
- [15] Mann, M., Jensen, O. N., Proteomic analysis of post-translational modifications. *Nature biotechnology* 2003, 21, 255-261.
- [16] Nielsen, M. L., Savitski, M. M., Zubarev, R. A., Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Molecular & cellular proteomics : MCP* 2006, 5, 2384-2391.
- [17] Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., *et al.*, A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature biotechnology* 2015, 33, 743-749.

- [18] Chiner, E., Signes-Costa, J., Andreu, A. L., Andreu, L., [Mycoplasma pneumoniae pneumonia: and uncommon cause of adult respiratory distress syndrome]. *An Med Interna* 2003, 20, 597-598.
- [19] Lluch-Senar, M., Delgado, J., Chen, W. H., Llorens-Rico, V., *et al.*, Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Molecular systems biology* 2015, 11, 780.
- [20] Guell, M., van Noort, V., Yus, E., Chen, W. H., *et al.*, Transcriptome complexity in a genome-reduced bacterium. *Science* 2009, 326, 1268-1271.
- [21] Guell, M., Yus, E., Lluch-Senar, M., Serrano, L., Bacterial transcriptomics: what is beyond the RNA horizon? *Nature reviews. Microbiology* 2011, 9, 658-669.
- [22] Kuhner, S., van Noort, V., Betts, M. J., Leo-Macias, A., *et al.*, Proteome organization in a genome-reduced bacterium. *Science* 2009, 326, 1235-1240.
- [23] Lluch-Senar, M., Luong, K., Llorens-Rico, V., Delgado, J., *et al.*, Comprehensive methylome characterization of Mycoplasma genitalium and Mycoplasma pneumoniae at single-base resolution. *PLoS genetics* 2013, 9, e1003191.
- [24] Maier, T., Schmidt, A., Guell, M., Kuhner, S., *et al.*, Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Molecular systems biology* 2011, 7, 511.
- [25] Schmidl, S. R., Gronau, K., Pietack, N., Hecker, M., *et al.*, The phosphoproteome of the minimal bacterium Mycoplasma pneumoniae: analysis of the complete known Ser/Thr kinome suggests the existence of novel kinases. *Molecular & cellular proteomics : MCP* 2010, 9, 1228-1242.
- [26] van Noort, V., Seebacher, J., Bader, S., Mohammed, S., *et al.*, Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. *Molecular systems biology* 2012, 8, 571.
- [27] Jaffe, J. D., Berg, H. C., Church, G. M., Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* 2004, 4, 59-77.
- [28] Koenig, T., Menze, B. H., Kirchner, M., Monigatti, F., *et al.*, Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. *Journal of proteome research* 2008, 7, 3708-3717.
- [29] Ma, B., Zhang, K., Hendrie, C., Liang, C., *et al.*, PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM* 2003, 17, 2337-2342.
- [30] Yadavalli, S. S., Ibba, M., Selection of tRNA charging quality control mechanisms that increase mistranslation of the genetic code. *Nucleic acids research* 2013, 41, 1104-1112.
- [31] Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A., *et al.*, ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology* 2014, 32, 223-226.
- [32] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, 20, 3551-3567.
- [33] Han, X., He, L., Xin, L., Shan, B., Ma, B., PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications. *Journal of proteome research* 2011, 10, 2930-2936.
- [34] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 1997, 25, 3389-3402.

- [35] Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., *et al.*, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic acids research* 2001, *29*, 2994-3005.
- [36] Nesvizhskii, A. I., Proteogenomics: concepts, applications and computational strategies. *Nature methods* 2014, *11*, 1114-1125.
- [37] Plamann, M. D., Stauffer, G. V., Characterization of the Escherichia coli gene for serine hydroxymethyltransferase. *Gene* 1983, *22*, 9-18.
- [38] Yang, Z., Shipman, L., Zhang, M., Anton, B. P., *et al.*, Structural characterization and comparative phylogenetic analysis of Escherichia coli HemK, a protein (N5)-glutamine methyltransferase. *Journal of molecular biology* 2004, *340*, 695-706.
- [39] Himeno, H., Nameki, N., Kurita, D., Muto, A., Abo, T., Ribosome rescue systems in bacteria. *Biochimie* 2014.
- [40] Castagnini, M., Picchianti, M., Talluri, E., Biagini, M., *et al.*, Arginine-specific mono ADP-ribosylation in vitro of antimicrobial peptides by ADP-ribosylating toxins. *PloS one* 2012, *7*, e41417.
- [41] Llorens-Rico, V., Serrano, L., Lluch-Senar, M., Assessing the hodgepodge of non-mapped reads in bacterial transcriptomes: real or artifactual RNA chimeras? *BMC genomics* 2014, *15*, 633.
- [42] Cao, W., Ma, D., Kapur, A., Patankar, M. S., *et al.*, RT-SVR+q: a strategy for post-Mascot analysis using retention time and q value metric to improve peptide and protein identifications. *Journal of proteomics* 2011, *75*, 480-490.
- [43] Wodke, J. A., Puchalka, J., Lluch-Senar, M., Marcos, J., *et al.*, Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling. *Molecular systems biology* 2013, *9*, 653.
- [44] Robinson, N. E., Robinson, A. B., Deamidation of human proteins. *Proceedings of the National Academy of Sciences of the United States of America* 2001, *98*, 12409-12413.
- [45] Li, X., Cournoyer, J. J., Lin, C., O'Connor, P. B., Use of 18O labels to monitor deamidation during protein and peptide sample processing. *Journal of the American Society for Mass Spectrometry* 2008, *19*, 855-864.
- [46] Li, L., Palencia, A., Lukk, T., Li, Z., *et al.*, Leucyl-tRNA synthetase editing domain functions as a molecular rheostat to control codon ambiguity in Mycoplasma pathogens. *Proceedings of the National Academy of Sciences of the United States of America* 2013, *110*, 3817-3822.
- [47] Schimmel, P., Mistranslation and its control by tRNA synthetases. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 2011, *366*, 2965-2971.
- [48] Wodke, J. A., Alibes, A., Cozzuto, L., Hermoso, A., *et al.*, MyMpn: a database for the systems biology model organism Mycoplasma pneumoniae. *Nucleic acids research* 2014.
- [49] Zaher, H. S., Green, R., Fidelity at the molecular level: lessons from protein synthesis. *Cell* 2009, *136*, 746-762.
- [50] Li, L., Boniecki, M. T., Jaffe, J. D., Imai, B. S., *et al.*, Naturally occurring aminoacyl-tRNA synthetases editing-domain mutations that cause mistranslation in Mycoplasma parasites. *Proceedings of the National Academy of Sciences of the United States of America* 2011, *108*, 9378-9383.
- [51] Lipman, R. P., Clyde, W. A., Jr., The interrelationship of virulence, cytoadsorption, and peroxide formation in Mycoplasma pneumoniae. *Proc Soc Exp Biol Med* 1969, *131*, 1163-1167.

[52] Arai, S., Munakata, T., Kuwano, K., Mycoplasma interaction with lymphocytes and phagocytes: role of hydrogen peroxide released from *M. pneumoniae*. *The Yale journal of biology and medicine* 1983, 56, 631-638.

[53] Marsavelski, A., Lesjak, S., Mocibob, M., Weygand-Durasevic, I., Tomic, S., A single amino acid substitution affects the substrate specificity of the seryl-tRNA synthetase homologue. *Molecular bioSystems* 2014, 10, 3207-3216.

## FIGURE LEGENDS

**Figure 1.** (A) Spectra assigned in the pilot search using Mascot software and the NCBI nr Mycoplasma database. (B) PSMs in the different analyses performed in this study using NCBI nr All as the database. Each black square represents the results obtained in the three different software analyses: *de novo*, *de novo* plus PTMs and *de novo*/PTMs/mutations by Spider. The first pie of each square indicates the number and percentage of spectra assigned (“A”, in dark red) and not assigned (“NA”, grey) to a peptide sequence. The second pie shows the number and percentage of these assigned spectra associated (“A”, dark red) and not associated (“NA”, grey) to a protein annotated in the NCBI nr All database. The third pie represents the number and percentage of spectra associated to a peptide that is found (“A”, dark red) or not found (“NA”, grey) in the *M. pneumoniae* database. Also, results of the BLAST search using peptide sequences not assigned to a protein in the *de novo*/PTMs/mutations by Spider search are represented in the right pie.

**Figure 2.** Functional study of peptides. (A) Pie diagram indicating the percentage of different PTMs identified in the analysis. (B) The graph represents the correlation between the number of alterations in the amino acid sequence and the protein abundances (protein copies per cell) for the 411 proteins quantified in the *M. pneumoniae* proteome [24]. (C) Analysis of errors. The

histogram represents the frequencies of amino acid changes normalized by the frequency of the amino acid in the genome. WT, wild-type, and Alt, changed amino acid found.

## TABLES

**Table 1.** PTMs studied.

Name	$\Delta$ Mass	#PSM	Position
Acetyl	42.01	62	HSTY
Acetylation	42.01	197	Protein N-term
Acetylation	42.01	109	K
Acetylation	42.01	58	N-term
Amidation	-0.98	23	C-term
Biotin	226.08	40	K,N-term
Carbamidomethyl	57.02	4832	C
Dimethylation	28.03	102	KNR,N-term
Dioxidation	31.99	21	M
Formylation	27.99	1724	K,N-term
Hep	192.06	30	KNQRST
Hex	162.05	22	KRT,N-term
Hex1HexNAc1	365.13	7	ST
HexN	161.07	34	KNTW
HexNAc	203.08	13	ST
HexNAc1dHex1	349.14	6	N
Hexose	162.05	12	NSY
Methylation	14.02	151	CDEHKRST,N-term
Oxidation	15.99	4725	M
Oxidation	15.99	169	HW
Phosphorylation	79.97	85	STY
Phosphorylation	79.97	12	DHR
Sulfation	79.96	44	STY
Ubiquitin	114.04	54	CKST



**Table 2.** Sequence alterations detected in the analysis.

WT	aa variant	Number Cases <sup>a</sup>	Number Cases/Proteome <sup>b</sup>	Average Area-Variant/ WT <sup>c</sup> ±SD
Ala	Ser	18 (8)	0.001	0.08±0.04
Phe	Tyr	32 (17)	0.002	0.06±0.03

<sup>a</sup>Total number of peptides identified with the corresponding aa alteration. In brackets are the aa substitutions represented with a  $\Delta RT < 0,5$ . <sup>b</sup>Ratio between the number of mutated amino acids found and the total number of the same amino acid in the proteome. <sup>c</sup>Average of the areas under the curve for those peptides where we identified the mutant and WT peptide (with  $\Delta RT < 0,5$  Table S11).