DOI 10.1002/pmic.200300511

Jacob D. Jaffe[1, 2]
Howard C. Berg[2, 3]
George M. Church[1]

[1]Harvard Medical School Dept. of Genetics, Boston, MA, USA
[2]Harvard University Department of Molecular and Cellular Biology,
Cambridge, MA, USA
[3]The Rowland Institute for Science,
Cambridge, MA, USA

Proteogenomic mapping as a complementary method to perform genome annotation

The accelerated rate of genomic sequencing has led to an abundance of completely sequenced genomes. Annotation of the open reading frames (ORFs) (*i.e.*, gene prediction) in these genomes is an important task and is most often performed computationally based on features in the nucleic acid sequence. Using recent advances in proteomics, we set out to predict the set of ORFs for an organism based principally on expressed protein-based evidence. Using a novel search strategy, we mapped peptides detected in a whole-cell lysate of *Mycoplasma pneumoniae* onto a genomic scaffold and extended these "hits" into ORFs bound by traditional genetic signals to generate a "proteogenomic map". We were able to generate an ORF model for *M. pneumoniae* strain FH using proteomic data with a high correlation to models based on sequence features. Ultimately, we detected over 81% of the genomically predicted ORFs in *M. pneumoniae* strain M129 (the originally sequenced strain). We were also able to detect several new ORFs not originally predicted by genomic methods, various *N*-terminal extensions, and some evidence that would suggest that certain predicted ORFs are bogus. Some of these differences may be a result of the strain analyzed but demonstrate the robustness of protein analysis across closely related genomes. This technique is a cost-effective means to add value to genome annotation, and a prerequisite for proteome quantitation and *in vivo* interaction measures.

Keywords: Mass spectrometry / *Mycoplasma pneumoniae* / Open reading frame determination / Proteogenomic mapping PRO 0511

1 Introduction

The explosion in genomic sequencing has, of late, produced over 80 publicly available, complete genomic sequences for unique organisms [1] and over 700 in the pipeline. Reliance on computational algorithms for prediction and annotation of genes has grown, and manual curation of data sets has diminished. However, direct observation of proteins *via* mass spectrometry is now feasible and may lend more confidence than inference of their expression from genomic sequence alone. This work was designed to unite the potential of proteomics with global genome annotation.

The majority of sequenced genomes belong to archea and eubacteria. The relative ease of sequencing smaller genomes is partially responsible for the abundance of completed microbial genomes relative to those of multicellular species. The bulk of genome primary annotations are achieved through computational tools such as BLAST homology searching, GLIMMER, and GeneMark, which are programs to perform automated sequence analysis and prediction [2–4]. Mass spectrometry has emerged recently as a popular technique for the detection of proteins. Signatures of ions derived from peptides belonging to cellular proteins provide direct molecular evidence of the existence of the protein in the living cell. As well, advances in instrumentation and analysis of complex mixtures have made mass spectrometry the principal technology of the emerging field of proteomics [5].

A potentially important goal of proteomics is the observation of the entire set of protein products synthesized and utilized by an organism. Such an analysis could help to more accurately determine the structure of the corresponding genome, including the boundaries and enumeration of its open reading frames (ORFs). It could additionally help to verify unknowns (URFs) that can not be well established on the basis of homology. It might also allow the detection of post-translational modifications not evident in the genomic sequence itself. To date, sev-

Correspondence: Dr. George M. Church, Harvard Medical School, Dept. of Genetics, 77 Avenue Louis Pasteur, NRB 238, Boston, MA 02115 USA
E-mail: http://arep.med.harvard.edu/gmc/email.html
Fax: +1-617-432-6513

eral groups have performed proteomic analysis in a variety of organisms. These previous efforts have detected 25% of *Saccharomyces cerevisiae,* 44% of *Mycoplasma pneumoniae*, and 60% of *Deinococcus radiodurans* predicted proteins, respectively [6–9]. However, none of these efforts have mainly focused on gene prediction based on proteomics. Rather, the data were used primarily to validate prior predictions. Here, we initially disregard prior gene predictions and generate a novel ORF model based on the detection of expressed proteins alone.

In the tradition of starting with a well-established model system, we chose *Mycoplasma pneumoniae* as a basis for this study. *M. pneumoniae* is a small, wall-less bacterium descended from Gram-positive bacteria with interesting biological properties such as motility and pathogenicity [10]. Its small genome and relatively limited growth environment made it an attractive candidate for proteomic studies. Moreover, these bacteria are not predicted to have regulation at the transcriptional level (*i.e.*, there are no predicted transcriptional regulatory proteins), and therefore we felt it would be likely that we could observe nearly all proteins regardless of genomic structure or growth conditions [10].

To aid this study, we developed new methods for the correlation of mass spectral data to the genome structure of *M. pneumoniae*, which we term “proteogenomic mapping.” Through this technique, we set out to build a set of gene predictions based on observations of peptides from expressed proteins, and then measure its correlation with the published genomic sequence annotation. In assessing this correlation , we discovered many new features in the genomic structure of *M. pneumoniae*, including new ORFs, extensions of existing ORFs, and have suggested removal of questionable predicted URFs. This is significant considering that this sequence has been annotated not once but twice, with the most recent annotation occurring in 2000 [11, 12]. It should be noted that we performed our analysis on a less virulent, but very closely related strain of *M. pneumoniae* (strain FH) than the one whose sequence has been determined (strain M129), but the genomic coordinates described herein refer to the sequenced strain. Therefore, precise coordinates may differ slightly between the two strains but the overall organization of the genome with respect to ORF order is substantially the same. We were able to verify the existence of many heretofore hypothetical proteins, but we also suspect that some others do not exist as translated protein products. In the course of this study, we have determined the most complete proteome (in percent coverage) to date for a single organism. We apply the data to generate a proteogenomic map and new ORF model for *M. pneumoniae.*

2 Materials and methods

2.1 Materials

Heart infusion broth and yeast extract were from Difco (Research Triangle Park, NC, USA). Inactivated horse serum and ampicillin were from Sigma (St. Louis, MO, USA). All other chemicals were of the highest possible commercial quality, HPLC-grade where possible.

2.2 Cell culture

M. pneumoniae strain FH (ATCC 15531, gift of M. Miyata) was cultured using Aluotto medium [13] in 175 cm^2 tissue culture flasks at 37°C until late phase (defined as $OD_{600} >$ 0.175 on a scraped culture). After 4 passages, cells were washed with and scraped into HS buffer (8 mM HEPES, 272 mM sucrose, pH 7.4). Scrapings from several flasks were spun at 20 000 × *g* for 20 min in a Sorvall RC5C centrifuge. The supernatant fraction was removed and pellets were frozen at −20°C.

2.3 Extract preparation

Extracts were prepared using modifications to a previously described method [14]. Frozen cell pellets were resuspended in lysis buffer (8 M urea, 0.05% SDS, 10 mM DTT, 10 mM Tris, pH 8.0) and sonicated for 1 min to disrupt the cells (Heat Systems – Ultrasonics, Inc Model 385 equipped with a microtip; settings: continuous, duty cycle 90%, setting 5). The lysate was reduced by DTT for 30 min at 37°C. Iodoacetamide was added to 50 mM, and the sample was alkylated for 30 min at 37°C in the dark. The extract was dialyzed (3500 Da molecular weight cutoff; Spectapore) against 3 × 1.0 L of dialysis buffer (2 M urea, 5 mM Tris, pH 8.0). The resulting dialysate (1.8 mL) was measured to be 3.3 mg/mL protein using the Coomassie Plus protein assay (Pierce Endogen, Rockford, IL, USA). 40 µg of sequencing-grade modified trypsin (Promega, Madison, WI, USA) was added to the extract and digestion was carried out overnight at 37°C with gentle rotation (final protein:trypsin ration ~150:1).

2.4 Chromatography and mass spectrometry

2.4.1 Strong cation exchange chromatography (SCX)

SCX was carried out on an HP1090 liquid chromatograph (Agilent Technologies, Palo Alto, CA, USA). Chromatography conditions: buffer A, 25% acetonitrile, 1% acetic acid, 1 mM ammonium acetate, pH 3.5; buffer B, 25%

acetonitrile, 1% acetic acid, 2.5% formic acid, 250 mM ammonium acetate, pH 3.5. Column: Partisphere SCX 4.6 × 250 mm (Whatman, Clifton, NJ, USA). Gradient: linear from 0–100% B over 120 min with a flow rate of 0.5 mL/min. The sample was adjusted to 25% acetonitrile/pH 3.5 before being loaded onto the column. 1100 μg of protein was injected. 500 μL fractions were collected. 200–250 μL of each SCX fraction was vacuum dried in a 96-well plate. Pellets were resuspended in 10 μL of 5% acetonitrile / 1% acetic acid.

2.4.2 Reversed-phase chromatography (RPC)

RPC was carried out using a nano-HPLC pump and autosampler (LC Packings, San Francisco, CA, USA). Chromatography conditions: buffer A, 0.5% acetic acid; buffer B, 0.5% acetic acid in acetonitrile. Column: 75 μm × 100 mm MAGIC C18 reversed phase (Michrom, Auburn, CA, USA) in a fritless fused-silica nanospray column pulled and packed in-house, as described in [15]. Gradient: 0–5 min, hold at 5% B, 1.0 μL/min; 5–215 min, from 5% to 35% B, 250 nL/min; 215–230 min, from 35% to 90% B, 250 nL/min; 230–240 min, hold at 90% B, 250 nL/min; 240–245 min, from 90% to 5% B, 250 nL/min; 245–285 min, hold at 5% B; 250 nL/min (except for 265–274 min, flow 1.0 μL/min). 5 μL of each sample was injected.

2.4.3 Mass spectrometry

The nanospray column was directly interfaced to the orifice of an LCQ Classic ion trap mass spectrometer (ThermoFinnigan, San Jose, CA, USA) and mass spectra were recorded using the following strategy. From a single parent scan (MS) spectrum, the five most abundant ions were selected for collision-induced dissociation (CID). MS/MS spectra were collected for each of these top five ions. If a particular parent ion was observed more than 3 times in a 2 min span, it was excluded from analysis for the subsequent 3 min (dynamic exclusion) in order to collect data on less abundant parent ions. Mass spectra were collected throughout the entire chromatography run.

2.5 Data analysis and protein identification

Mass spectra were analyzed by SEQUEST [16]. Multiple database search strategies were used, as summarized in Table 1. High scoring peptide matches were automatically identified and organized using in-house software. As an example, any trypsin-cleavage-derived peptide match with an assumed charge state of $z = 2$ and a SEQUEST XCorr score of > 2.5, or charge state of $z = 3$ and a SEQUEST XCorr score of > 3.75, was automatically accepted as valid (note that these criteria were above those set forth in [17]. Since the SEQUEST algorithm gives higher XCorr scores to longer theoretical peptide matches, a new metric (NormCorr) was used to assess borderline peptide match candidates. The NormCorr score is simply the XCorr divided by the number of potential ions (b series + y series) derived from the peptide. Any putative top-ranked match with a NormCorr of > 0.12, XCorr > 1.5, and satisfying DelCN > 0.25 or RSp =1 was held for further review. These borderline candidates were examined manually and either rejected or accepted for inclusion into the proteome data set. Criteria for manual acceptance required a readily observable series of at least 4 y-ions, or a proline cleavage feature that might confound SEQUEST scoring (especially doubly charged fragment ions at proline). After all data sets were collected, processed, and screened as described above, proteins represented by a small number of observed peptides were checked by hand using the following criteria: 1–2 peptides/protein, all putative peptide spectra were checked; 3–4 peptides/protein, at least 2 putative peptide spectra were checked; 5+ peptides, spectra were not necessarily checked. Any unacceptable spectra were removed from the dataset for accuracy. It should be noted, however, that a small number of peptides could sometimes result in very good coverage for a protein, with the extreme case that 1 observed peptide represented $> 26\%$ of a particular protein's sequence (gi|13508279|).

2.6 Proteogenomic mapping and new ORF detection

The complete nucleotide sequence of *M. pneumoniae* was translated *in silico* (using the mycoplasmal substitution of Trp for the codon UGA) in all 6 frames and chunked into 50% overlapping 80-mer oligopeptides. The result was a FASTA-style database (searchable by SEQUEST) that had genomic position and frame information embedded into each header tag for a given sequence. All data files were analyzed by SEQUEST against this new database, using no enzyme specificity. Peptides were automatically accepted using the criteria stated above (using different thresholds for completely tryptic and partially tryptic peptides), and no borderline reviews were performed for this purpose. Detected peptides were graphically mapped onto the *M. pneumoniae* genome which also displayed the predicted ORFs as defined by the most recent NCBI release for this genome (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Mycoplasma_pneumoniae/NC_000912.ptt). The results were generated and visualized using novel, web-based software tools

Table 1. Data Search Methods

Strategy	Enzyme	Modification	Auto accept	Hold for review	Notes
Tryptic	Trypsin (K,R)	Cys + 57 Met ?+ 16 Lys ?+ 43	z = 2, XCorr = 2.5 *z = 3, XCorr = 3.75*	normcorr > 0.12 AND XCorr > 1.5 AND (Rsp = 1 OR delcn > 0.25)	If criteria were met for z = 2 and z = 3 for the same spectrum, primary data were used to determine the correct charge state
Phospho	Trypsin (K,R)	Cys + 57 Met ?+ 16 Ser,Thr,Tyr ?+ 80	z = 2, XCorr = 2.5 z = 3, XCorr = 3.75 AND Neutral loss of phospho group present in spectrum with > X abundance	If any of the top 3 ranked peptides contained a putative phosphorylation, it was checked by hand if: XCorr was below the threshold AND top ranked peptide had low confidence AND the delcn of the candidate was within 0.2 of the top ranked peptide	All putative phosphopeptides reviewed by hand
Nontryptic	None; cleavage after any amino acid is possible	Cys + 57 Met ?+ 16 Lys ?+ 43	z = 2, XCorr = 3.0 z = 3, XCorr = 4.25 (nontryptic and tryptic peptides were then sorted)	No borderline candidates were held for review	Peptides were favored if 1 of the cleavages was tryptic

+ indicates a SEQUEST search strategy with a static modification of the specified number of Da (the modification always was considered to be present).
?+ specifies a SEQUEST search strategy with a differential modification of the specified number of Da (the modification may or may not be present).

that include embedded URLs to primary mass spectral data and genomic sequence information. The current software is designed specifically for the *M. pneumoniae* proteome project, but we are in the process of generalizing it and will release it as an open source package at a later date. Peptide hits that fell outside of annotated ORFs were automatically detected using another in-house software package and dynamic HTML was generated to provide links to their positions on the proteogenomic map. This enabled rapid screening of potential "new" ORFs by providing direct access to the primary mass spectral data and raw genomic sequence. An ORF model for *M. pneumoniae* was generated computationally but manual review was employed for potential ORFs with less than 5 peptide hits (not necessarily unique). Briefly, each peptide hit was assigned to a bin bounded by the two adjacent stop codons in the same translational frame. Then, for each of these bins, the 3'-most boundary (with respect to translational frame) was termed the "stop" codon for the potential ORF. Two possibilities were considered for the "start" site of each ORF: (i) the upstream start codon closest to the most 5' peptide observed, and (ii) the downstream start codon closest to the 5'-most bin boundary (the maximally extensible ORF). The proteogenomic map shown here is based on the latter method. Start codons were chosen with the following preference: ATG > TTG > GTG > the first 3 bases after the 5' bin boundary. 135 of the potential ORFs were manually reviewed (due to low coverage) by inspection of the raw spectra, with occasional searching against the nonredundant protein database available from NCBI (as of 11/15/02, see [18]) to prove that a spectra was not derived from a contaminant protein. 83 potential ORFs based on obviously spurious hits were manually removed from the ORF model. Peptides with degenerate hits were allowed to serve as the basis for more than one potential ORF. This scenario was limited almost exclusively to fragments of the cytadherence protein P1 and associated cytadherence proteins. A difference map was computed between the current NCBI annotation and the ORF map we generated. An interactive proteogenomic map can be viewed online at http://massive.med.harvard.edu/MPProteoGenomics/index.html. Although spurious ORFs were removed as described above, spurious hits were not edited out of the database. Therefore, one may observe some peptide hits that we felt did not justify the existence of an ORF, and these hits will most likely appear spurious to the trained eye. However, we left these data as part of our map to ensure transparency to those wishing to inspect our results.

3 Results

3.1 Proteogenomic mapping and a proteome-derived ORF model

We developed the technique "proteogenomic mapping" to correlate interpreted mass spectral data with genome sequence information. Briefly, peptide matches generated by SEQUEST were rapidly mapped back to their original genomic locations through a novel database encoding strategy that considered all six possible translational frames (see Section 2.6 for more information) [16]. By locating stop codons adjacent to peptide matches and determining the extensibility of ORFs to a possible start codon based on peptide data alone we were able to generate a proteome-derived ORF model for *M. pneumoniae*. The proteogenomic map and corresponding ORF model can be viewed in Fig. 1 and interactively, online at http://massive.med.harvard.edu/MPProteoGenomics/index.html.

Our ORF model consists of 573 possible ORFs. There is a high degree of correlation between the ORF model we derived from peptides alone with the current gene predictions based on computational algorithms, as depicted in the difference map in Fig. 1. Of the ORFs we predicted through proteogenomic mapping, 504 have exactly the same boundaries as those in the current genome annotation. 39 additional ORFs differ only in location of start codon based on our observations. 19 of these instances resulted in *N*-terminal extensions of genes from their current boundaries based on observation of peptides from regions 5' to their currently assigned start codon. The remaining 20 differences are due to the simplistic nature of start codon assignment of our algorithm, which we arbitrarily designed with the preference ATG $>$ TTG $>$ GTG $>$ NONE and did not consider CTG as an initiation codon, and are probably not real differences. Therefore, we did not include them in Table 4.

44 of the ORFs are derived from fragments of the cytadherence operon proteins such as P1 that are known to be present in multiple copies of the genome (30 match previous genome predictions). Because the amino acid sequences of these copies are repetitive, it is hard to discern whether more than one copy of the protein is expressed. However, the presence of a cytadherence protein fragment would probably make the likelihood of another overlapping ORF in the same region small. For instance, MPN 091, MPN 371, and MPN 465 all are partially overlapping with a newly detected cytadherence operon-derived fragment (by 23 bp, 164 bp, and 481 bp, respectively), and we propose that they should be deleted from the current annotation. That is to say, peptides derived from a cytadherence gene could be mapped to these locations. These peptides are redundant within the genome and we are not proposing to add a novel gene in these locations. Rather, we have detected vestiges of a cytadherence protein in these locations that would make a competing, overlapping unlikely to be expressed given the rarity of overlapping genes in bacteria. Therefore, we have included these repeated fragments as part of the ORF model as a suggestion that some currently predicted ORFs that overlap them may be bogus.

Through the use of proteogenomic mapping, we were able to discover 16 new ORFs for *M. pneumoniae* (Table 4, and Fig. 2a as an example) not previously predicted by computational methods or related to cytadherence proteins. It should be noted that the strain we used for this study (*M. pneumoniae* strain FH) differs from the strain that was originally sequenced and annotated (*M. pneumoniae* strain M129). These strains share almost perfect coding identity (R. Herrmann, personal communication), but it is possible that the new ORFs are specific to strain FH even though we detected them based on an M129-derived database (*i.e.*, mutations to generate novel initiation codons). We consider this an especially good test for proteogenomic mapping: the task of detecting genes where slight sequence polymorphism may exist.

Many of these are completely novel genes, although some had homology to other mycoplasma genes using a BLASTsearch [3]. In one instance, our detection of a novel ORF (from genome position 250021 to 250293) overlaps the specification of an existing ORF, MPN 206 by 274 bp (Fig. 2b). The new ORF is likely correct because (i) it agrees with the general direction of transcription in that region of the genome whereas MPN 206 does not, and (ii) it would serve to complete an operon structure between MPN205 and ptsG, and (iii) MPN 206 was not detected in the general proteome survey. We note that actual protein ORFs overlapping by more than a few codons are extremely rare in nonmobile bacterial genes. Another new ORF from 536168 to 535005 (reverse reading frame, starting with TTG) obviates MPN 441 due to a complete overlap of all 309 bp. Like the new ORF from

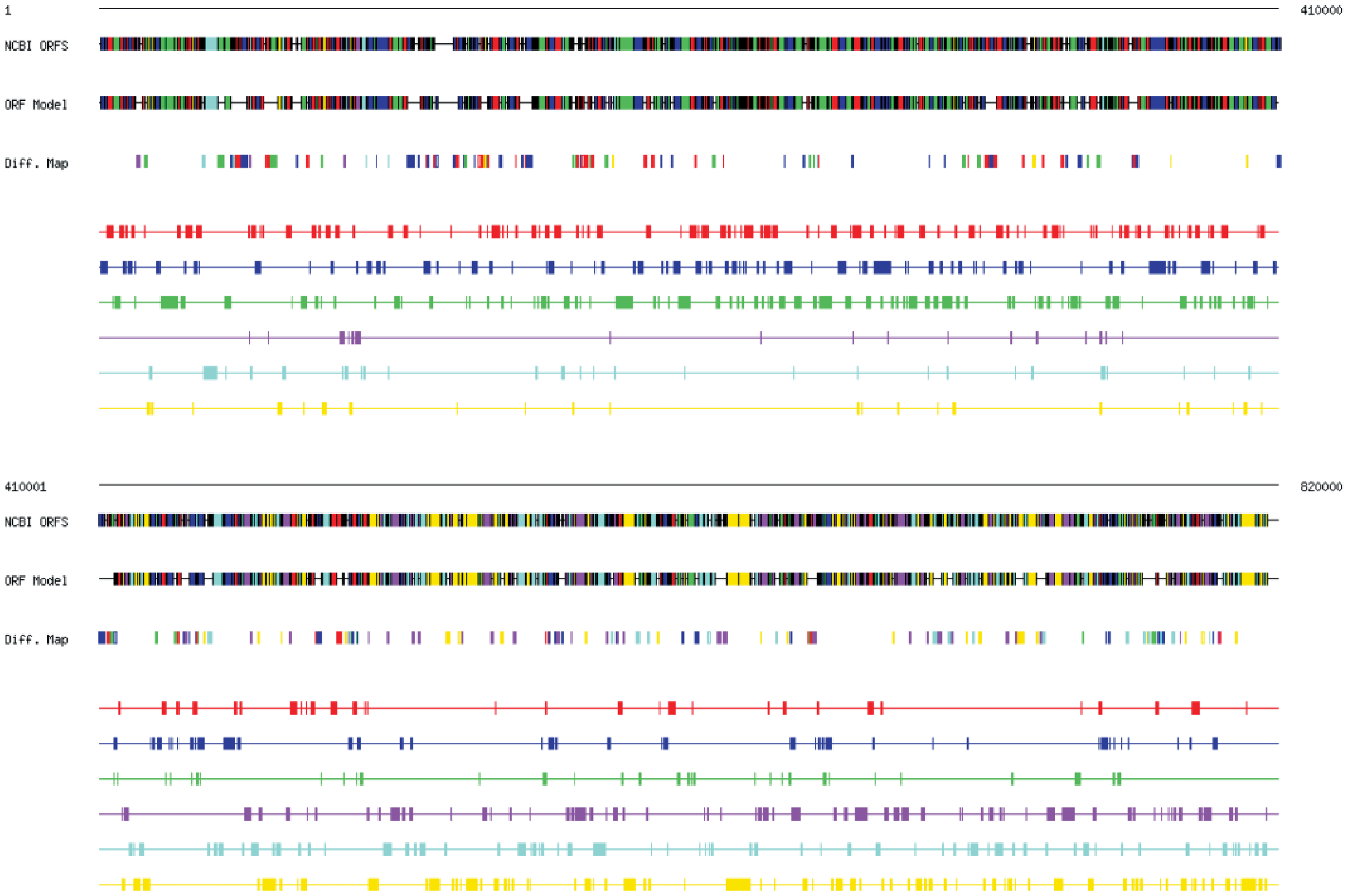

Figure 1. Proteogenomic map of *M. pneumoniae*. Line 1, genome ruler; 2, previously predicted ORFs; 3, proteome-derived ORF model; 4, difference map of lines 2 and 3, closed boxes are present in line 2 but not line 3, while open boxes are the opposite. Lines 2–3 are color-coded by frame (corresponding to lines 5–10), while line 4 is a numerical subtraction of the colors; Lines 5–10: the 6 possible frames of genome translation with rectangles indicating coverage for that genomic region through observation of a peptide. Red, blue, and green are forward frames. Magenta, cyan, and yellow are reverse frames. The pattern repeats twice in the figure to cover the entire genome. This map can be viewed interactively online at http://massive.med.harvard.edu/MPProteoGenomics/.

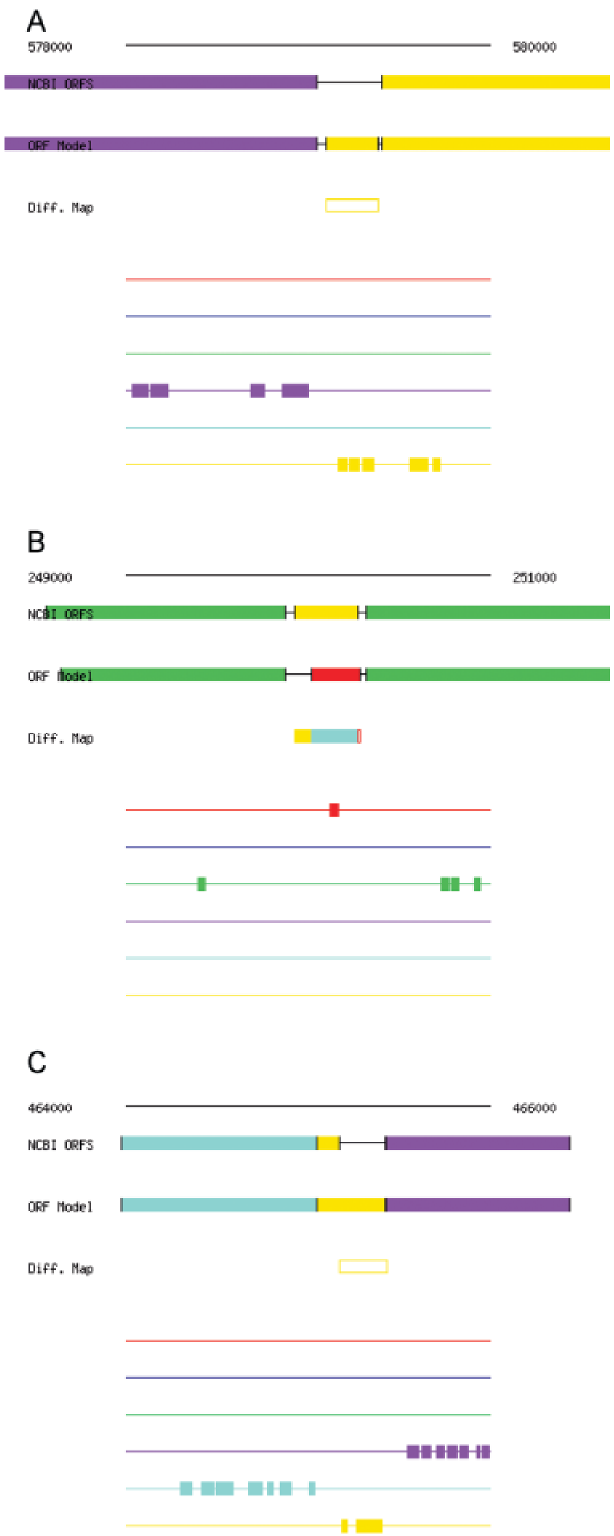

Figure 2. Selected details of the proteogenomic map. (A) Use of proteogenomic mapping to discover a new ORF. (B) Use of proteogenomic mapping to discover a new ORF and delete an inaccurately predicted ORF. (C) *N*-Terminal extension of an existing ORF. Color codes as in Fig. 1.

250021 to 250293, it is in the general direction of transcription in that region of the genome and would complete an operon structure between MPN 440 and MPN 442.

Another interesting example of the sensitivity of proteogenomic mapping is the ability to detect a potential translational frameshift. We observed peptides from the intergenic region that is transcriptionally 3’ to the gene for isoleucyl tRNA ligase (ileRS). These peptides share homology to the 3’-region of ileRS from *M. genitalium.* Notably, this portion of the gene is divergent to the current *M. pneumoniae* annotation for IleRS. We believe that there is either a translational frameshift for this gene or a difference from the published genomic sequence in this region. The frame shift would occur at approximately residue 830 of the protein, and lengthen it from 861 to 895 amino acids.

Several *N*-terminal extensions of current ORFs were detected, as depicted in Table 4. For MPN 388, the *N*-terminal extension that we discovered resulted in addition of this protein to the list of ORFs detected (marked with a ‡ in Table 2, Fig. 2c). Although we did not observe peptides from MPN 388 as defined by the current genome annotation, proteogenomic mapping reveals its existence. Twelve out of the 14 extensions (cytadherence fragments excluded) now predict that genes start with alternative start codons, such as TTG and GTG. Note that this proposed extension verifies the results of another study [8]. While some genes in *M. pneumoniae* are predicted to start with these codons, these discoveries illustrate a possible pitfall of some computational prediction algorithms’ bias toward ATG as the initiation codon. These examples also demonstrate the unbiased nature of proteogenomic mapping.

3.2 Towards a complete proteome of *M. pneumoniae*

When we manually reviewed “borderline” spectra (see Section 2 for criteria), we were able to detect 14 more of the previously predicted ORFs in *M. pneumoniae*. All together, we found 9709 unique peptides corresponding to 557 of the 689 predicted ORFs for *M. pneumoniae* (= 81% coverage) (Tables 2 and 3), plus an additional 61 peptides corresponding to the 16 newly proposed ORFs. This represents the highest degree of proteomic coverage for an organism to date. Amino acid sequence coverage for the 557 detected ORFs averaged 31%. We observed 3 or more peptides for 470 of the 557 detected ORFs, and any protein with less than 5 supporting mass spectra was verified by manual inspection of the primary data.

Table 2. Detection of predicted ORFs

GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage	GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage
13507740	dnaN	MPN001	42	60.3%	13508076†	–	MPN337	19	25.0%
13507741	xdj1	MPN002	14	33.0%	13508077†	–	MPN338	19	26.1%
13507742	gyrB	MPN003	37	48.8%	13508078†	–	MPN339	2	7.1%
13507743	gyrA	MPN004	34	29.6%	13508079	pcrA	MPN340	24	35.5%
13507744	serS	MPN005	12	39.3%	13508080	mutB1	MPN341	1	1.8%
13507745	–	MPN006	12	40.0%	13508081	hsdM	MPN342	5	12.5%
13507746	holB	MPN007	7	32.0%	13508083†	–	MPN344	1	3.7%
13507747	thdF	MPN008	17	31.0%	13508085	–	MPN346	1	7.8%
13507748	yabD	MPN009	6	28.4%	13508086	hsdR	MPN347	1	2.1%
13507749*†	–	MPN010	1	10.7%	13508087	–	MPN348	2	5.5%
13507752*†	–	MPN013	1	5.4%	13508088†	–	MPN349	8	24.6%
13507754†	–	MPN015	15	36.5%	13508089†	ygiH	MPN350	2	9.6%
13507755	rimK	MPN016	10	33.0%	13508090	–	MPN351	10	47.4%
13507756	mtd1	MPN017	13	45.0%	13508091	sigA	MPN352	29	39.7%
13507757	pmd1	MPN018	8	12.4%	13508092	dnaE	MPN353	7	16.6%
13507758	msbA	MPN019	19	24.6%	13508093	grs1	MPN354	23	49.9%
13507759	yb95	MPN020	55	39.2%	13508094	yacO	MPN355	10	37.2%
13507760	dnaJ	MPN021	29	49.5%	13508095	cysS	MPN356	8	18.3%
13507761	pip	MPN022	40	62.1%	13508096	lig	MPN357	23	35.9%
13507762	metS	MPN023	29	36.9%	13508097†	–	MPN358	8	11.2%
13507763	rpoE	MPN024	23	45.9%	13508099	rpmE	MPN360	8	47.4%
13507764	tsr	MPN025	36	71.5%	13508100	prfA	MPN361	14	31.5%
13507765	yyaF	MPN026	12	22.7%	13508101	–	MPN362	12	26.7%
13507766†	–	MPN027	1	5.8%	13508102*†	–	MPN363	1	9.8%
13507767	trsB	MPN028	8	22.4%	13508103*†	–	MPN364	7	5.2%
13507768	efp	MPN029	12	47.9%	13508105	–	MPN366	4	14.7%
13507769†	–	MPN030	2	20.2%	13508107†	–	MPN368	2	8.3%
13507770†	–	MPN031	4	23.6%	13508109†	–	MPN370	21	15.1%
13507772	upp	MPN033	11	42.7%	13508111	–	MPN372	41	44.7%
13507773	polC	MPN034	45	30.6%	13508115†	–	MPN376	30	27.2%
13507775†	–	MPN036	14	22.4%	13508116†	–	MPN377	10	54.1%
13507782	glpF	MPN043	4	12.9%	13508117	dnaE	MPN378	26	25.7%
13507783	tdk	MPN044	19	64.9%	13508118	polA	MPN379	25	55.7%
13507784	hisS	MPN045	12	20.5%	13508119	fpg	MPN380	22	39.4%
13507785	aspS	MPN046	26	35.7%	13508120	yidA	MPN381	18	42.1%
13507786	–	MPN047	13	26.8%	13508121†	–	MPN382	6	26.5%
13507788	–	MPN049	1	1.4%	13508122	yidA	MPN383	21	50.7%
13507789	glpK	MPN050	33	51.6%	13508123	leuS	MPN384	33	34.2%
13507790	glpD	MPN051	30	52.3%	13508125	yaaF	MPN386	17	34.5%
13507791	–	MPN052	46	46.9%	13508126†	–	MPN387	14	38.8%
13507792	ptsH	MPN053	10	70.5%	13508127† ‡	–	MPN388	16	38.3%
13507794	potA	MPN055	41	43.2%	13508128	lplA	MPN389	50	64.0%
13507796	potI	MPN057	3	10.5%	13508129	pdhD	MPN390	45	49.7%
13507797†	–	MPN058	16	29.7%	13508130	pdhC	MPN391	34	57.7%
13507798	gcp	MPN059	11	26.6%	13508131	pdhB	MPN392	59	64.8%
13507799	metX	MPN060	16	33.9%	13508132	pdhA	MPN393	54	75.7%
13507800	ffh	MPN061	31	45.1%	13508133	nox	MPN394	61	48.6%
13507801	deoD	MPN062	19	53.8%	13508134	apt	MPN395	10	33.9%
13507802	deoC	MPN063	15	53.1%	13508135	–	MPN396	24	19.4%
13507803	deoA	MPN064	20	35.2%	13508136	spoT	MPN397	22	28.2%
13507804	cdd	MPN065	8	39.8%	13508137†	–	MPN398	7	23.4%
13507805	cpsG	MPN066	30	39.4%	13508138†	–	MPN399	11	31.0%
13507806	nusG	MPN067	38	50.3%	13508139†	–	MPN400	41	53.3%

Table 2. Continued

GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage	GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage
13507807	SecE	MPN068	3	10.4%	13508140	greA	MPN401	36	83.8%
13507808	rpmG2	MPN069	1	18.8%	13508141	proS	MPN402	26	44.9%
13507809†	–	MPN070	1	7.9%	13508145	–	MPN406	7	45.2%
13507810†	yabC	MPN071	7	21.4%	13508146	–	MPN407	1	1.7%
13507811	yabF	MPN072	4	26.4%	13508147†	–	MPN408	21	27.1%
13507812	prs	MPN073	24	38.1%	13508149*†	–	MPN410	2	13.5%
13507813†	–	MPN074	1	9.5%	13508150†	–	MPN411	1	2.8%
13507814	ywdF	MPN075	3	15.7%	13508151*	–	MPN412	9	23.7%
13507815†	uhpT	MPN076	28	24.5%	13508153*†	–	MPN414	8	13.2%
13507816†	–	MPN077	12	15.2%	13508154	P37	MPN415	5	22.6%
13507817	fruA	MPN078	23	25.8%	13508155	P29	MPN416	5	20.1%
13507818	fruK	MPN079	9	23.7%	13508156	P69	MPN417	1	1.8%
13507819	–	MPN080	18	12.8%	13508157†	MPN419	MPN419	7	51.4%
13507820	glnQ	MPN081	17	35.7%	13508158	alaS	MPN418	57	44.3%
13507821	tklB	MPN082	29	34.3%	13508159	glpQ	MPN420	22	58.1%
13507822†	–	MPN083	7	12.2%	13508160	–	MPN421	2	4.0%
13507823†	–	MPN084	9	24.8%	13508161	–	MPN422	11	41.4%
13507828†	hsdS	MPN089	1	4.5%	13508162†	–	MPN423	1	5.4%
13507829	–	MPN090	4	10.6%	13508163†	ylxM	MPN424	1	6.9%
13507831*	–	MPN092	2	12.1%	13508164	ftsY	MPN425	26	56.0%
13507832	–	MPN093	7	22.3%	13508165	–	MPN426	72	53.2%
13507833*†	–	MPN094	1	5.0%	13508166	yidA	MPN427	22	52.4%
13507836*†	–	MPN097	5	8.5%	13508167	pta	MPN428	45	63.4%
13507838*†	–	MPN099	6	11.0%	13508168	pgk	MPN429	56	58.2%
13507839†	–	MPN100	4	17.5%	13508169	gap	MPN430	62	82.2%
13507840*†	–	MPN101	12	17.5%	13508171	artP	MPN432	4	10.7%
13507844	pheS	MPN105	5	14.1%	13508173	dnaK	MPN434	89	68.7%
13507845	pheT	MPN106	32	36.6%	13508174	–	MPN435	2	4.8%
13507847†	–	MPN108	3	9.2%	13508175	–	MPN436	62	51.3%
13507848	–	MPN109	2	7.9%	13508176*	–	MPN437	2	5.1%
13507849†	–	MPN110	6	8.2%	13508178	–	MPN439	1	5.9%
13507850*†	–	MPN111	1	3.3%	13508179	–	MPN440	6	9.6%
13507854	infC	MPN115	18	41.3%	13508182	deaD	MPN443	3	8.3%
13507855	rpmI	MPN116	1	18.6%	13508183†	–	MPN444	54	43.2%
13507856	rpLT	MPN117	11	43.3%	13508184	lip3	MPN445	2	12.5%
13507857	–	MPN118	6	24.6%	13508185	rpsD	MPN446	32	51.2%
13507858	–	MPN119	14	14.7%	13508186	hmw1	MPN447	39	25.1%
13507859	grpE	MPN120	24	40.6%	13508188†	orf8	MPN449	14	23.4%
13507860†	–	MPN121	6	42.1%	13508189†	orf7	MPN450	3	7.3%
13507861	parB	MPN122	21	27.6%	13508191	hmw3	MPN452	48	37.8%
13507862	parC	MPN123	24	23.2%	13508192	–	MPN453	8	30.7%
13507863	yqxE	MPN124	12	34.8%	13508193†	–	MPN454	9	32.6%
13507864	uvrC	MPN125	14	21.2%	13508195†	–	MPN456	77	48.3%
13507865	–	MPN126	1	8.2%	13508196*†	–	MPN457	7	11.9%
13507867*†	–	MPN128	3	20.1%	13508197*†	–	MPN458	2	11.5%
13507870*†	–	MPN131	1	5.4%	13508198*†	–	MPN459	12	14.6%
13507872	–	MPN133	3	8.0%	13508199	ktrB	MPN460	3	5.5%
13507873	ugpC	MPN134	40	43.7%	13508200	ktrA	MPN461	10	41.6%
13507874	ugpA	MPN135	4	10.6%	13508201†	–	MPN462	5	18.1%
13507876†	–	MPN137	2	8.8%	13508203*	–	MPN464	19	27.9%
13507877*†	–	MPN138	1	4.2%	13508205*†	–	MPN466	3	12.1%
13507878†	–	MPN139	2	11.0%	13508208†	–	MPN469	5	14.4%
13507879	orf4	MPN140	26	37.3%	13508209	pepX	MPN470	54	58.2%

Table 2. Continued

GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage	GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage
13507880	P1	MPN141	96	42.1%	13508211†	degV	MPN472	13	39.9%
13507881	orf6	MPN142	93	37.4%	13508212	lip2	MPN473	7	22.0%
13507883*†	–	MPN144	6	15.5%	13508213	–	MPN474	60	41.3%
13507885†	–	MPN146	1	3.8%	13508214	–	MPN475	22	41.4%
13507886†	–	MPN147	2	7.0%	13508215	cmk	MPN476	6	31.8%
13507887*†	–	MPN148	4	20.0%	13508216†	–	MPN477	2	9.1%
13507888*†	–	MPN149	6	13.4%	13508217	–	MPN478	12	25.5%
13507889*†	–	MPN150	1	4.9%	13508218	–	MPN479	25	58.4%
13507891†	–	MPN152	26	31.1%	13508219	valS	MPN480	29	29.7%
13507892†	–	MPN153	87	48.2%	13508220	yihA	MPN481	5	23.3%
13507893	nusA	MPN154	52	54.4%	13508221†	MPN482	MPN482	2	20.3%
13507894	infB	MPN155	34	37.4%	13508222	yibD	MPN483	9	21.1%
13507895	rbfA	MPN156	2	15.5%	13508224*	–	MPN485	8	24.7%
13507896†	–	MPN157	11	19.9%	13508226	nifS	MPN487	15	35.0%
13507897	yaaC	MPN158	10	34.2%	13508227	–	MPN488	6	25.7%
13507898	hlyC	MPN159	13	24.8%	13508228	–	MPN489	52	34.0%
13507899†	–	MPN160	1	1.9%	13508229	recA	MPN490	2	5.1%
13507900†	–	MPN161	23	36.4%	13508230	–	MPN491	7	12.7%
13507901†	–	MPN162	13	18.8%	13508231	yjfW	MPN492	8	29.2%
13507902†	–	MPN163	1	12.5%	13508232	yjfV	MPN493	13	56.0%
13507903	rpsJ	MPN164	20	44.4%	13508234	MPN495	MPN495	6	35.8%
13507904	rplC	MPN165	25	34.8%	13508235	yjfS	MPN496	3	5.2%
13507905	rplD	MPN166	26	49.5%	13508237	araD	MPN498	2	5.0%
13507906	rplW	MPN167	17	40.1%	13508238†	–	MPN499	3	14.7%
13507907	rplB	MPN168	22	33.1%	13508239†	–	MPN500	17	23.5%
13507908	rpsS	MPN169	12	67.8%	13508240*†	–	MPN501	3	10.7%
13507909	rplV	MPN170	7	31.5%	13508241†	–	MPN502	24	48.3%
13507910	rpsC	MPN171	22	43.2%	13508242	–	MPN503	15	35.2%
13507911	rplP	MPN172	8	33.8%	13508244*†	–	MPN505	1	4.3%
13507912	rpmC	MPN173	9	37.8%	13508245†	–	MPN506	17	16.3%
13507913	rpsQ	MPN174	8	34.1%	13508247	–	MPN508	2	4.3%
13507914	rplN	MPN175	7	28.7%	13508248	–	MPN509	3	7.5%
13507915	rplX	MPN176	1	8.1%	13508249	–	MPN510	2	6.1%
13507916	rplE	MPN177	24	62.8%	13508254	rpoC	MPN515	137	54.5%
13507917	rpsN	MPN178	4	52.5%	13508255	rpoB	MPN516	136	52.1%
13507918	rpsH	MPN179	7	37.3%	13508256	–	MPN517	18	52.4%
13507919	rplF	MPN180	22	62.5%	13508257†	–	MPN518	20	34.2%
13507920	rplR	MPN181	8	38.8%	13508258	lip3	MPN519	3	12.5%
13507921	rpsE	MPN182	15	42.0%	13508259	ileS	MPN520	36	34.6%
13507922	rplO	MPN183	8	31.1%	13508260	ygl3	MPN521	7	45.2%
13507923	secY	MPN184	3	11.1%	13508261	–	MPN522	6	31.0%
13507924	adk	MPN185	18	54.4%	13508262	–	MPN523	6	15.1%
13507925	map	MPN186	1	7.3%	13508264†	–	MPN525	1	2.7%
13507926	infA	MPN187	6	39.7%	13508265†	–	MPN526	14	27.4%
13507927	rpmJ	MPN188	1	21.6%	13508267	ppa	MPN528	14	44.6%
13507928	rpsM	MPN189	7	29.0%	13508268	–	MPN529	4	22.0%
13507929	rpsK	MPN190	8	34.7%	13508269†	–	MPN530	19	55.1%
13507930	rpoA	MPN191	43	62.4%	13508270	clpB	MPN531	88	64.1%
13507931	rplQ	MPN192	7	22.6%	13508271	licA	MPN532	19	37.6%
13507932	CbiO	MPN193	11	32.5%	13508272	ackA	MPN533	52	53.8%
13507933	hisP	MPN194	20	45.5%	13508277	rplJ	MPN538	15	44.1%
13507934	–	MPN195	10	18.4%	13508278	rplL	MPN539	14	81.1%
13507935	hisT	MPN196	5	23.0%	13508279	rpmF	MPN540	3	26.3%

Table 2. Continued

GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage	GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage
13507936	pepF	MPN197	62	53.5%	13508280	rpsT	MPN541	6	19.5%
13507937	mte1	MPN198	8	22.3%	13508281†	–	MPN542	7	21.6%
13507938†	–	MPN199	6	9.6%	13508282	fmt	MPN543	7	23.5%
13507939†	–	MPN200	49	49.9%	13508283†	–	MPN544	13	17.6%
13507941*†	–	MPN202	3	6.1%	13508284	rnc	MPN545	7	19.5%
13507943*†	–	MPN204	3	18.2%	13508285	plsX	MPN546	20	41.8%
13507944*†	–	MPN205	8	9.1%	13508286	–	MPN547	51	58.4%
13507946	ptsG	MPN207	76	48.3%	13508287	–	MPN548	4	12.0%
13507947	rpsB	MPN208	18	44.6%	13508288	–	MPN549	16	34.5%
13507948	mgtA	MPN209	14	17.4%	13508289	–	MPN550	6	22.7%
13507949	secA	MPN210	59	42.5%	13508290†	–	MPN551	8	32.4%
13507950	uvrB	MPN211	37	41.9%	13508291†	–	MPN552	10	25.7%
13507952†	–	MPN213	40	34.0%	13508292	thrSv	MPN553	38	40.4%
13507953†	–	MPN214	3	17.4%	13508293†	–	MPN554	3	20.2%
13507954	oppB	MPN215	15	23.1%	13508294†	–	MPN555	20	46.1%
13507955	amiD	MPN216	21	30.3%	13508295	argS	MPN556	38	50.1%
13507956	oppD	MPN217	19	36.2%	13508296	gidA	MPN557	17	33.5%
13507957	oppF	MPN218	60	45.7%	13508297	gidB	MPN558	18	58.6%
13507958	rplK	MPN219	12	47.4%	13508298†	–	MPN559	2	9.8%
13507959	rplA	MPN220	22	48.7%	13508299	arcA	MPN560	22	43.2%
13507960	pth	MPN221	7	33.0%	13508300	udk	MPN561	6	21.6%
13507961	yacA	MPN222	7	30.1%	13508301	outB	MPN562	17	42.7%
13507962	–	MPN223	15	38.5%	13508302	obg	MPN563	14	38.3%
13507963	lgt	MPN224	6	18.0%	13508303	adh	MPN564	3	13.7%
13507964	rpsL	MPN225	8	28.1%	13508305	glpQ	MPN566	11	33.8%
13507965	rpsG	MPN226	20	52.9%	13508306	P200	MPN567	14	8.0%
13507966	fus	MPN227	79	64.5%	13508307	spg	MPN568	14	37.8%
13507967	rpsF	MPN228	18	50.2%	13508308	–	MPN569	2	18.5%
13507968	ssb	MPN229	9	38.6%	13508310	lcnDR3	MPN571	1	2.0%
13507969	rpsR	MPN230	9	49.0%	13508311	–	MPN572	66	66.1%
13507970	rplI	MPN231	9	34.9%	13508312	groEL	MPN573	104	69.4%
13507971	dnaB	MPN232	32	38.7%	13508313	groES	MPN574	30	82.8%
13507972	–	MPN233	31	65.2%	13508315	glyA	MPN576	36	44.3%
13507973	–	MPN234	3	8.4%	13508320*†	–	MPN581	1	3.8%
13507974	ung	MPN235	3	24.2%	13508321†	–	MPN582	2	7.1%
13507975	–	MPN236	12	16.7%	13508324†	–	MPN585	3	9.3%
13507976	–	MPN237	30	35.4%	13508325†	–	MPN586	1	5.5%
13507977	PET112	MPN238	31	43.5%	13508326†	–	MPN587	1	11.3%
13507978†	–	MPN239	15	48.2%	13508327†	–	MPN588	3	5.6%
13507979	trxB	MPN240	21	49.8%	13508329*†	–	MPN590	2	8.3%
13507980†	–	MPN241	5	24.3%	13508330†	–	MPN591	5	16.1%
13507982	vacB	MPN243	58	46.4%	13508331†	–	MPN592	18	33.0%
13507983†	–	MPN244	5	17.3%	13508334	lacA	MPN595	6	29.6%
13507984	def	MPN245	13	48.1%	13508335†	–	MPN596	3	4.6%
13507985	gmk	MPN246	10	28.9%	13508336	atpC	MPN597	8	43.6%
13507986	ptc1	MPN247	15	38.2%	13508337	atpD	MPN598	54	69.1%
13507987	–	MPN248	7	21.1%	13508338	atpG	MPN599	7	21.1%
13507989	pgiB	MPN250	15	22.3%	13508339	atpA	MPN600	36	38.8%
13507990	cfxE	MPN251	3	17.7%	13508340	atpH	MPN601	4	18.0%
13507991	asnS	MPN252	22	34.7%	13508341	atpF	MPN602	12	42.0%
13507992	pgsA	MPN253	1	4.0%	13508342	atpE	MPN603	2	15.2%
13507993†	MPN254	MPN254	16	38.9%	13508343	atpB	MPN604	3	4.4%
13507995†	–	MPN256	6	24.7%	13508345	eno	MPN606	63	75.0%

Table 2. Continued

GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage	GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage
13507996	galE	MPN257	3	8.6%	13508346	pmsR	MPN607	15	51.0%
13507998	–	MPN259	1	1.4%	13508347	phoU	MPN608	6	23.1%
13508000	topA	MPN261	61	44.7%	13508350	–	MPN611	16	37.7%
13508001†	–	MPN262	17	27.2%	13508351†	–	MPN612	2	2.2%
13508002	trx	MPN263	16	60.8%	13508355	rpsI	MPN616	8	37.9%
13508003	–	MPN264	15	38.1%	13508356	rplM	MPN617	10	33.6%
13508004	trpS	MPN265	11	29.5%	13508357	dnaX	MPN618	30	37.3%
13508005	ygl1	MPN266	18	68.3%	13508358	uvrA	MPN619	48	48.4%
13508006†	–	MPN267	20	47.1%	13508359†	–	MPN620	26	24.1%
13508007	–	MPN268	6	35.9%	13508360	–	MPN621	31	39.2%
13508008	ysr1	MPN269	15	26.0%	13508361	rpsO	MPN622	1	9.3%
13508010*†	–	MPN271	1	4.4%	13508362	deaD	MPN623	22	30.7%
13508011†	–	MPN272	6	29.0%	13508363	rpmB	MPN624	3	36.9%
13508012	hit1	MPN273	17	63.2%	13508364	–	MPN625	11	56.7%
13508014†	yaaK	MPN275	4	45.0%	13508366	ptsI	MPN627	47	51.2%
13508015†	–	MPN276	2	12.3%	13508367	pgm	MPN628	35	44.5%
13508016	lysS	MPN277	23	32.3%	13508368	tim	MPN629	27	53.7%
13508017	yefE	MPN278	6	15.3%	13508369†	yfiB	MPN630	3	6.2%
13508018	lepA	MPN279	12	20.7%	13508370	tsf	MPN631	33	62.8%
13508019	–	MPN280	45	53.1%	13508371	pyrH	MPN632	10	50.6%
13508020†	–	MPN281	4	10.6%	13508372†	–	MPN633	15	55.9%
13508021†	–	MPN282	1	5.4%	13508373†	–	MPN634	5	26.0%
13508023†	–	MPN284	99	58.6%	13508375	frr	MPN636	23	62.5%
13508025*†	–	MPN286	1	2.6%	13508376	cdsA	MPN637	2	5.1%
13508027†	–	MPN288	21	17.0%	13508377	–	MPN638	28	46.9%
13508030†	–	MPN291	10	37.2%	13508378†	–	MPN639	6	19.5%
13508031	yceC	MPN292	5	17.8%	13508379†	–	MPN640	15	32.7%
13508032	lsp	MPN293	2	10.3%	13508381†	–	MPN642	20	51.6%
13508033	–	MPN294	8	36.4%	13508382†	–	MPN643	5	17.2%
13508034†	–	MPN295	24	66.8%	13508383†	–	MPN644	2	8.5%
13508035	rpsU	MPN296	3	18.3%	13508384†	–	MPN645	10	30.7%
13508036†	–	MPN297	15	53.7%	13508385†	–	MPN646	5	19.9%
13508037†	–	MPN298	1	9.2%	13508386†	–	MPN647	4	21.7%
13508038	plsB	MPN299	12	43.6%	13508389†	–	MPN650	1	7.9%
13508039	dyr	MPN300	24	36.0%	13508391	mtlD	MPN652	4	12.9%
13508040†	ypuH	MPN301	18	54.3%	13508395	–	MPN656	6	31.0%
13508041	pfk	MPN302	24	59.8%	13508396†	–	MPN657	5	10.5%
13508042	pyk	MPN303	54	51.4%	13508397	rplS	MPN658	6	47.1%
13508046	arcC	MPN307	11	39.5%	13508398	trmD	MPN659	1	3.8%
13508047	–	MPN308	1	2.3%	13508399	rpsP	MPN660	8	39.8%
13508048†	P65	MPN309	16	23.5%	13508400†	–	MPN661	2	7.3%
13508049	–	MPN310	141	44.5%	13508401†	pilB	MPN662	19	70.2%
13508050†	–	MPN311	29	45.1%	13508402†	–	MPN663	11	40.2%
13508051†	–	MPN312	7	40.8%	13508403†	degV	MPN664	16	35.0%
13508053†	yabB	MPN314	20	63.1%	13508404	tuf	MPN665	67	66.5%
13508054	yabC	MPN315	13	27.6%	13508405†	–	MPN666	6	17.5%
13508055†	–	MPN316	2	7.9%	13508406	gtaB	MPN667	16	37.1%
13508056	ftsZ	MPN317	4	11.3%	13508407†	osmC	MPN668	8	32.1%
13508057	–	MPN318	6	7.9%	13508408	tyrS	MPN669	12	30.1%
13508058	gap1	MPN319	6	8.5%	13508409†	–	MPN670	28	50.7%
13508059	thyA	MPN320	5	13.7%	13508410	ftsH	MPN671	64	57.8%
13508060	dhfr	MPN321	8	34.4%	13508411	hpt	MPN672	4	20.6%
13508061	nrdF	MPN322	29	43.7%	13508412†	–	MPN673	11	29.0%

Table 2. Continued

GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage	GenBankID	Gene name	MPN (as in [11])	Unique supporting peptides	% Sequence coverage
13508062	–	MPN323	11	54.2%	13508413	ldh	MPN674	66	74.7%
13508063	nrdE	MPN324	26	34.0%	13508416	–	MPN677	15	27.1%
13508064	rplU	MPN325	9	47.0%	13508417	gltX	MPN678	32	38.4%
13508065†	ysxB	MPN326	3	34.0%	13508418	ksgA	MPN679	10	25.9%
13508066	rpmA	MPN327	4	36.5%	13508419†	–	MPN680	8	10.4%
13508067	nfo	MPN328	9	32.5%	13508420	rnpA	MPN681	1	8.5%
13508068	–	MPN329	3	19.6%	13508421	rpmH	MPN682	1	18.8%
13508069†	–	MPN330	7	24.8%	13508422	devA	MPN683	5	19.2%
13508070	tig	MPN331	60	51.4%	13508423†	–	MPN684	70	35.8%
13508071	lon	MPN332	82	59.1%	13508424	cysA	MPN685	24	56.0%
13508072†	–	MPN333	4	7.1%	13508425	dnaA	MPN686	36	50.6%
13508073	bcrA	MPN334	2	3.7%	13508426†	–	MPN687	9	24.4%
13508074†	–	MPN335	2	4.0%	13508427	soj	MPN688	28	51.9%
13508075	–	MPN336	10	27.0%					

Entries marked with * are supported only by degenerate peptides (see text).
Entries marked with † are confirmed hypothetical proteins.
Entries marked with ‡ were detected only after proteogenomic mapping was used to extend a reading frame.

Table 3. Predicted ORFs not detected

GenBankID	Gene name	MPN	Detected paralog	GenBankID	Gene name	MPN	Detected paralog
13507750	–	MPN011		13508114	–	MPN375	
13507751	–	MPN012		13508124	–	MPN385	
13507753	dnaE	MPN014	MPN353	13508142	–	MPN403	
13507771	–	MPN032		13508143	–	MPN404	
13507774	–	MPN035		13508144	–	MPN405	
13507776	–	MPN037		13508148	–	MPN409	
13507777	–	MPN038		13508152	–	MPN413	
13507778	–	MPN039		13508170	–	MPN431	
13507779	–	MPN040		13508172	cbiO	MPN433	MPN432
13507780	–	MPN041		13508177	–	MPN438	
13507781	–	MPN042		13508180	–	MPN441	
13507787	–	MPN048		13508181	–	MPN442	
13507793	–	MPN054		13508187	–	MPN448	
13507795	potB	MPN056		13508190	come3	MPN451	
13507824	–	MPN085		13508194	ctaD	MPN455	
13507825	–	MPN086		13508202	–	MPN463	
13507826	–	MPN087		13508204	–	MPN465	
13507827	–	MPN088		13508206	–	MPN467	
13507830	–	MPN091		13508207	–	MPN468	
13507834	–	MPN095		13508210	rpmG	MPN471	MPN069
13507835	–	MPN096		13508223	–	MPN484	
13507837	–	MPN098		13508225	–	MPN486	
13507841	–	MPN102		13508233	yjfU	MPN494	
13507842	–	MPN103		13508236	–	MPN497	
13507843	–	MPN104		13508243	–	MPN504	
13507846	–	MPN107		13508246	–	MPN507	
13507851	–	MPN112		13508250	–	MPN511	

Table 3. Predicted ORFs not detected

GenBankID	Gene name	MPN	Detected paralog	GenBankID	Gene name	MPN	Detected paralog
13507852	–	MPN113		13508251	–	MPN512	
13507853	cpt2	MPN114		13508252	–	MPN513	
13507866	–	MPN127		13508253	–	MPN514	
13507868	–	MPN129		13508263	–	MPN524	
13507869	–	MPN130		13508266	–	MPN527	
13507871	–	MPN132		13508273	–	MPN534	
13507875	ugpE	MPN136		13508274	ruvA	MPN535	
13507882	–	MPN143		13508275	ruvB	MPN536	
13507884	–	MPN145		13508276	mucB	MPN537	
13507890	–	MPN151		13508304	–	MPN565	
13507940	–	MPN201		13508309	–	MPN570	
13507942	–	MPN203		13508314	–	MPN575	
13507945	–	MPN206		13508316	–	MPN577	
13507951	–	MPN212		13508317	–	MPN578	
13507981	–	MPN242		13508318	–	MPN579	
13507988	yjeQ	MPN249		13508319	–	MPN580	
13507994	–	MPN255		13508322	–	MPN583	
13507997	yjcW	MPN258		13508323	–	MPN584	
13507999	rbsC	MPN260		13508328	–	MPN589	
13508009	–	MPN270		13508332	–	MPN593	
13508013	–	MPN274		13508333	–	MPN594	
13508022	–	MPN283		13508344	–	MPN605	
13508024	prrB	MPN285	MPN089	13508348	pstB	MPN609	
13508026	–	MPN287		13508349	pstA	MPN610	
13508028	hsdS1B	MPN289	MPN089	13508352	–	MPN613	
13508029	–	MPN290		13508353	–	MPN614	
13508043	arcA	MPN304	MPN560	13508354	hsdS	MPN615	
13508044	arcA	MPN305	MPN560	13508365	–	MPN626	
13508045	argI	MPN306		13508374	–	MPN635	
13508052	–	MPN313		13508380	–	MPN641	
13508082	–	MPN343		13508387	–	MPN648	
13508084	hsdR	MPN345	MPN347	13508388	–	MPN649	
13508098	–	MPN359		13508390	mtlA	MPN651	
13508104	–	MPN365		13508392	mtlF	MPN653	
13508106	–	MPN367		13508393	–	MPN654	
13508108	–	MPN369		13508394	–	MPN655	
13508110	–	MPN371		13508414	–	MPN675	
13508112	–	MPN373		13508415	–	MPN676	
13508113	–	MPN374		13508428	–	MPN528a	

If a homolog of a “named” gene was detected in the proteome survey, its MPN code is listed to the right of the undetected gene’s MPN code.

Figure 3a shows a breakdown by functional category for the proteins we observed. The rate of detection was quite high (90%) for “well-annotated” genes, *i.e.*, genes that have homologs outside of the mycoplasma family or whose function is well documented. Interestingly, the detection rate was significantly lower (66%, *p*-value = 1.7×10^{-15}) for “poorly-annotated” genes, *i.e.*, genes derived from *Mycoplasma*-specific homology search results or functionally unannotated ORFs. We included any ORF with the word “hypothetical” in its NCBI annotation to compose this category. Figure 3b illustrates this discontinuity between “well-annotated” and “poorly annotated” genes. We hypothesize that the ability to detect members of either of these classes of proteins in this organism should be roughly equal given its apparent lack of transcriptional regulation. We therefore propose that not all of the 259 proteins falling into the “hypothetical” category are *bona fide* ORFs. Based on these estimates, we believe that the true number of ORFs in *M. pneumoniae* is closer to 622 rather than 689. We were able to verify that up to 172 hypothetical URFs exist as translated protein products (marked with a † in Table 2).

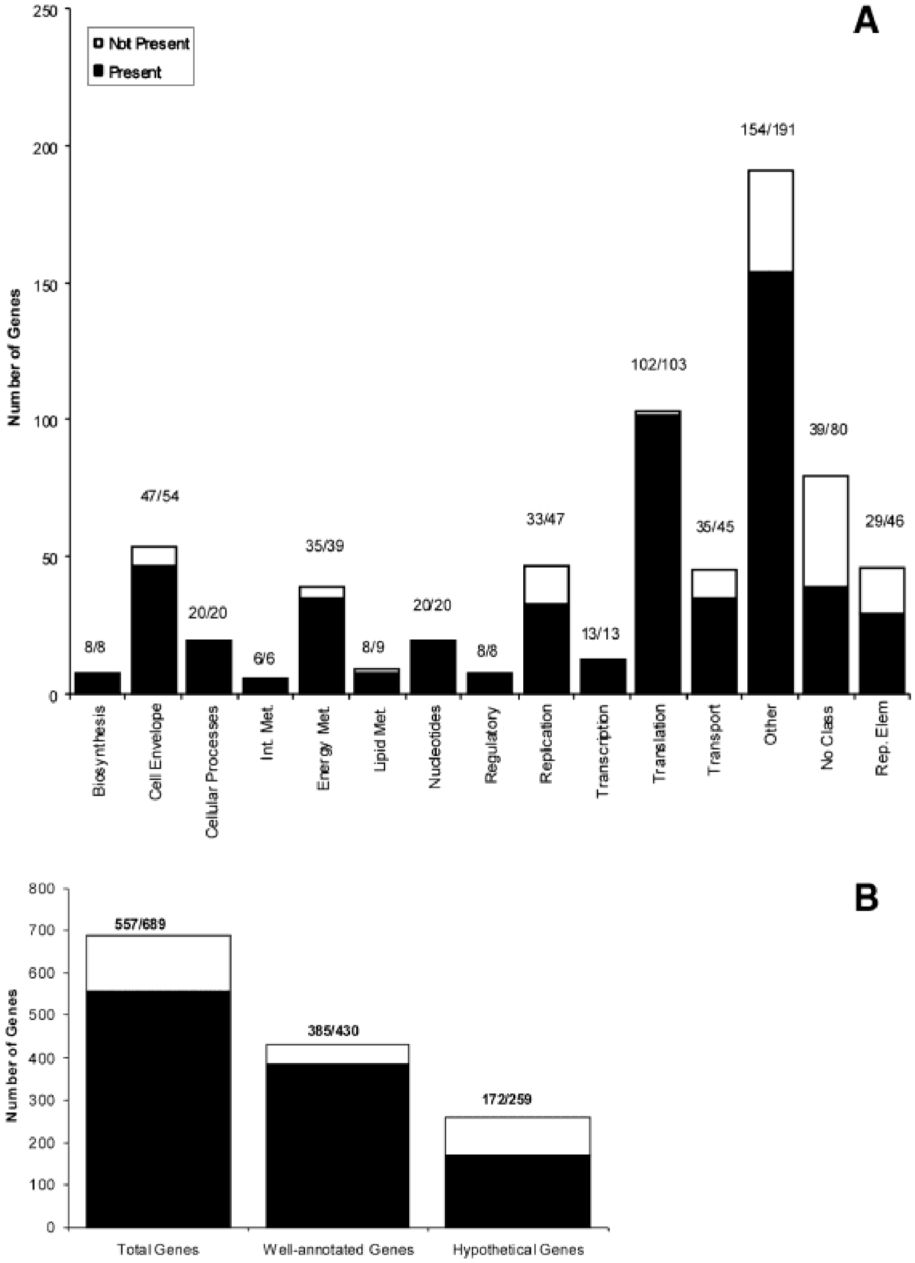

Figure 3. (A) Functional category breakdown of detected proteins, as assigned in [11, 12]. The number found/total predicted is shown over each column. (B) Detection rates of proteins by annotation strength, labeled as in (A).

In addition, 36 of the 557 ORFs that we detected have supporting evidence derived solely from degenerate peptides (marked with a * in Table 2). That is to say, the observed spectra that support these ORFs match a peptide that is encoded in more than one ORF in the genome. The majority of these proteins (30/36) are derived from the multiple P1 cytadherence operon fragments that are scattered throughout the genome. It has been demonstrated that *M. pneumoniae* makes use of these repetitive sequences to effect antigenic switching through homologous recombination at the primary cytadherence locus [19]. This makes it difficult to determine whether more than one of the P1 operons is actually expressed, but we can not rule it out. Despite starting from a single colony expansion, we also may have been sampling from a mixed population in which some members expressed one version of the P1 operon while other members expressed a variant. Therefore, we have included the detection of degenerate repetitive-element derived ORFs in our total, although it is most likely that only one or a few versions of the P1 operon are expressed at any given time. However, we did observe evidence that more than one cytadherence operon set was being expressed in the culture we studied (data available in web supplement).

Table 4. Proteogenomic mapping novel findings

New features

	Start	Stop	Frame	Peptides	BLAST	Notes
1	52515	52399	REV	1	N	
2	77313	77594	FWD	6	N	
3	135094	135360	FWD	1	N	
4	100609	100493	REV	1	N	Start codon would be TTG, overlapping another ORF
5	167508	167735	FWD	1	N	
6	207448	207717	FWD	2	N	
7	250021	250293	FWD	2	MG	
8	415161	415295	FWD	5	N	
9	~415289	415489	FWD	5	N	Start would be GTG
10	~415490	416032	FWD	7	N	No observable start codon
11	536168	535005	REV	2	MPN 436	
12	579389	579105	REV	18	MG	
13	592479	592201	REV	3	MG	
14	640277	640017	REV	1	IleRS-MG	Probably frameshift
15	794318	793737	REV	3	N	
16	796591	796418	REV	3	N	

BLAST result codes:
N, no significant similarity to any protein (E value < = 0.01)
MG, similar to a hypothetical protein in *M. genitalium*
MPN 436, homolog to MPN 436 (also detected in the survey)
IleRS-MG, homologous to the *C*-terminal portion of IleRS from *M. genitalium*

N-Terminal extensions

	Gene	New start	Old start	Frame	New start	Notes
1	MPN069	85227	85066	REV	TTG	Ribosomal protein L33 type 2
2	MPN101	130406	130466	FWD	NON	(Cytadherence fragment)
3	MPN111	144997	145021	FWD	NON	Conserved hypothetical protein
4	MPN128	166282	166483	FWD	TTG	(Cytadherence fragment)
5	MPN131	169954	170068	FWD	NON	(Cytadherence fragment)
6	MPN144	190273	190621	FWD	TTG	hypothetical protein
7	MPN148	195539	195875	FWD	TTG	Conserved hypothetical protein
8	MPN163	217150	217198	FWD	TTG	Conserved hypothetical protein
9	MPN367	437377	437563	FWD	NON	(Cytadherence fragment)
10	MPN388	465434	465176	REV	ATG	Conserved hypothetical protein
11	MPN412	496535	496634	FWD	GTG	Conserved hypothetical protein
12	MPN462	565135	565237	FWD	GTG	KtrA, Na^+, K^+ uptake
13	MPN464	566207	566891	FWD	TTG	(Cytadherence fragment)
14	MPN485	590475	589980	REV	TTG	Species specific lipoprotein
15	MPN509	622498	621844	REV	TTG	Membrane export protein family
16	MPN569	691880	691742	REV	TTG	Predicted metalloenzyme
17	MPN591	714283	713908	REV	ATG	Conserved hypothetical protein
18	MPN634	760322	760403	FWD	NON	Conserved hypothetical protein
19	MPN664	788141	787982	REV	GTG	hypothetical protein degV2

Deletions

	Gene	Start	Stop	Frame	Notes
1	MPN091	113838	114254	FWD	P1 fragment overlap
2	MPN206	249933	250274	REV	New orf detected here instead
3	MPN371	443552	444187	REV	P1 fragment overlap
4	MPN441	535468	535776	FWD	New orf detected here instead
5	MPN465	568645	569244	REV	P1 fragment overlap
6	MPN486	589922	590365	FWD	N-term ext of MPN485 overlap

Of the peptides we observed, roughly half were completely tryptic, with the vast majority of the remainder being at least "half-tryptic" (*i.e.*, ending in K or R but the residue preceding the cleavage was not K or R, and *vice versa*). This result is most probably due to the presence of a combination of nonspecific trypsin activity and endogenous *Mycoplasma* proteases. However, it directly illustrates the utility of searching mass spectra with no enzyme specificity with a gain in identified mass spectra at the cost of computing time. As noted in the Methods section, our criteria for accepting a non-tryptic peptide were stricter than that for accepting a completely tryptic peptide.

We were also able to detect one phosphoprotein through alternative search strategies of the data. We detected a peptide NH_2-SIINLMSLGIK-COOH from the HPr Phosphocarrier protein (MPN223) that was phosphorylated on the first serine residue. This interpretation was strengthened by our detection of the HPr(Ser) kinase (Genbank ID 13507962) in our proteomic survey [20, 21]. Although the SWISS-PROT (P75548) entry for this protein does not report phosphorylation of this residue, the homologous residue in *B. subtilis*, a Gram-positive relative is reported to be phosphorylated. We also observed the corresponding peptide in an unphosphorylated state (data available in web supplement). We were unable to observe phosphopeptides from HMW1 or HMW2 which are known to be phosphorylated *in vivo* [22].

As a "common-sense" check of the survey, we looked for any well-characterized proteins that we thought should be essential for viability but remained undetected. Of the "named" genes among the ORFs not detected (Table 3), we thought that dnaE (DNA primase – MPN 014), rpmG (ribosomal protein L33 – MPN 471), and arcA (arginine deiminase – MPN 304 and MPN 305) would likely be essential for viability. For each of these, we detected a homolog that could serve to carry out the function of the gene. For instance, MPN014 was annotated as DNA primase – a seemingly required protein – in the original annotation, but it is noted that *M. pneumoniae* has a second copy of DNA primase that was detected in this experiment, MPN 353. We also note that the original annotators of the genome noticed this duplication, and MPN 353 has also been detected previously [8, 12]. The absence of MPN 014 also agrees with the revision of is annotation to "conserved hypothetical protein" by Bork and co-workers [11] and our much lower detection rate for proteins that fall into this class (see above). *M. pneumoniae* also seems to have two copies of rpmG, as pointed out in [11]. We did detect the ribosomal protein L33 Type 2 (MPN 069). This protein is also more closely related to the *B. subtilis* ortholog of ribosomal protein L33 than MPN 471 (BLAST e-value of 3×10^{-5} *vs.* 0.007), and is located in an orthologous operon to its *B. subtilis* counterpart. We therefore believe that the majority of ribosomes include the L33 Type 2 protein rather than the L33 encoded by MPN 471, based on abundance detection. Finally, the copy of arcA that we failed to detect is actually broken down into two fragments of the gene that may not be functional as the intact version. We were able to detect a single, full-length ORF that corresponds to arcA functionality, MPN 560. This also corresponds with a previous finding [8]. Therefore, we believe that we have detected a set of proteins consistent with the current biological knowledge of required functionalities for a self-sustaining organism.

4 Discussion

Mass spectrometry provides an independent and complementary means of protein detection than inference from genomic sequence. Detection of a protein with mass spectrometry allows one to remove the "hypothetical" tag associated with many currently annotated ORFs in biological databases. Paired with multidimensional separations, it is an extremely powerful aid in biological studies of complex mixtures of proteins. Here, we have coupled mass spectrometry with several new data mining tools to generate an independent model of the ORFs in the small bacterium *M. pneumoniae*. We have developed proteogenomic mapping as an automated computational and graphical method for representing mass spectral data in the context of a corresponding genome. It is extremely useful for the elucidation of the primary ORF structure of the genome of an organism, and has demonstrated a high correlation with existing annotation methods despite its orthogonal approach. Proteogenomic mapping quickly enables discovery of new ORFs, validation of existing ORFs, modifications to existing annotation architecture, and discovery of discrepancies between existing annotation and mass spectral data. Feedback from this method might be used to refine existing gene prediction algorithms for more accuracy. This method provides a natural complement and enhancement to traditional sequence-based annotation methods.

In the course of this work, we have determined the most complete proteome to date for a single organism on a percent-wise basis. Of course, the relatively low complexity of *M. pneumoniae* has greatly aided our coverage. As with genome sequencing and crystalography, it is very important to push closer to 100% coverage (and methods to assess coverage) as this ultimately allows us to build better system models and make firmer statements about the absence of molecules. We consider this system to be a valuable test bed for technology that will be widely applicable to other organisms whose genomic sequence has been determined. We chose *M. pneumoniae* because we believed the genome structure was well understood after six years of study and two annotation efforts. Again, we should stress that our strain differed slightly from the sequenced strain, and some new features detected may be strain-specific. However, given the high degree of overlap at the protein level and the fact that all our analysis was done based on the sequenced strain's genome, we feel confident that this approach adds value to *M. pneumoniae* genome annotation. Given the number of new discoveries and modifications for the small genome of *M. pneumoniae* (816 kb), we anticipate that similar efforts in organisms with larger, less well-studied genomes will yield even more revelations about their genome structures. It would be reasonable in terms of time, scale, and cost to apply this technique to all currently sequenced bacteria. It would probably require that more environmental conditions be explored for bacteria with more complex gene regulation, but the experiment that we have described here could be completed in less than one month using a single mass spectrometer, or possibly even more rapidly using the "MuDPIT" method of Yates and colleagues [7]. Moreover, it offers rapid, large-scale contributions to the fundamental goal of having the most comprehensive understanding of an organismal system as possible.

Somewhat analogous data may be obtained by sequencing a library of cDNA clones of mRNA isolated from an organism by random priming [23]. While it has been noted that detection of an mRNA species is not proof-positive evidence for a protein product [24], such a study of RNA offers advantages for the determination of RNA termini, splice junctions, untranslated RNAs, and quantitative processes at the RNA level that are independent of the protein level. However, complex cellular processes governing the half-lives and bioavailabilities of proteins are not represented in nucleic acid-based studies. As well, cDNA sequencing alone is not capable of detecting post-translational modifications, and previous studies have indicated the utility in searching for such events [25]. Without specific enrichment strategies, we have detected at least one phosphorylation event in *M. pneumoniae*. Direct observation of this phosphorylation event in the living cell has not been reported before, and only a limited number of phosphoproteins (perhaps up to 9) are believed to exist in *M. pneumoniae* [22]. That we found only a single phosphoprotein without specific enrichment may be reflective of our coverage rate in general.

The prospects for the future of proteomics are also encouraging. In the current study, we utilized two dimensions of separation to partition peptides from hundreds of proteins prior to mass spectrometry. By introducing more orthogonal separation techniques, we can expect to increase the capacity to thousands or tens-of-thousands of proteins, mainly at the expense of time. However, the automated high-throughput nature of the current generation of mass spectrometers makes processing of large numbers of samples easy. In addition, quantitation methods for proteomic-scale data continue to be developed that may offer an informative complement to data from the now ubiquitous microarray experiment ([26]; Leptos *et al.*, in preparation).

An important problem for proteomics remains in finding methods to increase protein coverage. While we consider our coverage rate to be fairly good (31% amino acid sequence coverage for detected ORFs), it could certainly be desirable to increase this figure. The inherent biases in ionization efficiency and protein separation technology need to be addressed, and instruments with increased dynamic range would also help to detect low-abundance species in mixtures where a few dominant analytes may be present in excess (*i.e.*, albumin in serum). Improved coverage would strengthen our ability to draw a comprehensive proteogenomic map and derive an ORF model from peptide data alone. It would also strengthen our ability to detect post-translational modifications in a "shotgun" manner, as described by MacCoss *et al.* [27].

In summary, we have developed novel methods and applied existing technologies to further elucidate the genomic and proteomic structures of the small bacterium, *M. pneumoniae*. We observed that current methods for genome annotation can be significantly validated, enhanced, and complemented by the addition of proteomic data. The particular techniques used here are relatively inexpensive and therefore available to a wide range of researchers. We expect that these techniques will be extended to other organisms with increasing scale in the near future, yielding similar sets of novel discoveries.

We would like to thank Dr. Makoto Miyata of Osaka City University, Japan, for his gift of M. pneumoniae strains and expertise in their cultivation. We would also like to thank the staff of the Biopolymer Facility at Harvard Medical School (especially Mr. Eric Spooner) for use of their instrumentation and advice. This work was generously supported by United States Department of Energy Grant DE-FG02-87ER60565 to GMC and grants from the National Institutes of Health and the Rowland Institute for Science to HCB.

Received April 25, 2003
Revised May 29, 2003
Accepted May 29, 2003

5 References

[1] http://wit.integratedgenomics.com/GOLD/completegenomes.html (as of April 8, 2002).

[2] Lukashin, A. V., Borodovsky, M., *Nucleic Acids Res.* 1998, *26*, 1107–1115.

[3] Altschul, S. F. *et al.*, *J. Mol. Biol.* 1990, *215*, 403–410.

[4] Delcher, A. L. *et al.*, *Nucleic Acids Res.* 1999, *27*, 4636–4641.

[5] Peng, J., Gygi, S. P., *J. Mass Spectrom.* 2001, *36*, 1083–1091.

[6] Smith, R. D. *et al.*, *Proteomics* 2002, *2*, 513–523.

[7] Washburn, M. P., Wolters, D., Yates III, J. R., *Nat Biotechnol* 2001, *19*, 242–247.

[8] Ueberle, B., Frank, R., Herrmann, R., *Proteomics* 2002, *2*, 754–764.

[9] Lipton, M. S. *et al.*, *Proc. Natl. Acad. Sci. USA* 2002, *99*, 11049–11054.

[10] Razin, S., Yogev, D., Naot, Y., *Microbiol. Mol. Biol. Rev.* 1998, *62*, 1094–1156.

[11] Dandekar, T. *et al.*, *Nucleic Acids Res.* 2000, *28*, 3278–3288.

[12] Himmelreich, R. *et al.*, *Nucleic Acids Res.* 1996, *24*, 4420–4449.

[13] Aluotto, B. B. *et al.*, *Int. J.* 1970, *20*, 35–58.

[14] Gygi, M. P., Licklider, L. J., Peng, J., Gygi, S. P., in: Simpson, R. (Ed.), *Protein Analysis: A Laboratory Manual*, Cold Spring Harbor Press, New York 2002.

[15] Link, A. J. *et al.*, *Nat. Biotechnol.* 1999, *17*, 676–682.

[16] Eng, J. K., McCormack, A. L., Yates, J. R., *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[17] Peng, J. *et al.*, *J. Proteome Res.* 2002, *1*, 47–54.

[18] ftp://ftp.ncbi.nih.gov/blast/db/nr.Z.

[19] Dorigo-Zetsma, J. W. *et al.*, *Infect. Immun.* 2001, *69*, 5612–5618.

[20] Allen, G. S. *et al.*, *J. Mol. Biol.* 2003, *326*, 1203–1217.

[21] Steinhauer, K. *et al.*, *Microbiology* 2002, *148*, 3277–3284.

[22] Dirksen, L. B., Krebes, K. A., Krause, D. C., *J. Bacteriol.* 1994, *176*, 7499–7505.

[23] Adams, M. D. *et al.*, *Nat. Genet.* 1993, *4*, 373–380.

[24] Gygi, S. P. *et al.*, *Mol. Cell Biol.* 1999, *19*, 1720–1730.

[25] Link, A. J., Robison, K., Church, G. M., *Electrophoresis* 1997, *18*, 1259–1313.

[26] Gygi, S. P. *et al.*, *Nat. Biotechnol.* 1999, *17*, 994–999.

[27] MacCoss, M. J. *et al.*, *Proc. Natl. Acad. Sci. USA* 2002, *99*, 7900–7905.