Proteome Organization in a Genome-Reduced Bacterium

Sebastian Kühner, *et al.*

Science **326**, 1235 (2009);
DOI: 10.1126/science.1176343

opened the door to large-scale screens. At the same time, limitations of this approach are increasingly apparent, such as the induction of off-target effects that complicate genome-wide screens in particular (*29*, *30*) and the inability to completely switch off gene expression. When similar small interfering RNA screens are conducted independently in mammalian cells, the lack of concordance between them is an additional complicating factor (*31*, *32*). Finally, mammals are rather robust in their tolerance to partial loss of gene function: Haploinsufficiency appears to be the exception rather than the rule, because inactivation of one gene copy, as in heterozygous knockout mice, rarely leads to severe phenotypes.

Although we have focused on host-pathogen biology, similar screens could in principle be applied to any phenotype that can be recognized in a population of mutant cells, such as modulation of a genetically encoded reporter. In the future, haploid genetic screens could be used to generate comprehensive compendia of host factors that are used by different pathogens and may yield new strategies to combat infectious disease. In conclusion, the haploid genetic screens described here expand mutagenesis-based screens in model organisms by providing a window on disease-associated molecular networks that can be studied in cultured human cells.

References and Notes

1. H. J. Muller, *Science* **66**, 84 (1927).
2. A. L. Brass *et al.*, *Science* **319**, 921 (2008); published online 10 January 2008 (10.1126/science.1152725).
3. J. A. Philips, E. J. Rubin, N. Perrimon, *Science* **309**, 1251 (2005); published online 14 July 2005 (10.1126/science.1116006).
4. L. Hao *et al.*, *Nature* **454**, 890 (2008).
5. R. Salomon, R. G. Webster, *Cell* **136**, 402 (2009).
6. A. Moscona, *N. Engl. J. Med.* **360**, 953 (2009).
7. M. Kotecki, P. S. Reddy, B. H. Cochran, *Exp. Cell Res.* **252**, 273 (1999).
8. Materials and methods are available as supporting material on *Science* Online.
9. S. Nagata, *Cell* **88**, 355 (1997).
10. B. Luo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 20380 (2008).
11. F. Cong *et al.*, *Mol. Cell* **6**, 1413 (2000).
12. M. Lara-Tejero, J. E. Galán, *Science* **290**, 354 (2000).
13. D. Nesic, Y. Hsu, C. E. Stebbins, *Nature* **429**, 429 (2004).
14. M. Wistrand, L. Kall, E. L. L. Sonnhammer, *Protein Sci.* **15**, 509 (2006).
15. M. Miyaji *et al.*, *J. Exp. Med.* **202**, 249 (2005).
16. L. Guerra *et al.*, *Cell. Microbiol.* **7**, 921 (2005).
17. B. Wollscheid *et al.*, *Nat. Biotechnol.* **27**, 378 (2009).
18. H. Sprong *et al.*, *Mol. Biol. Cell* **14**, 3482 (2003).
19. R. J. Collier, *Toxicon* **39**, 1793 (2001).
20. S. Liu, G. T. Milne, J. G. Kuremsky, G. R. Fink, S. H. Leppla, *Mol. Cell. Biol.* **24**, 9487 (2004).
21. J. C. Milne, S. R. Blanke, P. C. Hanna, R. J. Collier, *Mol. Microbiol.* **15**, 661 (1995).
22. H. M. Scobie, G. J. A. Rainey, K. A. Bradley, J. A. T. Young, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5170 (2003).
23. J. G. Naglich, J. E. Metherall, D. W. Russell, L. Eidels, *Cell* **69**, 1051 (1992).
24. L. C. Mattheakis, W. H. Shen, R. J. Collier, *Mol. Cell. Biol.* **12**, 4026 (1992).
25. S. Liu, S. H. Leppla, *Mol. Cell* **12**, 603 (2003).
26. J. Y. Chen, J. W. Bodley, *J. Biol. Chem.* **263**, 11692 (1988).
27. M. E. Hillenmeyer *et al.*, *Science* **320**, 362 (2008).
28. S. L. Forsburg, *Nat. Rev. Genet.* **2**, 659 (2001).
29. Y. Ma, A. Creanga, L. Lum, P. A. Beachy, *Nature* **443**, 359 (2006).
30. C. J. Echeverri *et al.*, *Nat. Methods* **3**, 777 (2006).
31. S. P. Goff, *Cell* **135**, 417 (2008).
32. F. D. Bushman *et al.*, *PLoS Pathog.* **5**, e1000437 (2009).

33. We thank D. Sabatini, S. Nijman, J. Roix, and J. Pruszak for discussion and critical review of the manuscript; C. Y. Wu and G. Fink for yeast deletion strains; J. Kaper for the CDT expression plasmid; J. Collier, R. Moon, and M. Wernig for plasmids; and E. Guillen for help with influenza infections. C.P.G. has a fellowship from Fundacao Ciencia Tecnologia, Portugal. T.R.B. was funded by the Kimmel Foundation and the Whitehead Institute Fellows Program. The Whitehead Institute has filed a patent on the application of gene-trap mutagenesis in haploid or near-haploid cells to identify human genes that affect cell phenotypes, including host factors used by pathogens.

Supporting Online Material

www.sciencemag.org/cgi/content/full/326/5957/1231/DC1
Materials and Methods
Figs. S1 to S9
References

10 July 2009; accepted 5 October 2009
10.1126/science.1178955

Proteome Organization in a Genome-Reduced Bacterium

Sebastian Kühner,[1]* Vera van Noort,[1]* Matthew J. Betts,[1] Alejandra Leo-Macias,[1] Claire Batisse,[1] Michaela Rode,[1] Takuji Yamada,[1] Tobias Maier,[2] Samuel Bader,[1] Pedro Beltran-Alvarez,[1] Daniel Castaño-Diez,[1] Wei-Hua Chen,[1] Damien Devos,[1] Marc Güell,[2] Tomas Norambuena,[3] Ines Racke,[1] Vladimir Rybin,[1] Alexander Schmidt,[4] Eva Yus,[2] Ruedi Aebersold,[4] Richard Herrmann,[5] Bettina Böttcher,[1]† Achilleas S. Frangakis,[1] Robert B. Russell,[1] Luis Serrano,[2,6] Peer Bork,[1]‡ Anne-Claude Gavin[1]‡

The genome of *Mycoplasma pneumoniae* is among the smallest found in self-replicating organisms. To study the basic principles of bacterial proteome organization, we used tandem affinity purification–mass spectrometry (TAP-MS) in a proteome-wide screen. The analysis revealed 62 homomultimeric and 116 heteromultimeric soluble protein complexes, of which the majority are novel. About a third of the heteromultimeric complexes show higher levels of proteome organization, including assembly into larger, multiprotein complex entities, suggesting sequential steps in biological processes, and extensive sharing of components, implying protein multifunctionality. Incorporation of structural models for 484 proteins, single-particle electron microscopy, and cellular electron tomograms provided supporting structural details for this proteome organization. The data set provides a blueprint of the minimal cellular machinery required for life.

Biological function arises in part from the concerted actions of interacting proteins that assemble into protein complexes and networks. Protein complexes are the first level of cellular proteome organization: functional and structural units, often termed molecular machines, that participate in all major cellular processes. Complexes are also highly dynamic in the sense that their organization and composition vary in time and space (*1*), and they interact to form higher level networks; this property is central to whole-cell functioning. However, general rules concerning protein complex assembly and dynamics remain elusive.

The combination of affinity purification with mass spectrometry (MS) (*2*) has been applied to several organisms to provide a growing repertoire of molecular machines. Genome-wide screens in *Saccharomyces cerevisiae* (*3*–*5*) captured discrete, dynamic proteome organization and revealed higher-order assemblies with direct connections between complexes and frequent sharing of common components. To date these exhaustive analyses have been applied only in yeast. In bacteria, genome-wide yeast two-hybrid analyses have been reported (*6*, *7*), but only a few biochemical analyses on selected sets of complexes are available (*8*–*11*). The understanding of proteome organization in these organisms concerns thus the binary interaction networks.

Here, we report a genome-scale analysis of protein complexes in the bacterium *Mycoplasma pneumoniae*, a human pathogen that causes atypical

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany. [2]Centro Regulacion Genomica–Universidad Pompeu Fabra, Dr Aiguader 88, 08003 Barcelona, Spain. [3]Pontificia Universidad Catolica de Chile, Alameda 340, Santiago, Chile. [4]ETH (Eidgenössische Technische Hochschule) Zürich, Wolfgang-Pauli-Strasse 16, 8093 Zürich, Switzerland; Faculty of Science, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland, and Institute for Systems Biology, Seattle, WA 98013, USA. [5]ZMBH (Zentrum für Molekulare Biologie der Universität Heidelberg), Im Neuenheimer Feld 282, 69120 Heidelberg, Germany. [6]ICREA (Institució Catalana de Recerca i Estudis Avançats), 08010 Barcelona, Spain.

*These authors contributed equally to this work.
†Present address: University of Edinburgh, Kings Buildings, Mayfield Road, Edinburgh EH9 3JR.
‡To whom correspondence should be addressed. E-mail: gavin@embl.de (A.-C.G.); bork@embl.de (P.B.)

pneumonia (*12*). This self-replicating organism has one of the smallest known genomes (689 protein-encoding genes) (*13*, *14*), making it an ideal model organism for the investigation of absolute essentiality (*15*). This analysis and the integration with other consistently derived large-scale data sets provide a blueprint of the proteome organization in a minimal cell and reveal principles underlying adaptation to a reduced genome.

Genome-wide screen for protein complexes in *M. pneumoniae*. We adapted the tandem affinity purification–mass spectrometry (TAP-MS) protocol (*2*) to *M. pneumoniae* M129 (Fig. 1) (*16*). We processed all 689 *M. pneumoniae* protein-coding genes, of which 617 were successfully cloned [90% of the genome (*14*)]. With use of a transposon-based expression system, we constructed a total of 456 *M. pneumoniae* strains. They carry a stable genomic integration of carboxy-terminal TAP fusions under transcriptional control of the *M. pneumoniae* clpB (mpn531) promoter. From this collection, all 352 individual strains expressing soluble TAP fusions were grown to confluence in 2 liters of adherent culture, leading to 212 successful purifications. The components of the purified complex were separated by denaturing gel electrophoresis, and individual bands were trypsin-digested and analyzed by MS (table S1). We processed a total of 10,447 MS samples and identified proteins by using a new approach that integrates the Mascot (*17*) and Aldente (*18*) search algorithms (*19*). This increased the identification of known complex components by ~20% compared with either method alone (fig. S1, A and B). The procedure also scores the quality of individual identifications by considering all peptide profiles that we observed for each protein, including our purification data set and a PeptideAtlas, a comprehensive set of tryptic peptides (*20*) measured with Fourier transform–MS from whole *M. pneumoniae* lysates (table S2). We removed protein identifications with overlapping peptide profiles (3%) (fig. S1, C and D). When applied to the entire purification data set, this approach uncovered 411 distinct proteins from 5899 identifications (table S2).

The 411 proteins identified with 212 tagged proteins correspond to 60% of the annotated open reading frames (ORFs) and 85% of the predicted soluble proteome (fig. S2). They cover all cellular functions, although low abundant, small, or trans-membrane segment–containing proteins are notably underrepresented (fig. S2). Membrane proteins purification requires separate biochemical protocols, so they were not included in this screen. The proportion of new proteins identified per purification dropped asymptotically as the screen progressed, implying that the procedure was near saturation (fig. S3). This entails recurring protein complex retrieval through reverse tagging and is important both to confirm novel interactions and to identify dynamic complexes (*3*).

To define complexes in a quantitative way, we first calculated socio-affinity indices that measure the frequency with which pairs of proteins were found associated in our set of biochemical purifications (*3*, *16*). We improved the concept by integrating predicted interactions from the STRING database (*21*) and the relative abundance of a given prey when associated with different baits (i.e., across different purifications) (*22*). We used the MS scores that measure the probability for a peptide mass fingerprint to characterize each protein based on spectral counting. A reduced score for a prey in a purification, when compared to the same prey in other purifications, reflects identifications by a smaller number of peptides (lower spectral counts); it is indicative of a spurious interaction and is therefore down-weighted (fig. S4A). We applied this new scoring scheme to the entire data set and calculated a list of 10,083 interactions. A cut-off was defined at an accuracy, that is, a fraction of true interactions (*23*), of more than 80%, which gave a set of 1058 high-confidence interactions (fig. S4, B and C; also table S3). We also measured the overall experimental reproducibility on a set of 18 experiments that we performed twice; duplicates included growth of adherent cultures, biochemical purifications, and MS analyses (*16*). For protein pairs with socio-affinity scores ≥0.8, the overall reproducibility is 73%; for those scoring below it is 43% ($P = 10^{-13}$, χ^2 test). For comparison, the reproducibility calculated on the duplicated MS measurements of 72 MS samples is 97%. We then applied cluster analysis by using a procedure called clique percolation that allows proteins to be part of different complexes. We varied the clustering parameters over reasonable ranges. The best conditions in terms of coverage (see below) generated a collection of 116 heteromultimeric complexes. They are organized into densely (>one link) and loosely interconnected (one link) components we called "core" and "attachment," respectively (fig. S4D and table S4). Generally, *M. pneumoniae* proteins within complexes and cores are more often co-expressed (*24*) and conserved between species than average; proteins within complexes appear on average in 244 species compared with 173 for the entire proteome (median = 190). Comparison to a set of 31 known complexes, described in other species (table S5), revealed a coverage

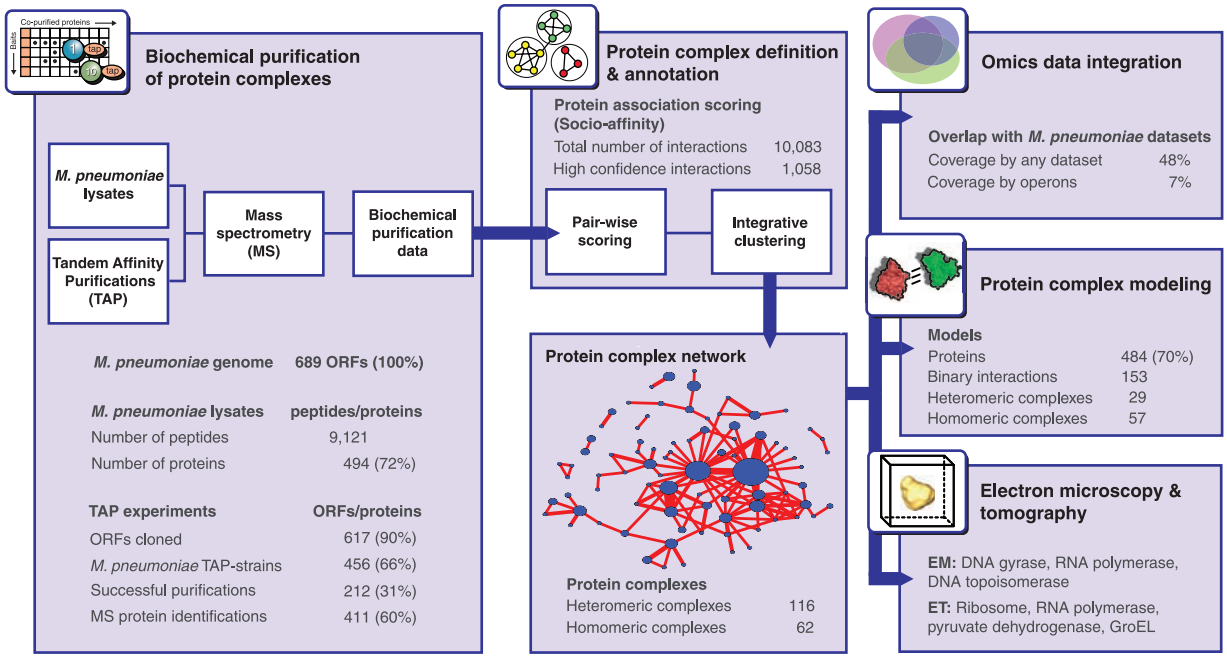

Fig. 1. Synopsis of the genome-wide screen of complexes in *M. pneumoniae*.

of 61%, which is similar to results from previous screens in yeast and *Escherichia coli* (coverage ~60%) (*3*, *4*, *9*, *25*).

Systematic detection of homomultimeric protein complexes. The TAP fusions were expressed from exogenous loci and promoter and are therefore present together with the untagged wild-type allele. It was thus common to observe both TAP-tagged and -untagged versions of the bait in the same purification, which is an indication of homomultimerization (fig. S5). Careful scrutiny of the purification data set revealed evidence for 62 homomultimeric complexes (table S4) covering 62% of those previously seen either in *M. pneumoniae* or in another species by orthology (table S5). Fourteen homomultimeric complexes were previously unknown, and for 12 of these we could find supporting structural evidence from homologs of known structure (*26*) (table S6). An example is Mpn266, a protein of previously unknown function that we found associated to RNA polymerase (complex 49, table S4) as a dimer. Its binding to the polymerase is consistent with its similarity to SpxA, an RNA polymerase–binding protein that regulates transcription initiation in Gram-positive bacteria (*27*, *28*). Comparative modeling of structure and single-particle electron microscopy (fig. S6) (*16*) show that *M. pneumoniae* RNA polymerase resembles that of *Thermus aquaticus* (*29*) with the exception of a substantially bigger stalk at the position of the sigma factor, RpoD (Mpn352), consistent with *M. pneumoniae* RpoD being 200 amino acids longer than its *T. aquaticus* ortholog (fig. S7A). The models also further support the idea that each Mpn266 in the dimer binds one of the two α subunits of the polymerase, as do other transcription factors (fig. S7A).

From the number of baits used (212) and from the effectiveness of the method in recovering known complexes (62% coverage), we estimate that as many as 47% of all soluble proteins form homomultimers in *M. pneumoniae*. This is in agreement with a recent analysis of more than 5000 protein structures (*30*). Lastly, considering both homo- and heteromultimers, almost 90% of soluble proteins were found to be part of at least one complex, a figure similar to values estimated in yeast (*3*, *4*). This further consolidates the view that exhaustive organization into complexes is a general property of proteomes in bacteria and eukaryotes.

Characteristics of *M. pneumoniae* protein complexes. Overall, more than half of the identified complexes were not previously described. We also found new components in previously known complexes: The data set contains 126 proteins with previously unknown or conflicting functional annotation. For example, complex membership identifies Mpn426, previously annotated as a P115 homolog, as the missing Smc (structural maintenance of chromosomes) DNA-binding subunit of the cohesin-like complex (complex 40, fig. S7B and table S4) (*28*). This complex also contains the adenosine triphosphate (ATP)–dependent protease Lon (Mpn332) that binds DNA and regulates chromosome replication (*31*). The observed physical association between Lon and Smc and the observation that Lon expression increases concomitant with Smc degradation at the onset of the stationary phase (fig. S7B) (*28*) suggest that Smc might be a target of this protease. The existence of a native complex including Lon, ScpA (Mpn300), and P115 is further supported by the observation that these three proteins co-elute during gel filtration chromatography (fig. S7B). We also identified known eukaryotic complexes such as those including several glycolytic enzymes (GEs) that have been discovered at eukaryotic plasma membrane, where they locally produce ATP (table S5). We observed similar assemblies in *M. pneumoniae* (complexes 12 and 45; table S4), which suggests that this function is conserved in bacteria.

Comparison of methods for estimation of proteome organization. We overlaid the protein complex data with complementary large-scale data sets that have been previously used to deduce physical interactions (Fig. 2A). Only 48% of the TAP interactions within complexes were found in any existing data set; 359 associations were only identified by TAP-MS (Fig. 2A). Even in the worst-case scenario, where we consider the upper limit of the estimated false-positives rate (20% = 100% to 80% accuracy) and assume that false positives are completely excluded from the other data sets, we estimate at least 220 previously unknown true associations were identified here. Overlap with interactions inferred from genome organization or gene expression was particularly low: Only 7% of the high-confidence interactions are between gene products from the same operon, and only 18% were consistently

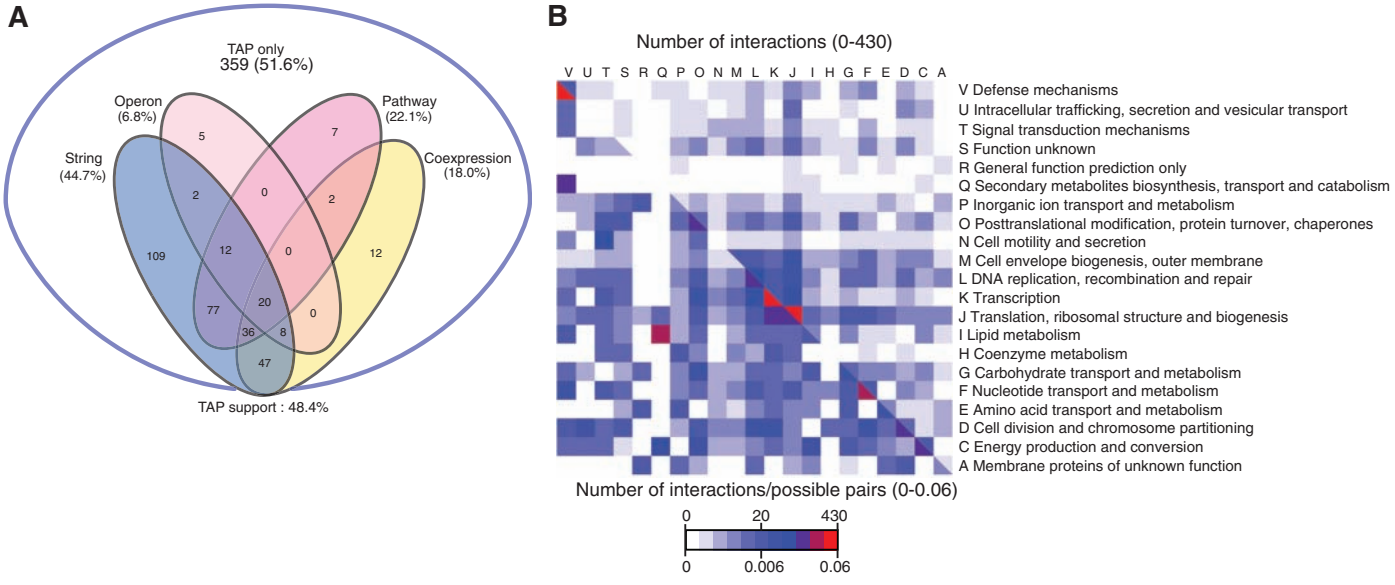

Fig. 2. Proteome organization is only partially reflected by other biological data sets. (**A**) General overlap between TAP and interactions inferred from other data sets: coexpression (*24*, *28*), operons (*24*), STRING (*21*), and pathways (*48*). Numbers refer to the interacting pairs within the different data sets. The fraction of TAP interactions that cluster into complexes and are covered by other data sets is given between brackets. For TAP-interacting protein pairs the cutoff was set at 80% accuracy. Cutoffs for other data sets were optimized for coverage (accuracies from 40 to 100%). (**B**) Frequent functional cross-talk in the protein complex data set. All proteins within high confidence pairs were functionally annotated according to the COG (Clusters of Orthologous Groups of Proteins) database (*49*). Boxed areas are colored proportionally to the number of interactions linking two functional classes. The scales represent the total (top) and normalized (bottom) number of interactions (*23*). Category Q (secondary metabolites) contains only two proteins. The category most frequently linked is J (translation) with itself; however, it contains the highest number of proteins. The highest proportion of interactions is between proteins within category K (transcription).

coexpressed (*24*). This implies that temporal or conditional regulation of complex formation is analogous to that for eukaryotes, in which different components are expressed at different times (*1*). For example, the four known subunits of the RNA polymerase are in three operons, and their transcription profiles correlate with two different gene expression groups along the growth curve (*24*, *28*). With current knowledge, only a small fraction of proteome organization can be inferred from analysis of the genomes or transcriptional data, making proteomics studies critical for understanding prokaryotic systems.

The *M. pneumoniae* protein complex network reveals substantial cross-talk. About a third of the heteromultimeric complexes in *M. pneumoniae* have extensive physical interconnections that suggest proteins participate in different cellular processes (Fig. 2B). These reflect protein multifunctionality (see below) and organization into at least 35 larger assemblies, sometimes hinting at physical, possibly temporal, associations of sequential steps in biological processes (table S4). For example, we reconstituted major parts of the ribosome from the interaction screen and saw extensive cross-talk with RNA polymerase (Fig. 3A). This higher-level association was unaffected by ribonuclease (RNAse) and deoxyribonuclease (DNAse) treatments, which suggests that protein-protein rather than protein–nucleic acid interactions were involved (fig. S5). The TAP-MS data were consistent with gel filtration results showing that the RNA polymerase α subunit, RpoA (Mpn191), and the ribosomal protein RpsD (Mpn446) co-elute with high apparent molecular sizes (Fig. 3A). These observations are further supported by the genome organization, where the rpoA gene is localized in and co-regulated with a ribosomal operon (*24*). This network provides a molecular model for the coupling of transcription and translation proposed in bacteria (*32*) and the direct involvement of ribosomal proteins in transcriptional regulation (*33*). The same assembly also includes translational initiation factors InfA (Mpn187), InfB (Mpn155), and InfC (Mpn115), which are part of the 30*S* initiation complex, as well as elongation factors Tuf (Mpn665) and Tsf (Mpn631), suggesting that we have captured sequential steps in a pathway running from transcription to translation.

Functional reuse and modularity of protein complexes. Genome-wide screens in eukaryotes show that proteins often participate in more than one complex, an attribute that has been proposed to account for protein multifunctionality, pleiotropy, and moonlighting (*34*). We defined a multifunctionality index that measures the tendency of proteins to associate with more than one complex (*16*). This index is based on frequency with which pairs of proteins were found associated in our set of purifications and is insensitive to the clustering parameters. We found 156 multifunctional proteins (table S7), covering 54% of *M. pneumoniae* proteins that are currently known to be multifunctional in the literature (table S8). We also compared our results with a set of multifunctional enzymes that catalyze different enzymatic reactions (*28*), and the overlap was smaller (32%). Our analysis captured distinct mechanisms for

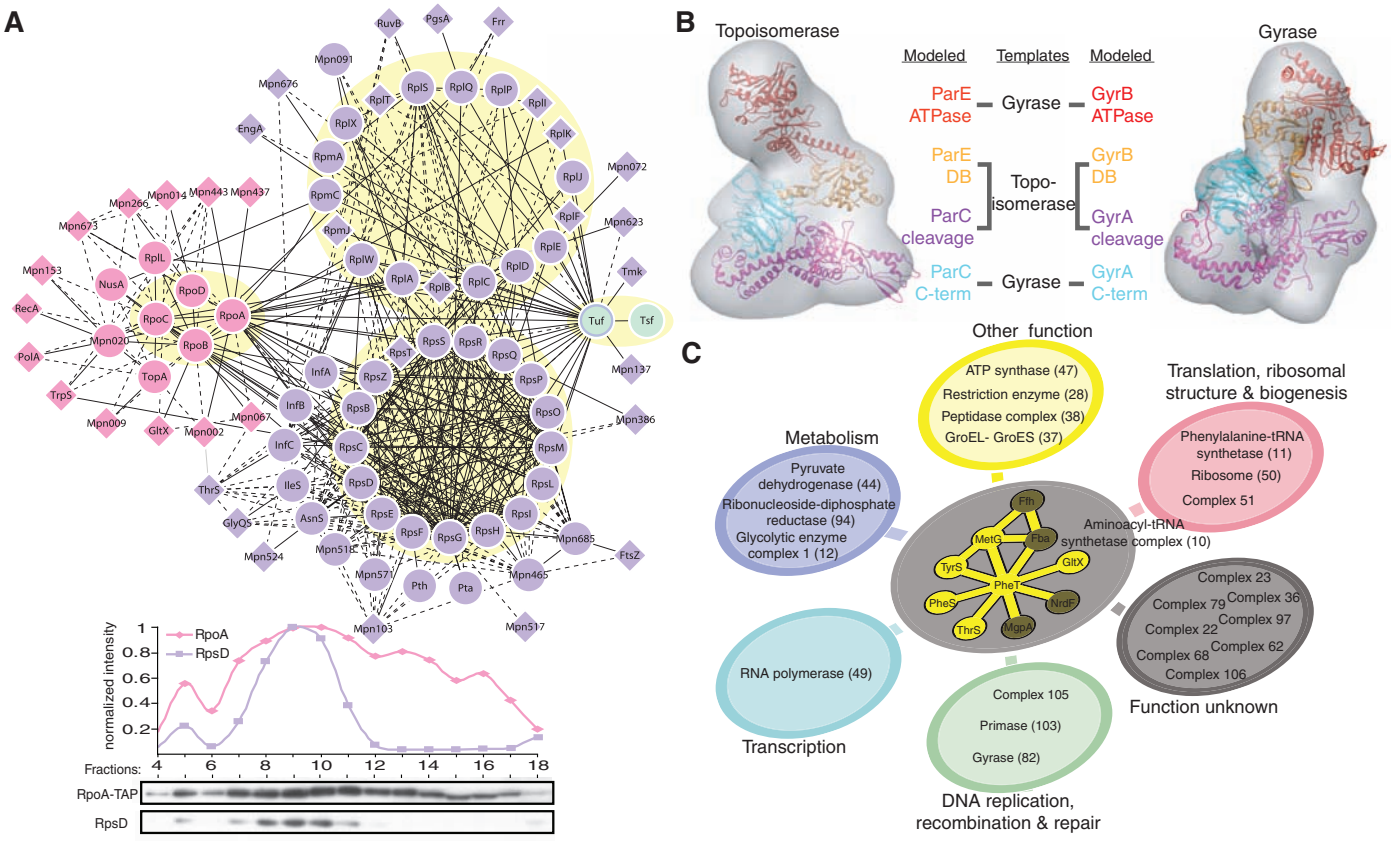

Fig. 3. Higher level of proteome organization. (**A**) The RNA polymerase–ribosome assembly. Core components are represented by circles, attachments by diamonds. The line attribute corresponds to socio-affinity indices: dashed lines, 0.5 to 0.86; plain lines, >0.86. Color code and shaded yellow circles around groups of proteins refer to individual complexes: RNA polymerase (pink), ribosome (purple), and translation elongation factor (green). The bottom graph shows that the ribosomal protein RpsD (23 kD) and the α subunit of the RNA polymerase, RpoA-TAP (57 kD), co-elute in high molecular weight fractions (MD range) during gel filtration chromatography. (**B**) DNA topoisomerase (diameter ~ 12 nm) is a heterodimer in bacteria: ParE (ATPase and DNA binding domains) and ParC (cleavage and C-terminal domains). The interaction between ParE-DNA–binding and ParC–cleavage domains was modeled by using yeast topoisomerase II as a template [Protein Data Bank (PDB) code 2rgr], and ParE-ATPase and ParC–C-terminal domains were modeled separately on structures of gyrase homologs (PDB 1kij and 1suu). All four domains were fitted into the electron microscopy density. Gyrase (~12 nm) is similarly split in bacteria into GyrA/GyrB, which are paralogs of ParE/ParC, and was modeled and fitted by using PDB 1bjt as a template for the GyrB-DNA–binding and GyrA-cleavage domains interaction. (**C**) Protein multifunctionality in *M. pneumoniae* illustrated with the AARS complexes.

multifunctionality that imply the combinatorial use of gene products in different contexts, for different functions.

For example, GyrA (Mpn004) is a component of the DNA gyrase complex that introduces negative supercoils into DNA, and ParE (Mpn122) is a member of the topoisomerase IV complex, which decatenates DNA (*35*). Besides well-documented interactions within their respective complexes (complexes 17 and 82, table S4), GyrA and ParE were also found to stably associate with each other (complex 102, table S4). Single-particle electron microscopy and comparative modeling (fig. S6) showed that DNA topoisomerase and DNA gyrase have related overall shapes, as expected from their functional similarity, and also support the notion that they might be able to interchange subunits (Fig. 3B). In eukaryotes, ParE and ParC (Mpn123) are fused into one single polypeptide. In bacteria, the possibility for the split ParE and ParC to contribute to different complexes might represent a parsimonious way of generating functional diversity and also robustness to mutations with a set of paralogous proteins.

Another example is a complex containing a cluster of five different aminoacyl transfer RNA (tRNA) synthetases (AARSs) (complex 10, Fig. 3C and table S4). In eukaryotes and archaea, AARSs form macromolecular complexes that improve aminoacylation efficiency by channeling substrates to ribosomes (*36*, *37*). These assemblies also act as reservoirs of AARSs that additionally exert a range of noncanonical regulatory functions in transcription, metabolism, and signaling (*38*). The existence in bacteria of big multi-AARS complexes is controversial; the most recent review advocates assembly in binary complexes that are functionally involved in tRNA metabolism and editing (*39*). Our results suggest that higher-order multi-AARS complexes might also exist in bacteria. We also found several AARSs in other complexes involved in functions as diverse as translation, transcription, DNA replication, and metabolism (Fig. 3C).

Structural anatomy of *M. pneumoniae*. Because of their small genome size, bacteria from the genus *Mycoplasma* have attracted attention as model organisms for structural genomics (*40*). We used these data to populate our protein complex network with structural information. Sequence

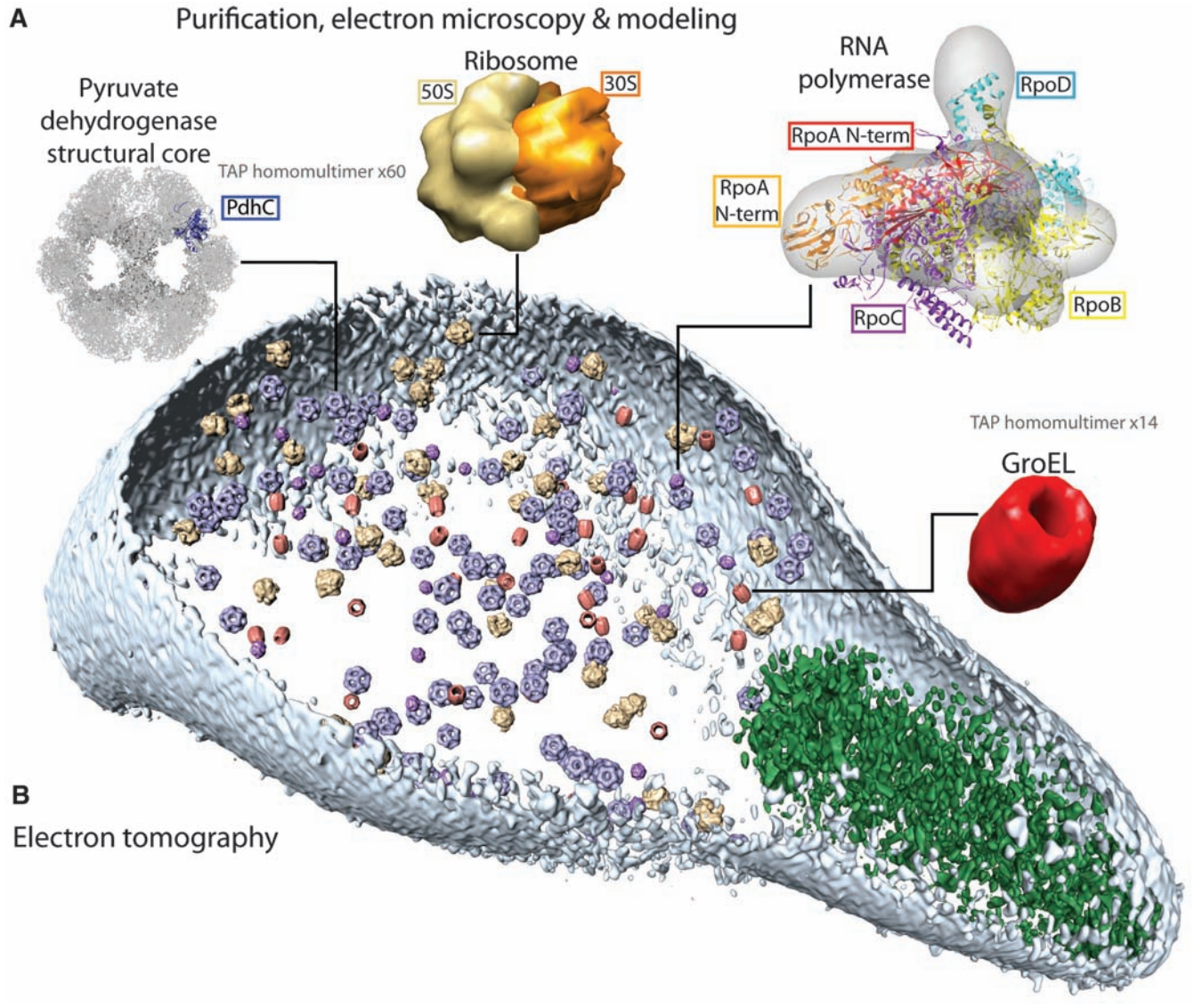

Fig. 4. From proteomics to the cell. By a combination of pattern recognition and classification algorithms, the following TAP-identified complexes from *M. pneumoniae*, matching to existing electron microscopy and x-ray and tomogram structures (**A**), were placed in a whole-cell tomogram (**B**): the structural core of pyruvate dehydrogenase in blue (~23 nm), the ribosome in yellow (~26 nm), RNA polymerase in purple (~17 nm), and GroEL homomultimer in red (~20 nm). Cell dimensions are ~300 nm by 700 nm. The cell membrane is shown in light blue. The rod, a prominent structure filling the space of the tip region, is depicted in green. Its major structural elements are HMW2 (Mpn310) in the core and HMW3 (Mpn452) in the periphery, stabilizing the rod (*42*).The individual complexes (A) are not to scale, but they are shown to scale within the bacterial cell (B).

similarity searches and comparative modeling provided structures for 484 *M. pneumoniae* proteins (70% of the genome) and 340 proteins in the network. There were also structural templates to construct models for 153 binary interactions (Fig. 1) covering 29 heteromultimeric and 57 homomultimeric complexes (table S6). These data can be used both to study particular interactions or complexes (Fig. 3B and fig. S7A) and to infer general correlations. Structural interfaces are particularly illuminating for the multifunctional proteins. When structural models are available for multiple interactions with a common protein, analysis of the interfaces can suggest whether the interactions are mutually exclusive (same binding sites) or compatible (different sites) (*41*). We observed that multifunctional proteins generally tend to accommodate more ligands per interacting interface ($P = 0.003$), consistent with the view that multifunctionality engages mutually exclusive interactions. For example, the protein P115 (Mpn426) has six distinct interfaces, each of which has several mutually exclusive interaction partners.

Having assembled a repertoire of structural information, the next logical step is to map these networks and protein complexes in their native environment, the cell. For this purpose, we performed cryogenic electron tomography of 26 entire *M. pneumoniae* cells (*42*) (fig. S8). We used pattern recognition techniques to generate probability maps for complexes selected from the larger ones in *M. pneumoniae* (Fig. 4) because larger complexes are more likely to be identified. After a thorough classification considering missing data, low signal-to-noise ratio, and known spatial proximities of different subcomplexes, we generated maps for the ribosome, the chaperone GroEL (Mpn573), the structural core of the pyruvate dehydrogenase (PdhC, Mpn391, homomultimer), and RNA polymerase, with a minimal number of false positives (Fig. 4). These large complexes are excluded from the tip, an organelle required for the attachment to epithelial cells, illustrating that even in a simple, minimal bacteria the proteome is spatially organized (*42*). Within the cell bodies, we could not find substantial proximities or patterns among the different complexes. In contrast to *E. coli* that contains a compact nucleoid forming an exclusion area in the cell center (*43*), circular DNA in *M. pneumoniae* is apparently uniformly distributed (*44*). We estimated the average number of complexes per cell to be 140 for the ribosome, 100 for GroELs, 100 for pyruvate dehydrogenase, and 300 for RNA polymerase. For the ribosome and GroEL, we also quantified complex abundances by Western blotting (fig. S9). For both, the numbers derived from Western blot were in the range of those estimated from the tomograms. This adds to the emerging view that the mapping of macromolecular structures into entire-cell tomograms (*45*), even though still challenging, is a powerful strategy when combined with unbiased large-scale complex purification. It opens the way to more general charting of cellular networks in entire-cell tomograms.

Conclusions. Our genome-scale screen for soluble complexes in a bacterium provides a valuable resource for the functional annotation of many genes whose biological roles in prokaryotic or parasitic cells are elusive. The coverage of known complexes leads to an estimate of some 200 molecular machines in *M. pneumoniae*. The study allows estimation of unanticipated proteome complexity for an apparently minimal organism that could not be directly inferred from its genome composition and organization or from extensive transcriptional analysis. Organisms with small genomes are the most tractable for systems biology, and the biochemical data set, proteome-wide spectra, ORFome, and collection of TAP-expressing *M. pneumoniae* strains will provide an extremely useful resource for this community. Comparison to both more complex bacteria and to even smaller ones, such as *M. genitalium* with 485 annotated protein-coding genes (*46*), should reveal additional systemic features associated with genome streamlining.

With protein structures available for about three-quarters of its ORFs, either directly from structural genomics efforts (*40*) or indirectly inferred by homology, *M. pneumoniae* has been extensively studied. We demonstrated that we can integrate data sets of biochemically determined complexes with structural information to approximate the three-dimensional organization of proteins into functional molecular machines. These models can then be mapped in entire cell tomograms, providing a three-dimensional view of cellular proteomes and interactomes (*47*); ultimately whole-cell models will benefit studies of biological function and disease.

References and Notes

1. U. de Lichtenberg, L. J. Jensen, S. Brunak, P. Bork, *Science* **307**, 724 (2005).
2. G. Rigaut *et al.*, *Nat. Biotechnol.* **17**, 1030 (1999).
3. A. C. Gavin *et al.*, *Nature* **440**, 631 (2006).
4. N. J. Krogan *et al.*, *Nature* **440**, 637 (2006).
5. K. Tarassov *et al.*, *Science* **320**, 1465 (2008); published online 7 May 2008 (10.1126/science.1153878).
6. J. C. Rain *et al.*, *Nature* **409**, 211 (2001).
7. J. R. Parrish *et al.*, *Genome Biol.* **8**, R130 (2007).
8. L. Terradot *et al.*, *Mol. Cell. Proteomics* **3**, 809 (2004).
9. G. Butland *et al.*, *Nature* **433**, 531 (2005).
10. M. Arifuzzaman *et al.*, *Genome Res.* **16**, 686 (2006).
11. P. Hu *et al.*, *PLoS Biol.* **7**, e96 (2009).
12. K. B. Waites, D. F. Talkington, *Clin. Microbiol. Rev.* **17**, 697 (2004).
13. R. Himmelreich *et al.*, *Nucleic Acids Res.* **24**, 4420 (1996).
14. T. Dandekar *et al.*, *Nucleic Acids Res.* **28**, 3278 (2000).
15. J. I. Glass *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 425 (2006).
16. Materials and methods are available as supporting material on *Science* Online.
17. D. N. Perkins, D. J. Pappin, D. M. Creasy, J. S. Cottrell, *Electrophoresis* **20**, 3551 (1999).
18. E. Gasteiger *et al.*, in *The Proteomics Protocols Handbook*, J. M. Walker, Ed. (Humana, Totowa, NJ, 2005), pp. 571–607.
19. K. A. Resing *et al.*, *Anal. Chem.* **76**, 3556 (2004).
20. F. Desiere *et al.*, *Nucleic Acids Res.* **34**, D655 (2006).
21. L. J. Jensen *et al.*, *Nucleic Acids Res.* **37**, D412 (2009).
22. M. E. Sowa, E. J. Bennett, S. P. Gygi, J. W. Harper, *Cell* **138**, 389 (2009).
23. C. von Mering *et al.*, *Nature* **417**, 399 (2002).
24. M. Güell *et al.*, *Science* **326**, 1268 (2009).
25. A. C. Gavin *et al.*, *Nature* **415**, 141 (2002).
26. H. Berman, K. Henrick, H. Nakamura, *Nat. Struct. Biol.* **10**, 980 (2003).
27. P. Zuber, *J. Bacteriol.* **186**, 1911 (2004).
28. E. Yus *et al.*, *Science* **326**, 1263 (2009).
29. K. S. Murakami, S. Masuda, S. A. Darst, *Science* **296**, 1280 (2002).
30. E. D. Levy, E. Boeri Erba, C. V. Robinson, S. A. Teichmann, *Nature* **453**, 1262 (2008).
31. R. Wright, C. Stephens, G. Zweiger, L. Shapiro, M. R. Alley, *Genes Dev.* **10**, 1532 (1996).
32. J. Gowrishankar, R. Harinarayanan, *Mol. Microbiol.* **54**, 598 (2004).
33. M. Torres, C. Condon, J. M. Balada, C. Squires, C. L. Squires, *EMBO J.* **20**, 3811 (2001).
34. J. Hodgkin, *Int. J. Dev. Biol.* **42**, 501 (1998).
35. E. L. Zechiedrich, N. R. Cozzarelli, *Genes Dev.* **9**, 2859 (1995).
36. M. Praetorius-Ibba, C. D. Hausmann, M. Paras, T. E. Rogers, M. Ibba, *J. Biol. Chem.* **282**, 3680 (2007).
37. S. V. Kyriacou, M. P. Deutscher, *Mol. Cell* **29**, 419 (2008).
38. S. G. Park, P. Schimmel, S. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 11043 (2008).
39. C. D. Hausmann, M. Ibba, *FEMS Microbiol. Rev.* **32**, 705 (2008).
40. S. H. Kim *et al.*, *J. Struct. Funct. Genomics* **6**, 63 (2005).
41. P. M. Kim, L. J. Lu, Y. Xia, M. B. Gerstein, *Science* **314**, 1938 (2006).
42. A. Seybert, R. Herrmann, A. S. Frangakis, *J. Struct. Biol.* **156**, 342 (2006).
43. M. Thanbichler, L. Shapiro, *Nat. Rev. Microbiol.* **6**, 28 (2008).
44. S. Seto, G. Layh-Schmitt, T. Kenri, M. Miyata, *J. Bacteriol.* **183**, 1621 (2001).
45. A. Al-Amoudi, D. C. Diez, M. J. Betts, A. S. Frangakis, *Nature* **450**, 832 (2007).
46. D. G. Gibson *et al.*, *Science* **319**, 1215 (2008); published online 23 January 2008 (10.1126/science.1151721).
47. P. Bork, L. Serrano, *Cell* **121**, 507 (2005).
48. M. Kanehisa *et al.*, *Nucleic Acids Res.* **36**, D480 (2008).
49. R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (1997).

50. We are grateful to M. Wilm, T. Franz, F. Thommen, E. Dalton, M. Schulz, E. Sawa, M. Diepholz, E. Pirkl, A. Seybert, C. Davis, J. Stülke, Gavin's and Bork's groups, and the EMBL Proteomic and Gene Core Facilities for expert help and discussion. This work is in part supported by the European Commission 6th and 7th Framework Integrated Projects "3D-Repertoire" and "Prospects," respectively; SystemsX.ch, the Swiss initiative for systems biology; the Netherlands Organization for Scientific Research (NWO); the Foundation Marcelino Botín; the Spanish Ministry of Education and Science (MEC)–Consolider; and the European Research Council. The data set has been submitted to the International Molecular Exchange Consortium (http://imex.sf.net) through IntAct (pmid is 17145710; identifier is IM-11644). The electron microscopy maps have been submitted to the Electron Microscopy Data Bank (www.ebi.ac.uk/pdbe-srv/emsearch/) (identification codes EMD-1637, EMD-1638, and EMD-1639).

Supporting Online Material

www.sciencemag.org/cgi/content/full/326/5957/1235/DC1
Materials and Methods
Figs. S1 to S9
Tables S1 to S8

15 May 2009; accepted 2 October 2009
10.1126/science.1176343