Big complexity in a minimal bacterium

Marc Güell Cargol

TESI DOCTORAL UPF / ANY 2010

DIRECTOR DE LA TESI

Dr. Luís Serrano Pubull (Centre de Regulació Genòmica)

UNIVERSITAT POMPEU FABRA

The research in this thesis has been carried out at the European Molecular Biology Laboratory (EMBL) and at the Centre for Genomic Regulation (CRG).

The thesis has been supported by a FPU fellowship:

Als meus pares,

Agraïments

Aquest document recull part del treball realitzat durant els últims tres anys i mig sota la direcció del Dr. Luís Serrano, començat a l'EMBL (European Molecular Biology Laboratory, Heidelberg) i continuat al CRG (Centre de Regulació Genòmica, Barcelona). Físicament, aquest projecte ha estat dut a terme entre Heidelberg i Barcelona. De fet, la posada a punt i bona part de les mesures de les dades tant de microarrays com de tiling arrays les vaig dur a terme al laboratori de l'Anne-Claude Gavin a l'EMBL. La continuació de les mesures i el corresponent anàlisis es va fer a Barcelona, al laboratori del Luis Serrano al CRG. Per tal d'ajustar alguns punts fou necessari fer dues visites curtes el laboratori del Peer Bork a l'EMBL, grup que va participar intensament en l'anàlisi de les dades obtingudes. Vull destacar que els resultats obtinguts són fruit d'una col·laboració multidisciplinària amb la contribució de diferents persones. Treballar en aquests dos instituts ha estat un plaer i un orgull enorme. Totes dues institucions, l'EMBL consolidada, i el CRG una mica més emergent, representen dos llocs extremadament preparats per dur a terme investigació de primer nivell i competir amb els millors. Ambdós presenten una estructura moderna i amb recursos. I més important, amb moltíssimes persones altament preparades i brillants, amb molta voluntat de interacció i col·laboració. Tanmateix, m'agradaria agrair molt especialment el suport d'algunes persones concretes. En primer lloc el director de tesi, el Luis Serrano, i els membres del comitè de tesi doctoral, el Mark Isalan, Lauro Sumoy i el Peer Bork. També m'agradaria agrair de forma molt especial l'Anne-Claude Gavin, la supervisora que vaig tenir a l'inici de tot aquest treball. A més dels supervisors, ha estat essencial l'ajuda i suport dels diferents companys de laboratori. Tant a l'EMBL com a Barcelona. Per una banda, m'agradaria destacar el grup de Mycoplasma de Barcelona (Eva Yus, Tobias Maier, Judith Wodke, Konstantinos Michalodimitrakis, Sira Martínez, Ronan Bourgeois, Paolo Ribeca, Emanuele Raineri, Bernard Paetzold). I per l'altra, les primeres persones que vaig rebre suport als inicis de la tesi a Heidelberg (Oriol Gallego, Tobias Maier entre d'altres), no només en els aspectes professionals. També he après moltíssim dels responsables de les plataformes d'ajut a la recerca. En la primera part, el Vladimir Benes i la Sabine Schmidt de la Genomics Core Facility de l'EMBL, i durant la darrera etapa, l'Ana Vivancos i el Heinz Himmelbauer a la unitat d'ultraseqüenciació del CRG. No puc acabar aquesta part sense tenir un record per les persones que m'han ensenyat i descobert el món de la ciència ja durant la carrera en química. Per una banda, molts professors del IQS (Institut Químic de Sarrià) com poden ser l'Antoni Planas, la Magda Faijes o el Santi Nonell, entre d'altres. I per la l'altra, el Gregori València i la Gemma Arsequell, del CSIC (Consejo Superior de Investigaciones Científicas) amb qui vaig tenir un fantàstic primer contacte amb la recerca.

Abstract

With only 689 genes *Mycoplasma pneumoniae* (*M. pneumoniae*) is among the simplest known organisms. Because of this simplicity, mycoplasma represents an attractive organism for systems-wide analyses. Such approaches aiming at the whole quantitative understanding of an entire organism are expected to illustrate the basic principles of life. Strand-specific tiling arrays complemented by transcriptome sequencing, were combined with more than 252 spotted arrays to study *M. pneumoniae* transcriptional organization. An important presence of alternative transcripts (42%) within operons and a high frequency of antisense RNA (89) were detected.
Metabolism was also studied in detail. A manually curated metabolic network allowed the definition of a minimal medium with 19 essential nutrients. This has been complemented with measurements of biomass indicators, metabolites and fluxes. Integration with transcriptional profiling has provided keys in the metabolic regulation.
Protein organization and interactions have been addressed systematically by Tandem affinity purification-mass spectrometry (TAP-MS) in a proteome-wide screen. The biochemical analysis revealed 178 protein complexes which have been complemented by structural models, single-particle electron microscopy and electron tomography.
By integrating the datasets from these different approaches, we show that this small bacterium harbors an unexpected complexity with features such as the frequent occurrence of alternative transcripts and antisense RNA, a small but tightly controlled metabolic network and a high level of proteome organization.

Resum

Amb només 689 gens *Mycoplasma pneumoniae* es troba entre els organismes més simples que es coneixen. Degut a aquesta simplicitat, mycoplasma representa un organisme atractiu per dur a terme estudis a nivell genòmic. S'espera d'aquests treballs que pretenen descriure de manera quantitativa l'organisme sencer que ajudin a entendre els principis bàsics de la vida.
Per tal l'estudiar amb profunditat del transcriptoma, s'ha fet ús d'una combinació de dades de "tiling arrays" amb especificitat de cadena, ultraseqüenciació i més de 252 microarrays. Després d'analitzar els resultats s'ha detectat una alta presència de transcrits alternatius (42%) dintre operons i una alt contingut de ARN de tipus "antisense" (89).
També s'ha realitzat un estudi detallat del metabolisme. S'ha revisat i completat manualment el mapa metabòlic de *M. pneumoniae*, fet que ha permès el disseny d'un medi mínim amb l'ús de 19 ingredients essencials. El mapa s'ha completat amb diferents mesures d'indicadors de biomassa, metabòlits i fluxos. També s'ha estudiat la regulació de metabolisme mitjançant microarrays.
Per altra banda, s'han mesurat sistemàticament les interaccions proteïna-proteïna mitjançant "Tandem affinity purification-mass spectrometry (TAP-MS)". Aquest anàlisis ha detectat 178 complexes diferents, els quals han estat complementats amb models estructurals, microscòpia electrònica i tomografia electrònica.
Mitjançant la integració d'aquestes col·leccions de dades, es pot mostrar que aquest petit bacteri amaga un inesperada complexitat amb característiques com la freqüència de transcrits alternatius i ARN "antisense", una xarxa metabòlica petita però fortament controlada i una alta organització del proteoma.

Preface

During the last years a new scientific discipline has emerged. Systems biology has been defined as the coordinated study of biological systems (*1*). It is basically composed by high-throughput and whole-genome techniques, and integrates computational methods with experimental efforts. The systems biology project where this thesis is framed has two interrelated goals.

The first, describing quantitatively a living system with the highest precision and completeness. For this, one of the simplest free-living organisms has been selected: *M. pneumoniae*. The most modern experimental techniques have been used to measure mycoplasma's molecular reality and combined with intense data analysis and computation to extract and integrate information about such an organism. DNA chips and deep sequencing techniques provided a map of the transcripts and their dynamics. A combination of deep literature search and experimental validation permitted the reconstruction of the metabolic network. And, the different protein complexes where screened by systematic protein purification. All these information and their integration conform one of the most complete datasets ever build for a single organism.

Second, a natural continuation of this project is to evolve towards Synthetic Biology. Once understood much better *M. pneumoniae* biology and having a complete list of its parts, the next challenge it is to make use of this information in guiding rational design of the bacterium. The first very initial steps have been started, but during the following years it will be time to develop this second phase.

The research carried out in this thesis has been mainly involved in the transcriptional mapping but it has also participated in the other parts. Special emphasis will be given when describing this part.

This thesis has involved experimental work, bioinformatic analysis and methods development. For instance, the transcriptional mapping provided in Güell et al (*2*), involved clearly the three phases. An initial stage of methods development, when setting up for the first time a prokaryotic strand specific tiling array and DSSS (Direct Strand Specific Sequencing); a second phase with intensive bioinformatic preparation of the data, including mapping, data annotation and statistical analysis; and eventually the extraction of the biological information such as operons or antisense transcription. In addition, DSSS, one of the methods used in Güell et al (*2*), was documented and tested in prokaryotic and eukaryotic organisms and submitted for publication independently (*3*).

The data were produced and analyzed under the highest standards and it represents an important effort from several groups at the EMBL and the CRG. The dataset is not only an amazingly complete amount of information centered in a single organism but also uncovers important aspects of bacterial biology. I really wish you a pleasant reading.

Contents

1 LIST OF PUBLICATIONS

1.1 Articles

(*2*) Transcriptome Complexity in a Genome-Reduced Bacterium Marc Güell, Vera van Noort, Eva Yus, Wei-Hua Chen, Justine Leigh-Bell, Konstantinos Michalodimitrakis, Takuji Yamada, Manimozhiyan Arumugam, Tobias Doerks, Sebastian Kühner, Michaela Rode, Mikita Suyama, Sabine Schmidt, Anne-Claude Gavin, Peer Bork, and Luis Serrano Science 27 November 2009: 1268-1271.

(*3*) Strand-specific deep sequencing of the transcriptome Ana P. Vivancos‡, Marc Güell‡, Juliane C. Dohm, Luis Serrano, and Heinz Himmelbauer Submitted for publication ‡Coauthors

(*4*) Proteome Organization in a Genome-Reduced Bacterium Sebastian Kühner, Vera van Noort, Matthew J. Betts, Alejandra Leo-Macias, Claire Batisse, Michaela Rode, Takuji Yamada, Tobias Maier, Samuel Bader, Pedro Beltran-Alvarez, Daniel Castaño-Diez, Wei-Hua Chen, Damien Devos, Marc Güell, Tomas Norambuena, Ines Racke, Vladimir Rybin, Alexander Schmidt, Eva Yus, Ruedi Aebersold, Richard Herrmann, Bettina Böttcher, Achilleas S. Frangakis, Robert B. Russell, Luis Serrano, Peer Bork, and Anne-Claude Gavin Science 27 November 2009: 1235-1240.

(*5*) Impact of Genome Reduction on Bacterial Metabolism and Its Regulation Eva Yus, Tobias Maier, Konstantinos Michalodimitrakis, Vera van Noort, Takuji Yamada, Wei-Hua Chen, Judith A. H. Wodke, Marc Güell, Sira Martínez, Ronan Bourgeois, Sebastian Kühner, Emanuele Raineri, Ivica Letunic, Olga V. Kalinina, Michaela Rode, Richard Herrmann, Ricardo Gutiérrez-Gallego, Robert B. Russell, Anne-Claude Gavin, Peer Bork, and Luis Serrano Science 27 November 2009: 1263-1268.

(*6*) Correlation of mRNA and protein in complexn biological samples Maier T, Güell M, Serrano L.. FEBS Lett. 2009 Oct 20. [Epub ahead of print]

1.2 Posters

From the chip to the network: Setting the basis for the design Marc Güell, Anne-Claude Gavin and Luis Serrano. Synthetic Biology 3.0, Zurich (Switzerland) 2007

Reconstruction of *M.pneumoniae* transcripional network Marc Güell, Anne Claude Gavin and Luis Serrano. EMBL PhD Symposium, Heidelberg (Germany) 2007

Reconstruction of *M. pneumoniae* transcripional network Marc Güell, Anne Claude Gavin and Luis Serrano. ESF conference, Sant Feliu de Guíxols (Spain) 2008

Functional, Spatial and Temporal characterization of *Mycoplasma pneumoniae* transcriptome Marc Güell, Eva Yus, Anne Claude Gavin and Luis Serrano. Internacional Conference in Systems Biology, Göteborg (Sweden) 2008

1.3 Oral presentations

Frequency of alternative transcripts and antisense RNA in bacteria reveal a high regulatory complexity, Marc Güell, Vera van Noort, Eva Yus, Wei-Hua Chen, Justine Leigh-Bell, Konstantinos Michalodimitrakis, Takuji Yamada, Manimozhiyan Arumugam, Tobias Doerks, Sebastian Kühner, Michaela Rode, Mikita Suyama, Sabine Schmidt, Anne-Claude Gavin, Peer Bork, and Luis Serrano. Systems Biology of Microorganisms, Paris (France) 2009

2 INTRODUCTION

2.1 Systems Biology

Life is among the most complex phenomena in the universe. It has been systematically studied by different classical disciplines such as botany or zoology at a macroscopic level but also at a microscopic level by cell biology or microbiology. In parallel, advances in biochemistry and molecular biology have contributed to uncover the life molecular landscape since many years. An impressive amount of knowledge has been generated but fragmented and not homogeneous.
Now the time has come to integrate different fields of natural sciences in order to understand better how cells work. The development of such a systematic view has been triggered by a technological and methodological revolution. Systems Biology, still an emergent discipline, aims to systematize the investigation of life (Fig. 1).

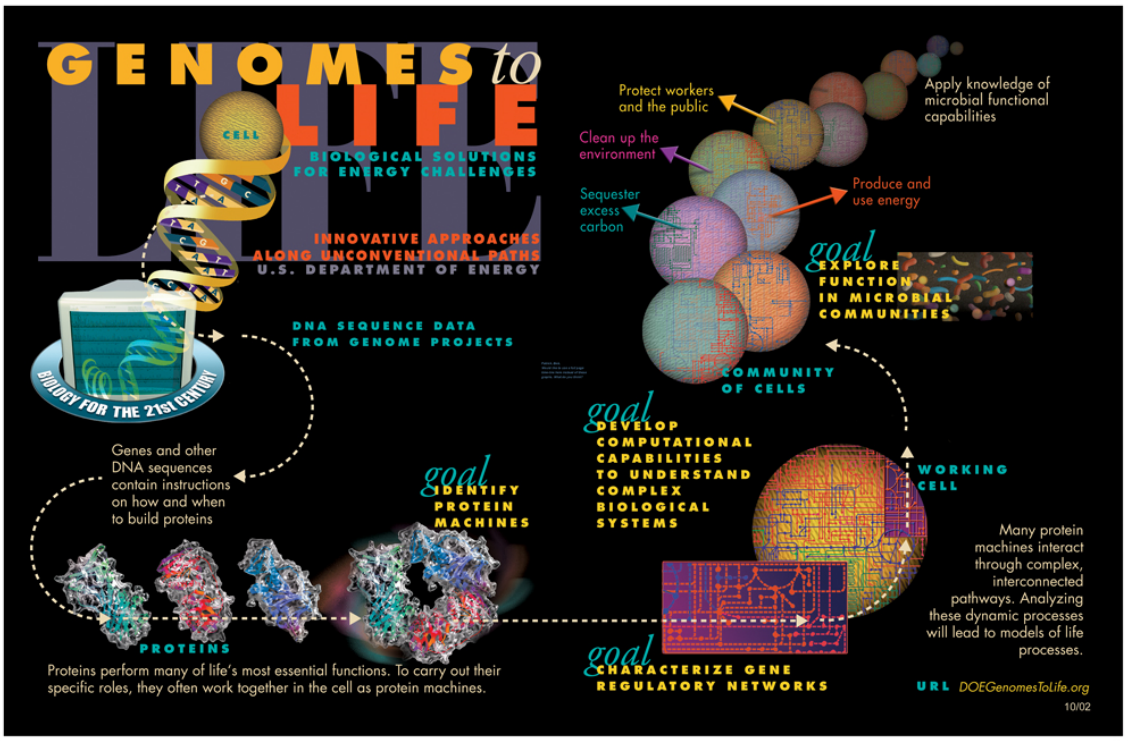


Figure 1. Example of systems biology research (obtained from Wikipedia.org).

Systems Biology has been driven by the curiosity of the scientists but even more by the high potential of its applications. Biotechnological processes can be optimized. Classical examples are found, for instance in metabolic engineering (*7*), where bacteria are redesigned to produce with efficiency certain compounds. Pharmaceutical industry has been also revolutionized, systems studies are expected to guide more effective drugs (*8*). A more detailed model of cell regulation may provide a guide to rationally design drugs and predict site-effects. Furthermore, it has driven a trend in health care towards individualized and advanced medicine. High-throughput technologies for analysis of genomes, transcriptomes, proteomes and metabolomes have provided the foundations for deciphering the structure, variation and function of the human genome and relating

them to health and disease states. Even, at the individual level (1000 genomes project)(*9*). This trend has been summarized under a new concept called P4 medicine (*10, 11*). Systems Biology will enable medicine to become predictive, personalized, preventive and participatory. Medicine today is mainly reactive and general, but it will move towards predictive and personalized modes. New technologies are allowing personal genome sequencing, and advanced molecular diagnostics will provide predictive and personalized diagnostics. How the acquisition, storage, mining and integration will occur, is one of the great challenges of Systems Biology. This P4 medicine has the potential to catalyze fundamental changes in virtually every aspect of the healthcare system and it will require rethinking the educational requirements for physicians.

a) Systems Biology as a technological revolution

A technological revolution has provided the tools needed to carry out such a big step ahead. One of the most important contributions has been provided by the large sequencing projects (*12, 13*). This knowledge provides the digital information necessary to deduce the coding sequence of the genes and their regulatory elements. After, it is possible to design exact sequence probes for monitoring the transcriptional level of genes using DNA chips or annotate proteomics datasets. Proteomics and transcriptomics are probably the two most influent post-genomic disciplines. Proteomics is the large-scale study of proteins and their function and transcriptomics aims to characterize the entire RNA content of a cell.

	454 GS FLX*	AB SOLiD	Illumina GAII
Chemistry	Pyrosequencing	Ligation based	Reversible terminators
	Standard	**Fragment**	**Fragment**
Run Time	7 hours	3-6.5 days	3 days
Read Lengths (bp)	250	25, 50	35, 50
Ave. Reads per Run	400K	$150x10^6$	$85x10^6$
Data per run	100MB	up to 7GB	up to 4.3GB
Throughput	100MB	1.1GB/day	1.4GB/day
	Titanium	**Mate-Paired**	**Mate-Paired**
Run Time	10 hours	7-13 days	6.5 days
Read Lengths (bp)	400+	2x25, 2x35	2x50
Ave. Reads per Run	$1x10^6$	$250x10^6$	$90x10^6$ pairs
Data per run	400MB	up to 8.75GB	9Gb
Throughput	400MB	900MB/day	1.3GB/day

*Metrics apply to both Fragment and Mate-Paired runs.

Figure 2. Brief summary of the current performance of the most popular ultrasequencing platforms (Obtained from Agencourt® Bioscience Corporation).

Modern transcriptomics has benefited from an emergent technological revolution (Fig. 2). It has evolved from a nucleic acids chips based discipline to a sequencing-based discipline (*14*). DNA chips technology and next generation sequencing have provided tools to massively survey the transcriptome at organism-wide level. Mapping all transcribed regions of an entire organism transcribed regions has become possible. Tiling chips and RNA-seq can provide information of UTRs, transcripts architecture and ncRNAs in a single experiment (*14-16*). When modern transcriptomics is combined with chromatin immunoprecipitation (ChIP) technologies, it offers great promise for the advancement of our understanding of transcription factors. ChIP-Seq (*17*) and ChIP-Chip (*18*) reveal the genome wide location of DNA-bound proteins. All these technologies have opened an important field of research. In Vivancos et al. (*3*)(included in the thesis), a method for strand specific transcriptome sequencing is described in detail. This method is applied in Güell et al. (*2*)(included in the thesis).
Mass spectrometry based proteomics has developed into a rapidly growing field providing researchers with opportunities to study biological systems on a variety of scales. Specific goals of proteomics studies include: determining the identity and modification state of the proteins present in a cell, measurement of protein abundance, changes over time, and elucidating the interactions of the proteins and their localization. An extensive revision of quantitative proteomics and quantitative transcriptomics is provided in Maier et al. (*6*)(included in the thesis). In Kühner et al (*4*)(included in the thesis), protein-protein interactions are studied at the systemic level, with the aid of biochemical purification and mass spectrometry.
Metabolomics is described as the systematic study of all metabolites in a biological organism. Different tools have become available for the integration of heterogeneous biological information and large-scale datasets, computational analysis and simulation of dynamical systems, and computational predictions of molecular interactions. In Yus et al. (*5*)(included in the thesis), the metabolic network of *M. pneumoniae* is reconstructed together with measurements of important parameters such as fluxes, biomass indicators and small metabolite concentrations.

b) Systems Biology as philosophy

Probably since a long time ago, systems level understanding has been discussed in biological science (*8*). However, until very recently, organism-wide studies were technologically not possible. An impressive data acquisition power has led to a paradigm change in Biology.
To understand biology, it is essential to examine the system as a whole. Properties such as robustness appear as emergent properties of the systems which can be only understood from a general perspective. With accumulating knowledge of many molecular entities at the same time, the natural question is what are the mechanisms dictating the systems behavior. Observation of the real world confronts us with many simple and complex processes that cannot be explained with common sense. Mathematical modeling and computer simulations can help us to understand the internal nature and dynamics of these processes. Simulation of biological systems is a broad field which already delivered interesting applications (*19*) and computing and statistical analysis has become a recurrent tool essential in dealing with most of the incoming datasets (mass spectrometry, microarrays, DNA sequencing,…).
It seems that we are well inside of a new era of Biomedical Science, where boundaries of disciplines are more difficult to set. Technology and computing are well rooted in our research day-life. However, their potential is limited to our hypothesis space, which is

always based on biological knowledge. Thus, interaction of different fields of science is more required than ever. Only with an open-minded attitude and interdisciplinary philosophy it will be possible to push forward our understanding of life.

2.2 Transcription in bacteria

Transcription is the first step leading to gene expression. Transcription begins with the binding of the RNA polymerase complex to a special sequence at the beginning of the gene known as the promoter. Activation of the RNA polymerase complex enables transcription initiation, and this followed by elongation of the transcript. The DNA sequence is read by RNA polymerase (RNAP), which produces a complementary antiparallel RNA strand. DNA is read from 3' to 5' during transcription (Fig. 3). Meanwhile, the complementary RNA is created from 5' to 3'. The RNAP covers about 30 base pairs (bp) of the template DNA, including the transcription bubble of 12-14 bp. The growing transcript is held to the template strand of the DNA by approximately eight RNA-DNA base pairs (Fig. 4).

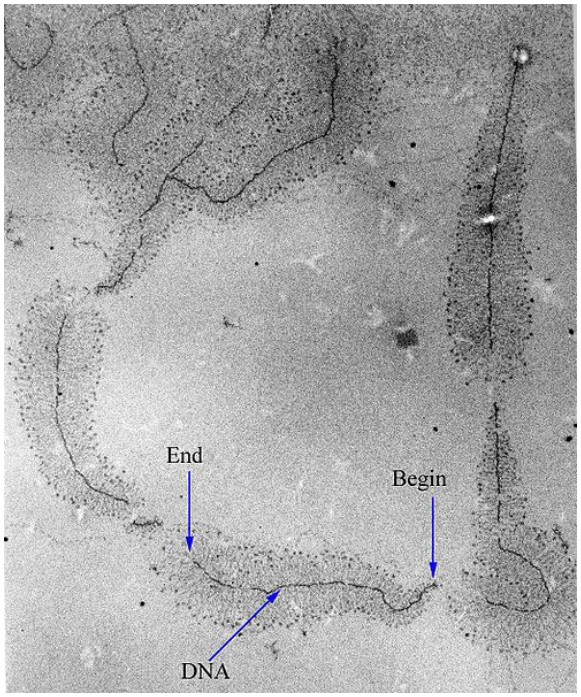


Figure 3. Micrograph of gene transcription of rRNA illustrating the growing primary transcripts. "Begin" indicates the 5' end of the DNA, where the new RNA synthesis begin; "End"indicates the 3' end, where the primary transcripts are almost complete (obtained from Wikipedia.org).

Bacterial transcription occurs in the cytoplasm and is coupled to translation. Prokaryotic mRNA involves less processing than in Eukaryotes and is divided into 4 stages: initiation, promoter clearance, elongation and termination.

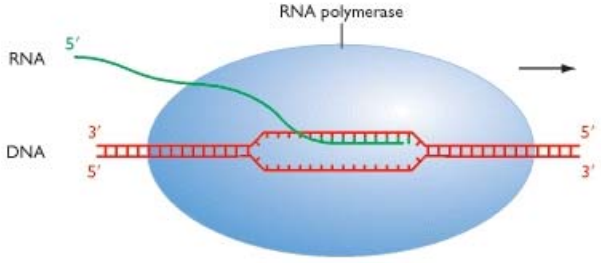


Figure 4. Schematic representation of the transcription elongation.

a) Transcription initiation and promoter clearance in prokaryotes

RNA synthesis is accomplished under the direction of DNA by the RNA polymerase (RNAP) holoenzyme. RNAP consists of the core enzyme and the sigma factor. A RNA core polymerase is multi-subunit complex with a general structure $\alpha_2\beta\beta'$ that undertakes the elongation of RNA. The sigma factor is needed for the initiation of RNA transcription, and it is a major influence on the selection of promoters (*20*). A transcription factor (TF) is a protein needed to activate or repress the transcription of a gene, but is not itself a part of the holoenzyme. Some TF bind to *cis*-acting DNA sequences only; others bind to DNA as well as to other TFs. When a TF binds to a specific promoter, it can either activate or repress transcription initiation. For instance, *Escherichia coli* is estimated to have 314 TFs, composed by 35% activators, 43% repressors and 22% dual regulators (*21*).
Transcription initiation begins with the binding of RNAP to the promoter in DNA. At the start of initiation, the core enzyme is associated with a sigma factor that aids in finding the appropriate -35 and -10 base pairs downstream of the promoter sequences (Fig. 5). After the first bond is synthesized, the RNAP must clear the promoter. During this action there is tendency to release the RNA transcript and to produce truncated transcripts. Abortive initiation continues to occur until the sigma rearranges (*22*). The sigma factor is released before 80 nucleotides of mRNA are synthesized. Once the transcript reaches approximately 23 nucleotides, it no longer slips and elongation can occur.
Promoter clearance coincides with phosphorylation of Serine 5 on the C-terminal domain of RNA polymerase.

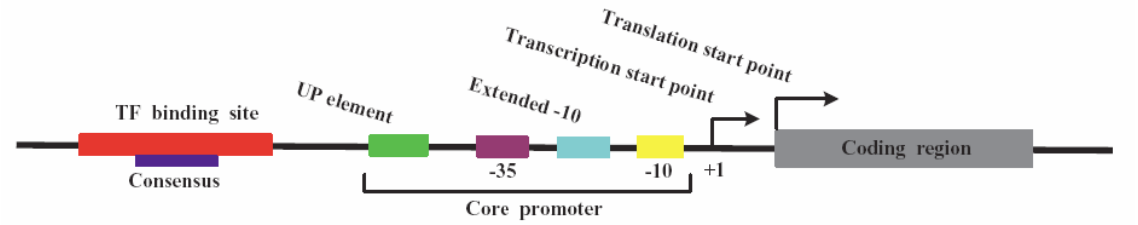


Figure 5. Scheme of a prokaryotic promoter. Adapted from Zhou et al. (*23*).

b) Transcription elongation in prokaryotes

One strand of DNA, the template strand, is used as the template for RNA synthesis. As transcription proceeds, RNAP traverses the template strand and uses base pairing complementarity with the DNA template to create an RNA copy. An RNA molecule is produced as an exact copy of the coding strand, except that thymines (T) are replaced by uracils (U). Ribonucleotides are added one after another to the growing 3' end of the transcript. A number of proteins are known to act during elongation and termination stages of transcription by direct modification of RNAP properties. Among others, Nus factors, ribosomal protein S4 or Gre factors influence elongation (*24*). These proteins affect RNAP processivity by modulation of transcription pausing, arrest, termination or anti-termination. Gre factors suppress RNAP pausing and arrest by stimulating the intrinsic nucleolytic activity of RNAP. When RNAP encounters a roadblock during elongation and backtracks, the 3'-end of the nascent transcript gets backpedaled into the

secondary channel of RNAP. Gre-induced cleavage of this extruded portion generates a new 3' terminus, giving a second chance to transcribe over the roadblock and resume elongation (*25*). Nus proteins are multifunctional transcription elongation factors, and depending on the situation may induce opposite effects on transcription. By itself, NusA stimulates certain types of pausing. In cooperation with other Nus factors, it can induce anti-termination at both ρ-dependent and ρ-independent terminators (*26*). During transcription of many genes, NusA induced pausing plays an important role in synchronizing transcription and translation (*27*).

c) Transcription termination in prokaryotes

Bacteria use two different strategies for transcription termination. In ρ-independent transcription termination, RNA transcription stops when the newly synthesized RNA molecule forms a GC rich hairpin loop followed by a run of Us, which makes it detach from the DNA template (Fig. 6)(*28*). In the ρ-dependent type of termination, a protein called ρ destabilizes the interaction between the template and the mRNA, thus releasing the newly synthesized mRNA from the elongation complex (*29*).
In some bacteria, such as *E. coli*, a large fraction is mediated by the ρ protein or its homologs. In others, such as *Bacillus subtilis*, ρ homologs play a smaller role, and ρ-independent termination is the norm.

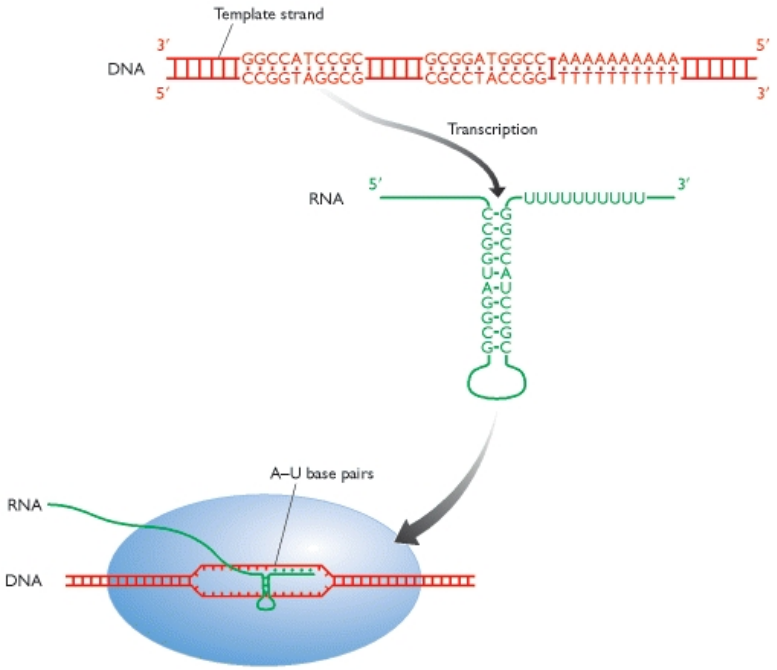


Figure 6. Scheme illustrating ρ-independent transcription termination (obtained from NCBI).

d) Prokaryotic transcriptome organization: Operons

Genes in prokaryotic organisms tend to be clustered in operons. An operon (Fig. 7) is composed by several genes with the potential to be transcribed within the same transcript. In classic biology, an operon has been considered an adjacent group of genes arranged under a unique common promoter and regulated by a common operator. Often,

a polycistronic operon encodes functionally related products, for example, enzymes of a metabolic pathway or subunits of a complex. This is a mechanism to co-express related genes. However, operons generating alternative transcripts or combining unrelated functions have been recently described to be abundant (*30*). For instance, an analysis of the DBTBS database of transcriptional regulation in *B. subtilis* revealed that more than 20% of its genes in known polycistronic operons are transcribed from more than one promoter. These additional promoters are often located downstream of the first gene, such that only part of the operon is transcribed from the internal promoter (*31*). Similarly, almost 6% of the known polycistronic operons contain an internal read-through terminator, at which partial continuation of the transcription occurs (*32*).

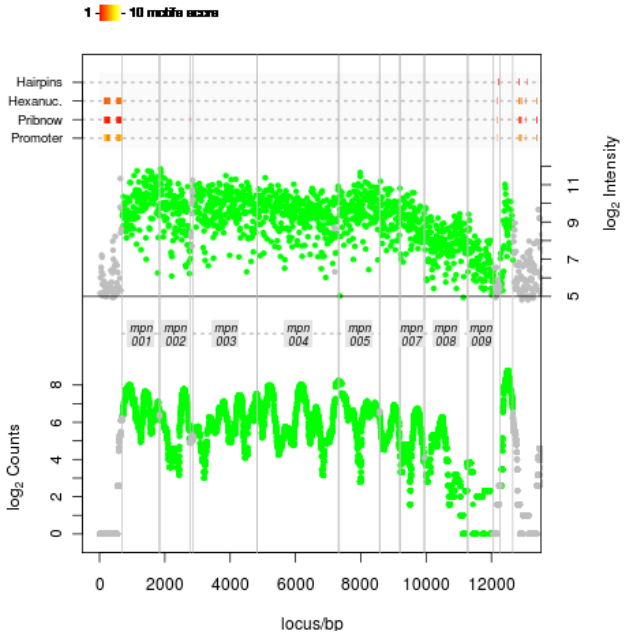


Figure 7. Expression across reference operon 001 (genes Mpn001-Mpn009) in *M. pneumoniae*. Upper panel: tiling array data/Lower panel: ultrasequencing data.

Multiple publications have shown suboperonic regulatory signals. Different examples in the literature and in *M. pneumoniae* are discussed extensively in Güell et al. (*2*).

e) Regulatory RNAs

RNA regulators in bacteria are a heterogeneous group of molecules that act by various mechanisms to modulate a wide range of physiological responses (*14*). The first class, binds to proteins. Some have housekeeping functions (i.e. Rnase P), whereas others act in a regulatory fashion by mimicking the structure of other nucleic acids (i.e. CsrB (*33*)). Another class is composed by riboswitches, which are part of the mRNAs that they regulate. A conformational switch in the mRNA molecule is induced by environmental changes, and which leads to a switch in gene regulatory function. Recently, it was found that some could bind small molecules and adopt different conformations acting as molecular sensors. In this case they directly regulate the genes involved in the uptake or use of the metabolite (*34*). The largest class and most studied

is the small RNAs (sRNAs) that act through base pairing with RNAs. They are usually involved in modulating the translation and stability of mRNAs.
Different general examples are provided in the discussion of Güell et al. (*2*) and specific examples affecting translation efficiency are discussed in Maier et al. (*6*), both publications included in thesis.

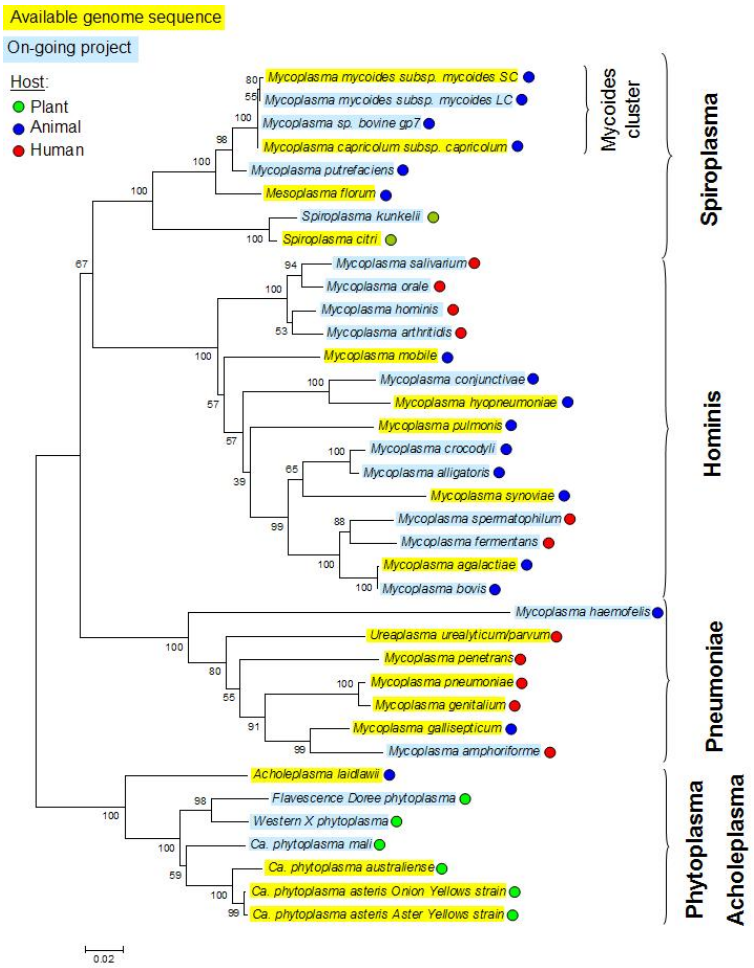
2.3 Mycoplasma pneumoniae

As systems and synthetic biology model, we have chosen to study one of the smallest bacteria that can live outside a host cell, *M. pneumoniae* accounting for 15-50% of all human pneumonia patients (Medline Plus, http://medlineplus.gov/). The M129 strain of this small bacterium has 689 annotated protein-coding genes and 44 classified RNAs (http://www.ncbi.nlm.nih.gov/). It is closely related to *Mycoplasma genitalium* (*M. genitalium*)(485 genes) which has previously been selected as a model for a minimal cell in synthetic biology (*35, 36*). In fact, almost all *M. genitalium* genes have an orthologue in its larger relative. There is ample literature regarding the essentiality of genes in *M. genitalium* and *M. pneumoniae* (*37, 38*) as well as on biochemical analysis of enzymatic reactions they can catalyze (*39*). Proteomics studies have been carried out with the aim of identifying the proteins expressed by this organism (*40, 41*). Interestingly, *M. pneumoniae* does not have cell wall but it has a complex cellular structure including a cytoskeleton and a differentiated attachment organelle which can be used for gliding (*42*). It is also able to survive in at least two environments, the extracellular matrix of the lung and on the surface and/or interior of the lung cells (*43*).

Figure 8. *M. pneumoniae* microscopy image. D Kunkel/Photolibrary.

a) Mollicutes

M. pneumoniae belongs to the Mollicutes (Fig. 9), which is a class of bacteria distinguished by the absence of a cell wall, their small genome size, a low G+C content and reduced metabolic capabilities. More than 200 species have been described. Since they lack a cell wall, they are osmotically fragile and pleomorphic. The small genome size places a restriction on the number of the regulatory elements present and proteins that can be coded. As a consequence, mollicutes possess limited metabolic activities and are dependent on a vast array of nutrients from their environment.
Mollicutes are widely distributed as pathogens or comensal organisms of a wide range of plant and animal hosts (Fig. 9). Taxonomically, they are considered sufficiently distinct from cell-walled bacteria to be placed in a separate division, *Tenericutes*, which has four orders and eight recognized genera. Traditionally, the major impetus for work with mollicutes has been their association with diseases of humans and economically important disease of other animals and plants. Because they lack a cell wall and their membrane composition may be modified by alterations in the growth medium, they are valuable organisms in which to study cell membrane and function. However, recently they have become target of major attention in attempts to obtain a complete understanding between genome sequence and cellular structure and function (*2, 4, 5*) or model organisms in synthetic biology (*36*). A major difficulty in purifying proteins from these organisms by expressing them in other organisms (i.e. *E. coli*) is that in Mycoplasma, Ureoplasma and Spiroplasma, the UGA codon is used to encode tryptophan.
The most important human pathogen of this group is *M. pneumoniae*, which causes respiratory infections, but also complications of the central nervous system. Interestingly, with the exception of a potential ADP ribosylating enzyme (*44*), there are not obvious classical virulence factors in the *M. pneumoniae* genome. Anyway, in general the mollicutes do cause harm to their hosts. At least in part, in many mycoplasma species this is due to the formation of hydrogen peroxide which is mainly formed during the utilization of glycerol (*45*).



Phylogenetic tree inferred from 16S rDNA sequences

Figure 9. Phylogeny of Mollicutes (obtained from Molligen 2.0).

b) Transcription

Some genes and operons have been studied at the transcriptional level (*46-48*) by a variety of methodologies and nylon arrays were used to study transcription under different conditions (*49*). These analyses have suggested that *M. pneumoniae* has a poorly defined -35 promoter region and a better defined -10 region, that many genes lack a canonical ribosome binding site and that in some cases there is heterogeneity at the transcription initiation point (*47*). *M. pneumoniae* contains most of the transcriptional machinery found in larger bacteria, including Nus proteins and other regulatory factors, only missing the ω subunit of the RNA polymerase. Transcriptional regulation in *M. pneumoniae* can be as complex as shown for lactate dehydrogenase and includes other elements than just the -10 and -35 boxes (*46*). Quantitative PCR analysis on the ftsZ operon in *M. pneumoniae* has shown a 'staircase/decay' expression behaviour where the first gene of the operon exhibits high transcription that progressively decreases with each gene towards the end of the transcript (*50*). The 'staircase/decay' behaviour has also been reported for some other operons in several bacteria (*51, 52*), but it is not clear whether this is a rule or an exception. It seems that transcription termination is mainly achieved by the placement of hairpin structures with a polyU tail at the end of operons as *M. pneumoniae* does not have a ρ termination factor (*53*). Taken together, *M. pneumoniae* is complex enough to be an attractive target for systems biology and at the same time simple enough to allow the refinement of large scale annotation by detailed analysis of its transcriptome.

c) Genome structure

Mycoplasmas have evolved from more classical bacteria of the firmicutes taxon by so-called regressive evolution that resulted in massive genome reduction. Erosion of bacterial genomes is more prone to occur in populations that are spatially isolated and sexually deficient. In restricted habitats, the environment is rather steady and natural selection tends to be reduced, resulting in gene inactivation by genetic drift. Reductive evolution has shaped the genome of Mycoplasmas.
Mycoplasmas have a small genome size with a marked A+T bias and a low number of genes involved in recombination and repair. Only one copy of the rRNA operon is present and most of the anabolic pathways have been eliminated.
Transcription as well as translation control elements are present in the *M. pneumoniae* genome. Comparing different promoters (this work), a strong -10 consensus region can be identified. While it remains difficult to describe a consensus -35 box. However, the most abundant -35 motif is equivalent to the most abundant in *E. coli* (*2*). A clear Shine-Dalgarno motif (GGAGG) is observed for certain genes, but the majority lack such motif (*47*). Novel mechanisms for regulation of gene expression are likely to be elucidated by further studies.

d) Metabolism

The reductive evolution of the *M. pneumoniae* is reflected in its metabolic properties. The metabolism of *M. pneumoniae* has been studied biochemically (*54*) and computationally (*55*). Its genome has undergone a massive genome reduction and most of the anabolic pathways have been lost. It lacks cytochromes and a tricarboxylic acid cycle, produces ATP by substrate phosphorylation and pyruvate kinases and cannot synthesize purines. Guanine, guanosine, uracil, thymine, cytidine, adenine and

adenosine may serve as precursors for nucleic acids and coenzymes. It seems to have ability to metabolize glucose and arginine (*56*).
Lipid metabolism is largely unknown in *M. pneumoniae*. Mycoplasmas are reported to uptake lipid compounds from the growth medium (*57*). However, different lipids are detected that are not present in the medium. Thus, at least it has ability to modify uptaken lipids. An important diversity of lipoid molecules has been detected including glycolipids, phophatidylglycerols, cholesterols, phospholipids, aminolipids, aminoglycolipids or unusual lipids (*58*). Specifically, membrane lipid biosynthesis has been shown using cell extracts: three glycolipids, five phosphoglycolipids and six phospholipids have been identified (*59*). Membrane lipid composition has been shown to be essential in the host immune response (*59*).

e) Signal transduction

Only few regulatory mechanisms are present in *M. pneumoniae*. They act in different layers of the cellular machinery.

i. Postranslational modifications

M. pneumoniae contains two protein kinases, the HPr kinase that phosphorylates the protein HPr involved in sugar import and a Ser/Thr protein kinase (PrkC). It also has a Ser/Thr phosphatase (Pp2C). Only a very few protein regulatory modifications have been documented in *M. pneumoniae*. One well studied example is the phosphorylation of HPr of the phosphotransferase system (PTS) triggered by glycerol. The phosphorylation of HPr on Ser-46 leads to carbon catabolyte repression (*45*). Another very recently described example, is the posttranslational modification of a set of cytadherence proteins by PrkC. A *M. pneumoniae* mutant affected in PrkC, results in nonadherent growth and loss of cytotoxicity (*60*). It is quite possible that acetylation and methylation of proteins could play a role in signal transduction since there are acetylases and methylases present in the genome. However, further research is required to elucidate their possible role as regulators.

ii. Gene transcription

The HrcA heat shock transcription factor has been shown to be active and mediate the upregulation of DnaK, ClpB, DnaJ and Lon (*49, 61*).

iii. Chemical messengers

Different chemical messengers are described in prokaryotes. Genes for the synthesis of (p)ppGpp, AppppA or c-di-AMP (*5*) are present in the *M. pneumoniae* genome. (p)ppGpp is an alarmone involved in the stringent response, causing the inhibition of RNA synthesis when there is a shortage of amino acids present (*62*), AppppA is synthesized in bacterial cells under stress and is proposed to be a signaling alarmone for oxidative stress (*63*) and c-di-AMP signals genomic DNA integrity in *B. subtilis* (*64*).
Their presence shows that despite the massive genome reduction some signaling mechanisms remain.

iv. GTPases.

A set of GTPase like proteins (ie: Mpn008, Mpn249) are encoded in the *M. pneumoniae* genome. Typically these proteins participate in signal transduction mechanisms; further research is expected to elucidate their roles. Based on sequence similarity to other bacteria, Mpn008 is suggested to be involved in tRNA modification (*65*) and Mpn249 in ribosome biogenesis (*66*).

f) Protein architecture

Protein architecture has been studied mainly at two different levels. First, at the single protein level, the Berkeley Structural Genomics Center (http://www.strgen.org/) has solved 84 protein structures encoded in the *M. pneumonia* genome. This program intends to obtain a near-complete structural complement of minimal life. Second, important research has been carried out at the full-organism level. *M. pneumoniae* has been analyzed by cryo-electron tomography. The small size of this organism makes it suitable to fit all its size into a single tomogram. This technique offered the possibility to visualize the cells three-dimensionally and resolve the structure of the attachment organelle. The tip surface proteins could be observed with a remarkable resolution (*67*). This study has been extended in one of the publications included in this thesis (*4*) to capture protein complexes in solution. Thus, it delivers a structure-based network of protein complexes in the cell.

g) Cell division, motility and structure

The small genome of *M. pneumoniae* does not contain genes for cell wall production. Nevertheless, it has a remarkable level of structural complexity. Experimental evidence indicated the presence of cytoskeletal-like structure and function in *M. pneumoniae* well before than in walled bacteria (*68*). Furthermore, *M. pneumoniae* possesses a complex and differentiated terminal organelle that mediates adherence to host cells and gliding motility (*67, 69, 70*). This attachment organelle and the polar filamentous cell shape of *M. pneumoniae* are thought to be stabilized by the intracellular cytoskeletal-like structures. The most remarkable architectural feature of the cytoskeleton-like structures is the electron-dense core, a rod-like structure that exists at the attachment organelle (Fig. 10). A network of fibrous structures is also observed in the cytoplasm.
Mycoplasma cells are divided by binary fission. The current model of *M. pneumoniae* division is that duplication of the attachment organelle structure precedes the division of the cell (*71*). The duplication is shown in Fig. 10. FtsZ protein, a bacterial homolog of tubuline, is found in *M. pneumoniae* and it is essential for cell divisiont. However, other genes (with the exception of a very divergent FtsA and DivIVA genes) working at the constriction site of walled bacteria cannot be found. Further research is required to uncover the details of mycoplasma cell division.

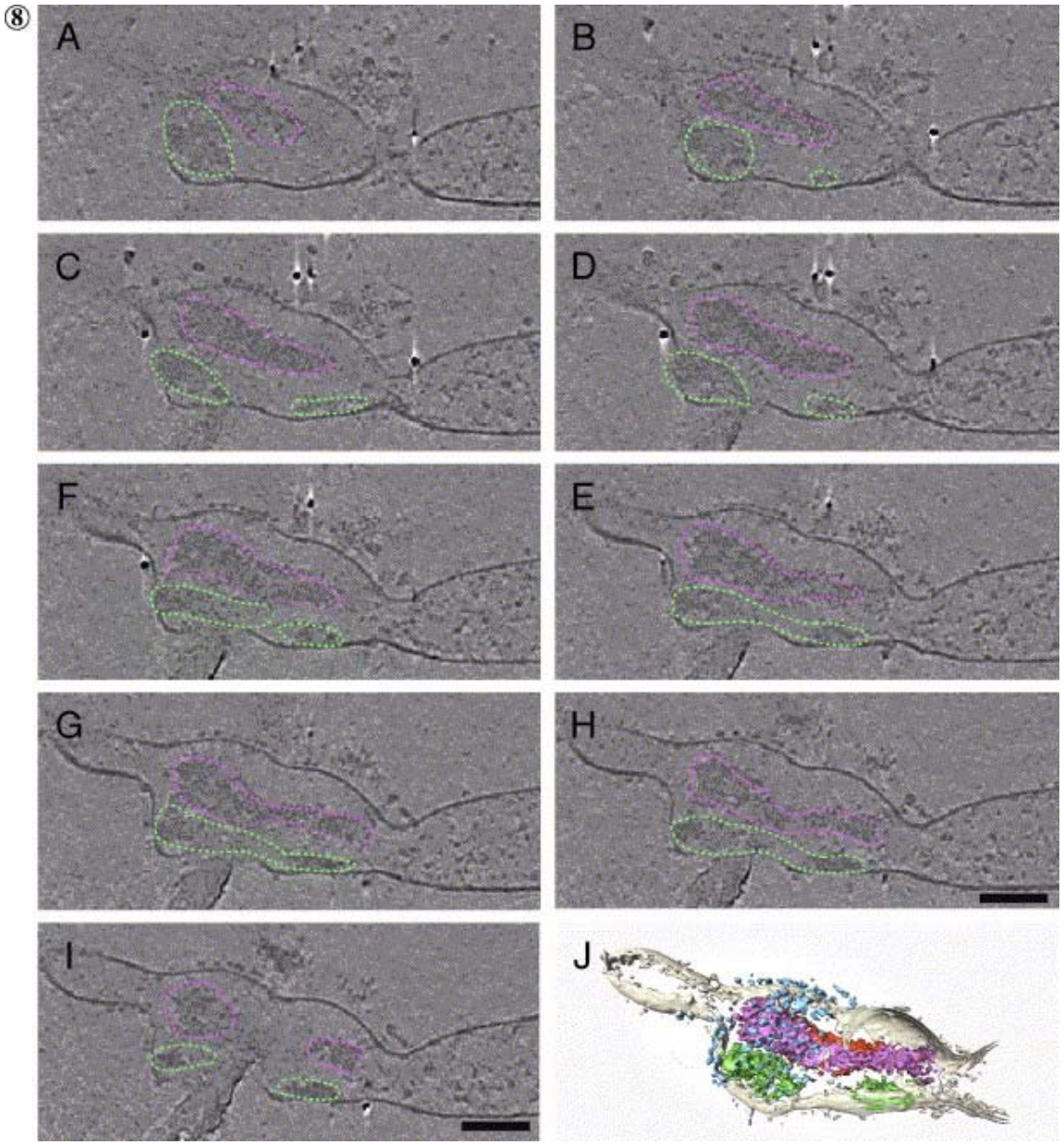


Figure 10. Tip structure in the process of duplication. The order of images is from A to I. The parental rod is surrounded by a magenta line, the nascent rod by a green line. (J) A surface rendered view of the reconstruction is shown. The nascent rod is colored green, membrane in beige, surface proteins in blue, rod in magenta and complexes accompanying the rod in red. Scale bars, 50 nm. (Adapted from Seybert et al., (*67*))

3 PUBLICATIONS

3.1 Transcriptome complexity in a Genome-Reduced Bacterium

Güell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, et al. Transcriptome complexity in a genome-reduced bacterium. Science. 2009; 326(5957): 1268-71.

3.2 Strand-specific deep sequencing of the transcriptome

Vivancos AP, Güell M, Dohm JC, Serrano, L, Himmelbauer H. Strand-specific deep sequencing of the transcriptome. Genome Res. 2010; 20(7): 989-99.

3.3 Impact of Genome Reduction on Bacterial Metabolism and Its Regulation

Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, Chen WH, et al. Impact of genome reduction on bacterial metabolism and its regulation. Science. 2009; 326(5957): 1263-8

3.4 Proteome Organization in a Genome-Reduced Bacterium

Kühner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, et al. Proteome organization in a genome-reduced bacterium. Science. 2009; 326(5957): 1235-40.

3.5 Correlation of mRNA and protein in complex biological samples

Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. FEBS Lett. 2009; 583(24): 3966-73.

4 DISCUSSION

The simplest cells are bacteria, but they generally contain thousands of genes. Natural science has always wanted to understand life. This desire to understand has long attracted biologists to work with simple, near-minimal cells. The simplest cells known to grow axenically are mycoplasmas.

This thesis is composed by 5 related publications. Three of them are centered in *M. pneumoniae.* They analyze with an unprecedented detail its transcriptional regulation (*2*), proteome organization (*4*) and metabolism (*5, 72*) (Fig. 11). Two more technical papers are also linked, one detailing one of methods used to measure the transcriptome and another reviewing protein-mRNA copy number correlation (*6*).

All this papers should be read as chapters within a larger story. The power can only be envisaged considering them as a whole. It is an attempt to apply the best tools available to quantify all possible parts of a minimal cell. This work probably provides one of the most complete datasets ever build for a single organism.

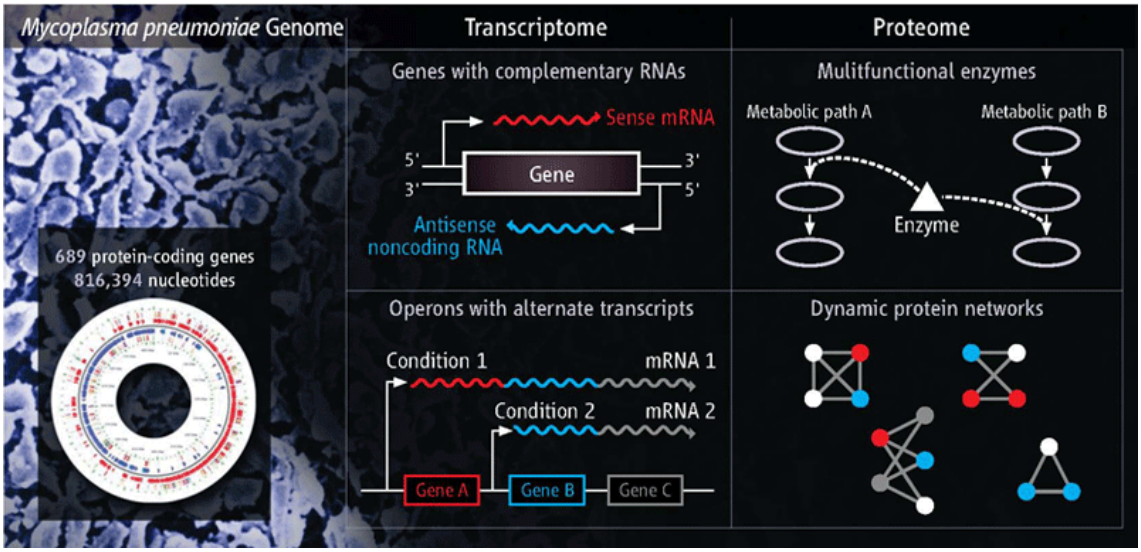


Figure 11. Scheme of the different aspects of *M. pneumoniae* addressed. Adapted from Ochman et al. (*73*).

4.1 Strand-specific deep sequencing of the transcriptome

In the last few years, different ultrasequencing platforms have been developed. These technologies have provided an impressive array of successful applications. In this article, DSSS (Direct Strand-Specific Sequencing), a method for transcriptome sequencing using next generation sequencing is described. The performance of the method is tested in RNA prepared from *M. pneumoniae* and from *Mus musculus*. DSSS is also compared to strand-specific tiling arrays. The method offers single base resolution, a superior dynamic range and a reduced 5' to 3' bias.
DSSS is proposed as a simple and efficient strategy for strand-specific transcriptome sequencing. It does not require a complex sample preparation and it is compatible with long fragments and sequences prone to generate secondary structure. Another advantage is that it can be used with the widely available platforms Illumina Solexa and Roche 454.

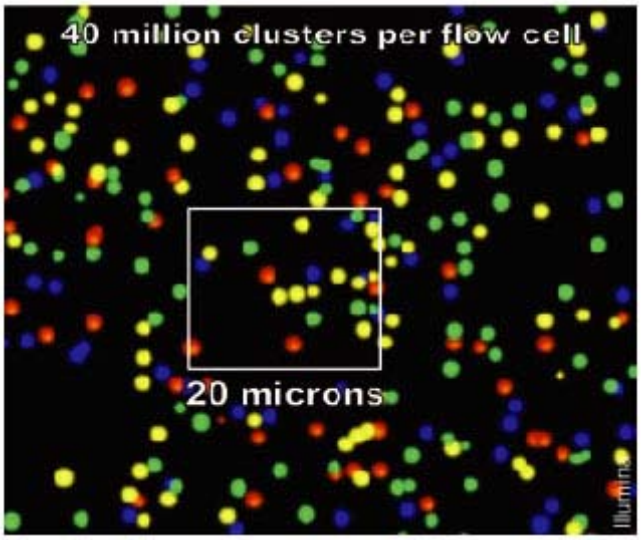


Figure 12 . Color dots have replaced bands as a sequence is read with reversible terminators on a Solexa (Illumina) sequencer.

DSSS provides a much higher dynamic range than tiling arrays. This is an important advantage since transcriptomics experiments face the challenge of interrogating RNA species ranging different orders of magnitude. In addition, array experiments tend to be noisy due to cross-hybridization and background issues. DSSS shows a much higher signal/noise ratio than arrays.
Single base resolution, big dynamic range and low background permit to fulfill the requirements of the most demanding transcriptomics experiments addressing precise characterization of the transcripts and antisense ncRNAs. This technology in combination with tiling arrays and microarrays has been used to study the transcriptome of *M. pneumoniae*.

4.2 Transcriptome Complexity in a Genome-Reduced Bacterium

In this study, *M. pneumoniae* transcription is analyzed with an unprecedented depth. A combination of spotted arrays, strand-specific tiling arrays and transcript sequencing provided an exhaustive description of the transcriptome including operons structure and more than 100 previously un-annotated transcripts. Most of the new transcripts are in antisense with annotated genes. *M. pneumoniae* transcriptional regulation shows an unexpected complexity more similar to what is present in conventional bacteria (i.e. *E. coli*) and even Eukaryotes.

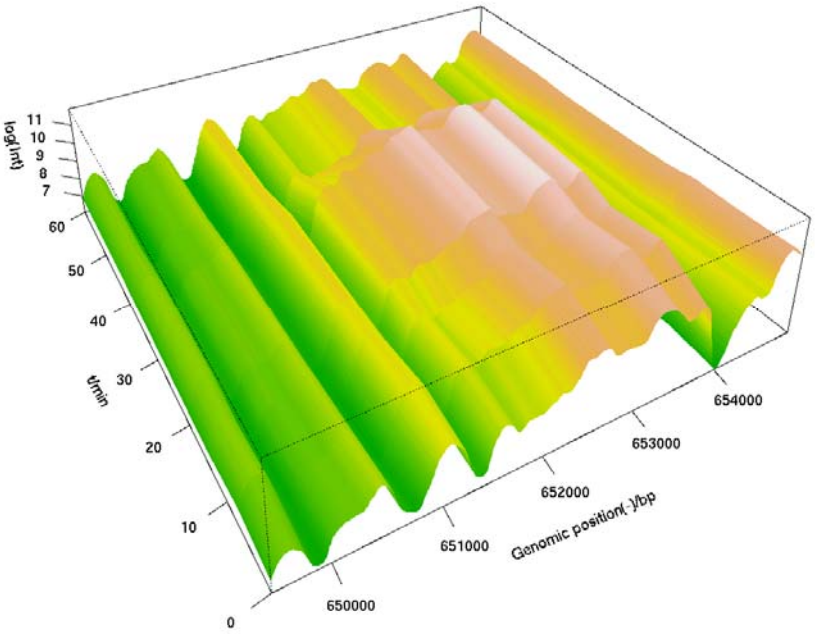


Figure 13. 3D representation of the transcriptional activity of the genomic area near gene *clpB* in a time series where heat shock is induced and later released. We observe a controlled upregulation in time and space.

a) Operon architecture

In bacterial genomes, genes of related functions often are localized in operons. But the fact that consecutive genes within operons do not have the same expression level has been observed many times in many bacterial species (*30*), although sometimes under the suspicion that the results could be due to experimental artifact. Most of the observed reference operons show natural polarity, where the first gene of the operon exhibits transcription that progressively decreases with each gene towards the end of the transcript. We also observed operons which do not present polarity, that is, the expression of the cistrons was equimolar, and operons showing complex patterns of expression (Fig. 1B in Güell et al. (*2*)). Interestingly, specific polarity became condition dependent for most of the operons (Fig. 14). Analysis of the 43 tiling arrays and integration with 252 spotted arrays representing 173 independent conditions, some of them from time-series, revealed context-dependent modulation of operon structures involving repression/activation of operon internal genes, as well as of genes located at the beginning, or end (Fig. 2 in Güell et al. (*2*), Fig. 14). In some cases this modulation can be assigned to specific environmental changes. Drop of the four first genes of the ftsZ operon involved in initiation of cell division corresponds to entry into stationary phase (Fig. 2 in Güell et al. (*2*)). The observed increase in arginine fermentation genes (arcA, arcI, arcC*)* (Fig. 2 in Güell et al. (*2*)) in stationary phase could be a mechanism to cope with acidification (*74*). We found formal evidence for a total of 447 transcriptional units (336 monocistronic and 111 polycistronic), implying a high rate of alternative transcripts (42%) in this bacterium in the conditions studied, similar to that in eukaryotes (40%, although still under debate. Interestingly, we found that genes that are split into different suboperons tend to belong to different functional categories. Thus, although genome reduction leads to longer operons accommodating genes with different functions (*75*), the latter can still retain internal transcription and termination sites under certain conditions. From the operons that were found in the reference condition but were condition dependent, 97 % were supported by DSSS reads (i.e. intergenic regions are covered by sequencing reads from the reference condition).

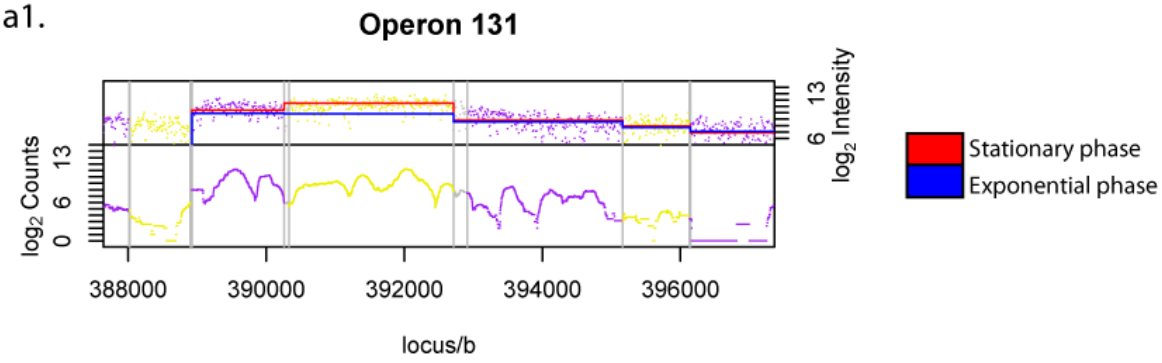
Single cases of alternative transcripts have been reported for bacteria (*30*) and have been attributed to a plethora of different mechanisms. A subset of genes inside the operon may have its own promoter (*76*), terminator or both, regulated individually. An interesting example is the variable operon polarity established of the operon gal in *E.coli*. Gal cistrons are expressed in equimolar amounts when cells are grown in glycerol versus succinate plus galactose, but the distal cistrons reduce their expression when grown in glucose plus galactose (*77*). Natural polarity variation is induced through a complex effect. The concentration gradient is established due to a not fully understood coupling between transcription initiation and termination. An analogue effect is also present in eukaryotes, where transcription by RNA polymerase II is coupled to mRNA splicing (*78*). These observations strengthen the emerging view that gene expression in prokaryotes is under the control of complex regulatory mechanisms which considerably deviate from the traditional operon concept. Condition specific polarity allows a tight regulation of gene expression. In Fig. 14 and Fig. 2 in Güell et al. (*2*), we show three examples of how specific transcriptional units can be regulated under certain conditions. A DNA repair machinery suboperon was also specifically upregulated upon DNA damage within the reference operon 1. The first two ORFs were more expressed under DNA damage than in cells growing at stationary phase whereas the other genes of the operon remained constantly expressed. These two genes are coding for the DNA polymerase III beta subunit (Mpn001) subunit and the DnaJ

(Mpn002) like protein, both involved in DNA repair (Fig. 2 in Güell et al. (*2*)). Metabolic genes were also controlled in this way. Reference operon 126 contains the 4 members of the arginine fermentation pathway plus a putative amino acid permease (Fig. 2 in Güell et al. (*2*)). Gene expression of the whole arginine pathway was increased in stationary phase, when the medium was more acidic due to the fermentation products. This has also been observed in *Lactococcus lactis* where the ammonia produced by the pathway may compensate the decrease of the intracellular pH (*74*). Other examples of condition specific polarity are provided in Fig. 13. Furthermore, two detailed independently validated operons are studied in detail (Fig. 13B). We provide the estimated promoter predictive score (see Methods in Güell et al. (*2*)) whose maximum values coincide with the beginning of the different suboperons.

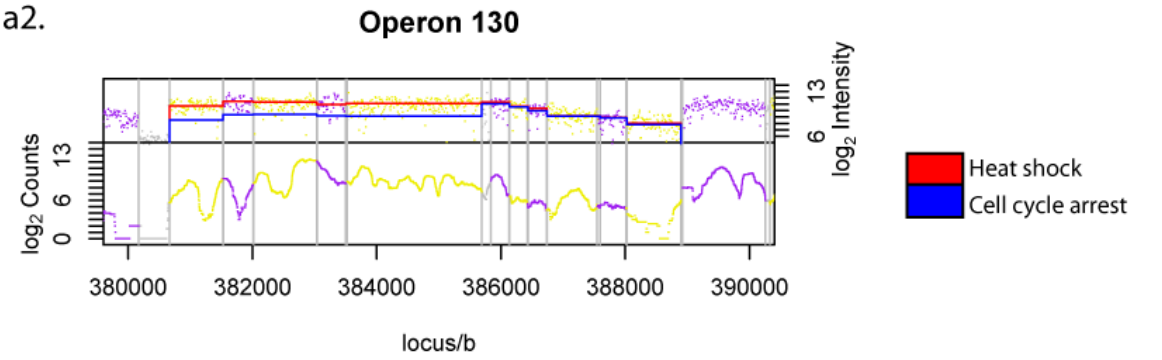
A recent study carried out in the Archea in *Halobacterium salinarum* detected 40% of condition dependent operons (*79*). Thus, together with what we observe in *M. pneumoniae* a complex relationship between gene expression and genomic organization seems to be well rooted in the tree of life.

A.

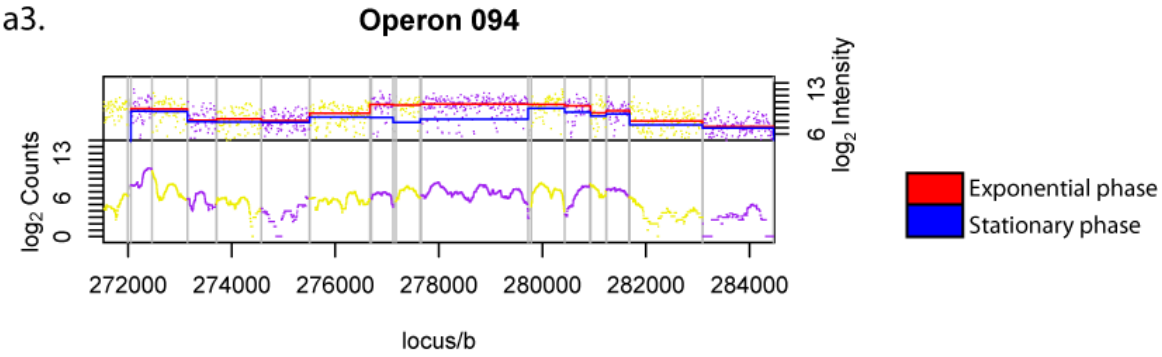
a1.


Operon 131
log2 Counts
0 6 13
6 13 log2 Intensity
388000 390000 392000 394000 396000
locus/b
Stationary phase
Exponential phase

a2.


Operon 130
log2 Counts
0 6 13
6 13 log2 Intensity
380000 382000 384000 386000 388000 390000
locus/b
Heat shock
Cell cycle arrest

a3.


Operon 094
log2 Counts
0 6 13
6 13 log2 Intensity
272000 274000 276000 278000 280000 282000 284000
locus/b
Exponential phase
Stationary phase

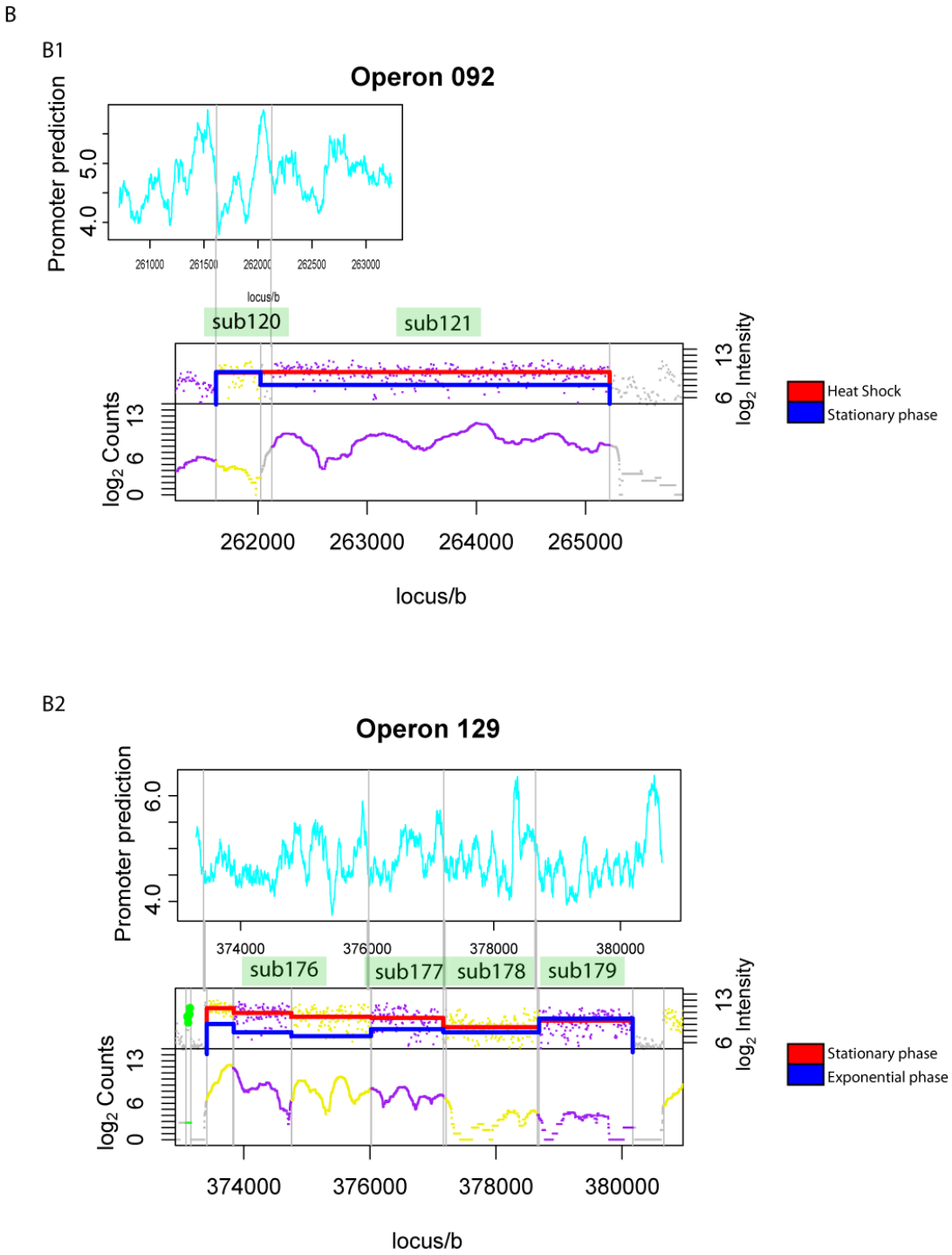


Figure 14. Examples of operons having alternative transcripts. A. The tiling array data obtained under different experimental perturbations was fitted as indicated in methods (*2*). The continuous lines indicate the fitted data. The colour corresponds to the specific condition analyzed. Top panel, The gene corresponding to the heat shock Lon protease (Mpn332) is specifically activated by heat shock. Medium panel, the first fine genes (Mpn320 to 324) involved in nucleotide metabolism and redox are activated under active growth. Bottom panel. Internal genes (Mpn225 to 227) encoding for two ribosomal proteins and elongation factor EF-G are induced specifically during active growth. B. Two more examples are provided. In this case, the promoter predicion score (see methods section) is presented. Green boxes indicate suboperons. All different suboperon start coincide with maximums in the promoter score. Both full length transcript have been validated (*50, 80*). Operon 092 encodes for a ADP ribosilating toxin (Mpn212) and the second a hypothetical protein (Mpn213). Operon 129 it is described in Fig. 2, Güell et al. (*2*) .

b) Antisense transcription

For *de novo* transcript discovery, a segmentation algorithm (*81*) was used to identify putatively transcribed regions. We identified 117 regions with signal above background and transcript-like shape in the tiling array (i.e.: quadratic pulse, 5' to 3' decay and no previous annotation, see Methods in Güell et al. (*2*)). These regions were further analyzed by DSSS reads in order to complement for some of the limitations in the tiling arrays, i. e. lower resolution, low dynamic range and the weakness in discriminating expressed regions from background signals at low levels of expression. It should be noted that DSSS alone is not sufficient to define transcripts because RNA fragmentation and subsequent amplification lead to large, periodic variations in base coverage within genes as well as gaps in genes expressed at low level. Thus, changes in transcript levels from DSSS can be used to define transcript boundaries only in combination with tiling arrays.

Sequence similarity with known proteins revealed the presence of two new protein-coding genes, a pseudogene, as well as one N-terminal truncation, and five 5'-extensions of known genes. The remaining 108 transcripts are likely regulatory rather than structural RNAs as comparison of their predicted secondary structures with the ones of coding genes do not show any significant difference. Eighty-nine of them are antisense with respect to previously annotated genes. In total, 13% of the coding genes are covered by antisense; this is two times more than in yeast (7%) (*82*), and about half of what was reported for plants (22.2%)(*83, 84*), or human (22.6%)(*85*). Surprisingly, they account for a significant fraction of transcripts (15%) under reference condition. For example, Mpn418 which codes for a Holliday junction resolvase, has a completely overlapping antisense RNA near its end (Fig 1C, Güell et al. (*2*)). Previous studies already detected non-coding RNA transcription in *M. genitalium (86)* (in an indirect way by transposon analysis) and in *M. pneumoniae (50)* but here we could extensively map and quantify these non-coding transcripts.

The relatively low number of detected antisense RNA indicates that the actinomycin D treatment during preparation of the tiling array samples, was efficient in preventing artefactual apparent transcription from the opposite strand (*87*). The DSSS protocol is also proved to be strictly strand specific (*88*). As a further indicator of their biological relevance, we find that the RNA expression level changes under different conditions differently than that of neighboring genes (Fig. 15). As final proof for these eukaryote-like elements we confirmed expression of four new transcripts by qPCR.

The effects of antisense on their sense RNAs have not yet been clearly established (*89, 90*). In mammals, perturbation of an antisense RNA may alter the expression of the messenger RNA (*91*). Antisense transcripts may affect expression of the overlapping functional sense transcripts through several molecular mechanisms, and these can be inferred from the expression pattern of the two transcripts (*92*). Double-stranded RNA-dependent mechanisms require co-expression with their target, whereas transcriptional interference rather implies mutual exclusion of sense and antisense transcripts.

In prokaryotes, individual cases of antisense RNAs were reported a long time ago (*93, 94*) but only very recently, it has been suggested that some antisense RNAs could be involved in gene expression regulation (*95*). For instance, *Clostridium acetobutylicum* uniGmccBA operon, involved in methionine to cysteine conversion, is regulated via antisense mediated transcriptional interference. When a sulfur source is available, the abundance of ubiG transcript is anticorrelated with the antisense transcripts (*95*). In *B. subtilis*, ratA is a small untranslated RNA overlapped 75 nucleotides with the toxic peptide txpA, such that the 3' of ratA is complementary to the 3'region of txpA.

Deletion of ratA leads to increased levels of txpA mRNA and lysis of the cells (*96*). On the other hand, mRNA stabilizing effects have also been reported. In *E. coli*, gadY and gadX, a mRNA encoding a transcription factor involved in acid response overlap at their 3'regions. A gadY overexpressing strain displays 20-fold increase in levels of gadX-mRNA levels (*97*). It has been suggested that base-pairing between gadY and the gadX mRNA simulates cleavage of a longer gadXW mRNA resulting in two products that are more stable than the full-length transcript (*98*). Such discoveries indicate the importance of sRNA as regulators of bacterial expression. In this study, we mapped and quantified genome-wide these non-coding transcripts. As observed in mammals (*91*), most of them (47%) were positively correlated with the sense transcript and only 2% were anticorrelated. These observations indicate a predominance of double-stranded RNA-dependent mechanism and that they generally decrease the stability of the sense gene. We confirmed and extensively described the wide-spread presence of antisense transcription in *M. pneumoniae.* We provided an exhaustive ncRNA catalog but the range of molecular mechanisms and functions will require further research. However at least in two cases, NEW87 and NEW8, we could see specific regulation of non-overlapping ncRNAs that could indicate functionality. In the case of NEW8 we saw a perfect anticorrelation in expression with the neighbor opp peptide importer operon, suggesting some relationship to metabolism regulation (Fig. 16). In the case of NEW87 its location after the origin of replication and its specific regulation by DNA damage, suggested a potential role in DNA replication and repair (Fig. 17). Both of them are conserved in *M. genitalium* (Fig. 16 and 17).

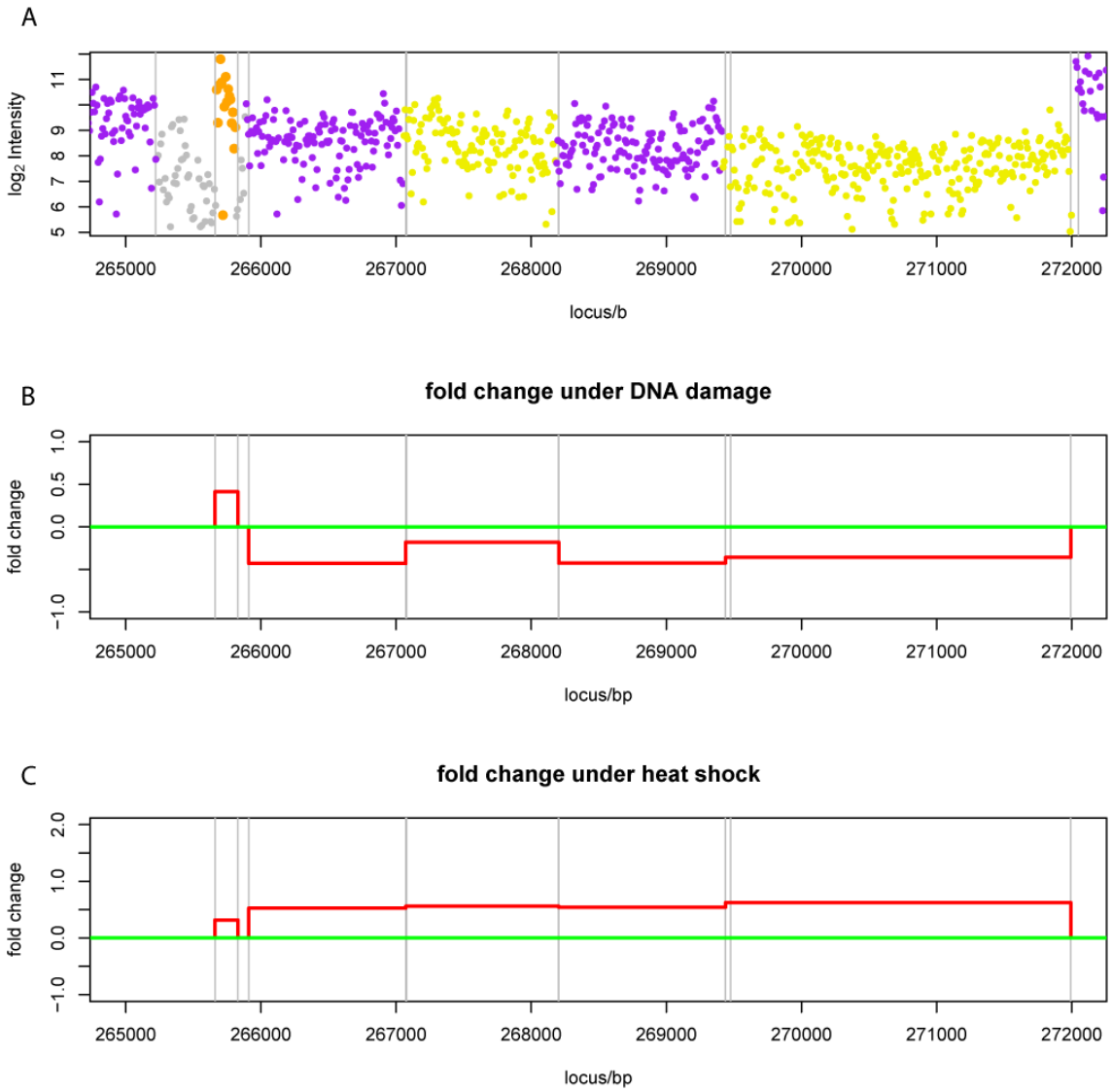


Figure 15. The expression of some small RNA were sometimes independent from that of neighboring genes. A) The tiling array profile of the region containing the small RNA NEW8 (orange dots) and its neighboring genes (alternating purple and yellow dots) under the reference condition. B) Log_2 of the fold change in the intensity at each position for Mitomycin C (MMC) treatment. NEW8 was up-regulated while the downstream genes were down-regulated. C) Log_2 of the fold change in the intensity at each position for a heat shock (HS) treatment. NEW8 is slightly upregulated as well as the downstream genes. Genes downstream of NEW8 encode different subunits of peptide transporter. Interestingly, a recent metatranscriptomics study (*99*) reported that the most predominant gene families flanking putative intergenic small RNAs included transporter genes involved in nutrient acquisition. Green reference condition normalized to zero. Red, relative changes in expression with respect to the reference.

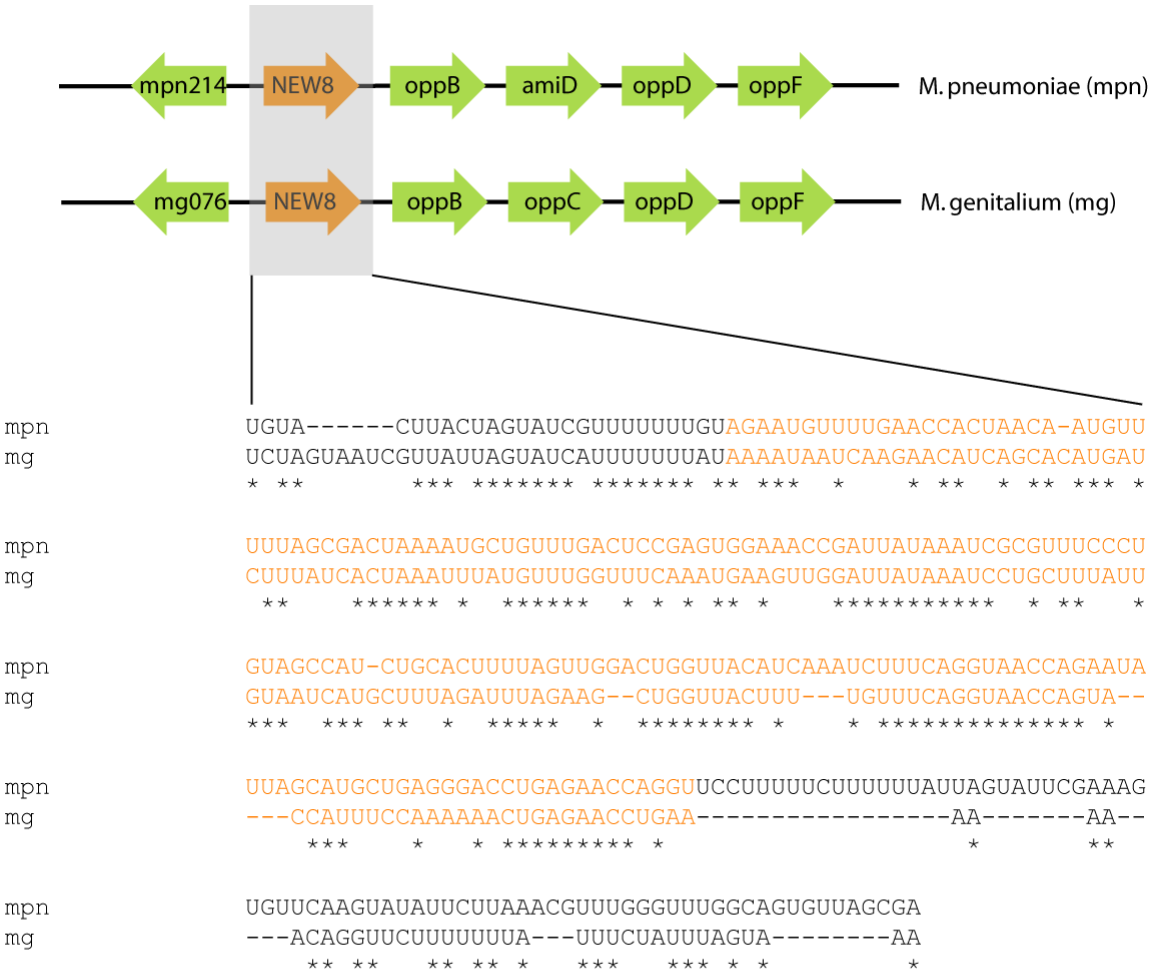
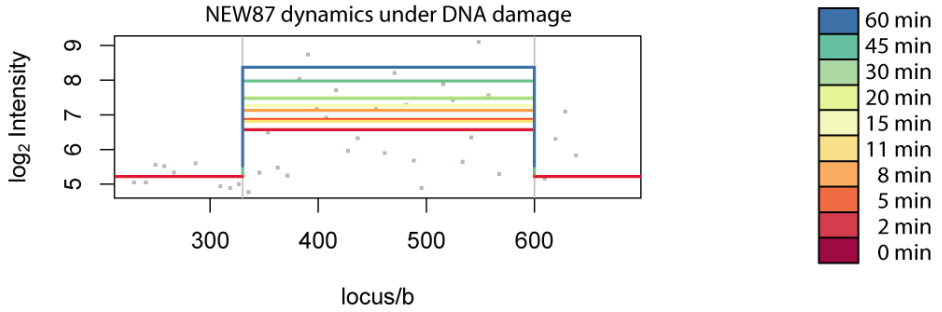


Figure 16. Conservation of a new ncRNA, NEW8, that is specifically regulated under different conditions (Fig. 14). Sequence similarity analysis revealed that NEW8 (in orange) is also present in *M. genitalium*. NEW8 genomic flanking genes are similar in both species. Genes located downstream encode different subunits of a peptide transporter system. Alignment has been carried out with T-Coffee Version 7.71, mode RCOFFEE.

a



b

```
mpn             ACUAUAAUUAAGUAUAAAUACU-UAAUUAUAGAUUAAUUUGGCUUUUAUUUCUGGAUGCC
mg              --UAUAAG-AACUAAAAUGUCUAUAAUUUUGUUCAUAAAAAGCUUAUAAUUAAGCAUAAA
                  *****  ** ** **   ** ***** *      *    **** ** **  * **

mpn             UA-UUAUAUAUAAUGUCAUGGUAGCCUUUCUUAUAUAUGUACUAUGUACAUAUA-UAGUA
mg              UGCUUAAUUAUAGUAAUAAUUAAUACUCUCUAAAAUAGAUACUAU-UAUAUAUAACAGAA
                *  ***  **** *   *    *  ** *** * ***  ****** ** *****  ** *

mpn             AAGGGCCAGAACAAACAACU-AAUUUUCUUCAUUGUUUAACCAUCUUGGUAUUAAAGGGA
mg              AUUAAUGAUAUUUUACAUAUGUACUAUGUACAU-AUGUAAUUAU--UACUGCUGAA--AA
                *      * *    ***  *  * * * * ***  * ***  **  *  *  * **   *

mpn             UAUUAAGUUAGAGCUAAGGGAAUA-GUGUUGGAGA-AU--ACUUUGAAUAAUACCCUAA-
mg              UAAUCAGUUCA-AUUGAUUAAAAUAUGUUUCAGAUAAAUAAAAUUUUUUUAAUCUGAUUAU
                ** * ****     * *   **** ** *  ** * **  * ***   ****    * *

mpn             -UAAUACUAACCACUUUAGUAUUGUCAUAAUAUCACCUA-UA
mg              UUAUUAUUCUACAC-AUAGUAUUAACUUAGUAUUAUGUAGAA
                 ** ** *   ***  *******  * ** *** *  **  *
```

Figure 17. Conservation and dynamics of NEW87. This non-coding RNA is located close to the origin of replication. We observe conservation in *M. genitalium* and regulation under DNA damage. The tiling array profile of the region containing the small RNA NEW87 (between blue vertical lines) under reference condition. Different adjusted continuous lines indicate changes in expression under mitomycin C treatment. We observe an increasing expression with the treatment. Alignment has been carried out with T-Coffee Version 7.71, mode RCOFFEE.

4.3 Proteome Organization in a Genome-Reduced Bacterium

In this article, *M. pneumoniae* protein architecture has been explored at different levels. Tandem affinity purification-mass spectrometry (TAP-MS) has been used to screen the protein complexes. A total of 178 complexes have been identified, of which the majority are novel. Different complementary structural information has also been provided which has complemented the biochemically purified complexes. Structural models for 484 proteins, single-particle electron microscopy and electron tomography linked the complexes to their structure and spatial organization.

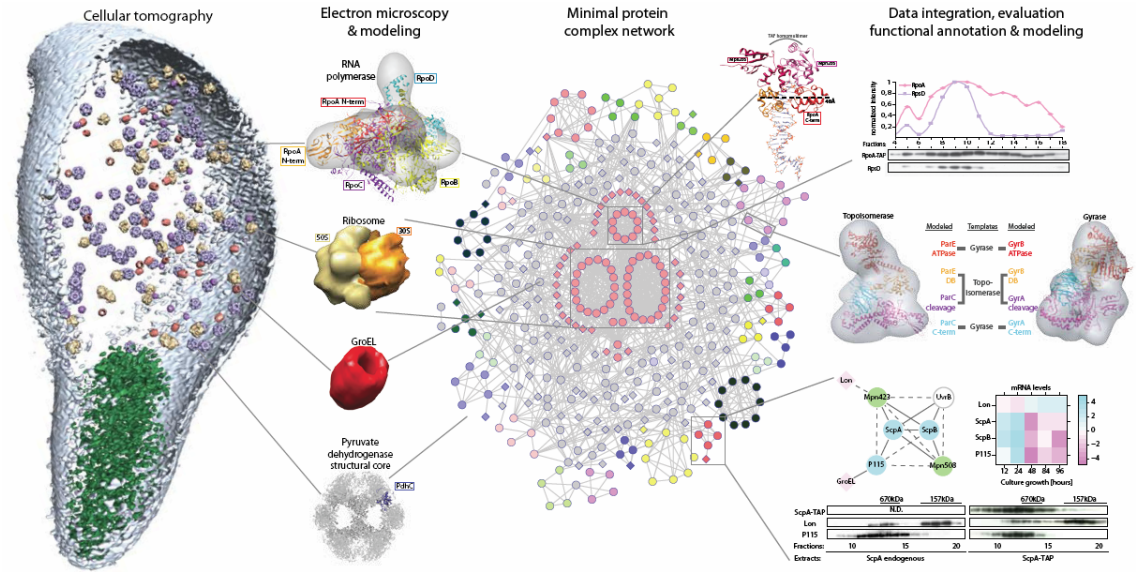


Figure 18. Scheme representing the workflow and results obtained in Kühner et al. (*4*)(obtained from Sebastian Kühner, EMBL).

The protein complement to *M. pneumonia* genome seems to build a complex network with modularity, cross-talk and a certain degree of reuse. The multifunctionality of one third of the complexes is reflected by the big number of physical interconnections. For instance, complexes reconstituting the ribosomes are intensively linked to RNA polymerase, proving the molecular bases of the described coupling between transcription and translation in bacteria (*100*). It is well established that an important number of proteins participate in different complexes (*101*). In mycoplasma, 156 are found to be multifunctional. An interesting example, involving two complexes with putative interchangeable units is described. GyrA, a component of the DNA gyrase complex, and ParE, a member of the topoisomerase IV complex are also both found in a complex together besides being also found in their respective canonic complexes. From a structural biology point of view, DNA topoisomerase and DNA gyrase complex have similar shapes (Fig. 18). Thus, they might be able to interchange units in certain situations.

Not only multifunctionality establishes some parallelisms with Eukaryotes, also some of the complexes have interesting analogies. The cohesin-like complex (Fig. 19), reproduces the yeast cohesin complex, essential in controlling the separation of chromosomes during mitosis (Fig. 19). This complex forms a ring that holds together the two chromosomes before the separase opens it and triggers anaphase (*102*). In *M. pneumoniae,* the cohesin-like complex includes, Mpn426, suggested as a putative Smc (structural maintenance of chromosomes) related to the cohesin (Smc1, Smc3 in *Saccharomyces cerevisiae*) protein in Eukaryotes, essential for the condensation of chromatin. Also includes ScpA and ScpB auxiliary proteins, which are related to the auxiliary proteins that close the ring in Eukaryotes (Scc1 and Scc3 in *S. cerevisiae*). Additionally, the Lon protease is also present. Lon could do the same role as the separase and permit the final separation of the chromosomes.

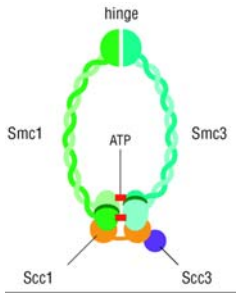


Figure 19. The yeast cohesin complex (obtained from Wikipedia.org)

This work significantly expands our knowledge of protein-protein interactions within bacterial cells. Protein-interaction networks are poorly correlated with genome architecture, suggesting several regulatory mechanisms in between these levels. Nevertheless, the main contribution of this work is the message that even in a minimal cell the proteome organization has important analogies to that in more complex organisms.

4.4 Impact of Genome Reduction on Bacterial Metabolism and Its Regulation

In this manuscript, *M. pneumoniae* metabolism is studied to an unprecedented level, since it goes far beyond the classical metabolic reconstructing papers. A manually curated metabolic network composed by 189 reactions catalyzed by 129 enzymes is provided. Analysis of such a network permitted the design of a defined medium which contains 19 essential nutrients. In comparison with more complex bacteria, *M. pneumoniae* metabolic network is shown to have a more linear topology and a higher frequency of multifunctional enzymes.

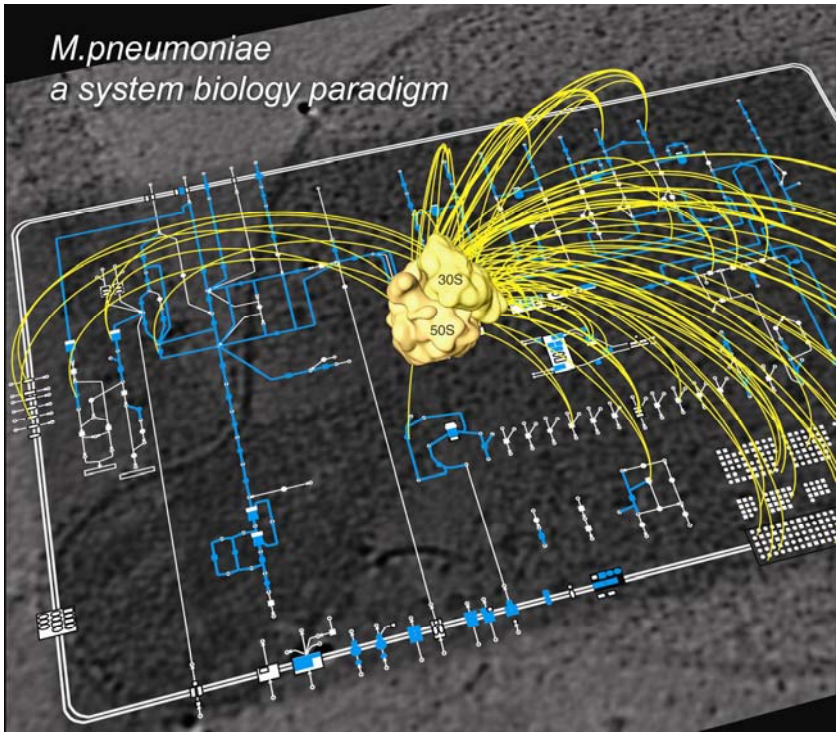


Figure 20. 3D representation of the interaction of the protein-protein interactions of ribosomal genes with metabolic enzymes (obtained from Takuji Tamada, EMBL).

As deduced during the minimal medium design, mycoplasma lacks most of the anabolic pathways. As carbon sources, glucose and glycerol are essential but not ribose, indicating that pentose phosphate pathway is active and it can provide sugars for nucleotides. Adenine and guanine need to be provided but cytidine is enough to deliver all other pyrimidines. Surprisingly, this indicates an important versatility in nucleotide metabolism. Aminoacids and peptides need to be also provided. The need of including peptides in the minimal medium is remarkable and can possibly be explained by the reduced number of amino acid transporters leading to competition between certain amino acids. Peptides could provide aminoacids by an independent route, namely the Opp ABC peptide transporter followed by intracellular digestion. Cholesterol and fatty acids are essential. In addition, a set of vitamins are also required for growth (nicotinate spermine, thiamine, pyridoxal, thioctic acid, riboblavin, choline, folic acid and coenzyme A/panthothenate).
In terms of regulation, specific metabolic responses have an important overlap to organisms with much bigger genomes. For instance, carbon starvation and stringent response show changes in anabolism and catabolism similar to those found in *L. lactis*

or *B. subtilis*. Although the individual changes under every different treatment are specific and show little overlap, the general response of a drastically genome-reduced bacterium is similar to much larger, more complex bacteria. Selection pressure acts at the phenotypic level and shapes the regulatory machinery in order to make the species efficient in its environment. Certainly, an equivalent regulation can be achieved by different circuitry and components. Yet, all the above indicate hitherto unknown additional regulatory mechanisms, perhaps via riboswitch-specific degradation of mRNAs, regulatory RNAs, changes in DNA supercoiling, or combinatorial role of a very few *cis*-acting elements.

Despite its intuitive simplicity, this works suggests that complex regulation can be achieved in a streamlined genome.

4.5 Correlation of mRNA and protein in complex biological samples

Preliminary work carried out by Tobias Maier (CRG) has determined the protein copy number of the majority of proteins in *M. pneumoniae*. Comparison with mRNA levels (Fig. 21) shows important divergences. This is an example of a more general trend in biology. The modest correlation between mRNA and protein in large scale datasets is explained by different biological and technical reasons. In this review, quantitative proteomics methods and quantitative transcriptomics are revised in detail. We also compile from the literature and discuss the different biological parameters influencing mRNA-protein correlation. It is shown that mRNA-protein correlation is complex since depends on various biological and technical factors.

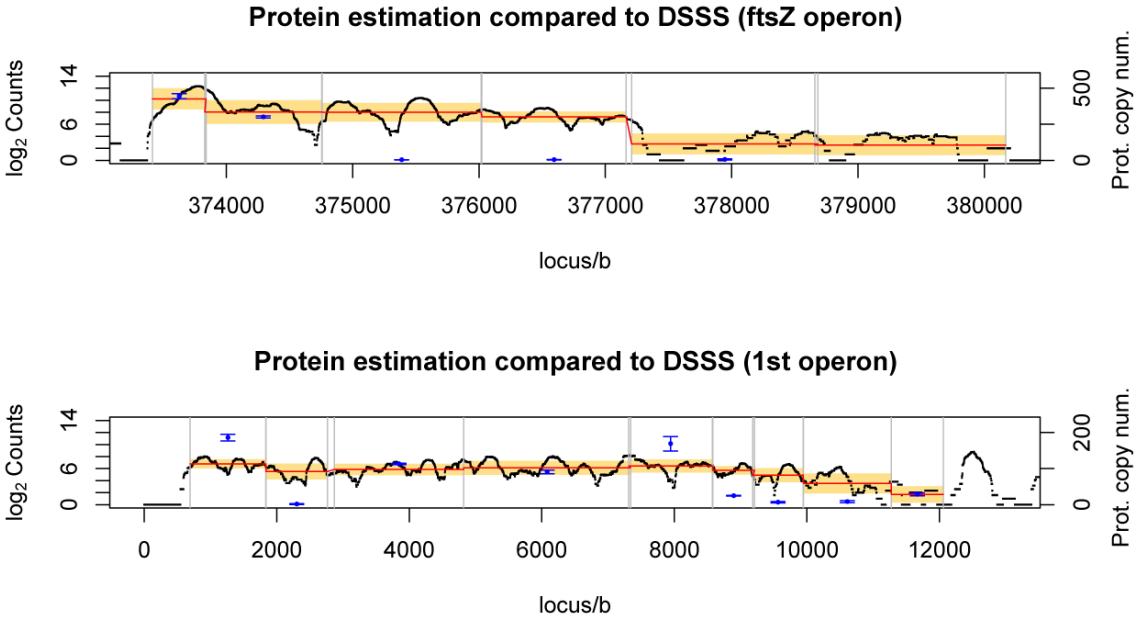


Figure 21. In orange RNA normalized copy number by DSSS (*3*) and in blue protein copy number estimated by Mass-Spectrometry (Data adquired by Tobias Maier, CRG). Proteins and the corresponding mRNA are not always correlated.

5 PERSPECTIVE

Even for one of the simplest organisms that can be grown axenically such as *M. pneumoniae,* our knowledge is far from being complete. As detailed along this document, the best technologies and analysis have been used to understand a simple organism. Most likely, we have provided one of the biggest datasets for a single organism ever build (*103*). In case of having the capacity to estimate the ratio of known information over unknown for all organisms, *M. pneumoniae* should be one of the highest. Nevertheless, this big step has delivered more open questions than we expected.

In case of focusing in transcription, we basically detailed the components and their dynamics. On one hand, for the first time, strand specific technologies were used to map the different transcripts present in the cell (operons, ncRNAs, …). On the other, a deep functional study was carried out so that all genes were interrogated in front of different insults. The first dataset provided information on the physical nature of the transcripts and the second about their phenomenological behavior. Afterwards, deep analysis was carried out in order to elucidate the molecular mechanisms dictating the observed behavior. In certain cases, we succeeded. For instance, we obtained a -10/-35 region motifs, Hrca binding motifs or a fur related motif. Different questions remain though.

Transcription termination remains obscure. It is clear that the main mechanism for transcription termination is the previously described ρ-independent mechanism, mediated by a hairpin followed by a polyU tail. However, we observe an important percentage of transcription termination independent of such a mechanism.
Transcription initiation also remains difficult to explain in certain cases. An important number of genes, mostly encoding lipoproteins or adhesins, possess a conserved promoter with absence of either -10 region motif or -35 region motif. With our current knowledge, we cannot explain the transcription of these highly expressed genes. These are only two examples of important aspects that we cannot understand yet.
We hypothesize that genomic neighborhood and genomic coordinates play an essential role in gene expression. Promoters and terminators are affected by near expression through molecular crowding, timing of DNA synthesis and alteration of DNA structure. Observing the strand specific expression data obtained, an obvious spatial correlation of forces is detected. Transcription seems to be significantly influenced by local genomic context and expression in one strand, induces important effects on the other inside a certain window. Also, most of the genes of the forward strand are located within the first half of the genome, and most of the reverse strand genes are in the second half. However, deconvoluting the different laws dictating gene expression still remains challenging.

Different experiments are planned to fill all these gaps. ChIP-Seq experiments are expected to help in identifying the targets of different putative transcription factors. A more detailed examination of the transcription starts is expected to provide more information on transcription initiation. Of special interest is the examination of the ‘abortome’. DSSS uncovered us the aspect of the abortive transcription which seems to be an excellent predictor of transcription initiation. Different cloning strategies are expected to indicate more details about the different mechanisms underlying transcription termination. An extension of the current dataset to *M. genitalium* should also reveal important aspects through comparison. In addition, we are planning to use

computational methods to analyze expression depending on the spatial location and neighborhood.

We definitely closed a chapter, but the conclusions obtained just boosted us to an amazing new horizon of research and challenges.

6 CONCLUSIONS

1. *M. pneumoniae* has been established as a model organism in Systems Biology.

2. Transcriptome analysis uncovered 117 previously undescribed, mostly non-coding transcripts in *M. pneumoniae*, 89 of them in antisense.

3. Based on tiling array data, 341 operons could be identified in *M. pneumoniae*, 139 of them polycistronic; almost half of the latter show decaying expression in a staircase-like manner. Under various conditions, operons can be divided into 447 smaller transcriptional units, resulting in many alternative transcripts.

4. Frequency of antisense transcription, alternative transcripts, and multiple regulators per gene imply a highly dynamic transcriptome, more similar to that of eukaryotes than previously thought.

5. DSSS, an strictly strand-specific protocol for transcriptome sequencing has been described. It provides a high dynamic range, single base resolution and low background enabling antisense ncRNA detection and precise transcript characterization.

6. DSSS is applicable both in prokaryotes and eukaryotes.

7. A manual curation of the metabolic network of *M. pneumoniae* has resulted in 189 reactions and 129 enzymes, allowing the design of a minimal medium with 19 essential ingredients.

8. Experimental characterization of metabolic network and its dynamics shows that its metabolic adaptation and responses are similar to other more complex bacteria.

9. *M. pneumoniae* show a high degree of proteome organization reflected by 62 homomultimeric complexes and 116 hereomultimeric soluble complexes.

10. Integration of the different datasets imply that *M. pneumoniae* harbors an unexpected complexity with frequency of alternative transcripts, antisense transcription, a small but tighly regulated metabolic network and a high level of proteome organization.

11. Different techniques are available to enable high-throughput quantitative measurements both at mRNA and protein level.

12. Protein abundances and their corresponding RNA are not directly correlated. The absence of correlation it is dictated by a regulatory layer composed by different regulatory mechanisms.

7 BIBLIOGRAPHY

1. E. Klipp, *Systems biology in practice : concepts, implementation and application*. (Wiley-VCH, Weinheim, 2005), pp. xix, 465 p.
2. M. Guell *et al.*, *Science* **326**, 1268 (Nov 27, 2009).
3. A. P. Vivancos, M. Güell, J. C. Dohm, L. Serrano, H. Himmelbauer, *Submitted*, (2009).
4. S. Kuhner *et al.*, *Science* **326**, 1235 (Nov 27, 2009).
5. E. Yus *et al.*, *Science* **326**, 1263 (Nov 27, 2009).
6. T. Maier, M. Guell, L. Serrano, *FEBS Lett*, (Oct 20, 2009).
7. V. J. Martin, D. J. Pitera, S. T. Withers, J. D. Newman, J. D. Keasling, *Nat Biotechnol* **21**, 796 (Jul, 2003).
8. H. Kitano, *Science* **295**, 1662 (Mar 1, 2002).
9. G. Project, http://www.1000genomes.org.
10. L. Hood, J. R. Heath, M. E. Phelps, B. Lin, *Science* **306**, 640 (Oct 22, 2004).
11. A. D. Weston, L. Hood, *J Proteome Res* **3**, 179 (Mar-Apr, 2004).
12. J. C. Venter *et al.*, *Science* **291**, 1304 (Feb 16, 2001).
13. E. S. Lander *et al.*, *Nature* **409**, 860 (Feb 15, 2001).
14. R. Sorek, P. Cossart, *Nat Rev Genet* **11**, 9 (Jan).
15. U. Nagalakshmi *et al.*, *Science* **320**, 1344 (Jun 6, 2008).
16. L. David *et al.*, *Proc Natl Acad Sci U S A* **103**, 5320 (Apr 4, 2006).
17. D. S. Johnson, A. Mortazavi, R. M. Myers, B. Wold, *Science* **316**, 1497 (Jun 8, 2007).
18. B. Ren *et al.*, *Science* **290**, 2306 (Dec 22, 2000).
19. B. Di Ventura, C. Lemerle, K. Michalodimitrakis, L. Serrano, *Nature* **443**, 527 (Oct 5, 2006).
20. M. M. Wosten, *FEMS Microbiol Rev* **22**, 127 (Sep, 1998).
21. E. Perez-Rueda, J. Collado-Vides, *Nucleic Acids Res* **28**, 1838 (Apr 15, 2000).
22. S. R. Goldman, R. H. Ebright, B. E. Nickels, *Science* **324**, 927 (May 15, 2009).
23. D. Zhou, R. Yang, *Cell Mol Life Sci* **63**, 2260 (Oct, 2006).
24. S. Borukhov, J. Lee, O. Laptenko, *Mol Microbiol* **55**, 1315 (Mar, 2005).
25. S. Borukhov, V. Sagitov, A. Goldfarb, *Cell* **72**, 459 (Feb 12, 1993).
26. E. Nudler, M. E. Gottesman, *Genes Cells* **7**, 755 (Aug, 2002).
27. C. L. Squires, D. Zaporojets, *Annu Rev Microbiol* **54**, 775 (2000).
28. I. Gusarov, E. Nudler, *Mol Cell* **3**, 495 (Apr, 1999).
29. V. Epshtein, D. Dutta, J. Wade, E. Nudler, *Nature* **463**, 245 (Jan 14).
30. S. Adhya, *Sci STKE* **2003**, pe22 (Jun 3, 2003).
31. Y. Makita, M. Nakao, N. Ogasawara, K. Nakai, *Nucleic Acids Res* **32**, D75 (Jan 1, 2004).
32. M. J. de Hoon, Y. Makita, K. Nakai, S. Miyano, *PLoS Comput Biol* **1**, e25 (Aug, 2005).
33. P. Babitzke, T. Romeo, *Curr Opin Microbiol* **10**, 156 (Apr, 2007).
34. M. Mandal *et al.*, *Science* **306**, 275 (Oct 8, 2004).
35. D. G. Gibson *et al.*, *Science* **319**, 1215 (Feb 29, 2008).
36. P. Ball, *Nature* **448**, 32 (Jul 5, 2007).
37. J. I. Glass *et al.*, *Proc Natl Acad Sci U S A* **103**, 425 (Jan 10, 2006).
38. C. A. Hutchison *et al.*, *Science* **286**, 2165 (Dec 10, 1999).
39. A. W. Rodwell, A. Mitchell, in *The Mycoplasmas: Cell Biology,* M. F. Barile, S. Razin, Eds. (Academic Press, New YHork, 1979), vol. I, pp. 103-140.

40. J. T. Regula *et al.*, *Microbiology* **147**, 1045 (Apr, 2001).
41. J. D. Jaffe, H. C. Berg, G. M. Church, *Proteomics* **4**, 59 (Jan, 2004).
42. D. C. Krause, M. F. Balish, *FEMS Microbiol Lett* **198**, 1 (Apr 20, 2001).
43. J. B. Baseman, M. Lange, N. L. Criscimagna, J. A. Giron, C. A. Thomas, *Microb Pathog* **19**, 105 (Aug, 1995).
44. T. R. Kannan, J. B. Baseman, *Proc Natl Acad Sci U S A* **103**, 6724 (Apr 25, 2006).
45. S. Halbedel, C. Hames, J. Stulke, *J Bacteriol* **186**, 7936 (Dec, 2004).
46. S. Halbedel *et al.*, *J Mol Biol* **371**, 596 (Aug 17, 2007).
47. J. Weiner, 3rd, R. Herrmann, G. F. Browning, *Nucleic Acids Res* **28**, 4488 (Nov 15, 2000).
48. K. M. Hallamaa, G. F. Browning, S. L. Tang, *J Bacteriol* **188**, 5393 (Aug, 2006).
49. J. Weiner, 3rd, C. U. Zimmerman, H. W. Gohlmann, R. Herrmann, *Nucleic Acids Res* **31**, 6306 (Nov 1, 2003).
50. G. A. Benders, B. C. Powell, C. A. Hutchison, 3rd, *J Bacteriol* **187**, 4542 (Jul, 2005).
51. E. Laing, V. Mersinias, C. P. Smith, S. J. Hubbard, *Genome Biol* **7**, R46 (2006).
52. K. M. Hallamaa, S. L. Tang, N. Ficorilli, G. F. Browning, *BMC Microbiol* **8**, 124 (2008).
53. M. D. Ermolaeva, H. G. Khalak, O. White, H. O. Smith, S. L. Salzberg, *J Mol Biol* **301**, 27 (Aug 4, 2000).
54. J. D. Pollack, *Trends Microbiol* **5**, 413 (Oct, 1997).
55. M. Pachkov, T. Dandekar, J. Korbel, P. Bork, S. Schuster, *Gene* **396**, 215 (Jul 15, 2007).
56. R. Himmelreich *et al.*, *Nucleic Acids Res* **24**, 4420 (Nov 15, 1996).
57. J. Dahl, *J Bacteriol* **170**, 2022 (May, 1988).
58. H. L. Worliczek, P. Kampfer, R. Rosengarten, B. J. Tindall, H. J. Busse, *Syst Appl Microbiol* **30**, 355 (Jul, 2007).
59. M. L. Klement, L. Ojemyr, K. E. Tagscherer, G. Widmalm, A. Wieslander, *Mol Microbiol* **65**, 1444 (Sep, 2007).
60. S. R. Schmidl *et al.*, *Infect Immun* **78**, 184 (Jan).
61. O. Musatovova, S. Dhandayuthapani, J. B. Baseman, *J Bacteriol* **188**, 2845 (Apr, 2006).
62. C. Condon, C. Squires, C. L. Squires, *Microbiol Rev* **59**, 623 (Dec, 1995).
63. B. R. Bochner, P. C. Lee, S. W. Wilson, C. W. Cutler, B. N. Ames, *Cell* **37**, 225 (May, 1984).
64. U. Romling, *Sci Signal* **1**, pe39 (2008).
65. D. Monleon *et al.*, *Proteins* **66**, 726 (Feb 15, 2007).
66. H. Himeno *et al.*, *Nucleic Acids Res* **32**, 5303 (2004).
67. A. Seybert, R. Herrmann, A. S. Frangakis, *J Struct Biol* **156**, 342 (Nov, 2006).
68. U. Gobel, V. Speth, W. Bredt, *J Cell Biol* **91**, 537 (Nov, 1981).
69. A. M. Collier, W. A. Clyde, Jr., *Am Rev Respir Dis* **110**, 765 (Dec, 1974).
70. U. Radestock, W. Bredt, *J Bacteriol* **129**, 1495 (Mar, 1977).
71. S. Seto, G. Layh-Schmitt, T. Kenri, M. Miyata, *J Bacteriol* **183**, 1621 (Mar, 2001).
72. J. I. Glass, C. A. Hutchison Iii, H. O. Smith, J. C. Venter, *Mol Syst Biol* **5**, 330 (2009).
73. H. Ochman, R. Raghavan, *Science* **326**, 1200 (Nov 27, 2009).

74. A. Budin-Verneuila, E. Maguin, Y. Auffraya, S. Dusko Ehrlich, V. Pichereaua, *le Lait* **84**, 8 (2004).
75. E. Yus *et al.*, *Accompanying manuscript*.
76. P. Alifano *et al.*, *Microbiol Rev* **60**, 44 (Mar, 1996).
77. H. J. Lee, H. J. Jeon, S. C. Ji, S. H. Yun, H. M. Lim, *J Mol Biol* **378**, 318 (Apr 25, 2008).
78. M. de la Mata, A. R. Kornblihtt, *Nat Struct Mol Biol* **13**, 973 (Nov, 2006).
79. T. Koide *et al.*, *Mol Syst Biol* **5**, 285 (2009).
80. M. F. Duffy, I. D. Walker, G. F. Browning, *Microbiology* **143 (Pt 10)**, 3391 (Oct, 1997).
81. W. Huber, J. Toedling, L. M. Steinmetz, *Bioinformatics* **22**, 1963 (Aug 15, 2006).
82. Z. Xu *et al.*, *Nature* **457**, 1033 (Feb 19, 2009).
83. X. J. Wang, T. Gaasterland, N. H. Chua, *Genome Biol* **6**, R30 (2005).
84. S. R. Henz *et al.*, *Plant Physiol* **144**, 1247 (Jul, 2007).
85. X. Ge, Q. Wu, Y. C. Jung, J. Chen, S. M. Wang, *Bioinformatics* **22**, 2475 (Oct 15, 2006).
86. M. Lluch-Senar, M. Vallmitjana, E. Querol, J. Pinol, *Microbiology* **153**, 2743 (Aug, 2007).
87. F. Perocchi, Z. Xu, S. Clauder-Munster, L. M. Steinmetz, *Nucleic Acids Res* **35**, e128 (2007).
88. A. Vivancos, M. Güell, L. Serrano, H. Himmelbauer, *Submitted for publication*, (2009).
89. R. Yelin *et al.*, *Nat Biotechnol* **21**, 379 (Apr, 2003).
90. G. G. Carmichael, *Nat Biotechnol* **21**, 371 (Apr, 2003).
91. S. Katayama *et al.*, *Science* **309**, 1564 (Sep 2, 2005).
92. M. Lapidot, Y. Pilpel, *EMBO Rep* **7**, 1216 (Dec, 2006).
93. J. Tomizawa, T. Itoh, *Proc Natl Acad Sci U S A* **78**, 6096 (Oct, 1981).
94. E. G. Wagner, R. W. Simons, *Annu Rev Microbiol* **48**, 713 (1994).
95. G. Andre *et al.*, *Nucleic Acids Res* **36**, 5955 (Oct, 2008).
96. J. M. Silvaggi, J. B. Perkins, R. Losick, *J Bacteriol* **187**, 6641 (Oct, 2005).
97. J. A. Opdyke, J. G. Kang, G. Storz, *J Bacteriol* **186**, 6698 (Oct, 2004).
98. S. Brantl, *Curr Opin Microbiol* **10**, 102 (Apr, 2007).
99. Y. Shi, G. W. Tyson, E. F. DeLong, *Nature* **459**, 266 (May 14, 2009).
100. J. Gowrishankar, R. Harinarayanan, *Mol Microbiol* **54**, 598 (Nov, 2004).
101. J. Hodgkin, *Int J Dev Biol* **42**, 501 (1998).
102. F. Uhlmann, F. Lottspeich, K. Nasmyth, *Nature* **400**, 37 (Jul 1, 1999).
103. http://mympn.crg.es.