Integrative understanding of transcription in a minimal cell model

Verónica Lloréns Rico

TESI DOCTORAL UPF / ANY 2016

THESIS DIRECTORS

Dr. Luis Serrano Pubul & Dra. Maria Lluch Senar

DEPARTMENT

EMBL/CRG Systems Biology Unit

Centre for Genomic Regulation (CRG)

Universitat
Pompeu Fabra
Barcelona

A mis padres y mi hermano

Acknowledgements

I would like to use these lines to express how grateful I am to those people that have contributed to the realization of this thesis, with their ideas, their time or their unlimited support (or all of them).

First of all, I want to thank my two supervisors, Luis and Maria, who first gave me the opportunity to join this lab and then guided me throughout my PhD, sharing their invaluable experience and advice. I can hardly describe my admiration for you. You both are such inspiring scientists and demonstrate an incredible passion for science; I truly hope that the path I will be starting after my graduation will turn me into such an inspiring scientist as you. Maria, thank you for being always positive and showing me the silver linings when things do not work as expected. I also want to acknowledge the members of my thesis advisory committee: Manuel Irimia, Toni Gabaldón, Javier Macía and Mark Isalan, for their always constructive feedback and valuable ideas, and for their advice on my future career.

I also want to thank every member of the Serrano lab for the ever-lasting discussions, beach volleyball games, crazy ideas and parties and video shootings, but especially for the unlimited support through the ups and downs of the PhD. Javi and Maria, you have been the two main pillars I had within the lab ever since I arrived. We have spent many hours together: in the lab, I which I have learned a lot from you, from Mycoplasma biology to C++ programming; but also outside the lab with so many italian dinners in Poblenou and Mario Kart games. Furthermore, I want to thank all (current and past) lab members with whom I have worked, especially Eva, with whom I have been closely collaborating over my entire thesis to decipher gene regulation in Mycoplasma; and Jonathan, with whom I could share two months in Stanford (besides countless hours of Skype) learning about the whole-cell model. Also Jaime, who did a great job in manually annotating sRNAs in several bacteria. I am also grateful to the entire Mycoplasma team for their fruitful ideas and feedback in discussions: Samuel (best beach-volley captain ever), Marc, Marie, Caro, Raúl, Sira, Dan, Adrià, Carlos and Tony. I want to acknowledge the signaling team as well, especially Christina for her trust when I collaborated in one of her projects, but also Violeta, Hannah, Claudia, Claire, Martin and Jae-Seong. Thanks to all of you for contributing to a great atmosphere in the lab, it is a pleasure to work with you.

Quiero también dar las gracias a las personas con las que he compartido tantas y tantas horas durante estos años y que han sido tanto fuente de inspiración como foco de muchos desahogos. En primer lugar, a mis compañeras de tesis Vicky, Lisa y Maria, por todo lo que hemos pasado juntas; a la familia del Tupper Club, por dejarme pertenecer a tan selecto grupo y hacerme comer cada día mucho mejor que en muchos restaurantes. Dentro de esta familia quiero hacer mención especial a Juan y Vicky, los mejores compañeros de piso que se pueden desear, mil gracias por aguantarme y entenderme cuando he estado desbordada; pero sin olvidarme del resto: Marc, Marco, Diego, Lisa, Alicia, Lara, Àlex, Nino, Marina, Pierre, Fede y Maria. Tampoco podía faltar aquí mi otra familia, la de la Barceloneta, la SUP family que tantas alegrías me ha dado en los dos últimos años en los que he descubierto un deporte, el Stand Up Paddle, que se ha convertido en mi otra gran pasión. Son muchas las horas que hemos compartido en el agua, remando, riendo o compitiendo, y también en tierra, comiendo y bebiendo. Gracias por hacer que no sólo haya descubierto un deporte sino una familia nueva: Miquel, Dah, Victor, Carlo, Maria, Cesc, Luis, Hellen, Sara, Tati, Raquel, Marta, Gwen, Bernat, Enric, Marce y Esther, entre muchos otros.

Finalment, vull donar les gràcies als que no puc veure molt sovint, però sé que em donen tot el suport del món des de la terreta, la meua família. En primer lloc, als meus pares, que sempre han estat a l'altre costat del telèfon per alegrar-se dels meus èxits i també per a preocupar-se quan les coses no anaven del tot bé. Vosaltres sou els que sempre haveu cregut en mi (inclús quan jo mateixa no ho feia), m'haveu demostrat que tot s'aconsegueix amb esforç i m'haveu ensenyat que l'únic límit que tindré a la vida és aquell que em pose jo mateixa. Moltíssimes gràcies. També al meu germà Diego, que m'alegra amb les seues visites de tant en tant i ha sigut company d'aventures els últims estius. No vull acabar sense donar les gràcies a la resta dels RiPé, que em donen i em donaran suport i ànims allà on vaja.

Abstract

One of the major challenges of biology is to understand how entire cells or organisms behave in homeostasis and in response to perturbations. Whole-cell modeling promotes this understanding by integrating different cellular processes in a single model that is able to predict emergent cellular behaviors. In this thesis, we have developed the first whole-cell model of the genome-reduced bacterium *Mycoplasma pneumoniae*, which encodes for less than 700 protein-coding genes. This model follows the structure of the previously described model in *Mycoplasma genitalium*. However, the lack of comprehensive knowledge of even these simple organisms limits the predictive power of these models. To address this problem and improve the model, we have focused in the process of transcription regulation, and we have studied the major determinants of transcript abundance in this bacterium. Therefore, we have characterized promoters and the role of small RNAs. We have also reconstructed the gene regulatory network, revealing that non-transcription factor regulation may have a large impact in coordinating RNA levels in *M. pneumoniae*. Furthermore, by analyzing the 'omics' data used to investigate the process of transcription and to fit the whole-cell model, we have found different biases of high-throughput profiling experiments, and we have described that chimeric RNAs identified in these datasets may be artifacts generated in RNA-sequencing experiments.

Resumen

Uno de los mayores retos de la biología es entender cómo células u organismos completos se comportan tanto en homeostasis como en respuesta a perturbaciones. El campo del modelado de células completas pretende comprender esto integrando diferentes procesos celulares en un único modelo capaz de predecir comportamientos celulares emergentes. En esta tesis, hemos desarrollado el primer modelo de célula completa de la bacteria de genoma reducido *Mycoplasma pneumoniae*, que codifica para menos de 700 proteínas. Este modelo sigue la estructura del descrito previamente en *Mycoplasma genitalium*. Sin embargo, la falta de conocimientos exhaustivos incluso para estos organismos simples limita el poder predictivo de estos modelos. Para hacer frente a este problema y mejorar el modelo, nos hemos centrado en el proceso de regulación de la transcripción, y hemos estudiando los principales factores que determinan la abundancia de tránscritos en esta bacteria. Así, hemos caracterizado los promotores y el papel de ARNs pequeños. También hemos reconstruido la red de regulación génica, observando que la regulación no debida a factores de transcripción puede tener un gran impacto en la coordinación de los niveles de ARN en *M. pneumoniae*. Además, analizando los datos procedentes de experimentos ómicos usados para investigar el proceso de la transcripción y ajustar el modelo, hemos encontrado diferentes sesgos en estos experimentos a gran escala u 'ómicos', y hemos descrito que ARNs quiméricos que se identifican en estos datos pueden ser artefactos generados en experimentos de secuenciación de ARN.

List of publications

reLluch-Senar, M., Luong, K., Lloréns-Rico, V., Delgado, J., Fang, G., Spittle, K., ... & Serrano, L. (2013). Comprehensive methylome characterization of Mycoplasma genitalium and Mycoplasma pneumoniae at single-base resolution. *PLoS Genet*, *9*(1), e1003191.

Lloréns-Rico, V., Serrano, L., & Lluch-Senar, M. (2014). Assessing the hodgepodge of non-mapped reads in bacterial transcriptomes: real or artifactual RNA chimeras?. BMC genomics, 15(1), 1.

Lloréns-Rico, V., Lluch-Senar, M., & Serrano, L. (2015). Distinguishing between productive and abortive promoters using a random forest classifier in Mycoplasma pneumoniae. *Nucleic acids research*, gkv170.

Lluch-Senar, M., Delgado, J., Chen, W. H., Lloréns-Rico, V., O'Reilly, F. J., Wodke, J. A., ... & Ferrar, T. (2015). Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Molecular systems biology*, *11*(1), 780.

Lluch-Senar, M., Cozzuto, L., Cano, J., Delgado, J., Llórens-Rico, V., Pereyre, S., ... & Serrano, L. (2015). Comparative “-omics” in Mycoplasma pneumoniae Clinical Isolates Reveals Key Virulence Factors. *PloS one*, *10*(9), e0137354.

Kiel, C., Benisty, H., Lloréns-Rico, V., & Serrano, L. (2016). The yin-yang of kinase activation and unfolding explains the peculiarity of Val600 in the activation segment of BRAF. *eLife*, e12814.

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W. H., ... & Lluch-Senar, M. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. *Science advances*, *2*(3), e1501363.

Junier, I., Unal, E. B., Yus, E., Lloréns-Rico, V., & Serrano, L. (2016). Insights into the mechanisms of basal coordination of transcription using a genome-reduced bacterium. *Cell Systems*.

Contents

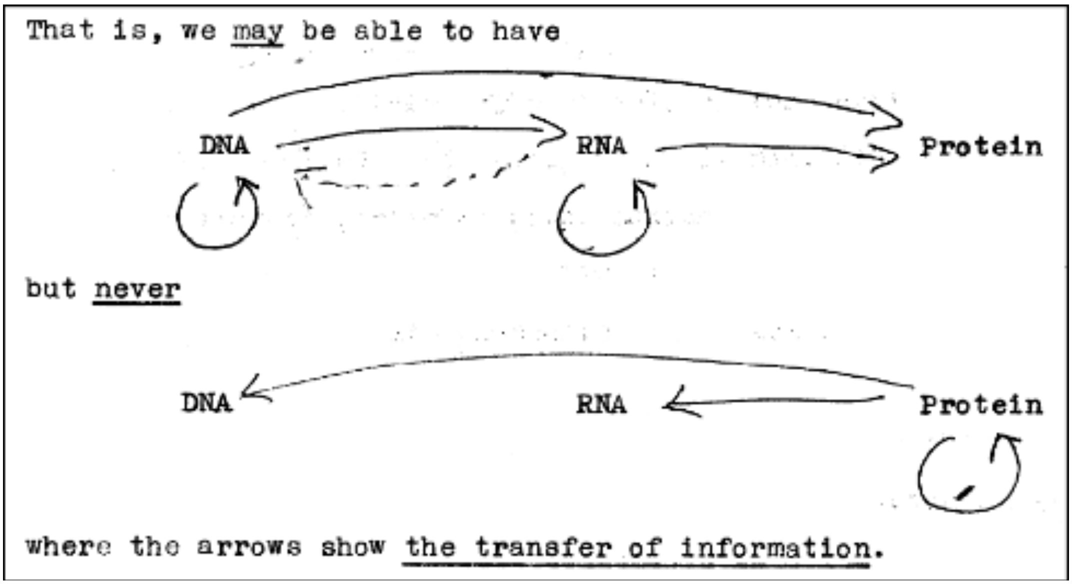
List of figures

List of tables

1. Introduction

1.1. Prokaryotic transcription

Transcription is the process by which a fragment of DNA is used as a template to be copied into a molecule of RNA. According to the Central Dogma of molecular biology, enunciated by Francis Crick for the first time in 1958 (*1*), transcription is the first step in the process of gene expression, by which the genetic information stored in the DNA is executed by proteins. The Central Dogma establishes a series of causal relationships among DNA, RNA and proteins (Figure 1.1). DNA can be replicated in a process that involves the action of different proteins, but not RNA. Also, DNA can serve as a template for the transcription of RNA molecules. These RNA molecules themselves will act as templates for the production of proteins, the final step of gene expression. The Central Dogma also establishes that RNA can self-replicate and that can be transcribed back to DNA. Both mechanisms have been shown to occur in different viruses (*2–5*). However, the direct production of proteins from DNA, also proposed by this model, has not been shown in any living organism. The Central Dogma also defines the transitions that can never happen within cells: from protein to either DNA or RNA, and from protein to protein.



That is, we <u>may</u> be able to have

but **never**

where the arrows show <u>the transfer of information</u>.

Figure 1.1. Excerpt of Francis Crick's paper 'On protein synthesis', describing the Central Dogma of Molecular Biology

With the exception of the direct transfer of information from DNA to proteins, the Central Dogma has remained unchanged over the years. However, new discoveries have recently posed some challenges to the absoluteness of this model. Some argue that prions, proteins in a misfolded conformation capable of misfolding other proteins present in a soluble conformation (*6–8*), would be transferring the information from protein to protein in a non-conventional manner (*9*), thus breaking one of the statements of the Dogma (Figure 1.1). Furthermore, there are bacterial proteins capable of synthesizing small peptides (*10*). This synthesis occurs in a ribosome-independent manner. Although not considered proteins, these peptides are biologically active; some of them are toxins, antibiotics, or surfactants, for instance (*11*). Despite these debates regarding the Central Dogma, the process of transcription from DNA to RNA is considered to be universal.

Being the first step of gene expression, transcription is a key cellular process in all kingdoms of life. However, there are differences in how it is performed in prokaryotes and in eukaryotes. In bacteria, transcription occurs in the cytoplasm, due to the absence of a membrane-delimited nucleus, and it is coupled to the process of translation into proteins (*12*). The transcription process in bacteria can be subdivided in three different events: initiation, elongation and termination. During initiation, the protein complex that synthesizes the RNA from the DNA template, the RNA polymerase, binds a sigma factor to form the RNA polymerase holoenzyme (see Chapter 1.1.2) and recognizes a region in the DNA, called the promoter. The RNA polymerase is formed by 4 polypeptides, named α (2 copies), β and β' (*13*, *14*). The structure formed by the RNA polymerase holoenzyme and the double stranded DNA is called the closed complex. After the binding event, the holoenzyme unwinds the double helix of the DNA to form the open complex (*15*). The polymerase starts synthesizing the RNA, but at this point the enzyme is not processive, as the sigma factor blocks the elongation. Eventually, the sigma factor dissociates from the complex and transcription elongation can occur. In transcription termination, RNA polymerase stops the elongation of the nascent transcript, which is released by various mechanisms (see Chapter 1.1.6).

Such a crucial process must be tightly regulated at different points, to guarantee that RNA levels are maintained in homeostasis and that cells are able to respond correctly and accurately to external perturbations. In the following subsections, different factors governing transcriptional regulation in bacteria are reviewed.

1.1.1. Gene organization in operons.

In bacteria, genes are not presented as isolated units of information. Instead, they are organized in blocks of related functions, termed operons. The name 'operon' and the first description of these blocks should be attributed to François Jacob and Jacques Monod (16, 17). The operon is the "genetic unit of co-ordinate expression", and can be defined as a set of genes of related function, located in the genome in a consecutive manner. These genes are transcribed together in the same RNA molecule, and are under the control of the same regulatory sequences (Figure 1.2).

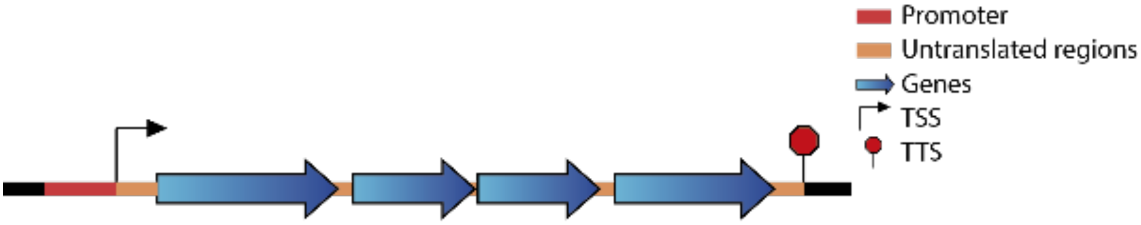


Figure 1.2. Scheme depicting the prototypical structure of a bacterial operon. Delimited by the transcription start site (TSS) and the transcription termination site (TTS), this operon contains four genes.

This organization in blocks implies a first level of gene regulation in bacteria. Genes of related function have been kept close during evolution, and many operons are conserved across different bacteria (*18*). This simplifies the need of specific regulation for each individual gene. However, expression of genes in the same operon is not always equitative. It has been shown that positional effects exist, and in operons with several genes, expression levels of the individual proteins are proportional to the distance to the transcription initiation site. That is, genes that are closer to the end of the operon show lower protein levels (*19*). This can be explained by the fact that transcription and translation occur simultaneously in bacteria (*12*). Translation of the first genes of the operon can occur while the last genes have not been transcribed yet, generating the differences in protein expression. This implies that the organization in blocks not only simplifies global regulation, but it also fine-tune controls the levels of the different proteins encoded in the same operon (*19*).

For many years, operons have been treated as static entities. However, recent research has shown that these structures are highly dynamic, being able to adapt in response to changing conditions. Accordingly, many operons in different bacteria have been divided in sub-operons, smaller groups within operons, whose genes are always expressed as a single unit (under the set of conditions tested) (*20*, *21*). The increased complexity and operon plasticity observed in the study of several conditions has led to the usage of the term 'transcription units' (*22*). Transcription units are RNAs containing one or more genes under the control of the same regulators, and they are usually condition-dependent. In some conditions they coincide with the classical 'operon' definition, but in others they differ. Plasticity of transcription units arises from the multiple transcription start sites and termination sites that exist within operons (*23*), but how are these entry and exit points of transcription regulated remains unknown in most conditions.

1.1.2. Bacterial promoters and transcription initiation.

Entry points of transcription are called 'transcriptional start sites', or TSS. These points are preceded by a short genomic region called 'promoter'. The RNA polymerase complex, the protein complex that transcribes a region of the genomic DNA into an RNA molecule, has to recognize and bind the promoter region prior to starting the transcription process. Promoter regions require certain features that make them recognizable by the RNA polymerase and other proteins (transcription factors) that may control the expression of the following transcription units.

One of the features that defines promoters is the exact sequence of nucleotides located upstream the TSS. This sequence is specific for a class of proteins, sigma factors (σ), that bind both this region and the RNA polymerase. There is a housekeeping sigma factor, that binds and controls the expression of the majority of exponential growth-related genes, and there are several alternative sigma factors that control the transcription of stress-related genes (*24*). According to the sequence of nucleotides upstream the TSS, a specific sigma factor will bind a given promoter. The housekeeping sigma factor (called σ^{70} in the model bacterium *Escherichia coli*) binds two regions: one located 10 bases upstream the TSS termed the -10 box or the Pribnow motif, and one located 35 bases before the TSS called the -35 box. The Pribnow motif has the consensus sequence TANAAT (where N is any base) (*25*), whilst the -35 box has the consensus motif TTGACA (*26*). Other sigma factors recognize different motifs. Furthermore, there are other sequences that, although non-essential for protein recognition, affect the binding of the RNA polymerase complex. These include the extended Pribnow motif (*27*, *28*) and the -45 region or UP element (*29*). The location of these motifs is represented in Figure 1.3, together with the regions of the RNA polymerase holoenzyme that recognize them. Besides these sequences, there are other motifs that are unique for different transcription factors, that regulate specific sets of genes. These additional sequences can be located upstream or downstream the TSS.

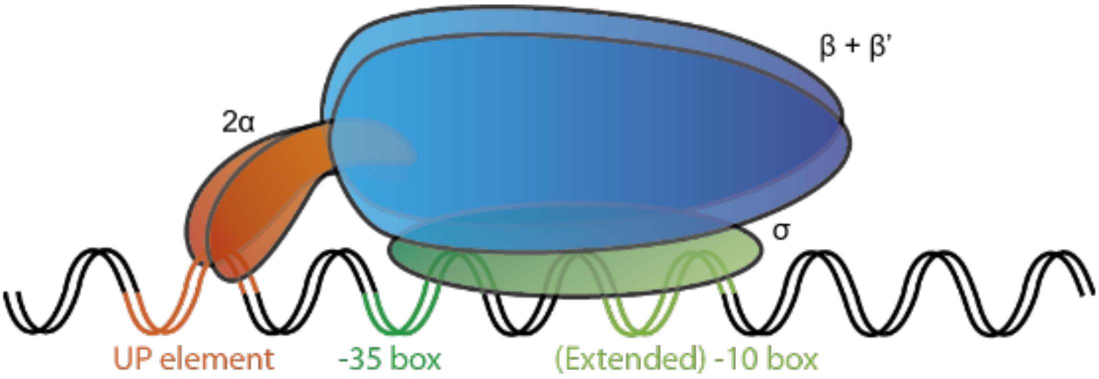


Figure 1.3. Schematic structure of the RNA polymerase holoenzyme, formed by the RNA polymerase complex (2αββ') and the sigma factor (σ). The alpha subunits bind the UP element, whilst the sigma factor recognizes the -35 and the (extended) -10 boxes

In addition to the specific motifs defining promoters, the structure of these regions is also important to trigger transcription. For example, the spacing between the -35 and the -10 motifs affects the expression levels of the downstream genes, and the optimal spacer length differs among bacteria (*30*–*32*). Furthermore, the double-stranded DNA needs to unwind specifically at the promoter region to accommodate the RNA polymerase complex prior to initiating transcription. Thus, the double helix should be less stable at this point, and the melting energy of the DNA at this point should be low (*33*, *34*).

Finally, promoter accessibility also influences transcript expression. This accessibility is controlled by the supercoiling and bending of the DNA at the promoter region, which is ultimately governed by the activity of various proteins: DNA gyrases and topoisomerases. Certain promoters require a specific degree of negative or positive supercoiling to be able to form a pre-initiation complex with the transcriptional machinery (*35*, *36*), and mutations in these promoters deactivate this requirement.

Mutations and changes in the the regions recognized by the RNA polymerase complex or transcription factors can affect the efficiency of the transcription process and alter the expression of the corresponding genes (*37*–*39*) (Yang et al, in preparation). In addition to mutations in the DNA sequence, epigenetic marks such as methylation of DNA could influence how the transcriptional machinery interprets promoter information. Little is known about the specific effects of DNA methylation, but it has been associated to stress response and virulence in many bacterial species (*40*). Indeed, an enrichment in methylation points has been observed in virulence-related genes in various bacterial species (*41*).

1.1.3. Transcription factors.

Transcription factors (TFs) are (DNABPs) that bind to a specific sequence of DNA, and upon that binding they regulate the expression of one or more genes, either activating or repressing their transcription. In bacteria, TFs bind to the promoter regions of transcription units, whilst in eukaryotes, they can also bind enhancers, regions located far from the genes they regulate. To describe the dual relationship between TFs and the regulatory sequences they bind, the former are also called trans-acting regulators, whilst the latters are named cis-acting regulators.

One important class of bacterial transcription factors corresponds to the aforementioned sigma factors, proteins needed to initiate the process of transcription (*24*, *42*). In *E. coli*, part from the housekeeping σ^{70}, other proteins of this class regulate genes that respond to heat stress (σ^{32} or σ^{24}) (*43*-*44*) or starvation (σ^{38} or σ^{54}) (*45*, *46*), or that control the expression of flagellar genes (σ^{28}) (*47*) or genes involved in iron transport (σ^{19}) (*48*). A single sigma factor binds the RNA polymerase complex to form the RNA polymerase holoenzyme and initiate transcription, so alternative sigma factors have to compete with the housekeeping σ^{70} for the binding (*49*). A peculiarity of sigma factors is that, although they bind the polymerase, once transcription is initiated and the first bases of the RNA have been polymerized, they dissociate from the transcription complex (*50*).

Apart from sigma factors, other TFs regulate the expression of specific subsets of genes. These proteins may activate or repress the transcription of the genes they regulate, and the regulatory sequences they bind can be either upstream or downstream the TSS (*51*). Some TFs of them form a complex with the RNA polymerase, whilst others do not. There are TFs that regulate the expression of other TFs. Altogether, they form a transcriptional regulatory network (Figure 1.4).

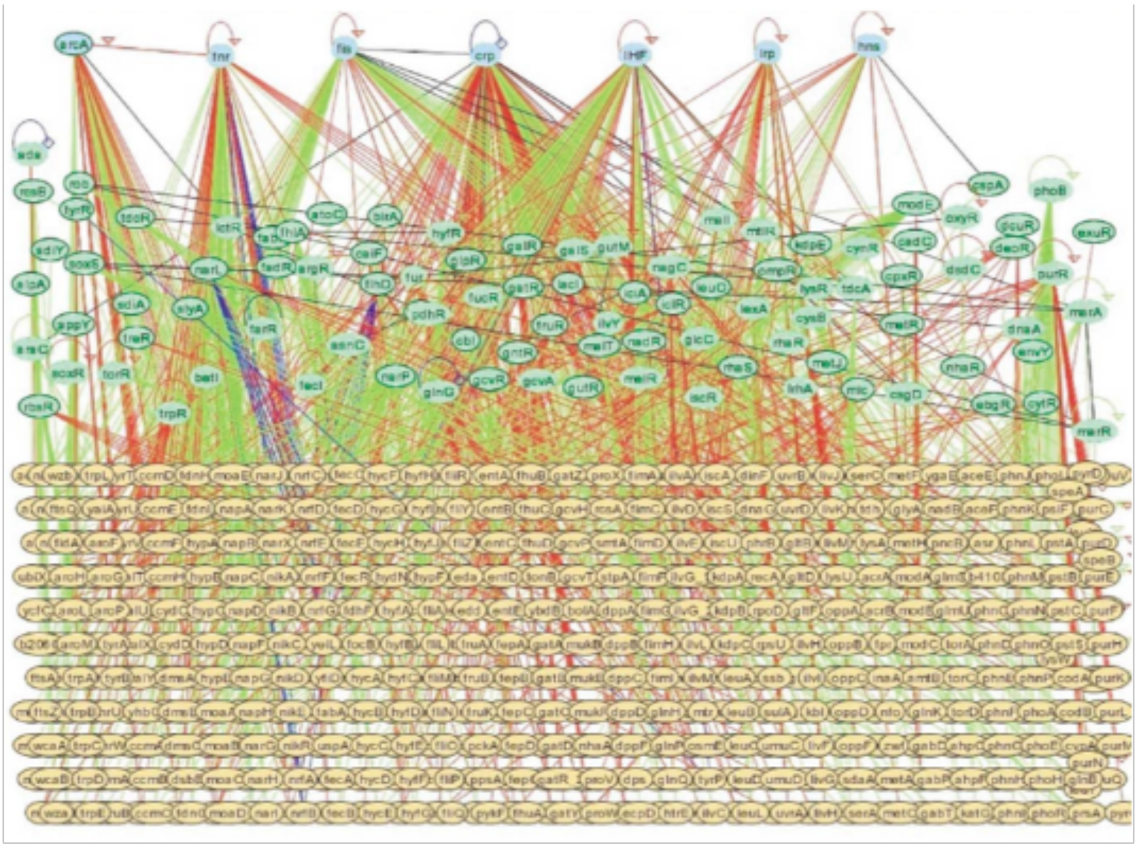


Figure 1.4. Schematic representation of the gene regulatory network of *E. coli*. Blue nodes represent global regulators, green nodes represent transcription factors, and yellow nodes represent target genes. Edges between the nodes represent the hierarchical relationships in the network. Figure extracted from Martínez-Antonio and Collado-Vives (2003)

Transcriptional regulatory networks (also called gene regulatory networks) are not random. Instead, they exhibit a notable hierarchy. Within these networks, it is possible to distinguish a small subset of TFs that are global regulators. These global regulators not only modulate the expression of a number of transcription units, but also control the transcription of other TFs and sigma factors, being on top of the hierarchy. For instance, in *E. coli* only 7 global regulators have a direct control over the transcription of 51% of the genes in this bacterium (*52*). This underscores the importance of TFs and the networks they form in regulating transcript expression.

Having such a large impact in transcriptional control, studying how these networks are wired not only provides a fundamental knowledge of how the transcript expression landscape is shaped, but also sets the basis for modification and engineering of the system.

1.1.4. The regulatory role of small RNAs.

The transcripts synthesized in bacterial cells not only encode for genes (mRNAs), tRNAs, rRNAs or other functional RNAs such as the 4.5S or 6S RNAs (*53*, *54*). There are RNAs that do not encode for any of these functions, and have been named 'small RNAs' (sRNAs). This term was initially limited to RNA molecules of ~50-200 nucleotides (*55*), but more recently this definition has expanded, including transcripts of thousands of nucleotides (*56*).

Initial studies reported the existence of a few dozens of sRNAs in different bacteria (*55*, *57*), but recent advances in computational prediction and especially in transcriptome profiling (see Chapter 1.2) have outnumbered initial estimations. Indeed, several authors suggest now that this non-protein-coding transcription is pervasive throughout the genome (*58*, *59*).

Two major categories of sRNAs have been described. *Cis*-encoded sRNAs are found overlapping functionally defined genes (that is, protein-coding genes, tRNAs or rRNAs). They can be found in the coding strand (sense sRNAs) or in the opposite strand, thus named antisense RNAs or asRNAs. The other major class corresponds to *trans*-encoded sRNAs, which are located in intergenic regions and do not overlap any gene.

Since their discovery, deciphering the function of these RNA molecules has become a major field of study, yet only a handful of sRNAs has been functionally characterized to date. Some participate in bacterial defense mechanisms against phage infections, such as the sRNAs of the CRISPR systems (*60*). In this setup, a non-protein-coding RNA molecule (termed crRNA), can target a specific sequence of the DNA/RNA of the phage, and direct an enzyme (Cas) to cleave this sequence. However, these RNAs are not naturally encoded in the bacterial genome, but acquired upon previous infections of the phage. Thus, the system works as a bacterial adaptive immunity (*61*).

Apart from this exception, the majority of sRNAs in bacteria is thought to have a regulatory function, acting at a transcriptional or post-transcriptional level and controlling RNA and protein levels of target genes. Thus, they could be key players in the gene regulatory networks shaping the transcriptional landscape (*62*). The majority of sRNAs act via complementary base-pairing with their mRNA targets. This binding is in some cases autonomous, nor requiring the participation of other elements than the RNAs involved, or dependent on the action of RNA chaperones, such as the Hfq protein (*63*, *64*).

Upon binding of the sRNA with its mRNA counterpart, multiple effects can be triggered. Some sRNAs affect the stability of their mRNA targets, either by favoring their degradation (*65*) or by preventing it (*66*). Sometimes, the sRNA binds the mRNA at the ribosome binding site (RBS), thus preventing its recognition and translation by the ribosome (*67*). There are other examples in which the binding of the sRNA induces a conformational change that exposes a previously hidden RBS, therefore facilitating translation (*68*). Some asRNAs regulate expression via transcription attenuation (*69*), by inducing a premature termination of transcription of the mRNA.

Another possible action mechanism does not involve the formation of the duplex mRNA-sRNA, but the sole act of transcribing the sRNAs. This is the case of transcriptional interference: here, two polymerases transcribing in opposite directions from convergent promoters may collide, resulting in one of them (or both) being released of the chromosome, generating a truncated transcript (*70*).

Despite the variety of mechanisms of action described, only a minority of the discovered sRNAs has been characterized, most of which correspond to the *trans*-encoded sRNAs. One of the challenges found in the characterization of sRNAs is that the many of them are present in very low copy numbers (*71*, *72*). They are usually overexpressed for their functional characterization, but this may give rise to artifactual responses, not seen at physiological sRNA levels where stochastic effects may predominate. Furthermore, it has recently been shown that some of these sRNAs are actually coding for small peptides (*73*). Further studies will be necessary to uncover the regulatory potential of these molecules and the role they play in modulating the gene regulatory networks of bacteria.

1.1.5. Regulation by metabolites.

In addition to transcription factors and sRNAs, metabolism also has an impact on transcription. The building blocks of RNA are nucleosides tri-phosphate, or nucleotides (NTPs), and their levels can determine the outcome of the process of RNA synthesis, thus establishing a link between metabolism and transcriptional regulation. This link is mediated by the promoter sequence and the stability of the complex formed between the promoter and the RNA polymerase. Unstable complexes require high concentrations of NTPs so that RNA synthesis can be launched immediately. Otherwise, the complex rapidly dissociates and the transcription event is not produced.

In contrast, very stable complexes require smaller concentrations of NTPs, as they will not easily dissociate (*74*).

More specifically, it is not the global NTP concentration the responsible for this regulation, but the NTP concentration of the initiating nucleotide in the RNA. The inclusion of this +1 nucleotide stabilizes the open complex and prevents the dissociation of the polymerase, allowing for transcription to continue (*74*). Later, it has been shown that also the +2 nucleotide participates in the sensing of NTP concentration (*75*). Concrete examples of this nucleotide-based regulation include the response to amino acid starvation (stringent response) in *Bacillus subtilis*. In this scenario, concentration of ATP increases whilst GTP decreases. Upregulated genes in this condition usually have adenosine in the +1 position, whilst downregulated promoters have guanosine. Mutations in the +1 base change the behavior of these genes, suggesting that the regulation of this condition relies on the NTP sensing (*76*). In *E. coli*, the rRNA operons (*rrn*) initiate with ATP or GTP, whose concentrations are dependent on the growth rate of the bacterium (*74*, *77*)

Besides NTPs, other metabolites can exert a regulatory effect on transcription. Some act as signalling molecules produced in certain stress conditions, called alarmones. The most common alarmone is guanosine tetraphosphate or pentaphosphate ((p)ppGpp). This molecule is involved in stringent response in bacteria. When a lack of aminoacyl-tRNAs causes the translating ribosome to stall, (p)ppGpp is synthesized (*78*). This molecule inhibits the transcription of a number of genes, and activates the synthesis of many others, changing the transcriptional program of the cell. In *E. coli*, (p)ppGpp is synthesized by the RelA protein in response to amino acid starvation, and degraded via the action of the protein SpoT. SpoT also has minor synthase activity (*79*). (p)ppGpp binds to the RNA polymerase (*80*) and inhibits the transcription of rRNA by competing with the GTP at position +1 of this RNA, thus preventing the open complex from stabilizing and preventing transcription initiation (*81*). Sometimes, the effect of (p)ppGpp can be mediated by other proteins, such as DksA in *E. coli (82)*. Also, it has been shown that (p)ppGpp is involved in the synthesis and usage of alternative sigma factors other than the housekeeping σ^{70} (*83*, *84*). The mechanism in Gram positive bacteria presents some major differences. Firstly, there is only one RelA/SpoT protein homolog, that performs both the synthetase and hydrolase activities (*85*), although more recently, additional small genes with synthetase activity only have been reported in *B. subtilis (86)*. Secondly, in Gram positive bacteria, the effect of (p)ppGpp seems to be mediated by a decrease in the precursors used for the alarmone synthesis, rather than an direct effect of the alarmone binding the RNA polymerase. For instance, in *B. subtilis*, the activation of sporulation genes caused by starvation, occurs as a consequence of the decrease of GTP or GDP, used to synthesize the molecule (*87*). The decrease of GTP also results in the inhibition of the rRNA synthesis (*88*). This highlights again the importance of nucleotide abundances in regulating transcription.

Finally, a third class of regulatory metabolites binds nascent RNA molecules, in untranslated regions termed riboswitches. This binding induces a conformational change in the RNA that can trigger different effects, such as induced/avoided premature termination (transcriptional attenuation) or altered translation (*89*, *90*). Transcription attenuation is caused by the formation of a termination hairpin, which releases the RNA polymerase. Usually, this is not an absolute mechanism, but there is some read-through of the RNA polymerase that continues transcribing. Thus, this mechanism generates both long, 'normal' transcripts and short, prematurely terminated RNAs (*89*). Translation initiation is regulated by exposing or hiding the RBS as a consequence of the riboswitch refolding (*91*). In some cases, the riboswitch exposes regions of the RNA recognized by RNAses, regulating degradation of the RNA molecule (*92*).

1.1.6. Transcriptional termination.

Transcriptional bacterial termination can occur in two different forms, termed intrinsic (or Rho-independent) and Rho-dependent and termination. Intrinsic termination occurs when the nascent RNA forms a GC-rich hairpin structure, followed by a poly-uridine tract. This structure makes the elongating polymerase to pause, and eventually to release the transcript (*93*). Rho-dependent termination relies on a DNA/RNA helicase, the protein Rho (*94*). This hexameric protein binds RNA and has ATPase activity (*95*, *96*). The hexamer binds to nascent RNA with unstructured regions, with little or no defined secondary structure (*97*). Once bound, Rho uses its ATPase activity to produce the energy necessary to move along the RNA molecule and reach the region of the transcribing RNA polymerase, where nascent RNA and DNA form a duplex. There, the helicase activity of Rho unwinds the RNA-DNA duplex and releases the RNA (*98*). RNA polymerase is more processive than Rho. Thus, in order for Rho to reach the position of the polymerase, the latter needs to pause. Secondary structure of the nascent RNA generates this pausing (*97*).

The process of transcriptional termination can be modulated in different forms. Mutations in the DNA can alter the binding of Rho or disrupt the secondary structures needed for termination (*99*). Other mutations in the RNA polymerase affecting its processivity can have similar effects (*100*). In addition, there are antitermination factors that make the polymerase override the termination signals, both in Rho-dependent and in intrinsic termination. Some RNA chaperones, such as the cold-shock protein CspA from *E. coli* and its homologues in *B. subtilis*, prevent the formation of RNA hairpins in low-temperature conditions, thus avoiding the formation of premature termination sites (*101*, *102*). As discussed above, termination can also be mediated by small metabolites in riboswitches, via exposing or hiding termination sites. Finally, some proteins interfere with Rho function, such as the Psu protein from the bacteriophage λ (*103*). Psu binds to Rho and prevents its translocation along the RNA molecule (*103*, *104*). There are other termination modulators, such as the proteins from the Nus family of termination/anti-termination factors, that bind the elongating RNA polymerase altering its processivity, changing therefore its sensitivity to terminators. Nus proteins, such as NusA, can have opposite effects: on the one side, NusA can also recognize hairpins and increase the sensibility to termination (*105*). On the other side, NusA is also part of a transcription antitermination complex acting in rRNA transcription. This complex recognizes sequences located in the 5’ end of the nascent rRNA transcript (nus boxes), and binds to the transcribing polymerase at these sites, increasing the elongation rate and yielding the RNA polymerase insensitive to termination signals occurring throughout the rRNA (*106*–*108*).

1.1.7. RNA processing and degradation control.

So far, the regulating mechanisms for RNA production have been reviewed. But RNA degradation is as important as production, as ultimately RNA levels in bacterial cells will be determined by the equilibrium between production and decay. RNA degradation is not an arbitrary process, but it is orchestrated by multiple ribonucleases (or RNases). Two groups of RNases can be distinguished: exoribonucleases, which degrade RNA from one of its extremes (either in direction 5’ → 3’ or 3’ → 5’); and endoribonucleases, which cleave RNA internally. There are different exo- and endoribonucleases with different specificities.

Among all RNases, it is of interest to describe the RNA degradosome complex. This is a multimeric protein complex involved in the degradation of mRNA. This complex is formed by the exoribonuclease polynucleotide phosphorylase PnpA, the endoribonuclease RNase E (*109*, *110*), the helicase RhlB (*111*), and the glycolytic enzyme enolase (*112*). RNase E performs the rate-limiting step of degradation, the mRNA cleavage. This enzyme cleaves the RNA at specific sites to prevent further binding of ribosomes. As the mRNA-bound ribosomes finish translating the genes in the mRNA, RNase E further cleaves the mRNA, generating small fragments. PnpA degrades these mRNA fragments in direction 3' → 5' (*113*). The role of RhlB is to unwind secondary structures in the mRNA, facilitating the action of the PNPase (*111*). The function of the enolase in the complex is less clear, although it has been shown that it can be part of a regulatory feedback loop, controlling the degradation of the glucose transporter mRNA *ptsG.* When glycolysis is blocked, enolase is responsible for the rapid degradation of this specific mRNA (*114*).

In *B. subtilis* and other Gram positive bacteria, there are remarkable differences regarding the degradosome complex. RNase E has been replaced by the complex formed by RNases J1 and J2. These enzymes have both 5' → 3' exonuclease and endonuclease activities (*115*), and have been reported to contribute to rRNA maturation (*116*). There is an extra endonuclease, RNase Y, involved in the degradation of mRNAs and riboswitches (*117*). There is also a different RNA helicase in the *B. subtilis* degradosome, CshA (*118*). PnpA is also present, and besides enolase, there is another glycolytic enzyme forming part of this complex: phosphofructokinase (*119*).

Besides these central components of the degradosome, other proteins have been co-purified with this complex in minor proportions. These can act as modulators of the function of the RNA degradosome (*120*). Some proteins, such as RraA and RraB, are capable to bind the RNase E and remodel the degradosome complex, for example by releasing some of its components. This remodeling of the degradosome results in the specific stabilization of some transcripts and the selective degradation of others (*121*).

RNase E has another crucial role, which involves the degradation of mRNAs targeted by a sRNA (see Chapter 1.1.4 above) and bound by the RNA chaperone Hfq (*67*). In addition to the aforementioned ribonucleases, involved in the decay of mRNAs, there are other RNases specialized in the maturation and degradation of tRNAs (such as ribonuclease P (*122*)) or rRNAs (RNase III (*123*)). Also, ribonuclease HI is involved in cleaving the RNA primer that is used to initiate DNA replication (*124*), and RNase HII in degrading the Okazaki fragment primers resulting from the replication of the lagging strand (*125*).

1.2. Profiling bacterial transcriptomes with '-omics' technologies

The majority of the studies on transcriptional regulation described in the previous section were performed using classical molecular biology and biochemistry techniques. However, the sequencing of bacterial genomes has allowed for a change of paradigm in the way experiments are performed. The sequencing of the first bacterial genomes, around 20 years ago (*126–128*), led to the development of a set of techniques that allowed the profiling of full genomes and transcriptomes faster than ever imagined. In parallel, techniques were also developed to identify and quantify proteins in biological samples.

In this Chapter, these techniques, grouped under the name of '-omics', are described together with their contributions to the study of transcription and its regulation in bacteria.

1.2.1. Microarrays and tiling arrays.

Microarrays are lab-chips that use the same principle of nucleic acids hybridization that Northern (*129*) or Southern (*130*) blots. These blots allow for the detection and (relative) quantification of specific sequences of RNA or DNA, respectively. Northern blots were widely used to quantify the expression of mRNA transcripts in different biological samples. The protocol involves the extraction of the RNA fraction from the biological sample, the separation of the RNA molecules by size in a denaturing agarose gel and the transfer of the separated fragments from the gel to a nitrocellulose membrane (*131*). The blotted RNA will be the substrate for the hybridization. Afterwards, probes that are complementary to the sequence of the RNA of interest, labelled with radioactive isotopes or chemiluminescence, are added to the membrane containing the transferred RNAs. These probes can be either RNA, DNA or cDNA fragments, or synthetic oligos. Probes are added in conditions that favor the hybridization of them with the complementary RNA, and not with others. After hybridizing and washing to remove the excess of labelled probe, the membrane can be developed to identify and quantify the RNA of interest (*131*).

Microarrays use the same hybridization principle, but in a reversed setup. In microarrays, thousands of DNA fragments -the probes-, are attached to a solid surface, arranged in microscopic spots (the chip; Figure 1.5). These fragments contain sequences of the genes whose expression needs to be profiled. In this case, the target is the RNA fraction of the sample (usually converted to cDNA), which has been previously labelled. The sample is hybridized with the probes in the array, and thus the expression of hundreds to thousands of genes can be measured in a single experiment.

There are two major types of microarrays used to profile transcriptomes, called one-channel and two-channel detection arrays. In one-channel arrays (Figure 1.5), one labelled sample is hybridized against the chip, and the readout is the intensity of the spots, each corresponding to a hybridization reaction with a specific RNA. Knowing the unique sequence of each of the spots allows to map the intensity of a signal to a given gene, and thus to determine its expression. This expression is relative, and it always needs to be compared to a control in a separate array. In two-channel detection arrays, sample and control are labelled with different fluorescent markers, and are hybridized simultaneously with the same microarray. Relative expression of the sample and control will be given by the ratio in the fluorescence intensity of the two markers. In both cases, proper bioinformatics and statistical analyses of the results is required to extract valuable information on which genes change their expression patterns in certain conditions.

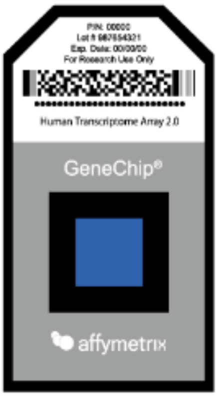


Figure 1.5. GeneChip microarray by Affymetrix, used to profile the human transcriptome. Source: Togo Picture Gallery. Togo Picture Gallery by DCBLS is licensed under a CC-BY 4.0 License

Microarrays became quite popular in the late 90s, after first genome sequences became available. In bacteria, they have been widely used to study changes in transcript abundances under a variety of conditions, in order to identify which genes regulate the responses to these conditions, and which are their targets (*132–134*). In contrast to their wide usage in identifying transcription factors and their targets, and under which conditions they are active, microarrays have had little or no application in the study of promoter strengths or transcription termination. However, they have been used to report RNA degradation in *E. coli*, measuring decay rates and RNA half-lives at a gene resolution (*135*), and to assess the effect of changes in the RNA degradosome in the decay of the different RNAs in the cell (*136*).

One particular inconvenient of microarrays is that they do not survey the entire genome, but instead the oligos arrayed in the chips only map to specific regions of protein-coding genes, or functional RNAs such as rRNAs or tRNAs. Therefore, small RNAs are not surveyed using this technology, and regulatory regions of the transcripts such as riboswitches are ignored. Also, probes have to be designed for each coding region, and this process may generate different biases. Factors such as probe length, hybridization temperature, and especially the avoidance of cross-hybridization need to be carefully considered in order to prevent problems derived from the probe design (*137*).

Tiling arrays appeared as an evolution of microarrays that solved some of these issues. Tiling arrays are also known as 'high density oligonucleotide arrays' (*138*). These arrays contain oligonucleotide probes mapping every few nucleotides in the genome, overlapping with each other ('tiled', hence their name). In these arrays, there is no distinction between coding and non-coding regions of the genome, allowing the identification and study of sRNAs and the characterization of untranslated regions (UTRs) of mRNAs (*139*). Bacterial genomes, smaller than eukaryotic genomes, are particularly well suited to these type of arrays, as one chip allows for the analysis of the entire genome. Some studies have surveyed large number of conditions and perturbations using tiling arrays in different bacteria. In *B. subtilis*, tiling arrays were used to characterize the transcriptional response in 269 conditions (*140*). This allowed to reconstruct the transcription unit architecture in this bacterium, as well as to identify condition-dependent sigma factors and their targets, leading to the reconstruction of part of the *B. subtilis* gene regulatory network. This work also identified large numbers of sRNAs, such as antisense RNAs responding to different perturbations. In *M. pneumoniae*, tiling arrays were used to characterize the transcriptome of this bacterium under a variety of conditions (*72*). One of the advantages of tiling arrays, compared to conventional arrays, is that with an appropriate experimental setup, they allow for absolute quantification of transcript levels (*141*). However, the challenges posed by the probe design remain unresolved in this technology.

One of the general principles of bacterial transcription uncovered thanks to tiling arrays is that the majority of the transcripts found in the cell are present in amounts slightly above the background noise levels (*142*). Thereby, the signal-to-noise ratio in these experiments is limited, hampering the quantification of lowly expressed transcripts in bacterial cells. Furthermore, their limit of detection is given by the amount of probe present in each spot. This quantity determines the maximum amount of cDNA that can hybridize. Over this point, saturation is reached and no more sample can be detected (*143*), thus limiting the dynamic range of arrays.

1.2.2. RNA-seq.

RNA sequencing (RNA-seq) was described for the first time in 2008 (*144*). It was derived from the whole genome shotgun DNA sequencing, the technique that was used to determine the sequence of the first bacterial genome, that of *Haemophilus influenzae (126)*. In this approach, genomic DNA is sheared and a specific size of fragment is selected. Fragments selected are amplified (originally this was done by cloning them into a vector) to obtain a large number of copies, and both ends of the fragments are sequenced. A bioinformatics analysis is then required to align the individual sequences generated, termed ‘reads’, and yield the assembled sequences. Currently, in the so-called next generation sequencing (NGS) protocols, fragments are not amplified in a vector, but in cell-free systems such as emulsion amplification (*145*), or solid-phase amplification (*146*). Also, the sequencing has changed. The original Sanger sequencing (*147*) was replaced by the newer pyrosequencing (*148*), sequencing-by-synthesis (*149*, *150*) or sequencing-by-ligation (*151*). The most extended technology today combines solid-phase amplification with sequencing-by-synthesis (Figure 1.6). Newer methods eliminate the amplification step to remove biases that may appear at this point, however they are not of general application yet (*152*, *153*).

Figure 1.6. Illumina HiSeq sequencer. Illumina technology combines solid-phase amplification with sequencing by synthesis. Source: Togo Picture Gallery. Togo Picture Gallery by DCBLS is licensed under a CC-BY 4.0 License

RNA-seq imitates the next generation techniques for sequencing DNA, but introducing a step of reverse transcription from RNA to cDNA. Once the cDNA is obtained from the sample, the rest of the protocol is similar. As with the DNA sequencing, the most used technology uses the solid-phase amplification and sequencing-by-synthesis. Originally, it was not possible to identify whether a read mapped to the plus or the minus strand of the genome, hampering the identification and quantification of antisense transcripts. Later, different modifications of the protocol allowed to identify and quantify transcripts in a strand-specific manner, allowing to decipher the polarity of each RNA (Figure 1.7) (*154–157*).

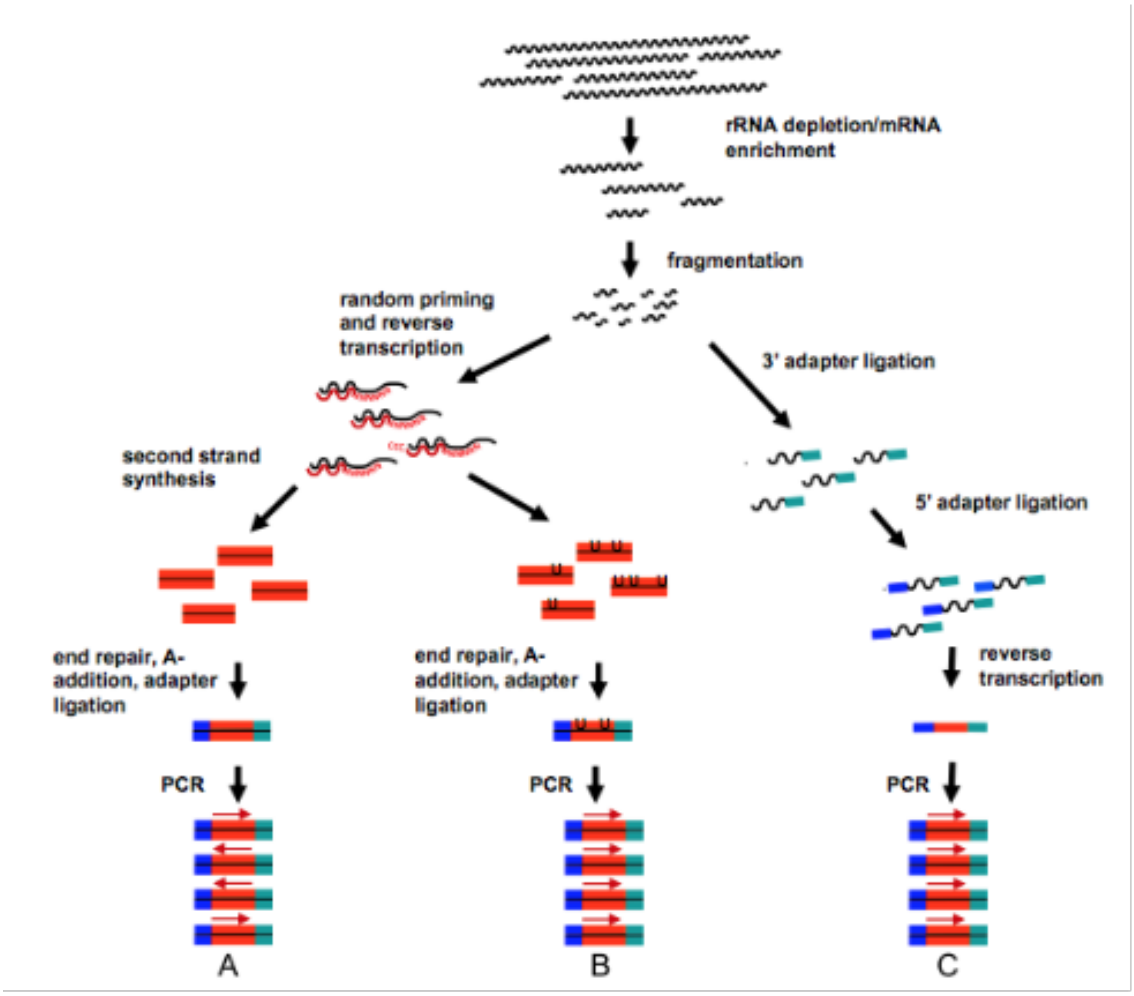

rRNA depletion/mRNA enrichment
fragmentation
random priming and reverse transcription
3' adapter ligation
second strand synthesis
5' adapter ligation
end repair, A-addition, adapter ligation
end repair, A-addition, adapter ligation
reverse transcription
PCR
PCR
PCR
A
B
C

Figure 1.7. Schematic view of different RNA-seq protocols. (A) Classical Illumina protocol, which does not hold the information of the polarity of the transcript. (B) Strand-specific protocol in which one strand is chemically modified with dUDP to retain the polarity of the transcript. (C) Strand-specific protocol in which the polarity of the transcript is kept by ligating different adapters in the 5' and 3' ends of the transcript. A key difference between B and C is the order in which reverse transcription and adapter ligation occurs; in B, reverse transcription happens first, whilst in C, adapters are ligated prior to reverse transcription. Image from van Dijk et al (2014).

Currently, RNA-seq has replaced arrays for transcriptome profiling, as it overcomes some of the disadvantages of these techniques. Firstly, it provides a single-base resolution, whilst resolution was limited to the spacing between probes in the tiling arrays. In addition, the amplification step yields a much larger dynamic range than in tiling arrays, permitting a more accurate quantification of transcripts (*158*). Also, the price of next generation sequencing has dropped dramatically, expanding the range of applications of these technologies (*159*).

RNA-seq has been used to study bacterial transcriptomes, with a variety of applications. As microarrays and tiling arrays, it has been used to reveal expression changes under varied conditions and perturbations (*160*), to study the process of RNA degradation (*161*) and to identify new sRNA species (*162*, *163*). However, the increased resolution of RNA-seq has allowed for its use in other applications, for example for the identification of transcriptional start sites or TSS (*157*, *164*). The mapping of these points in the genome has been used to define the transcription unit architecture in different bacteria, such as *E. coli (22)*, *Helicobacter pylori* (*164*) or *Mycoplasma pneumoniae* (*20*). In these works, the identification of alternative TSS within operons led to the description of condition-dependent transcription units. These transcription units provide the basis for the reconstruction of regulatory networks.

The accurate identification of TSS in bacterial genomes has also contributed to the study of naturally-occurring promoters, located upstream these points (*165*). In this study in *Sinorhizobium meliloti*, all TSS were identified and the sequences of their promoters were analyzed. In a different study, thousands of synthetic promoters were simultaneously screened using RNA-seq to evaluate the transcript levels yielded by each promoter (*166*), and assess the predictability of these levels based on sequence features of these promoters. In contrast to the study of TSS, little has been done for the high-resolution mapping of transcription termination sites in bacteria. Only very recently, a method has been described that allows the sequencing of the 3' ends of bacterial transcripts (*167*). One possible explanation is that termination, in contrast to initiation, does not occur at a single, defined point, but the mechanisms governing termination lead to individual transcripts terminating at different points located close after the termination signal (*168*). Finally, RNA-seq has also been used to describe regulation by riboswitches. In *Listeria monocytogenes*, the behavior of a riboswitch controlled by the abundance of vitamin B_{12} was described using RNA-seq (*169*). The activation of this riboswitch by vitamin B_{12} produces a shorter isoform of a sRNA, that ultimately regulates the expression of a set of proteins that use vitamin B_{12} as a cofactor.

Although RNA-seq has been widely used in studying transcription and the regulatory mechanisms that modulate this process, the fast and continuous development of the technique poses some challenges to researchers, especially related to the analysis of the data generated. The evolution of the sequencing technologies and protocols is usually faster than that of the analysis pipelines, which can lead to the obtention of artifactual results that are hard to identify. Bacterial genomes, smaller than eukaryotic ones, facilitate an increase in the sequencing coverage, but such an increase is accompanied by a risk of identifying artifactual transcripts or genomic DNA contaminants (*170*). Additionally, cDNA library preparation includes steps that may alter the sequencing output. For instance, biases in the amplification can generate large variability of RNA abundances across samples. This is especially notable in low-expressed transcripts (*171*). Also, the processes of RNA fragmentation and size selection, in which cDNA is selected according to its length in nucleotides, may cause the loss of sRNAs and other small transcripts such as tRNAs (*172*, *173*). Altogether, these challenges demand that computational analysis pipelines are designed as carefully as experimental and sequencing protocols.

1.2.3. ChIP

ChIP stands for 'chromatin immunoprecipitation'. This technique is used to isolate fragments of DNA that are bound by a specific protein (*174*, *175*). In more detail, DNA and its associated proteins are crosslinked by means of a crosslinking agent, for example formaldehyde. Then, chromatin is extracted from the biological sample and sheared by sonication or nuclease digestion, to obtain equally sized fragments of DNA. Some of these fragments are protein-free, whilst others are bound by a protein. An antibody specific to the protein of interest is then used to selectively bind and isolate it, together with the DNA fragments this protein is bound to (immunoprecipitation). Crosslinks are finally reversed to isolate the DNA fragment of interest. To characterize this fragment, the process of chromatin immunoprecipitation can then be coupled to sequence determination by means of arrays (ChIP-chip) (*176*, *177*) or, more commonly now, DNA-sequencing (ChIP-seq) (*178*). The distribution of signal/reads after ChIP-chip/seq experiments results in a pattern of peaks, each corresponding to a binding region. The binding can be direct, if the protein of interest interacts with the DNA, or indirect, if this interaction is mediated by other proteins. Different 'peak-calling' algorithms are used to identify and map the peaks resulting from the experiment (*179*, *180*).

ChIP-based techniques have been widely used to understand how proteins modulate the transcriptional function. In *E. coli*, ChIP coupled to tiling arrays was used to study the transition from transcription initiation to elongation in σ^{70} promoters at the genome scale (*181*). In this study, the authors identified a large variability in this transition, finding that in some cases this transition occurred rapidly, with a fast release of the σ^{70}

factor, whilst in others the polymerase was found stalled at the promoters. They suggested that the behavior of this transition was sequence-dependent. In many studies, ChIP-seq has been combined with RNA-seq or arrays to study both physical and regulatory interactions between transcription factors and their targets. In *Salmonella enterica*, ChIP-seq and RNA-seq were combined to uncover the targets of the regulator OmpR, involved in the virulence of this pathogen (*182*). In a broader study, ChIP-seq and tiling arrays were used to characterize the binding sites and the regulatory effect of 154 TFs in *Mycobacterium tuberculosis (183)*.

As the rest of high-throughput techniques, ChIP-based omics pose some challenges that need to be addressed. Firstly, there are some problems intrinsic to the chromatin immunoprecipitation. For instance, the availability of antibodies against the protein of interest can be a limiting factor. If an antibody is not available, ChIP can be performed expressing the protein of interest fused to a tag. This brings some issues related to the over-expression of the protein of interest: the effect of the tag on the binding and functionality of the protein is unknown *a priori*, and over-expression of the protein can cause off-target binding, leading to artifacts in the result. ChIP-chip resolution is limited by the array resolution, as occurred with arrays. Also, the hybridization step required in arrays may produce some bias in the results. Finally, arrays have a lower dynamic range that limits their use for quantification (*184*, *185*). These problems are overcome in ChIP-seq. However, sequencing protocols also suffer from biases in the amplification step that may produce artifacts, hard to distinguish by the peak-calling algorithms (*186*). More specifically, it has recently been described that active promoters produce artifactual ChIP-seq peaks in an unspecific manner, called ‘phantom peaks’ (*187*). These peaks are present even in conditions when the protein bound by the specific antibody is not present in the samples. Again, a careful statistical analysis is required to identify and address these possible biases and artifacts.

1.2.4. Proteomics and mass-spectrometry.

Proteomics refers to the study (i.e. identification, quantification and characterization) of all the proteins present in a biological sample. The term was chosen to parallel with the study of the genome (genomics) and that of the transcriptome (transcriptomics), being developed at the same time. Genome-scale proteomics has benefited enormously from mass-spectrometry (MS), an analytical technique. The basic principle underlying MS is the unequivocal identification and quantification of ionized molecules by their mass-to-charge ratio.

In a typical proteomics MS experiment, proteins from a biological sample are trypsin-digested in order to obtain smaller peptides. Trypsin is a protease that cleaves proteins after lysine or arginine residues. Alternatively, other proteases can be used, such as chymotrypsin, which cleaves peptides after tyrosine, phenylalanine or tryptophan residues. The peptides obtained are fractionated using high-performance liquid chromatography (HPLC). Fractionated samples are then ionized. Once ionized, peptides pass on to the mass analyzer. In the mass analyzer, the mass-to-charge ratio of each of the molecules entering is determined. From this, the exact masses of the different peptides are obtained and compared to a database using computational algorithms (*188*). Databases are generated by an *in silico* translation and trypsin/chymotrypsin-digestion of all the open reading frames (ORFs) in the genome, derived from the NCBI annotations. The masses of the *in silico* peptides are calculated and compared to the output of the mass-analyzer to identify those matching, corresponding to the proteins present in the sample. This protocol is known as protein fingerprinting or peptide mass fingerprinting (*189–191*). As the databases are derived from NCBI annotations, if some ORFs are not annotated in these databases, the corresponding peptides in the sample will not be assigned. If a peptide in the sample is not present in a database, it can be *de novo* sequenced to determine its amino acid composition. In such case, the selected peptide is fragmented in smaller units. These units are analyzed again by mass spectrometry (MS/MS) and the collection of subsequent masses can be compared to a database of predicted masses of peptides (*192*). If not present in the database, there are algorithms that match the mass differences among the individual fragments to the mass of single amino acids, to determine the exact sequence of the initial peptide (*193*).

For protein quantification, isotopic labelling is often used. A sample is labelled with heavy isotopes (usually ^{13}C and ^{15}N), whilst the other sample is labelled with the corresponding light isotopes (^{12}C and ^{14}N). They are mixed and analyzed together, as the isotopic mass difference allows to distinguish between the samples and permits a relative quantification (*194*). Label-free methods also exist, that avoid the treatment with heavy isotopes. In these methods, the relative signal of a peptide is compared across samples for relative quantification. The area of this peptide in the mass spectrum is calculated and used for the comparison. Currently, the relative areas of at least the three top peptides (the most abundant) identified in each protein are used for comparison (*195*). Using the three top peptides of each protein avoids biases in the quantification.

Although not related to the transcription process in a straightforward manner, identification and quantification of protein levels is also relevant to the understanding of transcriptional regulation. Protein levels of different transcription factors in changing conditions can reveal how transcriptional responses to different perturbations are orchestrated. In *E. coli*, this approach has been used to identify proteins changing their expression profile in osmotic stress, both in aerobic and anaerobic conditions (*196*). Also, proteomics can provide an insight into the effects of post-transcriptional regulation, for example by sRNAs. In *E. coli*, mass-spectrometry was used to confirm the effect of an sRNA on the OmpA gene, decreasing the stability of mRNA and thus limiting translation to protein (*197*). In *M. pneumoniae*, a strain lacking the protein phosphatase PrpC was shown to have decreased levels of proteins from the cell-cycle operon, quantified by MS and confirmed by a decrease in the mRNA levels, and a strain lacking the protein kinase PknB had decreased levels of adhesion proteins (*198*). Furthermore, changes in protein levels can be used to confirm changes in the transcriptome upon certain perturbations (*199*). Finally, proteomics is not only used to quantify native proteins, but also to determine and quantify the prevalence of post-translational modifications (*198*, *200*, *201*) or to identify protein complexes that co-purify together (*202*, *203*).

Despite its numerous applications and a longer development than sequencing-based technologies, mass-spectrometry has some weaknesses that should be addressed here. The most notable one relates to a lack of robustness. The spectra obtained are dependent on the ionization properties of the peptides, which behave differently in the mass analyzer (*204*). Also, lack of reproducibility is an important issue, especially in low-expressed proteins in complex samples, and numerous replicates are required to address under-sampling (*205*). Finally, for a unequivocal identification of proteins, only unique peptides should be used (*206*). This hampers the identification of proteins with paralogs present in the genome, as their sequences tend to be highly similar and the number of unique peptides decreases. If no unique peptides are found, the identification and quantification can only be performed at the protein family level (*207*). Also, small proteins have less probability to be found in MS experiments, as at least three unique peptides should be considered for identification and quantification. These proteins, because of their size, have less tryptic peptides, which results in less chances to be identified.

1.3. Systems biology and the integrative study of transcription

In the last section, the applications of different ‘omics’ technologies in the study of transcription have been reviewed. In many of the aforementioned articles, different omics were combined to gain knowledge on a specific element of transcriptional regulation. For instance, transcriptomics has been combined with ChIP-seq (*182*, *183*), or with proteomics (*208*), mostly to reveal the function and targets of transcription factors. However, although important, transcription factors only represent one of the determinants of the RNA levels in the cell. As described above, there are other processes and factors intervening, such as the transcription unit structure, the promoter strength and its accessibility, determined by epigenetic modifications and supercoiling, the action of riboswitches or the effects of sRNAs, and the effect of RNA degradation. Little work has been done towards the integration of some or all of these elements. Indeed, there are only few examples concerning small systems (*209*). In order to gain a better knowledge of how all these elements interact or interfere with each other to give rise to determined levels of transcripts in a global manner, integrative approaches need to be considered.

In this section, the discipline of systems biology, and how it is the responsible of a change of paradigm in biological studies, is described. The applications of systems biology to the integrative study of transcription are reviewed, as well as the challenges and perspectives that this discipline poses.

1.3.1. Systems biology: the whole is greater than the sum of the parts.

“Systems biology” refers to the study of different biological systems, whose components may be molecules, cells, organisms or populations. This discipline not only describes each of the individual components of the system, but also the interactions among them. Biological systems are inherently complex, and their behavior cannot be explained or predicted only from the properties of the single components: “the whole is greater than the sum of parts”. This intrinsic complexity causes that, in this field, a number of disciplines converge, such as biology, chemistry, computer science, math, engineering and physics. The goal of this interdisciplinarity is to develop accurate models that allow to understand, describe and predict the behavior of the given biological system. Ultimately, the goal of systems biology is to understand how life is organized.

Originally, two broad directions emerged under the name of systems biology: 'mechanistic' and 'statistical' systems biology (*210*, *211*). Mechanistic systems biology focuses in understanding the dynamics of a certain system, and how the state of the system changes in time (*211*). This involves some sort of mathematical modeling to represent the evolution of the system over time. Different modeling formalisms can be chosen, depending on how much information is available on the system of study. One of the preferred choices when modeling transcriptional networks is the use of systems of differential equations. For instance, ordinary differential equations (ODEs) have been used to model feedback loops that generate oscillatory behaviors (*212*, *213*). Partial differential equations were used to account for diffusion in the cell to model the gap gene network of *Drosophila melanogaster*, which controls key steps of its embrionary development (*214*). In many cases, the stochastic nature of living systems requires different modeling formalisms to be able to capture their behavior. In transcription, it has to be considered that the number of regulators, RNAs, promoters, or any other molecules involved, might be low. In such cases, deterministic methods such as ODEs are unable to reproduce the behavior of the system and stochastic methods should be used (*215*). Spatial stochastic modeling can be chosen to address the movement of particles from different compartments within the cells (*216*). The challenges of this kind of models involve mainly unraveling which is the network and the parameters underlying the specific behavior, in other words, reverse engineering the network. Reverse engineering is feasible for small networks involving few components, but as the number of components grows up to hundreds or thousands, the large number of parameters to determine yields these models unscalable (*210*).

Statistical systems biology has arisen as a consequence of the development of omics technologies. The vast data generation from the past few years has led to the description of practically all the molecules present in a cell: RNAs, proteins, metabolites, and their concentrations. It has also allowed to map the interactions occurring within proteins or DNA (*217*), as well as between proteins and DNA (*183*) or proteins and RNA (*218*) among others. The application of thorough statistical analyses to these datasets has allowed to discern between real interactions and spurious or artifactual ones, and most importantly, it has allowed to describe important organizational principles of biological systems. For instance, the study of protein-protein interactions led to the observation that these follow the organization of scale-free networks, in which there is a small number of *hubs* and a large number of proteins with very few connections (*219*). Later, it has been proved that the connectivity of the proteins in the network is important to define their role in physiology and disease (*220*, *221*). Nevertheless, this description of the components of the cell and their interactions has little predictive power of the state of the system. Often, the context of the interactions is ignored, and it is crucial to consider it, as interacting proteins should be simultaneously present in the same tissue (*222*). Also, the networks are usually biased towards more studied proteins, leading sometimes that these appear as artificial hubs in the networks (*223*). Some perturbations in the network can be done, mimicking gene mutations and deletions, to test how the global structure of the network is affected, but this does not describe how the system can evolve in time. In words of Hiroaki Kitano, the description of the components and their association in networks is equivalent to listing all the parts of an airplane and drawing a diagram showing how they are assembled: it is insufficient to understand how the airplane works (*224*). What is missing is the information flow, this is, the dynamics of the system.

1.3.2. Bringing together the two views of systems biology. Genome-scale modeling.

The two different views of systems biology described above seemed initially irreconcilable. Nevertheless, over time there have been many efforts to reconcile these views. Currently, there is a continuum of modeling formalisms that goes from large qualitative models, with little or no kinetic parameters, to smaller fully-detailed kinetic models of differential equations (*225*) (Figure 1.8).

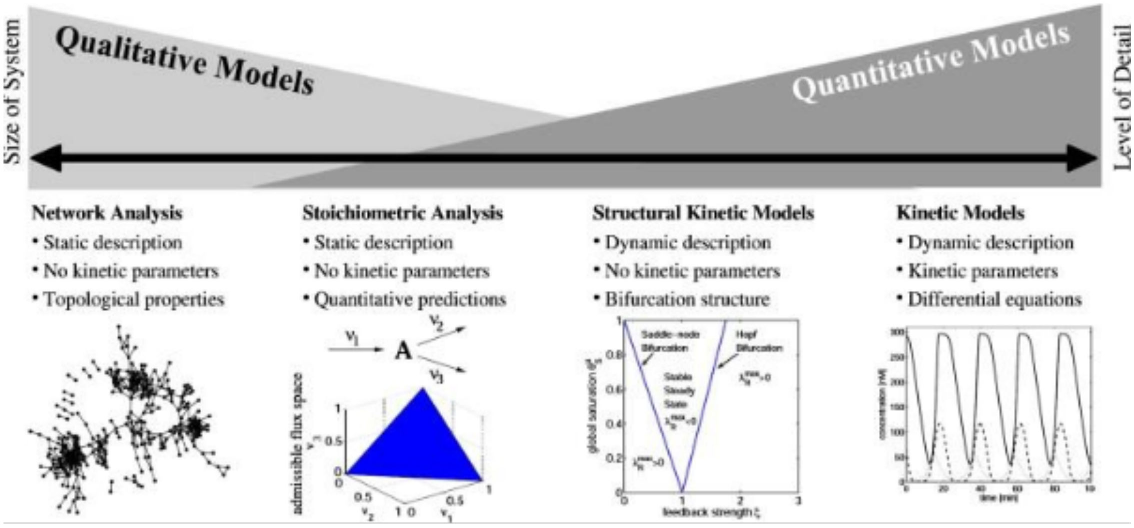


Figure 1.8. Different modeling formalisms used in metabolic modeling, showing the size of the systems studied and the level of detail of each formalism. Image from Steuer, R. (2007)

Apart from this continuum of modeling formalisms, the rapid development of omics technologies and the increasing amount of data available has allowed for the construction of larger kinetic models. Many of these efforts have been focused on modeling metabolic networks. For example, the entire central carbon metabolism of *E. coli* has been modeled using ODEs (*226*). However, kinetic models are still difficult to scale up, and for larger models such as entire metabolic reconstructions, constraint-based formalisms are the preferred choice (*227*). These models convert the metabolic network into a stoichiometric matrix that includes all reactions with their corresponding stoichiometric indices. This matrix has an associated solution space, that includes all the possible solutions of the system given a set of constraints. Known constraints (such as maximum fluxes, a steady state assumption, or network topologies) are included to limit this solution space and find values for the fluxes of the different reactions in the network (*228*). These models have been used to reconstruct the metabolic networks of different model organisms such as *E. coli (229, 230)* and *B. subtilis (231)*, as well as other bacteria (*232*, *233*). Although mainly used in the context of metabolism, constraint-based models have also been used to model transcriptional regulation (*234*).

These models pose the challenge of integrating information from different data sources, such as transcriptomics, proteomics, metabolomics or phenomics (*235*). This integration is not straightforward and it has been estimated that currently, a large proportion of resources is dedicated to data processing and integration (*236*). Also, these are not fully dynamic models, as they include the constraint of the steady state assumption. Some modifications exist that allow for the inclusion of dynamic behavior (*237*).

1.3.3. Going even further: multi-scale modeling and whole-cell modeling.

After the development of the aforementioned genome-scale models, that focus on a single biological process, the interest of the field is moving towards the integration of different cellular processes in single, multi-scale models. Different cellular processes occur at very different spatial and temporal scales (*238*), and their integration under a single modeling formalism is truly challenging.

Only focusing in the process of transcription and its regulation, the different time scales are evident. As an example, protein complexes such as those formed by the RNA polymerase holoenzyme, or by transcription factors, can form at the sub-second scale, while transcription elongation or translation into proteins may take minutes (*239*). In bacteria, spatial restrictions are not so evident, as transcription and translation occur simultaneously due to the lack of the physical barrier imposed by the nucleus. However, there are studies showing that transcription and translation occur in separate domains of the cell in some bacteria (*240*). Also, some authors claim that genes with related function are closely localized in the 3D conformation of the chromosome, to be co-transcribed in the so-called transcription factories (*241*). The different temporal and spatial scales are multiplied when integrating different biological processes, which hampers the development of suitable multiscale modeling formalisms.

Despite these challenges, some approaches have led to the obtention of models integrating different biological functions, mainly using constraint-based approaches that assume steady state. In *E. coli*, an integrated model of metabolism and transcriptional regulation was obtained (*242*). In this model, regulatory genes control the expression of metabolic enzymes, thus altering the fluxes of the metabolic model. In a different study also on *E. coli*, a model of transcription and translation was obtained (*243*). This model accounts for RNA synthesis, regulation by transcription factors or degradation, but misses other key determinants such as regulation by metabolites, sRNAs or the effect of promoters.

Most of these studies have used *E. coli* as a model system, due to the large availability of the data in this organism. However, the elevate complexity of this bacterium, with more than 4000 protein-coding genes, hampers the development of more detailed, integrative models. Indeed, the first 'whole-cell' computational model, that reproduces all processes occurring inside the cell, simulates the cell cycle of *Mycoplasma genitalium (244)*, a human pathogen whose genome encodes for only 525 genes (*127*). The consecution of this model has only been possible after the advancement of omics technologies, that has led to the obtention of vast amounts of quantitative data in non-model bacterial species; and the development of a mathematical framework that allows for the integration of processes occurring in a variety of time scales and in different compartments inside the cell. This mathematical framework consists in choosing an appropriate time step for the entire simulation; the different biological processes in the cell are then modeled and parameterized separately, in different 'modules', and simulated over this particular time step. Each of the modules uses the most appropriate modeling formalism for the specific process. After all the processes have been simulated, the output of each module is used to update the *state* of the cell. The input for the different modules for the next time step-long simulation will be extracted from the recently updated state, repeating the same process until the cell has divided (Figure 1.9). In the *M. genitalium* model, that includes 28 modules simulating different cellular processes, the chosen time step is 1 second (*244*). Additional considerations have to be taken to provide an equitative allocation of the cellular resources among the 28 processes.

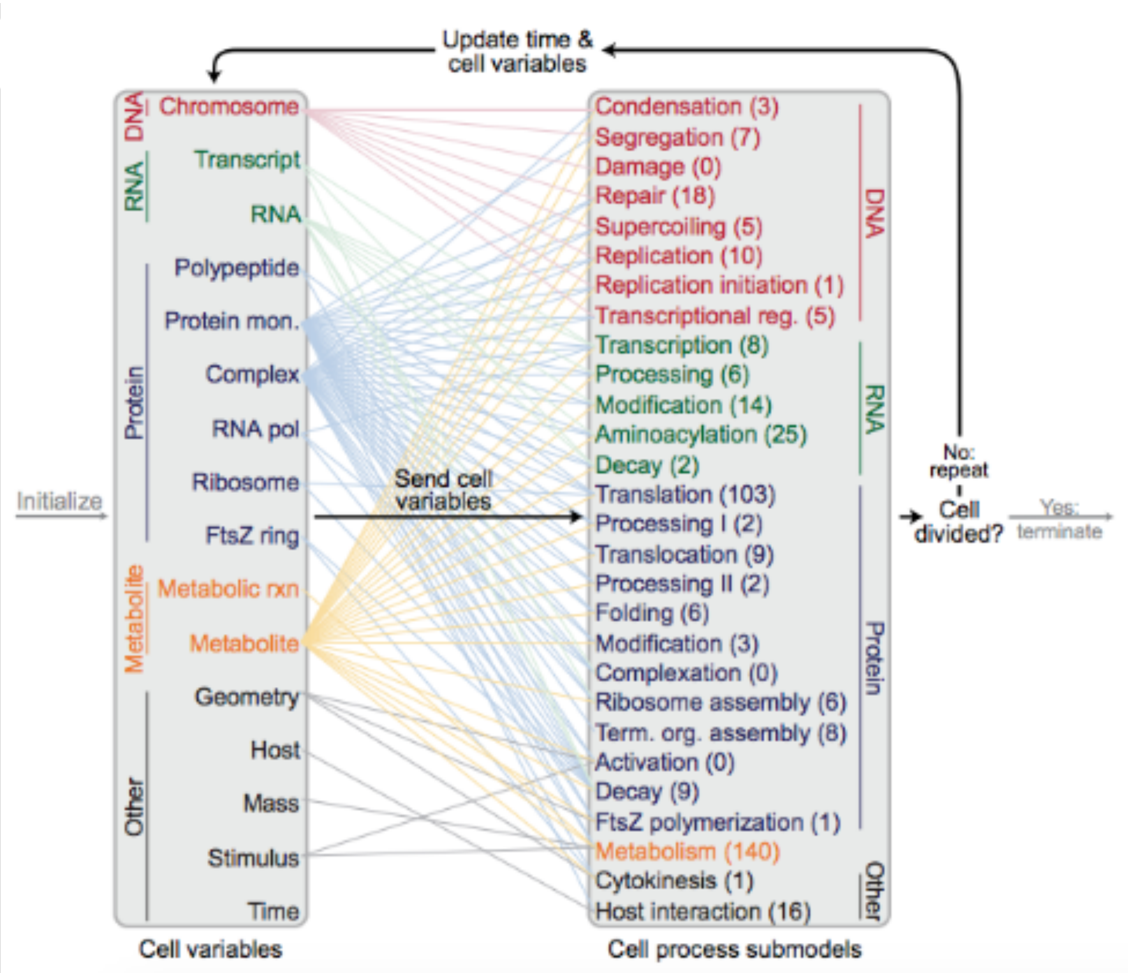


Figure 1.9. Structure of the whole-cell model of *M. genitalium*. In the left column, the state of the cell, representing the variables grouped in 16 categories, is listed. In the right column, the 28 cellular processes are listed. The different colors represent which categories of the state variables are updated after the simulation. Links between variables and submodels represent the input variables for each submodel. Image from Karr, J. et al (2002).

Two main criticisms to this model should be noted. Firstly, the data used to fit the model of *M. genitalium* was mostly coming from other bacterial species, sometimes closely related such as *M. pneumoniae*, sometimes very distant such as *E. coli (244)*. This may cause that inconsistencies between the predicted and the experimental data appear. Secondly, the selection of the time step is rather arbitrary and it has not been tested whether smaller time steps result in a different outcome. This would imply that some processes interact with each other at time scales shorter than the 1s time step, and thus some modules should be integrated together, or the chosen time step should be shorter. Apart from these general caveats, regarding the transcription process, the model of *M. genitalium* ignores the reported existence of sRNAs (*245*), and also it does not consider the possibility of regulation via riboswitches, supercoiling, or other factors described above. Furthermore, it does not provide mechanistic insights regarding promoter strength, which limits its predictive capacity on mutated or synthetic strains carrying non wild-type promoters. Also, it only accounts for 5 transcription factors and their binding sites are assigned based on homology with other species and not on experimental evidence in this bacterium. DNA-binding proteins other than transcription factors are not considered for alternative regulatory mechanisms.

Despite these caveats, the model represents a breakthrough in computational and systems biology. The development of omics and the generation of a vast amount of genome-scale datasets open the door to the expansion of the model to more complex species.

1.4. *M. pneumoniae* as a model organism in Systems Biology

As mentioned above, *E. coli* is a very complex bacterium encoding for more than 4000 genes in its genome (*246*). Therefore, although the majority of studies on bacterial transcription (and in many other processes) have utilized *E. coli* as a model organism, using this bacterium as a model for integrative models renders complicated. Simpler organisms with a smaller set of genes should be used, at least in this initial phase of integrative Systems Biology, where large scale dynamic models are still developing.

M. pneumoniae is a Gram positive bacterium, member of the Mollicutes class (*247*). This bacterial class is characterized by the lack of a cell wall, low GC content genomes and a small size. Mollicutes are parasites of plants (phytoplasmas) and animals (mycoplasmas), sometimes causing diseases to their hosts. Their parasitic lifestyle has led to their genome reduction (*248*), eliminating biosynthetic pathways and using nutrients taken from their hosts. Indeed, they are the smallest self-replicating organisms known to date. This feature makes them valuable model organisms for Systems Biology studies, an example of this is the use of *M. genitalium*, to develop the first computational model of an entire bacterial cell (*244*). Mollicutes, and in particular mycoplasmas, are interesting also from a Synthetic Biology point of view, as bacteria with reduced genomes are easier to manipulate and engineer. In fact, mycoplasmas tend to have linear metabolic pathways, with few crosstalks among them (*249*). This facilitates engineering, avoiding unwanted interferences with different natural pathways of the bacterium. Indeed, several laboratories are trying to eliminate non-essential genes in various mycoplasmas, such as *Mycoplasma mycoides*, *M. pneumoniae* or *M. genitalium,* in order to obtain a minimal chassis from to which add new functionalities (*250*).

M. pneumoniae is a human parasite that colonizes the respiratory tract, causing atypical pneumonia in immunocompromised patients (*251*). Therefore, it is interesting not only as a Systems Biology model organism but also to understand its role in disease. It is the closest relative of *M. genitalium,* and its genome encodes for 737 proteins and 311 sRNAs (*252*). It can be cultured in laboratory conditions, and a defined medium for growth of this bacterium has been proposed (*249*). In the last few years, a consortium of european laboratories has joined efforts to characterize this bacterium at the molecular level, taking advantage of the omics technologies being developed. They characterized the transcriptome (*72*), the proteome (*203*) and the metabolome (*249*) of *M. pneumoniae*. The characterization of the metabolome led to the posterior development of a flux balance analysis (FBA) model of its metabolism (*233*). The post-translational modifications of the proteins in this bacterium have been described (*198*), showing that there are functional relationships among them. Also, essentiality of all the components of the *M. pneumoniae* genome has been described (*73*). Finally, another study highlights the importance of assessing how transcriptional and post-transcriptional events are orchestrated in this bacterium. Understanding them is central to the study of the entire process of gene expression (this is, from DNA to RNA to protein), as it has been shown that majorly post-transcriptional, rather than post-translational mechanisms control the protein/mRNA ratios in this bacterium (*141*).

The aforementioned studies have generated a vast amount of information of *M. pneumoniae* at different levels, unique in a non-model system. All this information, together with the apparent simplicity of this bacterium, due to its reduced genome size, render *M. pneumoniae* as an ideal model organism for the construction of genome-scale models of different cellular processes, and also for the adaptation of the original whole-cell model of the close relative *M. genitalium*.

1.4.1. Transcription in *M. pneumoniae*.

Transcription in *M. pneumoniae* has been studied mainly using a combination of microarrays, tiling arrays and RNA-seq (*20*), in addition to previous studies that made use of classical molecular biology approaches. The latest genome annotation available indicates that its genome is organized in 1305 operons (*252*), although recent research establishes that transcription units in *M. pneumoniae* are highly plastic and can vary in a condition-dependent manner (*23*).

Promoters in *M. pneumoniae* have been accurately mapped by identifying the TSS throughout its chromosome (*157*). In this study, short RNAs, termed 'transcription start site-associated RNAs' or tssRNAs, were discovered. These short RNAs, of around 45 nucleotides, correspond to abortive transcripts, but whether they play a regulatory role in transcription remains unknown. The main features of promoters in this bacterium are the presence of a canonical Pribnow box, corresponding to the housekeeping σ^{70} factor, with the sequence 5'-TANAAT-3', where N stands for any nucleotide; and a degenerate -35 box (*72*, *253*, *254*). DNA methylation in *M. pneumoniae* has been associated to promoters, especially those of genes related to defense mechanisms and virulence (*41*), but a more specific role of methylation in regulating transcript expression could not be associated. Other promoter features have not been characterized in this bacterium.

Regarding transcription factors in *M. pneumoniae*, eight putative TFs were identified by sequence analysis and/or copurification with the RNA polymerase complex (*203*, *249*). These include the housekeeping σ^{70} factor and other two putative sigma factors: SigD and YlxM. The former has been recently validated as a true sigma factor in the close relative *M. genitalium (255)*. Despite this low number of regulators, *M. pneumoniae* displays complex and specific responses to a variety of perturbations, which suggests that there may be agents other than TFs playing a crucial role in regulating transcription in this bacterium. Metabolites and sRNAs may be responsible for this regulation, but there are no studies performed in *M. pneumoniae* to assess which is their regulatory role. Only the essentiality of sRNAs has been established in this bacterium, pointing that the majority of them are non-essential (*73*).

Concerning the transcription termination, the most notable feature of *M. pneumoniae* is the lack of a Rho termination factor, which causes that termination can only be Rho-independent. Therefore, termination is mediated by the presence of GC-rich hairpins followed by a poly-uridine tract. Although there is only intrinsic termination in this bacterium, three proteins from the Nus family of termination/antitermination factors are present in the genome of *M. pneumoniae*: NusA, NusB and NusG (*128*). The binding of these factors to the elongating RNA polymerase complex could alter its processivity mediating termination. A recent study in *M. gallisepticum*, closely related to *M.*

pneumoniae, describes the existence of two different kinds of transcription terminators: strong, hairpin-containing terminator sequences and weak terminators regulated by heat-shock stress (*256*).

Finally, there are no studies regarding RNA degradation in *M. pneumoniae*, but several components of the RNA degradosome have been identified, such as the J1 nuclease (Mpn280) as well as the RNase Y (Mpn269). PnpA and the helicase CshA have not been identified, although several RNA helicases are present in this bacterium that could replace the helicase from *B. subtilis*. Enolase (Mpn606) and phosphofructokinase (Mpn302) are also encoded in the *M. pneumoniae* genome. Other RNases identified are the RNase III (Mpn545), RNase R (Mpn243) and RNase P (Mpn681), involved in ribosomal and/or transfer RNA processing. Further ribonucleases in *M. pneumoniae* are involved in DNA replication or recombination. Although no studies focus directly on transcript degradation in *M. pneumoniae*, some have estimated an RNA half-life of 3 minutes (*141*).

Despite the already existing information on the process of transcription in *M. pneumoniae*, research is still needed to fill the knowledge gaps, in order to be able to compile a genome-scale model of transcription that can be part of a whole-cell model of this bacterium.

2. Objectives

The objectives of this PhD thesis are: 1) To adapt the whole-cell model of *M. genitalium* to obtain a first version of the whole-cell model of *M. pneumoniae*. 2) To critically assess the high-throughput profiling or ‘omics’ technologies, used to generate the data to fit the model, and the computational pipelines that accompany them. 3) To improve our knowledge of the process of transcription and what the determinants of transcriptional regulation are, taking advantage of all the ‘omics’ datasets regarding transcription that have been and are being generated in our laboratory. This objective corresponds to the central part of the thesis.

In Chapter 3, we detail the process of the whole-cell model adaptation. This involves the compilation of all the available data on *M. pneumoniae* in a knowledge base, from which the computational model will extract the model parameters, the adaptation of the code to the new data, and to be implemented in the computing facilities of the Centre for Genomic Regulation, and the process of parameter fitting to reproduce *in silico* the behavior of this bacterium. The future work towards the improvement of this model, and prospective applications are also described.

To build this model, we rely mostly on ‘omics’ datasets, generated usually in microarrays, deep sequencing or mass spectrometry experiments. In Chapter 4, we describe the case of an artifact occurring in RNA-sequencing experiments. This artifact led to the observation of chimeric RNAs in bacterial transcriptomes, and was dependent on the library preparation protocol. Furthermore, we describe how widely-used chimera detection algorithms fail to detect these artifacts. This raises a discussion on the pace at which these new technologies applied to molecular biology appear and evolve, and how we should be aware of their artifacts and biases.

Chapters 5, 6 and 7 are focused in the main objective of this thesis, the study of the key determinants of RNA abundance in *M. pneumoniae*. In Chapter 5, we characterize qualitatively the promoters of this bacterium, obtaining a classifier to distinguish among true promoters, non-productive promoters and non-promoter sequences. Future work in this field points to quantitative prediction of RNA levels based on the promoter sequences. In Chapter 6, we assess the function of small RNAs in bacteria in general and in *M. pneumoniae* in particular, focusing on antisense RNAs, to find that their low copy number in cells renders any functional effect (not due to enzymatic activity) as highly improbable. Finally, in Chapter 7, we describe the gene regulatory network of *M. pneumoniae*, describing its transcription factors and regulators, as well as their targets. Furthermore, we explore other mechanisms of transcriptional regulation, such as riboswitches, metabolites, supercoiling and RNA degradation, to conclude that these can be at least as important as transcription factors to regulate RNA levels in such a minimal bacterium.

3. Towards a whole-cell model of *Mycoplasma pneumoniae*

"*With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.*"

John von Neumann

3.1. Abstract

To understand how living organisms function, we should be capable of describing all their individual components, but also the interactions occurring among them, and their dynamics. The use of models has proven central to this understanding of all sorts of biological processes. In particular, multi-scale models address the problem of dealing with multiple spatial and temporal scales, naturally occurring in biology. However, the generation of such models poses a number of challenges, due to their high dimensionality, their high computational costs, and the lack of knowledge about some parts of the physiology of the modeled organisms. Here, we describe the construction of the first version of a computational whole-cell (WC) model of *M. pneumoniae*. This model is based on a previous WC model of *M. genitalium (*244*)*. We also describe our plans for increasing the accuracy of the model and the scientific questions that we intend to explore with the model. We anticipate that WC models will be valuable tools for guiding genome engineering.

3.2. Introduction

The ultimate goal of biology is to understand how living organisms function. This knowledge can then lead to the rational modification of these organisms for different purposes. To understand a living organism, it is necessary to know all of its components, how they interact, and the dynamics of their interactions with each other and with the environment (257).

Mathematical models are needed to make sense of the large lists of components, interactions and dynamic rules. These allow a representation of the events occurring at different levels. Also, mathematical models allow to reproduce the dynamic behavior of the studied systems by simulating their evolution in time. Additionally, these simulations, can be used to predict the result of different perturbations in the system, thus saving experimental time and resources to.

However, several considerations need to be taken into account when utilizing these mathematical models. The first one arises from the intrinsic modeling procedure: models are abstractions of the reality, and therefore simplifications. When constructing a mathematical model, it is the responsibility of the modeler to decide which assumptions and simplifications are acceptable in the context of the studied system. Secondly, deriving the mathematical models is not trivial. Our knowledge of biological systems, although profound, is usually insufficient to derive the mathematical equations that represent the dynamics of the reactions of the system (258). The modeler has to decide how to represent the system and what parameters to include in the model. This is not trivial and needs careful experimentation, as different models can yield the same output when tuned with the appropriate parameters (259). Conversely, it is possible that a given model can produce different outcomes depending on the set of parameters used. Besides considering these features of models, we have to consider that a common property of biological systems is robustness (260), and this property should be captured in mathematical models. For many biochemical reactions, changes of the parameters within reasonable limits should not produce significant effects. However, for some others, small changes in a parameter lead to dramatic physiological changes (261). A final consideration is that in many occasions, biological systems behave in a stochastic rather than in a deterministic manner. This stochasticity is intrinsic to the low copy numbers of some of their components, causing fluctuations that biological systems have evolved to overcome *(*262*)* or to take advantage of *(*263*)*. This randomness needs to be correctly handled in the mathematical models.

Despite these considerations, mathematical models have long been used in biology to study systems at different scales. From ecosystems, such as the Lotka-Volterra model for the evolution of predators and preys *(*264*,* 265*)*, to the biomolecular level, with models of protein folding *(*266*)*. There are also models of entire organs such as the heart *(*267*)*, groups of cells such as neurons *(*268*)*, and biomolecular pathways *(*269*)*. However, to the aforementioned considerations, we need to add the difficulties of devising models that are able to cope with different scales of complexity. One of these challenges consists of the reconstruction of models using heterogeneous data. The data used to build these models usually come from a number of different laboratories, and in a variety of sources (omics datasets, microscopy, low throughput experiments, computational predictions, etc.). This causes that sometimes, the identifiers used for the same gene, protein or metabolite may be quite different. Also, the data are usually incomplete and/or biased towards some parts of the physiology that are well understood. Additionally, currently there is no single mathematical framework that allows to model the multiple temporal and spatial scales of biological processes. Different formalisms adapt better to certain cellular processes, but no one fits all. Thus, multi-algorithm modeling is the only way to approach this problem.

These challenges hamper the development of comprehensive models of entire organisms, that span all the scales below this level. Indeed, the aforementioned models tend to focus in one or few scales and make different simplifications regarding the lower ones. To address this problem, the scientific community has set the focus on modelling entire bacterial cells. Bacteria offer various advantages for this purpose, compared to eukaryotic organisms. As unicellular systems, the number of scales to integrate for an entire organism is smaller than in multicellular eukaryotes. Also, the lack of internal compartmentalization of the cells with organelles reduces the multi-scale complexity, but we should not forget that in bacteria, different functions can occur in separate subcellular locations *(*240*)*. Within bacteria, especial attention has been given to genome-reduced species such as Mycoplasmas. With a close-to-minimal set of genes to sustain life, their simplicity compared to other organisms has attracted efforts to construct these multi-scale, integrative models. Indeed, the first attempts of performing WC simulations were performed using *M. genitalium* as a model system *(*270*)*. These first attempts involved the creation of a software platform for WC simulations, E-CELL, and the development of a model of a hypothetical cell containing

127 genes of *M. genitalium*, necessary for the basic functions of transcription, translation and metabolism *(*270*)*. This *in silico* cell was, however, unable to replicate its DNA or divide, due to the exclusion of the genes responsible for this functions in the reduced set of 127 genes. The E-CELL software helps researchers build and simulate systems models composed of multiple species which interact via reactions. E-CELL supports both deterministic and stochastic simulations *(*271*)*.

This first platform, however, did not account for diffusion or localization processes, which can be central to the outcome of the simulations. For instance, *M. genitalium* and its close relative *Mycoplasma pneumoniae* have a terminal organelle involved in cell adherence, virulence, motility and cell division *(*272*,* 273*)*. This organelle is located in one pole of the cell, rendering important to delimit subcellular locations even in bacterial models. To address these problems, there are now software platforms that account for subcellular localization, such as SmartCell *(*274*)*. This kind of software allows to consider different cell shapes and geometries, and approximates particle diffusion within the cell and transitions between the different subcellular compartments. This platform uses stochastic methods to simulate the models, and recreates the cell geometry by dividing the entire cell into multiple sub-volumes. Within these smaller volumes, particles are considered to be perfectly mixed, and there can be diffusion by translocating particles from some volumes to the adjacent ones.

The aforementioned platforms and the early models of *M. genitalium* have certain limitations. They are restricted to a few cellular functions, such as transcription, translation, and metabolism. Other relevant functions to the cell physiology are also not included, such as cell division, chromosome replication, regulation or response to external perturbations. To formulate a model of the entire cell physiology, a completely new approach is needed that is able to deal with different temporal scales as well as with the various subcellular locations. To deal with these circumstances, the idea of adopting a modular approach was proposed a decade ago *(*275*)*. This approach consisted in dividing the whole set of reactions occurring inside the cell in smaller subsets, and model these subsets independently, but linking them in a way that the dynamic behavior of the entire system is preserved. As an example, the authors link three different models covering different parts of the yeast metabolic network *(*276–278*)*.

Although in this work the idea of using a modular approach to large-scale modeling was restricted to metabolism, this concept was later used to generate the first comprehensive computational model of an entire organism, *M. genitalium* (244).This has been the first model aiming at modeling the complete physiology of a bacterium, considering all cellular processes and the function of each of its genes. In order to build such a model and deal with the complexity of the bacterium, the authors classified all cellular processes into 28 modules. Each of these modules is modeled independently, using the mathematical formalism that best suits our knowledge of the specific process and the data available. For instance, metabolism is modeled using flux balance analysis (FBA), while transcription initiation is modeled as a stochastic process. These independent modules are simulated separately for a timestep of one second. After this short timestep, the different modules interact among them by updating the state of the cell. They write the output of their simulations to different variables that account for the amount of DNA, RNAs, proteins, metabolites, etc. Prior to beginning the next round, the different modules read from these variables the input data for the next simulation.

The model was implemented in MATLAB, and it was accompanied by a knowledge base which compiled all the data used to train the model and fit each of the 1836 parameters that it uses *(*279*)*. This knowledge base is available online. The majority of the data used in the model was extracted from diverse databases or from previous literature. In some cases, this data referred to *M. genitalium*, but in many others, it was measured in other organisms. This point has been the focus of different criticisms *(*280*)*, and authors suggest that in the future, better characterized organisms should be used as models for these WC simulations. Another point which has been a focus of debate is the selection of a timestep of one second in order to integrate the different modules of the entire model. The assumption of module independence over such a timestep may not hold true for every single sub-model. Alternatively, using smaller timesteps poses a different problem: the computational cost in time and resources would increase dramatically. Given that the current simulations take around 20 hours to be completed, the selection of such a time interval represents a compromise between accuracy and efficiency of the model. However, further studies are needed to verify that the predictions of the model are not affected by the choice of a timestep, ensuring thus the independency of the different modules during these short periods.

Here, we present the work towards the generation of a WC model of the minimal bacterium *M. pneumoniae*. During the past decade, a consortium led by our lab has worked towards the detailed molecular characterization of this organism. To date, the transcriptome *(*72*)*, proteome *(*203*)* and metabolome *(*249*)* of this bacterium have been described. Also, the essentiality of all the genomic features of this bacterium has been recently characterized *(*73*)*. These initial studies have been later complemented with other works focusing on more specific aspects of the biology this bacterium. At the level of transcription, transcription start site associated RNAs (157) have been identified, and we have characterized promoters in *M. pneumoniae* ((281), Chapter 5). The regulation of transcription has been studied in depth (see Chapter 7), and we have recently described regulation at the level of RNA polymerase trafficking (23). At the level of translation, the major post-translational modifications have been determined (198), and a model integrating transcription and translation of all genes in this bacterium has been constructed (141). Regarding metabolism, a flux-balance analysis model of the entire metabolism of this bacterium has been generated (233) and several metabolites quantified (282). Furthermore, the DNA methylome of *M. pneumoniae* was also characterized (41). These studies, together with some others currently in preparation, account for the largest accumulation of information and data for a bacterium generated by a single consortium.

Therefore, *M. pneumoniae* constitutes the ideal organism to build a WC model of. Most of the data necessary to construct it s already available, and the fact that these data have been generated by a single consortium of laboratories reduces their variability. In the following, we detail the process of generating a knowledge base for *M. pneumoniae* that integrates all the information related to this bacterium, and the work towards constructing and improving the WC model of *M. pneumoniae,* using the original *M. genitalium* model as a platform.

3.3. Data curation and assembly in a knowledge-base

The first step towards the generation of a WC model of any organism consists of the compilation of all the available data from such organism in a dedicated knowledge base. The model will read the information from this database, which will include the information regarding model species, reactions and parameter values. For this purpose, WholeCellKB was created as a platform to create these knowledge bases for different model organisms (279). For the case of *M. pneumoniae*, this task was facilitated by the fact that most of the information was obtained via high-throughput experiments, covering the entire genome. Therefore, a few publications account for a large percentage of the data needed. Also, as these studies have been performed by a single consortium of laboratories, the same naming conventions for genes, proteins, metabolites, etc., were used throughout all datasets, facilitating the compilation and integration of the information.

Table 3-1. New experimental data from *M. pneumoniae* compiled in the WholeCellKB-MPN.

Data	Type	Source
DNA methylation sites	Experimental	*(*41*)*
DNA structural regions	Experimental	*(73)*
DNA DnaA boxes	Experimental	*(*41*)*
Gene annotation	Re-annotation	*(252)* Wodke et al, in preparation
Metabolites and reactions	Experimental	*(233, 249, 282)* and collaboration with U. Sauer's group at ETH, in preparation
Protein abundances	Experimental	*(*141*,* 283*)*
Protein complexes	Experimental	*(203)*
Protein half-lives	Experimental	*(*141*)*
Post-translational modifications	Experimental	*(*198*,* 203*)*
Secreted proteins	Experimental	Paetzold et al, in preparation
Transcription start sites	Experimental	*(*157*)*
Transcription termination sites	Manual curation from experimental data	*(252)*
RNA abundances	Experimental	*(72)*
RNA half-lives	Experimental	*(23)*; Yus et al, in preparation (see Chapter 7)
Transcription units	Manual curation from experimental data	*(252)*
Transcriptional regulation	Experimental	Yus et al, in preparation (see Chapter 7)

Table 3-1 summarizes the new experimental data for *M. pneumoniae* compiled in the WholeCellKB. The majority of this information came from automated analysis of raw experimental data, except for some cases in which lack of automated procedures required manual curation of the data. Such was the case for the determination of the transcription unit structure and the definition of transcription termination sites. Most of this data was ready to introduce into the database, or required minor modifications such as conversion from concentrations to copy numbers. However, complex transformations were required for some features. This was the case of the RNA abundances. In the previous model of *M. genitalium*, RNA expression was estimated on a gene-per-gene basis, as no information on transcription units was available. In contrast, transcription units (i.e. the operon and sub-operon structure) have been described in *M. pneumoniae* and thus RNA expression should be provided per transcription unit. Nevertheless, to facilitate the interpretation of transcriptomics experiments, this information had been always calculated and provided on a gene basis.

The problem of transforming expression data from genes to transcription units is not trivial. Operons are highly dynamic structures with different transcription start and termination sites, usually comprising various sub-operons that overlap partly or entirely with each other (Figure 3.1A). Therefore, it is complex to deconvolve the expression of a single gene in the different RNAs that include it. In order to solve this problem, we used an approach that is widely extended in determining eukaryotic RNA expression levels. For each operon, we considered each of the overlapping sub-operons as a different 'isoform'. We then considered the genes of each operon as different 'exons'. Thus, each sub-operon consists of an isoform containing some of the genes, or exons, of the entire operon (Figure 3.1B). Some exons are shared among different isoforms, but some are unique for given sub-operons. After this conversion, this problem is analogous to the problem of differential isoform expression from eukaryotic transcriptomics. We therefore used Cufflinks *(*284*)*, a widely-used software platform to determine isoform abundance in eukaryotes, to estimate the expression levels of each sub-operon. To verify the correct assignation of RNA abundances, we computed the expression of each gene by adding up the expressions of all sub-operons containing them, calculated with Cufflinks. The correlation between the original gene values and the newly computed ones is of 0.914.

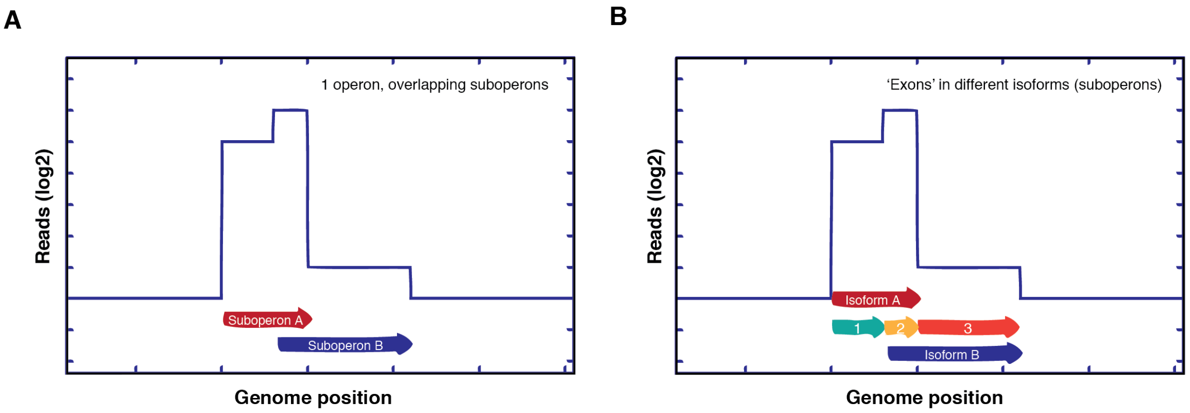
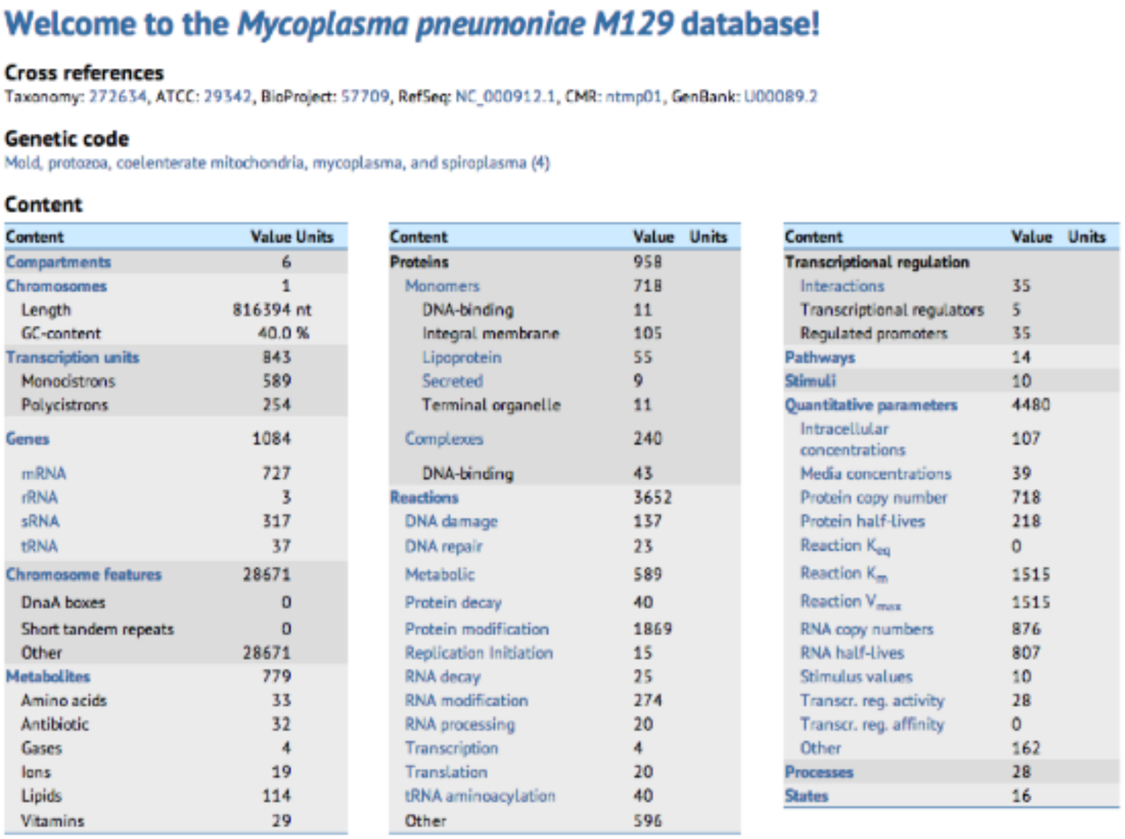


Figure 3.1. Transcript abundance determination. Scheme of the identification of overlapping regions between suboperons and exon mapping. (A) Transcripts (operons and suboperons) are annotated and the overlapping regions are identified. (B) Blocks of constant expression are identified as the regions between each TSS and TTS of the operon, and are categorizes as exons (1, 2, 3...). Each suboperon is then catalogued as an isoform including and excluding the corresponding exons. In the example, isoform A includes exons 1 and 2 and isoform B includes exons 2 and 3.

Besides the data obtained from prior publications or articles in preparation, the model also requires other types of information for which we do not have experimental values. For instance, the metal ions and other necessary co-factors of different enzymes in *M. pneumoniae* are, in the majority of cases, unknown. In this case, we extracted the information from the enzyme database BRENDA *(*285*)*. In some cases, data for *M. pneumoniae* or other Mycoplasmas was available in this database. In others, we had to retrieve data from other bacteria, with preference to closely related species. For other types of information, we needed to use bioinformatics predictions from different servers. For example, the signal peptides of all the proteins in *M. pneumoniae* were predicted using SignalP *(*286*)*, and the binding partners of protein chaperones such as the protein DnaK were predicted using Limbo *(*287*)*. The binding partners from the GroEL/GroES complex were inferred by homology with the binding partners in *E. coli (*288*)*.

All the data obtained either experimentally, through databases or prior publications, or from *in silico* predictions, was organized in a knowledge base, following the format of the *M. genitalium* WholeCellKB *(*279*)*. Figure 3.2 shows a summary of the data included in the knowledge base. It accounts for over 4400 quantitative parameters, compared to the 1836 included in the original *M. genitalium* model.



Welcome to the *Mycoplasma pneumoniae M129* database!

Cross references

Taxonomy: 272634, ATCC: 29342, BioProject: 57709, RefSeq: NC_000912.1, CMR: ntmp01, GenBank: U00089.2

Genetic code

Mold, protozoa, coelenterate mitochondria, mycoplasma, and spiroplasma (4)

Content

Content	Value	Units
Compartments	6	
Chromosomes	1	
Length	816394	nt
GC-content	40.0	%
Transcription units	843	
Monocistrons	589	
Polycistrons	254	
Genes	1084	
mRNA	727	
rRNA	3	
sRNA	317	
tRNA	37	
Chromosome features	28671	
DnaA boxes	0	
Short tandem repeats	0	
Other	28671	
Metabolites	779	
Amino acids	33	
Antibiotic	32	
Gases	4	
Ions	19	
Lipids	114	
Vitamins	29	

Content	Value	Units
Proteins	958	
Monomers	718	
DNA-binding	11	
Integral membrane	105	
Lipoprotein	55	
Secreted	9	
Terminal organelle	11	
Complexes	240	
DNA-binding	43	
Reactions	3652	
DNA damage	137	
DNA repair	23	
Metabolic	589	
Protein decay	40	
Protein modification	1869	
Replication Initiation	15	
RNA decay	25	
RNA modification	274	
RNA processing	20	
Transcription	4	
Translation	20	
tRNA aminoacylation	40	
Other	596	

Content	Value	Units
Transcriptional regulation		
Interactions	35	
Transcriptional regulators	5	
Regulated promoters	35	
Pathways	14	
Stimuli	10	
Quantitative parameters	4480	
Intracellular concentrations	107	
Media concentrations	39	
Protein copy number	718	
Protein half-lives	218	
Reaction K_{eq}	0	
Reaction K_m	1515	
Reaction V_{max}	1515	
RNA copy numbers	876	
RNA half-lives	807	
Stimulus values	10	
Transcr. reg. activity	28	
Transcr. reg. affinity	0	
Other	162	
Processes	28	
States	16	

Figure 3.2. Snapshot of the private website containing the *M. pneumoniae* knowledge base. The tables show a summary of all the data and parameters incorporated to the WholeCellKB-MPN

This difference arises mainly from the increased number of genes in *M. pneumoniae*, as well as the inclusion of the small RNAs (sRNAs), protein copy numbers, and protein half-lives. sRNAs were not accounted for in the original *M. genitalium* model. Another remarkable difference between both knowledge bases is related to the origin of the data used in the model (Figure 3.3). For *M. pneumoniae*, a large percentage of the information included was derived from studies in this bacterium, whilst in *M. genitalium*, lack of experimental data required the usage of data from other bacterial species.

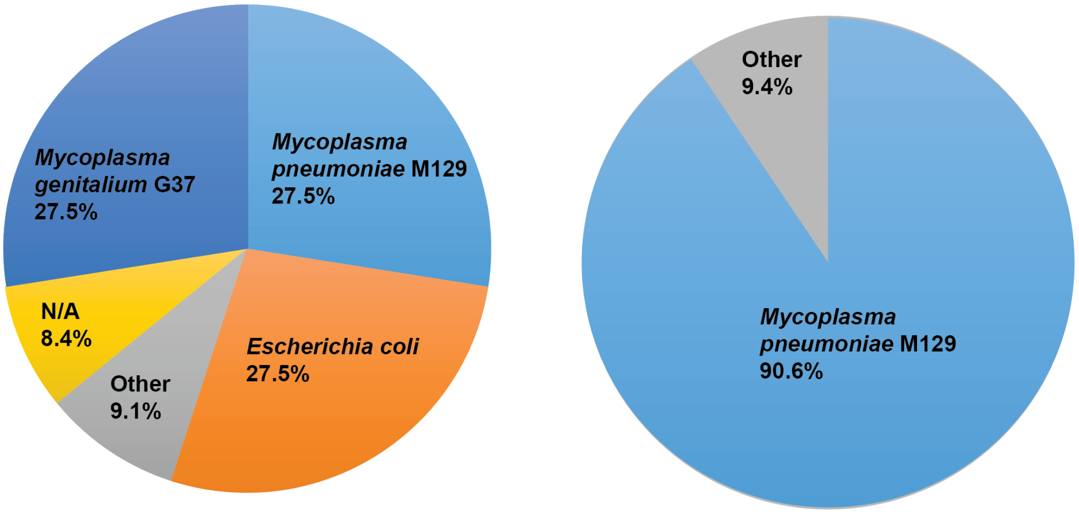


Figure 3.3. Sources of information for both the *M. genitalium* (left) and *M. pneumoniae* (right) models and knowledge bases. In *M. pneumoniae*, more than 90% the data used to train the model is extracted from studies on this bacterium. In *M. genitalium*, other bacteria were used, such as *M. pneumoniae* or *E. coli*.

3.4. Construction of the first version of the *M. pneumoniae* WC model

To generate the WC model of *M. pneumoniae,* we used the *M. genitalium* model as a starting point, given the similarity between these two closely related bacteria. The first step of this process consisted of replacing all the gene references of *M. genitalium* for their corresponding orthologs in *M. pneumoniae*. Then, we added all the remaining genes in *M. pneumoniae* without an ortholog in *M. genitalium*, and explicitly modeled their functions. The majority of the genes in this group correspond to hypothetical proteins or adhesins, proteins located in the membrane responsible for adhesion to the host cells and for generating antigenic variation. This simplified the modeling process, as few additional functions needed to be added. Only a minority of genes in this group has functions to be explicitly added to the WC model. One example is the gene *mpn372*, encoding for the CARDS (community-acquired respiratory distress syndrome) toxin, a unique pathogenicity determinant of *M. pneumoniae*. The protein encoded by this gene is thought to reach the host cells via endocytosis, and has an ADP-ribosyl transferase (ART) activity, catalyzing the addition of an ADP-ribose group to arginines in different proteins from the host. This modification causes protein inactivation, and this activity leads to vacuolation, disruption of cell homeostasis and eventually cell death. The function of this protein was included in the submodel dedicated to the host interaction. However, this module is rather simplified and it only considers that an inflammatory response is triggered if the CARDS toxin is expressed in the course of an infection.

Besides adding novel functions not described for the *M. genitalium* model, we also corrected genes that had been misannotated in the original model such as the alternative sigma factor MPN626, which was originally annotated as LuxR, a different transcriptional regulator.

Regarding metabolism, there are almost no differences between both species. However, in *M. pneumoniae* it has been reported that glycerol metabolism is another virulence determinant *(*289*)*. This is caused by the cytotoxicity of hydrogen peroxide (H_2O_2), a by-product of glycerol metabolism in *M. pneumoniae*, generated in the reaction that catalyzes the conversion of glycerol-3-phosphate to dihydroxyacetone-phosphate. The enzymes of glycerol metabolism are also present in *M. genitalium*, but this cytotoxic mechanism had not been reported. Therefore, a link between glycerol metabolism and virulence needed to be included. This was done by modifying the host interaction submodel. Thus, an inflammatory response is triggered in the model also when H_2O_2 is produced.

After modeling the function of new genes and establishing new links between different modules, the model had to be retrained in order to obtain the right parameters to simulate the cell cycle of *M. pneumoniae*. This is necessary as some of the parameters obtained from sources other than experimental data on this bacterium yielded simulation results that were incompatible with the observed growth of *M. pneumoniae*. Also, parameters obtained experimentally are measured by averaging over large populations of cells, and they are affected by both biological and experimental noise; so these also need to be optimized. Optimization of the model posed a number of challenges. Firstly, parameters of the different submodels need to be consistent with each other, so tuning the individual submodels separately is not appropriate. Secondly, using numerical optimization techniques in such a large model is also not feasible, as the problem becomes computationally intractable. Thus, a strategy similar to that used to train the *M. genitalium* model was used *(*244*,* 290*)*. This consisted in combining numerical optimization with model reduction. A reduced version of the WC model was created. This reduced model does not account for single-cell variation or temporal dynamics, but maintains all the parameters of the original model. Thus, it is much less expensive computationally, and numerical optimization can be applied to estimate the parameter values. In the reduced model, constraints were identified to determine upper and lower bounds for different parameters. These constraints were determined by studying the requirements of the different submodels, and considering that all cellular components must double in the course of a cell cycle. After setting upper and lower bounds, numerical optimization was used to estimate the values of the parameters that satisfied the constraints applied and deviated minimally from the experimentally determined values. This parameter set was then tested in the full WC model, and finally some of the parameters were manually tuned to improve the similarity of the predictions to the experimental data.

This first WC model of *M. pneumoniae* has been implemented in the cluster facilities of the Centre for Genomic Regulation (CRG). In this computing cluster, that consists of 160 computing nodes with 2620 cores, simulation of the cell cycle of a single bacterium lasts around 12 hours. Analysis of the output of the simulation can be automatized to facilitate interpretation. Figure 3.4 shows an example of the information that can be obtained from the simulations of *M. pneumoniae*.

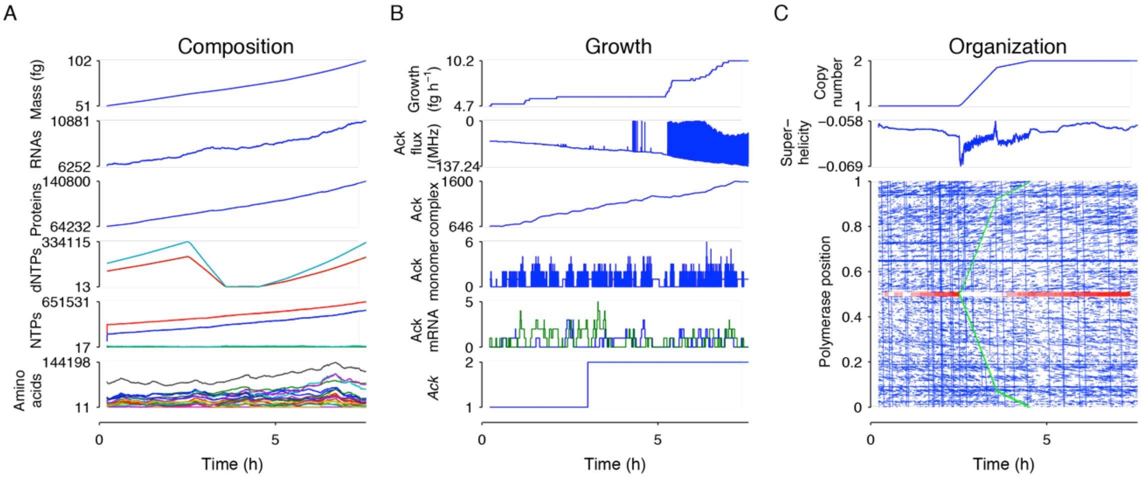


Figure 3.4. Example of the whole-cell model output. The panels show different aspects of the growth of a single wild-type *M. pneumoniae* cell over the duration of a cell cycle, 8 hours. (A) Cell composition, including the total mass, the number of RNAs proteins, nucleotides and amino acids. (B) Growth including mass growth rate, and the dynamics of the *ack* gene, as an example (DNA copies, mRNA copies, monomers, complexes and reaction fluxes)

3.5. Current development of the WC model of *M. pneumoniae*

One of the main applications of WC models is the discovery of ‘knowledge gaps’ in the biology of the simulated organisms. Comparison of the simulation results with experimental data can be used to highlight discrepancies between the model and the biology of the bacterium. These discrepancies may arise from inaccuracies in the parameters *(*291*)* or from mechanisms that are incorrectly described in the model. The chart depicted in Figure 3.5 illustrates the process of iteratively using the WC model of *M. pneumoniae* to discover these inconsistencies, improving our knowledge on a specific part of the physiology and updating the model.

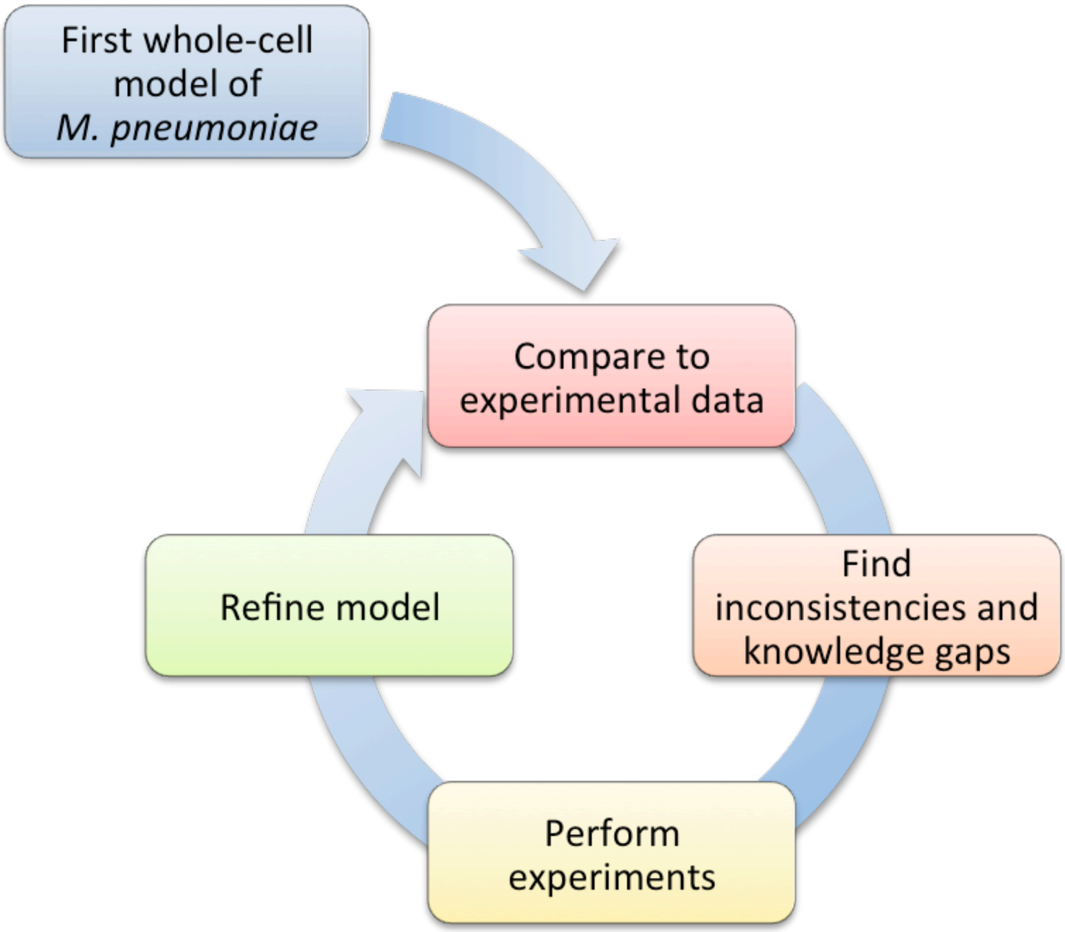

First whole-cell model of *M. pneumoniae*
Compare to experimental data
Find inconsistencies and knowledge gaps
Perform experiments
Refine model

Figure 3.5. Gap filling and whole-cell modeling for new discovery. Flow chart illustrating the iterative process that involves finding parts of the model that do not agree with the experimental data obtained in M. pneumoniae, performing new experiments to better understand these specific traits of the physiology of the bacterium, and then refining and improving the model using these new discoveries.

We compared the results of the simulations of wild-type *M. pneumoniae* with our experimental data and found a series of inconsistencies that are currently being addressed to obtain an improved second version of the *M. pneumoniae* model. The first one relates to the structure of the model and the selection of the timestep of 1 second. The problem arises when two of the submodels need to interact faster than what this one-second timestep allows. This occurred with the submodels of translation and tRNA aminoacylation. Translation occurs in bacterial cells at an average rate of ~12-21 amino acids per ribosome per second *(*292*)*. In *M. pneumoniae* there are around 140 ribosomes per cell *(*203*)*, which means that, if all ribosomes were active in the cell, the bacterium would need to have ~1500-3000 aminoacylated tRNAs per second to be able to translate proteins. Additionally, the simulated cells would need a similar number of uncharged tRNAs per second, to be charged in the tRNA aminoacylation process. These numbers predicted are much higher than the numbers of tRNAs observed in our RNA-sequencing experiments (600-700 tRNA copies). This discrepancy could be caused by multiple factors. It could be possible that only a minority of ribosomes are active at any moment during the growth of *M. pneumoniae*. This could be related to the fact that low translation efficiencies are required to compensate for the noise in RNA levels, to maintain stable protein levels *(*293*)*. This should be assessed in the WC model, as currently our simulations show that the majority of ribosomes are active throughout the entire cell cycle of the bacterium. It is also possible that in our RNA-sequencing experiments, we are underestimating the actual numbers of tRNAs. Due to their small size, they show a large variability across different experiments, and their actual copy numbers are difficult to estimate. The reason of this is that in the RNA-sequencing library preparation protocol, there is a step of RNA fragment size selection. Depending on the size selected, tRNAs might be depleted from the library pool and thus underestimated. Nevertheless, even the largest numbers observed in any of our experiments are far below those estimated by the model. A third possibility is that tRNA aminoacylation is a very fast process, occurring simultaneously with translation. In such case, with a fast recycling of tRNAs, the numbers of tRNAs needed would be much lower than those predicted by the model, and would become more similar to those observed in our experiments. We focused in this last possibility. Thus, the two processes, translation and tRNA aminoacylation, need to interact faster than the one second timestep. However, shortening this timestep in the WC model would increase a lot the computational cost of the simulations, which is already quite demanding. Therefore, we decided to fuse both processes in a single module of the WC model. To do so, all the reactions involved in each of the two processes were merged and implemented in the same module. In this way, the effect on the performance of the entire model is minimal, and the simultaneous translation and tRNA recycling can approximate better the observed tRNA numbers in our experiments.

The first version of the WC model found a large amount of unaccounted energy (ATP and GTP) produced by the cell, that was not used in any of the processes described. Similar predictions were made in the original model of *M. genitalium (*244*)*. In a previous study in our lab, the function of the ATPase in maintaining the intracellular pH and membrane potential, and keeping an optimal proton gradient for nutrient import, was identified as the most important energy sink *(*233*)*, confirming what had been previously observed in gram positive bacteria *(*294*,* 295*)*. This energy sink was not explicitly included in the first version of the WC, and is being currently parameterized and implemented in the model.

Other changes in the model that are currently being implemented relate to metabolic and transcriptional regulation. One example refers to the regulation of glucose import. We have observed experimentally that stress caused by low pH generates a phenotype similar to that of glucose starvation (Yus et al, in preparation; see Chapter 7). Further exploration revealed that glucose import stops at low pH, and can be restored upon re-buffering the pH of the culture medium *(*233*)*. Such a mechanism was not described in the model, but is needed to understand how the cell responds to external perturbations. We reviewed the literature on the topic, to find that the key to this regulation could lie in the phosphotransferase system of *M. pneumoniae*. A scheme of this system, centered in the HPr protein, is presented in Figure 3.6. It has been described that the histidine residue whose phosphorylation is needed for sugar import (His-15) can change its conformation according to its protonation state *(*296*)*. This histidine can be present in open and closed forms. The closed conformation predominates at pH 7, whilst at a lower pH the protonated, open form is prevalent. It was also observed that only the closed form can be phosphorylated. This histidine has a pKa value that is close to neutrality, meaning that small intracellular changes in the pH, due to acidification of the medium by lactate and acetate secretion, can change its protonation state and prevent phosphorylation. This suggests a possible mechanism of blocking the first step required for sugar import.

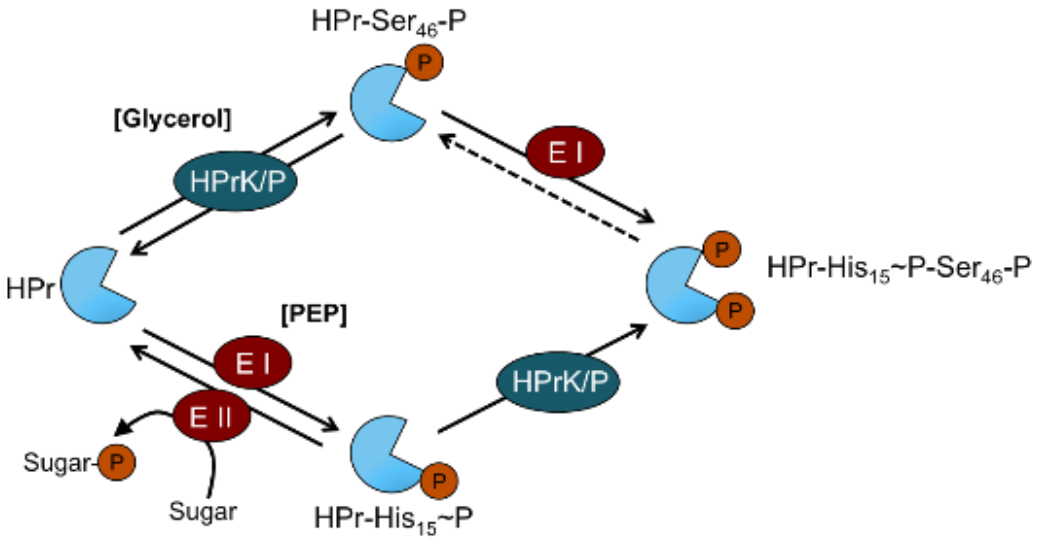


Figure 3.6. Phosphotransferase system of *M. pneumoniae*. Sugar transport is regulated via the HPr protein. This protein can be phosphorylated at two positions His15 and Ser46. Phosphorylation of His15 is mediated by the Enzyme I of the sugar transport system (E I), and it is favoured by high concentrations of phosphoenol-pyruvate (PEP). Enzyme II (E II) then catalyzes the transfer of the phosphate group from His15 to the imported sugar. Phosphorylation of Ser46 occurs via the HPrK/P protein and it is dependent on high concentrations of glycerol. Image adapted from Stülke and Halbedel (2005).

Regarding transcriptional regulation, the first version of the WC model included little information on transcription factors or specific promoter strengths. Instead, the differential transcription was modeled by explicitly taking RNA levels of the different transcripts, and converting these levels into binding probabilities for the RNA polymerase complex, and only a few transcription factors were included in the model, some of them with the wrong target specificity. We now know the targets of virtually all TFs in this bacterium (see Chapter 7) and we will include them in the modified model. However, these regulatory mechanisms included in the model could not reproduce specific transcriptional changes observed when the cells enter the stationary growth phase, as these could not be related to known transcription factor regulation (see Chapter 7). Therefore, we investigated alternative regulatory mechanisms, and found that the switch regulating the major transcriptional changes occurring between exponential and stationary phase could be based in nucleotide abundances (see Chapter 7). ATP and GTP concentrations can act as regulators of transcription initiation, by stabilizing the open complex of the RNA polymerase with the promoter. Unstable open complexes are sensitive to the concentrations of these nucleotides. If they are present in high concentrations, they can rapidly incorporate at the +1 position of the RNA stabilizing the open complex and preventing dissociation of the polymerase *(74)*. In contrast, if concentrations are low, the complex with the RNA polymerase is quickly dissociated and transcription is prevented. This mechanism has been demonstrated to regulate transcription in response to amino acid starvation in *Bacillus subtilis*. In this bacterium, the response to amino acid starvation is mediated by the production of (p)ppGpp by the RelA protein. The production of this signalling metabolite uses GTP, decreasing its availability inside the cell. The reduced levels of GTP causes a downregulation of RNAs whose first nucleotide is guanosine. This mechanism is accompanied by an increase in ATP and an upregulation of RNAs having adenine in the +1 position *(*76*)*. A similar mechanism has been experimentally characterized in *M. pneumoniae* (see Chapter 7) and will be implemented in a new version of the model.

Future improvements of the model will also include novel features, such as the structure of the chromosome of *M. pneumoniae*. A recent study in our lab has resolved the 3D structure of the chromosome (Trussart *et al*, under review). The structure of the chromatin can play a crucial role in various processes, such as replication, cell division and transcription *(297)*. More specifically, our study revealed that the chromosome of *M. pneumoniae* can be divided in regions, termed 'chromosome interacting domains' or CIDs. Genes located within the same CID tend to be more co-expressed than genes located in different CIDs (Trussart *et al*, under review). This could be implemented in the WC model by biasing the RNA polymerase binding. Concretely, after transcribing a certain gene, the binding affinities of the RNA polymerase to promoters located in the same CID could be increased, to increase the probability of moving to a promoter in the same region versus promoters located farther in the 3D structure of the chromosome. Additionally, we will explicitly model the error rates in transcription and translation. For translation, these error rates have been recently characterized, and were found to be much higher than in other bacteria *(*283*)*. Translation errors have been shown to have an impact on growth rate and fitness in bacteria *(*298*)*. The inclusion of these error rates poses the need to include protein structural information in the model, as it is strictly necessary to determine which errors will result in functional or nonfunctional proteins.

3.6. Potential applications of the WC model of *M. pneumoniae*

The applications of WC modeling are numerous. Besides pointing to gaps in our knowledge of the physiology of the modeled bacterium, they can be used as a tool to guide genome engineering. As an example of this, the WC model of *M. genitalium* was used to predict the effect of introducing exogenous genes in the cell *(*299*)*. Other possible applications of WC models include the prediction of the combinations of genes that can be eliminated in order to obtain a true minimal genome. Such a minimal genome could be then tested experimentally, and used as a chassis to add new biosynthetic pathways of interest. Also, in order to convert *M. pneumoniae* into a valuable model system for synthetic biology, we could use the WC model to predict which genes affect growth of this bacterium. *M. pneumoniae* is a slow-growing bacterium, with a doubling rate of 8 hours. This hampers its usage as a model organism for industrial applications, despite the advantages of such a reduced genome. Therefore, using the model to predict ways to boost growth of this bacterium could be largely beneficial to the synthetic biology community.

In order to develop these applications, we need to obtain highly accurate models. With the current models of *M. genitalium* or *M. pneumoniae*, we can make such predictions but we are still limited by the lack of knowledge in some of the areas of the physiology of these bacteria. In many cases, these predictions are qualitative, but not accurate quantitatively. We therefore anticipate that with few iterative rounds of 'knowledge gap filling', together with the mechanisms that we are currently incorporating in the model, this one can be used to make accurate and reliable predictions that can have an impact in the field of synthetic biology.

3.7. Author contributions

The *M. pneumoniae* WC model was developed in collaboration with Jonathan R. Karr (Icahn School of Medicine at Mount Sinai) and Markus W. Covert (Stanford University). VLR, JRK, MLS and LS compiled the experimental data and reconstructed the modeled species, reactions, and parameters. JRK constructed the model. VLR simulated the model at the CRG cluster facility. A second version of the model is currently under development.

Lloréns-Rico, V., Serrano, L., & Lluch-Senar, M. (2014). *Assessing the hodgepodge of non-mapped reads in bacterial transcriptomes: real or artifactual RNA chimeras?. BMC genomics*, *15*(1), 1.

4. Assessing the hodgepodge of non-mapped reads in bacterial transcriptomes. Real or artifactual RNA chimeras?

Lloréns-Rico V, Serrano L, Lluch-Senar M. Assessing the hodgepodge of non-mapped reads in bacterial transcriptomes: real or artifactual RNA chimeras? BMC Genomics. 2014 Jul 29;15:633. doi: 10.1186/1471-2164-15-633

Lloréns-Rico, V., Lluch-Senar, M., & Serrano, L. (2015). *Distinguishing between productive and abortive promoters using a random forest classifier in* Mycoplasma pneumoniae. *Nucleic acids research*.

5. Distinguishing between productive and abortive promoters using a random forest classifier in *Mycoplasma pneumoniae*.

Lloréns-Rico V, Lluch-Senar M, Serrano L. Distinguishing between productive and abortive promoters using a random forest classifier in Mycoplasma pneumoniae. Nucleic Acids Res. 2015 Apr 20;43(7):3442-53. doi: 10.1093/nar/gkv170

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.H., Bork, P., Glass, J.I., Serrano, L., and Lluch-Senar, M., (2016). *Bacterial antisense RNAs are mainly the product of transcriptional noise*. *Science Advances*.

6. Bacterial antisense RNAs are mainly the product of transcriptional noise

Lloréns-Rico V, Cano J, Kamminga T, Gil R, Latorre A, Chen WH, Bork P, Glass JI, Serrano L, Lluch-Senar M. Bacterial antisense RNAs are mainly the product of transcriptional noise. Sci Adv. 2016 Mar 4;2(3):e1501363. doi: 10.1126/sciadv.1501363

7. A comprehensive study of transcriptional control in a genome-reduced bacterium shows the importance of non-transcription factor regulation.

Yus, E.*, Lloréns-Rico, V.*, Martinez, S., Sevin, D., Sauer, U., Gallo, C., Lluch-Senar, M. & Serrano, L. A comprehensive study of transcriptional control in a genome-reduced bacterium shows the importance of non-transcription factor regulation. *In preparation*.

* Equal contribution

> *"We now possess enough reliable data to conclude that the world is certainly much stranger than expected"*
>
> Joseph Rain

7.1. Abstract

Determining the gene regulatory network is basic to have a global understanding of cell behavior. In general, studies of transcriptional regulation are limited to the annotated transcription factors (TFs), obviating other non-canonical regulators, or even unknown key players. Here, we describe the first systematic analysis of the protein-DNA interactome in a minimal bacterium, *Mycoplasma pneumoniae*. We have identified all potential DNA-binding proteins (DNABPs, 105 out of 689 annotated proteins) by DNA affinity chromatography, DNA pull-downs, and intact chromatin isolation. For each of them, together with some others added from the literature, we have determined their binding sites by ChIP-seq or biochemical assays. Also, we have studied the effect of overexpression and depletion of these putative DNA binding proteins by characterizing different *M. pneumoniae* strains using different 'omics' approaches. Strikingly, we found

new moonlighting functions for highly conserved proteins, that show DNA binding properties as well as other activities, like proteases and metabolic enzymes. Interestingly, for the majority of the proteins analyzed, we found no transcriptome or growth phenotype upon overexpression (64.8%, 81 out of 125 proteins with both experiments). This is indicative of the robustness of the system, despite its simplicity. This integrative approach revealed that metabolic control is a key regulatory element, highlighting a non-TF factor layer of regulation in bacteria. This layer would include, but is not limited to, the role of supercoiling and the genomic context, the RNA structure, forming riboswitches or condition-dependent terminators, the RNA regulated decay, and the abundances of certain metabolites.

7.2. Introduction

Unveiling the gene regulatory network (GRN) of an organism is the first and most important step to understand its physiology. Current studies focus on the search of protein-based regulatory factors. Such studies rely on genome annotation and comparative sequence analysis *(*421*,* 422*)*. The individual or systematic analysis of the genes targeted by these proteins, the so-called regulons, can be done using bottom-up approaches *(*183*,* 423*)*. This approximation has a strong limitation, due to the fact that many transcriptional regulators might not have been described/annotated yet, and there are many non-canonical regulators, such as metabolic enzymes *(*424*,* 425*)* or even structural proteins (Nucleoid Associated Proteins, NAPs; *(*426*)*) with dual roles that are usually not included in such studies.

Moreover, it is not clear if all transcriptional regulation in a bacterium is dependent only on classical TFs. Even in the case of well-studied bacterial models, such *as Escherichia coli* or *Bacillus subtilis*, less that 40 and 25% of the genes, respectively, are regulated by TFs *(*427–429*)*. This is even more remarkable in streamlined genomes such as endosymbionts *(*430*)*. Another player in transcriptional regulation are small RNAs, although their role in transcriptional regulation is still controversial *(*400*,* 401*)*. Indeed, a systematic study by our group has strengthened the hypothesis that many antisense RNAs detected by RNA sequencing are the result of spurious transcription (see Chapter 6; *(*431*)*), and lack a regulatory function. Nevertheless, even if some of them have regulatory roles, their expression must be also tightly regulated.

As bacteria rely on metabolism to adapt to many environmental stimuli, cell signalling is related in many cases to small metabolites and second messengers (432). Therefore, the overall physiology, growth rate or metabolic status of the cell can also be major contributors to transcriptional regulation (433, 434). Some of these regulatory mechanisms occur at the level of the RNA polymerase, such as the regulation by the alarmone (p)ppGpp (435), or the regulation of the transition from closed to open transcription complex by the concentration of certain NTPs (436, 437).

Genome organization also plays a role in gene regulation. Recently, our group has inquired the role of genomic organization in gene co-expression (23). We found that the degree of transcriptional co-expression between co-directional adjacent genes is tightly related to their capacity to be transcribed *en bloc*, into the same mRNA via RNA polymerase read-through, requiring a revision of the operon concept. Additionally, a report considering evolutionary conservation or synteny also challenges this concept and suggests that local domains can share a TF regulation even if not all the genetic components bear a binding site (438). Besides the linear organization of genes along bacterial chromosomes, the three dimensional structure of the chromosome may play a regulatory role. The chromosome structure of various bacteria has been elucidated, and the presence of 'chromosome interacting domains' or CIDs, with a role in transcriptional coordination has been pointed out ((439), Trussart et al, under review). Factors such as structural proteins and NAPs provide an additional regulatory level by influencing DNA topology (440, 441). DNA supercoiling also plays an important role in transcriptional regulation (442).

Besides, transcription termination attenuation by riboswitches and rho-independent terminators plays also an important regulatory role in many bacteria (89, 90). Finally, regulation of RNA half-life by bacterial RNAses could alter RNA levels without interfering in the transcription process. The RNA degradosome complex may contain different subunits depending on the metabolic status of the cell, changing the specificity for certain transcripts (120, 121).

With so many regulatory levels, the exact contribution of TFs to gene regulation has still to be clarified. Importantly, although many of these factors contributing to transcriptional regulation have been individually analyzed, the quantitative contribution of all these factors to the global transcriptional control and their hierarchy has not been addressed in bacteria yet.

M. pneumoniae has become a model organism in the study of minimal cells. This bacterium has undergone massive genome reduction due to its parasitic lifestyle, and it is thought to have maintained only the basic machinery to sustain autonomous life *(*73*)*. Regarding transcription, it was long believed to have little regulation. It was even postulated that transcription would happen in an autonomous non-regulated manner, due to their low GC content *(*443*)*. This was supported by the first genome annotation, that showed a lack of alternative sigma factors, low conservation of promoter regulatory regions and the presence of only a handful of canonical TFs. This was also consistent with the observation that cells living in an uniform environment have simpler and less hierarchical GRNs *(*444–446*)*. Nevertheless, recent studies have suggested transcriptional responses equivalent to those of more complex microbes, even in the absence of the cognate regulator, such as the SOS response *(*72*)*. Moreover, the variety of phenotypes observed cannot be explained only with the annotated TFs *(*72*,* 249*)*, and suggests hidden layers of regulation, and/or a limited knowledge of the real effectors. In this respect two recent works have shown the presence of an alternative sigma factor in *M. genitalium* (MG_428) that activates the recombination machinery *(*255*,* 447*)*, and is similar to MPN626 in *M. pneumoniae.*

Here, we have analyzed the transcriptional regulation of *M. pneumoniae* in a global, comprehensive manner to define the extent to which all the factors enumerated above determine transcription regulation. We have used classic methodologies such as DNA affinity chromatography and chromatin isolation to determine the entire protein-DNA interactome in an unbiased manner. With these approaches, we have identified 105 putative DNA binding proteins, and we have also considered the regulators annotated in the genome or described in other studies. We have characterized their function by overexpressing each of these proteins, and also in some cases by generating dominant negative mutants or using transposon insertion mutant strains. For those that exhibited DNA-binding properties or were annotated as DNA-binding proteins, we have identified their binding sites by chromatin immunoprecipitation and sequencing (ChIP-seq). Besides the ChIP-seq experiments, we performed a protein occupancy display (POD) experiment, and observed that, of all binding sites observed in our experiments, only 22% were covered only in the POD and not by any protein in our study. This demonstrates a large coverage of all the DNA-binding proteins in this bacterium. Furthermore, we have analyzed the changes in RNA and/or protein expression, as well as phenotypic changes in the growth rate.

Using this approach, we have identified 9 canonical TFs, and we have annotated their targets and their binding sites. We have also defined the targets for 30 regulators (that do not bind DNA directly). Moreover, we have combined the regulon information of TFs and regulators with the stress responses of *M. pneumoniae* to 111 different perturbations. We can thereby hypothesize on the involvement of the TFs in a particular response. Furthermore, we have discovered or validated 8 structural proteins, some of them involved in cell division (NAPs, *(*448*)*). Importantly, the majority of candidates validated by transcriptomics and growth rate measurements (55.88%; 76 out of 136) do not present any phenotype, indicating the robustness of the cell to such changes.

Overall, we can only associate a maximum of 50% of the expression variability in perturbation experiments to *bona fide* TFs and regulators, yet we have observed that experimental noise causes an underestimation of this value. The remaining variance may be related to the genome and chromosome organization, supercoiling, riboswitches, rho-independent terminators and specific NTP concentrations that are determined by the metabolic and growth status of the cell. This indicates that ancient, basal mechanisms of regulation exist and are revealed by the relative low complexity of this gene regulatory network.

7.3. Results

DNA-protein interactome comprises 105 preliminar candidates

M. pneumoniae has a reduced set of transcriptional regulators when compared with other species, even in percentage to the genome size *(*449*)*. We had previously reported an initial characterization of some of the *bona fide* annotated regulators, but found few targets regulated by them *(*72*)*. However, that same study indicated the existence of a variety of regulatory responses, exemplified in various gene expression clusters that could barely be explained by a handful of TFs *(*249*)*. This prompted us to expand our view and undertake non-biased approaches. We re-examined the genome annotation seeking for all TF candidate proteins (see Table S1 and Figure 7.1). We were able to identify possible candidates, including new putative TFs, structural DNABPs or moonlighting proteins such as metabolic enzymes or even proteases.

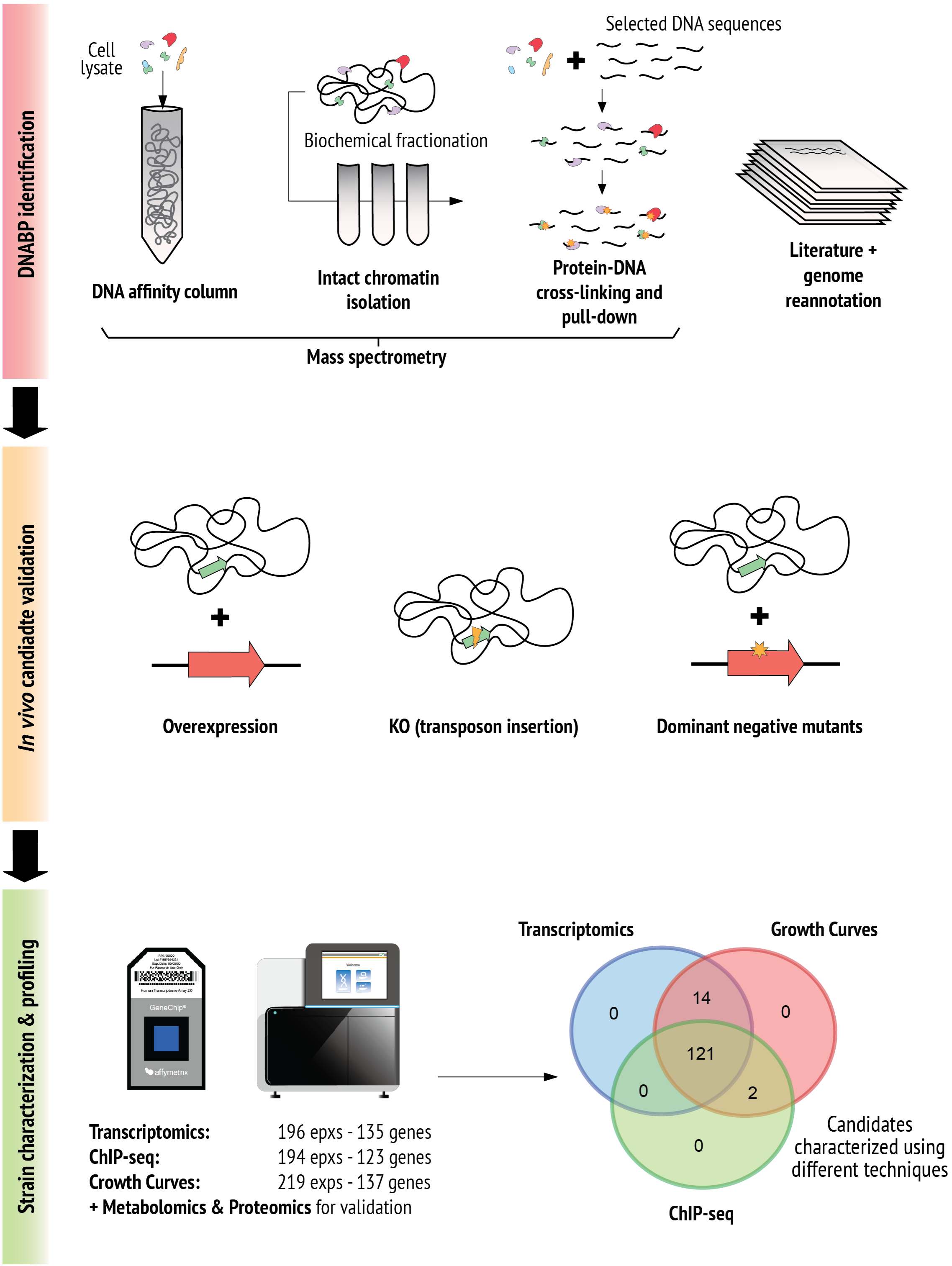

DNABP identification
Cell lysate
Selected DNA sequences
Biochemical fractionation
DNA affinity column
Intact chromatin isolation
Protein-DNA cross-linking and pull-down
Literature + genome reannotation
Mass spectrometry
In vivo candiadte validation
Overexpression
KO (transposon insertion)
Dominant negative mutants
Strain characterization & profiling
Transcriptomics: 196 epxs - 135 genes
ChIP-seq: 194 epxs - 123 genes
Crowth Curves: 219 exps - 137 genes
+ Metabolomics & Proteomics for validation
Transcriptomics
Growth Curves
ChIP-seq
0
14
0
121
0
2
0
Candidates characterized using different techniques

Figure 7.1. Scheme depicting the identification, validation and characterization of the candidate DNABPs. To reconstruct the DNA-protein interactome and the gene regulatory network of *M. pneumoniae*, first all candidate DNABPs were identified using a variety of mechods, including DNA affinity chromatography, chromatin isolation and protein-DNA cross-linking. Candidates reported in the scientific literature or given by the genome annotation were also included in the study. Candidates were either overexpressed or knocked-out by transposon insertion, and sometimes, dominant negatives were also overexpressed in *M. pneumoniae*. The strains generated were characterized by different omics to assess the function of each DNABP.

To confirm these candidates and find novel ones, we performed DNA affinity chromatography followed by mass spectrometry. As expected, with this technique we recovered all known TFs, the basic DNA replication and repair machinery, as well as the whole RNA polymerase (RNAP) complex, validating this methodology to identify DNABPs. However, RNA binding proteins were also detected, indicating that our approach cannot well distinguish between RNA- and DNA-binding proteins, or that some RNA-binding proteins have true moonlighting functions. Similar results were found after protein crosslinking and pull-down using selected DNA sequences from *M. pneumoniae* (Table S2), and by means of a biochemical fractionation using a sucrose gradient to isolate chromatin with all bound interactors.

We then calculated ROC curves using a predefined gold set of DNA- and RNA-binding proteins (Table S3), to set a threshold for each experiment that allowed identification of true DNABPs. With the results of this analysis, we established a consensus of putative DNABPs (Table S4), which includes all known TFs, DNA structural and replication-associated proteins, the RNAP complex, as well as previously described moonlighting proteins like the Leu and Lon proteases *(*450*,* 451*)*, metabolic enzymes *(*452*)*, and new putative DNA binding proteins. In total, this consensus of DNABPs is composed of 105 possible candidates. From this list, we removed DNA replication proteins, and additionally, for those cases in which several members of the same protein family were identified, we kept only one of them. We also added other proteins not passing our filters but reported as DNABPs in previous studies in *M. pneumoniae* or other bacteria, as some transcription factors are expressed in a transitory manner (i.e. MPN626, an alternative sigma factor) and they would not be identified in our previous experiments *(*447*)*. Table 7-1 shows the final list of all proteins included in the screening

Table 7-1. Final list of candidates characterized in our study

Candidate	Growth curve	Transcriptomics	ChIP-seq	Candidate	Growth curve	Transcriptomics	ChIP-seq
MPN002	Yes	Yes	Yes	**MPN329**	Yes	Yes	Yes
MPN004	Yes	Yes	Yes	**MPN330**	Yes	Yes	Yes
MPN015	Yes	No	Yes	**MPN332**	Yes	Yes	Yes
MPN020	Yes	Yes	Yes	**MPN348**	Yes	Yes	Yes
MPN024	Yes	Yes	Yes	**MPN349**	Yes	Yes	Yes
MPN027	Yes	Yes	No	**MPN352**	Yes	Yes	Yes
MPN030	Yes	Yes	Yes	**MPN368**	Yes	Yes	Yes
MPN032	Yes	Yes	Yes	**MPN372**	Yes	Yes	No
MPN038	Yes	Yes	Yes	**MPN397**	Yes	Yes	Yes

MPN051	Yes	Yes	Yes	**MPN400**	Yes	Yes	Yes
MPN055	Yes	Yes	Yes	**MPN401**	Yes	Yes	Yes
MPN063	Yes	Yes	Yes	**MPN420**	Yes	Yes	Yes
MPN064	Yes	Yes	Yes	**MPN421**	Yes	Yes	No
MPN067	Yes	Yes	Yes	**MPN424**	Yes	Yes	Yes
MPN069	Yes	Yes	Yes	**MPN426**	Yes	Yes	Yes
MPN076	Yes	Yes	No	**MPN428**	Yes	Yes	Yes
MPN077	Yes	Yes	No	**MPN430**	Yes	Yes	Yes
MPN081	Yes	Yes	Yes	**MPN440**	Yes	Yes	Yes
MPN082	Yes	Yes	Yes	**MPN443**	Yes	Yes	Yes
MPN106	Yes	Yes	Yes	**MPN446**	Yes	Yes	Yes
MPN114	Yes	Yes	No	**MPN473**	Yes	Yes	Yes
MPN122	Yes	Yes	Yes	**MPN475**	Yes	Yes	Yes
MPN124	Yes	Yes	Yes	**MPN478**	Yes	Yes	Yes
MPN127	Yes	Yes	Yes	**MPN481**	Yes	Yes	Yes
MPN133	Yes	Yes	No	**MPN482**	Yes	Yes	Yes
MPN140	Yes	Yes	Yes	**MPN484**	Yes	Yes	No
MPN148	Yes	Yes	Yes	**MPN485**	Yes	Yes	Yes
MPN154	Yes	Yes	Yes	**MPN487**	Yes	Yes	Yes
MPN159	Yes	Yes	Yes	**MPN490**	Yes	Yes	Yes
MPN164	Yes	Yes	Yes	**MPN499**	Yes	Yes	Yes
MPN165	Yes	Yes	Yes	**MPN507**	Yes	Yes	Yes
MPN166	Yes	Yes	Yes	**MPN516**	Yes	Yes	Yes
MPN168	Yes	Yes	Yes	**MPN518**	Yes	Yes	Yes
MPN173	Yes	Yes	Yes	**MPN525**	Yes	Yes	Yes
MPN178	Yes	Yes	Yes	**MPN526**	Yes	Yes	Yes
MPN191	Yes	Yes	Yes	**MPN529**	Yes	Yes	Yes
MPN192	Yes	Yes	Yes	**MPN545**	Yes	Yes	No
MPN194	Yes	Yes	Yes	**MPN547**	Yes	Yes	Yes
MPN197	Yes	Yes	Yes	**MPN549**	Yes	Yes	Yes
MPN208	Yes	Yes	Yes	**MPN554**	Yes	Yes	Yes
MPN222	Yes	Yes	Yes	**MPN555**	Yes	Yes	Yes
MPN223	Yes	Yes	Yes	**MPN559**	Yes	Yes	Yes
MPN229	Yes	Yes	Yes	**MPN563**	Yes	Yes	Yes
MPN239	Yes	Yes	Yes	**MPN566**	Yes	Yes	Yes
MPN241	Yes	Yes	Yes	**MPN568**	Yes	Yes	Yes
MPN243	Yes	Yes	No	**MPN569**	Yes	Yes	Yes
MPN244	Yes	Yes	Yes	**MPN572**	Yes	Yes	Yes
MPN246	Yes	Yes	No	**MPN574**	Yes	Yes	Yes
MPN247	Yes	Yes	Yes	**MPN576**	Yes	Yes	Yes
MPN248	Yes	Yes	Yes	**MPN590**	Yes	Yes	No
MPN250	Yes	Yes	Yes	**MPN606**	Yes	Yes	Yes
MPN252	Yes	Yes	Yes	**MPN608**	Yes	Yes	Yes
MPN255	Yes	Yes	Yes	**MPN615**	Yes	Yes	Yes
MPN265	Yes	Yes	Yes	**MPN617**	Yes	Yes	Yes

MPN266	Yes	Yes	Yes	**MPN621**	Yes	Yes	Yes
MPN269	Yes	Yes	No	**MPN626**	Yes	Yes	Yes
MPN273	Yes	Yes	Yes	**MPN627**	Yes	Yes	Yes
MPN275	Yes	Yes	Yes	**MPN633**	Yes	Yes	Yes
MPN280	Yes	Yes	Yes	**MPN634**	Yes	Yes	Yes
MPN284	Yes	Yes	No	**MPN635**	Yes	Yes	Yes
MPN287	Yes	Yes	Yes	**MPN638**	Yes	No	Yes
MPN294	Yes	Yes	Yes	**MPN663**	Yes	Yes	Yes
MPN295	Yes	Yes	Yes	**MPN667**	Yes	Yes	Yes
MPN300	Yes	Yes	Yes	**MPN673**	Yes	Yes	Yes
MPN301	Yes	Yes	Yes	**MPN674**	Yes	Yes	Yes
MPN303	Yes	Yes	Yes	**MPN677**	Yes	Yes	Yes
MPN314	Yes	Yes	Yes	**MPN683**	Yes	Yes	Yes
MPN315	Yes	Yes	Yes	**MPN686**	Yes	Yes	Yes
MPN316	Yes	Yes	Yes				

Physical genomic interactions by ChIP-Seq

To characterize the DNA-binding properties of the identified putative DNABPs (Table 7-1), the genomic regions recognized by them were identified by ChIP-Seq. To discard tag-artifacts, N- and/or C-terminal tagged forms of some of our protein candidates were expressed in *M. pneumoniae*. As positive control, ChIP-seq analysis was performed from four *M. pneumoniae* strains expressing RNAP subunits (RpoA and RpoB), as well as of the σ^{70} and the RNAP associated transcriptional regulator Spx (MPN266; *(*453*)*). As expected, these proteins were mainly bound to promoter regions. To assess the technique reproducibility, the ChIP-seq experiments were duplicated for a number of known DNABPs (MPN352, σ^{70}; MPN686, DnaA; MPN266, Spx) revealing quasi identical ChIP-seq profiles in both replicates, with correlations up to R=0.9 between replicates (Figure S1).

To assess our coverage of the protein-DNA interactome, we performed protein occupancy display (POD) experiments at 6 and 96h (exponential and stationary growth phases). In these experiments we identify all the chromosome regions bound by any protein, regardless of which. We compared the DNA regions bound in these experiments with the peaks from all our ChIP-seq experiments, and observed that 78% of the total protein-bound regions was covered by ChIP-seq and POD simultaneously, or by ChIP-seq only. This demonstrates a high coverage of all the DNA-binding proteins in *M. pneumoniae*. The regions not covered in our ChIP-seq analysis could correspond to DNA protected sites only at specific growth phases (not all experiments were performed at both 6h and 96h), to proteins forming part of complexes that bind DNA indirectly, or more interestingly, to DNA binding sites of small proteins (smORFs). These proteins, recently described in *M. pneumoniae*, were not included in our study. It has been shown that some of these smORFs can bind DNA *(*73*)*. Also, we may have missed binding sites of structural proteins like the ones forming the attachment organelle of *M. pneumoniae* or its cytoskeleton, as we did not explore all of these proteins systematically.

Also, to study possible DNA binding changes during the growth phase, different time points (exponential growth (6h) and stationary phases (96h)) were studied for some of the candidate TFs. We could detect a redistribution of the RNAP and RNAP associated proteins between 6h and 96h, with significant changes in the relative peak sizes (Figure 7.2). For instance, promoters with higher binding affinity for the RNAP at 6h, have higher RNA expression levels at 6h than at 96h (paired t-test, p-value=0.024). Similarly, promoters with higher binding affinity for the RNAP at 96h show higher expression levels at 96h (paired t-test, p-value=0.022) and they show a functional enrichment in COG category O, related to post-translational modification, protein turnover and chaperone functions (Fisher's enrichment test, p-value=0.017). Similar results were found analyzing other RNAP associated proteins such as σ^{70} and Spx.

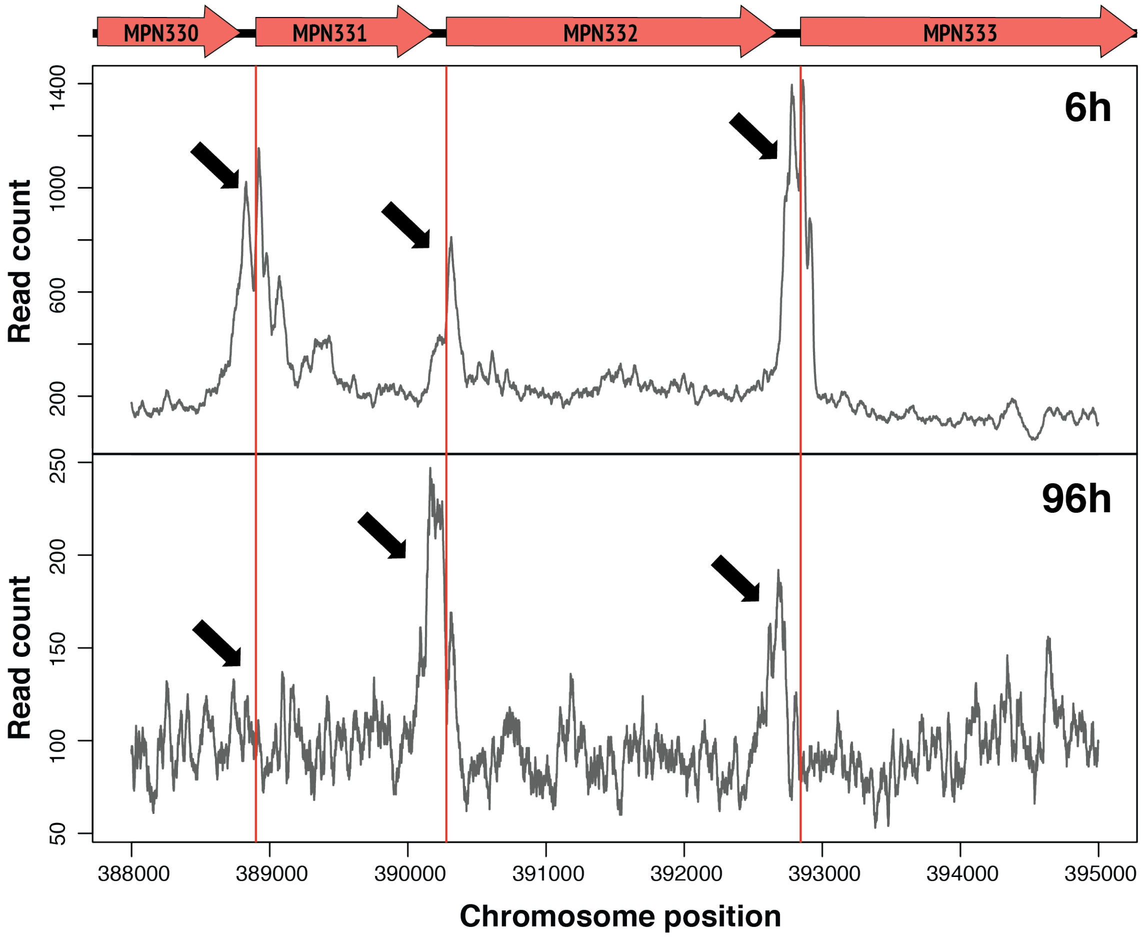


Figure 7.2. Changes in ChIP-seq of the RNA polymerase. Relative changes of promoter occupation of the alpha subunit of the RNA polymerase between 6h and 96h. For three consecutive genes, MPN331, MPN332 and MPN333, we can see polymerase occupationat their promoters at 6h. However, at 96h the promoter of MPN331 is no longer occupied by this protein, and the relative affinity for the promoters of MPN332 and MPN333 has changed, resulting in changes in the relative heights of the peaks.

Out of 194 ChIP-seq experiments of 123 candidates in different conditions (Table S5), we could detect binding for 83 experiments (42.78%)., as well as for the sigma factor and the RNAP polymerase. Out of these, 51 candidates did not show apparent specific binding with unique peaks (Table S6). The remaining 32 show at least some unique peaks, either associated to the RNA polymerase (putative TFs) or not (putative structural). For 9 experiments, the pattern was unclear between structural or TF-like. Regarding the structural proteins, in some cases, we detected few peaks and a rather specific binding to a broad region. An example of this is the Smc (Structural Maintenance of Chromosomes) protein which binds to the origin of replication, together with the complex formed by ScpA and ScpB (*(454)* and Figure 7.3).

For some proteins shown to bind DNA in other organisms and appearing as DNA-binding in our screening, like the leucine aminopeptidase (Leu, MPN572), we could not identify any specific chromosomal binding site, probably because they do not have a high sequence specificity (though we cannot exclude that the added tag is somehow affecting DNA binding, or that they bind DNA only under specific conditions). Other reasons why we could not find any DNA binding target for the remaining proteins are that either they may be false positives, they bind RNA or they need to form a complex to bind DNA.

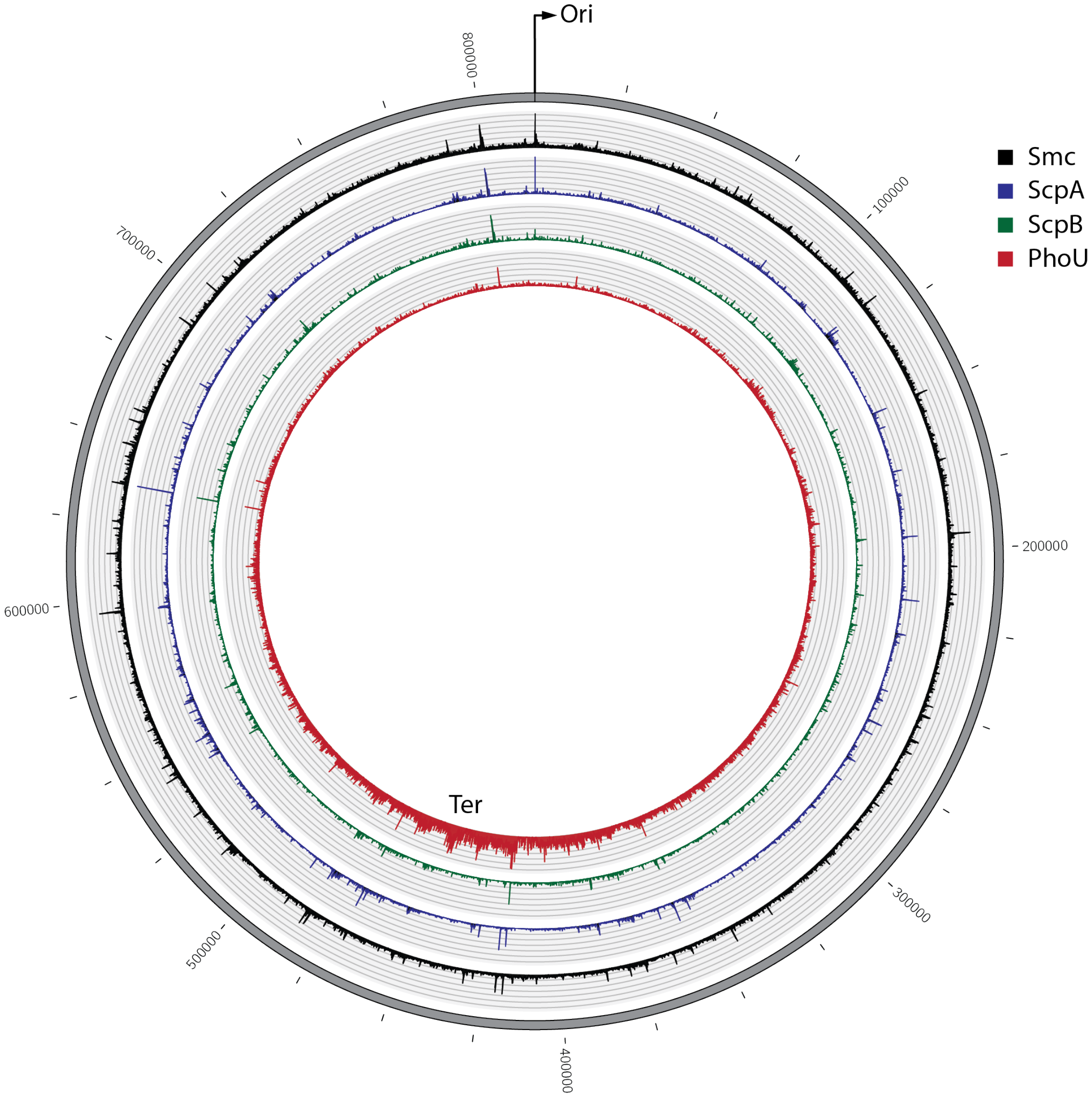


Figure 7.3. ChIP profile of different structural proteins from *M. pneumoniae*. Circos plot showing the ChIP-seq profile of 4 structural proteins. Smc, ScpA and ScpB have been shown to bind the Ori in other bacteria, something that is also observed in M. pneumoniae. Furthermore, we have found an enrichment for PhoU binding in the Ter region of the chromosome.

Regulatory roles of putative DNABPs

To study if physical interactions play a role in the regulation of gene expression, we analyzed the effect of the over-expression of these genes on transcription, and growth phenotype (Table S7), by means of microarrays and/or RNA-seq, and growth kinetics. In some cases, mass spectrometry was performed to study changes in the proteome to corroborate the results from transcriptomics at the protein level. It is important to note that in the overexpression experiments we are studying pools of transposon mutants to avoid the effect of the transposon insertion site. No preferential insertion sites were detected for any of the studied genes, and thus we do not expect any bias or artifact regarding the experimental setup. We confirmed the exogenous expression of the proteins by Western Blot, mass spectrometry and deep sequencing and/or microarrays. In most cases, we could see a good correlation between changes in RNA expression and protein levels of the candidate gene (Figure S2). However, there were some notable exceptions to this: for instance, for the gene encoding for the alternative sigma factor, MPN626 *(*255*,* 447*)*, the amount of protein was much smaller than the increase in RNA levels, corroborating its predicted toxicity. In total, we performed 196 gene expression experiments covering 136 genes, and we profiled their transcriptomes by microarrays or RNA-seq. The majority of these experiments focus in the overexpression of putative TFs. However, for some non-essential DNA-binding proteins *(*73*)*, transposon insertion mutants disrupting the gene of interest were isolated from the pools. Finally, for some essential putative TFs that did not show a transcription phenotype under the conditions tested, we introduced point mutations to create dominant negatives, or to constitutively activate their functionality. These mutations were designed using information from the literature, or by structural analysis of orthologs in other organisms (Figures S3-S9). For some predicted TFs, we could not detect specific ChIP-seq peaks or changes in the transcriptome when expressing N- and C-terminal flagged proteins. In those cases, we expressed the proteins without an epitope, which prevented doing ChIP-seq but could reveal their targets in transcriptomics experiments. This is the case for the known transcription factors GntR and the alternative sigma factor MPN626.

Out of the 196 experiments performed (Table S8), only 71 showed significant changes (36.22%). For some putative TFs for which we had several experiments run at the same time point, we calculated a consensus among all the available experiments, to obtain a unique result for each candidate at each growth stage (see Methods). In exponential growth (6h), we could detect significant expression changes in 51 experiments (out of 175), corresponding to the genetic perturbations of 37 candidates (Table 7-2). Of these, 25 experiments affected a reduced number of genes, which in some cases were also found as targets of the putative TF by ChIP-seq analysis. The remaining 26 experiments showed major perturbations (>20 genes significantly changing expression). For the remaining experiments we could not detect significant transcriptional changes in the conditions tested.

Table 7-2. Summary of the results of all transcriptomics experiments

Experiments	Growth phase	Changes
196 experiments	Exponential: 175 experiments (133 candidates)	25 changes
		26 major changes
		124 no changes
	Stationary: 21 experiments (13 candidates)	4 changes
		16 major changes
		1 no changes

Interestingly, we observed changes in 20 out of 21 experiments performed at stationary phase. In 7 of these experiments, we could see significant alterations at 24, 48 or 96h of growth but not in exponential growth (6h). However, in these cases it is difficult to determine which changes are specific and which ones are due to changes in growth rate with respect to the control. We assessed the phenotypic effects in growth rate by determining growth curves of the different strains and measuring their growth rates. In total, 220 growth curve experiments were performed, corresponding to genetic perturbations of 139 genes. 46 of them (20.9%) showed a differential growth phenotype, whilst the remaining 174 did not show changes respect to the controls. These percentages are similar to those found in the gene expression experiments (see above). This suggests that, despite its apparent simplicity, *M. pneumoniae* is a rather robust biological system. Indeed, it seems that overexpressing a variety of genes does not produce a detrimental metabolic load *(*455*)*, this only occurring in a minority of cases.

From data integration to network reconstruction

To reconstruct the gene regulatory network of *M. pneumoniae*, we combined our results from the ChIP-seq and the transcriptomics experiments. According to the results of these, we classified all the studied proteins in four different groups: TFs, those that show specific changes in both ChIP-seq and transcriptomics experiments, with common targets in both; regulators, those proteins that show specific changes in their transcription phenotype but no specific target in ChIP-seq; structural proteins, those proteins with specific peaks with a common motif in ChIP-seq, but no changes in transcription; and non-specific proteins, that do not show specific targets in any of the experiments. This classification was refined manually, and we added the category RNAP-like, to include those proteins forming part of the RNAP complex. The full classification can be found in Table S9.

By combining the information of the transcriptomics and the ChIP-seq experiments, we extracted motifs for each of the proteins classified as TFs. The motifs were used to curate the results, by discarding false positive targets without the motif, or adding new targets with the motif that did not pass the different filtering thresholds applied in the global analyses (see Methods). For the TFs DnaA (MPN686), Fur (MPN329) and HrcA (MPN686), we could extract a motif from the ChIP-seq data. The binding motifs for GntR (MPN239) and the alternative sigma factor (MPN626) were determined using their targets from the transcriptomics experiments. Finally, WhiA (MPN241) and MPN424 only have one target, corresponding to a distinct ribosomal operon regulated by each of them. Thus, we extracted a motif with a comparative analysis of the promoter region of the target operon of each TF in several bacterial species, using MEME *(*381*)*. For the transcriptional regulator Spx (MPN266), no motif was found in this analysis, as the ChIP-seq results showed that this regulator appears to bind the RNAP complex, as described in other bacteria *(*456*)*, and differential regulation may only occur with a change in conformation or oxidation state. Figure 7.4 shows the targets regulated by the different '*bona fide*' TFs after the analysis, and the motifs found for each of them.

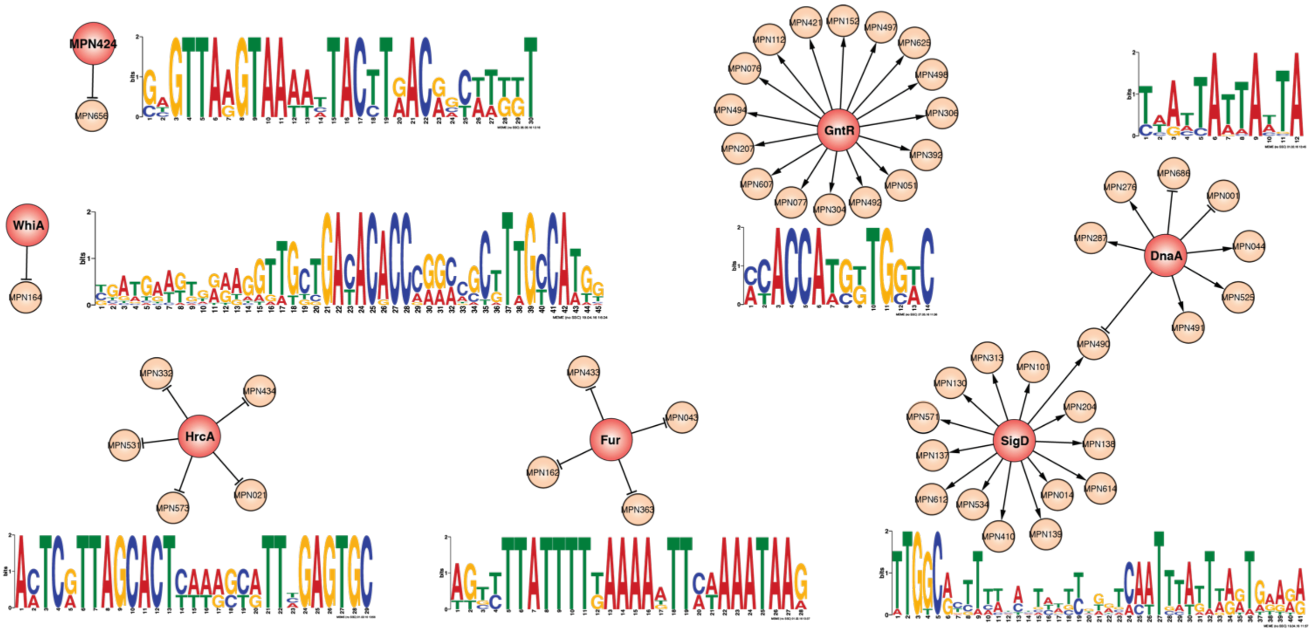


Figure 7.4. Targets and DNA-binding motifs for 7 out of the 9 proteins classified as transcription factors. Other known TFs, such as SigA as Spx, were classified also as RNAP-like as they bind the RNA polymerase complex and appear in the majority of promoters of M. pneumoniae. Arrows indicate activation, whilst T-shaped symbols indicate repression. Motifs were identified for all 7 TFs using the ChIP-seq and/or the transcriptomics targets and, in some cases, by comparative analysis with other bacterial species.

Regulators are an interesting group of proteins. For these, no specific DNA binding sites or motifs have been found by ChIP-seq, yet their overexpression or disruption by transposon insertion produces a clear phenotype in transcription. These proteins may act in signaling or metabolic pathways, and they should be linked either to a TF or to a different transcriptional regulatory mechanism, such as a riboswitch or the differential concentration of a metabolite. They could also function by binding TFs and sequestering them, preventing their binding to their target regions. We classified 30 proteins as regulators in *M. pneumoniae*.

Finally, among the *structural proteins* we found some that bind to a specific sequence motif rather than a broad genomic region. These proteins are DnaA (MPN686, also a TF in our analysis), MraZ (MPN314, which can also act as a TF; *(*457*)*) and a putative single strand binding protein (SSB, MPN554). For each of these, we identified their corresponding binding motif in our ChIP-seq analyses (Figure 7.5). Interestingly, for the DnaA protein we found peaks located in methylation-enriched regions in the chromosome of *M. pneumoniae*. We had previously hypothesized that these regions could act as checkpoints for DNA replication *(*41*)*. In our analysis we found other structural proteins, such as the condensin complex formed by the proteins Smc, ScpA and ScpB. For these we did not identify a specific binding motif, but rather found binding to broader regions of the chromosome, specifically the origin of replication. Another protein, PhoU (MPN608), showed preferential binding for the Ter region of the chromosome of *M. pneumoniae* (Figure 7.3). The histone-like protein Ihf (Integration host factor, MPN529) did not show many specific peaks but rather a non-uniform binding in broad regions of the chromosome of *M. pneumoniae*, as described in other bacteria. The only conserved peaks across the replicates of this TF showed a common TG-rich motif (Figure 7.5).

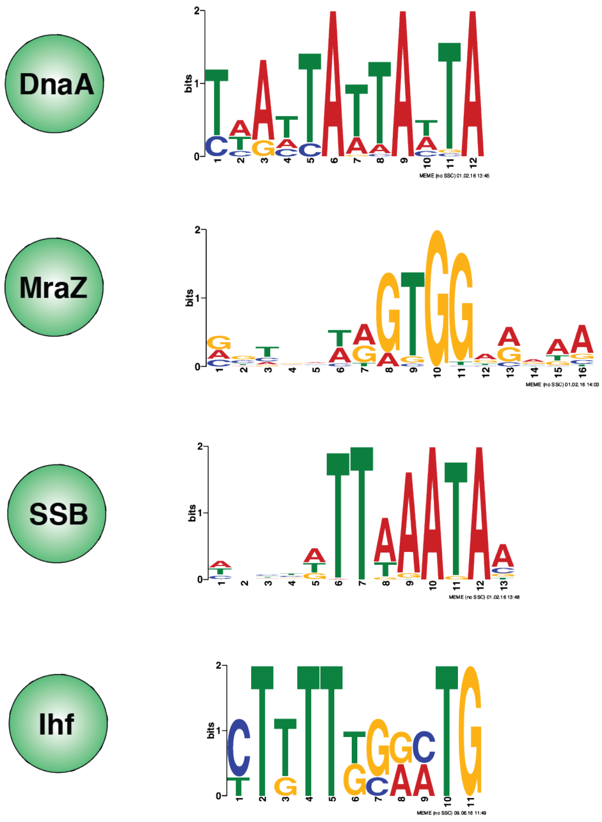


Figure 7.5. Structural DNABPs in *M. pneumoniae* with a defined ChIP-seq binding motif. A motif could be assigned to 4 out of the 8 structural proteins classified as such in *M. pneumoniae*: DnaA, MraZ, SSB and Ihf.

Taken altogether, the gene regulatory network formed by 7 TFs (excluding the two RNAP-like transcription factors SigA and Spx) is rather small and encompasses only 54 genes, which represent the 7.83% of the genome of *M. pneumoniae*. The targets of the TF Spx, which could be a major transcriptional regulator that binds the RNAP complex, have not been identified yet. Even so, the numbers contrast with those of other model organisms such as *E. coli*, with 208 TFs and over 3000 regulatory interactions *(*458*)*. To further investigate if other layers of transcriptional regulation exist, we performed a global correlation analysis of all the operons in *M. pneumoniae*, using 218 experiments of microarrays and RNA-seq (see Methods). Then, we constructed a network of operons, considering an edge in between any two operons with a global correlation higher than 0.5. We then clustered the network using the Girvan-Newman algorithm *(*459*)*, that identifies groups of nodes that are highly interconnected and is included in the GLay plugin for Cytoscape *(*460*)*. We identified meaningful clusters, as shown by the enrichment in different COG categories for many of them (Figure 7.6A). Interestingly, we could also identify correlations and anticorrelations between different clusters of the network. We find anticorrelations between clusters related to growth and those related to stress responses (for example, the cluster related to cell motility and that of amino acid metabolism have a negative correlation of -0.27). However, when superimposing the TFs to this network, the genes regulated are not always included in the same clusters (Figure 7.6B). This implies that there exists an underlying regulation layer in *M. pneumoniae*, not controlled by TFs, and that transcription factors (and regulators) act on top of this basal layers to regulate specific responses to certain stresses. Furthermore, not all operons are included in the network; there are operons that do not correlate with any of the clusters shown. This also points to the additional basal layers of regulation *(*23*)*.

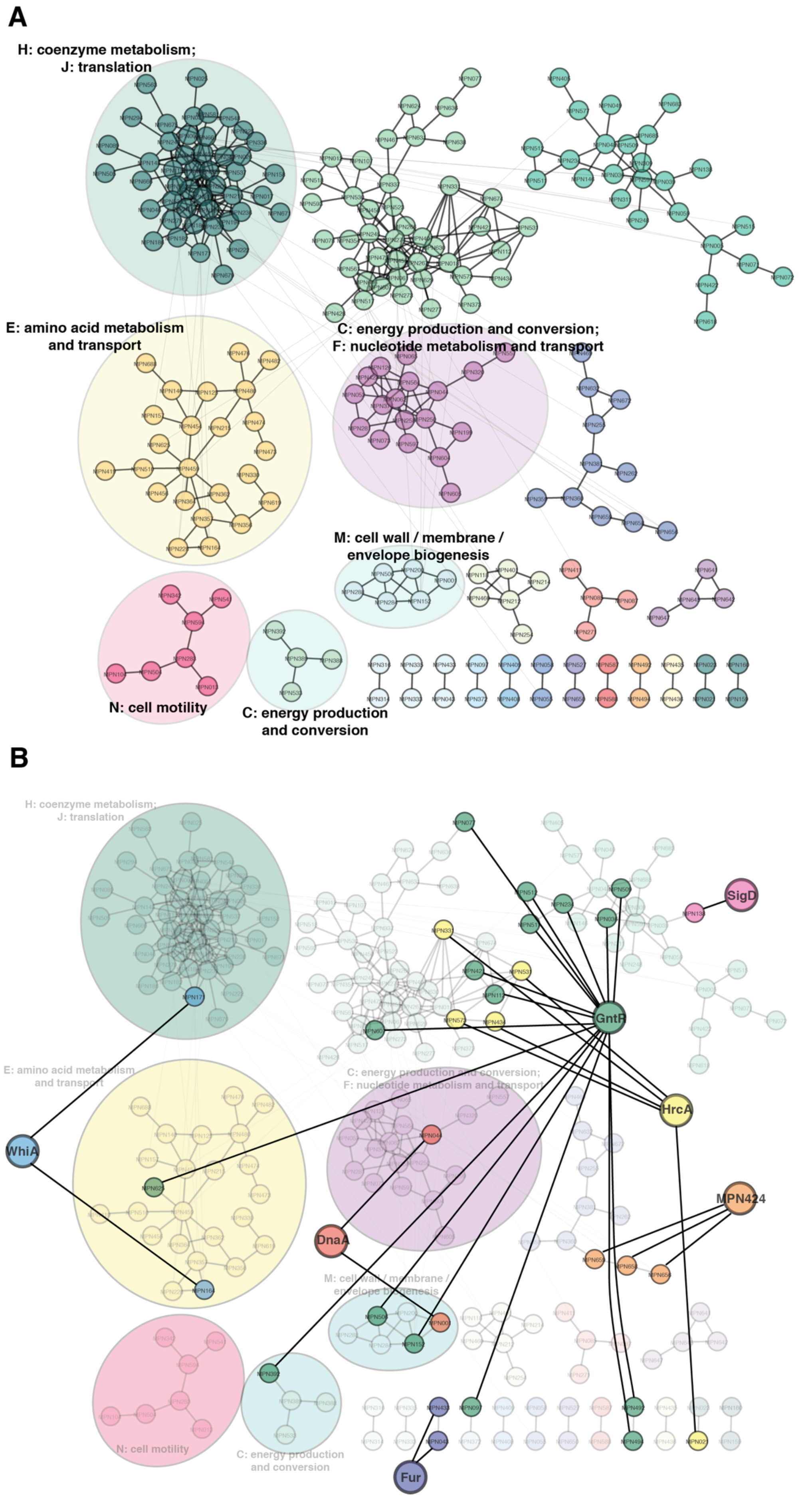

A
H: coenzyme metabolism;
J: translation
E: amino acid metabolism
and transport
C: energy production and conversion;
F: nucleotide metabolism and transport
M: cell wall / membrane /
envelope biogenesis
N: cell motility
C: energy production
and conversion
B
H: coenzyme metabolism;
J: translation
E: amino acid metabolism
and transport
C: energy production and conversion;
F: nucleotide metabolism and transport
M: cell wall / membrane /
envelope biogenesis
N: cell motility
C: energy production
and conversion
SigD
GntR
HrcA
WhiA
MPN424
DnaA
Fur

Figure 7.6. Network of operon-operon correlations. (A) Network of correlations across 218 transcriptomics experiments. Edges indicate correlations above 0.5. Clusters of highly interconnected operons are depicted with different node colors. Black edges indicate intra-cluster connections, whilst grey edges indicate inter-cluster connections. Significant enrichment for COG categories has been annotated for different clusters. (B) Overlay of TFs and their targets in the network. 7 TFs (all identified in this study except for the RNAP-like factors SigA and Spx) have been added with links to their targets. Targets for these TFs tend to be spread in different clusters.

Association of regulation to conditions: study of transcriptome changes in different perturbations

To gain more insight into the upstream effectors regulating TFs and regulators, we performed a number of perturbation experiments in *M. pneumoniae* cultures. These perturbations span a range of conditions that this bacterium can find in its natural niche, the respiratory tract epithelium, but also others that can happen in *in vitro* conditions, as well as various drugs affecting biological functions. In total, 111 experiments were performed that group into 42 different types of perturbations (Table S10). Some of these were performed both in wild-type *M. pneumoniae* and in strains with genetic perturbations (either overexpressing or having gene mutations). The results of the differential expression analysis of all these experiments can be found in Table S11.

Our first interest was to know how perturbations are related to each other, to assess if the response of *M. pneumoniae* to conditions not encountered in nature is similar to that of natural perturbations, and to “decompose” complex phenotypes like the one observed at the stationary growth phase. To do so, we created a bipartite graph of perturbations and operons, with edges connecting each perturbation and its regulated operons. Then we computed the perturbation-projection of the graph, to find the links between perturbations that share one or more co-regulated operons. We only considered operons changing in the same direction (either up- or down-regulated) to compute this graph. To further constrain the network, we applied two different filters: first, we removed those edges between perturbations with a correlation smaller than 0.3. Second, we applied an additional filter to remove those edges likely to be connected by chance (see Methods). Figure 7.7 shows the resulting perturbation network.

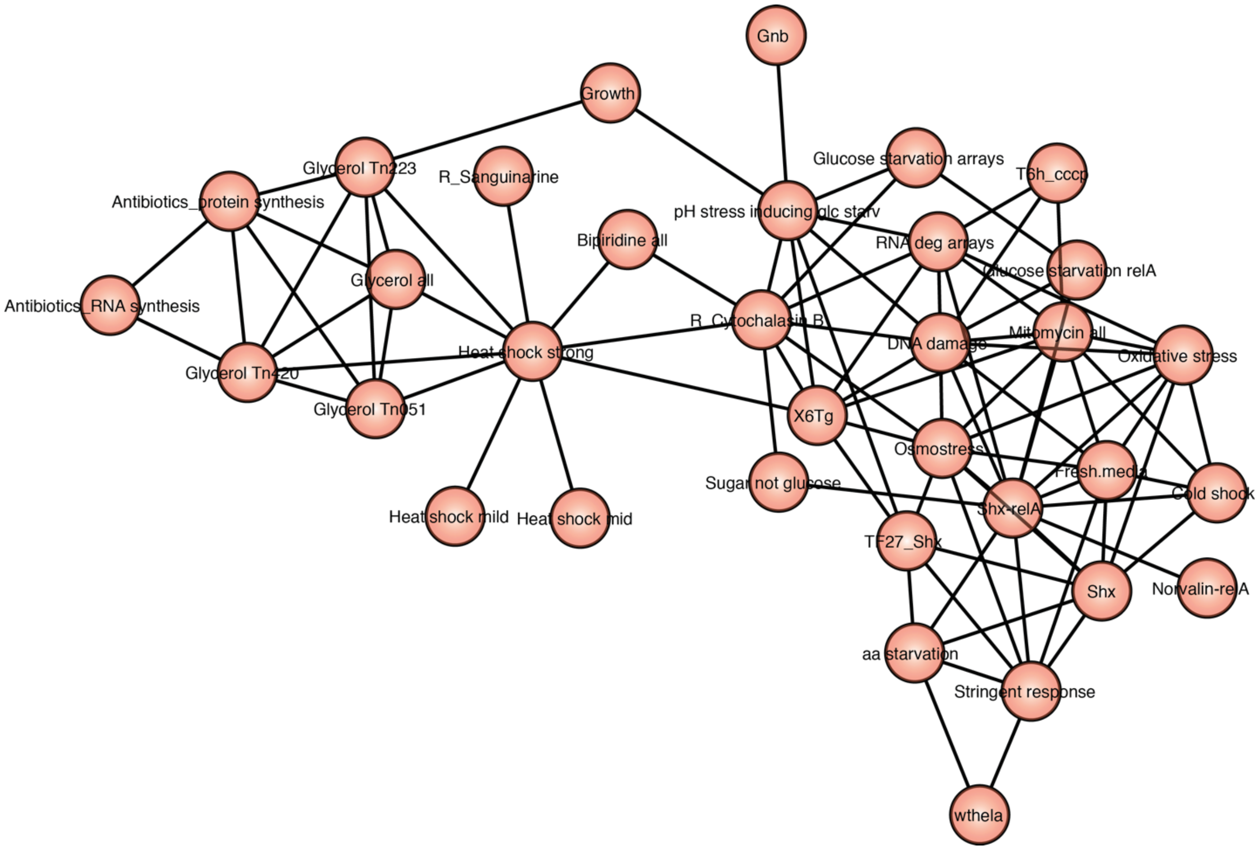


Figure 7.7. Network of perturbations. Edges are drawn between two perturbations if they share at least one gene in common, if the number of genes in common is hardly expected by chance (Probability<0.05) and if the expression changes between them are correlated (R>0.3).

As expected, different perturbations related to glycerol metabolism appear together in the network, as well as heat-shock perturbations, or conditions related to starvation and DNA damage. Interestingly, we found that experiments on infection of different types of cell cultures showed different behaviors. For instance, infection of HeLa cells results in a transcription phenotype that resembles that of amino acid starvation and stringent response. However, infection of erythrocytes presents a different phenotype, that does not correlate with any of the other conditions tested. This methodology can be generalized and used to incorporate new experiments and identify easily their transcription signatures.

Aside from identifying perturbations that are related to each other, we created a network of co-regulated operons in this set of experiments, and we assessed its similarity to the network created using only the genetic perturbations (putative TF overexpression and/or transposon insertions). To do so, we used the 111 initial experiments, and we discarded those experiments leading to global RNA degradation, such as glucose starvation and supercoiling inhibition by novobiocin. These experiments result in a general decrease of the total RNA and this effect may mask the underlying regulation. Indeed, a Principal Component Analysis (PCA) showed that the first principal component is dominated by experiments in which RNA degradation is involved, explaining 21.82% of the variance in the samples. After discarding these, we obtained a dataset of 98 experiments. Again, we found meaningful clusters, as represented by the enrichment for different COG categories, and the superimposition of the TFs does not explain the groups encountered.

We calculated the overlap of the two networks generated. Interestingly, they only share 132 nodes (out of 221 nodes in the overexpression and/or loss of function network and 231 nodes in the perturbation network). This suggests that, for a number of perturbations, the response observed is not mediated by transcription factors but by other mechanisms such as metabolite-based signalling, supercoiling, RNA degradation, and/or riboswitches. To further investigate this hypothesis, we combined the data from the different condition experiments with the data of genetic perturbations. To do so, we considered the 42 grouped conditions and the genetic perturbation experiments of the proteins classified as TFs or regulators (40 experiments, see Table S12). For each condition, we performed a regression analysis using random forests. We used the data on genetic perturbations to predict the expression levels of all the genes in that specific condition (see Methods). After running the random forest algorithm, we could estimate the percentage of the variance of the each condition that can be explained by TFs or regulators. Also, we can calculate the importance of each of these proteins for the prediction. Those proteins with higher importance will be directly linked to the transcription phenotype of that condition.

The variance explained by TFs and regulators in any of the conditions tested is rather limited. This may be caused by the large experimental noise observed in our transcriptomics datasets, but undoubtedly indicates that some determinants of transcription are missing in this analysis. To prove that experimental noise is partly responsible for these low percentages of variance explained, we simulated an experiment overexpressing the Fur transcription factor. This was done by using the results of a real experiment and adding experimental noise to each gene. Noise was generated by sampling a normal distribution with mean equal to zero and a standard deviation equal to that of each gene, for all the genetic perturbation experiments. 100 replicates were generated, which correlated well with the real experiment (R=0.754±0.018). Each of these replicates was merged with the entire genetic perturbations dataset, and we then used the random forest algorithm to predict the values of the real Fur experiment using the rest of the experiments (including the simulated one) as variables. We expected that a large percentage of the variance of the real Fur experiment could be explained, mostly by the simulated dataset. However, we found that although the simulated experiment holds the largest importance values, the variance explained is of 58.66±1.436%. In the dataset without the simulated experiment, this value is equal to 42.53%. This means that the addition of the simulated experiment helps explain the transcription phenotype of the Fur experiment, but noise still is responsible for an important percentage of the unexplained variance.

Considering that the experimental noise accounts for an important fraction of the variance, the largest percentage that could be explained corresponds to the growth condition, which compares wild type cells entering stationary growth phase (at 48h) with cells growing exponentially (6h). In this case, we could explain 49.89% of the variance only using the data from TFs and regulators. Interestingly, there are two experiments that have significantly higher importance in explaining this phenotype: the transposon insertion mutant of the lactate dehydrogenase (*ldh*, MPN674) gene, a *regulator*; and the overexpression of the GntR transcription factor (MPN239). It is known that during the transition from exponential to stationary phase in cultures of *M. pneumoniae*, the metabolism of this bacterium changes, transitioning from production of acetate to that of lactate. Thus, it is expected that the *ldh* gene plays a role in this transition. The TF GntR regulates genes related to lipid metabolism and lipoproteins, and could also be related to this phenotype.

An interesting example is that of *M. pneumoniae* cells growing on glycerol. In nature, *M. pneumoniae* cells live on the respiratory tract epithelium and rely on lipids such as phosphatidylcholine to sustain growth *(*461*)*, as *M. pneumoniae* has lost the metabolic pathways on lipids biosynthesis. Glycerol is essential for growth of *M. pneumoniae in vitro*, but in minimal concentrations *(*249*)*. An increase in the concentration of glycerol produces large transcriptional adjustments and deficiencies in growth (Figure S10). To study these changes, we exposed *M. pneumoniae* cells to concentrations of 0.1% and 1% of glycerol. We used both wild-type cells and strains deficient in genes known to be involved in glycerol metabolism: *Tn:mpn051* (GlpD, converts glycerol-3-P in DHAP and produces H_2O_2), *Tn:mpn223* (HprK, phosphorylates the protein Hpr at Ser-46, blocking glucose import) and *Tn:mpn420* (GlpQ, converts glycerophosphocholine into glycerol-3-phosphate and choline, *(*461*)*). Wild type cells incubated with glycerol showed the same behavior as two of the deletion strains, *Tn:mpn051* and *Tn:mpn223*. This indicates that the response to glycerol cannot be mediated by these proteins, or the metabolites they produce. However, the *Tn:mpn420* strain growing on high glycerol concentrations showed significant differences with respect to the wild-type. The random forest analysis showed that in this case, the most relevant TF whose overexpression partly reproduces the effect of glycerol is a mutant of the transcription factor Spx (MPN266). Spx can sense redox stress via two cysteines that, when oxidized, form a disulfide bond *(*462*)*. It binds the RNAP, and depending on the redox state of the cell, can change affinity for specific promoters *(*456*)*. The mutant of this experiment lacks one of these cysteines (C21S, equivalent to C10S in *B. subtilis*; *(*456*)*) and thus cannot form this disulfide bridge, preventing activation. Our findings suggest that in this case, glycerol metabolism, probably via the GlpQ protein (MPN420), would change the redox balance of the cell, affecting Spx and therefore transcription. In the *Tn:mpn420* strain, the redox balance is less affected and the phenotype resembles that of the Spx mutant. In fact, an analysis of the metabolic changes in cells exposed to glycerol showed similarities with the metabolome of cells exposed to oxidative stress. Nevertheless, in this case we can only explain 23.48% of the variance using the data on TFs and regulators, which suggest that other additional mechanisms participate in the glycerol response.

Taken altogether, these results suggest that the global regulation of gene expression in the minimal bacterium *M. pneumoniae* is not only controlled by transcription factors but other mechanisms largely contribute to gene expression coordination. Some TFs, such as Spx, have essential roles in the response to oxidative stress and could be the major regulators in *M. pneumoniae*, but the effect of other TFs would be added on top of other basal mechanisms to control the response of specific sets of genes under certain conditions.

Other factors that regulate transcription

Our data showed clearly that the sole contribution of transcription factors and other protein-based regulators is insufficient to explain the responses of *M. pneumoniae* observed in different conditions. Therefore, we studied the putative contribution to transcriptional regulation of mechanisms other than TFs.

Supercoiling: supercoiling may influence the accessibility of different promoters and thus modulate gene expression. Some promoters require a certain degree of supercoiling to be able to trigger transcription. To test the effect of supercoiling in regulating gene expression, we treated cultures of *M. pneumoniae* with increasing concentrations of the antibiotic novobiocin (see Methods). Novobiocin is an inhibitor of the DNA gyrase, and thus causes an imbalance between the DNA gyrases and topoisomerases present in the cells, which ultimately alters the chromosome supercoiling *(*463*)*. Upon this treatment, we observed that at high novobiocin concentrations, all RNAs present in *M. pneumoniae* decrease their concentration. This could be caused by an extreme change in the chromosome supercoiling, that provokes the release of the RNA polymerase, and prevents the formation of new initiation complexes, and arrests transcription, as occurring in eukaryotes *(*464*)*. This was confirmed by performing ChIP-seq of the RNAP in cells treated with the drug, as no ChIP-seq peaks were observed in this experiment (Figure S11).

At lower novobiocin concentrations, different behaviors coexist. The majority of the genes seem to be sensitive to the antibiotic, showing RNA decay at low concentrations, whilst others remain constant, or are upregulated at low novobiocin concentrations. A correlation analysis showed that we could distinguish up to 5 different global behaviors of genes in response to novobiocin (Figure 7.8A). By k-means clustering, we identified these 5 different behaviors and which genes are associated to each (Figure 7.8B-F). Interestingly, there is a group of genes that is upregulated at medium concentrations of novobiocin (Cluster 3). This cluster consists of 21 genes, and includes the operon of the DNA gyrase (MPN003-MPN004), confirming what has already been observed in other bacteria, that the gyrase operon is regulated by supercoiling *(*465*)*. The majority of these genes are located in the region proximal to the origin of replication. Interestingly, this cluster is also enriched in genes of the COG category L, implied in replication and repair (Fisher's enrichment test, p-value=9.47e-5). Supercoiling in this region is key to control the initiation of DNA replication in different bacteria, and this could therefore indicate a coupling between transcription and translation in *M. pneumoniae (*466*)*. Besides, Cluster 1, which shows a decay even at low concentrations is enriched in genes related to three COG categories (M, N and V) related to membrane proteins and virulence (Fisher's enrichment test, p-value < 0.05). In contrast, Cluster 4, which shows higher resistance to decay upon Novobiocin treatment, is enriched in genes related to the COG categories G and J, related to carbohydrate metabolism and translation (Fisher's enrichment test, p-value < 0.05). This points to supercoiling as a key regulator of transcription, that could be playing a role in the transition of gene expression from growth- to stress-related genes. The classification of all the genes in *M. pneumoniae* in the 5 clusters can be found in Table S13.

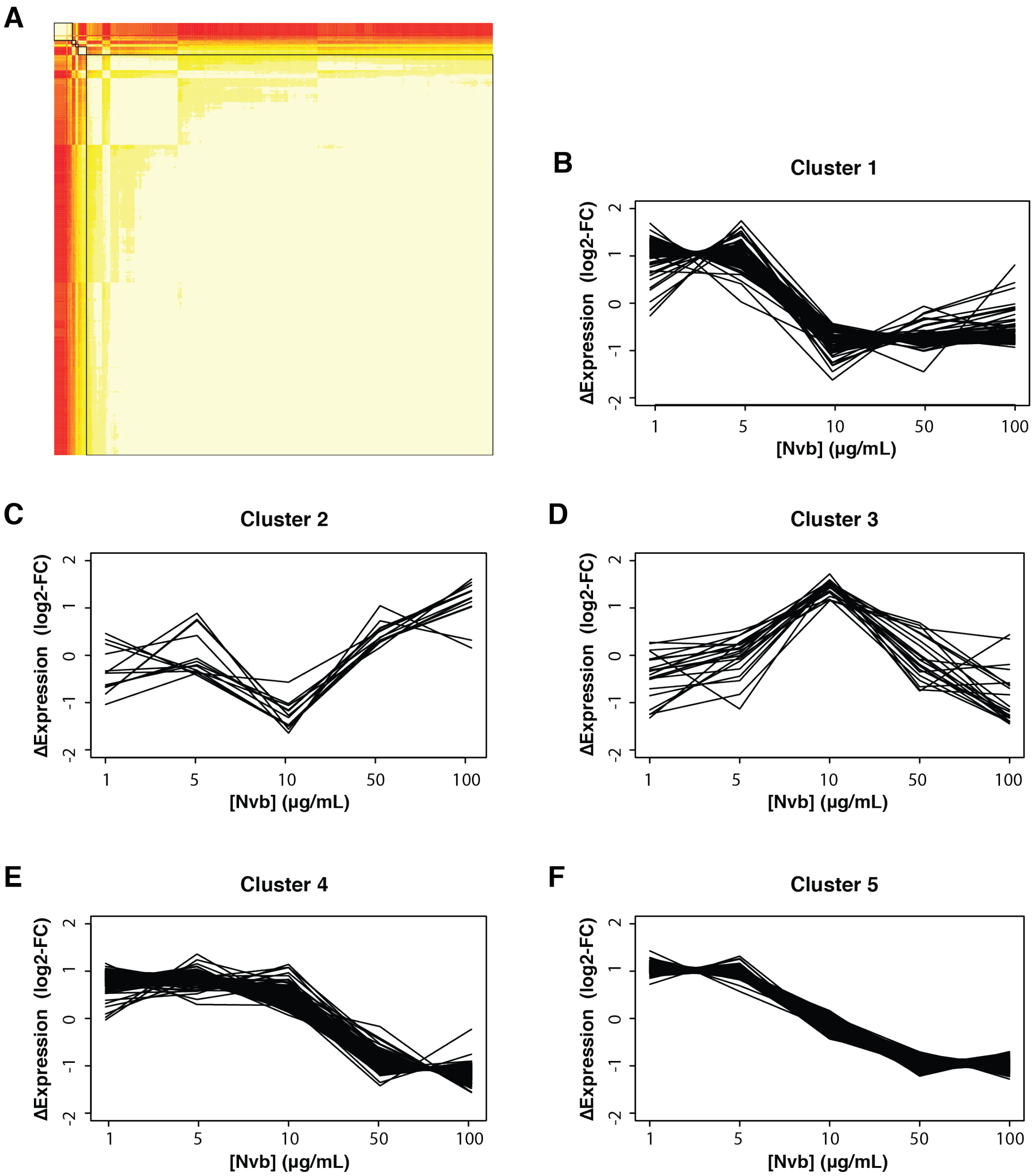


Figure 7.8. Novobiocin titration reveals different patterns of responses to supercoiling. (A) A correlation analysis revealed the existence of different groups of genes with distinct behaviors in response to increasing Novobiocin concentrations. (B-F) Behaviors of the 5 different clusters.

Riboswitches: riboswitches are untranslated regions of a nascent transcript that can fold differentially upon certain conditions, such as the binding of a metabolite *(*89, 90*)*. This differential folding can hide or expose transcription termination sites, or translation initiation regions, and thus affect those processes. Here, we focused on riboswitches regulating premature termination of transcription. To identify these riboswitches, we studied the profiles of RNA-seq experiments of *M. pneumoniae* in different conditions.

In total, 206 RNA-seq experiments on genetic or environmental perturbations were analyzed. For these experiments, we studied the region surrounding the annotated transcription start sites (TSS). We defined the region for the putative riboswitch as the first 150 bases after the TSS, coinciding with the average size of riboswitches described in bacteria *(*467*)*. For each TSS in each experiment, we compared the expression levels of the candidate riboswitch region with the expression of the following 300 bases (or with the rest of the gene in case it is shorter). We identified those cases in which the expression ratio riboswitch/gene was significantly higher than the average. Furthermore, we applied additional filters: first, the expression of the gene should be above a certain threshold, to avoid false positives due to the noise of the RNA-seq profiles. Second, the TSS should be active in the condition studied (i.e. the expression after the TSS should be significantly higher than the expression before). Third, each experiment was compared against its corresponding control. The ratio of expression riboswitch/gene should be different in the sample than in the control, to determine that in that specific condition the riboswitch was active.

With this analysis, we identified 36 TSSs with a putative riboswitch that was differentially expressed in at least one of the 206 conditions tested. Table S14 shows a list of all the riboswitches found in these analyses, and the conditions in which they are active. Figure 7.9A shows an example of a riboswitch that spans the 5'UTR of the *oppB* gene (MPN215), and changes upon overexpression of the *spoT* gene (MPN397), in glucose starvation. The SpoT protein produces and degrades ppGpp in *M. pneumoniae*, the metabolite responsible for the stringent response to amino acid starvation in different bacteria *(*81*,* 87*)*. The *oppBCDF* operon is involved in peptide import, deeming it reasonable to be regulated by a riboswitch in this condition. Indeed, this riboswitch shows a similar behavior in response to serine hydroxamate (Shx), a seryl-tRNA synthetase inhibitor, that causes stringent response *(*468*)*.

Other riboswitches are changing upon certain environmental perturbations. For example, a riboswitch located upstream the *atpB* gene, that encodes for a subunit of the F0F1 ATPase, changes its expression at high salt concentrations (Figure 7.9B). We used the Vienna RNAfold web server *(*469*)* to determine the secondary structures for these two riboswitches (Figure 7.9C-D) and their folding energies.

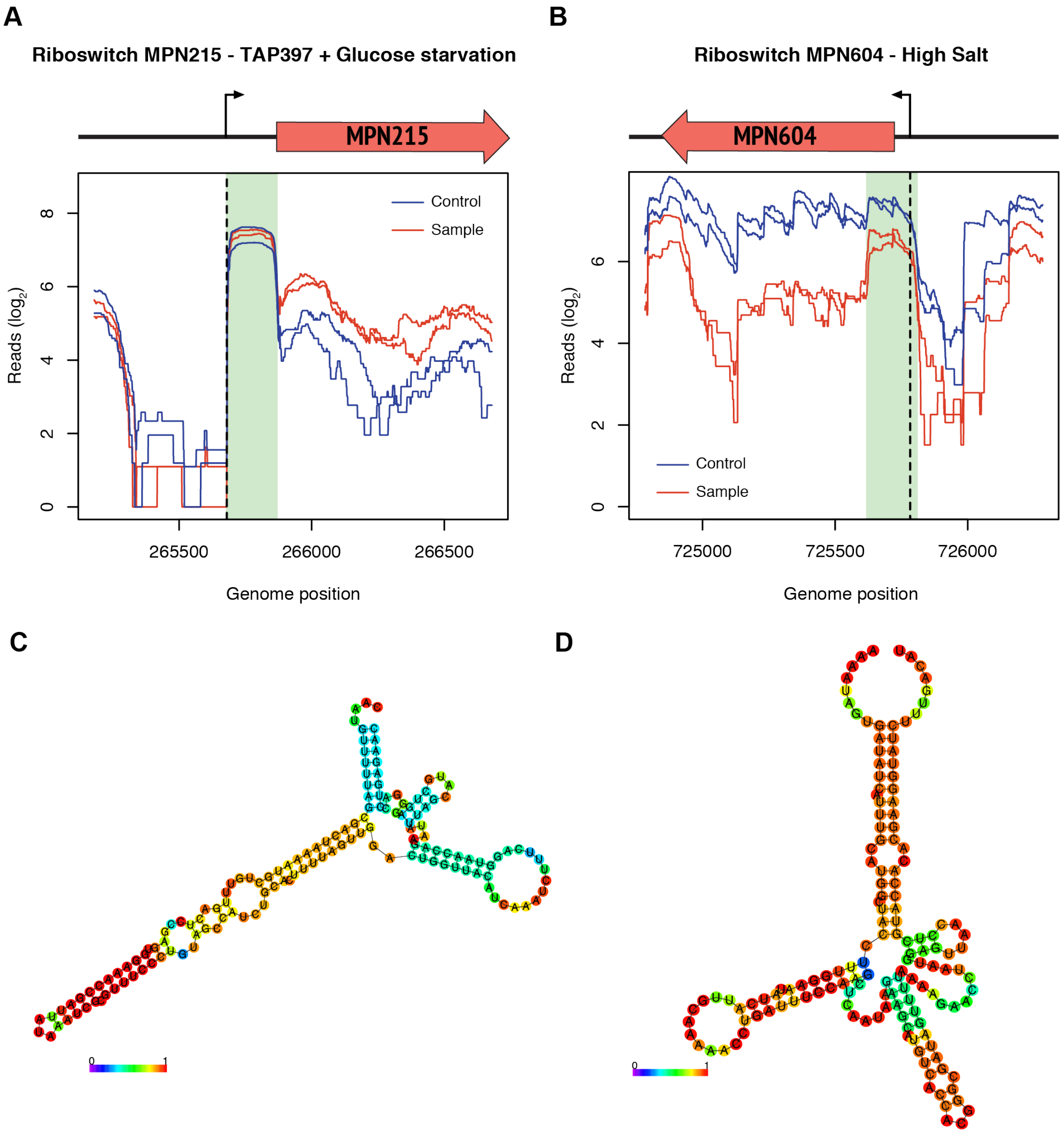


Figure 7.9. Riboswitches in *M. pneumoniae*. (A-B) RNA-seq profiles of two riboswitches in *M. pneumoniae*. Dashed lines indicate the transcription start site of the genes, whilst the green shaded area highlights the region of the riboswitch. Red lines represent the profile of the samples and blue lines represent the profiles of the controls (2 replicates each). (A) Riboswitch located upstream the *oppBCDF* operon. (B) Riboswitch located upstream the *atpB* gene. (B-C) Secondary structures of these two riboswitches in *M. pneumoniae*. Colors represent the probabilities of base pairing, from 0 to 1. These secondary structures were computed using the RNAfold web server. (C) Secondary structure of the riboswitch found upstream the *oppBCDF* operon, ΔG=-42.4 kcal/mol. (D) Secondary structure of the riboswitch located upstream the *atpB* gene, ΔG=-29.85 kcal/mol.

Transcriptional read-through: we have recently described that there exists a basal level of transcription coordination, that is mostly determined by structural properties of the genome organization, such as intergenic distances or presence of strong terminators *(*23*)*. We found that pairs of consecutive genes with an intergenic region smaller than

100 bps tend to have operon-like behaviors, co-transcribed in the same mRNA, and thus highly correlated. If the intergenic region is larger, the behavior of this pair of genes will depend on the structural properties of this region, and the trafficking of proteins (such as the RNAP) associated to it. According to this, we were able to classify pairs of genes in three groups: those with large probabilities of being co-transcribed (with an operon-like behavior), those that can be co-transcribed only in some conditions, and those that will never be transcribed in the same mRNA. Overall, there is an intrinsic stochasticity in transcription initiation, with the the RNAP able to initiate transcription at multiple entry points of an operon and to override termination signals with different probabilities. These observations support the idea that the traditional operon concept should be revisited, as transcription units are far more dynamical entities than it was thought before.

We became interested in the middle group, with pairs of genes co-transcribed only in certain conditions, as it represents a rather unexplored mechanism of gene expression regulation. We studied the transcriptional read-through (TRT), and in which conditions some termination signals could be overridden *(*23*)*. Interestingly, we found that during cold shock, global TRT is enhanced in *M. pneumoniae*. In other bacteria, this has also been described as a consequence of the anti-terminator function of cold shock-regulated proteins *(*102*,* 470*)*.

Regulation by metabolites and nucleotide concentrations: transcription regulation by different metabolites has been described in other bacteria. A classical example is that of (p)ppGpp, that is produced in response to amino acid starvation in bacteria *(*78*)*. However, the response to this metabolite is mediated by a different mechanism in Gram negative and Gram positive bacteria. In *E. coli*, (p)ppGpp binds to the RNAP and competes directly with GTP to occupy the +1 position of different RNAs, such as the rRNA *(*80*)*. In *B. subtilis*, however, (p)ppGpp does not compete for this position with the GTP. Instead, the production of (p)ppGpp uses GTP, decreasing its intracellular concentration. Thus, transcription initiation is stalled in those RNAs starting with this nucleotide *(*88*)*. Therefore, nucleotide abundances can also regulate RNA production. In *M. pneumoniae*, as in *B. subtilis*, nucleotide abundances change depending on the growth phase. In exponential growth, GTP is more abundant, whilst in stationary phase, levels of GTP decrease. A comparative analysis of the rRNA promoters in *B. subtilis*, *M. pneumoniae* and its close relative *M. genitalium* revealed that the +1 position is conserved among three species. Also, the +2 position is conserved between *M. pneumoniae* and *M. genitalium*. To investigate whether the concentration of GTP may affect RNA levels, we analyzed the first 2 positions of all the RNAs in *M. pneumoniae*. We did not find any functional enrichment considering only the first base, but when considering the +1 and +2 positions, we found interesting results. For the initial dinucleotide "GC" we found an enrichment in the COG category J, related to translation (Fisher's enrichment test, p-value=0.001). This finding fits with previous observations in other species such as *B. subtilis*, related to the response to amino acid starvation, affecting genes involved in the translation machinery *(*88*)*.

RNA degradation: finally, RNA homeostasis in the cell does not only depend on the regulation of transcript production, but also on the control of degradation. Indeed, we found that the first principal component of a PCA of the perturbation experiments was determined by degradation of RNA (see above). This is a highly regulated process, with a dedicated protein complex, the degradosome, whose composition may vary, changing its specificity for different transcripts *(*114*)*. To study half lives of RNAs in *M. pneumoniae,* we used the antibiotic novobiocin, as we observed that at high concentrations, it produces a change in the supercoiling that releases the RNAP from the chromosome and therefore stops transcription. We treated *M. pneumoniae* cells with novobiocin, using the highest concentration from our previous titration experiments. At different time points (0, control; 2; 4; 6; 8; 10; and 15 minutes) of treatment, we extracted total RNA from these samples and performed RNA-seq. After calculating gene expression in each of the samples, we determined the concentration of RNA for each gene and fitted the RNA degradation to an exponential decay curve (see Methods). With this, we could determine the half life for the majority of RNAs in *M. pneumoniae*, after discarding the cases with a suboptimal fit (adjusted R-squared < 0.6). The median half life of mRNAs in *M. pneumoniae* is around 5-6 minutes. We determined the RNA half lives in two growth stages: exponential and stationary phase. There was a general agreement in between both growth phases, but there were some notable differences: we identified a group of 25 genes with significantly larger half lives in exponential growth (5+ minutes longer in exponential than in stationary phase). Among these, we find 3 glycolytic enzymes (transketolase, MPN082; phosphofructokinase, MPN302; and pyruvate kinase, MPN303), and 6 genes related to translation (ribosomal proteins, tRNA synthases, etc.). Noteworthy, the phosphofructokinase (PFK) has been identified as one of the components of the degradosome in Gram positive bacteria, such as *B. subtilis (*119*)*. Therefore, it could be involved in a regulatory feedback loop, controlling the degradation of its own mRNA and the mRNA of other glycolytic enzymes, as previously described for the enolase *(*114*)*.

7.4. Discussion

Transcription, as the first step involved in gene expression, is a highly regulated process. RNA levels must be tightly controlled and maintained in homeostasis, and should be able to respond rapidly to external perturbations.

In model bacteria such as *E. coli*, more than 200 TFs have been described, together with 1999 regulatory interactions with strong evidence (from a total of 3231 total regulatory interactions) *(*458*)*. Indeed, it has been stated that only 7 global regulators in *E. coli* account for the expression control of 51% of the genes in its genome *(*52*)*. With such a large complexity in this layer, this has been the focus of numerous studies on transcription *(*471*,* 472*)*, with less attention being paid to other regulatory layers. However, other mechanisms have been described that can regulate RNA levels in specific situations. For instance, riboswitches can act as sensors of different metabolites, regulating translation from specific RNAs or premature termination of transcription *(*89*)*. Also, moderate changes in the supercoiling may expose or mask promoters, altering gene expression. This has been shown to depend on the spacing between the -35 and -10 promoter motifs in bacteria such as *E. coli (*465*)*. Furthermore, recent studies have shed light on the importance of the dynamics of transcription units *(*473*)* and also on the role of small RNAs (sRNAs) in regulating RNA levels *(*393*)*. Besides, regulation of transcript levels is not only determined by differential production, but also by their degradation rates, which are also regulated. The RNA degradosome composition can change depending on the growth phase or the external conditions, and thus it can change the specificity of the mRNAs being degraded to optimize translational resources *(*119*)*.

Here, we have studied several mechanisms involved in the regulation of this process in the minimal bacterium *M. pneumoniae*. For this bacterium, only 8 proteins had been previously annotated as putative transcription factors *(*72*,* 249*)*, yet it displays complex transcriptional phenotypes in response to different perturbations. To study all the possible gene expression regulators in an unbiased manner, we used different methods to capture all the DNABPs in *M. pneumoniae*. We recovered 105 proteins as DNA-binding candidates, and we curated this list manually by discarding some proteins involved in DNA replication and adding others reported in previous studies of transcription regulation. To characterize all these candidates, we overexpressed each of them in *M. pneumoniae* cells. Some of them were also knocked-out via transposon insertion, and for others, we overexpressed mutants acting as dominant negatives to study their function. The strains generated were characterized by assessing their growth phenotype and their transcriptome. Additionally, for the overexpressed candidates, we determined the DNA binding targets by ChIP-seq.

After generating all the different *M. pneumoniae* strains, we observed that most of them did not show a growth or transcription phenotype, rendering this organism very robust to genetic perturbations. A reduced gene regulatory network, with 9 TFs and 30 regulators identified in this study, and a streamlined metabolism, with few branching points, could be the reason of this robustness. These factors would limit the unwanted cross-talks occurring when overexpressing different genes. Also, the fact that the relative overexpression achieved is never very high, suggests that the small metabolic load imposed by these genetic perturbations is not detrimental to the cell.

Regarding ChIP-seq, one of the most striking results was that many of the candidates did not present specific ChIP-seq peaks, or did not present peaks at all (110 out of 194 experiments, 56.70%), whilst others present only a few of them with no common motif or apparent relationship. There is a variety of reasons why this can happen. First, it is possible that in our candidate identification we recovered some false positives, proteins that actually do not bind DNA. Also, some of these proteins may bind DNA but only in an indirect manner, through a protein complex. The addition of the tag in the overexpression of the protein may affect the formation of these complexes or even the direct binding of the candidate. Furthermore, proteins with disordered regions tend to be more sticky and bind other proteins in a non-specific manner. Overexpressing these proteins may lead to the formation of complexes with DNA-binding proteins and the recovery of false positive peaks. Indeed, we have observed the presence of 'Phantom Peaks' *(*187*)* in our experiments, coinciding with promoters of highly expressed genes. These phantom peaks can be the result of the overexpression of these disordered proteins binding the RNAP or the consequence of a low specificity of the antibodies used for the immunoprecipitation. All these effects, alone or combined, can mask the potential DNA-protein interactions, and require a custom-designed analysis to identify them.

In this work, we identified 8 potential structural DNABPs, some of them binding to specific motifs, others to broader regions of the chromosome. We also identified several potential transcription factors, which were later corroborated or discarded using the information from the transcriptomics experiments. 196 genetic perturbation experiments were used to identify transcription factors and regulators. In combination with the ChIP-seq experiments, we identified 9 TFs and 30 regulators. TFs were, in the majority of cases, validated by the ChIP-seq data. Among the TFs, we can distinguish at least one master regulator, Spx (MPN266). This master regulator would be implied in the response to glycerol, and linked to one of the regulators, GlpQ (MPN420). This regulator would be causing changes in the intracellular redox balance, which would be sensed by Spx via the formation of a disulfide bond. This conformational change in Spx would trigger its function as an activator, and would be regulating a large number of genes in *M. pneumoniae*. To verify our results about this TF, which are still preliminary, further data is required, including transcriptomics of cells expressing the dominant negative mutant of Spx in presence of glycerol. If our hypothesis holds true, we expect a phenotype similar to that of the *tn:mpn420* in presence of glycerol. Other interesting TFs are WhiA (MPN241) and MPN424. We have only found one target for each of them. These two targets correspond to different ribosomal operons in the genome of *M. pneumoniae*. It is interesting that a bacterium with such a reduced genome and a simplified regulation maintains two different TFs to regulate two different ribosomal operons.

Regulators are an interesting group of proteins. They exert an effect in the RNA levels of many genes, but in an indirect manner, as they do not bind DNA directly. Different mechanisms of action can be considered for these regulators. First, they can regulate gene expression via post-translational modifications. In *M. pneumoniae*, we have observed several transcriptional changes upon overexpression and/or knock-out of the protein kinase PrkC (MPN248) and the phosphatase PrpC (MPN247). These proteins control the balance of phosphorylation in the proteome of this bacterium, which has been shown to be key for the stability of these proteins *(*198*)*. Altering the abundances and stability of proteins can affect the effectors of transcriptional regulatory circuits and ultimately affect RNA levels too. Second, they can affect RNA stability and structure rather than RNA production. In *M. pneumoniae*, we have observed drastic changes of RNA levels in the knock-out of the gene *mpn545*, which encodes for an RNase III. This RNase is involved in the processing of ribosomal RNA and some mRNAs. Third, they can exert their effect by changing the metabolic state of the cell. For instance, we have observed that the activity of the protein GlpQ (MPN420) in presence of glycerol may induce some oxidative stress, inducing a conformational change in the TF Spx (MPN266), ultimately regulating the expression of a number of genes. Also, the SpoT (MPN397) protein acts as a regulator, as the production of ppGpp in *M. pneumoniae* decreases the GTP concentration, affecting the transcription initiation rates of many genes. Finally, regulators may form complexes with TFs that remain elusive because of experimental constraints such as the usage of a tag.

After defining the role of our candidate DNABPs as TFs, regulators or structural proteins, we drafted the gene regulatory network of *M. pneumoniae.* To date, this network comprises 7 TFs and 54 genes in total (7.83% of the genome of this bacterium). This is opposed to the GRN of other model organisms such as *E. coli*, which accounts for the majority of genes in this bacterium *(*458*)*. This observation suggests that some TFs may have been overlooked, or that other mechanisms controlling transcription may have a role that is more important than previously thought. Furthermore, after analyzing the transcriptome of *M. pneumoniae* cells under 111 different environmental perturbations, we could not explain more than 50% of the variance of any of these experiments by using only TFs and regulators. It is possible that this percentage has been underestimated because of the large experimental noise, and the actual variance explained is expected to be higher. Nevertheless, these results point to additional systems governing control of RNA abundances in *M. pneumoniae*.

Part of this non-assigned variance could be explained by other mechanisms as metabolites and riboswitches, supercoiling, and RNA degradation. We have described specific examples of each of these mechanisms in *M. pneumoniae*. However, further research is needed to understand to which extent is their contribution central to transcriptional regulation. One of the challenges to solve corresponds to the problem of parameterization of these variables. Experiments of transcriptomics with an expression value for each gene can be incorporated to our predictors in a straightforward manner. However, it is complicated to parameterize other factors involved in transcription such as riboswitches, which can be present or absent, and their behavior, which can be active or inactive in specific conditions. The same problem applies for the condition-dependent RNA degradation, transcriptional read-through or supercoiling. Nevertheless, our preliminary results show that supercoiling could be one of the major regulators of transcription, with genes from different categories and functions showing distinct behaviors in response to various degrees of supercoiling. In contrast, riboswitches would have a smaller role, as they would be delimited to specific conditions, and RNA degradation would be dependent on the degradosome composition, changing according to the growth phase or in some perturbations. Other studies within our group have also assessed the role of genome organization in operons, and how they are affected by transcriptional read through *(*23*)* and the role of small RNAs (*(*431*)*, see Chapter 6).

Here, we have shown that the control of RNA abundances in *M. pneumoniae* is not determined by a single regulatory layer, but that multiple processes in the cell intervene and there is feedback occurring among them. This suggests that genome-scale or whole-cell models will be required to integrate all these layers to fully understand transcription and to make reliable predictions of RNA levels upon different conditions.

7.5. Materials and Methods

Bacterial strains and cell cultures

M. pneumoniae M129 strain (passage 33-34) was grown in modified Hayflick medium and transformed by electroporation with pMT85 transposon as previously described *(*249*)*. The cell lines used in this study are detailed in Table S15. In general, proteins were flag-tagged (DYKDDDKG) in their N- or C-terminus, and expression was confirmed by Western Blot with M2 monoclonal anti-flag (Sigma). In some cases, when the tag was foreseen to interfere with the protein function, they were expressed without. In general, promoter from the *tuf* gene (MPN665) was used for overexpression, unless otherwise indicated (in the cases in which the protein was toxic, the endogenous promoter was used instead). Dominant negative mutants or deletions were done in some cases, as detailed in Table S15. In few cases, TAP-tagged clones from Anne-Claude Gavin's collection (EMBL) were used *(*203*)*. Transposon insertion mutants were obtained by haystack mutagenesis *(*474*)*.

DNA affinity column

M. pneumoniae cells were diluted 1:10 in Hayflick and grown for 3 days at 37ºC in a 300 cm^2 flask. Cells were washed twice with ice-cold phosphate buffer saline (PBS), collected in 5 ml of lysis buffer (50 mM Tris·HCl, 1 M NaCl, 1 mM $CaCl_2$, 1 mM EDTA, 0.1% Triton X-100, 1 mM DTT, pH 8), and supplemented with a protease inhibitors cocktail (Roche). High salt was used to release the proteins from the DNA. Cell extracts were centrifuged during 30 min at 100.000xg and 4ºC (Beckman ultracentrifuge) and the soluble fraction was diluted 10 times with 50 mM Tris·HCl, 1 mM $CaCl_2$, 1 mM EDTA, pH 8 (to dilute out salt and detergent). A DNA-Cellulose column was compacted and assembled (2 g, Sigma) and run in Äkta Xpress (GE Healthcare) in equilibration buffer (50 mM Tris·HCl, 0.1 M NaCl, 1 mM $CaCl_2$, 1 mM

EDTA, pH 8), before binding of the cleared cell lysate. After washing thoroughly with equilibration buffer, and was buffer (equilibration buffer plus 200 mM NaCl) nucleic acid binding proteins were eluted with 1 M NaCl in equilibration buffer, or 5 mg/ml yeast ribonucleic acid in TE (in order to elute proteins with affinity for RNA), and concentrated by TCA precipitation before submission for mass-spec identification (see below). A cellulose resin was used as a negative control for unspecific binding.

Chromatin isolation

We assessed DNA binding properties by ultracentrifugation employing a sucrose cushion following a previously described method *(*475*)* with modifications. Briefly, a 300 cm^2 flask was grow for 3 days and washed with PBS and lysis was performed using 2 ml of lysis buffer (10 mM Tris·HCl, 1 mM EDTA, 1% Nonidet P-40, pH 8 plus protease inhibitor cocktail from Roche). In order to follow the chromatin, 0.2 µl of Sybersafe (Invitrogen) was added and 1 ml of lysate was loaded on top of a of 20%-40% sucrose cushion (in TE: 10 mM Tris·HCl, 1 mM EDTA). Chromatin was fractionated by ultracentrifugation in a Ti45 rotor (Beckman) at 30000 rpm and 4ºC for 18 hours and collected from the interphase with the help of a UV light. After pelleting it by centrifugation at 100000g for 1h, supernatant was discarded and the pellet was resuspended in digestion buffer (50 mM Tris·HCl, 0.3 M NaCl, 1 mM $MgCl_2$, pH 7.5) plus 8U DNAse I (Ambion), for 1 hour at room temperature to release the DNABPs. After spinning down for 30 minutes at 14000 rpm at 4ºC in a table-top centrifuge, supernatant and pellet were analysed by SDS-electrophoresis.

Chromatin immunoprecipitation

Adapted from Buratowski's lab *(*476*)*. From a pre-culture, *M. pneumoniae* cells were split 1:10 in a 300 cm2-flask and grown for 4 days at 37ºC. When indicated (Table S7) cells were collected at this point (stationary phase), or they were scrapped and seed in 40 ml fresh Hayflick in a 150 cm^2 flask and incubated 6 hours more at 37ºC (exponential phase). Formaldehyde was added to 1% final (16% stock, Pierce) incubated for 10 min at room temperature and quenched by adding glycine to 100 mM, for 5 min at RT. Cells were washed twice with ice-cold PBS, scraped in 5 ml PBS and spun 5 min at 4ºC at 8000 rpm in a table-top centrifuge. The pellet was lysed by adding 1 ml of FA lysis buffer (50 mM Hepes·KOH, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, pH 7.5) plus a protease inhibitor cocktail

(Pierce) at 4ºC for 5 min. Chromatin was sheared by ultrasonication with Covaris (settings: Duty Cycle: 20%, Intensity: 5, Cycles/Burst: 200, Time: 15 min, Water level: 15) to ~200 base pair fragments and cell debris were removed by centrifugation at 16000g 10 min at 4ºC. Supernatant NaCl concentration was adjusted to 275 mM. 50 µl beads were preblocked with 0.5% bovine serum albumin (BSA) in PBS for 15 min at room temperature. Sepharose-protein-G was bound to either 10 µl of 1 mg/ml mouse IgG (control, Sigma), or 10 µl anti-Flag (M2 monoclonal) for Flag-tagged proteins. In the case of TAP-tagged proteins, 50 µl sepharose-IgG were used, and no control was included. Circa 0.5-1 mg chromatin per reaction was added and incubated over night at 4 ºC. The following washes were done: 1x of FA wash buffer 1 (FA lysis buffer with 275 mM NaCl), 1x with FA wash buffer 2 (FA lysis buffer with 500 mM NaCl), FA wash buffer 3 (10 mM Tris·HCl, 250 mM LiCl, 1 mM EDTA, 0.5% Nonidet P-40, 0.5% sodium deoxycholate, pH 8) and finally TE. Then, elute IPed material was extracted with 250 µl of FA elution buffer (50 mM Tris·HCl, 1% SDS, 10 mM EDTA, pH 7.5) and incubated 10 min at 65ºC. The beads were added to a micro-spin column (Bio-Rad) in order to collect the beads death volume by centrifugation. Then, 5 µl of 20 mg/ml proteinase K was added to elute the samples, tubes were incubated 15 min at 55 ºC and 10 min at 95 ºC before cooling at room temperature. To purify and extract the DNA, phenol/chloroform extraction protocol and ethanol precipitation was performed. Precipitated IPed DNA was resuspended in 10 mM Tris·HCl, pH 8 and measured in Qubit (High sensitivity kit, Invitrogen). At least 8 ng material was submitted for DNA ultra sequencing and a standard Illumina ChIP Sample Preparation protocol was used.

ChIP-seq analysis

Two curves were obtained for each ChIP-seq experiment, corresponding to the reads mapped to the plus and minus strand of the *M. pneumoniae* chromosome. Additionally, each experiment (IP) is accompanied by a control experiment (IgG), in which only the secondary antibody was used for the immunoprecipitation. For each of the experimental curves, we normalized the profile of the IP using the signal from the corresponding IgG to equate their baseline levels. After normalization, the control signal was subtracted from the experiment profile. With the resulting profile, noise was modeled to fit a Gaussian distribution centered on zero and with a standard deviation varying across experiments. To determine whether the experimental noise actually fits a Gaussian distribution, we obtained the ChIP-seq profiles of the wild-type *M. pneumoniae* strain, without the overexpression of any protein. After normalizing and subtracting the control signal, a Kolmogorov-Smirnov test for normality yielded a p-value of 0.5865. Therefore, with this data we fail to reject the null hypothesis of noise following a Gaussian distribution in our ChIP-seq data. According to this distribution, a threshold was set to reject all values with a probability larger than 1e-6 being noise.

The peak calling was done separately in each of the profiles for the plus and the minus strand, and was performed with the Matlab 'findpeaks' function. A custom R implementation of this function was used for our analyses. The parameters used in the peak calling were the following: slope threshold=0.0001 (minimum peak slope); amplitude threshold=5 (minimum peak width); smoothing width=15 (number of points to consider when smoothing the curve); and peak group=15 (number of points used to fit the peak). Additional filters were used to discard false positive peaks. We discarded those peaks that were detected in the IP-IgG curve, but were not found in the IP curve. Furthermore, we discarded those peaks in which the ratio between the IP-IgG peak and the IgG peak was smaller than 2, as those were likely to arise from experimental noise.

After the peak calling in each of the strands, the data from both of them was merged. With ChIP-seq, it is expected to find the same peaks in both strands. However, their positions tend to be displaced one from another and usually do not overlap. This is due to the fact that the DNA fragments whose ends are sequenced are usually larger than the sequencing read length. Thus, we associated each peak from the plus strand to a peak in the minus strand, provided that the distance between them was smaller than 300 bps. The actual peak position was inputted to the mid-point between the two partner peaks. The average distance between these partner peaks was calculated for each experiment. Single peaks with no partners in the opposite strand were discarded. Finally, a score was assigned to each of the peaks, describing how well the pair matches this average distance.

To analyze and classify the candidates according to their ChIP-seq profiles, the samples comprising the different subunits of RNA polymerase, sigA and mpn266 were selected as references for studying peaks that could be associated with RNA polymerase and thus to transcription factors.

Peaks located closer than 200bps of a RNAP-associated peak (encountered in one of the reference samples) were considered peaks associated to the RNA polymerase. The remaining peaks were considered as not associated to RNA polymerase transcribing events. Unique peaks were those only present in a given sample, without any peak located in the same region in the remaining experiments. We classified both RNAP and non-RNAP peaks as unique or non-unique. The numbers of total and unique peaks were estimated.

To classify the different experiments, the percentages of peaks from each category (RNAP, non-RNAP, unique and non-unique) were calculated for each sample. We also determined the specificity of the binding according to the percentage of unique peaks found in each sample (> 50%, HS=high specificity; <50% LS=low specific).

To classify eachprotein, we used the percentages of unique RNAP and unique non-RNAP peaks. In proteins regarded as transcription factors (TF) the percentage of unique RNAP peaks is equal to 100. If the percentage of non-unique RNAP peaks is 100 then the categrory assigned is structural (S). In between these two possibilities, proteins were classified as both putative structural and TF, indicating the category with the highest percentage of peaks (Table S6)

Gene expression analysis

M. pneumoniae cells on various stages of growth, overexpressing different regulators or being exposed to various perturbations (see Table S7 and Table S10) were washed with PBS and collected immediately in lysis buffer. In the case of cell lines, the antibiotic was omitted before the last inoculation to avoid unwanted phenotypes.

In the case of microarray analysis, cells were collected in RTL buffer + 143 mM beta-mercaptoethanol and RNA extraction, cDNA synthesis and labelling were performed as previously described *(*72*)*.

In the case of RNAseq, Qiazol was used to lyse the cells. RNA isolation was performed following the manufacturer's instructions (miRNeasy kit from Qiagen), and an in-column DNase treatment was included. RNA concentration was measured using a Nanodrop (Thermo) and its integrity was confirmed in a Bioanalyzer (Agilent). A paired-end directional strand–specific RNAseq protocol (Illumina) was applied for the library preparation at the CRG genomics facility. Briefly, 1 µg of total RNA was fragmented to ~100-150 nt using NEB Next Magnesium RNA Fragmentation Module (EB). Treatments with Antarctic and PNK (both from NEB) were performed in order to make the 5' and 3' ends of the RNA available for adapter ligation. Samples were further processed using the TruSeq small RNA Sample Prep Kit (ref. RS-200-0012, Illumina) according to the manufacturer's protocol. In summary, 3' adapters and subsequently 5' adapters were ligated to the RNA. cDNA was synthesized using reverse transcriptase (SuperScript II, Invitrogen) and a specific primer (RNA RT Primer) complementary to the 3' RNA adapter. cDNA was further amplified by PCR using indexed adapters supplied in the kit. Finally, size selection of the libraries pas performed using 6% Novex TBE Gels (Life Technologies). Fragments with insert sizes of 100 to 130 bp were cut from the gel, and cDNA was precipitated and eluted in 10 µl of elution buffer. dsDNA samples samples were cluster amplified and sequenced in the Hi-Seq 2000 platform (Illumina).

Resulting raw reads were mapped to the *M. pneumoniae* reference genome (NC_000912, NCBI) using MAQ (default parameters, and 1 mismatch allowed) *(*415*)*. Only reads mapping to a unique position of the genome were considered. Counts per gene were extracted from the pileups using our genome annotation and normalized with gene length. To compare the expression levels of each sample with its respective control, quantile normalization was used in the majority of cases, except of those in which we observed a global decay in RNA amounts. In this case, the assumptions of quantile normalization do not hold, and we normalized our data considering that the rRNA is stable and does not decay, compared to mRNAs.

The comparison of gene expression between each sample and its control led to the observation that even significant changes in expression, reproducible across several experiments, were small in magnitude. Thus, we used genes within the same transcription unit as biological replicates for the comparison, to gain statistical power. We did this under the assumption that genes in the same suboperon should be co-regulated across all conditions. To test whether this assumption held, we compared correlation of expression changes in 200 randomized pairs of genes within the same operon and 200 randomized pairs of genes in different operons. This correlation was calculated using 259 experiments of genetic perturbations, and revealed that genes in the same suboperon show much higher correlations (R=0.564) than genes in different operons (R=0.003), and this difference was statistically significant (Student's t-test, p-value<2.2e-16).

To consider a change in gene expression as significant, we filtered by both the absolute value of the $\log_2$ fold change and the p-value from a Student's t-test between the sample and the control. The threshold of the fold change was set empirically, and it varies depending on the number of genes in the operon. We observed that operons with one or few genes have higher variation than operons with more genes. We calculated the standard deviation for operons having 1, 2, …, up to N genes, and defined the threshold as twice the value of the standard deviation. We considered as significant those changes with a p-value smaller than 0.05 and an absolute fold-change larger than the corresponding threshold.

When there was more than one experiment for a gene (for instance, microarrays and RNA-sequencing of the same protein overexpression), we merged the experiments to generate a unique consensus for each putative transcription factor. Experiments were merged after the fold changes and p-values for each individual case were calculated. When merging experiments, we considered significant those results in which the p-value was smaller than 0.05 for any of the datasets, and the average of the fold changes was larger than the threshold determined for that operon.

Network reconstruction and data integration

To identify sets of genes that are tightly co-regulated under a broad set of conditions, we performed correlation analyses of the fold-changes observed in the set of genetic or environmental perturbations.

We reconstructed the network of co-regulated genes in genetic perturbations by establishing an edge between any two operons showing a correlation higher than 0.5 across this set comprising 218 experiments (RNA-seq and microarrays). The threshold was chosen as the value of 3 times the standard deviation of the all-versus-all correlations. Also, upon the observation that the correlation in randomized pairs of genes within the same operon was 0.562 (see above). Using this approach, we obtained a network comprising 221 nodes and 422 edges. To facilitate visualization and interpretation of this network, we clustered it to find groups of nodes highly interconnected. To do so, we used the Girvan-Newman algorithm, as implemented in the GLay plugin for Cytoscape *(*459*,* 460*,* 477*)*. This algorithm finds communities of nodes that are highly interconnected, and removes the edges between different communities. This is done by computing the betweenness centrality of all the edges in the network, and removing the edges with the largest values. We identified 24 clusters, ranging from 2 to 46 operons each. We performed Fisher's tests to determine if there were significant enrichments in any COG category in each cluster.

We followed a similar procedure to reconstruct the network based on environmental perturbations. In this case, we determined correlations across a set of 98 perturbations, after removing those experiments showing a general RNA decay. In this case, the standard deviation of the all-versus-all correlations was slightly larger (σ=0.2) than in the genetic perturbations. Thus, we set the threshold to establish an edge between two nodes as 0.6. With this threshold, we obtained a network of 231 nodes and 293 edges, in which we identified 37 clusters, ranging from 2 to 47 operons each. Again, we identified enrichments in COG categories by means of Fisher tests.

We also reconstructed a network of environmental perturbations, to identify similarities in the transcriptome responses to different stresses. To do so, first we grouped the original 111 perturbations in 41 groups, each containing different experiments with the same type of condition. In this case, to reconstruct the network of environmental perturbations, we first created a bipartite graph of conditions and operons. We established edges between each of the 41 conditions and the operons that change their expression in them. Afterwards, we extracted the condition-projection of this graph. In this projection, conditions that share one or more operons regulated in the same direction are linked. To further constrain this network, we imposed two conditions. First, for every two perturbations sharing an edge, we determined the number of common regulated operons *c*. Then, we determined the sets of regulated operons in each condition (of sizes X and Y respectively). With this, we computed the probability of finding an intersection of size *c* between the two sets by chance. This probability is calculated as:

$$P(|Set1_X \cap Set2_Y| = c) = \frac{\binom{T}{c}\binom{T-c}{X-c}\binom{T-X}{Y-c}}{\binom{T}{X}\binom{T}{Y}}$$

Where *T* is the total possible number of changes, calculated as twice the number of operons. We calculated the probability of finding *c* or more common regulated operons by chance, and chose only those pairs of conditions with values smaller than 0.05. This calculation is an approximation, as it assumes that all the changes are independent. Also, an operon cannot change in opposite directions in the same experiment. We also filtered the network using correlation coefficients, keeping pairs of experiments with a correlation coefficient higher than 0.3.

To combine the data from the different condition experiments with the data of genetic perturbations, we used the information from the 41 grouped conditions and the experiments regarding proteins classified as TFs or regulators, and only in exponential growth phase (40 experiments). For each condition, we performed a regression analysis using random forests. The data on genetic perturbations was used to predict the expression levels of all the genes in each specific condition. Each random forest was reconstructed with 500 trees, and randomly choosing 15 variables (i.e. genetic perturbation experiments) at each split. After running the random forest algorithm, we could estimate the percentage of the variance of the each condition that can be explained by TFs or regulators, and the importance of each putative TF or regulator in explaining the phenotype of the condition.

Growth curves

Cells were pre-cultured for 60-70 hours to achieve late exponential growth phase. Afterwards, cells were scraped and collected in 1 mL of Hayflick culture medium. Cell suspension was pipetted up and down five times to separate aggregated cells. From the cell suspension, 900 µL were aliquoted and stored at -80 ºC to use as inoculum for the growth curves and 100 µL were collected in a 1.5 mL microcentrifuge tube to measure protein concentration. Cells from the collected 100 µL were pelleted by centrifugation at 14100 xg for 10 minutes. Supernatant was discarded and the pellet was washed with 200 µL of PBS. This procedure was repeated twice more. In the last washing step, cells were lysed with 100 µL of lysis buffer (10 mM Tris·HCl, 6 mM $MgCl_2$, 1 mM EDTA, 100 mM NaCl, 0.1% Tx-100, pH 8, and protease inhibitors), pipetting up and down to complete lysis.

Protein lysates were then quantified by Pierce BCA Protein Assay Kit (Prod # 23225, Thermo Scientific), following manufacturer's instructions. Briefly, 25 µL of protein lysate was added to a well of a 96-well plate, by duplicate. The standards for reference were prepared with BSA at different concentrations, following manufacturer's instructions, and lysis buffer. Then, 200 µL of BCA working reagent were added to each well and mixed in a Tecan Infinite M200 plate reader for 30 seconds. Samples were incubated at 37ºC for 30 minutes, and after cooling down to room temperature, absorbance at 562 nm was measured using a Tecan Infinite M200 plate reader to determine protein concentrations. Concentrations from the standards were used to make the standard curve and extrapolate protein concentration from each sample.

After calculating the sample concentrations, 1 µg of protein, from the quantified pre-cultured aliquots, were used as inoculum to start the growth curves for all the mutants in a final volume of 200 µL of Hayflick culture medium in 96-well plates. Growth curves were determined in duplicate by using pH measurements, following the protocol previously described in *(*249*)*. Absorbance at 430 nm and 560 nm were taken once every 20 minutes in a Tecan Infinite M200 plate reader at 37 ºC for 5 days.

Motif pull down

Protocol was adapted from *(*478*)*. First, from a 3 days 300 cm2 culture flask, cells were washed with ice-cold PBS, scrapped in PBS plus 0.1% glucose at 4ºC, and spun 10 min on a tabletop centrifuge. Pellet was resuspended in 2 ml lysis buffer (1 M NaCl, 50 mM Hepes·NaOH, 0.1% NP 40, 6 mM $MgCl_2$, 1 mM EGTA, 1 mM EDTA, pH 7.5) plus a protease inhibitor Cocktail (Roche) and passed through a syringe G25 needle 10 times prior to clearance by spinning 30 min in a benchtop centrifuge at maximum speed and 4ºC. Supernatant was diluted 1:10 in dilution buffer (50 mM Hepes·NaOH, 1 mM EGTA, 6 mM $MgCl_2$ pH 7.5) and 4.4 ml were used per assay. Sepharose-streptavidin beads (M-280 from Invitrogen) were bound to biotinylated oligos as follows. First forward and reverse oligos at 50 µM were annealed in Annealing buffer (10 mM Tris·HCl, 50 mM NaCl, 1 mM EDTA, pH 8.0) in a PCR machine: 95ºC 2 min, 52ºC 10 min, 4ºC. Then 20 µl annealed oligos were mixed and incubated with equilibrated (TE) beads for 30 min at 4C in a roller. Beads were washed with binding buffer (50 mM Hepes·NaOH, 1 mM EGTA, 0.1 M NaCl, pH 7.5) and incubated with lysate 1h or O/N at 4C. Formaldehyde was added to 1% and proteins and DNA were fixed for 10 minutes at RT. Crosslinking was stopped with glycine (100 mM final) 5 minutes at RT. Beads with 1 ml of binding buffer, 3X with 1 ml wash buffer 1 (50 mM Hepes·NaOH, 1 mM EGTA, 0.2 M NaCl, 6 M Urea, 0.2% SDS, pH 7.5) and 3X with 1 ml wash buffer 2 (50 mM Hepes·NaOH, 1mM EGTA, 0.2 M NaCl, pH 7.5). Material was eluted/de-crosskinked with 50 µl of elution buffer (1% SDS, 10 mM Tris·HCl, 1 mM EDTA pH 8.0) at 65ºC 15 min and 95ºC 5 min and visualized on a SDS electrophoresis gel after staining with Instant Blue Coomassie (Expedeon). Optimal pull-downs were submitted to proteomics (see below).

Proteomics

Some gain or loss of function phenotypes were determined by subjecting the cell lines to protein quantification. Briefly *M. pneumoniae* cell lines overexpressing various putative regulators or bearing transposon insertions, were grown for 6 h and collected after washing twice with PBS with FASP buffer (100 mM Hepes·NaOH, 4% SDS pH 8.0). After total protein was determined by a BCA assay (Pierce), DTT was added to 100 mM. In some cases a sonication step was needed to disaggregate the lysate. After[EY3] trypsin digestion of 200 µl of each sample (amounts ranging from 20 to 486 µg) samples were desalted and dissolved in 300 µl and the 2.5 µl of each fraction was injected in an LTQ Velos Pro in the order of the chromatographic elution using a

MEDI_CID method. BSA controls were injected after each sample. 20 µg of the total extract was also digested and 1 µg injected in an LTQ Velos Pro using a LONG_CID method. The data has been searched using an internal version of the search algorithm Mascot against a database that contains all the putative proteins larger than19 amino acid (MPNHomoContTrans19). The data has been filtered using 5% FDR (False Discovery Rate). Protein grouping was not applied in the results and we have quantified the proteins using the following parameters: i) only peptides without miss cleavage; ii) only peptides with "Protein Group=1"; Top3 algorithm (it considers three best peptides of each protein); Top3 method (average of area of the three best peptides). Only peptides corresponding to ORFs for which we could identify an RNA transcript were considered.

Real time PCR

In order to monitor promoter regulation, reporter quimeras with YFP-Venus were built. As YFP seems to be a very stable protein in *M. pneumoniae*, gene expression was followed by real time PCR. Briefly, RNA was purified as above, and 1 µg was hybridized to 2 µg random hexamers (Invtrogen) by heating to 65 ºC for 5 min and quick chilling on ice in a 11 µl total volume. Retrotranscription was performed by adding 4 µl 5X first-strand buffer, 2 µl 0.1 M DTT, 1 µl RNase OUT (40 units/µl, Promega), 1 µl 10 mM dNTP mix, and 1 µl SuperScript II RT (200 units, Invitrogen) and incubating for 50 min at 42C before inactivation at 70C for 15 min. A 2x GoTaq qPCR mastermix was used (Promega) with 0.5 ng cDNA per 10 µl reaction and 0.5 µM oligos and run on a Lightcycler 480 (Roche). Venus oligos: F_qVen: ACGTAAACGGCCACAAGTTC, R_qVen: GGTCTTGTAGTTGCCGTCGT. Ribosomal RNA was used as reference, F_q16S: GCAGGTAATGGCTAGAGTTTGACT, R_q16S: GCCTTTAACACCAGACTTTTCAAT.

Novobiocin titration

We treated *M. pneumoniae* cells with increasing concentrations of novobiocin (0, control; 1; 5; 10; 50 and 100 µg/mL) for 30 minutes. Two replicates were used for each timepoint. After the treatment, total RNA was extracted and we performed transcriptome sequencing as detailed above. After read mapping and gene expression calculation, we normalized the data considering the rRNA does not change its expression in these experiments due to its high stability. We scaled the expression values of each gene by subtracting the mean value of the 5 experiments and dividing by their standard deviation. With the scaled values, we computed the correlation matrix for all the genes of *M. pneumoniae*. A heatmap of this correlation matrix (Figure 7.8A) showed a plaid pattern with 5 major groups of genes having differential behaviors. Therefore, we performed k-means clustering in our data with 5 centers to find the patterns corresponding to the different groups of genes.

RNA half-life determination

RNA half-lives were determined as previously described *(*23*)*. Briefly, we considered a simple scenario in which transcription is modeled as the continuous balance between RNA production and degradation, according to the following equation: d[RNA]dt=-k[RNA], where *α* and *k* are the production and degradation rates, respectively.

To determine the degradation rate *k*, the term of production (*α*) should be cancelled. Then, the differential equation can be solved to obtain that $[RNA]=[RNA]_0 \cdot e^{-kt}$.

To experimentally make the transcription rate *α* equal to zero, we used novobiocin, a DNA gyrase inhibitor, that causes the release of the RNAP complex of the chromosome of *M. pneumoniae*. We treated cells in exponential and stationary growth phases with 100 µg/mL of novobiocin and extracted total RNA at different timepoints after the addition: 0 (as a control), 2, 4, 6, 8, 10 and 15 minutes, with two biological replicates for each timepoint. We performed RNA-sequencing and transcript levels were calculated for each of the samples as detailed above. Normalization was performed assuming no degradation of the rRNA. Transcript levels were transformed to copy numbers per cell using an experimentally determined adjust function *(*141*)* and then to RNA concentrations, considering an approximate volume of $0.055µm^3$ for *M. pneumoniae* (479). We used these concentrations to fit an exponential decay curve, and determined the degradation rates (*k*) for each gene. Given the degradation rates, we determined the half-live of all genes in *M. pneumoniae* as $t_{1/2}$=log(2)/*k*.

Riboswitch scan

To find potential riboswitches in the genome of *M. pneumoniae*, we analyzed 206 experiments of genetic and environmental perturbations, with two replicates each. For each experiment, we identified the annotated TSSs and defined 3 different regions around each of them: the previous region, covering 100 bps upstream the TSS; the riboswitch region, spanning the first 150 bps of the transcript; and the gene region, including 300 bps starting at the end of the riboswitch region. If the transcript terminates before the end of the gene region, this was shortened to match the length of the transcript. We calculated the expression values for each of these regions in the 206 experiments and in their corresponding control samples (see above).

To annotate a putative riboswitch that regulates premature termination of transcription, we expect to find differential expression when comparing the riboswitch and the gene regions. Indeed, we expect that the riboswitch region has higher RNA levels than the gene region. Therefore, we calculated the ratio riboswitch/gene for all the annotated TSSs in all conditions. We removed the data points in which the expression of the riboswitch or the gene was smaller than 5 $\log_2$ CPKM, as in these, small fluctuations due to experimental noise lead to large changes in the ratios. We then scaled the ratios of each TSS, by subtracting the mean of all the experiments for that TSS.

After scaling, we selected as putative riboswitches those cases in which the scaled riboswitch/gene ratio was larger than 3 standard deviations of the whole distribution (considering all the experiments). We applied two further filters to increase the specificity of our search. First, the TSS should be active in the condition where the riboswitch is found. Therefore, we compared the riboswitch region with the previous region, and only kept those cases in which the riboswitch expression was significantly larger than the one of the previous region (using a t-test and filtering by fold-change and p-value). Second, the riboswitch should behave differently between the condition where it is identified and its corresponding control experiment. We applied a t-test to compare the riboswitch/gene ratio of the sample and the control, and only kept those cases in which the difference was significant.

After this filtering, we identified a set of riboswitches regulated in specific conditions. We then plotted the RNA-sequencing profiles of each of these riboswitches in all the conditions tested, to manually curate and validate our results, and to identify further conditions that did not pass our initial filtering criteria.

7.6. Author contributions

SM performed the identification of protein candidates by the different isolation techniques, and isolated the metabolomics samples. EY expressed each of the candidates in *M. pneumoniae* and perfomed the DNA extraction and ChIP sample preparation, the RNA extraction for the transcriptomics and the protein extraction for the proteomics experiments. D. Sevin performed the metabolomics experiments used for validation. EY and CG ran the growth curves of *M. pneumoniae*. VLR analyzed the ChIP-seq, transcriptomics and proteomics datasets and designed the pipeline for the custom analyses, integrated these data and reconstructed the gene regulatory network. EY performed the novobiocin experiment and VLR determined RNA half-lives and decay rates. VLR performed the riboswitch analysis. MLS performed the ChIP-seq classification analysis. EY, LS and US designed and supervised the full study. EY and VLR wrote the manuscript.

7.7. Supplementary data

Supplementary Material (Tables S1-S15, figures S1-S11 and their corresponding legends, as well as Figures 7.1-7.9 in high-resolution, are available at the following link: https://www.dropbox.com/sh/vi1a3545rx5rv5u/AAAQMRNFuks-WmUfMMeGKZyVa?dl=0).

8. Discussion

One of the challenges of current biology is to understand how entire cells or organisms function in homeostasis and under perturbations, and then develop computer models that will allow researchers to rationally engineer them. However, our limited knowledge of the mechanisms underlying many biological processes hampers the development of detailed predictive models at large scales. Even for the simplest self-replicating organisms, Mycoplasmas, with a reduced set of protein-coding genes, many of the functions encoded by these genes are still unknown. A striking example of our limited knowledge is that of the recent creation of a synthetic minimal Mycoplasma cell (*480*). In this work, a first approach was taken in which all the biological knowledge on the bacterium *M. mycoides* was used to decide on a set of genes to remove. However, this rational design approach failed, and authors recurred to an approach that tested individual regions of the genome separately in an iterative manner, until the obtention of the minimal cell (*480*). The genome of this cell contains 149 genes (out of 473) with an unknown function, many of which are conserved across several bacterial species.

This 'missing knowledge' clearly supposes a challenge for the emergent field of whole-cell (WC) modeling. The basis of this new discipline relies on explicitly modeling the function of every gene, to be able to reproduce and predict emergent behaviors, responses to perturbations, or the effect of introduction of novel functions for synthetic biology applications. This had been previously done for the smallest, naturally-occurring, self-replicating organism, *M. genitalium* (*244*). With only 525 protein-coding genes, this represents an ideal example for a proof-of-concept study for whole-cell models. The authors of this model proved that it was possible to integrate several cellular processes occurring at different time scales. However, the lack of biological knowledge in this bacterium resulted in the majority of predictions being only qualitative and not quantitative (*244*, *291*).

To test if an accurate knowledge of the biology of the bacterium could improve the predictions of these models, we have developed a whole-cell model of *M. pneumoniae* (Chapter 3). We took advantage of the structure of the original model, and of the phylogenetic proximity between *M. pneumoniae* and *M. genitalium*. Furthermore, *M. pneumoniae* is currently one of most characterized bacteria to date, with a large consortium led by this lab, devoted to its molecular characterization (*72*, *73*, *157*, *203*, *233*, *249*).

8.1. A critical view on 'omics' technologies

The vast majority of the biological knowledge in *M. pneumoniae* generated by this consortium has been obtained via high throughput or 'omics' experiments. In this thesis, we have included a paper (Chapter 4) that refers to the problems associated to the fast development of these 'omics' technologies. Concretely, in this work we present a problem related to the process of library preparation for RNA-seq experiments. The observation of large percentages of non-mapped reads in RNA sequencing experiments, together with the excitement over the existence of chimeric RNAs in eukaryotic genomes, brought by the success of large consortiums such as ENCODE, led to the question of whether these chimeric RNAs could exist in bacteria. If so, their existence would introduce a new paradigm in microbiology and explain part of the non-mapped reads present in these datasets.

To address this issue, an experiment was designed to assess whether chimeric reads exist, and if so, whether they were natural or artifactual, produced in the RNA extraction phase or in the library preparation prior to the sequencing. In this experiment we obtained RNA from two bacterial species independently, we mixed them and then extracted the RNA or we mixed the RNA after extraction.

We used the sequencing data to test three widely used pipelines to detect chimeric RNAs, and all of them yielded inter-species chimeric RNAs. Therefore, we designed a more stringent custom pipeline, that applied to our data resulted only in a small number of intra-species chimeras. All of these had in common that the two chimeric RNA fragments were derived from highly expressed transcripts, and that they formed very stable secondary structures. This led to the investigation of an alternative library preparation protocol, in which chimeric RNAs were not detected by any of the computational pipelines tested. The difference between protocols relies in the step of reverse transcription. In the original protocol, reverse transcription occurred after ligation of adapters, whilst in the second protocol studied, it occurred before adapter ligation. If the ligation step occurs before reverse transcription, some RNAs such as tRNAs and rRNAs might retain their secondary structure, and RNA fragments that are close in the space due to these structures may ligate together rather than with the adapters, giving rise to artifactual chimeric RNAs. The step of reverse transcription necessarily dissociates all secondary structures in the RNA, as the reverse transcriptase is a processive enzyme, thus eliminating these artifacts.

These findings are representative of the number of biases and challenges found in high-throughput profiling techniques, and they are not unique. Frequently, these problems are found long after the technologies, protocols and computational pipelines are developed, revealing that it is difficult to cope with the rapid evolution of these techniques. Another study focused in RNA-seq in prokaryotes also reports a high number of non-mapped transcripts that correspond to technical artifacts, and that contribute to the 'RNA-seq trash bin' (*327*). Some of the artifacts they report arise from the reverse transcription process (*481*), and these were accounted for in our analysis. Other steps of RNA-seq library preparation can also lead to biases in the results, such as rRNA depletion, fragmentation and fragment selection, as well as the PCR amplification (*172*).

Besides these artifacts, we have detected that, in all our transcriptomics experiments, noise levels are extremely large. By performing numerous replicates of the same experiment, we have been able to identify true differentially expressed genes. However, the fold changes observed are rarely larger than 1.5 (in $\log_2$), and are usually in the range of 0.7-1.2, which largely overlaps with the experimental noise. This hampers the application of standard statistical analyses and filtering thresholds in differential gene expression studies. Instead, custom pipelines need to be designed. Furthermore, the size selection steps usually performed in library preparation prior to RNA sequencing are very sensitive to small transcripts. These RNAs can be lost easily in this step, which results in an increase in the variability of small RNAs. We have observed that in some experiments, there are many less reads mapping to the 5S rRNA than mapping to the 23S or 16S rRNAs, whilst we expect similar numbers as these rRNAs are transcribed in a single operon. Sometimes, the observed difference between the levels of these rRNAs is up to 5-fold. Besides the possibility of a differential processing, the short length of the 5S rRNA in *M. pneumoniae* (108 bases) suggests that a large proportion of these molecules could be lost in the library preparation process. Improvements in library preparation protocols and in single cell sequencing techniques will be required to address these problems.

Aside from RNA-seq, other -omics suffer from similar biases. In chromatin immunoprecipitation, different steps can be sources of biases altering the experimental results. Fragmentation of DNA by sonication is a crucial step of this protocol, and it is not recommended to use controls that have not been sonicated together with the ChIP samples (*482*). Differences in sonication can lead to the observation of peaks with shifted position or different amplitudes, hindering normalization of the data. The step of immunoprecipitation can also be a source of biases. Frequently, antibodies targeting the protein of interest are used for this purpose. When an antibody is not available, the protein needs to be expressed exogenously, fused to a protein tag. This allows the usage of an antibody targeting this tag. However, the exogenous expression is associated to other problems. The protein needs to be expressed at levels sufficient to displace the endogenous protein from the target sites, but an excess of protein may bind DNA in a non-specific manner. Also, the addition of the tag may alter the binding of the protein or its functionality. Furthermore, it has recently been show that even with specific antibodies that do not require the addition of a tag to the protein of interest, some off-targets may appear in the analysis. Some DNA proteins such as the RNA polymerase have a disordered structure that is prone to interact with the antibodies in a non-specific manner. These proteins are the reason why some peaks appear even when the protein of interest is not expressed, the 'phantom peaks' (*483*). We have identified the presence of these peaks in our ChIP-seq experiments (Chapter 7), associated to promoters of highly-expressed genes.

Despite these challenges, high throughput profiling technologies have helped to get a global view of biological processes that was not possible previously. They also present a series of advantages, such as the usage of consensus naming conventions for all genes or proteins, and normalized values that can be used to compare among different features. Indeed, the whole-cell model of *M. pneumoniae* relies on a knowledge base that is built mostly upon omics datasets. Other data, generated via computational predictions or from global databases, was also included to complete this database.

8.2. Transcription in *M. pneumoniae*

The first version of the *M. pneumoniae* model was obtained using the model of *M. genitalium* as a starting point. Several unique aspects of the biology of *M. pneumoniae* were also included, and gene miss-annotations were corrected. However, after the first simulation runs of this model were completed, we identified a number of 'knowledge gaps'. These are inconsistencies between the model predictions and our validation data, and demonstrate that despite our broad knowledge of the biology of this bacterium, more experimentation is needed to learn about the aspects of the biology that remain unknown.

One of these aspects is related to transcriptional regulation, which is also one of the challenges of current molecular biology: to understand how transcription is orchestrated. There are many components contributing to promote or prevent transcription, and the interplay among them results in specific levels of RNAs for a certain condition. Although transcription is a universal process occurring in all forms of life, there are several differences between prokaryotic and eukaryotic transcription. Even within prokaryotes, differences arise between Gram positive and Gram negative bacteria, reviewed in the introduction of this thesis. Despite these differences, understanding the interplay of the different components of transcription in a minimal bacterium such as *M. pneumoniae* can provide the basis for comprehending how this process is orchestrated in more complex organisms. To contribute to this objective, and to improve our current whole-cell model of this bacterium, we have chosen to study in depth the process of transcription in *M. pneumoniae*. This organism, with a limited number of sigma and other transcription factors, facilitates the study of the core transcription determinants.

The main objective of this thesis is to assess how different determinants of transcription contribute to the different transcript abundances observed in the cells, both in physiological conditions and under several perturbations. Three articles are included in this thesis that respond to this objective. The first of them corresponds to the characterization of promoter sequences in *M. pneumoniae*, and the distinction between real promoter sequences and other non-promoter elements, either giving rise to abortive transcripts or not associated to any transcriptional event (Chapter 5). The second corresponds to the study of the functionality of sRNAs in *M. pneumoniae* that can be extended to other bacteria (Chapter 6). The third study refers to the transcription factors and the reconstruction of the gene regulatory network in this bacterium, and how elements other than transcription factors may be essential to control the transcriptional program of the cells (Chapter 7).

8.2.1. Getting quantitative predictions of RNA expression from promoter characterization

In this study, we presented a method to characterize and describe bacterial promoters. This method is based in six promoter features that had been previously reported to be important for promoter function. We integrated this six features in a random forest classifier, a machine learning algorithm that, properly trained, allows to distinguish among different classes of sequences. The random forest was trained with both promoter and non-promoter sequences of *M. pneumoniae*, and provided a useful distinction among the three classes of sequences found in the genome of this bacterium: actual promoters, unproductive promoters associated to tssRNAs, and sequences not related to any transcriptional event. This study also provided insights in which are the most important features for promoter definition in *M. pneumoniae*. In this case, the Pribnow motif is the most relevant factor, followed by the stacking energy of the nucleotides of the promoter region. The other sequence motifs have smaller contributions. Similar studies in other mycoplasmas also highlight the importance of the Pribnow motif in promoter determination (*256*).

A surprising finding in our work is that unproductive promoters are more similar to non-promoter sequences than to actual promoters, yet they are capable of producing short RNAs of around 50 nucleotides, the so-called tssRNAs (*157*). These unproductive promoters were not used in the training of the random forest not to bias its outcome. This finding implies that these sequences are capable of binding the RNA polymerase complex and induce transcription initiation, but there must be a physical impediment for transcription elongation to proceed, an impediment that does not exist in true promoter sequences. It is possible that the RNA polymerase complex fails to capture some necessary elongation factors (*484*).

A question arising from this study is whether it is possible to use this computational algorithm to obtain quantitative predictions of the RNA levels of a transcript, given its promoter sequence. Using the set of RNAs from *M. pneumoniae* for this purpose is not possible, as the RNA levels observed are the consequence of the interplay of different factors, one of which is the promoter strength and the other RNA degradation. This highlights the need of *in vivo* or *in vitro* systems that minimize the effect of the other elements involved. One such system that is being currently developed is 'Prot-seq' (Yang et al, in preparation). In Prot-seq, the bacterial DNA methylase Dam is expressed under the control of randomized promoter and/or 5'-UTR sequences. Assuming that no transcription factors other than the housekeeping sigma 70 bind these regions, and that accessibility of the polymerase and ribosome to their binding sequences is independent of the genomic location of the construct, it is possible to study the influence of the randomized sequences in the expression of the reporter gene. The readout of this methodology consists of the Dam methylase activity, proportional to the Dam concentration in the cell. This activity is measured by sequencing the GATC sites methylated by this protein, using DamID-sequencing (*485*).

We used Prot-seq to test thousands of randomized promoters controlling the transcription of the Dam methylase, in constructs randomly inserted in the chromosome of *M. pneumoniae*. All randomized promoters contained at least the canonical Pribnow motif 5'-TATAAT-3', and the 5'-UTR used was the same in all constructs. All these promoters were sequenced and evaluated using the random forest classifier described above. The results were compared to the Dam methylase activity, used as a proxy for the RNA expression. We found a positive correlation among the random forest score and the Dam activity.

Despite this correlation, there is a large variability in the promoter scores obtained. One possible explanation for this is that all randomized sequences contain a canonical Pribnow box. This is the major determinant of our promoter score, and it remains fixed in all the sequences, which causes an important loss of information for the classifier. To address this problem, we took 14000 randomized promoter sequences to re-train and test a new classifier, based on the same promoter features, except for the Pribnow score. 10000 sequences were used to train this classifier, and the remaining 4000 were used for testing its performance. Their expression was binned in three levels: high expression, medium expression and low expression. Only those with high and low expression were used for the training of this new classifier.

This new random forest was used to classify the remaining sequences from the test set. A significant separation between high and low expression promoters was accomplished. We built a ROC curve, and the area under it is 0.84, with precision of

0.71. We then decided to apply this classifier to all 4000 sequences in the test set (including those with medium expression levels) and find differences in their random forest scores. We found significant differences among the three groups of sequences (Mann-Whitney U test, p-value < 2.2 e-16 for all comparisons).

These results imply that this classifier can be used for a semi-quantitative prediction of transcript expression, given the promoter sequence. It is possible that including modifications of the canonical Pribnow motif in the sequences studied and in the classifier helps improving the resolution of the predictions obtained. The results also suggest that the features used for promoter classification and identification, can also be used to predict gene expression, and that promoters of different strengths lead to different transcript levels, regardless the rest of elements intervening in transcription. Intriguingly, we observed that natural promoters of *M. pneumoniae* have intermediate scores rather than high scores. This is interesting as it can provide room for both positive and negative regulation by different transcription factors.

These methods can be extended to other bacteria, as it has been a long standing question whether it is possible to predict gene expression from sequence features (*486–488*). This can have numerous applications in synthetic promoter design for genetic engineering purposes (*489*).

8.2.2. Most bacterial antisense RNAs are the product of transcriptional noise.

In this chapter, we present a striking correlation between the number of sRNAs and the genomic A+T content across 20 bacterial species, including also the chloroplast of *A. thaliana*. This correlation arises despite the fact that the sequencing and data analysis leading to the annotation of sRNAs in these organisms was done by different research teams. The correlation is maintained by antisense RNAs (asRNAs) but not by trans-encoded sRNAs or for protein-coding RNAs. This observation, together with the fact that the most important promoter determinant in bacteria is the Pribnow motif, an AT-rich element, led to the hypothesis of spurious promoter sequences arising by random mutations. These spurious promoters would give rise to asRNAs that would therefore be in general non-functional. In the paper we present evidence supporting this hypothesis, based on the low expression levels of these RNAs, their limited essentiality and also on stochastic computer simulations. Similar hypotheses have been presented by other authors, referring to the low expression levels and also to a limited conservation of asRNAs across species (*59*).

It is important to note that action mechanisms other than the ones presented in this work should not be discarded and therefore that some asRNAs have functionality. There are cases in which the effect of the asRNA does not occur via duplex formation. An example of this is transcription interference. In this scenario, opposing convergent promoters can cause incompatibilities in the transcription of both sense and antisense transcripts (*490*). According to this, the strength of the antisense promoter should correlate with the repression of transcription in the sense strand. Indeed, it has been possible tune gene expression using antisense promoters of different strengths (*491*). Another possible mechanism is that of transcription attenuation, in which an antisense RNA causes the formation of a premature termination site, preventing transcription elongation and therefore regulating the levels of sense transcripts (*69*).

Despite the existence of alternative mechanisms not considered in our work, it is necessary to consider the physiological copy numbers of asRNAs when considering any possible regulatory process. According to our simulations, none of the asRNAs detected in *M. pneumoniae* is present at levels high enough to trigger a non-enzymatic (ie Dicer type in eukaryotes) regulatory response. To consider the possibility of transcriptional interference, we should also take into account the number of RNA polymerases present in the cell. For *M. pneumoniae*, this number is estimated in ~150 (*203*). Given that there are more than 800 promoters in the genome of this bacterium (see Chapter 5), the events in which two polymerases collide, transcribing simultaneously from converging promoters, are expected to be rare. In *E. coli* and *Salmonella enterica*, only 3% of the asRNAs are expressed at high levels (*59*). Nevertheless, it is this small percentage of molecules that has received most of the attention, and studies reporting regulatory functions for asRNAs refer to this small group of highly expressed transcripts (*407*, *408*).

Finally, trans-encoded sRNAs should not be ignored. The fact that they do not accumulate in bacterial genomes proportionally to the A+T genomic content, behaving more similarly to protein-coding genes, suggests that these transcripts could have different regulatory functions, controlling processes such as virulence or quorum sensing (*492*, *493*). Indeed, many studies on sRNAs in bacteria focus on these intergenic transcripts. Furthermore, recent studies have found that some of these RNAs could actually be coding for small proteins (*73*). These varied functions could possibly explain the differences in the distribution of these molecules with respect to asRNAs.

8.2.3. The importance of non-TF regulation in *M. pneumoniae*

Chapter 7 of this thesis refers to the reconstruction of the gene regulatory network (GRN) of this genome-reduced bacterium. To do so, we identified all putative DNA-binding proteins (DNABPs) in *M. pneumoniae*. We overexpressed or knocked out each of them in *M. pneumoniae*, and tested their function by identifying their physical interactions with the chromosome (their binding sites) and also their genetic interactions (their targets in the GRN). We classified these proteins in TFs, regulators, structural, RNAP-like and non-specific. It is important to note that, unlike in similar studies performed in other bacteria such as *Mycobacterium tuberculosis* (*183*), we do not limit our search to annotated or predicted TFs. Instead, our work represents the first global unbiased study of all DNA-protein interactions performed in a bacterium.

Surprisingly, the GRN reconstructed with both TFs and regulators only covers a small percentage of the genome of *M. pneumoniae,* and includes 9 TFs and 30 regulators, yet we find several clusters of co-regulated operons across more than 100 conditions that cannot be explained by these TFs alone. Indeed, when comparing our results with data on different environmental perturbations, we found that considering both TFs and regulators, we could not explain more than 50% of the variance in any of these experiments. Although we have evidence suggesting that this could be an underestimation of the real percentage explained, due to the elevate noise in experimental data, this percentage is lower than the one reported in other species such as *B. subtilis* (66%; (*140*)). Given that we reported a high coverage of the DNABPs, we hypothesize that other layers of regulation must exist and play a central role in determining RNA levels. Thus, we investigated the role of DNA supercoiling, the presence of riboswitches, the regulation via metabolites and nucleotide concentration, and the differences in RNA degradation at distinct points of the growth curve. Also, in a recent study from our lab (*23*), we described transcriptional read-through, as another possible mechanism for transcriptional regulation.

Our study revealed that each of these factors could have a different role in regulating RNA levels in *M. pneumoniae.* Regarding supercoiling, previous works have reported that supercoiling could be acting at the top of the transcriptional regulation hierarchy (*426*, *494*). So far, we have discovered a group of genes that responds to supercoiling. Interestingly, the majority of these are located close to the origin of replication of the chromosome. However, besides this group of genes, we have not been able to unmask a global effect. Probably, this global effect could be explained in terms of the basal coordination of transcription observed in a recent work from our lab (*23*). Here, it has been observed that termination signals can be overridden in a condition-dependent manner, allowing for transcription *en bloc* to occur under certain circumstances. This could be a basal regulatory layer in bacteria, not unique to *M. pneumoniae*.

Another basal mechanism of transcriptional regulation is that mediated by metabolite concentrations. It has been described in other bacteria such as *B. subtilis* that concentrations of GTP can alter transcription initiation rates of those RNAs having a GTP in the +1 position (*88*). We observed that a similar mechanism could operate in *M. pneumoniae*. Furthermore, an exploratory analysis of the first positions of all RNAs in this bacterium revealed that RNAs starting with a GC dinucleotide are enriched in growth-related functions. Therefore, this mechanism could be involved in the transition from exponential to stationary growth phase, in which the GTP levels decrease. Therefore, this transition would not require the direct action of a protein master regulator.

Finally, other mechanisms could aid transcriptional regulation under specific circumstances. We described the existence of riboswitches that are activated under a limited number of conditions. However, unlike the aforementioned mechanisms, this would not represent a basal layer of regulation but an additional form acting on top of the elementary layers. Another specific form of regulating RNA levels is through degradation control. We have observed significant differences in transcript half-lives at various growth phases. However, we were not able to correlate this to the composition of the RNA degradosome. Further experimentation on this complex may reveal how the different components of the complex regulate the affinities for RNAs.

An example of how different layers of regulation interact to shape the transcriptional response is found in heat-shock and cold-shock perturbations. In gram-positive bacteria, there is a transcription factor, HrcA, responsible for the heat shock response. This is a repressor that controls the expression of chaperones and other response genes that have a common motif in their promoter, the CIRCE element *(*495*)*. In *M. pneumoniae*, this protein is encoded by the *mpn124* gene. However, only a fraction of all the genes whose expression is altered in heat shock is controlled by this mechanism. We found that a number of genes changing in heat shock and not under the control of the HrcA protein. Interestingly, these genes also change in the cold-shock perturbation, in the opposite direction. This points to a non-TF regulatory mechanism, involving probably the structure of the DNA. Changes in temperature may alter the properties of the DNA and make it more or less accessible at certain points, and thus regulating transcription initiation *(*496*)*. This regulation may also be due to differential transcriptional read-through, as we have recently described *(*23*)*.

Besides this example, our analyses have been preliminary so far, and further research is required to understand to which extent these mechanisms contribute to global transcriptional regulation and how they interact. Supercoiling, transcriptional read-through and nucleotide concentrations, together with some of the studied TFs, seem to be at the top of the regulatory hierarchy, coordinating basal transcription, whilst other factors such as riboswitches and TFs would be acting on top of this basal layer, coordinating more specific responses to certain conditions. In other bacteria with larger numbers of TFs, the remaining regulatory mechanisms remain largely unexplored. Thus, *M. pneumoniae* represents an ideal model organism to disentangle and understand these 'hidden' regulatory layers. Also, the existence of a whole-cell model in this bacterium will allow for the integration of all these layers to gain a global view of transcription in this organism.

8.3. Perspectives on the whole-cell model of *M. pneumoniae*

Our work in characterizing the key determinants in RNA abundance in *M. pneumoniae* can be extremely valuable for the development of the WC model of this bacterium. Future efforts in this direction will imply the integration of the effects of these determinants in a genome-scale model of transcription. This model should be able to reproduce the results observed in our transcriptomics experiments, and predict changes in different environmental conditions and growth phases.

Furthermore, we should be able to integrate this model in the WC model of *M. pneumoniae*, by replacing the current modules on transcription and transcriptional regulation. To do so, technical improvements in the model should be implemented. Its architecture should be redesigned to allow for a modular structure, facilitating the addition of new modules or replacement of existing ones by third parties. Current work in Dr. Karr's group points to this direction.

To further improve these models in the future, additional spatial considerations need to be taken, such as molecular crowding. Cellular components are not point particles, but they occupy a certain volume inside the cell. The whole set of components of the cell, with their respective volumes, should be considered, and the limitations that this molecular crowding imposes in free diffusion should be accounted for. A recent work in *M. genitalium* (497)

shows indeed that the chromosome and the entire proteome of this bacterium account for a large percentage of its volume. This work also highlights that although the aforementioned models are claimed to be 'whole-cell', an important part of the physiology of these bacteria, related to the structure of their components, is missing. Indeed, bridging the existing gap between protein structure and function will allow researchers to make more accurate predictions when including this information in a model. For instance, effects of specific mutations in different proteins could be *in silico* tested.

Altogether, the work in this thesis aims to provide the basis for future studies aiming at the improvement of WC models in general and the WC model of *M. pneumoniae* in particular, as well as to better understand the process of bacterial transcription. A first version of the WC model of this bacterium has been implemented, and work towards a second version including novel biological knowledge is currently ongoing. In parallel, we have taken advantage of all the transcriptomics datasets being generated in our lab to study in depth which are all the key components that shape the transcriptional landscape of this bacterium. We have found promoter features that help distinguishing true from non-productive promoters. Furthermore, we have shown that copy numbers matter, and that the extremely low abundances of the majority of asRNAs deem very unlikely a regulatory function for them. Additionally, we suggest that non-TF factor regulation, mediated by metabolites, riboswitches, supercoiling or RNA degradation, can be as important as TF factor based transcriptional control. Altogether, our studies suggest that regulation of transcription is mediated by multiple layers with cross-talk and feedback among them. Future work in this direction will imply the integration of all these layers and components in a genome-scale model of transcription that is able to simulate the results observed in transcriptomics experiments. Furthermore, this model could be integrated as a part of the current WC model of *M. pneumoniae*. Besides, taking advantage of all the 'omics' experiments used to feed the model and to study the process of transcription, we have assessed some of the biases and challenges that high-throughput profiling techniques have in bacteria. These biases do not prevent the usage of these technologies, whose advantages overcome their limitations, but need to be carefully considered when analyzing these experiments.

9. References

1. *F. H. Crick, On protein synthesis.* Symp. Soc. Exp. Biol. ***12**, 138–163 (1958).*
2. *D. Baltimore, RNA-dependent DNA polymerase in virions of RNA tumour viruses.* Nature*. **226**, 1209–1211 (1970).*
3. *H. M. Temin, S. Mizutani, RNA-dependent DNA polymerase in virions of Rous sarcoma virus.* Nature*. **226**, 1211–1213 (1970).*
4. *J. T. August, L. Shapiro, L. Eoyang, REPLICATION OF RNA VIRUSES I. CHARACTERIZATION OF A VIRAL RNA-DEPENDENT RNA POLYMERASE.* J. Mol. Biol. ***11**, 257–271 (1965).*
5. *S. Cooper, N. D. Zinder, The growth of an RNA bacteriophage: The role of DNA synthesis.* Virology*. **18**, 405–411 (1963).*
6. *J. S. Griffith, Self-replication and scrapie.* Nature*. **215**, 1043–1044 (1967).*
7. *I. H. Pattison, K. M. Jones, The possible nature of the transmissible agent of scrapie.* Vet. Rec. ***80**, 2–9 (1967).*
8. *D. A. Kocisko* et al.*, Cell-free formation of protease-resistant prion protein.* Nature*. **370**, 471–474 (1994).*
9. *E. V. Koonin, Does the central dogma still stand?* Biol. Direct*. **7**, 27 (2012).*
10. *P. Zuber, Non-ribosomal peptide synthesis.* Curr. Opin. Cell Biol. ***3**, 1046–1050 (1991).*
11. *S. Caboche* et al.*, NORINE: a database of nonribosomal peptides.* Nucleic Acids Res. ***36**, D326–31 (2008).*
12. *D. S. Oppenheim, C. Yanofsky, Translational coupling during expression of the tryptophan operon of Escherichia coli.* Genetics*. **95**, 785–795 (1980).*
13. *V. S. Sethi, S. V.Sagar, W. Zillig, H. Bauer, Dissociation of DNA-dependent RNA-polymerase from E. coli in lithium chloride.* FEBS Lett. ***6**, 339–342 (1970).*
14. *V. S. Sethi, S. V.Sagar, W. Zillig, H. Bauer, Dissociation and reconstitution of active DNA-dependent RNA-polymerase from E. coli.* FEBS Lett. ***8**, 236–239 (1970).*
15. *M. J. Chamberlin, The Selectivity of Transcription.* Annu. Rev. Biochem. ***43**, 721–775 (1974).*
16. *F. Jacob, J. Monod, On the Regulation of Gene Activity.* Cold Spring Harb. Symp. Quant. Biol. ***26**, 193–211 (1961).*
17. *F. Jacob, J. Monod, Genetic regulatory mechanisms in the synthesis of proteins.* J. Mol. Biol. ***3**, 318–356 (1961).*
18. *G. Moreno-Hagelsieb, V. Treviño, E. Pérez-Rueda, T. F. Smith, J. Collado-Vides, Transcription unit conservation in the three domains of life: a perspective from Escherichia coli.* Trends Genet. ***17**, 175–177 (2001).*

19. *H. N. Lim, Y. Lee, R. Hussein, Fundamental relationship between operon organization and gene expression.* Proc. Natl. Acad. Sci. U. S. A. ***108**, 10626–10631 (2011).*

20. *M. Güell* et al.*, Transcriptome Complexity in a Genome-Reduced Bacterium.* Science*. **326**, 1268–1271 (2009).*

21. *S. Li, X. Dong, Z. Su, Directional RNA-seq reveals highly complex condition-dependent transcriptomes in E. coli K12 through accurate full-length transcripts assembling.* BMC Genomics*. **14**, 520 (2013).*

22. *B.-K. Cho* et al.*, The transcription unit architecture of the Escherichia coli genome.* Nat. Biotechnol. ***27**, 1043–1049 (2009).*

23. *I. Junier, E. B. Unal, E. Yus, V. Lloréns-Rico, L. Serrano, Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium.* Cell Systems *(2016), doi:10.1016/j.cels.2016.04.015.*

24. *T. M. Gruber, C. A. Gross, Multiple sigma subunits and the partitioning of bacterial transcription space.* Annu. Rev. Microbiol. ***57**, 441–466 (2003).*

25. *D. Pribnow, Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter.* Proceedings of the National Academy of Sciences*. **72**, 784–788 (1975).*

26. *H. Schaller, C. Gray, K. Herrmann, Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd.* Proc. Natl. Acad. Sci. U. S. A. ***72**, 737–741 (1975).*

27. *M. Djordjevic, Redefining Escherichia coli σ(70) promoter elements: -15 motif as a complement of the -10 motif.* J. Bacteriol. ***193**, 6305–6314 (2011).*

28. *A. G. Sabelnikov, B. Greenberg, S. A. Lacks, An extended -10 promoter alone directs transcription of the DpnII operon of Streptococcus pneumoniae.* J. Mol. Biol. ***250**, 144–155 (1995).*

29. *W. Ross* et al.*, A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase.* Science*. **262**, 1407–1413 (1993).*

30. *K. Mazouni, S. Bulteau, C. Cassier-Chauvat, F. Chauvat, Promoter element spacing controls basal expression and light inducibility of the cyanobacterial secA gene.* Mol. Microbiol. ***30**, 1113–1122 (1998).*

31. *N. Agarwal, A. K. Tyagi, Mycobacterial transcriptional signals: requirements for recognition by RNA polymerase and optimal transcriptional activity.* Nucleic Acids Res. ***34**, 4245–4257 (2006).*

32. *J. E. Stefano, J. D. Gralla, Spacer mutations in the lac ps promoter.* Proc. Natl. Acad. Sci. U. S. A. ***79**, 1069–1072 (1982).*

33. *R. P. Bandwar, S. S. Patel, The energetics of consensus promoter opening by T7 RNA polymerase.* J. Mol. Biol. ***324**, 63–72 (2002).*

34. *H. Margalit, B. A. Shapiro, R. Nussinov, J. Owens, R. L. Jernigan, Helix stability in prokaryotic promoter regions.* Biochemistry*. **27**, 5179–5188 (1988).*

35. *N. Figueroa, N. Wills, L. Bossi, Common sequence determinants of the response of a prokaryotic promoter to DNA bending and supercoiling.* EMBO J. ***10**, 941–949 (1991).*

36. *S. M. Richardson, C. F. Higgins, D. M. Lilley, DNA supercoiling and the leu-500 promoter mutation of Salmonella typhimurium.* EMBO J. ***7**, 1863–1869 (1988).*

37. *M. L. Berman, A. Landy, Promoter mutations in the transfer RNA gene tyrT of Escherichia coli.* Proc. Natl. Acad. Sci. U. S. A. ***76**, 4303–4307 (1979).*

38. *N. Caroff, E. Espaze, D. Gautreau, H. Richet, A. Reynaud, Analysis of the effects of -42 and -32 ampC promoter mutations in clinical isolates of Escherichia coli hyperproducing ampC.* J. Antimicrob. Chemother. ***45**, 783–788 (2000).*

39. *S. Inouye, M. Inouye, Up-promoter mutations in the lpp gene of Escherichia coli.* Nucleic Acids Res. ***13**, 3101–3110 (1985).*

40. *D. M. Heithoff, R. L. Sinsheimer, D. A. Low, M. J. Mahan, An essential role for DNA adenine methylation in bacterial virulence.* Science*.* ***284**, 967–970 (1999).*

41. *M. Lluch-Senar* et al.*, Comprehensive methylome characterization of Mycoplasma genitalium and Mycoplasma pneumoniae at single-base resolution.* PLoS Genet. ***9**, e1003191 (2013).*

42. *A. Ishihama, Functional modulation of Escherichia coli RNA polymerase.* Annu. Rev. Microbiol. ***54**, 499–518 (2000).*

43. *A. D. Grossman, D. B. Straus, W. A. Walter, C. A. Gross, Sigma 32 synthesis can regulate the synthesis of heat shock proteins in Escherichia coli.* Genes Dev. ***1**, 179–184 (1987).*

44. *J. W. Erickson, C. A. Gross, Identification of the sigma E subunit of Escherichia coli RNA polymerase: a second alternate sigma factor involved in high-temperature gene expression.* Genes Dev. ***3**, 1462–1471 (1989).*

45. *M. P. McCann, J. P. Kidwell, A. Matin, The Putative Sigma Factor KatF Has a Central Role in Development of Starvation-Mediated General Resistance in* Escherichia coli. J. Bacteriol. *(1991).*

46. *J. Hirschman, P. K. Wong, K. Sei, J. Keener, S. Kustu, Products of nitrogen regulatory genes ntrA and ntrC of enteric bacteria activate glnA transcription in vitro: evidence that the ntrA product is a sigma factor.* Proc. Natl. Acad. Sci. U. S. A. ***82**, 7525–7529 (1985).*

47. *D. N. Arnosti, M. J. Chamberlin, Secondary sigma factor controls transcription of flagellar and chemotaxis genes in Escherichia coli.* Proceedings of the National Academy of Sciences*.* ***86**, 830–834 (1989).*

48. *M. A. Lonetto, K. L. Brown, K. E. Rudd, M. J. Buttner, Analysis of the Streptomyces coelicolor sigE gene reveals the existence of a subfamily of eubacterial RNA polymerase sigma factors involved in the regulation of extracytoplasmic functions.* Proc. Natl. Acad. Sci. U. S. A. ***91**, 7573–7577 (1994).*

49. *M. Mauri, S. Klumpp, A Model for Sigma Factor Competition in Bacterial Cells.* PLoS Comput. Biol. ***10**, e1003845 (2014).*

50. *S. S. Daube, P. H. von Hippel, Interactions of Escherichia coli 70 within the transcription elongation complex.* Proceedings of the National Academy of Sciences*.* ***96**, 8390–8395 (1999).*

51. *J. D. Gralla, J. Collado-Vides, in* Escherichia coli and Salmonella: Cellular and Molecular Biology*, F. C. Neidhardt, E. C. Lin, R. Curtiss, Eds. (American Society for Microbiology, Washington, DC, 1996), pp. 1232–1245.*
52. *A. Martínez-Antonio, J. Collado-Vides, Identifying global regulators in transcriptional regulatory networks in bacteria.* Curr. Opin. Microbiol. ***6**, 482–489 (2003).*
53. *B. E. Griffin, Separation of 32P-labelled ribonucleic acid components. The use of polyethylenimine-cellulose (TLC) as a second dimension in separating oligoribonucleotides of "4.5 S" and 5 S from E. coli.* FEBS Lett. ***15**, 165–168 (1971).*
54. *J. Hindley, Fractionation of 32P-labelled ribonucleic acids on polyacrylamide gels and their characterization by fingerprinting.* J. Mol. Biol. ***30**, 125–136 (1967).*
55. *J. Livny, M. K. Waldor, Identification of small RNAs in diverse bacterial species.* Curr. Opin. Microbiol. ***10**, 96–101 (2007).*
56. *M. K. Thomason, G. Storz, Bacterial antisense RNAs: how many are there, and what are they doing?* Annu. Rev. Genet. ***44**, 167–188 (2010).*
57. *S. Gottesman, Micros for microbes: non-coding regulatory RNAs in bacteria.* Trends Genet. ***21**, 399–404 (2005).*
58. *M. Lybecker, I. Bilusic, R. Raghavan, Pervasive transcription: detecting functional RNAs in bacteria.* Transcription*.* ***5**, e944039 (2014).*
59. *R. Raghavan, D. B. Sloan, H. Ochman, Antisense transcription is pervasive but rarely conserved in enteric bacteria.* MBio*.* ***3** (2012), doi:10.1128/mBio.00156-12.*
60. *P. Horvath, R. Barrangou, CRISPR/Cas, the immune system of bacteria and archaea.* Science*.* ***327**, 167–170 (2010).*
61. *R. Barrangou* et al.*, CRISPR provides acquired resistance against viruses in prokaryotes.* Science*.* ***315**, 1709–1712 (2007).*
62. *C. L. Beisel, G. Storz, Base pairing small RNAs and their roles in global regulatory networks.* FEMS Microbiol. Rev. ***34**, 866–882 (2010).*
63. *D. D. Sledjeski, C. Whitman, A. Zhang, Hfq is necessary for regulation by the untranslated RNA DsrA.* J. Bacteriol. ***183**, 1997–2005 (2001).*
64. *A. Zhang, K. M. Wassarman, J. Ortega, A. C. Steven, G. Storz, The Sm-like Hfq Protein Increases OxyS RNA Interaction with Target mRNAs.* Mol. Cell*.* ***9**, 11–22 (2002).*
65. *T. Morita, K. Maki, H. Aiba, RNase E-based ribonucleoprotein complexes: mechanical basis of mRNA destabilization mediated by bacterial noncoding RNAs.* Genes Dev. ***19**, 2176–2186 (2005).*
66. *J. A. Opdyke, J. G. Kang, G. Storz, GadY, a Small-RNA Regulator of Acid Response Genes in Escherichia coli.* J. Bacteriol. ***186**, 6698–6705 (2004).*
67. *H. Aiba, Mechanism of RNA silencing by Hfq-binding small RNAs.* Curr. Opin. Microbiol. ***10**, 134–139 (2007).*

68. *J. H. Urban, K. Papenfort, J. Thomsen, R. A. Schmitz, J. Vogel, A conserved small RNA promotes discoordinate expression of the glmUS operon mRNA to activate GlmS synthesis.* J. Mol. Biol. ***373***, *521–528 (2007).*

69. *S. Brantl, E. G. H. Wagner, An antisense RNA-mediated transcriptional attenuation mechanism functions in Escherichia coli.* J. Bacteriol. ***184***, *2740–2747 (2002).*

70. *N. Crampton, W. A. Bonass, J. Kirkham, C. Rivetti, N. H. Thomson, Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy.* Nucleic Acids Res. ***34***, *5416–5425 (2006).*

71. *R. Raghavan, E. A. Groisman, H. Ochman, Genome-wide detection of novel regulatory RNAs in E. coli.* Genome Res. ***21***, *1487–1497 (2011).*

72. *M. Güell* et al.*, Transcriptome Complexity in a Genome-Reduced Bacterium.* Science*.* ***326***, *1268–1271 (2009).*

73. *M. Lluch-Senar* et al.*, Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium.* Mol. Syst. Biol. ***11***, *780–780 (2015).*

74. *T. Gaal, M. S. Bartlett, W. Ross, C. L. Turnbough Jr, R. L. Gourse, Transcription regulation by initiating NTP concentration: rRNA synthesis in bacteria.* Science*.* ***278***, *2092–2097 (1997).*

75. *C. M. Lew, J. D. Gralla, Mechanism of stimulation of ribosomal promoters by binding of the +1 and +2 nucleotides.* J. Biol. Chem. ***279***, *19481–19485 (2004).*

76. *L. Krásný, H. Tiserová, J. Jonák, D. Rejman, H. Sanderová, The identity of the transcription +1 position is crucial for changes in gene expression in response to amino acid starvation in Bacillus subtilis.* Mol. Microbiol. ***69***, *42–54 (2008).*

77. *D. A. Schneider, T. Gaal, R. L. Gourse, NTP-sensing by rRNA promoters in Escherichia coli is direct.* Proc. Natl. Acad. Sci. U. S. A. ***99***, *8602–8607 (2002).*

78. *M. Cashel, Regulation of Bacterial ppGpp and pppGpp.* Annu. Rev. Microbiol. ***29***, *301–318 (1975).*

79. *L. U. Magnusson, A. Farewell, T. Nyström, ppGpp: a global regulator in Escherichia coli.* Trends Microbiol. ***13***, *236–242 (2005).*

80. *D. Chatterji, N. Fujita, A. Ishihama, The mediator for stringent control, ppGpp, binds to the beta-subunit of Escherichia coli RNA polymerase.* Genes Cells*.* ***3***, *279–287 (1998).*

81. *L. Jöres, R. Wagner, Essential steps in the ppGpp-dependent regulation of bacterial ribosomal RNA promoters can be explained by substrate competition.* J. Biol. Chem. ***278***, *16834–16843 (2003).*

82. *B. J. Paul, M. B. Berkmen, R. L. Gourse, DksA potentiates direct activation of amino acid promoters by ppGpp.* Proc. Natl. Acad. Sci. U. S. A. ***102***, *7823–7828 (2005).*

83. *D. R. Gentry, V. J. Hernandez, L. H. Nguyen, D. B. Jensen, M. Cashel, Synthesis of the stationary-phase sigma factor sigma s is positively regulated by ppGpp.* J. Bacteriol. ***175***, *7982–7989 (1993).*

84. *M. Jishage, Regulation of sigma factor competition by the alarmone ppGpp.* Genes Dev. ***16***, *1260–1270 (2002).*

85. *T. M. Wendrich, M. A. Marahiel, Cloning and characterization of a relA/spoT homologue from Bacillus subtilis.* Mol. Microbiol. ***26****, 65–79 (1997).*

86. *H. Nanamiya* et al.*, Identification and functional analysis of novel (p)ppGpp synthetase genes in Bacillus subtilis.* Mol. Microbiol. ***67****, 291–304 (2008).*

87. *K. Ochi, J. Kandala, E. Freese, Evidence that Bacillus subtilis sporulation induced by the stringent response is caused by the decrease in GTP or GDP.* J. Bacteriol. ***151****, 1062–1065 (1982).*

88. *L. Krásný, R. L. Gourse, An alternative strategy for bacterial ribosome synthesis: Bacillus subtilis rRNA transcription regulation.* EMBO J. ***23****, 4473–4483 (2004).*

89. *A. S. Mironov* et al.*, Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria.* Cell. ***111****, 747–756 (2002).*

90. *J. K. Soukup, G. A. Soukup, Riboswitches exert genetic control through metabolite-induced conformational change.* Curr. Opin. Struct. Biol. ***14****, 344–349 (2004).*

91. *A. G. Vitreschak, D. A. Rodionov, A. A. Mironov, M. S. Gelfand, Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation.* Nucleic Acids Res. ***30****, 3141–3151 (2002).*

92. *M.-P. Caron* et al.*, Dual-acting riboswitch control of translation initiation and mRNA decay.* Proc. Natl. Acad. Sci. U. S. A. ***109****, E3444–53 (2012).*

93. *P. J. Farnham, T. Platt, Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro.* Nucleic Acids Res. ***9****, 563–577 (1981).*

94. *C. A. Brennan, A. J. Dombroski, P. Terry, Transcription termination factor rho is an RNA-DNA helicase.* Cell. ***48****, 945–952 (1987).*

95. *J. P. Richardson, Rho-dependent transcription termination.* Biochim. Biophys. Acta. ***1048****, 127–138 (1990).*

96. *T. Platt, Transcription Termination and the Regulation of Gene Expression.* Annu. Rev. Biochem. ***55****, 339–372 (1986).*

97. *W. Morgan, D. G. Bear, B. L. Litchman, P. H. von Hippel, RNA sequence and secondary structure requirements for rho-dependent transcription termination.* Nucleic Acids Res. ***13****, 3739–3754 (1985).*

98. *J. P. Richardson, R. Conaway, Ribonucleic acid release activity of transcription termination protein rho is dependent on the hydrolysis of nucleoside triphosphates.* Biochemistry. ***19****, 4293–4299 (1980).*

99. *B. Cisneros, D. Court, A. Sanchez, C. Montañez, Point mutations in a transcription terminator, λtI, that affect both transcription termination and RNA stability.* Gene. ***181****, 127–133 (1996).*

100. *W. S. Yarnell, J. W. Roberts, Mechanism of intrinsic transcription termination and antitermination.* Science. ***284****, 611–615 (1999).*

101. *S. Phadtare, Recent developments in bacterial cold-shock response.* Curr. Issues Mol. Biol. ***6****, 125–136 (2004).*

102. *W. Bae, B. Xia, M. Inouye, K. Severinov, Escherichia coli CspA-family RNA chaperones are transcription antiterminators.* Proc. Natl. Acad. Sci. U. S. A. ***97**, 7784–7789 (2000).*

103. *N. A. Linderoth, R. L. Calendar, The Psu protein of bacteriophage P4 is an antitermination factor for rho-dependent transcription termination.* J. Bacteriol. ***173**, 6722–6731 (1991).*

104. *A. Ranjan, S. Sharma, R. Banerjee, U. Sen, R. Sen, Structural and mechanistic basis of anti-termination of Rho-dependent transcription termination by bacteriophage P4 capsid protein Psu.* Nucleic Acids Res. ***41**, 6839–6856 (2013).*

105. *M. C. Schmidt, M. J. Chamberlin, nusA protein of Escherichia coli is an efficient transcription termination factor for certain terminator sites.* J. Mol. Biol. ***195**, 809–818 (1987).*

106. *S. Prasch* et al.*, RNA-binding specificity of E. coli NusA.* Nucleic Acids Res. ***37**, 4736–4742 (2009).*

107. *K. B. Arnvig, S. Pennell, B. Gopal, M. J. Colston, A high-affinity interaction between NusA and the rrn nut site in Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences*. **101**, 8325–8330 (2004).*

108. *U. Vogel, K. F. Jensen, NusA is required for ribosomal antitermination and for modulation of the transcription elongation rate of both antiterminated RNA and mRNA.* J. Biol. Chem. ***272**, 12265–12271 (1997).*

109. *B. Py, H. Causton, E. A. Mudd, C. F. Higgins, A protein complex mediating mRNA degradation in Escherichia coli.* Mol. Microbiol. ***14**, 717–729 (1994).*

110. *A. J. Carpousis, Copurification of E. coli RNAase E and PNPase: Evidence for a specific association between two enzymes important in RNA processing and degradation.* Cell*. **76**, 889–900 (1994).*

111. *B. Py, C. F. Higgins, H. M. Krisch, A. J. Carpousis, A DEAD-box RNA helicase in the Escherichia coli RNA degradosome.* Nature*. **381**, 169–172 (1996).*

112. *V. Chandran, B. F. Luisi, Recognition of enolase in the Escherichia coli RNA degradosome.* J. Mol. Biol. ***358**, 8–15 (2006).*

113. *A. J. Carpousis, The RNA degradosome of Escherichia coli: an mRNA-degrading machine assembled on RNase E.* Annu. Rev. Microbiol. ***61**, 71–87 (2007).*

114. *T. Morita, H. Kawamoto, T. Mizota, T. Inada, H. Aiba, Enolase in the RNA degradosome plays a crucial role in the rapid decay of glucose transporter mRNA in the response to phosphosugar stress in Escherichia coli.* Mol. Microbiol. ***54**, 1063–1075 (2004).*

115. *S. Even, Ribonucleases J1 and J2: two novel endoribonucleases in B.subtilis with functional homology to E.coli RNase E.* Nucleic Acids Res. ***33**, 2141–2152 (2005).*

116. *R. A. Britton* et al.*, Maturation of the 5' end of Bacillus subtilis 16S rRNA by the essential ribonuclease YkqC/RNase J1.* Mol. Microbiol. ***63**, 127–138 (2007).*

117. *K. Shahbabian, A. Jamalli, L. Zig, H. Putzer, RNase Y, a novel endoribonuclease, initiates riboswitch turnover in Bacillus subtilis.* EMBO J. ***28**, 3523–3533 (2009).*

118. *M. Lehnik-Habrink* et al.*,* Mol. Microbiol.*, in press.*

119. *F. M. Commichau* et al.*, Novel activities of glycolytic enzymes in Bacillus subtilis: interactions with essential proteins involved in mRNA processing.* Mol. Cell. Proteomics*.* ***8**, 1350–1360 (2009).*

120. *V. R. Kaberdin, L.-C. Sue, Unraveling new roles for minor components of the E. coli RNA degradosome.* RNA Biol. ***6**, 402–405 (2009).*

121. *J. Gao* et al.*, Differential modulation of E. coli mRNA abundance by inhibitory proteins that alter the composition of the degradosome.* Mol. Microbiol. ***61**, 394–406 (2006).*

122. *S. G. Svärd, L. A. Kirsebom, Several regions of a tRNA precursor determine the Escherichia coli RNase P cleavage site.* J. Mol. Biol. ***227**, 1019–1031 (1992).*

123. *A. W. Nicholson, in* Nucleic Acids and Molecular Biology *(2011), pp. 269–297.*

124. *T. Itoh, J. Tomizawa, Initiation of replication of plasmid ColE1 DNA by RNA polymerase, ribonuclease H, and DNA polymerase I.* Cold Spring Harb. Symp. Quant. Biol. ***43 Pt 1**, 409–417 (1979).*

125. *A. Sato, A. Kanai, M. Itaya, M. Tomita, Cooperative regulation for Okazaki fragment processing by RNase HII and FEN-1 purified from a hyperthermophilic archaeon, Pyrococcus furiosus.* Biochem. Biophys. Res. Commun. ***309**, 247–252 (2003).*

126. *R. D. Fleischmann* et al.*, Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.* Science*.* ***269**, 496–512 (1995).*

127. *C. M. Fraser* et al.*, The minimal gene complement of Mycoplasma genitalium.* Science*.* ***270**, 397–403 (1995).*

128. *R. Himmelreich* et al.*, Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae.* Nucleic Acids Res. ***24**, 4420–4449 (1996).*

129. *J. C. Alwine, D. J. Kemp, G. R. Stark, Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.* Proc. Natl. Acad. Sci. U. S. A. ***74**, 5350–5354 (1977).*

130. *E. M. Southern, Detection of specific sequences among DNA fragments separated by gel electrophoresis.* J. Mol. Biol. ***98**, 503–517 (1975).*

131. *L. G. Davis, M. D. Dibner, J. F. Battey, in* Basic Methods in Molecular Biology *(1986), pp. 143–146.*

132. *M. Zheng* et al.*, DNA microarray-mediated transcriptional profiling of the Escherichia coli response to hydrogen peroxide.* J. Bacteriol. ***183**, 4562–4570 (2001).*

133. *M. P. DeLisa, C. F. Wu., L. Wang, J. J. Valdes, W. E. Bentley, DNA Microarray-Based Identification of Genes Controlled by Autoinducer 2-Stimulated Quorum Sensing in Escherichia coli.* J. Bacteriol. ***183**, 5239–5247 (2001).*

134. *T. Kaan, G. Homuth, U. Mäder, J. Bandow, T. Schweder, Genome-wide transcriptional profiling of the Bacillus subtilis cold-shock response.* Microbiology*.* ***148****, 3441–3455 (2002).*

135. *J. A. Bernstein, A. B. Khodursky, P.-H. Lin, S. Lin-Chao, S. N. Cohen, Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays.* Proc. Natl. Acad. Sci. U. S. A. ***99****, 9697–9702 (2002).*

136. *J. A. Bernstein, P.-H. Lin, S. N. Cohen, S. Lin-Chao, Global analysis of Escherichia coli RNA degradosome function using DNA microarrays.* Proc. Natl. Acad. Sci. U. S. A. ***101****, 2758–2763 (2004).*

137. *S. Rimour, D. Hill, C. Militon, P. Peyret, GoArrays: highly dynamic and efficient microarray probe design.* Bioinformatics*.* ***21****, 1094–1103 (2005).*

138. *D. W. Selinger* et al.*, RNA expression analysis using a 30 base pair resolution Escherichia coli genome array.* Nat. Biotechnol. ***18****, 1262–1268 (2000).*

139. *B. Tjaden, Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays.* Nucleic Acids Res. ***30****, 3732–3738 (2002).*

140. *P. Nicolas* et al.*, Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis.* Science*.* ***335****, 1103–1106 (2012).*

141. *T. Maier* et al.*, Quantification of mRNA and protein and integration with protein turnover in a bacterium.* Mol. Syst. Biol. ***7****, 511–511 (2011).*

142. *J. M. Johnson, S. Edwards, D. Shoemaker, E. E. Schadt, Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments.* Trends Genet. ***21****, 93–102 (2005).*

143. *C. P. Scott, V. Jeff, M. Danielle McDonald, D. L. Crawford, Technical Analysis of cDNA Microarrays.* PLoS One*.* ***4****, e4486 (2009).*

144. *B. T. Wilhelm* et al.*, Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.* Nature*.* ***453****, 1239–1243 (2008).*

145. *R. Williams* et al.*, Amplification of complex gene libraries by emulsion PCR.* Nat. Methods*.* ***3****, 545–550 (2006).*

146. *C. Adessi* et al.*, Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms.* Nucleic Acids Res. ***28****, E87 (2000).*

147. *F. Sanger, A. R. Coulson, A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.* J. Mol. Biol. ***94****, 441–448 (1975).*

148. *M. Ronaghi, M. Uhlén, P. Nyrén, A sequencing method based on real-time pyrophosphate.* Science*.* ***281****, 363, 365 (1998).*

149. *T. S. Seo* et al.*, Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides.* Proc. Natl. Acad. Sci. U. S. A. ***102****, 5926–5931 (2005).*

150. *J. Ju* et al.*, Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators.* Proc. Natl. Acad. Sci. U. S. A. ***103****, 19635–19640 (2006).*

151. *V. Pandey, R. C. Nutter, E. Prediger, in* Next Generation Genome Sequencing *(2008), pp. 29–42.*

152. *H. Bayley, Sequencing single molecules of DNA.* Curr. Opin. Chem. Biol. ***10**, 628–637 (2006).*

153. *D. Branton* et al.*, The potential and challenges of nanopore sequencing.* Nat. Biotechnol. ***26**, 1146–1153 (2008).*

154. *N. J. Croucher* et al.*, A simple method for directional transcriptome sequencing using Illumina technology.* Nucleic Acids Res. ***37**, e148 (2009).*

155. *D. Parkhomchuk* et al.*, Transcriptome analysis by strand-specific sequencing of complementary DNA.* Nucleic Acids Res. ***37**, e123 (2009).*

156. *A. P. Vivancos, M. Guell, J. C. Dohm, L. Serrano, H. Himmelbauer, Strand-specific deep sequencing of the transcriptome.* Genome Res. ***20**, 989–999 (2010).*

157. *E. Yus* et al.*, Transcription start site associated RNAs in bacteria.* Mol. Syst. Biol. ***8** (2012), doi:10.1038/msb.2012.16.*

158. *Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics.* Nat. Rev. Genet. ***10**, 57–63 (2009).*

159. *E. L. van Dijk, A. Hélène, J. Yan, T. Claude, Ten years of next-generation sequencing technology.* Trends Genet. ***30**, 418–426 (2014).*

160. *D. R. Yoder-Himes* et al.*, Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing.* Proc. Natl. Acad. Sci. U. S. A. ***106**, 3976–3981 (2009).*

161. *S. M. Kristoffersen* et al.*, Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a Gram-positive bacterium.* Genome Biol. ***13**, R30 (2012).*

162. *C. Kröger* et al.*, The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium.* Proc. Natl. Acad. Sci. U. S. A. ***109**, E1277–86 (2012).*

163. *D. Pellin, P. Miotto, A. Ambrosi, D. M. Cirillo, C. Di Serio, A genome-wide identification analysis of small regulatory RNAs in Mycobacterium tuberculosis by RNA-Seq and conservation analysis.* PLoS One*. **7**, e32723 (2012).*

164. *C. M. Sharma* et al.*, The primary transcriptome of the major human pathogen Helicobacter pylori.* Nature*. **464**, 250–255 (2010).*

165. *J.-P. Schlüter* et al.*, Global mapping of transcription start sites and promoter motifs in the symbiotic α-proteobacterium Sinorhizobium meliloti 1021.* BMC Genomics*. **14**, 156 (2013).*

166. *S. Kosuri* et al.*, Composability of regulatory sequences controlling transcription and translation in Escherichia coli.* Proc. Natl. Acad. Sci. U. S. A. ***110**, 14024–14029 (2013).*

167. *D. Dar* et al.*, Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria.* Science*. **352**, aad9822–aad9822 (2016).*

168. *J. M. Peters, A. D. Vangeloff, L. Robert, Bacterial Transcription Terminators: The RNA 3'-End Chronicles.* J. Mol. Biol. ***412**, 793–813 (2011).*

169. *J. R. Mellin* et al.*, Sequestration of a two-component response regulator by a riboswitch-regulated noncoding RNA.* Science*. **345**, 940–943 (2014).*

170. *B. J. Haas, M. Chin, C. Nusbaum, B. W. Birren, J. Livny, How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes?* BMC Genomics*. **13**, 734 (2012).*

171. *P. P. Łabaj* et al.*, Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling.* Bioinformatics*. **27**, i383–91 (2011).*

172. *E. L. van Dijk, Y. Jaszczyszyn, C. Thermes, Library preparation methods for next-generation sequencing: tone down the bias.* Exp. Cell Res*.* ***322**, 12–20 (2014).*

173. *S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, A. Conesa, Differential expression in RNA-seq: a matter of depth.* Genome Res*. **21**, 2213–2223 (2011).*

174. *M.-H. Kuo, C. D. Allis, In Vivo Cross-Linking and Immunoprecipitation for Studying Dynamic Protein:DNA Associations in a Chromatin Environment.* Methods*. **19**, 425–433 (1999).*

175. *M. J. Solomon, P. L. Larsen, V. Alexander, Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene.* Cell*. **53**, 937–947 (1988).*

176. *Y. Blat, N. Kleckner, Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region.* Cell*. **98**, 249–259 (1999).*

177. *C. E. Horak, M. Snyder, ChIP-chip: a genomic approach for identifying transcription factor binding sites.* Methods Enzymol*.* ***350**, 469–483 (2002).*

178. *G. Robertson* et al.*, Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat. Methods*. **4**, 651–657 (2007).*

179. *S. Pepke, B. Wold, A. Mortazavi, Computation for ChIP-seq and RNA-seq studies.* Nat. Methods*. **6**, S22–32 (2009).*

180. *H. Ji* et al.*, An integrated software system for analyzing ChIP-chip and ChIP-seq data.* Nat. Biotechnol*.* ***26**, 1293–1300 (2008).*

181. *N. B. Reppas, J. T. Wade, G. M. Church, S. Kevin, The Transition between Transcriptional Initiation and Elongation in E. coli Is Highly Variable and Often Rate Limiting.* Mol. Cell*.* ***24**, 747–757 (2006).*

182. *T. T. Perkins* et al.*, ChIP-seq and transcriptome analysis of the OmpR regulon of Salmonella enterica serovars Typhi and Typhimurium reveals accessory genes implicated in host colonization.* Mol. Microbiol*.* ***87**, 526–538 (2013).*

183. *K. J. Minch* et al.*, The DNA-binding network of Mycobacterium tuberculosis.* Nat. Commun*.* ***6**, 5829 (2015).*

184. *P. J. Park, ChIP–seq: advantages and challenges of a maturing technology.* Nat. Rev. Genet*.* ***10**, 669–680 (2009).*

185. *T. Waldminghaus, K. Skarstad, ChIP on Chip: surprising results are often artifacts.* BMC Genomics*. **11**, 414 (2010).*

186. *J. K. Pickrell, D. J. Gaffney, Y. Gilad, J. K. Pritchard, False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions.* Bioinformatics*.* ***27***, *2144–2146 (2011).*

187. *D. Jain* et al.*, Active promoters give rise to false positive "Phantom Peaks" in ChIP-seq experiments.* Nucleic Acids Res. ***43***, *6959–6968 (2015).*

188. *D. N. Perkins, D. J. Pappin, D. M. Creasy, J. S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis*.* ***20***, *3551–3567 (1999).*

189. *D. J. C. Pappin, P. Hojrup, A. J. Bleasby, Rapid identification of proteins by peptide-mass fingerprinting.* Curr. Biol. ***3***, *327–332 (1993).*

190. *M. Mann, P. Højrup, P. Roepstorff, Use of mass spectrometric molecular weight information to identify proteins in sequence databases.* Biol. Mass Spectrom. ***22***, *338–345 (1993).*

191. *P. James, M. Quadroni, E. Carafoli, G. Gonnet, Protein Identification by Mass Profile Fingerprinting.* Biochem. Biophys. Res. Commun. ***195***, *58–64 (1993).*

192. *J. A. Taylor, R. S. Johnson, Sequence database searches via de novo peptide sequencing by tandem mass spectrometry.* Rapid Commun. Mass Spectrom. ***11***, *1067–1075 (1997).*

193. *V. Dančík, T. A. Addona, K. R. Clauser, J. E. Vath, P. A. Pevzner, De Novo Peptide Sequencing via Tandem Mass Spectrometry.* J. Comput. Biol. ***6***, *327–342 (1999).*

194. *A. P. L. Snijders, M. G. J. de Vos, P. C. Wright, Novel approach for peptide quantitation and sequencing based on 15N and 13C metabolic labeling.* J. Proteome Res. ***4***, *578–585 (2005).*

195. *W. M. Old* et al.*, Comparison of label-free methods for quantifying human proteins by shotgun proteomics.* Mol. Cell. Proteomics*.* ***4***, *1487–1502 (2005).*

196. *A. Weber, S. A. Kögl, K. Jung, Time-dependent proteome alterations under osmotic stress during aerobic and anaerobic growth in Escherichia coli.* J. Bacteriol. ***188***, *7165–7175 (2006).*

197. *A. A. Rasmussen* et al.*, Regulation of ompA mRNA stability: the role of a small regulatory RNA in growth phase-dependent control.* Mol. Microbiol. ***58***, *1421–1429 (2005).*

198. *V. van Noort* et al.*, Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium.* Mol. Syst. Biol. ***8***, *571 (2012).*

199. *D. K. Thompson* et al.*, Transcriptional and proteomic analysis of a ferric uptake regulator (fur) mutant of Shewanella oneidensis: possible involvement of fur in energy metabolism, transcriptional regulation, and oxidative stress.* Appl. Environ. Microbiol. ***68***, *881–892 (2002).*

200. *B. Macek* et al.*, Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation.* Mol. Cell. Proteomics*.* ***7***, *299–307 (2008).*

201. *N. Gupta* et al.*, Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation.* Genome Res. ***17**, 1362–1377 (2007).*

202. *D. Gully, D. Moinier, L. Loiseau, E. Bouveret, New partners of acyl carrier protein detected in Escherichia coli by tandem affinity purification.* FEBS Lett. ***548**, 90–96 (2003).*

203. *S. Kuhner* et al.*, Proteome Organization in a Genome-Reduced Bacterium.* Science*.* ***326**, 1235–1240 (2009).*

204. *M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, B. Kuster, Quantitative mass spectrometry in proteomics: a critical review.* Anal. Bioanal. Chem. ***389**, 1017–1031 (2007).*

205. *T. Nilsson* et al.*, Mass spectrometry in high-throughput proteomics: ready for the big time.* Nat. Methods*.* ***7**, 681–685 (2010).*

206. *J. Rappsilber, M. Mann, What does it mean to identify a protein in proteomics?* Trends Biochem. Sci. ***27**, 74–78 (2002).*

207. *C. Kiel* et al.*, Quantification of ErbB network proteins in three cell types using complementary approaches identifies cell-general and cell-type-specific signaling proteins.* J. Proteome Res. ***13**, 300–313 (2014).*

208. *S. H. Yoon, M.-J. Han, S. Y. Lee, K. J. Jeong, J.-S. Yoo, Combined transcriptome and proteome analysis of Escherichia coli during high cell density culture.* Biotechnol. Bioeng. ***81**, 753–767 (2003).*

209. *H. Giladi, S. Koby, M. E. Gottesman, A. B. Oppenheim, Supercoiling, integration host factor, and a dual promoter system, participate in the control of the bacteriophage λ pL promoter.* J. Mol. Biol. ***224**, 937–948 (1992).*

210. *I. I. Goryanin, A. B. Goryachev,* Advances in Systems Biology *(Springer Science & Business Media, 2011).*

211. *H. M. Lodhi, S. H. Muggleton,* Elements of Computational Systems Biology *(John Wiley & Sons, 2010).*

212. *S. Bernard, B. Cajavec, L. Pujo-Menjouet, M. C. Mackey, H. Herzel, Modelling transcriptional feedback loops: the role of Gro/TLE1 in Hes1 oscillations.* Philos. Trans. A Math. Phys. Eng. Sci. ***364**, 1155–1170 (2006).*

213. *J. R. Pomerening, S. Y. Kim, J. E. Ferrell Jr, Systems-level dissection of the cell-cycle oscillator: bypassing positive feedback produces damped oscillations.* Cell*.* ***122**, 565–578 (2005).*

214. *T. J. Perkins, J. Jaeger, J. Reinitz, L. Glass, Reverse engineering the gap gene network of Drosophila melanogaster.* PLoS Comput. Biol. ***2**, e51 (2006).*

215. *W. Koole, M. Tijsterman, Mosaic analysis and tumor induction in zebrafish by microsatellite instability-mediated stochastic gene expression.* Dis. Model. Mech. ***7**, 929–936 (2014).*

216. *C. Lemerle, B. Di Ventura, L. Serrano, Space as the final frontier in stochastic simulations of biological systems.* FEBS Lett. ***579**, 1789–1794 (2005).*

217. *E. Lieberman-Aiden* et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science. ***326****, 289–293 (2009).*

218. *A. Castello* et al., *Comprehensive Identification of RNA-Binding Proteins by RNA Interactome Capture.* Methods Mol. Biol. ***1358****, 131–139 (2016).*

219. *R. Albert, Scale-free networks in cell biology.* J. Cell Sci. ***118****, 4947–4957 (2005).*

220. *T. Ideker, R. Sharan, Protein networks in disease.* Genome Res. ***18****, 644–652 (2008).*

221. *H. Jeong, S. P. Mason, A. L. Barabási, Z. N. Oltvai, Lethality and centrality in protein networks.* Nature. ***411****, 41–42 (2001).*

222. *M. H. Schaefer* et al., *Adding protein context to the human protein-protein interaction network to reveal meaningful interactions.* PLoS Comput. Biol. ***9****, e1002860 (2013).*

223. *M. H. Schaefer, L. Serrano, M. A. Andrade-Navarro, Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types.* Front. Genet. ***6****, 260 (2015).*

224. *H. Kitano, Systems Biology: A Brief Overview.* Science. ***295****, 1662–1664 (2002).*

225. *R. Steuer, Computational approaches to the topology, stability and dynamics of metabolic networks.* Phytochemistry. ***68****, 2139–2151 (2007).*

226. *C. Chassagnole, N. Noisommit-Rizzi, J. W. Schmid, K. Mauch, M. Reuss, Dynamic modeling of the central carbon metabolism of Escherichia coli.* Biotechnol. Bioeng. ***79****, 53–73 (2002).*

227. *D. A. Beard, Q. Hong, in* Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics *(2005).*

228. *A. Bordbar, J. M. Monk, Z. A. King, B. O. Palsson, Constraint-based models predict metabolic and associated cellular functions.* Nat. Rev. Genet. ***15****, 107–120 (2014).*

229. *J. S. Edwards, B. O. Palsson, The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities.* Proc. Natl. Acad. Sci. U. S. A. ***97****, 5528–5533 (2000).*

230. *J. L. Reed, T. D. Vo, C. H. Schilling, B. O. Palsson, An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).* Genome Biol. ***4****, R54 (2003).*

231. *Y.-K. Oh, B. O. Palsson, S. M. Park, C. H. Schilling, R. Mahadevan, Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data.* J. Biol. Chem. ***282****, 28791–28799 (2007).*

232. *P. F. Suthers* et al., *A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189.* PLoS Comput. Biol. ***5****, e1000285 (2009).*

233. *J. A. H. Wodke* et al.*, Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling.* Mol. Syst. Biol. ***9****, 653 (2013).*

234. *E. P. Gianchandani, A. R. Joyce, B. Ø. Palsson, J. A. Papin, Functional states of the genome-scale Escherichia coli transcriptional regulatory system.* PLoS Comput. Biol. ***5****, e1000403 (2009).*

235. *A. R. Joyce, B. Ø. Palsson, The model organism as a system: integrating "omics" data sets.* Nat. Rev. Mol. Cell Biol. ***7****, 198–210 (2006).*

236. *B. Palsson, K. Zengler, The challenges of integrating multi-omic data sets.* Nat. Chem. Biol. ***6****, 787–789 (2010).*

237. *R. Mahadevan, J. S. Edwards, F. J. Doyle 3rd, Dynamic flux balance analysis of diauxic growth in Escherichia coli.* Biophys. J. ***83****, 1331–1340 (2002).*

238. *J. O. Dada, M. Pedro, Multi-scale modelling and simulation in systems biology.* Integr. Biol. ***3****, 86 (2011).*

239. *G. Fuchs* et al.*, 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells.* Genome Biol. ***15****, R69 (2014).*

240. *P. J. Lewis, S. D. Thaker, J. Errington, Compartmentalization of transcription and translation in Bacillus subtilis.* EMBO J. ***19****, 710–718 (2000).*

241. *P. R. Cook, Predicting three-dimensional genome structure from transcriptional activity.* Nat. Genet. ***32****, 347–352 (2002).*

242. *M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, B. O. Palsson, Integrating high-throughput and computational data elucidates bacterial networks.* Nature*.* ***429****, 92–96 (2004).*

243. *I. Thiele, N. Jamshidi, R. M. T. Fleming, B. Ø. Palsson, Genome-Scale Reconstruction of Escherichia coli's Transcriptional and Translational Machinery: A Knowledge Base, Its Mathematical Formulation, and Its Functional Characterization.* PLoS Comput. Biol. ***5****, e1000312 (2009).*

244. *J. R. Karr* et al.*, A Whole-Cell Computational Model Predicts Phenotype from Genotype.* Cell*.* ***150****, 389–401 (2012).*

245. *M. Lluch-Senar, M. Vallmitjana, E. Querol, J. Piñol, A new promoterless reporter vector reveals antisense transcription in Mycoplasma genitalium.* Microbiology*.* ***153****, 2743–2752 (2007).*

246. *F. R. Blattner* et al.*, The complete genome sequence of Escherichia coli K-12.* Science*.* ***277****, 1453–1462 (1997).*

247. *D. G. F. F. Edward, E. A. Freundt, Proposal for Mollicutes as name of the class established for the order Mycoplasmatales.* Int. J. Syst. Bacteriol. ***17****, 267–268 (1967).*

248. *S. Razin, D. Yogev, Y. Naot, Molecular biology and pathogenicity of mycoplasmas.* Microbiol. Mol. Biol. Rev. ***62****, 1094–1156 (1998).*

249. *E. Yus* et al.*, Impact of Genome Reduction on Bacterial Metabolism and Its Regulation.* Science*.* ***326****, 1263–1268 (2009).*

250. *M. K. Cho, D. Magnus, A. L. Caplan, D. McGee, Policy forum: genetics. Ethical considerations in synthesizing a minimal genome.* Science*.* ***286***, *2087, 2089–90 (1999).*

251. *K. B. Waites, D. F. Talkington, Mycoplasma pneumoniae and Its Role as a Human Pathogen.* Clin. Microbiol. Rev. ***17***, *697–728 (2004).*

252. *J. A. H. Wodke* et al.*, MyMpn: a database for the systems biology model organism Mycoplasma pneumoniae.* Nucleic Acids Res. ***43***, *D618–23 (2015).*

253. *S. Halbedel* et al.*, Transcription in Mycoplasma pneumoniae: analysis of the promoters of the ackA and ldh genes.* J. Mol. Biol. ***371***, *596–607 (2007).*

254. *J. Weiner III, R. Herrmann, G. F. Browning, Transcription in Mycoplasma pneumoniae.* Nucleic Acids Res. ***28***, *4488–4496 (2000).*

255. *S. Torres-Puig, A. Broto, E. Querol, J. Piñol, O. Q. Pich, A novel sigma factor reveals a unique regulon controlling cell-specific recombination in Mycoplasma genitalium.* Nucleic Acids Res. ***43***, *4923–4936 (2015).*

256. *P. V. Mazin* et al.*, Transcriptome analysis reveals novel regulatory mechanisms in a genome-reduced bacterium.* Nucleic Acids Res. ***42***, *13254–13268 (2014).*

257. *M. E. Csete, J. C. Doyle, Reverse Engineering of Biological Complexity.* Science*.* ***295***, *1664–1669 (2002).*

258. *P. D. W. Kirk, A. C. Babtie, M. P. H. Stumpf, Systems biology (un)certainties.* Science*.* ***350***, *386–388 (2015).*

259. *J. Cotterell, J. Sharpe, An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients.* Mol. Syst. Biol. ***6***, *425 (2010).*

260. *H. Kitano, Biological robustness.* Nat. Rev. Genet. ***5***, *826–837 (2004).*

261. *C. V. Rao, D. M. Wolf, A. P. Arkin, Control, exploitation and tolerance of intracellular noise.* Nature*.* ***420***, *231–237 (2002).*

262. *Y. Dublanche, K. Michalodimitrakis, N. Kümmerer, M. Foglierini, L. Serrano, Noise in transcription negative feedback loops: simulation and experimental analysis.* Mol. Syst. Biol. ***2***, *41 (2006).*

263. *Z. Zhang, W. Qian, J. Zhang, Positive selection for elevated gene expression noise in yeast.* Mol. Syst. Biol. ***5***, *299 (2009).*

264. *A. J. Lotka, Fluctuations in the Abundance of a Species considered Mathematically.* Nature*.* ***118***, *558–560 (1927).*

265. *V. Volterra, Variations and Fluctuations of the Number of Individuals in Animal Species living together.* ICES J. Mar. Sci. ***3***, *3–51 (1928).*

266. *M. Levitt, A. Warshel, Computer simulation of protein folding.* Nature*.* ***253***, *694–698 (1975).*

267. *D. Noble, Modeling the heart--from genes to cells to the whole organ.* Science*.* ***295***, *1678–1682 (2002).*

268. *R. Brette* et al.*, Simulation of networks of spiking neurons: a review of tools and strategies.* J. Comput. Neurosci. ***23***, *349–398 (2007).*

269. *J. Macia* et al.*, Dynamic signaling in the Hog1 MAPK pathway relies on high basal signal transduction.* Sci. Signal. ***2****, ra13 (2009).*

270. *M. Tomita* et al., *E-CELL: software environment for whole-cell simulation.* Bioinformatics*.* ***15****, 72–84 (1999).*

271. *K. Takahashi, K. Kaizu, B. Hu, M. Tomita, A multi-algorithm, multi-timescale method for cell simulation.* Bioinformatics*.* ***20****, 538–546 (2004).*

272. *S. Seto, G. Layh-Schmitt, T. Kenri, M. Miyata, Visualization of the attachment organelle and cytadherence proteins of Mycoplasma pneumoniae by immunofluorescence microscopy.* J. Bacteriol. ***183****, 1621–1630 (2001).*

273. *M. Lluch-Senar, E. Querol, J. Piñol, Cell division in a minimal bacterium in the absence of ftsZ.* Mol. Microbiol. ***78****, 278–289 (2010).*

274. *M. Ander* et al.*, SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: analysis of simple networks.* Syst. Biol. *.* ***1****, 129–138 (2004).*

275. *J. L. Snoep, F. Bruggeman, B. G. Olivier, H. V. Westerhoff, Towards building the silicon cell: a modular approach.* Biosystems*.* ***83****, 207–216 (2006).*

276. *B. Teusink* et al.*, Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry.* Eur. J. Biochem. ***267****, 5313–5329 (2000).*

277. *G. R. Cronwright, J. M. Rohwer, B. A. Prior, Metabolic Control Analysis of Glycerol Synthesis in Saccharomyces cerevisiae.* Appl. Environ. Microbiol. ***68****, 4448–4456 (2003).*

278. *A. M. Martins, P. Mendes, C. Cordeiro, A. P. Freire, In situ kinetic analysis of glyoxalase I and glyoxalase II in Saccharomyces cerevisiae.* Eur. J. Biochem. ***268****, 3930–3936 (2001).*

279. *J. R. Karr, J. C. Sanghvi, D. N. Macklin, A. Arora, M. W. Covert, WholeCellKB: model organism databases for comprehensive whole-cell models.* Nucleic Acids Res. ***41****, D787–92 (2013).*

280. *D. N. Macklin, N. A. Ruggero, M. W. Covert, The future of whole-cell modeling.* Curr. Opin. Biotechnol. ***28****, 111–115 (2014).*

281. *V. Lloréns-Rico, M. Lluch-Senar, L. Serrano, Distinguishing between productive and abortive promoters using a random forest classifier in Mycoplasma pneumoniae.* Nucleic Acids Res. ***43****, 3442–3453 (2015).*

282. *T. Maier* et al.*, Large-scale metabolome analysis and quantitative integration with genomics and proteomics data in Mycoplasma pneumoniae.* Mol. Biosyst. ***9****, 1743–1755 (2013).*

283. *M. Lluch-Senar* et al.*, Rescuing discarded spectra: Full comprehensive analysis of a minimal proteome.* Proteomics*.* ***16****, 554–563 (2016).*

284. *C. Trapnell* et al.*, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat. Biotechnol. ***28****, 511–515 (2010).*

285. *A. Chang* et al.*, BRENDA in 2015: exciting developments in its 25th year of existence.* Nucleic Acids Res. ***43****, D439–46 (2015).*

286. *T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nat. Methods*.* ***8****, 785–786 (2011).*

287. *J. Van Durme* et al.*, Accurate prediction of DnaK-peptide binding via homology modelling and experimental data.* PLoS Comput. Biol. ***5****, e1000475 (2009).*

288. *M. J. Kerner* et al.*, Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli.* Cell. ***122****, 209–220 (2005).*

289. *C. Hames, S. Halbedel, M. Hoppert, J. Frey, J. Stülke, Glycerol metabolism is important for cytotoxicity of Mycoplasma pneumoniae.* J. Bacteriol. ***191****, 747–753 (2009).*

290. *J. R. Karr* et al.*, Summary of the DREAM8 Parameter Estimation Challenge: Toward Parameter Identification for Whole-Cell Models.* PLoS Comput. Biol. ***11****, e1004096 (2015).*

291. *J. C. Sanghvi* et al.*, Accelerated discovery via a whole-cell model.* Nat. Methods*.* ***10****, 1192–1195 (2013).*

292. *R. Young, H. Bremer, Polypeptide-chain-elongation rate in Escherichia coli B/r as a function of growth rate.* Biochem. J*.* ***160****, 185–194 (1976).*

293. *E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, A. van Oudenaarden, Regulation of noise in the expression of a single gene.* Nat. Genet. ***31****, 69–73 (2002).*

294. *H. Kobayashi, A proton-translocating ATPase regulates pH of the bacterial cytoplasm.* J. Biol. Chem. ***260****, 72–76 (1985).*

295. *R. W. Hutkins, N. L. Nannen, pH Homeostasis in Lactic Acid Bacteria.* J. Dairy Sci. ***76****, 2354–2365 (1992).*

296. *N. Homeyer, T. Essigke, G. M. Ullmann, H. Sticht, Effects of histidine protonation and phosphorylation on histidine-containing phosphocarrier protein structure, dynamics, and physicochemical properties.* Biochemistry*.* ***46****, 12314–12326 (2007).*

297. *M. A. Umbarger* et al.*, The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation.* Mol. Cell. ***44****, 252–264 (2011).*

298. *C. G. Kurland, Translational accuracy and the fitness of bacteria.* Annu. Rev. Genet. ***26****, 29–50 (1992).*

299. *O. Purcell, B. Jain, J. R. Karr, M. W. Covert, T. K. Lu, Towards a whole-cell modeling approach for synthetic biology.* Chaos*.* ***23****, 025112 (2013).*

300. *E. Park, B. Williams, B. J. Wold, A. Mortazavi, RNA editing in the human ENCODE RNA-seq data.* Genome Res. ***22****, 1626–1633 (2012).*

301. *T. Melcher* et al.*, A mammalian RNA editing enzyme.* Nature*.* ***379****, 460–464 (1996).*

302. *M. Schaub, W. Keller, RNA editing by adenosine deaminases generates RNA and protein diversity.* Biochimie*.* ***84****, 791–803 (2002).*

303. *S. Djebali* et al.*, Evidence for Transcript Networks Composed of Chimeric RNAs in Human Cells.* PLoS One*.* ***7****, e28213 (2012).*

304. *T. R. Gingeras, Implications of chimaeric non-co-linear transcripts.* Nature. ***461**, 206–211 (2009).*

305. *O. Delattre* et al., *Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours.* Nature. ***359**, 162–165 (1992).*

306. *M. Soda* et al., *Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer.* Nature. ***448**, 561–566 (2007).*

307. *C. A. Maher* et al., *Transcriptome sequencing to detect gene fusions in cancer.* Nature. ***458**, 97–101 (2009).*

308. *J. Sugahara, N. Yachie, K. Arakawa, M. Tomita, In silico screening of archaeal tRNA-encoding genes having multiple introns with bulge-helix-bulge splicing motifs.* RNA. ***13**, 671–681 (2007).*

309. *S. R. Salgia, S. K. Singh, P. Gurha, R. Gupta, Two reactions of Haloferax volcanii RNA splicing enzymes: joining of exons and circularization of introns.* RNA. ***9**, 319–330 (2003).*

310. *L. Randau, D. Söll, Transfer RNA genes in pieces.* EMBO Rep. ***9**, 623–628 (2008).*

311. *L. Randau, RNA processing in the minimal organism Nanoarchaeum equitans.* Genome Biol. ***13**, R63 (2012).*

312. *R. H. Herai, M. E. B. Yamagishi, Detection of human interchromosomal trans-splicing in sequence databanks.* Brief. Bioinform. ***11**, 198–209 (2010).*

313. *K. Kannan* et al., *Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing.* Proc. Natl. Acad. Sci. U. S. A. ***108**, 9172–9177 (2011).*

314. *G. Parra* et al., *Tandem chimerism as a means to increase protein complexity in the human genome.* Genome Res. ***16**, 37–44 (2006).*

315. *P. Akiva* et al., *Transcription-mediated gene fusion in the human genome.* Genome Res. ***16**, 30–36 (2006).*

316. *M. Frenkel-Morgenstern* et al., *Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts.* Genome Res. ***22**, 1231–1242 (2012).*

317. *M. Frenkel-Morgenstern, A. Valencia, Novel domain combinations in proteins encoded by chimeric transcripts.* Bioinformatics. ***28**, i67–74 (2012).*

318. *R. W. Francis* et al., *FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data.* PLoS One. ***7**, e39987 (2012).*

319. *H. Ge* et al., *FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution.* Bioinformatics. ***27**, 1922–1928 (2011).*

320. *C. Liu, J. Ma, C. J. Chang, X. Zhou, FusionQ: a novel approach for gene fusion detection and quantification from paired-end RNA-Seq.* BMC Bioinformatics. ***14**, 193 (2013).*

321. *Y. Li, J. Chien, D. I. Smith, J. Ma, FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq.* Bioinformatics. ***27**, 1708–1710 (2011).*

322. *K. Wang* et al., *MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.* Nucleic Acids Res. ***38**, e178 (2010).*

323. *A. Dobin* et al.*, STAR: ultrafast universal RNA-seq aligner.* Bioinformatics*. **29**, 15–21 (2013).*

324. *M. K. Iyer, A. M. Chinnaiyan, C. A. Maher, ChimeraScan: a tool for identifying chimeric transcription in sequencing data.* Bioinformatics*. **27**, 2903–2904 (2011).*

325. *D. Kim, S. L. Salzberg, TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.* Genome Biol. ***12**, R72 (2011).*

326. *B. J. Haas* et al.*, Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons.* Genome Res. ***21**, 494–504 (2011).*

327. *G. Doose* et al.*, Mapping the RNA-Seq trash bin: unusual transcripts in prokaryotic transcriptome sequencing data.* RNA Biol. ***10**, 1204–1210 (2013).*

328. *C. K. Stover* et al.*, Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen.* Nature*. **406**, 959–964 (2000).*

329. *D. D. Jima* et al.*, Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs.* Blood*. **116**, e118–e127 (2010).*

330. *S. Pfeffer, M. Lagos-Quintana, T. Tuschl, in* Current Protocols in Molecular Biology *(2003).*

331. *Z. Li* et al.*, RNA-Seq improves annotation of protein-coding genes in the cucumber genome.* BMC Genomics*. **12**, 540 (2011).*

332. *T. R. Mercer* et al.*, Targeted RNA sequencing reveals the deep complexity of the human transcriptome.* Nat. Biotechnol. ***30**, 99–104 (2012).*

333. *T. T. Perkins* et al.*, A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi.* PLoS Genet. ***5**, e1000569 (2009).*

334. *B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol. ***10**, R25 (2009).*

335. *S. Marco-Sola, M. Sammeth, R. Guigó, P. Ribeca, The GEM mapper: fast, accurate and versatile alignment by filtration.* Nat. Methods*. **9**, 1185–1188 (2012).*

336. *M. Kircher, S. Sawyer, M. Meyer, Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform.* Nucleic Acids Res. ***40**, e3 (2012).*

337. *J. Cocquet, A. Chong, G. Zhang, R. A. Veitia, Reverse transcriptase template switching and false alternative transcripts.* Genomics*. **88**, 127–131 (2006).*

338. *H. Edgren* et al.*, Identification of fusion genes in breast cancer by paired-end RNA-sequencing.* Genome Biol. ***12**, R6 (2011).*

339. *F. Abate* et al.*, Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model.* Bioinformatics*. **28**, 2114–2121 (2012).*

340. *E. V. Koonin, Y. I. Wolf, Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world.* Nucleic Acids Res. ***36**, 6688–6719 (2008).*

341. *M. Güell, E. Yus, M. Lluch-Senar, L. Serrano, Bacterial transcriptomics: what is beyond the RNA horiz-ome?* Nat. Rev. Microbiol. ***9**, 658–669 (2011).*

342. *K. D. Passalacqua* et al.*, Structure and Complexity of a Bacterial Transcriptome.* J. Bacteriol. ***191**, 3203–3211 (2009).*

343. *F. Jacob, D. Perrin, C. Sanchez, J. Monod, [Operon: a group of genes with the expression coordinated by an operator].* C. R. Hebd. Seances Acad. Sci. ***250**, 1727–1729 (1960).*

344. *Q.-Z. Li, H. Lin, The recognition and prediction of sigma70 promoters in Escherichia coli K-12.* J. Theor. Biol. ***242**, 135–141 (2006).*

345. *T. J. Todt* et al.*, Genome-wide prediction and validation of sigma70 promoters in Lactobacillus plantarum WCFS1.* PLoS One*.* ***7**, e45097 (2012).*

346. *H. Jarmer* et al.*, Sigma A recognition sites in the Bacillus subtilis genome.* Microbiology*.* ***147**, 2417–2424 (2001).*

347. *J. J. Gordon, M. W. Towsey, J. M. Hogan, S. A. Mathews, P. Timms, Improved prediction of bacterial transcription start sites.* Bioinformatics*.* ***22**, 142–148 (2006).*

348. *B. Demeler, G. W. Zhou, Neural network optimization for E. coli promoter prediction.* Nucleic Acids Res. ***19**, 1593–1599 (1991).*

349. *S. de Avila e Silva, S. Echeverrigaray, G. J. L. Gerhardt, BacPP: Bacterial promoter prediction—A tool for accurate sigma-factor specific assignment in enterobacteria.* J. Theor. Biol. ***287**, 92–99 (2011).*

350. *S. Burden, Y.-X. Lin, R. Zhang, Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences.* Bioinformatics*.* ***21**, 601–607 (2005).*

351. *P. B. Horton, M. Kanehisa, An assessment of neural network and statistical approaches for prediction of E. coli promoter sites.* Nucleic Acids Res. ***20**, 4331–4338 (1992).*

352. *R. N. Kalate, S. S. Tambe, B. D. Kulkarni, Artificial neural networks for prediction of mycobacterial promoter sequences.* Comput. Biol. Chem. ***27**, 555–564 (2003).*

353. *A. de Jong, P. Hilco, C. Martijn, O. P. Kuipers, K. Jan, PePPER: a webserver for prediction of prokaryote promoter elements and regulons.* BMC Genomics*.* ***13**, 299 (2012).*

354. *G. Z. Hertz, G. D. Stormo, Escherichia coli promoter sequences: analysis and prediction.* Methods Enzymol. ***273**, 30–42 (1996).*

355. *A. M. Huerta, J. Collado-Vides, Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals.* J. Mol. Biol. ***333**, 261–278 (2003).*

356. *T. Aoyama* et al.*, Essential structure of E. coli promoter: effect of spacer length between the two consensus sequences on promoter function.* Nucleic Acids Res. ***11**, 5855–5864 (1983).*

357. *D. K. Hawley, W. R. McClure, Compilation and analysis of Escherichia coli promoter DNA sequences.* Nucleic Acids Res. ***11**, 2237–2255 (1983).*

358. *M. Voskuil, The -16 region of Bacillus subtilis and other gram-positive bacterial promoters.* Nucleic Acids Res. ***26****, 3584–3590 (1998).*

359. *M. I. Voskuil, K. Voepel, G. H. Chambliss, The -16 region, a vital sequence for the utilization of a promoter in Bacillus subtilis and Escherichia coli.* Mol. Microbiol. ***17****, 271–279 (1995).*

360. *M. J. Kazmierczak, M. Wiedmann, K. J. Boor, Alternative Sigma Factors and Their Roles in Bacterial Virulence.* Microbiol. Mol. Biol. Rev. ***69****, 527–543 (2005).*

361. *C. J. Benham, Energetics of the strand separation transition in superhelical DNA.* J. Mol. Biol. ***225****, 835–847 (1992).*

362. *D. Zhabinskaya, C. J. Benham, Theoretical Analysis of Competing Conformational Transitions in Superhelical DNA.* PLoS Comput. Biol. ***8****, e1002484 (2012).*

363. *S. Lisser, H. Margalit, Determination of common structural features in Escherichia coli promoters by computer analysis.* Eur. J. Biochem. ***223****, 823–830 (1994).*

364. *H. J. Vollenweider, M. Fiandt, W. Szybalski, A relationship between DNA helix stability and recognition sites for RNA polymerase.* Science*.* ***205****, 508–511 (1979).*

365. *N. Olivares-Zavaleta, R. Jáuregui, E. Merino, Genome analysis of Escherichia coli promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated.* Genomics*.* ***87****, 329–337 (2006).*

366. *P. Meysman* et al.*, Structural properties of prokaryotic promoter regions correlate with functional features.* PLoS One*.* ***9****, e88717 (2014).*

367. *V. Rangannan, M. Bansal, Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition.* Mol. Biosyst. ***5****, 1758–1769 (2009).*

368. *V. Rangannan, M. Bansal, Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability.* J. Biosci. ***32****, 851–862 (2007).*

369. *A. Kanhere, M. Bansal, A novel method for prokaryotic promoter prediction based on DNA stability.* BMC Bioinformatics*.* ***6****, 1 (2005).*

370. *V. Rangannan, M. Bansal, High-quality annotation of promoter regions for 913 bacterial genomes.* Bioinformatics*.* ***26****, 3043–3050 (2010).*

371. *H. Wang, Stress-Induced DNA Duplex Destabilization (SIDD) in the E. coli Genome: SIDD Sites Are Closely Associated With Promoters.* Genome Res. ***14****, 1575–1584 (2004).*

372. *R. R. Mallios, D. M. Ojcius, D. H. Ardell, An iterative strategy combining biophysical criteria and duration hidden Markov models for structural predictions of Chlamydia trachomatis sigma66 promoters.* BMC Bioinformatics*.* ***10****, 271 (2009).*

373. *C. Bland, A. S. Newsome, A. A. Markovets, Promoter prediction in E. coli based on SIDD profiles and Artificial Neural Networks.* BMC Bioinformatics*.* ***11****, S17 (2010).*

374. *A. Askary* et al.*, N4: a precise and highly sensitive promoter predictor using neural network fed by nearest neighbors.* Genes Genet. Syst. ***84****, 425–430 (2009).*

375. *C. Bustamante, S. B. Smith, J. Liphardt, D. Smith, Single-molecule studies of DNA mechanics.* Curr. Opin. Struct. Biol. ***10****, 279–285 (2000).*

376. *M. Rief, H. Clausen-Schaumann, H. E. Gaub, Sequence-dependent mechanics of single DNA molecules.* Nat. Struct. Biol. ***6****, 346–349 (1999).*

377. *U. Ohler, H. Niemann, Liao Gc, G. M. Rubin, Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition.* Bioinformatics*.* **17 Suppl 1***, S199–206 (2001).*

378. *H. Wang, C. J. Benham, Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress.* BMC Bioinformatics*.* ***7****, 248 (2006).*

379. *L. Breiman, Random Forest.* Mach. Learn. ***45****, 5–32 (2001).*

380. *T. G. Dietterich, in* Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21-23, 2000 Proceedings*, J. Kittler, F. Roli, Eds. (Springer Science & Business Media, 2000), pp. 1–15.*

381. *T. L. Bailey, C. Elkan, University of California, San Diego. Dept. of Computer Science and Engineering,* Fitting a mixture model by expectation maximization to discover motifs in bipolymers *(1994).*

382. *T. Ishii, K. Yoshida, G. Terai, Y. Fujita, K. Nakai, DBTBS: a database of Bacillus subtilis promoters and transcription factors.* Nucleic Acids Res. ***29****, 278–280 (2001).*

383. *J. SantaLucia, D. Hicks, The Thermodynamics of DNA Structural Motifs.* Annu. Rev. Biophys. Biomol. Struct. ***33****, 415–440 (2004).*

384. *M. C. Graves, J. C. Rabinowitz, In vivo and in vitro transcription of the Clostridium pasteurianum ferredoxin gene. Evidence for "extended" promoter elements in gram-positive organisms.* J. Biol. Chem. ***261****, 11409–11415 (1986).*

385. *C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif.* Bioinformatics*.* ***27****, 1017–1018 (2011).*

386. *T. M. Oshiro, P. S. Perez, J. A. Baranauskas, in* Lecture Notes in Computer Science *(2012), pp. 154–168.*

387. *J. T. Robinson* et al.*, Integrative genomics viewer.* Nat. Biotechnol. ***29****, 24–26 (2011).*

388. *A. Mendoza-Vargas* et al.*, Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. coli.* PLoS One*.* ***4****, e7526 (2009).*

389. *S. Altuvia, Identification of bacterial small non-coding RNAs: experimental approaches.* Curr. Opin. Microbiol. ***10****, 257–261 (2007).*

390. *F. Repoila, F. Darfeuille, Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects.* Biol. Cell. ***101***, *117–131 (2009).*

391. *J. T. Wade, D. C. Grainger, Pervasive transcription: illuminating the dark matter of bacterial transcriptomes.* Nat. Rev. Microbiol. ***12***, *647–653 (2014).*

392. *S. Gottesman, G. Storz, Bacterial small RNA regulators: versatile roles and rapidly evolving variations.* Cold Spring Harb. Perspect. Biol. ***3*** *(2011), doi:10.1101/cshperspect.a003798.*

393. *G. Storz, J. Vogel, K. M. Wassarman, Regulation by small RNAs in bacteria: expanding frontiers.* Mol. Cell. ***43***, *880–891 (2011).*

394. *J. Gripenland* et al.*, RNAs: regulators of bacterial virulence.* Nat. Rev. Microbiol. ***8***, *857–866 (2010).*

395. *K. Papenfort, J. Vogel, Regulatory RNA in Bacterial Pathogens.* Cell Host Microbe. ***8***, *116–127 (2010).*

396. *S. Chabelskaya, O. Gaillot, B. Felden, A Staphylococcus aureus small RNA is required for bacterial virulence and regulates the expression of an immune-evasion molecule.* PLoS Pathog. ***6***, *e1000927 (2010).*

397. *J. Georg, W. R. Hess, cis-antisense RNA, another level of gene regulation in bacteria.* Microbiol. Mol. Biol. Rev. ***75***, *286–300 (2011).*

398. *M. Tijsterman, R. H. A. Plasterk, Dicers at RISC; the mechanism of RNAi.* Cell. ***117***, *1–3 (2004).*

399. *Z. Zebec, A. Manica, J. Zhang, M. F. White, C. Schleper, CRISPR-mediated targeted mRNA degradation in the archaeon Sulfolobus solfataricus.* Nucleic Acids Res. ***42***, *5280–5288 (2014).*

400. *A. Hüttenhofer, P. Schattner, N. Polacek, Non-coding RNAs: hope or hype?* Trends Genet. ***21***, *289–297 (2005).*

401. *F. F. Costa, Non-coding RNAs: lost in translation?* Gene. ***386***, *1–10 (2007).*

402. *R. Hershberg, D. A. Petrov, Evidence that mutation is universally biased towards AT in bacteria.* PLoS Genet. ***6***, *e1001115 (2010).*

403. *P. Mehta, S. Goyal, N. S. Wingreen, A quantitative comparison of sRNA-based and protein-based gene regulation.* Mol. Syst. Biol. ***4***, *221 (2008).*

404. *P. Ortet, G. De Luca, D. E. Whitworth, M. Barakat, P2TF: a comprehensive resource for analysis of prokaryotic transcription factors.* BMC Genomics. ***13***, *628 (2012).*

405. *E. Pérez-Rueda, J. Collado-Vides, L. Segovia, Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea.* Comput. Biol. Chem. ***28***, *341–350 (2004).*

406. *Y. Taniguchi* et al.*, Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells.* Science. ***329***, *533–538 (2010).*

407. *S. Legewie, D. Dienst, A. Wilde, H. Herzel, I. M. Axmann, Small RNAs establish delays and temporal thresholds in gene expression.* Biophys. J. ***95***, *3232–3238 (2008).*

408. *Y. Shimoni* et al.*, Regulation of gene expression by small non-coding RNAs: a quantitative view.* Mol. Syst. Biol. ***3***, *138 (2007).*

409. *E. Levine, T. Hwa, Small RNAs establish gene expression thresholds.* Curr. Opin. Microbiol. ***11****, 574–579 (2008).*

410. *A. F. Palazzo, E. S. Lee, Non-coding RNA: what is functional and what is junk?* Front. Genet. ***6****, 2 (2015).*

411. *A. Lamelas* et al.*, Serratia symbiotica from the Aphid Cinara cedri: A Missing Link from Facultative to Obligate Insect Endosymbiont.* PLoS Genet. ***7****, e1002357 (2011).*

412. *D. G. Gibson* et al.*, Creation of a bacterial cell controlled by a chemically synthesized genome.* Science*.* ***329****, 52–56 (2010).*

413. *C. Lartigue* et al.*, Genome transplantation in bacteria: changing one species to another.* Science*.* ***317****, 632–638 (2007).*

414. *J. Z. Levin* et al.*, Comprehensive comparative analysis of strand-specific RNA sequencing methods.* Nat. Methods*.* ***7****, 709–715 (2010).*

415. *H. Li, J. Ruan, R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res. ***18****, 1851–1858 (2008).*

416. *J. E. Dornenburg, A. M. Devita, M. J. Palumbo, J. T. Wade, Widespread antisense transcription in Escherichia coli.* MBio*.* ***1*** *(2010), doi:10.1128/mBio.00024-10.*

417. *E. Levine, Z. Zhang, T. Kuhlman, T. Hwa, Quantitative Characteristics of Gene Regulation by Small RNA.* PLoS Biol. ***5****, e229 (2007).*

418. *S. Hoops* et al.*, COPASI--a COmplex PAthway SImulator.* Bioinformatics*.* ***22****, 3067–3074 (2006).*

419. *V. Chelliah* et al.*, BioModels: ten-year anniversary.* Nucleic Acids Res. ***43****, D542–8 (2015).*

420. *O. Q. Pich, R. Burgos, R. Planell, E. Querol, J. Pinol, Comparative analysis of antibiotic resistance gene markers in Mycoplasma genitalium: application to studies of the minimal gene complement.* Microbiology*.* ***152****, 519–527 (2006).*

421. *M. Hecker, U. Völker, in* Advances in Microbial Physiology *(2001), pp. 35–91.*

422. *A. Y. Mitrophanov, E. A. Groisman, Signal integration in bacterial two-component regulatory systems.* Genes Dev. ***22****, 2601–2611 (2008).*

423. *T. I. Lee* et al.*, Transcriptional Regulatory Networks in Saccharomyces cerevisiae.* Science*.* ***298****, 799–804 (2002).*

424. *F. M. Commichau, J. Stülke, Trigger enzymes: bifunctional proteins active in metabolism and in controlling gene expression.* Mol. Microbiol. ***67****, 692–702 (2008).*

425. *C. J. Jeffery, Why study moonlighting proteins?* Front. Genet. ***6****, 211 (2015).*

426. *S. C. Dillon, C. J. Dorman, Bacterial nucleoid-associated proteins, nucleoid structure and gene expression.* Nat. Rev. Microbiol. ***8****, 185–195 (2010).*

427. *H. Salgado* et al.*, RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more.* Nucleic Acids Res. ***41****, D203–D213 (2013).*

428. *N. Sierro, Y. Makita, M. de Hoon, K. Nakai, DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information.* Nucleic Acids Res. ***36***, *D93–6 (2008).*

429. *S. A. Leyn* et al.*, Genomic reconstruction of the transcriptional regulatory network in Bacillus subtilis.* J. Bacteriol. ***195***, *2463–2473 (2013).*

430. *L. Brinza, F. Calevro, H. Charles, Genomic analysis of the regulatory elements and links with intrinsic DNA structural properties in the shrunken genome of Buchnera.* BMC Genomics. ***14***, *73 (2013).*

431. *V. Llorens-Rico* et al.*, Bacterial antisense RNAs are mainly the product of transcriptional noise.* Science Advances*.* ***2***, *e1501363–e1501363 (2016).*

432. *K. Shimizu, Metabolic Regulation of a Bacterial Cell System with Emphasis on Escherichia coli Metabolism.* ISRN Biochem*.* ***2013***, *645983 (2013).*

433. *S. Berthoumieux* et al.*, Shared control of gene expression in bacteria by transcription factors and global physiology of the cell.* Mol. Syst. Biol. ***9***, *634 (2013).*

434. *S. Klumpp, T. Hwa, Bacterial growth: global effects on gene expression, growth feedback and proteome partition.* Curr. Opin. Biotechnol. ***28***, *96–102 (2014).*

435. *K. Potrykus, M. Cashel, (p)ppGpp: still magical?* Annu. Rev. Microbiol. ***62***, *35–51 (2008).*

436. *D. A. Schneider, R. Wilma, R. L. Gourse, Control of rRNA expression in Escherichia coli.* Curr. Opin. Microbiol. ***6***, *151–156 (2003).*

437. *L. Sojka* et al.*, Rapid changes in gene expression: DNA determinants of promoter regulation by the concentration of the transcription initiating NTP in Bacillus subtilis.* Nucleic Acids Res. ***39***, *4598–4611 (2011).*

438. *I. Junier, O. Rivoire, Conserved Units of Co-Expression in Bacterial Genomes: An Evolutionary Insight into Transcriptional Regulation.* PLoS One*.* ***11***, *e0155740 (2016).*

439. *J. Dekker, M. A. Marti-Renom, L. A. Mirny, Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data.* Nat. Rev. Genet. ***14***, *390–403 (2013).*

440. *G. W. Hatfield, C. J. Benham, DNA topology-mediated control of global gene expression in Escherichia coli.* Annu. Rev. Genet. ***36***, *175–203 (2002).*

441. *A. Travers, G. Muskhelishvili, Bacterial chromatin.* Curr. Opin. Genet. Dev. ***15***, *507–514 (2005).*

442. *L. Baranello, D. Levens, A. Gupta, F. Kouzine, The importance of being supercoiled: how DNA mechanics regulate dynamic processes.* Biochim. Biophys. Acta. ***1819***, *632–638 (2012).*

443. *K. L. Knudtson, F. C. Minion, Construction of Tn4001lac derivatives to be used as promoter probe vectors in mycoplasmas.* Gene*.* ***137***, *217–222 (1993).*

444. *N. Molina, E. van Nimwegen, Scaling laws in functional genome content across prokaryotic clades and lifestyles.* Trends Genet. ***25***, *243–247 (2009).*

445. *E. Pérez-Rueda, S. C. Janga, A. Martínez-Antonio, Scaling relationship in the gene content of transcriptional machinery in bacteria.* Mol. Biosyst. ***5****, 1494–1501 (2009).*

446. *J. Grilli, B. Bassetti, S. Maslov, M. Cosentino Lagomarsino, Joint scaling laws in functional and evolutionary categories in prokaryotic genomes.* Nucleic Acids Res. ***40****, 530–540 (2012).*

447. *R. Burgos, P. A. Totten, MG428 is a novel positive regulator of recombination that triggers mgpB and mgpC gene variation in M ycoplasma genitalium.* Mol. Microbiol. ***94****, 290–306 (2014).*

448. *C. J. Dorman, in* Advances in Applied Microbiology *(2009), pp. 47–64.*

449. *R. C. Souza, D. F. de Almeida, Z. Arnaldo, D. A. de Lima Morais, A. T. R. de Vasconcelos, In search of essentiality: Mollicute-specific genes shared by twelve genomes.* Genet. Mol. Biol. ***30*** *(2007), doi:10.1590/s1415-47572007000200002.*

450. *M. J. Smith, Bacterial Protease Lon Is a Site-specific DNA-binding Protein.* J. Biol. Chem. ***272****, 534–538 (1997).*

451. *D. Charlier* et al.*, Mutational analysis of Escherichia coli PepA, a multifunctional DNA-binding aminopeptidase.* J. Mol. Biol. ***302****, 409–424 (2000).*

452. *E. N. Kozhevnikova* et al.*, Metabolic enzyme IMPDH is also a transcription factor regulated by cellular state.* Mol. Cell. ***47****, 133–139 (2012).*

453. *S. Nakano, M. M. Nakano, Y. Zhang, M. Leelakriangsak, P. Zuber, A regulatory protein that interferes with activator-stimulated transcription in bacteria.* Proc. Natl. Acad. Sci. U. S. A. ***100****, 4233–4238 (2003).*

454. *S. Gruber, J. Errington, Recruitment of Condensin to Replication Origin Regions by ParB/SpoOJ Promotes Chromosome Segregation in B. subtilis.* Cell. ***137****, 685–696 (2009).*

455. *B. R. Glick, Metabolic load and heterologous gene expression.* Biotechnol. Adv. ***13****, 247–261 (1995).*

456. *M. M. Nakano* et al.*, Promoter Recognition by a Complex of Spx and the C-Terminal Domain of the RNA Polymerase α Subunit.* PLoS One*.* ***5****, e8664 (2010).*

457. *G. Y. Fisunov* et al.*, Binding site of MraZ transcription factor in Mollicutes.* Biochimie *(2016), doi:10.1016/j.biochi.2016.02.016.*

458. *S. Gama-Castro* et al.*, RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond.* Nucleic Acids Res. ***44****, D133–D143 (2015).*

459. *M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks.* Phys. Rev. E Stat. Nonlin. Soft Matter Phys. ***69****, 026113 (2004).*

460. *G. Su, A. Kuchinsky, J. H. Morris, D. J. States, F. Meng, GLay: community structure analysis of biological networks.* Bioinformatics*.* ***26****, 3135–3137 (2010).*

461. *S. Großhennig, S. R. Schmidl, G. Schmeisky, J. Busse, J. Stülke, Implication of glycerol and phospholipid transporters in Mycoplasma pneumoniae growth and virulence.* Infect. Immun. ***81****, 896–904 (2013).*

462. *S. Nakano, K. N. Erwin, M. Ralle, P. Zuber, Redox-sensitive transcriptional control by a thiol/disulphide switch in the global regulator, Spx.* Mol. Microbiol. ***55***, *498–510 (2005).*

463. *A. Maxwell, The interaction between coumarin drugs and DNA gyrase.* Mol. Microbiol. ***9***, *681–686 (1993).*

464. *M. L. Webb, S. T. Jacob, Inhibition of RNA polymerase I-directed transcription by novobiocin. Potential use of novobiocin as a general inhibitor of eukaryotic transcription initiation.* J. Biol. Chem. ***263***, *4745–4748 (1988).*

465. *W. Ahmed, S. Menon, P. V. Karthik, V. Nagaraja, Autoregulation of topoisomerase I expression by supercoiling sensitive transcription.* Nucleic Acids Res. ***44***, *1541–1552 (2015).*

466. *D. Magnan, D. Bates, Regulation of DNA Replication Initiation by Chromosome Structure.* J. Bacteriol. ***197***, *3370–3377 (2015).*

467. *S. Zhang, I. Borovok, Y. Aharonowitz, R. Sharan, V. Bafna, A sequence-based filtering method for ncRNA identification and its application to searching for riboswitch elements.* Bioinformatics*.* ***22***, *e557–65 (2006).*

468. *T. Tosa, L. I. Pizer, Biochemical bases for the antimetabolite action of L-serine hydroxamate.* J. Bacteriol. ***106***, *972–982 (1971).*

469. *A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neubock, I. L. Hofacker, The Vienna RNA Websuite.* Nucleic Acids Res. ***36***, *W70–W74 (2008).*

470. *J. Stülke, Control of transcription termination in bacteria by RNA-binding proteins that modulate RNA structures.* Arch. Microbiol. ***177***, *433–440 (2002).*

471. *D. Thieffry, A. M. Huerta, E. Pérez-Rueda, J. Collado-Vides, From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli.* Bioessays*.* ***20***, *433–440 (1998).*

472. *S. S. Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transcriptional regulation network of Escherichia coli.* Nat. Genet. ***31***, *64–68 (2002).*

473. *U. Mäder* et al.*, Staphylococcus aureus Transcriptome Architecture: From Laboratory to Infection-Mimicking Conditions.* PLoS Genet. ***12***, *e1005962 (2016).*

474. *S. Halbedel, J. Stulke, Tools for the genetic analysis of Mycoplasma.* Int. J. Med. Microbiol. ***297***, *37–44 (2007).*

475. *S. C. Prasad, A. Dritschilo, High-resolution two-dimensional electrophoresis of nuclear proteins: a comparison of HeLa nuclei prepared by three different methods.* Anal. Biochem. ***207***, *121–128 (1992).*

476. *M.-C. Keogh, S. Buratowski, in* mRNA Processing and Metabolism *(2004), pp. 001–016.*

477. *M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, T. Ideker, Cytoscape 2.8: new features for data integration and network visualization.* Bioinformatics*.* ***27***, *431–432 (2011).*

478. *K. Masternak* et al.*, CIITA is a transcriptional coactivator that is recruited to MHC class II promoters by multiple synergistic interactions with an enhanceosome complex.* Genes Dev. ***14****, 1156–1166 (2000).*

479. *B. M. Hasselbring, J. L. Jordan, R. W. Krause, D. C. Krause, Terminal organelle development in the cell wall-less bacterium Mycoplasma pneumoniae.* Proc. Natl. Acad. Sci. U. S. A. ***103****, 16478–16483 (2006).*

480. *C. A. Hutchison 3rd* et al.*, Design and synthesis of a minimal bacterial genome.* Science*.* ***351****, aad6253 (2016).*

481. *S. W. Roy, M. Irimia, When good transcripts go bad: artifactual RT-PCR "splicing" and genome analysis.* Bioessays*.* ***30****, 601–605 (2008).*

482. *C. A. Meyer, X. Shirley Liu, Identifying and mitigating bias in next-generation sequencing methods for chromatin biology.* Nat. Rev. Genet. ***15****, 709–721 (2014).*

483. *D. Jain, S. Baldi, A. Zabel, T. Straub, P. B. Becker, Active promoters give rise to false positive "Phantom Peaks" in ChIP-seq experiments.* Nucleic Acids Res. ***43****, 6959–6968 (2015).*

484. *S. Borukhov, J. Lee, O. Laptenko, Bacterial transcription elongation factors: new insights into molecular mechanism of action.* Mol. Microbiol. ***55****, 1315–1324 (2005).*

485. *B. van Steensel, S. Henikoff, Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase.* Nat. Biotechnol. ***18****, 424–428 (2000).*

486. *M. E. Mulligan, D. K. Hawley, E. Robert, W. R. McClure, Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity.* Nucleic Acids Res. ***12****, 789–800 (1984).*

487. *V. A. Rhodius, V. K. Mutalik, Predicting strength and function for promoters of the Escherichia coli alternative sigma factor, E.* Proceedings of the National Academy of Sciences*.* ***107****, 2854–2859 (2010).*

488. *V. A. Rhodius, V. K. Mutalik, C. A. Gross, Predicting the strength of UP-elements and full-length E. coli σE promoters.* Nucleic Acids Res. ***40****, 2907–2924 (2012).*

489. *R. C. Brewster, D. L. Jones, R. Phillips, Tuning promoter strength through RNA polymerase binding site design in Escherichia coli.* PLoS Comput. Biol. ***8****, e1002811 (2012).*

490. *Y. Liu, I. Kobayashi, Negative regulation of the EcoRI restriction enzyme gene is associated with intragenic reverse promoters.* J. Bacteriol. ***189****, 6928–6935 (2007).*

491. *J. A. Brophy, C. A. Voigt, Antisense transcription as a tool to tune gene expression.* Mol. Syst. Biol. ***12****, 854 (2016).*

492. *D. H. Lenz* et al.*, The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae.* Cell*.* ***118****, 69–82 (2004).*

493. *B. Mann* et al.*, Control of Virulence by Small RNAs in Streptococcus pneumoniae.* PLoS Pathog. ***8****, e1002788 (2012).*

494. *C. J. Dorman, DNA supercoiling and bacterial gene expression.* Sci. Prog. ***89****, 151–166 (2006).*

495. *A. Schulz, W. Schumann, hrcA, the first gene of the Bacillus subtilis dnaK operon encodes a negative regulator of class I heat shock genes.* J. Bacteriol. ***178****, 1088–1093 (1996).*

496. *M. Falconi, B. Colonna, G. Prosseda, G. Micheli, C. O. Gualerzi, Thermoregulation of Shigella and Escherichia coli EIEC pathogenicity. A temperature-dependent structural transition of DNA modulates accessibility of virF promoter to transcriptional repressor H-NS.* EMBO J. ***17****, 7033–7043 (1998).*

497. *M. Feig* et al.*, Complete atomistic model of a bacterial cytoplasm for integrating physics, biochemistry, and systems biology.* J. Mol. Graph. Model. ***58****, 1–9 (2015).*