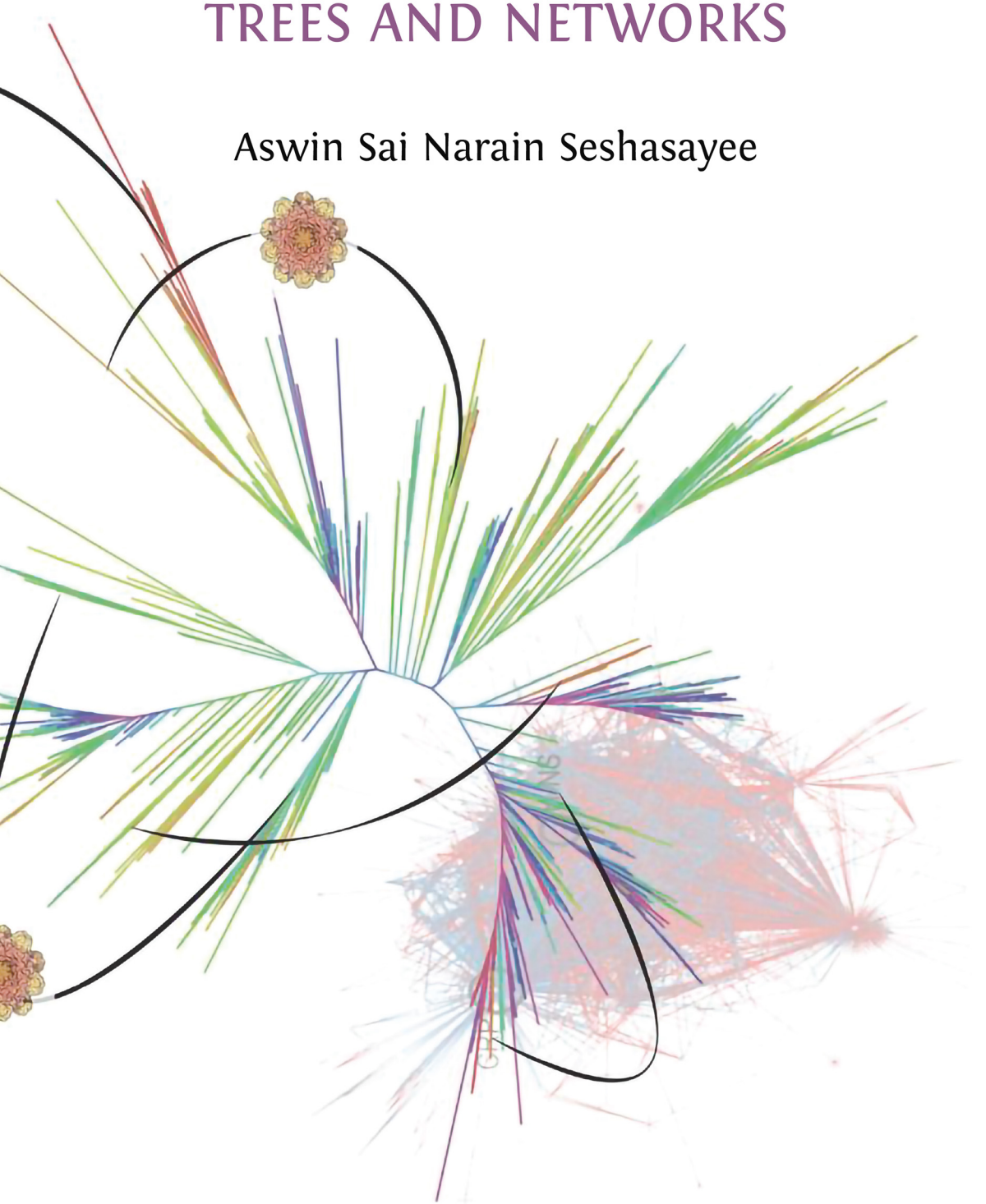
BACTERIAL GENOMES

TREES AND NETWORKS

Aswin Sai Narain Seshasayee



https://www.openbookpublishers.com

Aswin Sai Narain Seshasayee, *Bacterial Genomes: Trees and Networks*. Cambridge, UK: Open Book Publishers, 2025, https://doi.org/10.11647/OBP.0446

ISBN Paperback: 978-1-80511-495-6
ISBN Hardback: 978-1-80511-496-3
ISBN Digital (PDF): 978-1-80511-497-0
ISBN Digital ebook (EPUB): 978-1-80511-498-7
ISBN HTML: 978-1-80511-499-4
DOI: 10.11647/OBP.0446

Cover images by Prerana Sudarshan and David Goodsell, assembled by Ashitha Arun, Inder Raj Singh and Madhumitha K. Cover design: Jeevanjot Kaur Nagpal.

5. Reading and organising the genome

5.1. Expressing the genome and decision making

The genome is a blueprint,[1] and does not by itself get its host cell up and running. The genome must be read and interpreted before it can set in motion many series of connected events that somehow create life. The first step in this process is transcription, during which a gene sequence, a small part of the genome, is read and an RNA transcript with a sequence corresponding to that of the transcribed DNA is produced. Many of these RNA molecules serve as messengers (mRNA), and are further read to create proteins during translation; other RNA molecules, such as the rRNA and tRNA which help the ribosome perform translation, play direct roles in cell function without being translated. The discussion in this chapter will focus on transcription, how it is regulated, how regulators of transcription evolve and the role played by genomics in our understanding of these processes. We will conclude by asking how transcription and the manner in which genes are strung together to form a genome are linked.

Transcription is essentially an enzymatic process that is constrained by the sequence of the DNA being transcribed. The process minimally requires a DNA template, free ribonucleotides that can be linked together to form the RNA chain and an enzyme that can polymerise ribonucleotides to create an RNA sequence that is complementary to the sequence of the DNA template. In addition, the mechanics of transcription requires additional enzymes that help unwind the DNA in front of the machinery that performs transcription, and a host of other proteins that ensure that the process doesn't stall in the middle of a gene and terminates at the right place; these will not be described much in this book. The discovery of the enzyme and that of the fact that transcription is tightly regulated in bacterial cells played important roles in the series of epiphanies that led to the explosion of molecular biology in the 1960s.[2]

In order to transcribe a gene, RNA Polymerase (RNAP), the enzyme that performs transcription, should specifically bind somewhere near the start of the gene. Once this happens, the double-stranded DNA must unwind and the unwound DNA must move

1 A flexible one at that, such that the same sequence can be interpreted in different ways to produce different trait outcomes.

2 J. Hurwitz, 'The Discovery of RNAP', *Journal of Biological Chemistry* 280 (2005), 42477–42485. https://doi.org/10.1074/jbc.x500006200; R.R. Burgess, 'What is in the black box? The discovery of the sigma factor and the subunit structure of *E. coli* RNAP', *Journal of Biological Chemistry* 297 (2021), 101310. https://doi.org/10.1016/j.jbc.2021.101310; M. Lewis, 'A tale of two repressors – a historical perspective', *Journal of Molecular Biology* 409 (2011), 14–27. https://doi.org/10.1016/j.jmb.2011.02.023
See Chapter 2.

https://doi.org/10.11647/OBP.0446.05

base-by-base relative to the RNAP. As the enzyme reads the DNA bases, ribonucleotides complementary to the base being read should be assembled and attached to the growing, nascent RNA chain. The DNA in front of the RNAP must be kept unwound throughout the process. Finally, the RNAP should drop off the gene and terminate transcription at the end of the gene. The focus of this chapter will be on transcription initiation.

The RNAP is a multi-subunit protein,[3] i.e., it comprises several proteins that assemble together to form a functional enzyme. These subunit proteins include those that perform the enzymatic reaction of linking ribonucleotides together, proteins that ensure that the enzyme stays on the DNA through the length of the gene and assembly factors.[4] The core RNAP, which in *E. coli* has five subunits, is perfectly capable of performing transcription but cannot specifically recognise and initiate transcription at the start of genes. Specific recognition of these transcription start sites requires an exchangeable subunit called the σ-factor (sigma factor; Fig. 5.1). The σ-factor binds to the core RNAP, forming what is called the RNAP holoenzyme. The RNAP holoenzyme then specifically recognises DNA sequences upstream of the start of genes. The σ-factor also helps the RNAP unwind the DNA, thus initiating transcription. The σ-factor usually dissociates from the RNAP complex after initiation.

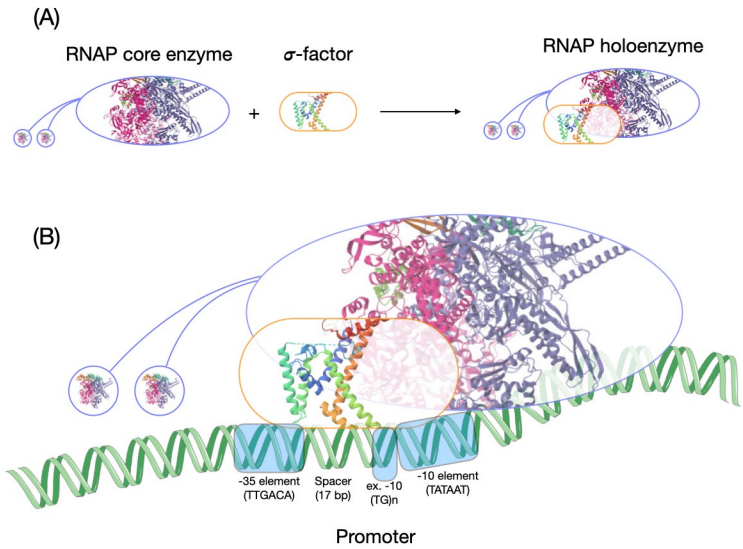


Fig. 5.1. Transcription **initiation**. (A) This figure shows the formation of an RNAP holoenzyme by the binding of the RNAP core enzyme with a σ-factor. The structure of the RNAP inside the oval is from PDB: 7MKP, and that of a fragment of a σ-factor is from PDB: 1SIG. (B) This figure shows the interaction of an RNAP holoenzyme with the promoter. The image of the DNA is from SMART-Servier Medical Art, part of Laboratoires Servier, via Wikimedia Commons, available freely under CC BY-SA 3.0.

3 Note that RNA polymerase from bacteriophage T7 comprises a single subunit. It was discovered about a decade or so after the discovery of the bacterial multi-subunit RNA polymerase.

4 Eukaryotes have multiple types of RNAPs. The eukaryotic RNAP transcribing messenger RNA is larger than the prokaryotic RNAP and has many more subunits.

The DNA sequence that the RNAP holoenzyme recognises is called the promoter. The promoter region is usually A+T-rich. Each gene has its own promoter sequence, but taken together many promoters show some common properties. For example, the bacterial promoter—based on the paradigm established in *E. coli* but shown to be applicable to many other bacterial genomes[5]—is bipartite. There is a six-base −10 element (minus 10) and a six-base −35 element (minus 35). The −10 element is centred 10 bases upstream of the transcription start site (the site at which mRNA synthesis begins) of a gene, and the −35 element is centred 35 bases upstream. The −10 element, when analysed across many genes, has a consensus sequence TATAAT, whereas the −35 element shows a consensus of TTGACA. The specificity-determining σ-factor, when bound to the RNAP, recognises these elements on the DNA. The sequence of the stretch of DNA between the two elements is immaterial. However the length of this spacer is critical to ensure that the −10 and the −35 elements are oriented correctly for the RNAP to bind to the promoter. The precise consensus sequence is not necessary to produce a functional promoter. It is merely a construct that represents the most common base found at each site.

Natural promoters usually differ from the consensus at one or more sites, and the more divergent it is from the consensus element the weaker is its affinity to the RNAP. Therefore, each gene, on the basis of its promoter sequence alone, has its own unique ability to attract RNAP and initiate its own transcription. This creates cross-gene variation in the extent to which a gene can be transcribed. Some promoters do not contain a −35 element, and these sequences carry what is an extended −10 element, which is a slightly longer version of the −10 sequence motif.

Though the sequence of the promoter itself can determine to some extent the expression level of a gene, this does not vary within the lifetime of a cell. Changes in the promoter sequence can happen over generations and, similar to mutations within a gene sequence, its fate can be determined by selection or drift. However, a cell often needs to make decisions in a matter of minutes about which gene to express, and when, within its lifetime. Many bacterial cells experience conditions that change from time to time. Even if their genetic repertoire is sufficient to handle all these environmental conditions, only a subset of their genes would be required under any given condition. Expressing the rest can be costly. As we noted in Chapter 3, expressing a gene under conditions in which the gene offers no selective advantage to the cell can be very costly, especially in bacteria with large population sizes. In addition, there is a constraint that arises from resource availability. The number of free RNAP molecules available to initiate transcription is often limited, because ~80% of all RNAP molecules are involved in transcribing a very small number of genes coding for rRNAs.[6] Therefore,

5 A.M. Huerta, M.P. Francino, E. Morette, and J. Collado-Vides, 'Selection for unequal densities of σ70 promoter-like signals in different regions of large bacterial genomes', *PLoS Genetics* 2 (2006), 185. https://doi.org/10.1371/journal.pgen.0020185

6 D.F. Browning and S.J.W. Busby, 'The regulation of bacterial transcription initiation', *Nature Reviews Microbiology* 2 (2004), 57–65. https://doi.org/10.1038/nrmicro787; I. Bervoets and D. Charlier, 'Diversity, versatility and complexity of bacterial gene regulation mechanisms: opportunities and

the number of RNAP molecules available for transcription is often far less than the number of genes in the genome. Finally, in complex bacterial genomes, the expression of one gene may counteract that of another and such conflicts should necessarily be contained. Thus, regulation of gene expression, or, in other words, taking decisions on which gene to express at any point in time, is important.

Various regulatory systems, or networks, help the cell achieve gene regulation. First, though the sequence of DNA is relatively static, its structure is not. The DNA double helix is usually in what is called a B-form, in which each turn of the DNA has ~10 base pairs. The double helix can unwind or overwind such that the number of base pairs per turn is less than or greater than 10, and, in *E. coli*, this is to a large extent determined by the energy levels available to the cell.[7] DNA that is unwound is said to be negatively supercoiled. As the degree of negative supercoiling decreases and approaches the standard B-form twist, the DNA is said to be more relaxed. *E. coli* DNA is rarely, if ever, positively supercoiled, though this is known to happen in other bacteria. Enzymes under the umbrella name topoisomerase help modulate supercoiling states of DNA. In *E. coli* a topoisomerase called DNA gyrase negatively supercoils DNA, whereas DNA topoisomerase 1 helps relax DNA. When the cellular energy levels are high, the DNA is negatively supercoiled due to high DNA gyrase activity and this permits rapid transcription; during starvation, the DNA becomes relaxed, which can globally suppress transcription.[8] However, this overarching link between DNA supercoiling and transcription does not apply equally to all genes (Fig. 5.2). It has been observed that genes whose expression is preferentially reduced during starvation (or whose expression is high specifically during rapid growth) have a G+C-rich region in their promoters. This might make the promoter harder to unwind because G-C base pairs are more stable than A-T base pairs.[9] Unwinding of such promoters might be facilitated by negative supercoiling, which is favoured during high growth states. This mechanism appears to affect the expression of many genes involved in translation, including that of rRNA, whose transcription at high levels under nutrient stress can be hugely wasteful and damaging. Under nutrient-replete conditions, however, high transcription of such genes is necessary to support growth. On the other hand, promoters that are extraordinarily A+T-rich may be preferentially transcribed during starvation, when the genome in general is less negatively supercoiled.[10] Therefore, the

drawbacks for applications in synthetic biology', *FEMS Microbiol Rev*. 43 (2019), 304–339. https://doi.org/10.1093/femsre/fuz001

7 The number of bases per turn is called the twist. It represents how one strand of DNA winds around the other. There is a second component called writhe. This represents the coiling of the entire double helix around itself. We will not discuss this in any detail here.

8 C.J. Dorman, 'DNA supercoiling and transcription in bacteria: a two-way street', *BMC Molecular and Cell Biology* 20 (2019), 26. https://doi.org/10.1186/s12860-019-0211-6

9 R. Forquet, M. Pineau, W. Nasser, S. Reverchon, and S. Meyer, 'Role of the discriminator sequence in the supercoiling sensitivity of bacterial promoters', *mSystems* 6 (2021), e00978–21. https://doi.org/10.1128/msystems.00978-21

10 B.J. Peter, J. Arsuaga, A.M. Brier, A. Khodursky, P.O. Brown, and N. Cozzarelli, 'Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*', *Genome Biology* 5 (2004), R87. https://doi.org/10.1186/gb-2004-5-11-r87

fact that the structure of the DNA can respond to some cellular conditions and in turn affect the extent to which different genes are transcribed makes the DNA itself an important regulator of gene expression.

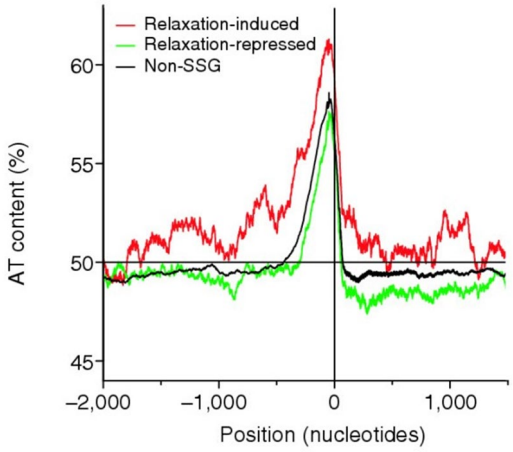


Fig. 5.2. Base composition upstream of genes regulated by DNA supercoiling in *E. coli*. This figure shows that genes that are induced by DNA relaxation are more A+T-rich than the average gene, whereas the reverse holds for genes that are induced by negative supercoiling. Along the x-axis, values to the left of '0' indicate positions upstream of genes, and positions to the right indicate the gene body and further beyond. Originally published as Figure 5B in B.J. Peter, J. Arsuaga, A.M. Brier, A. Khodursky, P.O. Brown, and N. Cozzarelli, 'Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*', *Genome Biology* 5 (2004), R87, CC BY 2.0.

Yet another component of the core transcriptional machinery that plays a regulatory role is the σ-factor, which helps the RNAP recognise promoters. Many bacterial genomes code for multiple σ-factors. Depending on the relative abundance of the core RNAP and σ-factors, the multitude of σ-factors can all be bound to abundant RNAP molecules, or they compete for the limited real estate presented by an insufficient number of core RNAP molecules. Though protein quantification by different labs support different scenarios, the most comprehensive and recent analysis (to my knowledge) supports the latter.[11] Thus, we now accept that different σ-factors compete with each other for binding to the core RNAP. The outcome of this competition will be determined by the relative abundance or availability and the affinity of each σ-factor to the core RNAP.

Different σ-factors recognise different promoter types. The standard bipartite promoter structure that we described earlier is best recognised by what is called the σD σ-factor, following the nomenclature used for *E. coli*. σD is a 'housekeeping' σ-factor that, by recognising the standard promoter, helps initiate transcription of a majority of genes involved in growth and metabolism that operate in nutrient-rich conditions. A second σ-factor, σ^S in *E. coli*, becomes available in sufficient concentrations as nutrients

11 S.E. Piper, J.E. Mitchell, D.J. Lee, and S.J.W. Busby, 'A global view of *Escherichia coli* Rsd protein and its interactions', *Molecular Biosystems* 5 (2006), 1943–1947. https://doi.org/10.1039/b904955j

deplete and cells enter a period of starvation and stress. This σ-factor helps the RNAP bind to promoters of genes underlying the bacterial response to a variety of stresses, which together form the general stress response. There is evidence that several promoters bound by σ^S-bound RNAP are recognised by this σ-factor when the DNA is relaxed, as it is during starvation, pointing to how DNA structure can contribute to differential promoter recognition by various σ-factors.[12] Because different σ-factors, when bound to the RNAP, can recognise their own set of promoters, the outcome of the competition between these σ-factors for binding to the core RNAP is a major determinant of which genes are expressed. We will return to this aspect of regulation later in this chapter.

Whereas the structure of the DNA and σ-factors (as regulatory molecules) are still part of the machinery that performs transcription, several other 'outside' players fulfil important roles in gene regulation.[13] The most prominent among these are transcription factors (TFs). TFs are DNA-binding proteins. They bind to specific DNA sequence motifs often present around the promoter region. The DNA sites to which TFs bind are sometimes called operators, or simply TF-binding sites. When bound to these sites, TFs can either activate or repress transcription (Fig. 5.3).[14] TFs repress transcription usually by binding close enough to the promoter that they block access to the RNAP; in other words, they sterically hinder RNAP-promoter interactions. By binding to one site near the promoter and another further upstream, they can also loop the intervening DNA and form a strongly repressive structure that prevents RNAP activity. Sometimes they do not block the initial interaction between the enzyme and the DNA, but instead prevent further progress.

The discovery by Arthur Pardee, Francois Jacob and Jacques Monod of a repressor of the set of genes that help *E. coli* metabolise sugar lactose, published in 1959, played a central role in the discovery of messenger RNA.[15] This repressor was isolated a few years later by Benno Muller-Hill[16] and shown to bind specifically to its operator site on DNA. While Monod was working on the induction of lactose metabolism genes, Andre Lwoff was demonstrating the phenomenon of bacteriophage lysogeny. Mark Ptashne's work revealing the central role of repressors of transcription in the maintenance of lysogeny was yet another landmark in the history of gene regulation.[17]

---

12 S. Kusano, Q. Ding, N. Fujita, and A. Ishihama, 'Promoter selectivity of Escherichia coli RNAP E sigma 70 and E sigma 38 holoenzymes. Effect of DNA supercoiling', *Journal of Biological Chemistry* 271 (1996), 1998–2004. https://doi.org/10.1074/jbc.271.4.1998

13 Small molecules such as guanosine tetraphosphate are produced in response to starvation and can bind to the RNAP and repress transcription of growth-related genes. We do not discuss this regulatory arm beyond brief mentions in this book.

14 Browning and Busby, 2004.

15 A.B. Pardee, F. Jacob, and J. Monod, 'The genetic control and cytoplasmic expression of inducibility in the synthesis of b-galactosidase in *E coli*', *Journal of Molecular Biology* 1 (1959), 165–178. https://doi.org/10.1016/b978-0-12-131200-8.50004-6

16 W. Gilbert and B. Mueller-Hill, 'Isolation of the lac repressor', *Proceedings of the National Academy of Sciences USA* 56 (1966), 1891–1898. https://doi.org/10.1073/pnas.56.6.1891

17 Reviewed and described in retrospect in M. Ptashne, *A Genetic Switch: Phage Lambda Revisited* (Plainview, NY: Cold Spring Harbor Laboratory Press, 2004).

The discovery of the repressor-based regulation of lactose metabolism genes also showed that in bacteria several genes encoded in tandem on the genome can be expressed from a single promoter. Such a series of co-transcribed genes is referred to as an operon, a fundamental feature of bacterial genomes: the ~4,000 genes in the *E. coli* genome may be organised into ~2,000 operons. Not all genes are organised into operons; many are singletons. Some operons are short, comprising not more than two or three genes whereas other uber-operons can encompass tens of genes.

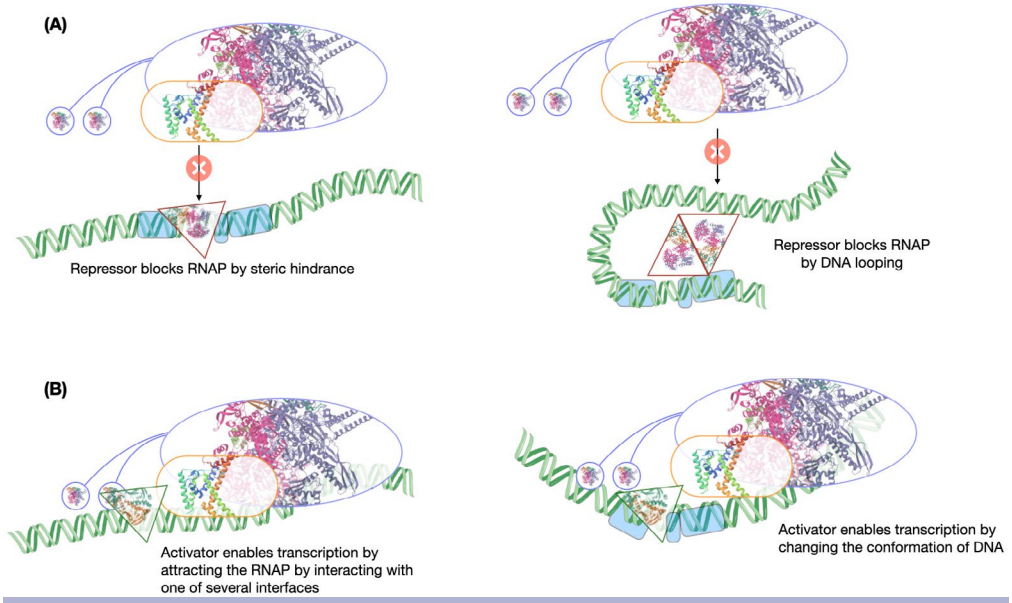


Fig. 5.3. Activation and repression by transcription factors. This figure shows a sample of simple ways by which (A) repressors (red-bordered triangles) and (B) activators (green-bordered triangles) act to regulate transcription initiation. The structure of the lac repressor filling the red triangle is from PDB: 1LB1, and that of CRP filling the green triangle is from PDB: 4N9H. The images of the DNA are from SMART-Servier Medical Art, part of Laboratoires Servier, via Wikimedia Commons, available freely under CC BY-SA 3.0.

Not all regulators of transcription are repressors. Activators normally bind upstream of the promoter and have elements that can attract the RNAP, interacting either with the core RNAP components or with the σ-factor. Ellis Engelsberg, along with his colleagues Joseph Irr, Joseph Power and Nancy Lee, discovered in 1965 that the genes for utilisation of the sugar arabinose in *E. coli* came under the control of an activator.[18] This discovery was initially met with much scepticism because of the deeply entrenched repressor-based model of gene regulation espoused by work of Pardee, Jacob and Monod. In the words of Steven Hahn, "(though) the evidence in 1965 for positive control by AraC was as good or better than the data used to formulate

18 Reviewed from a scientific and historical perspective in S. Hahn, 'Ellis Englesberg and the discovery of positive control in gene regulation', *Genetics* 198 (2014), 455–460. https://doi.org/10.1534/genetics.114.167361

the negative control model, Englesberg needed to accumulate much additional data to answer his critics."[19] Nevertheless, it was only a matter of time before other activator systems were described and Engelsberg stood vindicated. Many TFs can activate transcription of one gene but repress that of another, and whether they activate or repress transcription depends on where they bind relative to the promoter.[20] Some TFs, including Engelsberg's activator, can even perform dual actions on the same target gene by changing its binding site upstream of the gene. One can expect any activator to be able to repress transcription as long as it binds the DNA in such a manner that the RNAP cannot bind to the promoter. The reverse need not be true—a pure repressor cannot activate transcription just because it binds upstream of the promoter—it may not possess an interface to attract the RNAP.

The activity of TFs themselves is often determined by the presence or absence of a signal. For example, a TF that activates transcription of genes responsible for metabolising a sugar as a nutrient will be activated by the presence of the sugar. Such a TF, in addition to being able to bind DNA, will also be able to bind to the sugar to which it responds. The binding of the sugar to the protein will then activate (if the TF is an activator of the sugar metabolism genes) or hinder the TF's ability to bind to the DNA (if the activator is a repressor). Many TFs in bacteria possess such a property. The repressor of lactose metabolism binds to allolactose (similar to lactose). When not bound to allolactose, the repressor binds to the DNA and blocks RNAP activity. The binding of allolactose causes the TF to release the DNA, thus allowing transcription. The activator of arabinose metabolism binds to the DNA both in the presence and absence of the sugar. In the former situation, it acts as an activator but switches to being a repressor in the latter. Other TFs may not directly bind to a signal, but may be activated following a series of reactions that are initiated by a separate signal-sensing protein that, for example, may be located on the cell membrane. Each TF regulates its own set of target genes and the set of TF-target gene interactions constitutes a transcriptional regulatory network. Thus, TFs are proteins whose activities are usually determined by the presence of certain environmental or cellular conditions, in response to which they regulate the transcription of other genes.

5.2. The transcriptional regulatory network of *E. coli*

The *E. coli* genome encodes ~300 TFs for its total complement of ~4,400 genes. Even in this well-studied organism, we do not know all the regulatory connections these TFs make. Over half of these TFs have at least one known target gene,[21] as discovered through biochemical or genetic experiments; the others are predicted to be TFs based on their sequences. In

19 Hahn, 2014.

20 M. M. Babu and S.A. Teichmann, 'Functional determinants of transcription factors in Escherichia coli: protein families and binding sites', *Trends in Genetics* 19 (2003), 75–79. https://doi.org/10.1016/s0168-9525(02)00039-2

21 A target gene of a TF is a gene whose expression is regulated by the TF. Usually, it refers to genes that are directly regulated by the TF which binds to a site upstream of the gene's promoter.

addition to these TFs, *E. coli* has seven σ-factors competing to bind to the core RNAP. These TFs and their target genes or operons together constitute the transcriptional regulatory network. For *E. coli* there are publicly available databases such as RegulonDB[22] and Ecocyc,[23] from which the currently known transcriptional regulatory network can be downloaded. These networks comprise of data from a variety of experiments—from small-scale, detailed studies on how a particular TF binds to an operator to regulate a target gene, to large-scale, bird's-eye view studies that catalogue the list of all genes or operons that are regulated by one or more TFs under a set of growth conditions.

The targets of a TF can be defined in several ways. A gene can be called a target of a TF if the regulator binds upstream of the gene and, when bound, alters the expression state of the gene. This would define direct targets of a TF. Sometimes, the mere binding of a TF to an operator is used to define a target irrespective of whether there is evidence that the binding affects the expression of the gene. This may be appropriate when there is reason to believe that absence of evidence (of an effect on gene expression) is not evidence of absence. On the flip side, some TF-DNA interactions may also be non-functional. Alternatively, genes that change in expression when a TF is deleted can be called targets of the regulator. However, the targets defined may, therefore, not always be bound by the TF, and may change in expression as a result of a cascade of effects initiated far upstream by the direct regulation of a different gene(s) by the TF.

Given such complications in the ways in which a regulatory network can be defined, are such networks even useful to define on a genome-wide scale? In other words, does a regulatory network—built by aggregating data from hundreds to thousands of experiments together encompassing a cocktail of approaches—predict gene expression: the defining, measurable output of the regulatory network? Xin Fang and colleagues recently showed that a transcriptional regulatory network built from data on where TFs bind on the genome agrees well with genes that change in expression when a TF is deleted, and that the regulatory network is good enough to predict the gene expression states of over 85% of operons.[24] Earlier work by Gabor Balazsi and colleagues had shown that groups of genes that are expressed together under a given condition often belonged to coherent, closely-linked parts of the then known regulatory network.[25] Thus, the transcriptional regulatory network—despite being incomplete even for a well-studied model organism such as *E. coli*—serves as a good predictor of gene expression. However, though groups of genes regulated in the same manner may be expressed together, the expression level of a TF may not correlate well with that of its targets, in part because the activity of a TF is not defined entirely by its expression level.[26]

22 https://regulondb.ccg.unam.mx/

23 https://www.ecocyc.org/

24 X. Fang, A. Sastry, N. Mih, D. Kim, J. Tan, et al., 'Global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to TF activities', *Proceedings of the National Academy of Sciences USA* 114 (2017), 10286–10291. https://doi.org/10.1073/pnas.1702581114

25 G. Balazsi, A.-L. Barabasi, and Z.N. Oltvai, 'Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*', *Proceedings of the National Academy of Sciences USA* 102 (2005), 7841–7846. https://doi.org/10.1073/pnas.0500365102

26 S.J. Larsen, R. Rottger, H.H.H.W. Schmidt, and J. Baumbach, '*E. coli* gene regulatory networks

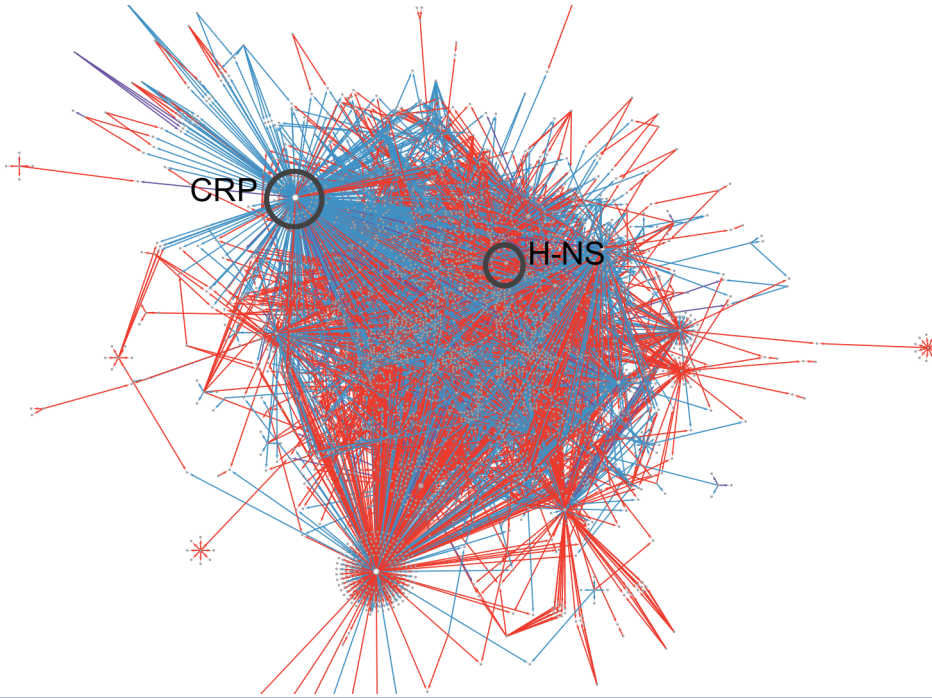


Fig. 5.4. The *E. coli* regulatory network. This figure shows a representation of the *E. coli* transcriptional regulatory network. Each line indicates a regulatory interaction between a regulator (mostly, but not necessarily, TFs) and a target gene. Red lines mark repressive interactions, whereas blue lines indicate activating interactions. Global regulators CRP and FNR are marked. Figure produced by Ganesh Muthu using the regulatory network available in the RegulonDB database (https://regulondb.ccg.unam.mx/) and Cytoscape (y-Force layout; https://cytoscape.org/ and https://www.yworks.com/products/yfiles-layout-algorithms-for-cytoscape).

The first decade of this century saw the publication of several papers describing graph theoretical studies of biological networks. Among these networks are transcriptional regulatory networks. To some extent, these studies were spurred by genome-scale studies of the eukaryote *Saccharomyces cerevisiae*, a yeast. One major work identified binding sites for over 150 TFs encoded by this organism,[27] triggering a large number of studies curating and analysing the vast amounts of data produced by this work. Any network is a graph that draws *edges* connecting points called *nodes*. In what is called a protein-protein interaction network, nodes are proteins and an edge is drawn between two proteins that physically interact with each other. The edges in such a network are not directional: if protein *A* interacts with protein *B*, then *B* also interacts with *A* and there is no direction to how the two proteins interact with each other. A transcriptional regulatory network, which connects TFs to their target genes or their binding sites, is directional: each edge is directed from the TF to a target gene because the TF *regulates*

are inconsistent with gene expression data', *Nucleic Acids Research* 47 (2019), 85–92. https://doi.org/10.1093/nar/gky1176

27 C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macissac, et al., 'Transcriptional regulatory code of a eukaryotic genome', *Nature* 431 (2004), 99–104. https://doi.org/10.1038/nature02800

the expression of the target gene (Fig. 5.4). Some representations of the transcriptional regulatory network also include σ-factors as regulatory proteins similar to TFs, others do not.

As one would immediately guess, such networks are not exclusive to biology. One can envisage a whole host of networks—such as the internet, electricity grids, and postal networks—and all of these can be directional. In, say, a postal network, which connects two post offices with an edge if letter bundles are sent from one to the other, not all nodes are connected equally. For example, the general or hub post office of a city will receive all letters sent to the city and then forward it to various local offices. The local post offices, despite being in the same city, may not be connected to each other directly. Thus, many post offices will have low connectivity, often being connected both ways with only the city's general post office. The general post office on the other hand is highly connected, but the number of such offices is very small compared to the number of local offices.

Similar trends have been described for biological networks as well. For example, in the *E. coli* transcriptional regulatory network, most TFs regulate only a few genes;[28] the repressor of lactose metabolism regulates only what is called the *lac operon*. On the other hand, a few TFs regulate hundreds of genes! The former, with their limited sphere of influence, are often called *local* TFs, and the latter, in contrast, are referred to as *global* TFs. That said, however, the distribution of the number of targets a TF has is continuous, and therefore it is not entirely obvious where a line demarcating local from global TFs should be drawn. As a result, there have been several definitions of what constitutes a global TF.

Often, an arbitrary threshold number of targets is used to separate global and local TFs. But is the number of targets the only parameter that defines global TFs? Some studies, primarily by Julio Collado-Vides and colleagues, argue otherwise. To follow this line of thinking, we must first understand and visualise the network itself a bit better. The regulatory network is not a disconnected set of TFs regulating their target genes in a one-on-one or a one-on-many manner. Just as a TF can regulate multiple genes, many genes are regulated by multiple TFs. For example, the operon for lactose metabolism is regulated not only by the lactose-responsive repressor but also by a TF CRP that responds indirectly to glucose availability in such a way that CRP becomes active when glucose is limiting. The same CRP acts as a second regulator of arabinose metabolism as well. The lactose operon is expressed only when the repressor is not bound and CRP *is* bound; the arabinose metabolism genes are expressed when the arabinose-responsive TF is bound to its operator sites in an activating configuration *and* CRP is also bound. Genes that determine the decision of the *E. coli* cell to move or to stay put are regulated by several TFs, and biochemical experiments with purified

28 A recent study has suggested that very few TFs regulate only a single target gene: T. Shimada, H. Ogasawara, I. Kobayashi, N. Kobayashi, and A. Ishihama, 'Single-target regulators constitute the minority group of TFs in *Escherichia coli* K-12', *Frontiers in Microbiology* 12 (2021), 697803. https://doi.org/10.3389/fmicb.2021.697803

protein and DNA sequences suggest that several tens of regulators can bind to regions upstream of these genes.[29] It is well-nigh impossible for all these regulators to bind simultaneously to the small stretch of DNA upstream of these genes determining motility/adhesion. However, different small sets of regulators may be active and involved in the regulation of these genes under different conditions.

Next, TFs can regulate genes for other TFs. This can set up a variety of network motifs. It can create cascades in which a series of regulatory events ultimately determines the expression of a non-TF target gene. For example, a dimeric TF called FlhDC activates the expression of a σ-factor called FliA. FliA then regulates the expression of a host of genes that allow the bacterial cell to move in particular ways. But then, the cascade is not purely linear along a single path. FlhDC, in addition to regulating the gene encoding the σ-factor FliA, also directly activates genes that allow the cell to move. This creates what is called a feed-forward loop: TF *A* regulates the expression of TF *B*, and *A* and *B* together regulate the expression of a non-transcription-factor target gene *C*. Depending on whether the effect of *A* on *C* is the same as the composite effect of *A* and *B* on *C*, the feed-forward loop is either coherent or incoherent.

Most feed-forward loops in the *E. coli* regulatory network appear to be coherent.[30] FliA, being a σ-factor, activates all its targets and FlhDC activates its target genes. Therefore, both routes to the motility-determining genes are activating, so this represents a coherent feed-forward loop. The two activating regulatory arms leading to *C* may represent an AND gate, in which both arms are required for the full expression of *C*, or they may form a SUM gate in which the effect on *C* is the sum of the effects of the two individual arms. The regulatory system for motility forms a SUM gate.[31] The arabinose system also includes an AND feed-forward loop in which CRP sits as a top-level TF that regulates the expression of the arabinose-responsive TF as well as that of the enzymes that metabolise arabinose. OR gates between the two arms of a feed-forward loop are also possible. For example, a coherent feed-forward loop forming an OR gate regulates the expression of a negative regulator of an adhesive structure called holdfast in a bacterium *Caulobacter crescentus*.[32] Whereas the activating SUM input of a coherent feed-forward loop keeps the expression of motility genes on for a long time, a coherent OR input structure helps to decrease the effect of environmental fluctuations on the expression of the ultimate target gene. Other types of local network structures include one in which the same TF regulates multiple genes, but with varying binding affinities such that some targets are prioritised for regulation over others.[33]

29 A. Ishihama, 'Prokaryotic genome regulation: a revolutionary paradigm', *Proceedings of the Japan Academy B* 88 (2012), 485–508. https://doi.org/10.2183/pjab.88.485

30 S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, 'Network motifs in the transcriptional regulation network of Escherichia coli', *Nature Genetics* 31 (2002), 64–68. https://doi.org/10.1038/ng881

31 S. Kalir, S. Mangan, and U. Alon, 'A coherent feed-forward loop with a SUM input function prolongs flagella expression in *Escherichia coli*', *Molecular Systems Biology* 1 (2005), 2005.0006. https://doi.org/10.1038/msb4100010

32 M. McLaughlin, D.M. Hershey, L.M.R. Ruiz, A. Fiebig, and S. Crosson, 'A cryptic TF regulates Caulobacter adhesin development', *PLoS Genetics* 18 (2022), e1010481. https://doi.org/10.1371/journal.pgen.1010481

33 A. Zaslaver, A.E. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M.G. Surette, and U. Alon,

Finally, feedback loops also occur. If one were to include the signal molecule, such as lactose (or allolactose), in our representation of the regulatory network, the regulation of the lactose operon would be an example of a positive feedback loop. In this positive feedback loop, activation of the lactose operon in a cell by a small quantity of the signal would result in more of the signal molecule being taken up by this cell. When lactose is limiting such that not all cells see lactose first, a bistable state, in which two sub-populations of cells with very distinct levels of expression of the *lac* operon will be established. The lucky few cells that came into contact with lactose early would be primed to take up most of the lactose available in the local environment, whereas others would be oblivious to the availability of this sugar. Here we immediately notice a situation where genetically identical cells display two distinct traits because of how their regulatory network has interpreted the environment. This highlights the important argument that the genome sequence is not a dictatorial directive but, as the science writer Philip Ball puts it, it helps to establish some "flexible rules" from which life can emerge.[34] Negative feedback loops also exist, and there are also instances where one TF regulates another and the latter returns the favour. In other words, even if we do not include the signal molecule in our representation of the regulatory network, feedback loops involving TFs and no other types of molecules exist.[35] Finally, TFs can auto-regulate their own expression, often negatively, but also positively. Whereas the former helps to reduce response time and decrease fluctuations, the latter does the opposite. Thus, each type of network motif has its own kinetic properties.[36] Taken together, the take home message from this short discussion is that the transcriptional regulatory network is a complex set of highly interconnected nodes that form a variety of converging, diverging, and even circular loops.

Back to global regulators: do some regulators with a large number of targets possess additional properties that distinguish them from local TFs that regulate a small set of genes on demand? Agustino Martinez-Antonio and Julio Collado-Vides found that some TFs with a large number of targets share certain properties which together qualify them as global regulators. Firstly, they regulate genes belonging to distinct functions.[37] For example, CRP regulates genes involved in carbohydrate metabolism as well as the

'Just-in-time transcription program in metabolic pathways', *Nature Genetics* 36 (2004), 486–91. https://doi.org/10.1038/ng1348

34 P. Ball, 'How life really Works', *Nautilus*, 6 November 2023. https://nautil.us/how-life-really-works-435813/.

35 J.A. Freyre-Gonzalez, J.A. Alonso-Pavon, L.G. Trevino-Quintanilla, and J. Collado-Vides, 'Functional architecture of *Escherichia coli*: new insights provided by a natural decomposition approach', *Genome Biology* 9 (2008), 2008. https://doi.org/10.1186/gb-2008-9-10-r154

36 These have been reviewed elsewhere. See, for example, U. Alon, 'Network motifs: theory and experimental approaches', *Nature Reviews Genetics* 8 (2007), 450–461. https://doi.org/10.1038/nrg2102

I have also summarised some of this in chapter 5 of my previous book. A.S.N. Seshasayee, *Bacterial Genomics: Genome organisation and gene expression tools* (Cambridge: Cambridge University Press, 2016). https://doi.org/10.1017/cbo9781139942225.005

37 A. Martinez-Antonio and J. Collado-Vides, 'Identifying global regulators in transcriptional regulatory networks in bacteria', *Current Opinion in Microbiology* 6 (2003), 482–489. https://doi.org/10.1016/j.mib.2003.09.002

genes encoding FlhDC, the master regulator of motility, among others. In contrast, TFs with only a few targets often regulate genes which all or mostly belonging to a single pathway or function. For instance, the TF TyrR regulates a small number of operons, all encoding genes involved in the metabolism of aromatic amino acids. In a more recent work, Julio Freyre-Gonzalez and colleagues suggested that global regulators integrate multiple *modules* within the transcriptional regulatory network.[38] Modules are groups of highly interconnected nodes, such that nodes within a module are more connected to each other than nodes from across modules. Modules are often determined by one or more TFs with local scope whereas global regulators sit astride multiple modules, presumably priming the expression of a large and diverse set of genes. Each module would, in some ways, represent a functionally coherent set of genes, and therefore Freyre-Gonzalez and colleagues' work might be taken as an independent validation of Martinez-Antonio's suggestion that global TFs regulate genes from multiple functions.

The *E. coli* genome encodes seven σ-factors. Each σ-factor helps to transcribe its own set of genes. Though there is some overlap between the sets of genes regulated by different σ-factors, one can say that the transcriptional space of *E. coli*, or for that matter that of many bacteria with large genomes, is *partitioned* among σ-factors.[39] Global TFs often regulate genes from different σ-factor partitions. According to the data analysed by Martinez-Antonio and co-workers, CRP regulates genes from as many as four σ-factor partitions! Global TFs do not usually regulate a gene as its sole regulator; they often act in concert with other TFs. Again, the regulation by CRP of its targets in sugar metabolism in concert with local TFs is a good example. Global TFs often regulate other TFs, something that TFs with a local scope rarely do. Finally, global TFs are also active in multiple conditions, whereas local regulators are usually activated by highly specific signals. Based on these parameters, Martinez-Antonio and Collado-Vides concluded that the *E. coli* genome encodes seven global TFs, and that a majority of target genes have at least one of these seven TFs as their regulators.

Years ago, I was involved in a piece of work that attempted to study how genes involved in metabolism are regulated in the transcriptional regulatory network of *E. coli*, and how different segments of the metabolic network might be regulated differently by global and local TFs.[40] The metabolic network can be visualised as an hourglass. A great diversity of nutrient breakdown pathways converge down to what is called central metabolism, which eventually produces energy. And a variety of biosynthetic pathways diverge away from these central metabolic pathways. Usually, nutrient breakdown pathways are regulated by TFs that respond to the nutrient itself, i.e., these are regulated by supply levels. On the other hand, biosynthetic pathways are regulated

38 Freyre-Gonzalez et al., 2008.

39 T.M. Gruber and C.A. Gross, 'Multiple sigma subunits and the partitioning of bacterial transcription space', *Annual Review of Microbiology* 57 (2003), 441–466. https://doi.org/10.1146/annurev.micro.57.030502.090913

40 A.S.N. Seshasayee, G.M. Fraser, M.M. Babu, and N.M. Luscombe, 'Principles of transcriptional regulation and evolution of the metabolic system in *E. coli*', *Genome Research* 19 (2009), 79–91. https://doi.org/10.1101/gr.079715.108

by TFs that bind to final product molecules, i.e., these are controlled by demand. These TFs are often local. Central metabolic pathways, in contrast, are regulated by a multitude of global TFs. This makes sense in light of Martinez-Antonio and Collado-Vides' work showing that global TFs respond to a wide range of environmental and cellular conditions, as well as the notion that central metabolic pathways are the culmination of a whole range of metabolic cues.[41] Whereas biosynthetic pathways are usually regulated by a single TF, the regulation of breakdown pathways often involves combinations of global and local TFs, as exemplified by CRP acting as a major node regulating a whole variety of breakdown pathways in concert with local TFs responding to particular nutrient molecules.

Finally, global TFs are usually expressed at high levels in the cell, and this presumably is to ensure their availability in sufficient concentrations to bind to all their targets. In contrast, local TFs are expressed at low levels; in fact, there is a positive correlation between the expression level of a TF and the number of targets it regulates. As shown by Grigory Kolesov and colleagues, for a TF present at low concentrations, finding one or a few target binding sites in the cell can be inefficient.[42] To counter this, local TFs are often encoded adjacently to their target sites. This logic works for bacteria in which transcription and translation occur more or less simultaneously, but not for eukaryotic cells in which transcription occurs within the nucleus whilst proteins are synthesised outside and so the TF will have to be transported back into the nucleus before it can bind to and regulate its target. There is also an evolutionary explanation for why local TFs are encoded close to their cognate targets. Horizontal transfer of a stretch of DNA carrying genes for a metabolic pathway is more likely to be successful if it also carries its own TF!

In summary, the transcriptional regulatory network—as exemplified in *E. coli*—is a complex structure with highly interconnected regulators and their targets. This structure ultimately determines which genes are expressed in the cell and when. Many regulators act in an intuitive manner, responding to a signal and regulating genes that should respond precisely to the inducing signal. In contrast, there are global TFs that integrate multiple segments of the metabolic network, and whose significance is a lot harder to understand and rationalise. In the next few sections, we will describe and try to understand the functioning of two regulatory networks determined by global TFs.

5.3. Driving the stress response: σ^S and its competition with σ^D

In *E. coli*, the major σ-factor σ^D regulates the expression of most genes involved in growth. It is the most abundant σ-factor protein in the cell and, out of the seven σ-factors encoded by the *E. coli* genome, it has the highest affinity for binding to the core RNAP. Its function contrasts with that of σ^S, the stress responsive σ-factor, whose production

41 Martinez-Antonio and Collado-Vides, 2003.

42 G. Kolesov, Z. Wunderlich, O.N. Laikova, M.S. Gelfand, and L.A. Mirny, 'How gene order is influenced by the biophysics of transcription regulation', *Proceedings of the National Academy of Sciences USA* 104 (2007), 13948–53. https://doi.org/10.1073/pnas.0700672104

increases as nutrients deplete and growth starts to cease. σ^S is also expressed during growth, but under conditions that are not ideal for growth of the bacterium. These include conditions of suboptimal osmolarity, temperature, and pH.[43] Like the global regulator CRP, σ^S does not respond nor activate cognate responses to specific stresses but plays a "preventive"[44] role, priming the cell to tolerate a wide range of stresses. Given its ability to contribute positively to cellular responses to such stresses, it is not surprising that the gene for σ^S was discovered independently multiple times by several researchers in the 1980s–early 1990s[45] before it was recognised that all these researchers had been referring to the same protein doing its job in different contexts![46]

The expression of σ^S is regulated by a plethora of mechanisms at every conceivable step in gene expression, from transcription through translation to protein stability.[47] Transcription of *rpoS*, the gene encoding σ^S, increases during down-shifts in growth, most notably as an *E. coli* culture transitions from exponential growth to a stationary phase.[48] There is evidence that the global TF CRP is involved in the up-regulation of σ^S during the stationary phase. The RegulonDB database for transcriptional regulatory interactions in *E. coli* also includes an acid stress regulator called GadX as a regulator of the σ^S gene. The small molecule guanosine tetraphosphate, which is produced during transition to states of starvation, also up-regulates the transcription of the σ^S gene. The concentration of the mRNA for σ^S also decreases in the absence of the cytosine methyltransferase Dcm (see Chapter 4). Various other regulators of σ^S expression have been described in other bacteria as well. However, the amount of σ^S mRNA is not a good predictor of the amount of σ^S protein: σ^S protein has been reported to be hardly detectable under some conditions in which the σ^S mRNA is abundant.[49] This suggests regulation of this gene beyond transcription. Some RNA binding proteins such as Hfq play roles in enabling translation of the σ^S mRNA in concert with other proteins and RNA that act as specific stress signals.[50] Even the protein HU, best known as a non-

43 R. Hengge-Aronis, 'Signal Transduction and Regulatory Mechanisms Involved in Control of the σS (RpoS) Subunit of RNAP', *Microbiology and Molecular Biology Reviews* 66 (2002), 373–395. https://doi.org/10.1128/mmbr.66.3.373-395.2002

44 Ibid., p. 374.

45 For example, P.C. Loewen and B.L. Triggs, 'Genetic mapping of katF, a locus that with katE affects the synthesis of a second catalase species in *Escherichia coli*', *Journal of Bacteriology* 160 (1984), 668–675. https://doi.org/10.1128/jb.160.2.668-675.1984; E. Touati, E. Dassa, and P.L. Bouquet, 'Pleiotropic mutations in appR reduce pH 2.5 acid phosphatase expression and restore succinate untilization in CRP-deficient strains of *Escherichia coli*', *Molecular and General Genetics* 202 (1986), 257–64. https://doi.org/10.1007/bf00331647

46 R. Lange and R. Hengge-Aronis, 'Identification of a central regulator of stationary-phase gene expression in Escherichia coli', *Molecular Microbiology* 5 (1991), 49–59. https://doi.org/10.1111/j.1365-2958.1991.tb01825.x

47 Hennge-Aronis, 2002.

48 Recall from Chapter 4 that a small inoculum of *E. coli* cells in fresh media, after a brief period of adaptation would start growing 'exponentially', or double in number at regular intervals until nutrient exhaustion and accumulation of toxic byproducts of growth metabolism cause cessation of cell multiplication in what is referred to as 'stationary phase'. During the stationary phase, any low rate of population growth is offset by cell death.

49 Hengge-Aronis, 2002.

50 D.D. Sledjeski and C. Whitman, 'Hfq is necessary for regulation by the untranslated RNA DsrA',

specific DNA binding protein, may bind to the σ^S mRNA and affect its translation.[51] The stability of σ^S protein is also regulated by complex signal cascades involving multiple proteins. Protein stability is a pivot point in σ^S expression that is targeted by adaptive strategies that use genetic evolution of σ^S, which we will return to later in this chapter. Thus, regulation at multiple stages of gene expression and protein stability seems to play a role in determining σ^S levels under different conditions. Whereas regulation at the level of transcription seems to predominate during steady transition to slow growth rates, stresses such as high osmolarity and low temperatures seem to particularly stimulate σ^S translation.

Harald Weber and colleagues identified genes regulated by σ^S in *E. coli* growing in three different conditions (stationary phase, osmotic, and acid stress) by measuring changes in the levels of the mRNA of all genes encoded by the genome in the presence and absence of an intact *rpoS* gene.[52] Of the ~500 genes that changed in expression in a σ^S-dependent manner in at least one of the three conditions, a third were defined as the 'core' σ^S *regulon*, being responsive to σ^S in all three conditions. The members of the core σ^S regulon typically contained an extended −10 element in their promoters, whereas the rest did not and might be expressed from sub-optimal promoter elements by σ^S-containing RNAP holoenzyme, in concert with other stress-responsive regulatory proteins. Byung-Kwan Cho and colleagues, while assembling a network of regulatory interactions between all σ-factors and their targets in *E. coli* (Fig. 5.5), showed that σ^S was bound to over 1,000 promoters in *E. coli*, but a majority of these interactions did not result in a change in gene expression when σ^S was removed.[53] The role of these binding interactions, if any, is yet to be understood. Whereas most of the genes whose expression changes in response to σ^S are activated by the σ-factor, the rest are in fact less expressed when σ^S is present. How could this be possible? Cho et al. showed that in many of these genes, the presence of σ^S reduced the binding of σ^D, the major housekeeping σ-factor. This suggests that competition between σ-factors, through which the presence of one σ-factor affects the influence of another, plays a role in determining the mRNA levels of several genes.

Journal of Bacteriology 183 (2001), 997–2005. https://doi.org/10.1128/jb.183.6.1997-2005.2001

51 A. Balandina, L. Claret, R. Hengge-Aronis, and J. Rouviere-Yaniv, 'The *Escherichia coli* histone-like protein HU regulates rpoS translation', *Molecular Microbiology* 39 (2001), 1069–1079. https://doi.org/10.1046/j.1365-2958.2001.02305.x

52 H. Weber, T. Polen, J. Heuveling, V.F. Wendisch, and R. Hengge, 'Genome-wide analysis of the general stress response network in Escherichia coli: σS-dependent genes, promoters, and sigma factor selectivity', *Journal of Bacteriology* 187 (2005), 1591–1603. https://doi.org/10.1128/jb.187.5.1591-1603.2005

53 B.-K. Cho, D. Kim, E.M. Knight, K. Zengler, and B.O. Palsson, 'Genome-scale reconstruction of the sigma factor network in *Escherichia coli*: topology and functional states', *BMC Biology* 12 (2014), 4. https://doi.org/10.1186/1741-7007-12-4

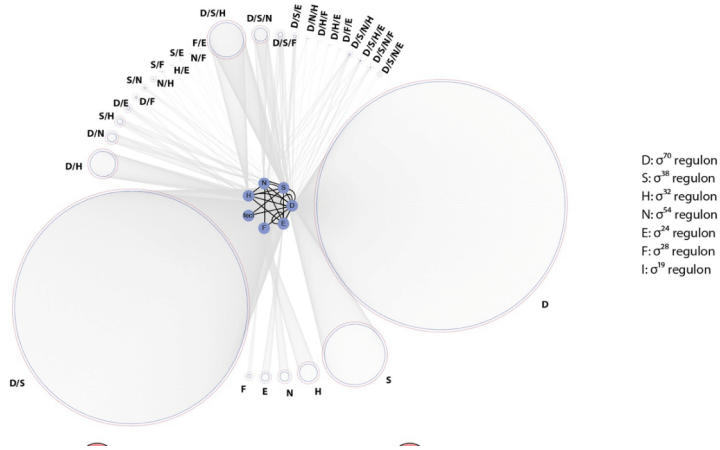


Fig. 5.5. Targets of various σ-factors in *E. coli*. This figure shows the sizes of various regulons (sets of targets of a regulatory protein) for *E. coli* σ-factors. σ^{38} is an alternative name for σ^{S}, and σ^{70} for σ^{D}. Originally published as Figure 2D in B.-K. Cho, D. Kim, E.M. Knight, K. Zengler, and B.O. Palsson, 'Genome-scale reconstruction of the sigma factor network in *Escherichia coli*: topology and functional states', *BMC Biology* 12 (2014), 4, CC BY 2.0.

For a σ-factor to activate the transcription of its targets, it first has to outcompete with other σ-factors for binding to the core RNAP; only then do holoenzymes bind to their respective promoters. Competition between σ-factors for a limited core RNAP—particularly the RNAP between σ^{D} and σ^{S}, both of which regulate large numbers of genes—can present some interesting problems. Though the concentration of σ^{S} increases as *E. coli* cells transition from exponential growth to a stationary phase, the absolute concentration of σ^{S} stays well below that of σ^{D}. In addition, the affinity of σ^{S} to the core RNAP is much less than that of σ^{D}. Therefore, even in stationary phase, σ^{S}, on its own, would be hard-pressed to compete effectively with σ^{D} to form the RNAP holoenzyme. According to calculations using the known total concentrations of σ^{S}, σ^{D}, and the RNAP in free and complexed forms, as well as the affinities of σ^{S} and σ^{D} to the RNAP, the concentration of the σ^{S}-holoenzyme (henceforth called $E\sigma^{S}$, with 'E' standing for the RNAP core enzyme) would be much smaller than that of $E\sigma^{D}$ even in stationary phase.[54] This immediately suggests the need for additional players that modulate σ-factor competition to favour the formation of $E\sigma^{S}$ in the stationary phase. Indeed, several such factors have been identified and described. One such protein, Crl, binds to σ^{S} and increases its affinity for the core RNAP, thus favouring the formation of $E\sigma^{S}$. This promotes the transcription of several σ^{S}-dependent genes.[55] There is evidence again that the small molecule guanosine tetraphosphate,

54 Comparing numbers from M. Mauri and S. Klumpp, 'A model for sigma factor competition in bacterial cells', *PLoS Computational Biology* (2014) e1003845, and those from Lal et al. 2018 quoted below. https://doi.org/10.1371/journal.pcbi.1003845

55 A. Typas, C. Barembruch, A. Possling, and R. Hengge, 'Stationary phase reorganisation of the *Escherichia coli* transcription machinery by Crl protein, a fine-tuner of sigmas activity and levels',

which signals starvation, also favours alternative σ-factors including σ^S in its competition with σ^D for holoenzyme formation.[56]

Two other molecules influence σ-factor competition by negatively targeting σ^D and $E\sigma^D$ activity. One is the protein Rsd, whose level increases during slow growth,[57] including in the stationary phase.[58] This protein binds to σ^D and sequesters it away from the core RNAP, thus removing a proportion of σ^D from the σ-factor competition. This should favour $E\sigma^S$ formation during the stationary phase. The other molecule influencing σ-factor competition is a non-coding RNA called 6S RNA.[59] 6S RNA is produced by the transcription of a gene called *ssrS*, but the RNA is not translated into protein. The 6S RNA adopts a looped structure that forms a base-paired motif resembling a standard *E. coli* promoter. This attracts $E\sigma^D$, thus reducing its availability for transcription while also removing part of σ^D during the stationary phase when the 6S RNA level is at its highest. Thus, while Rsd removes σ^D from the equation, 6S RNA reduces the availability of $E\sigma^D$ for transcription. The effect of 6S RNA therefore would also reduce the amounts of core RNAP available for σ-factors to bind to. Avantika Lal in my lab asked what effect each of these two methods of modulating the partitioning of transcription space across different Eσ holoenzymes would have on global gene expression in *E. coli*.[60] She used a combination of genome-scale gene expression measurements and mathematical modelling of σ-factor competition to approach this.

Experiments measuring mRNA expression levels of genes in *E. coli* cells lacking Rsd, in comparison with those that have the protein, showed that very few genes changed substantially in their expression between these two conditions. However, many σ^S target genes showed small but consistent decreases in expression levels in the absence of Rsd. Though Rsd operates by sequestering σ^D, σ^D targets did not show any change in their mRNA levels when Rsd was removed from the system. On the other hand, deletion of 6S RNA resulted in large changes in the expression levels of several genes, but the sets of genes responding to 6S RNA availability depended on the growth phase that the cells were in. But like the Rsd deletion, removal of 6S RNA also caused decreases in the expression of many σ^S target genes. In contrast to the Rsd deletion, removal of 6S RNA caused an increase in the expression of several σ^D target genes. Many of these σ^D targets showing elevated expression in the absence of 6S RNA were expressed at below average levels in the presence of 6S RNA. This suggests that one role for 6S RNA is in suppressing

EMBO Journal 26 (2007), 1569–1578. https://doi.org/10.1038/sj.emboj.7601629

56 M. Jishage, K. Kvint, V. Shingler, and T. Nystrom, 'Regulation of sigma factor competition by the alarmone ppGpp', *Genes and Development* 16 (2002), 1260–1270. https://doi.org/10.1101/gad.227902

57 R. Balakrishnan, M. Mori, I. Segota, Z. Zhang, R. Aebersold, C. Ludwig, and T. Hwa, 'Principles of gene regulation quantitatively connect DNA to RNA and proteins in bacteria', *Science* 378 (2022), eabk2066. https://doi.org/10.1126/science.abk2066

58 M. Jishage and A. Ishihama, 'Transcriptional organization andin vivo role of the Escherichia coli rsd gene, encoding the regulator of RNAP sigma D', *Journal of Bacteriology* 181 (1999), 3768–3776. https://doi.org/10.1128/jb.181.12.3768-3776.1999

59 K.M. Wassarman and G. Storz, '6S RNA regulates E. coli RNAP activity', *Cell* 101 (2000), 613–623. https://doi.org/10.1016/s0092-8674(00)80873-9

60 A. Lal, S. Krishna, and A.S.N. Seshasayee, 'Regulation of Global Transcription in *Escherichia coli* by Rsd and 6S RNA', *Genes Genomes Genetics* 8 (2018), 2079–2089. https://doi.org/10.1534/g3.118.200265

the expression of genes with relatively weak promoters. The removal of both 6S RNA and Rsd produced much greater effects on gene expression than would be expected from the product of these individual deletions.

These findings raised some important questions. First, Rsd acts by sequestering σ^D and any effect it has on σ^S function should be indirect. Yet, the aforementioned gene expression experiments showed that the removal of Rsd has an effect on the expression of σ^S targets but little, if any, on genes regulated by σ^D. How does this work? To answer this, Lal and colleagues used a mathematical model of holoenzyme formation and transcription that incorporated core RNAP, σ^D, σ^S, Rsd, 6S RNA, and DNA (Fig. 5.6). This model showed that in the presence of 6S RNA, increasing Rsd concentration results in the freeing up of $E\sigma^D$ from its complex with 6S RNA, leading to the release of some core polymerase which can then bind to σ^S and form $E\sigma^S$. The end result is an increase in transcription of $E\sigma^S$-dependent promoters. The model also showed that an increase in Rsd concentration also causes a decrease in transcription of σ^D target genes, something that was not apparent in the gene expression data. The gene expression experiments however showed that Rsd regulates 6S RNA: when Rsd is removed, there is a ~2.5-fold increase in the expression of 6S RNA. This could potentially decrease the availability of $E\sigma^D$ for transcription. Incorporation of this regulation of 6S RNA by Rsd into the mathematical model showed that it preserves the effect of Rsd on σ^S-dependent transcription but abolishes its effect on σ^D-dependent gene expression.

The second question that arises pertains to the effect of 6S RNA on transcription. 6S RNA not only sequesters σ^D but, by binding to $E\sigma^D$, also reduces the amount of the core RNAP available for σ^S to interact with. This should result in an overall decrease in transcription, though this effect would be more pronounced for σ^D-dependent genes. The theoretical model also supports this view. How then does the removal of 6S RNA selectively decrease σ^S-dependent transcription while increasing the expression of many σ^D-dependent genes? The gene expression data showed that the absence of 6S RNA reduced the expression of Rsd and increased that of σ^S itself. Further, it also decreased the expression of the catalytic subunit of the RNAP, which reduces the overall availability of this enzyme. These effects together help reduce σ^S-dependent transcription; in fact, a ~20% reduction in Rsd appears to be sufficient to reduce σ^S-dependent transcription in the absence of 6S RNA. 6S RNA also appears to affect the expression of other regulators of σ-factor competition such as Crl, all of which should contribute to σ^S competing effectively with σ^D for binding to core RNAP.

Thus, Rsd and 6S RNA regulate each other and also other players involved in transcription and σ-factor competition. These combined effects appear to be necessary for these molecules to modulate σ-factor competition in a manner that favours σ^S-dependent gene expression. Why such a complex web of interactions to modulate the interplay between σ^D and σ^S? We will try to answer this question when we explore how transcription regulatory networks evolve a little later in this chapter.

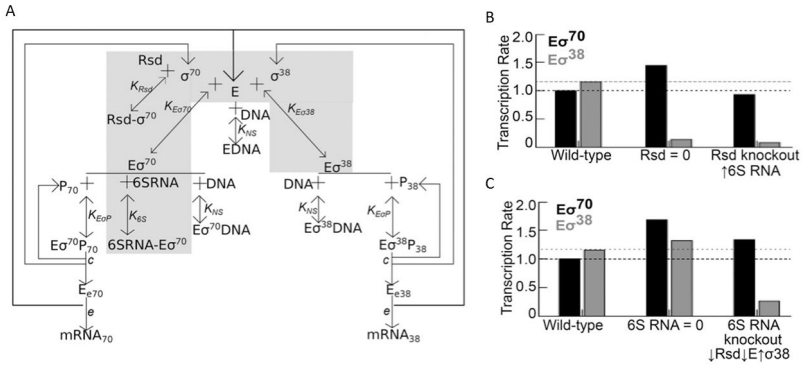


Fig. 5.6. Complex regulation of σS-σD competition. (A) A schematic showing the set of interactions involved in the regulation of σ^S-σ^D competition by the protein Rsd and the non-coding RNA 6S RNA. (B) Figure shows calculated transcription rate when Rsd is removed and when 6S RNA levels are increased in the absence of Rsd. (C) Figure shows calculated transcription rate when 6S RNA is removed, and the effect of the observed decrease in Rsd and core RNAP and increase in σ^S in the absence of 6S RNA. (B) and (C) show how functional connections between 6S RNA and Rsd appear to be important in determining the outcome of the competition between σ^S and σ^D. Originally published as part of Figure 6 in A. Lal, S. Krishna, and A.S.N. Seshasayee, 'Regulation of Global Transcription in *Escherichia coli* by Rsd and 6S RNA', *Genes Genomes Genetics* 8 (2018), 2079–2089, CC BY 4.0.

5.4. Managing the costs of horizontally acquired DNA: the 'genome sentinel'[61]

Horizontal transfer is a major mode of gene acquisition in bacteria (see Chapter 4), leading to genome expansion. As with any other segment of DNA, whether a piece of horizontally acquired DNA is maintained in a genome is a function of the selective pressure in its favour. This is especially so in bacteria with such large population sizes that the cost of merely maintaining and expressing a non-functional or neutral piece of DNA is enough for this DNA to be lost from the population (see Chapter 3). As described in Chapter 4, the selective pressure in favour of maintaining a piece of DNA in a genome could be conventional in that the DNA includes genes that enhance the growth and survival of the organism in its niche—for example, by allowing it to resist antibiotics in an antibiotic-rich environment. Or selection could arise from addiction, in which the loss of a piece of DNA once acquired proves toxic to the cell—a concept demonstrated by what are called toxin-antitoxin systems, which are often horizontally acquired. The cost that a piece of DNA presents to its host arises from the metabolic expense of maintaining it, as well as the possibility that its expression could prove

61 C.J. Dorman, 'H-NS, the genome sentinel', *Nature Reviews Microbiology* 5 (2007), 157–161. https://doi.org/10.1038/nrmicro1598

toxic to the cell or interfere with the functioning of molecules already well-established in the host. For example, Rotem Sorek and colleagues analysed gene fragments, from nearly 80 prokaryotic genomes, that could not be successfully transferred into *E. coli*.[62] They presented evidence that the expression of such genes, even at low levels, could not be tolerated by the host cell, arguing that these genes are toxic to *E. coli*. This toxicity prevents their successful establishment in the *E. coli* genome. For other genes, their failure to transfer into *E. coli* appeared to be best explained by increased dosage—or, in other words, their high expression levels. That high gene expression is a barrier to horizontal transfer of some genes was reaffirmed very recently by Rama Bhatia and colleagues, who studied the transfer of genes from *E. coli* into the closely-related *Salmonella*.[63] Thus, both toxicity and inappropriately high expression levels can act as barriers to horizontal gene transfer.

Many organisms encode dedicated mechanisms to control or even 'silence' horizontally-acquired genes at the level of their expression. A prominent example among eukaryotes is the silencing of selfish transposable elements by small regulatory RNA molecules in many plants. An example in bacteria involves a TF called H-NS, which in *E. coli* and related bacteria represses the expression of a variety of horizontally-acquired genes. H-NS, best known for its DNA-binding activities,[64] recognises A+T-rich sequences and binds extensively to such stretches of DNA. While doing so, it can form highly rigid or tightly looped structures that can either block the binding of RNAP to promoters or the relative movement of DNA and RNAP when the latter is already bound to a promoter. Now, it turns out that many horizontally-acquired genes in *E. coli* and related free-living bacteria tend to be A+T-rich. The genomes of *E. coli* and many related bacteria are usually nearly 50% A+T on average. However, the distribution of the A+T content of genes in these genomes is skewed towards the right. This means that very few genes are G+C-rich, but a larger proportion are A+T-rich, or more A+T-rich than the mean. Many such genes are believed, with good reasons, to have been acquired horizontally. These are often poorly conserved even across closely-related strains and species, and also include genes of bacteriophage origin. The high A+T content of many horizontally-acquired genes makes them attractive to H-NS for binding. Where H-NS binds, it represses transcription. Therefore, H-NS, which preferentially binds A+T-rich genes, emerges as a 'silencer' or repressor of the transcription of horizontally-acquired genes.

H-NS, as a protein that modifies the structure of DNA and also regulates gene

62 R. Sorek, Y. Zhu, C.J. Creevey, F.M. Pilar, P. Bork, and E. Rubin, 'Genome-wide experimental determination of barriers to horizontal gene transfer', *Science* 318 (2007), 1449–1452. https://doi.org/10.1101/2022.06.29.498157

63 R.P. Bhatia, H.A. Kirit, C.M. Lewis Jr, K. Sankaranarayanan, J.P. Bollback, 'Evolutionary barriers to horizontal gene transfer in macrophage-associated Salmonella', *Evolution Letters* 7 (2023), 227–239. https://doi.org/10.1093/evlett/qrad020

64 There is some evidence that H-NS can also bind to RNA. See C.C. Brescia, M.K. Kaw, and D.D. Sledjeski, 'The DNA binding protein H-NS binds to and alters the stability of RNA in vitro and in vivo', *Journal of Molecular Biology* 339 (2004), 505–514. https://doi.org/10.1016/s0022-2836(04)00382-1

expression, has been known since the 1980s.[65] However, its major role as a silencer of horizontally-acquired genes was not recognised until 2006, when two papers described the binding of H-NS to the chromosome of *Salmonella*—a close relative of *E. coli*—and the impact this binding has on gene expression on a genomic scale. The two pieces of work, one by William Navarre et al.[66] and the other by Sacha Lucchini et al.,[67] showed—by isolating and identifying chromosomal DNA regions bound by H-NS in *Salmonella* cells—that H-NS binds to hundreds of genes, including many coding for proteins that help the bacteria cause disease (Fig. 5.7). When H-NS was removed from cells by genetic means, genes bound by the protein greatly increased in expression, showing that H-NS represses the expression of genes it binds to. Many genes bound and regulated by H-NS are poorly conserved, are often specific to *Salmonella*, and show higher A+T-content than is typical of the average gene in this bacterium. Many of these genes have a role to play in the virulence of *Salmonella* and are normally expressed only during specific stages of infection and not during normal growth. The uncontrolled expression of these virulence-associated genes in the absence of H-NS is detrimental to the host bacterium. In fact, in the absence of additional, compensating mutations in the stress responsive σ-factor σ^S, the removal of H-NS is lethal to *Salmonella*. These findings clearly emphasised the importance of gene silencing to the fitness and evolutionary success of these bacteria.

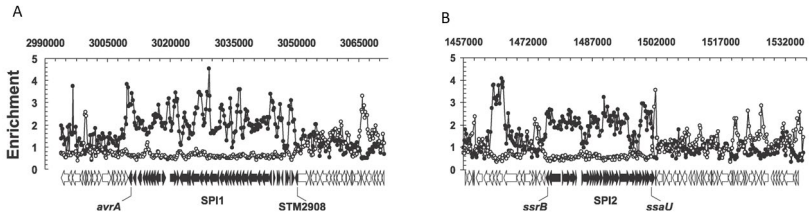


Fig. 5.7. Binding of H-NS to virulence genes. This figure shows the binding of H-NS (enrichment, on the y-axis) to pathogenicity-determining genes belonging to (A) SPI-1 and (B) SPI-2 pathogenicity islands in *Salmonella enterica* Typhimurium. Originally published as Figures 3D and 3E in S. Lucchini, G. Rowley, M.D. Goldberg, D. Hurd, M. Harrison, and J.C. Hinton, 'H-NS Mediates the Silencing of Laterally Acquired Genes in Bacteria', *PLoS Pathogens* 2 (2006), e81, Creative Commons Attribution License.

65 For an early review of H-NS, see C.F. Higgins, J.C. Hinton, C.S. Hulton, T. Owen-Hughes, G.D. Pavitt, and A. Seirafi, 'Protein H1: a role for chromatin structure in the regulation of bacterial gene expression and virulence?', *Molecular Microbiology* 4 (1990), 2007–2012. https://doi.org/10.1111/j.1365-2958.1990.tb00559.x

66 W.W. Navarre, S. Porwollik, Y. Wang, M. McClelland, H. Rosen, S.J. Libby, and F.C. Fang, 'Selective Silencing of Foreign DNA with Low GC Content by the H-NS Protein in Salmonella', *Science* 313 (2006), 236–238. https://doi.org/10.1126/science.1128794

67 S. Lucchini, G. Rowley, M.D. Goldberg, D. Hurd, M. Harrison, and J.C. Hinton, 'H-NS Mediates the Silencing of Laterally Acquired Genes in Bacteria', *PLoS Pathogens* 2 (2006), e81. https://doi.org/10.1371/journal.ppat.0020081

A few years later, Christina Kahramanoglou and colleagues, in a piece of work I was involved in, showed—using genome-scale techniques again—that H-NS binds to A+T-rich horizontally-acquired genes in *E. coli* and represses their expression.[68] However, unlike typical TFs, H-NS binding regions on the DNA extend over long stretches. The longer the stretch of DNA bound by H-NS, the more likely that a gene located at or proximal to the bound DNA would be transcriptionally silenced. Thus, the silencing of horizontally-acquired genes by H-NS requires many molecules of the protein to bind adjacently, essentially covering long stretches of DNA like beads on a string.

Studies of H-NS in *E. coli*, unlike those in *Salmonella*, rarely reported lethality when the protein was removed. Some studies observed a slight reduction in growth rates, and others none. Despite performing similar functions in *Salmonella* and *E. coli*, H-NS is essential for bacterial survival in one and not the other. These observations suggest that there is something else in the genome of *E. coli*, presumably differing in some way from the contents of the *Salmonella* genome, that minimises the impact of the loss of H-NS on survival. What might this be? The *E. coli* genome encodes a protein called StpA, which is similar in sequence to H-NS. StpA also binds to A+T-rich DNA sequences, but is produced by *E. coli* cells in much smaller quantities than H-NS. Removing StpA from *E. coli* cells has little, if any, impact on growth, at least in laboratory conditions. Ebru Uyar and colleagues identified sites on the chromosome within *E. coli* cells that are bound by H-NS and StpA in the presence and absence of the other protein.[69] They first found that the binding of H-NS to the *E. coli* chromosome is unaffected by the presence or absence of StpA. StpA binds to the same locations as H-NS when the latter is also present. However, when H-NS is removed from cells, StpA loses its ability to bind to as many as two-thirds of its sites. Uyar and co-workers further suggest that this reduction in binding of StpA to the chromosome in the absence of H-NS probably reflects the intrinsic binding properties of StpA. They also propose that H-NS can induce changes in the structure of the DNA that it binds to, which might enable StpA to bind to these regions of the chromosome. The loss of such DNA structural features in the absence of H-NS might reduce the ability of StpA to bind to it.

Does the binding of StpA to a small subset of its targets in the absence of H-NS help maintain the repression of these genes? If so, is there some basis to *which* subset of H-NS regulated genes are 'chosen' for repression by the StpA-dependent backup regulatory system? Rajalakshmi Srinivasan in my lab attempted to address this question. She first asked what effect the removal of one or both proteins will have on the expression levels of horizontally-acquired genes in *E. coli*.[70] Consistent with the

68 C. Kahramanoglou, A.S. Seshasayee, A.I. Prieto, D. Ibberson, S. Schmidt, J. Zimmermann, V. Benes, G.M. Fraser, and N.M. Luscombe, 'Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli', *Nucleic Acids Research* 39 (2011), 2073–2091. https://doi.org/10.1093/nar/gkq934

69 E. Uyar, K. Kurokawa, M. Yoshimura, S. Ishikawa, N. Ogasawara, and T. Oshima, 'Differential binding profiles of StpA in wild-type and hns mutant cells: a comparative analysis of cooperative partners by chromatin immunoprecipitation-microarray analysis', *Journal of Bacteriology* 191 (2009), 2388–2391. https://doi.org/10.1128/jb.01594-08

70 R. Srinivasan, D. Chandraprakash, R. Krishnamurthi, P. Singh, V. Scolari, S. Krishna, and A.S.N.

findings of Uyar and colleagues, she found that loss of StpA affects the expression of very few genes. However, when H-NS is removed, the expression of many of its targets—including horizontally-acquired genes—greatly increases. Most importantly, when both H-NS and StpA are removed, many other genes—including those bound by H-NS but unaffected by the loss of H-NS alone—increase in expression. Srinivasan and colleagues compared their gene expression data with Uyar and colleagues' chromosome binding data for H-NS and StpA. In doing so, they found that genes whose expression increases when StpA is removed from *E. coli* already lacking H-NS are often those to which StpA remains bound in the absence of H-NS.

In an earlier study, Blair Gordon and colleagues had measured the affinity of H-NS to tens of thousands of 8-mer DNA sequences.[71] Using these data, Srinivasan and colleagues noted that regions to which StpA binds in the absence of H-NS tend to have a high density of 8-mers that display high-affinity binding to H-NS—and, by inference, to StpA. This finding offers a biophysical rationale for how StpA is able to retain its ability to bind to some but not all its wildtype sites. Srinivasan and co-workers also found that while the loss of H-NS alone has very little effect on growth of *E. coli* under the conditions they had tested in the lab, the loss of both H-NS and StpA resulted in a large growth impairment. Thus, StpA binds to a subset of H-NS-repressed, horizontally-acquired genes in the absence of H-NS and dampens their over-expression when H-NS is lost from the system. This backup function of StpA also helps soften the adverse effect of the loss of H-NS on bacterial growth. Thus, these results lead to the suggestion that keeping horizontally-acquired genes—to which StpA binds in the absence of H-NS—transcriptionally silent is important to ensure that the loss of H-NS on its own does not severely impair the growth of *E. coli*.

Srinivasan and co-workers also observed that genes that are repressed by StpA in the absence of H-NS: (a) are expressed at very low levels, lower than genes repressed by H-NS but not by StpA in the absence of H-NS; (b) transition to very high expression levels when H-NS and StpA are removed from the system. These two findings suggest that these horizontally-acquired genes that are silenced by both H-NS and StpA have a high intrinsic ability for transcription. Transcription at these genes may not even produce full length mRNA.[72] Instead, the high A+T content of these genes, in the absence of H-NS/StpA, exposes many promoter-like elements within gene sequences. This attracts RNAP, causing it to waste resources by performing useless transcription. Going by the bioenergetic cost calculations by Lynch and Marinov that we discussed in Chapter 3, these genes—if left unregulated—would carry a very large negative

Seshasayee, 'Genomic analysis reveals epistatic silencing of ''expensive'' genes in Escherichia coli K-12', *Molecular Biosystems* 9 (2013), 2021–2033. https://doi.org/10.1039/c3mb70035f

71 B.R. Gordon, Y. Li, A. Cote, M.T. Weirauch, P. Ding, T.R. Hughes, W.W. Navarre, B. Xia, and J. Liu, 'Structural basis for recognition of AT-rich DNA by unrelated xenogeneic silencing proteins', *Proceedings of the National Academy of Sciences USA* 108 (2011), 10690–10695. https://doi.org/10.1073/pnas.1102544108

72 S.S. Singh, N. Singh, R.P. Bonocora, D.M. Fitzgerald, J.T. Wade, and D.C. Grainger, 'Widespread suppression of intragenic transcription initiation by H-NS', *Genes and Development* 28 (2014), 214–219. https://doi.org/10.1101/gad.234336.113

selection coefficient ($s << 0$) and be detrimental to the cell if not eliminated. And the role of the silencing system orchestrated by H-NS and StpA is to ensure that they do not get expressed inappropriately when not required. In fact, as demonstrated by Marie Doyle and co-workers, some horizontally-acquired plasmids—accessory genetic elements found in many copies in cells—encode their own version of H-NS to ensure that the expression of plasmid-borne genes is kept under control without syphoning off the host cell's endogenous H-NS reserves.[73] This function, by ensuring that the use of H-NS to silence plasmid-encoded genes does not compromise its function on chromosomal DNA, enables the maintenance of the plasmid in its host.

Though many horizontally-acquired genes may be detrimental to the cell if allowed to be transcribed under favourable growth conditions, there is evidence that some of these genes might in fact be beneficial ($s >> 0$) under stress. For example, several genes normally repressed by H-NS in *Salmonella* are required for virulence, but can negatively impact growth when expressed under benevolent laboratory conditions. Positive *s* under certain conditions could ensure that these genes are maintained in the bacterial population. Additional regulatory mechanisms, such as anti-H-NS proteins that displace H-NS from the DNA, can relieve repression by H-NS and StpA precisely when necessary.[74]

Salmonella, in which the deletion of H-NS is lethal, also encodes an StpA. However, it does not seem to be capable of supporting bacterial survival and growth in the absence of H-NS. Why would this be so? Sabrina Ali and colleagues allowed *Salmonella* lacking H-NS[75] to grow in the laboratory in such a way that these slow-growing populations could accumulate additional mutations.[76] Some of these mutations would be adaptive, allowing the bearer to grow faster than its parent. Such adaptive mutations would soon dominate in the population, as predicted by natural selection. This would then allow investigators to discover mechanisms by which *Salmonella* can compensate for growth defects caused by the loss of H-NS. Ali et al. found that the loss of pathogenicity islands—which are usually kept silent by H-NS but are expressed at high levels in an inappropriate manner in the absence of the repressor—allows *Salmonella* lacking H-NS to adapt to the loss of the repressor. This finding reinforces the idea that improper expression of horizontally-acquired virulence genes, under conditions in which virulence has no role to play in the organism's lifestyle, can be costly. It also shows that evolution would quickly result in the loss of such expensive pieces of DNA if they happened to reside and be expressed inside bacterial cells when not required. In addition, these researchers found that mutations in StpA also allowed

73 M. Doyle, M. Fookes, A. Ivens, M.W. Mangan, J. Wain, and C.J. Dorman, 'An H-NS-like stealth protein aids horizontal DNA transmission in bacteria', *Science* 315 (2007), 251–252. https://doi.org/10.1126/science.1137550

74 D.M. Stoebel, A. Free, and C.J. Dorman, 'Anti-silencing: overcoming H-NS-mediated repression of transcription in Gram-negative enteric bacteria', *Microbiology* 154 (2008), 2533–2545. https://doi.org/10.1099/mic.0.2008/020693-0

75 As well as σS. *Salmonella* lacking H-NS can survive in the absence of σS activity.

76 S.S. Ali, J. Soo, C. Rao, A.S. Leung, D.H. Ngai, A.W. Ensminger, and W.W. Navarre, 'Silencing by H-NS potentiated the evolution of Salmonella', *PLoS Pathogens* 10 (2014), e1004500. https://doi.org/10.1371/journal.ppat.1004500

the H-NS-negative *Salmonella* to adapt to the absence of H-NS (Fig. 5.8A). StpA in *Salmonella* is not identical in sequence to that in *E. coli*. A quick look at the two StpA sequences showed that some of the mutations which allowed *Salmonella* to adapt to the loss of H-NS targeted amino acid positions at which the *Salmonella* StpA differed from the *E. coli* StpA.[77] An open question from this analysis is whether the Salmonella *StpA*, in its normal form, is unable to act as an effective backup for H-NS, and whether the mutations discovered by Ali and co-workers allow it to do so!

Rajalakshmi Srinivasan in my lab performed an experiment similar to that by Sabrina Ali and colleagues, but for *E. coli* lacking both H-NS and StpA.[78] She expected to see losses of segments of horizontally-acquired DNA in populations displaying adaptation to the absence of H-NS and StpA. However, this did not happen. Instead, she first observed that mutations that inactivate σS emerged; this was not surprising in light of prior evidence linking H-NS and σS. Yet, this did highlight an important point which we will examine shortly: that mutations which perturb portions of the transcriptional regulatory network can be adaptive.

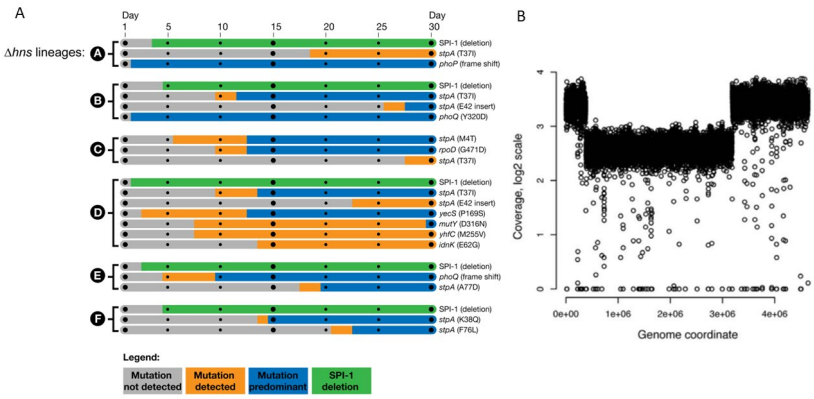


Fig. 5.8. Compensation for the loss of the H-NS gene silencing system. (A) This figure shows that mutations that change StpA and those that delete clusters of pathogenicity-related genes compensate for the loss of H-NS in *Salmonella*. Originally published as Figure 3 in S.S. Ali, J. Soo, C. Rao, A.S. Leung, D.H. Ngai, A.W. Ensminger, and W.W. Navarre, 'Silencing by H-NS potentiated the evolution of Salmonella', *PLoS Pathogens* 10 (2014), e1004500, Creative Commons Attribution License. (B) This figure shows that a duplication of nearly 40% of the chromosome centred around *ori* (which is located at ~3.9e06 on the x-axis) partially compensated for the lack of H-NS and StpA in *E. coli*. Originally published as Figure 3E in R. Srinivasan, V.F. Scolari, M.C. Lagomarsino, and A.S.N. Seshasayee, 'The genome-scale interplay amongst xenogene silencing, stress response and chromosome architecture in *Escherichia coli*', *Nucleic Acids Research* 43 (2005), 295–308, CC BY 4.0.

77 An analysis I had quickly performed when Ali et al., 2014 was published.

78 R. Srinivasan, V.F. Scolari, M.C. Lagomarsino, and A.S.N. Seshasayee, 'The genome-scale interplay amongst xenogene silencing, stress response and chromosome architecture in *Escherichia coli*', *Nucleic Acids Res*. 43 (2015), 295–308. https://doi.org/10.1093/nar/gku1229

A second mutation that emerged, partly compensating for the loss of H-NS and StpA, was a duplication of nearly 40% of the chromosome (Fig. 5.8B). This mutation, while increasing the expression of many genes in the duplicated segment of the chromosome, also caused a strong reduction in the expression of horizontally-acquired genes that had been derepressed by the loss of H-NS and StpA. This underlined the idea that rearrangements of large parts of the chromosome can also be adaptive, and can compensate in part for the loss of a global regulatory network. Keeping in mind the fact that the H-NS-StpA system primarily represses horizontally-acquired genes, we can now ask the following questions: are genes of different functions and of different evolutionary origins positioned differently on the chromosome, and how does this arrangement interplay with gene expression? We will address these questions in the final section of this book. But before that, we make a detour and ask how transcriptional regulatory networks evolve.

5.5. Evolving regulation

Evolution, via changes in the sequence of the genome, is central to adaptation and to the continued existence of life on our planet. Over shorter timescales, gene regulation is a physiological response that allows an organism to react to fast-changing circumstances. The repertoire and function of the machinery responsible for gene expression and its regulation are also subject to change through genetic evolution, even as organisms explore new niches and lifestyles. Even the RNAP, despite being a highly conserved and essential multi-subunit protein, shows variation in the sequences of its component subunits across bacteria, and some of these variations are adaptive. Even closely-related bacteria, and members of the same species, show such variations!

As discussed in Chapter 4, mutations in the main enzymatic subunit of the RNAP confer resistance to the antibiotic rifampicin, most notably in the pathogen *Mycobacterium tuberculosis*. A catalogue of known antibiotic resistance mutations in this pathogen, made publicly available by the World Health Organisation on their website,[79] includes nearly 25 entries (plus several hundred more whose significance for resistance is unclear) for the core subunits of the RNAP. These mutations, in all likelihood, inhibit the binding of the antibiotic to its target protein, the catalytic component of the RNAP, or make the RNAP impervious to interactions with the antibiotic. In addition, the presence of such mutations, at least in one example, is associated with the elevated expression of transporter proteins that throw the antibiotic out of the cell.[80] It is not by any means clear that this effect on gene expression is a direct and specific consequence of the mutation in RNAP and not merely a feedback mechanism arising from extended exposure of these resistant bacteria to the antibiotic. Nevertheless, this does support a role for gene expression changes, which

79 https://www.who.int/publications/i/item/9789240028173

80 G.J. de Knegt, O. Bruning, M.T. ten Kate, M. de Jong, A. van Belkum, H.P. Endtz, T.M. Breit, I.A. Bakker-Woudenberg, and J.E. de Steenwinkel, 'Rifampicin-induced transcriptome response in rifampicin-resistant Mycobacterium tuberculosis', *Tuberculosis* 93 (2013), 96–101. https://doi.org/10.1016/j.tube.2012.10.013

may or may not be directly linked to mutations in the RNAP, in resistance to rifampicin.

In *E. coli*, mutations in the core or the σ subunits of the RNAP are not uncommon. Yasmin Cohen and Ruth Hershberg found that *E. coli* populations adapting to new, uncomfortable environments—most commonly exposure to antibiotics or high temperatures—found mutations in core RNAP subunits (Fig. 5.9).[81] These mutations often affected amino acid residue positions that are otherwise conserved across RNAPs from thousands of different *E. coli* types. These residues also happen to be present close to the active centre of the RNAP enzyme. With conserved amino acid residues usually being important for some crucial aspect of protein function, one can assume with reasonable confidence that most of these RNAP core mutations change RNAP activity in some way, and presumably in a manner that allows the bacterium to improve in the environment concerned. A particular example of a circumstance in which mutations in both the core and the σ subunits of the RNAP confer a selective advantage on the bacterium is late in the stationary phase, during which the environment is inimical to bacterial population growth.

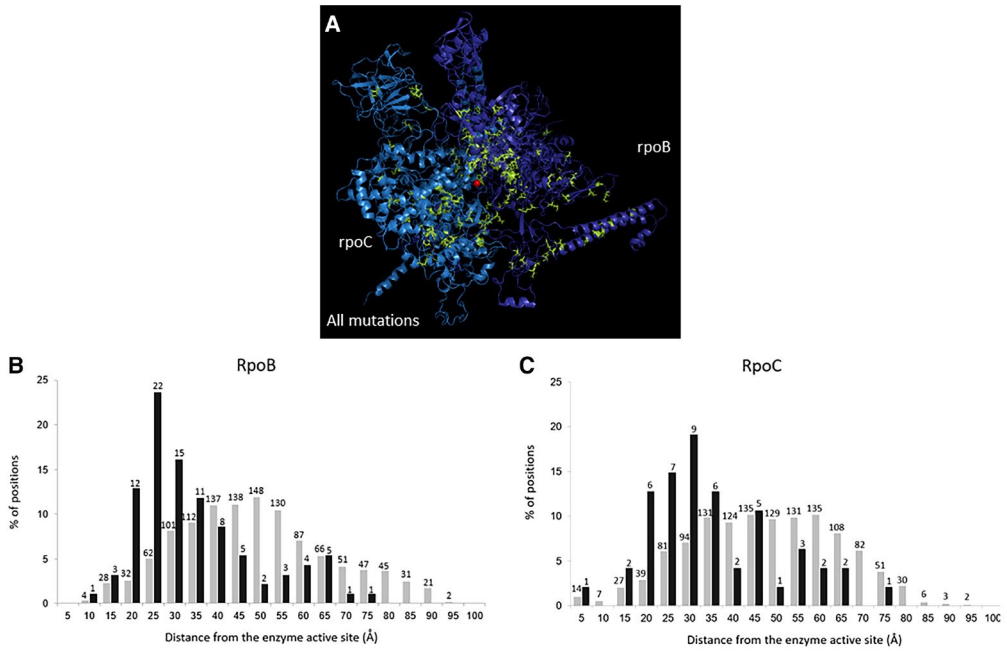


Fig. 5.9. Adaptive mutations in the RNAP. (A) This figure shows the structure of the RNAP, highlighting sites showing adaptive mutations in light green. (B) and (C) These figures show that adaptive mutations in the RNAP occur close to the active site of the enzyme, which is involved in performing the transcription reaction. Originally published as Figure 2 in Y. Cohen and R. Hershberg, 'Rapid Adaptation Often Occurs through Mutations to the Most Highly Conserved Positions of the RNAP Core Enzyme', *Genome Biology and Evolution* 14 (2022), evac105, CC BY 4.0.

81 Y. Cohen and R. Hershberg, 'Rapid Adaptation Often Occurs through Mutations to the Most Highly Conserved Positions of the RNAP Core Enzyme', *Genome Biology and Evolution* 14 (2022), evac105. https://doi.org/10.1093/gbe/evac105

Many bacteria survive long in the stationary phase, and many do so by developing into 'spores' or forms that are dormant but hardy and ready to explode into normal, rapid growth when the environment changes. *E. coli*, despite lacking the ability to form spores, can survive for years in a stationary phase! A few days after entering stationary phase, non-growing *E. coli* cells start to die. However, this phase of death does not fully wipe out the population. Some 1% of the population manages to survive. Maria Zambrano and co-workers[82] showed that after 10 days of stationary phase, the survivors include mutants of the stress-responsive σ-factor σ^S, a mutation we will refer to here as σ^S -GASP1.[83] The mutation is a duplication of a portion of the gene for σ^S, which produced a variant protein with much lower expression and, therefore, host cells with reduced σ^S activity. The activity of σ^S-GASP1 in these cells, though reduced, was not fully abolished. Bacteria carrying this mutation multiplied in the nutrient-depleted stationary phase environment and outcompeted their parents lacking the mutation. Thus, the variant σ^S-GASP1 protein allowed the desperate *E. coli* population to adapt to its trying circumstances. The phenomenon in which a newly emerging variant/mutant outcompetes its parent while growing its population deep in the stationary phase has been termed *GASP*, an acronym for Growth Advantage in Stationary Phase.

Deactivating mutations in σ^S also emerge in *E. coli* lacking H-NS and StpA introduced earlier in this chapter, and partially offset the growth defect caused by the loss of these proteins. In fact, the lethality of the loss of H-NS in *Salmonella* can be reduced by de-activating mutations in σ^S. Thus, σ^S mutations appear to emerge and provide adaptive benefits to bacteria in multiple contexts, often those that cause prolonged slow, very suboptimal, growth.[84] As argued by Thomas Ferenci, such circumstances call for a delicate balance between growth-promoting and stress-responsive functions,[85] which are enabled by opposing σ-factor activities—σ^D and σ^S respectively, as described earlier. Full σ^S activity may not permit even the slow growth that a non-benevolent environment might still allow. Complete abolition of σ^S activity, on the other hand, might make *E. coli* extremely vulnerable to the hazards presented by the same environment. Thus, it might be careful fine-tuning of σ^S activity and balance in the competition between σ^D and σ^S that is called for under such circumstances. This might very well occur in *E. coli* populations entering a deep stationary phase.[86]

Considerable work, published in the years following the publication of Zambrano and colleagues' work, showed that as the stationary phase progresses, fresh variants

82 M.M. Zambrano, D.A. Siegele, M. Almiron, A. Tormo, and R. Kolter, 'Microbial competition: *Escherichia coli* mutants that take over stationary phase cultures', *Science* 259 (1993), 1757–1760. https://doi.org/10.1126/science.7681219

83 This mutation is often referred to in the published literature as *rpoS819*.

84 T. Ferenci, 'What is driving the acquisition of mutS and rpoS polymorphisms in Escherichia coli?', *Trends in Microbiology* 11 (2003), 457–461. https://doi.org/10.1016/j.tim.2003.08.003

85 T. Ference, 'Maintaining a healthy SPANC balance through regulatory and mutational adaptation', *Molecular Microbiology* 57 (2005), 1–8. https://doi.org/10.1111/j.1365-2958.2005.04649.x

86 P. Nandy, 'The role of sigma factor competition in bacterial adaptation under prolonged starvation', *Microbiology* 168 (2022), 001195. https://doi.org/10.1099/mic.0.001195

keep emerging, each trying to thrive while outcompeting their parents and driving them to extinction.[87] Thus, the stationary phase for *E. coli* is not as stationary as its name would indicate. Instead, it is characterised by the aggregate lives of ever-emerging and disappearing variants of *E. coli* engaging in various interactions among themselves while adapting, surviving, and growing in the dynamic, yet increasingly forbidding environment. Till recently, however, the genetic composition or diversity of the *E. coli* population inhabiting this changing environment remained uncharacterised. Savita Chib and Farhan Ali in my lab sought to fill this gap.[88]

Chib and Ali sequenced the genomes of five *E. coli* populations maintained in stationary phase for four weeks at several time points over the lifetime of these populations.[89] They identified several mutations appearing in these populations and found that most waxed and waned over these four weeks. All in all, the genetic diversity of these populations—defined by the number of different mutations seen in a population—increased consistently over time. This increase in genetic diversity was not random and occurred under selection. A more recent study by Sophia Katz and colleagues, interrogating the genomes of multiple *E. coli* populations kept in stationary phase for a staggering three years, also showed that some populations acquired the ability to mutate at a faster rate over time.[90] Both studies found that the same mutations appeared across multiple populations. This suggested that the same genetic strategies enable adaptation to deep stationary phases in independent populations, pointing to the repeatability of evolutionary strategies in these environments. Importantly in the context of the present discussion, both studies identified several mutations in the RNAP subunits emerging in these populations. In fact, Katz et al. showed that over 90% of all genomes sequenced in their study, across multiple independent populations over several years, carried a mutation in a core RNAP subunit.

Pabitra Nandy in my lab, following up on the work by Chib and Ali, noticed that *E. coli* with a mutation in core RNAP had appeared after around three weeks of maintained stationary phase in the population.[91] This mutation was found alongside another in σ^S, which we will call σ^S-GASP2 for it is a variant of σS-GASP1 described above. This mutant was hardy and slow-growing, and yet able to outcompete its faster-growing (in rich media) ancestors in highly limiting stationary phase media—but not in media from, say, *E. coli* cultures grown for only a few days in stationary

87 Reviewed in S.E. Finkel, 'Long-term survival during stationary phase: evolution and the GASP phenotype', *Nature Reviews Microbiology* 4 (2006), 113–120. https://doi.org/10.1038/nrmicro1340

88 S. Chib, F. Ali, and A.S.N. Seshasayee, 'Genomewide mutational diversity in Escherichia coli population evolving in prolonged stationary phase', *mSphere* 2 (2017), e00059–17. https://doi.org/10.1128/msphere.00059-17

89 Ibid.

90 S. Katz, S. Avrani, M. Yavneh, S. Hilau, J. Gross, and R. Hershberg, 'Dynamics of adaptation during three years of evolution under long-term stationary phase', *Molecular Biology and Evolution* 38 (2021), 2778–2790. https://doi.org/10.1093/molbev/msab067

91 P. Nandy, S. Chib, and A. Seshasayee, 'A Mutant RNA Polymerase Activates the General Stress Response, Enabling *Escherichia coli* Adaptation to Late Prolonged Stationary Phase', *mSphere* 5 (2020), e00092–20. https://doi.org/10.1128/msphere.00092-20

phase. The σS-GASP2 mutant showed higher σ^S activity than σ^S-GASP1, as measured by the extent to which genes known to be expressed under the action of RNAP- σ^S holoenzyme changed in expression between the two σ^S mutants. Now, the RNAP core mutation somehow enhanced the degree to which known σ^S targets are expressed, irrespective of whether σ^S-GASP1 or σ^S-GASP2 was present. This ability of the RNAP core mutation, however, required some residual σ^S activity and was lost in the complete absence of σ^S. This suggested that the RNAP core mutation differentially affected gene expression via σ^S, presumably by tilting the σ^S–σ^D competition in favour of σ^S. The RNAP mutation did not clearly enhance the stationary phase growth of *E. coli* carrying σ^S-GASP2; however, the σ^S-GASP1 + RNAP core double mutant considerably outperformed the σ^S-GASP1 single mutant, suggesting that the RNAP core mutation might have emerged in a σ^S-GASP1 background and that σ^S-GASP2 developed later. Together, these findings present the argument that the balance between growth and stress response shifts in favour of the latter as stationary phase progresses, and that this balance can be tuned by mutations not only in the relevant σ-factor genes but also those in the core RNAP.

The bacterial regulatory network, though constrained by the capabilities of the RNAP and the σ-factors, is driven by a plethora of regulatory proteins such as TFs and their interactions with the DNA. Mutations in any of these components can influence gene expression and, in some instances, do so in an adaptive manner. A few mutations in a TF can easily change the binding properties of a TF, as demonstrated by Ryan Schultzaberger and colleagues using artificially-generated sequence variants of a bacterial TF.[92] Let us, to begin with, examine the phenomenon of GASP once more. Chib and Ali, while analysing their data on mutations emerging over four weeks in stationary phase *E. coli* populations, found several mutations in proteins involved directly or indirectly in the regulation of gene expression. This was in addition to mutations in core RNAP and σ-factors. They found that regulatory proteins were in fact more likely to be altered by adaptive mutations than non-regulatory proteins. These included an enzyme responsible for the degradation of cyclic AMP, a small molecule whose levels depend on the availability of glucose. Mutations in this gene were observed in several independent stationary phase *E. coli* populations, pointing to an important role for this modification in stationary phase growth. Cyclic AMP binds to and activates the global TF CRP, which in turn regulates the expression of hundreds of genes. CRP, being required for the metabolism of many unusual carbon sources, can be expected to play a role in gene expression affecting stationary phase survival. It is therefore not unreasonable to expect that this mutation in the enzyme that degrades cyclic AMP will result in an increase in the levels of this small molecule and will thereby change the expression of some genes under the control of CRP. In fact, a very recent study by Shira Zion and

92 R.K. Shultzaberger, S.J. Maerkl, J.F. Kirsch, and M.B. Eisen, 'Probing the informational and regulatory plasticity of a transcription factor DNA-binding domain', *PLoS Genetics* 8 (2012), e1002614. https://doi.org/10.1371/journal.pgen.1002614

colleagues reported that multiple *E. coli* populations gain mutations in CRP, which further potentiate the emergence of many other secondary mutations, deep into stationary phase.[93] Consistent with the findings of Chib and Ali, Nicole Ratib and co-workers also found many mutations in regulatory proteins in *E. coli* maintained in stationary phase for nearly three years.[94]

In an early work, Erik Zinser and Roberto Kolter showed that a mutation in the global TF Lrp provides a growth advantage during stationary phase.[95] Lrp is primarily a regulator of genes involved in amino acid metabolism.[96] In particular, it represses the expression of genes that help in the breakdown of certain amino acids, and those involved in the uptake of peptides (short chains of amino acid residues). The GASP mutation in Lrp, which abolishes the DNA binding activity of the protein, may enable increased growth in nutrient-deprived stationary phase conditions by increasing amino acid breakdown, leading to their increased utilisation as sources of nutrition and energy instead of mere building blocks of proteins.

In a more recent piece of work, Savita Chib and Subramony Mahadevan showed that a mutation in H-NS, the silencer of horizontally-acquired genes introduced earlier, conferred the GASP phenotype.[97] The mutation they discovered reduced the activity of H-NS. Among genes whose expression was derepressed by decreased H-NS activity are those involved in the utilisation of unusual sugars as carbon sources. These sugars may be unusual in standard laboratory media, but may become available as cells metabolise and excrete their way into a deep stationary phase. Taken together, mutations in Lrp and in H-NS allow the simultaneous expression of genes that help cells find unusual sources of nitrogen and carbon respectively, something that they would not do during normal, rapid growth, and are selected for during deep stationary phases. Since these genes are typically under the control of σ^D, their activation is further enabled in σ-factor mutations that tune the balance between growth-promoting nutrient utilisation programmes and stress responses.

Mutations in TFs enable adaptation in other circumstances as well, such as in antibiotic resistance. A commonly cited example of a TF with a role in antibiotic resistance is MarR.[98] MarR indirectly, through the action of another TF whose

93 S. Zion, S. Katz, and R. Hershberg, '*Escherichia coli* adaptation under prolonged resource exhaustion is characterized by extreme parallelism and frequent historical contingency', *PLoS Genetics* 20 (2024), e1011333. https://doi.org/10.1101/2024.03.21.586114

94 N.R. Ratib, F. Seidl, I.M. Ehrenreich, and S.E. Finkel, 'Evolution in Long-Term Stationary-Phase Batch Culture: Emergence of Divergent Escherichia coli Lineages over 1,200 Days', *mBio* 12 (2021), e03337–20. https://doi.org/10.1128/mbio.03337-20

95 E.R. Zinser and R. Kolter, 'Prolonged stationary phase incubation selects for lrp mutations in *Escherichia coli* K12', *Journal of Bacteriology* 182 (2000), 4361–4365. https://doi.org/10.1128/jb.182.15.4361-4365.2000

96 J.M. Calvo and R.M. Matthews, 'The leucine responsive regulatory protein, a global regulator of metabolism in *E. coli*', *Microbiology Reviews* 1994 (1994), 466–490. https://doi.org/10.1128/mmbr.58.3.466-490.1994

97 S. Chib and S. Mahadevan, 'Involvement of the global regulator H-NS in the survival of *Escherichia coli* in stationary phase', *Journal of Bacteriology* 194 (2012), 5285–5293. https://doi.org/10.1128/jb.00840-12

98 G.A. Beggs, R.G. Brennan, and M. Arshad, 'MarR family proteins are important regulators of clinically relevant antibiotic resistance', *Protein Science* 29 (2020), 647–653. https://doi.org/10.1002/pro.3769

expression it directly controls, represses the expression of efflux pumps, which eject antibiotics and other toxic molecules out of the cell with broad specificity. Tens of mutations that reduce MarR activity to different extents have been associated with resistance to unrelated antibiotics such as ciprofloxacin, trimethoprim, tetracycline, and several more, in the laboratory as well as in clinically-relevant contexts.

Mutations in TFs mediate antibiotic resistance, albeit indirectly by affecting the expression of some other protein which may then eliminate the antibiotic in some manner. In contrast, high levels of resistance to antibiotics such as ciprofloxacin can be readily attained via precise mutations in the protein that the antibiotic targets, resulting in reduced binding of the antibiotic to its target. We have already noted how mutations in RNAP can confer resistance to rifampicin, which acts by binding to RNAP. The set of such mutations is small however, for these affect highly conserved, essential proteins. Mutations that inadvertently reduce the activity of such an essential target protein while protecting it from an antibiotic could have severe consequences for the cell. This again does not mean that the alternative mutations—which affect efflux pump expression by inactivating a TF—are without a cost, though the number of mutations that have the same outcome of inactivating a protein can be quite large and therefore these mutations can be discovered more easily. As discussed earlier in this text, inappropriate expression of any given gene can potentially be expensive to bacterial lineages that are part of large populations. Thus, it is often *combinations* of mutations walking the tightrope between antibiotic resistance and general fitness that establish themselves in a population. For instance, Lisa Alzrigat and colleagues showed that clinical samples of *E. coli* resistant to ciprofloxacin carry a combination of mildly-inactivating mutations in MarR in addition to mutations in ciprofloxacin's direct target, DNA gyrase.[99] The conclusion arising from this study is that constitutive, high expression of efflux pumps may impose a cost that effectively selects against strongly inactivating mutations in the TF MarR.

One can surmise that environments which present a range of diverse toxic substances would favour robust inactivation of TFs such as MarR, for high, persistent expression of a broad-specificity efflux pump may be a more efficient solution to the problem presented by such environments than mutations in all possible target proteins. However, an alternative possibility is that large genetic elements carrying multiple antibiotic resistance genes are acquired horizontally. The very evolution, let alone the spread, of such a complex series of genes would require persistent selection by consistent exposure to multiple antimicrobial agents. Unfortunately, in response to antibiotic abuse that has created environments replete with antibiotics and other toxins, these genetic modules are becoming increasingly common in bacterial populations.

Richard Lenski's pioneering long-term experimental evolution (LTEE, see Chapter 4), during which several independent populations of *E. coli* grew and evolved

99 L.P. Alzrigat, D.L. Huseby, G. Brandis, and D. Hughes, 'Fitness cost constrains the spectrum of marR mutations in ciprofloxacin-resistant *Escherichia coli*', *Journal of Antimicrobial Agents and Chemotherapy* 72 (2017), 3016–3024. https://doi.org/10.1093/jac/dkx270

for tens of thousands of generations, has also provided us with instances of TF evolution. An example involves the global TF and chromosome shaping protein FIS. FIS is expressed at high levels as cells transition from a period of physiological adaptation to a new medium to one of exponential population multiplication, providing a boost to the transcription of growth-enabling genes such as those involved in protein synthesis. Estelle Crozat and colleagues noticed that FIS, alongside the topoisomerase protein topoisomerase 1, had mutated in several LTEE populations.[100] These mutations altered chromosome supercoiling states. This finding suggested that mutations affecting chromosome structure can be adaptive.

In a later study, Crozat again—with other co-workers—discovered that a FIS mutation that had emerged during LTEE had additional, more subtle effects relating to its function as a TF rather than as a chromosome-shaping protein.[101] They found that FIS activates the transcription of a protein—a *porin*, that forms pores on the cell surface, allowing the exchange of material between a cell and its environment—called OmpF. This regulatory interaction had not been described previously. In fact, this interaction is known to be absent in some other varieties of *E. coli*. These *E. coli* varieties have a second porin OmpC in addition to OmpF. The overlapping nature of the functions of OmpC and OmpF means that the high level of OmpF achieved by the activation of its transcription by FIS is not necessary in these *E. coli*. It can be hypothesised that an ancestor of the *E. coli* variety that was used in LTEE lost OmpC and instead co-opted FIS as a transcriptional activator of OmpF. However, this FIS-dependent activation of OmpF was being compromised by the FIS mutation that had evolved during LTEE. Despite this FIS mutation, the levels of OmpF did not quite decrease in the evolved lines, suggesting that additional compensatory mutations elsewhere had kicked in, reducing the dependence of OmpF on FIS. In addition, the expression level of FIS itself had decreased over time during LTEE, suggesting that evolution had reduced the requirement of FIS for the rapid growth of *E. coli*. Why might this be the case? FIS is an extraordinarily abundant TF. At its peak, there are some 60,000 molecules of FIS in the *E. coli* cell, which is several times more than the expression levels of other global TFs such as CRP and H-NS, and comparable to the levels of non-specific DNA binding proteins such as HU that coat the chromosome. Are evolving populations of *E. coli* attempting to minimise the cost of expressing FIS to such high levels by discovering combinations of mutations, in FIS and in other parts of the genome, that compensate for reduced FIS availability in other ways?

It has been assumed in the past that if an ortholog of a TF and that of its target gene in one organism is present in another organism, the regulatory interaction between

100 E. Crozat, N. Philippe, R.E. Lenski, J. Geiselmann, and D. Schneider, 'Long-Term Experimental Evolution in *Escherichia coli*. XII. DNA topology as a key target of selection', *Genetics* 169 (2005), 523–532. https://doi.org/10.1534/genetics.104.035717

101 E. Crozat, T. Hindre, L. Kuhn, J. Garin, R.E. Lenski, and D. Schneider, 'Altered regulation of the OmpF porin by Fis in *Escherichia coli* during an evolution experiment and between B and K-12 strains', *Journal of Bacteriology* 193 (2011), 429–440. https://doi.org/10.1128/jb.01341-10

the TF and the target gene is also conserved.[102] However, the FIS example shows that orthologous TFs can regulate different genes in different organisms. Is this a rule or an exception? Morgan Price and co-workers attempted to answer this question using the regulatory network of *E. coli* as a reference to predict gene expression patterns in *E. coli* as well as in other bacteria.[103] Their work was based on the premise that genes regulated by the same set of TFs will be expressed in the same manner across conditions. As mentioned earlier, this is a good assumption that holds in many cases, despite the incompleteness of the known regulatory network, as demonstrated in the first instance by Price and co-workers, and can lead to useful inferences.[104]

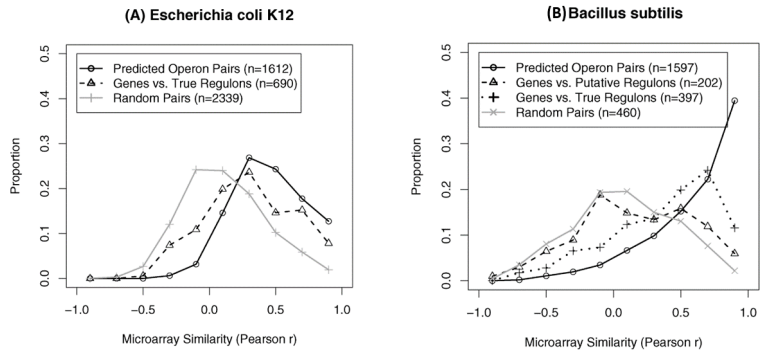


Fig. 5.10 Transcriptional regulatory interactions are not conserved. (A) This figure shows correlation in gene expression between pairs of genes belonging to the same operon and therefore expected to be co-regulated, those regulated by the same TFs, and random pairs; (B) This figure measures the same correlations as A, but for *B. subtilis*. Compare the distribution of co-expression measures between genes that are known to be regulated by the same TFs in *B. subtilis* ('true regulons') and those that are predicted to be co-regulated based on the *E. coli* regulatory network ('putative regulons'). Originally published as Figures 5A and 5E in M. Price, P.S. Dehal, and A.P. Arkin, 'Orthologous transcription factors in bacteria have different functions and regulate different genes', *PLoS Computational Biology* 3 (2007), e175, Creative Commons Public Domain declaration.

Price and colleagues identified orthologs of *E. coli* TFs and their target genes in the evolutionarily distant bacterium *Bacillus subtilis* and asked if pairs of genes[105] predicted to be co-regulated in the latter based on data from the former showing

102 M.M. Babu, S.A. Teichmann, and L. Aravind, 'Evolutionary dynamics of prokaryotic transcriptional regulatory networks', *Journal of Molecular Biology* 358 (2006), 614–633. https://doi.org/10.1016/j.jmb.2006.02.019

103 M. Price, P.S. Dehal, and A.P. Arkin, 'Orthologous Transcription Factors in Bacteria Have Different Functions and Regulate Different Genes', *PLoS Computational Biology* 3 (2007), e175. https://doi.org/10.1371/journal.pcbi.0030175

104 And by other work such as Balazsi et al., 2005.

105 This is a simplified description of Price et al., 2007's work. What they did is the following: for each gene belonging to a *regulon*, i.e., the set of genes regulated by the same TF, they measured the correlation in its expression with the average expression across genes in the same regulon.

correlated expression patterns. They found, across ~200 examples, that predicted groups of co-regulated genes in *B. subtilis* were not necessarily co-expressed, implying that regulatory interactions between TFs and target genes in *E. coli* are not conserved in evolutionarily distant bacteria (Fig. 5.10). This held true even when regulatory interactions known in *E. coli* were used to predict expression patterns in more closely related bacteria such as *Vibrio cholerae*. Thus, orthologs of both a TF and its target gene may be present in another organism, but the regulatory interaction between the two need not be. This is exemplified by the FIS-OmpF pair. Both FIS and OmpF are conserved in two different strains of *E. coli*, whilst FIS regulates the transcription of OmpF in one strain but not the other. Price and colleagues' work suggests that this may not be an exception. Note here though that such differences in regulatory networks can arise not only through mutations in TFs, which can alter the DNA recognition specificity and/or activity of TFs, but also mutations in inter-genic DNA sequences to which these TFs bind.

We have seen several examples of adaptive mutations in TFs. The fact that some TFs regulate several genes implies that mutations in such TFs would affect the expression of all or most of these targets. If the adaptation driven by a TF mutation is determined by only a small subset of the TF's targets, then the change in expression of all the other genes that the TF regulates should be considered as a side-effect and potentially one with negative consequences. Thus, consequential mutations in TFs can be a double-edged sword. Though genetic alteration of TF function might offer early adaptation, would it be maintained over longer timescales? One can surmise that the side-effects of an otherwise beneficial TF mutation may cause such mutations to be selected against as evolution progresses, allowing populations to discover and select for other beneficial combinations of mutations with fewer detrimental side-effects.

To test this, Farhan Ali in my lab investigated sequence variation in *E. coli* TFs at two distinct timescales:[106] a very short timescale represented by ~30 years of LTEE and a longer timescale covering ~100 million years that have elapsed since the divergence of *E. coli* and its relative *Salmonella*. Novel mutations in TFs, especially those that regulate a large number of genes, emerged early during the LTEE. As time progressed and as *E. coli* populations adapted to the environment imposed on them by the experiment, the frequency of new mutations in TFs declined. At the ~100 million year timescale represented by the diversity of *E. coli*, TFs show significantly smaller sequence diversity than the genes that they regulate. This was especially apparent in TFs regulating many genes. Thus, the pleiotropic nature of TF mutations might, over the time of divergence and diversification of a single bacterial species, reduce the extent of variation in TF sequences (Fig. 5.11).

106 F. Ali and A.S.N. Seshasayee, 'Dynamics of genetic variation in transcription factors and its implications for the evolution of regulatory networks in Bacteria', *Nucleic Acids Research* 48 (2020), 4100–4114. https://doi.org/10.1093/nar/gkaa162

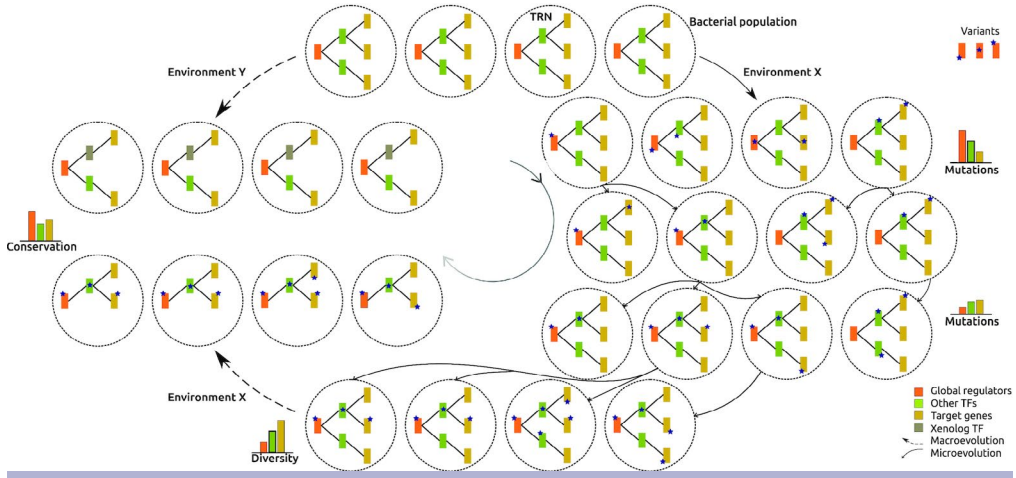


Fig. 5.11 TF evolution in bacteria. A population of bacteria beginning to adapt to a new environment accumulates mutations in TFs, especially global TFs. Over longer timescales, of the order of divergence of a whole species like *E. coli*, TFs show low sequence diversity. As the niche changes to *Y*, adaptation may proceed through TF repertoire changes created by gene loss and horizonal gene transfer. Originally published as Figure 8 in F. Ali and A.S.N. Seshasayee, 'Dynamics of genetic variation in transcription factors and its implications for the evolution of regulatory networks in Bacteria', *Nucleic Acids Research* 48 (2020), 4100–4114, CC BY 4.0.

In the examples we have seen so far, regulatory changes were effected by mutations in TFs. Even if a mutation in a TF is strong enough to inactivate it completely, the gene for the TF is still present. In such cases, there is a chance for TF activity to recover quickly even in the absence of horizontal re-acquisition of a lost TF gene. For example, much of σ^S activity lost in σ^S-GASP1 was recovered in σ^S-GASP2. In many cases, such as TF σ^S,[107] mutations in GASP, or the FIS mutations in LTEE, mutations do not fully abolish the protein's activity and will instead subtly alter it. However, over longer evolutionary distances, even within the period required for the diversification of a species, whole TF genes can be gained or lost. Ali found that TFs showing high sequence divergence within *E. coli* are often lost in related species.[108] These may represent examples of TFs that are lost in some lineages as a result of the inexorable process of sequence decay by mutation. Or these might merely be TFs with high tolerance for mutation that were acquired specifically in *E. coli* and related lineages.

Bacterial genomes encode fairly large numbers of TFs. The numbers of TFs coded for by a bacterial genome increases with increase in genome size, or more precisely, the total number of genes. However, the relationship is not linear, but is closer to being quadratic. Very small bacterial genomes such as those of obligate parasites and endosymbionts code for hardly any TFs and only a single σ-factor. Bacteria like *E. coli* with, say, 4,000 genes in total, contain ~300 TFs—similar if not more than the number of TFs encoded by the genome of the eukaryote budding yeast. Large bacterial genomes

107 We assume that σ-factors are TFs here, although I usually do not assume this.
108 Ali and Seshasayee, 2020.

with ~10,000 genes code for as many as ~1,000 TFs, a number that is comparable to that for higher eukaryotes. The lack of TFs in parasitic or endosymbiotic genomes likely arises from a preferential loss of TFs. As genomes grow, where do new TFs come from? The DNA binding portion of bacterial TFs, called the DNA-binding *domain* (DBD), belong to a fairly small set of sequence *families* that in many cases are variants within a common theme known as the helix-turn-helix (HTH) motif. These DBDs are often found alongside other domains that may help to sense a signal. Proteins belonging to the same family show fairly high sequence similarities with each other. When a family of TFs shows an expansion—i.e., an increase in the number of protein sequences belonging to the family—in a clade of bacteria, gene duplications initiated from a progenitor family member become available as an option. Or, new family members may be gained independently of pre-existing relatives through horizontal gene transfer. Which of these two is more predominant? This is merely a special case of the broader duplication vs horizontal gene transfer debate we had examined in Chapter 4.

While horizontal gene acquisition is believed to be the more dominant force responsible for bacterial genome/gene repertoire expansion overall, arguments in favour of both forces have been advanced in the special case of TFs. Early work by Sarah Teichmann and Madan Babu suggested that duplication dominates TF evolution.[109] Many proteins are modular, comprising of multiple 'domains'. Each domain is usually responsible for one function. For example, a bacterial TF may have one domain that binds to the DNA and another that binds to a signal molecule that directs it to bind to the DNA. Teichmann and Madan Babu assumed that any pair of proteins with the same 'domain architecture'—namely the same set of domains arranged in a particular order from one end of the sequence to the other—are likely to have arisen by duplication. Thus, they concluded that a majority of TFs in *E. coli* have arisen by duplication. Unlike other studies that have compared rates of duplication and horizontal gene transfer in other contexts (see Chapter 4), this study did not compare *E. coli* TF sequences with those from other organisms.

A few years later, Morgan Price and co-workers disagreed with Teichmann and Madan Babu. They built phylogenetic trees of *E. coli* TFs, comparing sequences from *E. coli* with those of similar proteins from other bacterial species.[110] They compared these trees with species trees. Based on the incongruity between species trees and gene trees, they concluded that nearly two-thirds of all *E. coli* TFs had been acquired by horizontal gene transfer (Fig. 5.12). Farhan Ali had also noticed that several TFs with low sequence divergence in *E. coli* are poorly conserved in related species; these are probably useful TFs horizontally acquired in the *E. coli* lineage.[111] Horizontal gene transfer appears

109 S.A. Teichmann and M.M. Babu, 'Gene regulatory network growth by duplication', *Nature Genetics* 36 (2004), 492–496. https://doi.org/10.1038/ng1340

110 M.N. Price, P.S. Dehal, and A.P. Arkin, 'Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*', *Genome Biology* 9 (2008), R4. https://doi.org/10.1186/gb-2008-9-1-r4

111 Ali and Seshasayee, 2020.

to be rare for global TFs with a large number of targets, but common for TFs that are encoded adjacently to their target genes on the genome. The latter group of TFs might have been transferred into *E. coli* as modules comprising of the TF as well as its target genes. Thus, TF repertoires do change over certain phylogenetic distances. These may reflect niche divergence between the species concerned, and the primary means by which TF repertoires expand is likely to be horizontal gene acquisition.

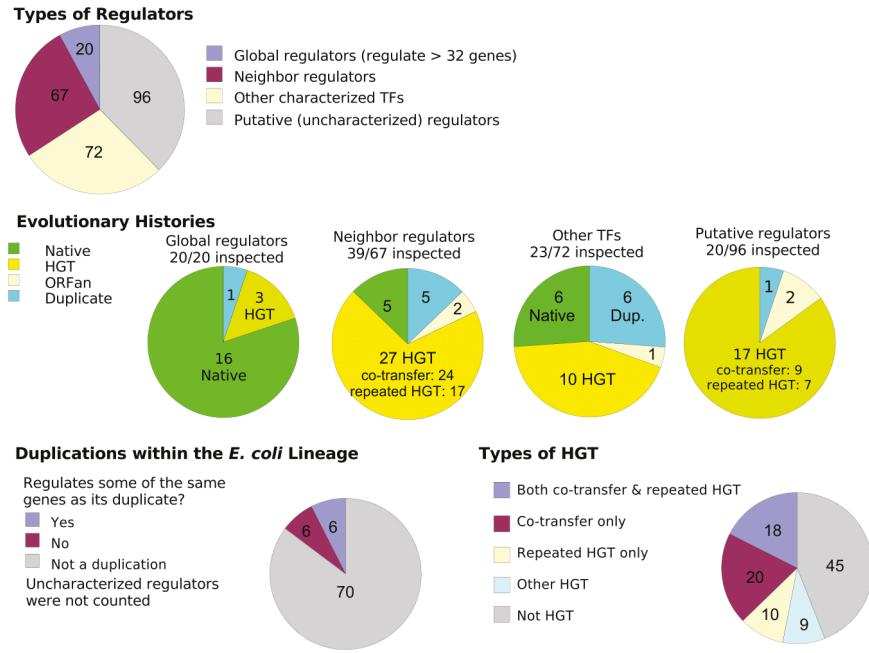


Fig. 5.12 Evolution of TFs by horizontal gene transfer. This figure shows that a majority of TFs in *E. coli* were likely acquired by horizontal gene transfer. This does not hold true for global regulators, most of which are 'native' or not recently acquired. This was published as Figure 6 in Price et al. (2008). *Genome Biology*, CC BY 2.0.

Acquiring even a single regulatory system can have major consequences for an organism's lifestyle. This is exemplified by some beautiful work done some time ago by Mark Mandel and colleagues.[112] They investigated the biology of two varieties of the bacterial species *Vibrio fischeri*. One variety is a symbiont of fish whereas the other colonises squid. Comparing the genomes of the two symbionts revealed that the fish symbiont, which was unable to colonise squid, lacked a protein called RscS that was encoded in the genome of the squid symbiont. RscS is a protein that activates a TF by modifying it,[113] which in turn activates the expression of genes that help the

112 M.J. Mandel, M.S. Wollenberg, E.V. Stabb, K.L. Visick, and E.G. Ruby, 'A single regulatory gene is sufficient to alter bacterial host range', *Nature* 458 (2009), 215–218. https://doi.org/10.1038/nature07660

113 RscS is what is called a sensor kinase that phosphorylates a response regulator protein. The response regulator is a TF.

bacterium colonise surfaces inside squid. The fish symbiont, when engineered by the addition of the gene for RscS, gained the ability to colonise squid. Further analysis of a collection of *V. fischeri* isolates from fish and squid showed that all squid colonisers encoded RscS, whereas only a subset of fish symbionts did, and the RscS found in fish symbionts was very divergent in sequence from that encoded by squid colonisers. The phylogeny of *V. fischeri* suggested that the ancestor of this species likely lacked the ability to colonise squid, and that the acquisition of RscS allowed its descendants to do so. If this acquisition was achieved through horizontal transfer, its source remains unknown. The fact that the addition of a single regulator allowed a bacterium to access a new niche indicated that genes responsible for this colonisation were already part of the ancestral bacterium. The addition of a single master regulator, an organiser or pied-piper, was sufficient to activate these genes in a coordinated manner, enabling the colonisation of a new environment.

We have so far examined how TFs evolve by mutation and how the repertoire of TFs encoded in a genome can change. We finally ask how TFs evolved in the first place. What is the ultimate origin of transcription regulation by TFs? This question is motivated by the argument, articulated by Sandhya Visweswaraiah and Stephen Busby, that "transcription regulation is a 'luxury' for a bacterium."[114] No TF, unless one considers the major σ-factor σD a TF, is part of the hypothetical minimal bacterial genome (Chapter 3). Bacteria with small genomes code for few TFs, and endosymbiont genomes encode hardly any that we know of. As we had briefly mentioned earlier, properties inherent to the chromosome such as DNA supercoiling can regulate the expression of genes,[115] and can probably play a key role in gene regulation in bacteria with highly reduced genomes such as the *Mycoplasma*. The so-called contingency loci in bacteria such as *Helicobacter pylori* and *Campylobacter jejuni* can mutate, reversibly, at rates as high as ~1/20 and help to achieve rapid phenotypic switches at a population level.[116] Thus, TF-based regulation might become important only as bacteria evolve to adopt 'complex' lifestyles—complexity defined as, say, the diversity of environmental and cellular situations the organism has to deal with.

In contrast to transcription regulation, DNA shaping and compaction are likely to be fundamental for survival to even a primitive cell. Even a genome coding for ~100 genes will need to be compacted by two orders of magnitude to be packed inside a tiny cell. Unlike TFs, which recognise particular sequence motifs on the DNA to make sequence-specific interactions with operator sites, most chromosome shaping proteins bind non-specifically. Often, both sequence-specific and nonspecific DNA binding

114 S. Visweswaraiah and S.J.W. Busby, 'Evolution of bacterial transcription factors: how proteins take on new tasks, but do not always stop doing the old ones', *Trends in Microbiology* 23 (2015), 463–467, p. 465. https://doi.org/10.1016/j.tim.2015.04.009

115 For a specific example, see W. Zhang and J.B. Baseman, 'Transcriptional regulation of MG_149, as osmoinducible lipoprotein gene from *Mycoplasma genitalium*', *Molecular Microbiology* 81 (2011), 327–339. https://doi.org/10.1111/j.1365-2958.2011.07717.x

116 J. Parkhill, B.W. Wren, K. Mungall, J.M. Ketley, C. Churcher, et al., 'The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences', *Nature* 403 (2000), 665–668. https://doi.org/10.1038/35001088

proteins adopt similar structural folds, belonging to the same protein 'superfamily', and likely share a common ancestor.[117] At the level of protein function, the boundary between a chromatin-shaping protein (called nucleoid-associated proteins in bacteria) and a TF is not all that clear.[118] For example, the protein FIS, which we had encountered earlier, is not only a TF regulating the expression of many genes by making precise contacts with operator sites near promoters, it also affects topological properties of the DNA such as supercoiling. Several TFs, including FIS and CRP—the latter being an important subject of extended discussion, in Visweswaraiah and Busby's opinion—bind to thousands of sites on the genome, a vast majority of which appear to have nothing to do with the regulation of transcription of a proximal gene. Do these other interactions have anything to do with functions such as chromosome shaping? Likely yes for FIS and maybe for CRP. Maybe chromosome shaping interactions represent the 'original' function of these proteins, with features required for transcription regulation evolving subsequently? These excess interactions, however, are not non-specific contacts per se but are likely sequence-specific contacts with lower affinity targets. Yet this begs the following extrapolation: did an ancestral non-specific or weakly-specific DNA binding protein gain the ability of sequence-specific DNA recognition to evolve into a TF? We can only guess at the moment, and explore whether sequence data publicly available for extant TFs and nucleoid-associated proteins allows us to answer this question systematically.

The evolution of a TF involves much more than just evolving DNA recognition specificity. Take, for example, the protein CRP. This protein not only has the ability to recognise a target motif and bind to it, it also has the capacity to bind to the signal molecule cyclic AMP and then present interfaces that attract the RNAP.[119] Though the DNA-binding and the cyclic AMP-binding domains are distinct and lie in different segments of the protein sequence, one residue in the latter contributes to DNA binding. The binding of cyclic AMP somewhere on the protein should translate to modulation of its DNA binding properties which are encoded elsewhere on the same protein. These point to some correlated evolution between the two domains. Further, CRP has the ability to activate transcription by more than one mechanism depending on which part of the RNAP it interacts with. That means that this protein has evolved multiple RNAP binding interfaces. Assuming that TFs evolved from a nucleoid-associated protein ancestor, all these elements had to have evolved on a non-specific DNA binding protein backbone from scratch. This would have involved a combination of domain sequence evolution and the fusion of distinct domains. The evolutionary

117 N.M. Luscombe, S.E. Austin, H.M. Berman, and J.M. Thornton, 'An overview of the structures of protein-DNA complexes', *Genome Biology* 1 (2000), REVIEWS001. https://doi.org/10.1186/gb-2000-1-1-reviews001

118 C.J. Dorman, M.A. Schumacher, M.J. Bush, R.G. Brennan, and M.J. Buttner, 'When is a transcription factor a NAP?', *Current Opinion in Microbiology* 55 (2020), 26–33. https://doi.org/10.1016/j.mib.2020.01.019

119 Visweswaraiah and Busby, 2015; A. Kolb, S. Busby, H. Buc, A. Garges, and S. Adhya, 'Transcriptional regulation by cAMP and its receptor protein', *Annual Review of Biochemistry* 62 (1993), 749–795. https://doi.org/10.1146/annurev.bi.62.070193.003533

trajectories leading to the evolution of such a complex set of coordinated activities remain to be explored.

In summary, there exists an intriguing hypothesis that proteins involved in fundamental processes of chromosome shaping and compaction might have evolved the ability to bind to specific sites on the DNA and regulate the transcription of nearby genes. The repertoire of TFs is in flux and presumably responds to selection imposed by the environment and is probably driven, in large part, by genome reduction and horizontal gene acquisition. Finally, evolution often allows mutations that alter TF activity early in adaptation; the pleiotropic nature of TF mutations, which can potentially cause much collateral damage, might eventually select against sequence variation in conserved TFs.

5.6. Building to read and reading to build

Evolving a bacterial genome is complicated business. Its construction reflects the fine balance of a host of selection pressures within a relatively short stretch of DNA. This is unlike the genomes of higher eukaryotes, in which relaxed selection allows the accommodation of whole stretches of non-functional or 'junk' DNA (see Chapter 3). In bacteria, selection appears to decide not only the gene repertoire but also where each gene is positioned on the chromosome. We will call this 'gene order', 'gene organisation', or 'chromosome organisation' interchangeably. Here we will discuss the interplay between transcription and gene organisation; how gene organisation helps enable efficient transcription, and how transcription drives gene organisation. Note here that transcription is by no means the only driver behind chromosome organisation, but in the context of this book and this chapter it is merely the most relevant factor.

But first, a brief reiteration and description of some features of a bacterial genome. Most known bacteria contain a single circular chromosome. There are several exceptions to both 'single' and 'circular'. Bacteria such as *Vibrio cholerae* carry more than one chromosome and members of *Streptomyces* have linear chromosomes. Our discussion will apply to the single, circular bacterial genome and some conclusions might apply to any chromosome identifiable as primary in bacteria with multiple chromosomes.

The bacterial genome encodes genes for a variety of functions, starting from those that are required minimally for any cell to function to those needed for defence against the most unusual threats. Many of these functions emerge or become better represented in the genetic repertoire in larger bacterial genomes than in smaller ones, whereas others show no such relationship.[120] For example, the number of genes coding for proteins that are part of the ribosome would be more or less constant irrespective of genome size, for the bacterial ribosome is a highly conserved structure that is grossly

120 E. van Nimwegen, 'Scaling laws in the functional content of genomes', *Trends in Genetics*. 19 (2009), 479–484. https://doi.org/10.1007/0-387-33916-7_14

the same for most, if not all, bacteria. On the other hand, the number of genes involved in small molecule metabolism—the breakdown of nutrients, biosynthesis of monomeric building blocks among other molecules, and energy generation—increases linearly with genome size. This suggests that an increase in genome size in bacteria reflects an expansion of metabolic capabilities, which in turn is in response to an increase in the complexity and diversity of its habitats. Curiously, as we had briefly discussed in the previous section, the number of TF-encoding genes (and other regulatory protein genes) increases more or less quadratically with genome size. This brings to the table the idea that as genome size and gene function repertoire increase, the regulatory 'overhead' for ensuring their optimal function increases more than linearly—to the extent that Juan Ranea and colleagues argued that regulatory cost can impose a ceiling on how large a bacterial genome can grow.[121]

Genes representing this vast diversity of functions are arranged fairly tightly along the bacterial chromosome, with very little intergenic DNA separating adjacent genes. Adjacent genes can be transcribed in the same direction, being encoded on the same strand of DNA. They can be convergent, with the end of one gene being next to that of the other. Or they can be divergent, with the starts of the two genes adjacent to each other. Both convergent and divergent pairs of genes are encoded on opposite strands of DNA. As mentioned earlier, groups of co-directional, adjacent genes are often organised as operons. Genes forming part of the same operon are transcribed together from a single promoter as a single mRNA, with each protein-coding gene within the operon translated from its own ribosome loading site.

When genes are close to each other, the transcription of one gene can affect that of neighbouring genes. This arises from the interplay between DNA supercoiling and transcription. When RNAP is transcribing a gene, the mechanics of the process is such that the DNA in front is overwound, or positively supercoiled, while the DNA behind the RNAP is underwound, or hyper-negatively supercoiled (Fig. 5.13).[122] The progress of the RNAP, and therefore transcription, would require topoisomerases to act and stabilise supercoiling states. This imposes additional constraints that can impede idealised, smooth progress of transcription. A recent study by Ihab Boulas et al. showed that, in artificial DNA constructs introduced into a bacterial cell, the expression of a downstream gene decreases when that of another gene upstream increases, unless an 'insulator'—an element that blocks the diffusion of supercoiling states along the length of a DNA molecule—is introduced between the two genes.[123]

Patrick Sobetzko, in an earlier study, had performed an analysis of the *E. coli* genome and asked how genes whose expression is sensitive to supercoiling states are organised

121 J.A. Ranea, A. Grant, J.M. Thornton, and C.A. Orengo, 'Microeconomic principles explain an optimal genome size in bacteria', *Trends in Genetics* 21 (2005), 21–25. https://doi.org/10.1016/j.tig.2004.11.014

122 C.J. Dorman, 'DNA supercoiling and transcription in bacteria: a two-way street', *BMC Molecular Cell Biology* 20 (2019), 26. https://doi.org/10.1186/s12860-019-0211-6

123 I. Boulas, L. Bruno, S. Rimsky, O. Espeli, I. Junier, and O. Rivoire, 'Assessing in vivo the impact of gene context on transcription through DNA supercoiling', *Nucleic Acids Research* 51 (2023), 9509–9521. https://doi.org/10.1093/nar/gkad688

on the chromosome in terms of their adjacency properties.[124] He found that genes that respond to hyper-negative supercoiling are enriched among divergently oriented gene pairs. When two genes are divergently encoded, the transcription of one would leave the other highly negatively supercoiled. In contrast, genes whose expression is favoured by less negatively supercoiled DNA are encoded in a convergent manner. RNAP activity at one gene would cause the other gene to be more tightly wound and would thus favour its expression. This work suggested that local DNA organisation is such that it enables transcription in the face of constraints imposed by transcription-induced DNA supercoiling.

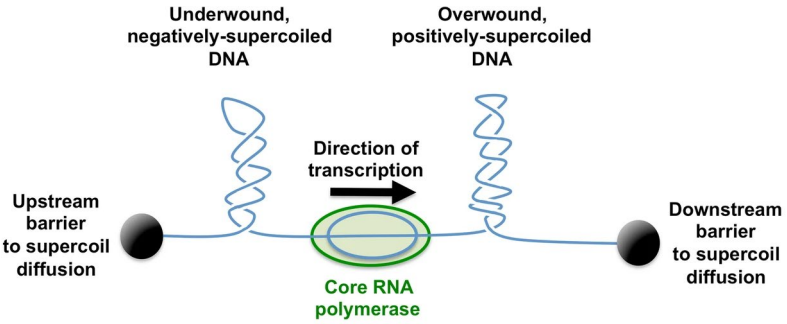


Fig. 5.13 Model of transcription and supercoiling. This figure shows the process of transcription, highlighting positive supercoils forming in front of and negative supercoils behind of the elongating RNAP. Originally published as Figure 1 in C.J. Dorman, 'DNA supercoiling and transcription in bacteria: a two-way street', *BMC Molecular Cell Biology* 20 (2019), 26, CC BY 4.0.

Transcription and gene organisation appear intertwined even if we are to zoom out and take a bird's-eye view of large chunks of the bacterial chromosome. To investigate this, we must first understand the effect of DNA replication on gene dosage, i.e., the number of copies of each gene present in the cell as a result of chromosome replication.[125] The typical bacterial chromosome has a single origin of replication (*ori*). This is the site at which DNA polymerase (DNAP), the enzyme that replicates the chromosome, binds. Replication proceeds bidirectionally outwards from *ori* and ends at a series of terminus sites (*ter*) located more or less at a diametrically opposite location on the circular chromosome. Consider the chromosome as a perfect circle and draw its diameter from *ori* to *ter*. We will call this line the *ori-ter axis* and the two semicircles thus formed as *replichores*. The two replichores, on either side of the *ori-ter* axis, would be nearly equal in length. The DNA polymerase in *E. coli* replicates the DNA at the rate of ~1,000 bp per second. Assuming that the average *E. coli* chromosome is ~5 Mbp long, and that the two replichores are being replicated simultaneously, it will take over 40 minutes for the chromosome to be replicated completely. If, in a minimal growth medium, *E.*

124 P. Sobetzko, 'Transcription-coupled DNA supercoiling dictates the chromosomal arrangement of bacterial genes', *Nucleic Acids Research* 44 (2016), 1514–1524. https://doi.org/10.1093/nar/gkw007

125 E.P.C. Rocha, 'The replication-related organization of bacterial genomes', *Microbiology* 150 (2004), 1609–1627. https://doi.org/10.1099/mic.0.26974-0

coli populations double, say, every hour, then we can expect *ori*-proximal genes to be present in two copies for at least two-thirds of a bacterium's life cycle. On the other hand, *ter*-proximal genes would be replicated, producing a second copy, only a short while before the cell divides. Imagine the situation of *E. coli* growing in rich media that supports the population's doubling every 20–30 minutes on average, much less than the time required to make two copies of the chromosome to be partitioned between the two daughter cells. In such situations, replication initiates at *ori* more than once per life cycle such that the DNA copy number or dosage between *ori* and *ter*-proximal genes can be much higher than two—say, four, or possibly even eight.

The more copies of a gene, the higher the availability of promoters for its transcription. This is especially true when a gene promoter in one copy of the chromosome is saturated with RNAP. In such a situation, creating a second copy would pretty much be the only way to increase transcription even if the cell's physiology absolutely requires it. Thus, this replication-dependent difference in gene dosage between *ori*- and *ter*-proximal genes can make additional promoters of *ori*-proximal genes available for RNAP to access and bind to. To what extent does this aspect of the interplay between replication and transcription affect gene organisation?

Patrick Sobetzko and colleagues, by analysing gene organisation in the *E. coli* genome, showed that genes under σD control are relatively more frequently encoded proximally to *ori* whereas those regulated by σS are located closer to *ter*[126]. This applies equally well to both replichores. σD regulated genes are expressed primarily during exponential growth during which chromosome replication prominently occurs. One can hypothesise that some part of the increased expression of σD-regulated genes during exponential growth is facilitated by their higher dosage, which in turn is a consequence of their presence in *ori*-proximal regions of the chromosome and ongoing DNA replication. Rajalakshmi Srinivasan and co-workers in my lab found further evidence that gene expression coherence extends well beyond the confines of the operon.[127] A gene encoded in one *half* of the chromosome, centred around *ori* or *ter* and thus comprising one half of each replichore, is more likely to be expressed under similar conditions as another present in the same half than with one found in the opposite half. Further, if a gene in one half of the chromosome is activated in one condition, then a gene from the other half tends to be repressed in the same condition. Thus, genes that are expressed together tend to be found in the same half of the chromosome whereas mutually exclusive or antagonistic pairs of genes are encoded in opposite halves. This could well be explained at least in part by the differential localisation of σD- and σS-regulated genes in *ori*-proximal and *ter*-proximal parts of the chromosome respectively.

126 P. Sobetzko, A. Travers, and G. Muskhelishvili, 'Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle', *Proceedings of the National Academy of Sciences USA* 109 (2011), E42–E50. https://doi.org/10.1073/pnas.1108229109

127 Srinivasan et al., 2015.

In the results published by Sobetzko et al., the relative preference of σD genes to be encoded in *ori-proximal*- over *ter*-proximal regions of the chromosome was relatively small. This suggests that only a small, though significant, subset of σD genes are preferentially encoded close to *ori*. Is this a random subset or does it represent particular sets of gene functions? The answer to this question has already been provided by Etienne Couturier and Eduardo Rocha, who asked where genes involved in different functions and expressed at different levels are encoded on a couple of hundred different bacterial genomes, and how these patterns might change across bacteria capable of growing at different rates.[128] We know that the growth rate of a bacterial population is determined to a large extent by the nutrition available to it. Nutrient availability is dynamic, but gene organisation is not nearly as dynamic. So, what is the rationale behind trying to correlate the two? It may well be the case that richer media conditions support faster growth, but this relationship cannot hold indefinitely. The genomic content of a bacterium would impose a ceiling on how fast its population can grow. After a point, one can keep adding better and better nutrients, but growth rates would saturate. This ceiling is a product of evolution and probably a reflection of the bacterium's ecology. Couturier and Rocha used an experimentally determined dataset of the highest known growth rates for ~200 bacteria and made the reasonable assumption that these correspond to the maximum growth rate possible for these bacteria. Given that this is a product of evolutionary optimisation, they asked whether gene organisation is in any way linked to maximum growth rate as contained in data they had assembled. As we had discussed earlier, a higher growth rate can result in higher copy number differences between *ori* and *ter* and presumably stronger selective pressures arising from such a difference. In fact, instead of growth rate, Couturier and Rocha sought to find the relationship between gene organisation and a measure of the gene dosage difference between *ori* and *ter*, which can be estimated from growth rates and the rate of progress of replication.

Couturier and Rocha first predicted highly expressed protein-coding genes from their sequence.[129] A prediction, as opposed to an experimentally-determined set of highly expressed genes, was necessitated by the fact that appropriate experimental data were available for only a small set of bacteria and not across the broad phylogenetic spread these researchers studied. The prediction was based on the degeneracy of the genetic code. Most amino acids are encoded by multiple codons, and in most organisms one or a smaller subset of codons for each amino acid is preferentially used. This may reflect the relative availability of the different types of tRNAs, each of which recognises a particular codon and brings its respective amino acid to the translating ribosome. The presence of a rare codon results in translation slowing down, because the appropriate tRNA is not immediately available. Thus, highly expressed protein-

128 E. Couturier and E.P.C. Rocha, 'Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes', *Molecular Microbiology* 59 (2006), 1506–1518. https://doi.org/10.1111/j.1365-2958.2006.05046.x

129 Ibid.

coding genes can be expected to often use a preferred codon for most amino acids. This tendency of any gene to utilise preferred codons can be measured by calculating what is known as the '*codon adaptation index*' (CAI) from its sequence. We expect genes with high CAI to be expressed at high levels.

Couturier and Rocha[130] found that genes with high CAI are preferentially encoded closer to *ori*, but only in fast-growing bacteria. In particular, highly expressed genes responsible for fast growth, including those encoding RNAP, rRNA, and ribosomal proteins, are encoded in *ori*-proximal regions in fast-growing bacteria. rRNA genes are of particular relevance here. In fast-growing bacteria, some 80–90% of all transcription is diverted to the synthesis of rRNA. In fact, growth rate and rRNA expression levels are tightly correlated. For the same bacteria, increasing growth rate—for example by the provision of better nutrition—is associated with an increase in rRNA expression. Across bacterial species, the number of rDNA copies per chromosome increases with increasing maximum growth rates. Overall, high translation supplies fast growth. This requires both rRNA and ribosomal proteins. The supply of the latter is provided by a product of transcription and translation. The former, however, lack the luxury of amplification provided by translation and should be entirely supplied by transcription. The high levels of rRNA transcription required for fast growth cannot be provided by a single gene copy. Thus, fast growing bacteria code for multiple copies of rRNA genes per chromosome. Their being encoded near *ori* will further increase their gene dosage, making more of their promoters available for transcription, during rapid growth. The limiting factor then is the supply of RNAP. The encoding of RNAP genes close to the *ori*, closer than rDNA, should help the cell beat this constraint. Thus, genes required at high levels for fast growth are encoded in *ori*-proximal regions in bacteria capable of rapid population growth.

In a much more recent work, Supriya Khedkar in my lab reiterated the findings of Couturier and Rocha showing that essential genes encoding proteins involved in translation are present in *ori*-proximal regions in fast-growing bacteria (Fig. 5.14A).[131] She also showed that horizontally-acquired genes are depleted from regions close to *ori* in both fast- and slow-growing bacteria, but more so in the former (Fig. 5.14B). There are two possible explanations for this. The first is that regions around the *ori* are rich anyway in essential genes involved in crucial information processes such as transcription and translation. In gene-rich bacterial genomes, a random insertion of a horizontally-acquired gene will more often than not split and disrupt a gene already present in the chromosome. The successful maintenance or loss of such a disruptive insertion will depend on the relative contributions of the inserted and the disrupted gene to growth and survival. When this occurs in *ori*-proximal regions, the chance that it will disrupt an essential gene and cause lethality is relatively high.

130 Ibid.

131 S. Khedkar and A.S.N. Seshasayee, 'Comparative genomics of interreplichore translocations in bacteria: a measure of chromosome topology?', *Genes, Genomes, Genetics* 6 (2016), 1597–1606. https://doi.org/10.1534/g3.116.028274

Such insertions will be purged out by selection, leading to the under-representation of horizontally-acquired genes in regions near *ori* in many extant bacterial genomes. A second explanation, not mutually exclusive with the first, is that some aspect of chromosome structure, such as the protection of bound DNA by nucleoid-associated proteins, disallows insertions in regions near *ori* in the first place. The evidence for this is complex.

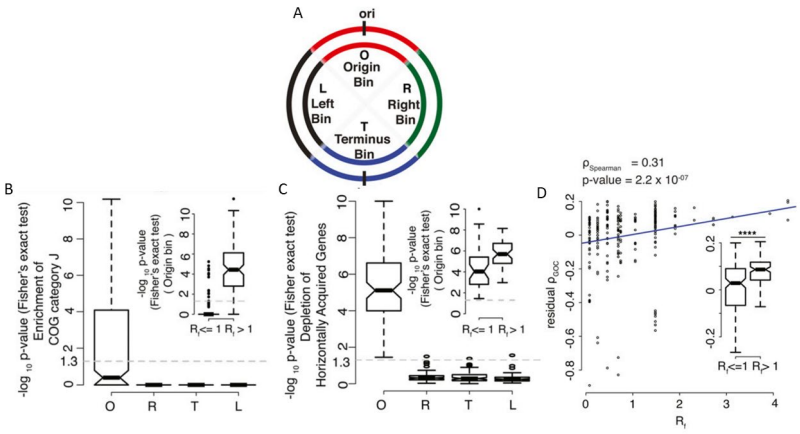


Fig. 5.14 Gene organisation in fast- and slow-growing bacteria. (A) This figure shows the division of the bacterial chromosome into four bins referred to in B and C. (B) This figure shows the enrichment of translation related genes (referred to as 'COG category J') in the *ori*-proximal region across bacterial genomes; inset shows the difference between fast-growing ($R > 1$) and slow-growing ($R <= 1$) bacteria. (C) As in B, but this figure shows the depletion of horizontally-acquired genes in *ori*-proximal regions in both fast- and slow-growing organisms but more so in the former. (D) This figure shows that gene order conservation (ρ_{GOC}) between pairs of closely-related bacteria after correcting for phylogenetic relatedness is correlated with the growth rate of bacteria. Originally published as part of S. Khedkar and A.S.N. Seshasayee, 'Comparative genomics of interreplichore translocations in bacteria: a measure of chromosome topology?', *Genes, Genomes, Genetics* 6 (2016), 1597–1606, CC BY 4.0.

In a recent study, Malikmohammed Yousuf and colleagues found that insertions of transposons may occur more frequently near *ori* than *ter*.[132] Thus, it is reasonable to posit that the lack of horizontally-acquired genes near *ori* is more likely a result of selection than any insertion bias. Yousuf et al. also noticed that insertion sites are enriched in locations bound by H-NS. Though H-NS binding sites are more often in the *ter*-half of the chromosome, this relationship between H-NS binding sites and insertion was not strong enough to overcome the overall balance favouring insertions in the *ori*-half. Because Yousuf et al. did not look for insertions in *E. coli* lacking H-NS, it is unclear whether the protein influences where insertions happen. Transposon insertions are not entirely random, and many prefer inserting in A+T-rich loci. This can also explain

132 M. Yousuf, I. Iuliani, R.T. Veetil, A.S.N. Seshasayee, B. Sclavi, and M.C. Lagomarsino, 'Early fate of exogenous promoters in *E. coli*', *Nucleic Acids Research* 48 (2020), 2348–2356. https://doi.org/10.1093/nar/gkz1196

why insertions are more common in H-NS binding sites, which themselves are A+T-rich. It is probably more likely that H-NS inhibits integration. N. Sharadamma and colleagues had shown—using purified protein from *Mycobacterium tuberculosis* and chemically synthesised DNA—that H-NS inhibits biochemical processes underlying integration.[133] Further, a genome-wide study of transposon insertion in *Vibrio cholerae* by Satoshi Kimura and co-workers showed that H-NS binding sites were depleted for such insertions, and this was no longer observed in bacteria lacking H-NS.[134] Thus, the balance of evidence is in favour of H-NS, which helps to keep horizontally-acquired DNA transcriptionally silent, operating one step earlier by discouraging the insertion of foreign DNA. As a result, mechanistic processes can create non-uniformities in the insertion of foreign DNA, but the evidence presented here suggests that these would not block insertions near *ori*, thus further strengthening the case of selection acting to minimise insertions in *ori*-proximal DNA.

Khedkar also showed that gene order is usually more stable in fast-growing bacteria,[135] consistent with similar findings made earlier by Couturier and Rocha using a different analytical approach (Fig. 5.14C).[136] In other words, the replication-dependent dosage of a gene is better conserved in fast-growing than in slow-growing bacteria. Khedkar was, in particular, interested in measuring long-range translocations in bacterial genomes—i.e., is a gene located at a position *p* in one bacterial genome positioned elsewhere at *q*, distant from *p*, in another, related genome? Genes encoded in *ori*-proximal regions rarely translocate to distant *ter*-proximal parts of the genome. However, genes encoded on one replichore in one bacterium have in several instances moved to the opposite replichore in a related bacterium. These translocations from one replichore to the other do not usually disrupt gene dosage; in other words, the distance of a translocated gene from *ori* remains more or less the same, within reasonable limits, in whichever replichore it is found in (Fig. 5.15A). Or, inter-replichore translocations are often symmetric about the *ori-ter* axis, thus conserving distance from the *ori*. This helped to generalise findings made years earlier based on very few genomes.[137] Assuming translocations can occur randomly (symmetrically or asymmetrically at more or less equal frequencies), then selection imposed by the gene dosage gradient can eliminate a good proportion of asymmetric inter-replichore translocations.

An alternative explanation for the predominance of symmetric translocations is that asymmetric translocations that disrupt gene dosage do not happen at all, or happen at

133 N. Sharadamma, Y. Harshavardhana, P. Singh, and K. Muniyappa, 'Mycobacterium tuberculosis nucleoid-associated DNA-binding protein H-NS binds with high-affinity to the Holliday junction and inhibits strand exchange promoted by RecA protein', *Nucleic Acids Research* 38 (2010), 3555–3569. https://doi.org/10.1093/nar/gkq064

134 S. Kimura, T.B. Hubbard, B.M. Davis, and M.K. Waldor, 'The Nucleoid Binding Protein H-NS Biases Genome-Wide Transposon Insertion Landscapes', *mBio* 7 (2016), e01351–16. https://doi.org/10.1128/mbio.01351-16

135 Khedkar and Seshasayee, 2016.

136 Couturier and Rocha, 2006.

137 M. Suyama and P. Bork, 'Evolution of prokaryotic gene order: genome rearrangements in closely related species', *Trends in Genetics* 17 (2001), 10–13. https://doi.org/10.1016/s0168-9525(00)02159-4

very low frequencies. Multiple factors may contribute to the symmetry of translocations. One is that these events often require single-stranded DNA. This happens in replication forks—places where the DNAP is replicating the chromosome. The two replication forks, one on each replichore, move at more or less the same speed. If such forks promote translocations, then the manner in which replication occurs can ensure that translocations are usually symmetric around the *ori-ter* axis. Further, in some bacteria, the chromosome is structured in 3D space such that the two replichores are intertwined about each other, more or less symmetrically about the *ori-ter* axis.[138] Translocations would require the two regions of the chromosome to lie in close proximity. Khedkar showed that in one such bacterium, *C. crescentus*, translocation events appear to have occurred more frequently between pairs of regions that are often in contact with one another. But any such effect is likely to be amplified by selection that ensures that disadvantageous translocations are lost. Therefore, long-range translocations of bacterial genes are not uncommon, but these, along with other mechanisms of gene rearrangements, minimally disrupt gene dosage. This, in part, reflects selection acting to eliminate rearrangements that detrimentally disrupt gene dosage.

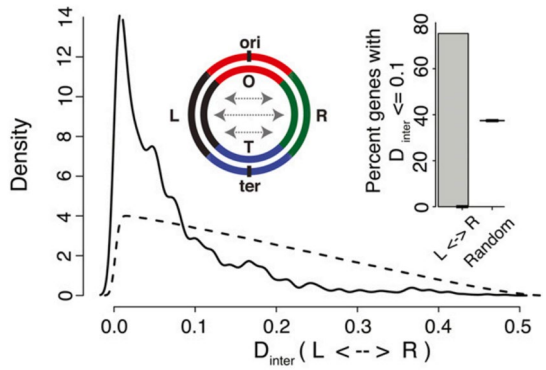


Fig. 5.15 Chromosome rearrangements maintaining gene dosage along the *ori-ter* axis. (A) This figure shows that inter-replichore translocations tend to be symmetric about the *ori-ter* axis. D_{inter} refers to the distance of the translocated pair of genes on either replichore from *ori*. The smaller the distance, the more symmetric the translocation. The dotted line shows what would be expected if translocations occurred between random sites across replichores. Originally published as Figure 5A in S. Khedkar and A.S.N. Seshasayee, 'Comparative genomics of interreplichore translocations in bacteria: a measure of chromosome topology?', *Genes, Genomes, Genetics* 6 (2016), 1597–1606, CC BY 4.0.

In another work published shortly after Khedkar's study, Jelena Repar and Tobias Warnecke showed that the tendency for inter-replichore translocations to be symmetric was not uniform across bacterial clades.[139] In different clades, different

138 T.B. Le, M.V. Imakaev, L.A. Mirny, and M.T. Laub, 'High- resolution mapping of the spatial organization of a bacterial chromosome', *Science* 342 (2013), 731–734. https://doi.org/10.1126/science.1242059

139 J. Repar and T. Warnecke, 'Non-random inversion landscapes in prokaryotic genomes are shaped by heterogeneous selection pressures', *Molecular Biology and Evolution* 8 (2017), 1902–1911. https://doi.org/10.1093/molbev/msx127

selection pressures seemed to be associated with symmetric translocations. In some, it was the presence of translation-associated genes close to *ori*, which can be interpreted as a measure of growth rate. In other clades, this relationship did not hold. Instead, correlations with what is known as '*gene strand bias*' (*GSB*) were observed. What is GSB and how is this related to replication?

Replication and transcription are two essential processes that engage the bacterial chromosome simultaneously. While replication is ongoing, RNAP is also going about doing its job transcribing genes. It therefore becomes inevitable that the two polymerases with collide, either codirectionally or in a head-on fashion. The addition of a new nucleotide to a growing DNA chain during replication is directional; the phosphate group of an incoming nucleotide is attached to a hydroxyl group at the end of the growing, nascent DNA chain. One strand, called the *leading strand*, is that which is synthesised continuously, in the direction in which new nucleotides are added to the growing DNA chain. The opposite strand, or the lagging strand, is synthesised discontinuously as fragments that are later glued together. Transcription, because it produces single stranded RNA, is free from such concerns despite being just as directional as replication. A gene that is encoded on the leading strand is transcribed in the same direction as replication. Therefore, DNAP and RNAP move codirectionally over a leading strand gene. Note that when we say that a gene is encoded on the leading strand, we mean that this strand acts as the coding strand whose sequence is identical[140] to that of the RNA chain being synthesised; for such a gene, the *lagging strand* serves as the template for transcription. Any meeting between the DNAP and RNAP transcribing a lagging strand gene will be head-on.

We now understand that codirectional collisions between RNAP and DNAP are largely inconsequential, whereas head-on collisions lead to several problems, from the mere slowing down of transcription[141] to genotoxicity arising from the stalling of replication.[142] In particular, head-on conflicts between the two polymerases increase local mutation rates around the collision site.[143]

One can expect that highly expressed genes would usually be encoded on the leading strand to minimise the chance of detrimental head-on collisions between RNAP and DNAP. A gene that is highly transcribed is more likely to see an RNAP meet a DNAP than one that is less transcribed, with replication being equal for all genes. As pointed out earlier, over 80% of all transcription in growing cells is reserved for rRNA synthesis. Thus, even the most highly expressed mRNA would not be as highly transcribed as rRNA genes, and so rRNA genes—across bacteria—are almost always

140 But for uracil replacing thymine.

141 B. Liu and B.M. Alberts, 'Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex', *Science* 267 (1995), 1131–1137. https://doi.org/10.1126/science.7855590

142 E.V. Mirkin and S.M. Mirkin, 'Mechanisms of Transcription-Replication Collisions in Bacteria', *Molecular and Cellular Biology* 25 (2005), 888–895. https://doi.org/10.1128/mcb.25.3.888-895.2005

143 S. Paul, S. Million-Weaver, S. Chattopadhyay, E. Sokurenko, and H. Merrikh, 'Accelerated gene evolution through replication–transcription conflicts', *Nature* 495 (2013), 512–515. https://doi.org/10.1038/nature11989

encoded on the leading strand. More broadly, in most bacteria, a majority of genes are encoded on the leading strand. Now, are highly expressed genes preferentially encoded on the leading strand? For example, if 55% of all genes in a bacterial genome are present on the leading strand, are, say, 75% or 80% of highly expressed genes so encoded? Eduardo Rocha and Antoine Danchin showed that it is the essentiality of a gene, and not its expression level as measured by its CAI, that determines which strand a gene is encoded on.[144] Whereas a large proportion of essential genes are encoded on the leading strand irrespective of whether they are highly or less expressed in *Bacillus subtilis* and *E. coli*, the same proportion for non-essential genes drops ~20% again independently of expression level.

In a more recent study, Christopher Merrikh and Houra Merrikh showed that, in contrast to essential genes, genes involved in processes like antibiotic resistance and virulence are encoded on the lagging strand and show elevated mutation rates.[145] It is well established that there are more guanines than cytosines on the leading strand, and by definition the reverse is true for the lagging strand. This bias can be measured by what is called as '*GC skew*' ((G-C)/(G+C)). This value is positive for the leading strand. The skew arises from differences in mutational pressures between the two strands. It is a reflection of the long-term evolutionary history of which strand the piece of DNA sequence—for which the skew is measured—has been encoded on. However, not every stretch of leading strand DNA sequence exhibits a positive skew, and not every segment of lagging strand DNA exhibits a negative skew. Local variations in GC skew along a strand can tell us whether a segment of DNA in an extant genome under study has stayed on the same strand over long periods or has switched strands in more recent times. Using this reasoning, Merrikh and Merrikh showed that several highly mutable genes have inverted or switched from the leading to the lagging strand recently. This led them to propose that these genes—often involved in antibiotic resistance and virulence—might undergo accelerated evolution by deploying head-on DNAP-RNAP collisions to cause mutations.

Though certain genes may undergo high rates of evolution through head-on polymerase collisions, it is known that a majority of genes in most bacteria are encoded on the leading strand. For example, most of the inter-replichore translocations present in Khedkar's analysis are inversions. In an inversion, a gene that is on one strand of the DNA switches to the other. Now, because of the bidirectional nature of replication, the leading strand on one replichore becomes lagging in the other. Thus, inversions *within* a replichore would flip a leading strand gene to the lagging strand and vice-versa. On the other hand, inversions across replichores preserve gene strandedness. Thus, inter-replichore translocations analysed by Khedkar not only kept gene dosage disruptions to a minimum, but also maintained GSB. In a very recent study, Malhar

144 E.P.C. Rocha and A. Danchin, 'Essentiality, not expressiveness, drives gene-strand bias in bacteria', *Nature Genetics* 34 (2003), 377–378. https://doi.org/10.1038/ng1209

145 C.N. Merrikh and H. Merrikh, 'Gene inversion potentiates bacterial evolvability and virulence', *Nature Communications* 9 (2018), 4662. https://doi.org/10.1038/s41467-018-07110-3

Atre and colleagues performed a detailed analysis of inversions and GSB in over 2,000 bacterial genomes.[146] Consistent with the idea that intra-replichore inversions disrupt GSB whereas inter-replichore inversions do not, the latter are more common in most bacterial genomes.

Now, GSB is not uniform across bacterial clades. Some genomes code for only a slight excess of leading strand genes whereas in others, as many as 80–90% of all genes are on the leading strand. This indicates that different bacterial clades have different mechanisms to resolve head-on DNAP-RNAP conflicts and thus manage them differently. For example, if the mechanism of replication itself causes high rates of detrimental head-on DNAP-RNAP conflicts, it stands to reason that selection would keep a large proportion of genes in such bacteria to be encoded on the leading strand. There is a correlation between GSB and the nature of DNAP utilised by a bacterium for replication.[147] Again, an evolutionary argument for the variation in GSB is that different bacteria have selective pressures, independent of any differences in mechanisms of replication, against the detrimental effects of such collisions. An example would be growth rate: Anjana Srivatsan and colleagues demonstrated in *B. subtilis*, a genome with high GSB, that a large inversion near the *ori*—which causes many rRNA genes to shift from the leading to the lagging strand—has a stronger negative effect on growth, specifically during fast growth.[148] Atre and co-workers discovered that genomes with high GSB display very low frequencies of inversions overall. Further, in high GSB genomes, whatever inversions there are tend to be of the non-disruptive inter-replichore type. The authors argue that differences in inversion frequencies and type may be a factor underlying variation in GSB. Alternatively, if differences in the DNAP cause genomes with high GSB to be less tolerant of head-on DNAP-RNAP conflicts, then disruptive inversions would be much more strongly selected in such genomes. Thus, there is a strong mechanistic basis for the variation in GSB, and any evolutionary factor may arise from these mechanistic differences. Growth rate may also be a player, operating at a level distinct from the mechanistic factor, but its strength across bacteria needs to be clarified.

Chromosome rearrangements such as duplications, deletions, and inversions are often facilitated by repetitive sequences or just repeats. Bacterial genomes, unlike eukaryotic genomes, are relatively poor in repeats but are not entirely devoid of them. Pairs of repeats are called *direct repeats* when they are encoded in the same orientation. Rearrangements mediated by interactions between direct repeats are duplications and deletions. On the other hand, repeat pairs that are coded for on opposite strands are

146 M. Atre, B. Joshi, J. Babu, S. Sawant, S. Sharma, and T.S. Sankar, 'Origin, evolution and maintenance of gene-strand bias in bacteria', *Nucleic Acids Research* 52 (2024), 3493–3509. https://doi.org/10.1093/nar/gkae155

147 Ibid.; E.P.C. Rocha, 'Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes?', *Trends in Microbiology* 10 (2002), 393–395. https://doi.org/10.1016/s0966-842x(02)02420-4

148 A. Srivatsan, A. Tehranchi, D.M. MacAlpine, and J.D. Wang, 'Co-orientation of replication and transcription preserves genome integrity', *PLoS Genetics* 6 (2010), e1000810. https://doi.org/10.1371/journal.pgen.1000810

called *inverted repeats*, and these promote inversions. Repetitive sequences are usually horizontally acquired, and in many instances are transposable elements that can jump or copy themselves around the chromosome. One can expect these sequences to be randomly distributed, subject to constraints arising from the mechanisms by which they are generated. Nitish Malhotra in my lab showed that repeats are non-randomly distributed across bacterial genomes, especially in fast-growing bacteria.[149] He found that inverted repeats are more commonly present inter-replichore than intra-replichore, in fast-growing bacteria in particular (Fig. 5.16A). Thus, inversions mediated by a majority of inverted repeat pairs would be inter-replichore and would not affect GSB. Further, inter-replichore inverted repeat pairs tend to be positioned more or less symmetrically about the *ori-ter* axis, implying that inversions that they promote would disrupt gene dosage of inverted genes to a reduced extent. This was again prominent in fast-growing bacteria (Fig. 5.16B). Direct repeat pairs are often intra-replichore and are positioned closer to each other than would be predicted by random chance. This ensures that deletions and duplications caused by such repeats affect only relatively short stretches of DNA. Direct repeats are usually generated by duplications that create tandem copies of the duplicated sequence.[150] Thus, one can expect the mere processes that generate repeats to keep direct repeat pairs closer to each other than expected by chance. However, Malhotra also noted that the distance between direct repeat pairs was significantly shorter in fast-growing than in slow-growing bacteria, thus invoking an argument in favour of selection against large deletions and duplications that distant direct repeat pairs might cause. Therefore, the organisation of repetitive elements—among the drivers of chromosome rearrangements—on the chromosome is non-random and probably set up such that the rearrangements which can promoted do not drastically alter favourable gene organisation.

Finally, given that the location of *ori* on the genome is a central player in gene organisation, what will be the consequences for bacterial growth and gene organisation of the *ori* shifting elsewhere on the same chromosome? Xindan Wang and colleagues engineered *E. coli* to carry an origin of replication[151] at a non-native site, and removed the native *ori*. They called the new origin of replication *oriZ*. This newly engineered *oriZ* was placed about 1 Mbp away from the native *ori*. They found that *oriZ* was fully functional, causing initiation of chromosome replication normally, in this engineered *E. coli*. They also noticed that the initiation of replication from *oriZ*, instead of from the native *ori*, had a minimal effect on time to cell doubling. This is a curious observation. The sequence between *oriZ* and where the native *ori* originally

149 N. Malhotra and A.S.N. Seshasayee, 'Replication-Dependent Organization Constrains Positioning of Long DNA Repeats in Bacterial Genomes', *Genome Biology and Evolution* 14 (2022), evac102. https://doi.org/10.1093/gbe/evac102

150 G. Achaz, E.P.C. Rocha, P. Netter, and E. Coissac, 'Origin and fate of repeats in bacteria', *Nucleic Acids Research* 30 (2002), 2987–2994. https://doi.org/10.1093/nar/gkf391

151 X. Wang, C. Lesterlin, R. Reyes-Lamothe, G. Ball, and D.G. Sherratt, 'Replication and segregation of an *Escherichia coli* chromosome with two replication origins', *Proceedings of the National Academy of Sciences USA* 108 (2011), E243–E250. https://doi.org/10.1073/pnas.1100874108

was includes several highly expressed rRNA genes. These rRNA genes, which would have been on the leading strand in relation to replication initiated from the native *ori*, are now on the lagging strand. Thus, DNAP initiating replication from the newly-introduced *oriZ* would engage in head-on conflicts with RNAP transcribing rRNA genes. So how come bacteria facing such detrimental conflicts are replicating and growing just fine?

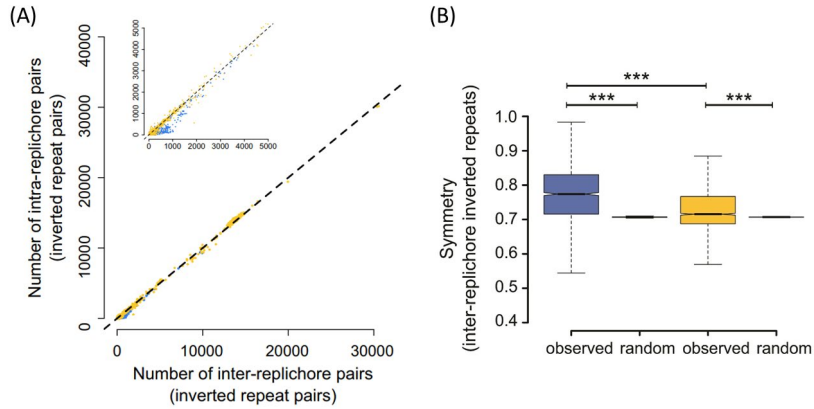


Fig. 5.16 Non-random organisation of repeats in bacterial genomes. (A) This figure shows that fast growing genomes contain more inter-replichore inverted repeats. This is especially clear in the zoomed-in version shown in the inset. Blue dots are for fast-growing bacteria and yellow for slower growing ones. Originally published as Figure 6E in N. Malhotra and A.S.N. Seshasayee, 'Replication-Dependent Organization Constrains Positioning of Long DNA Repeats in Bacterial Genomes', *Genome Biology and Evolution* 14 (2022), evac102, CC BY-NC 4.0; copyright held by the author of this book. (B) Inter-replichore inverted repeat pairs are more symmetric about the *ori-ter* axis in fast-growing bacteria. Originally published as Figure 7H in Malhotra and Seshasayee (2022).

Ivanova and co-workers answered this question.[152] They created an *E. coli* strain similar to that engineered by Wang and colleagues, but—unlike Wang and co-workers—found that this experienced a severe growth defect when grown in rich media. They also observed that problems in replication arose near rRNA genes, suggesting that head-on DNAP-RNAP conflicts contribute to the growth defect these bacteria suffer from. They provided further evidence for this argument by showing that a mutation in RNAP that allows it to bypass head-on conflicts alleviated the growth defect. They also noticed that the *E. coli* strain generated by Wang et al. had acquired a chromosome rearrangement that masked the growth defect that replication initiating at *oriZ* would have otherwise caused. This solution is simple. A fairly long stretch of DNA containing several rRNA genes had simply inverted, thus returning them onto the leading strand given replication initiation from the new *ori*.

152 D. Ivanova, T. Taylor, S.L. Smith, J.U. Dimude, A.L. Upton, et al., 'Shaping the landscape of the *Escherichia coli* chromosome: replication-transcription encounters in cells with an ectopic replication origin', *Nucleic Acids Research* 43 (2015), 7865–7877. https://doi.org/10.1093/nar/gkv704

Reshma Veetil and others in my lab found a similar adaptation evolving in an *E. coli* mutant which was defective in initiating replication at the native *ori* but managed to do so from elsewhere on the chromosome.[153] They found that these *E. coli* initiated replication from a site called *oriX*, 0.4–0.7 Mbp away from the native *ori*, creating a situation similar to the strains used by Wang et al. and Ivanova and colleagues. Again, an inversion of a stretch of DNA including several rRNA genes was one adaptive strategy discovered by these *E. coli* under selection to multiply faster. Such inversions are easy to achieve. rRNA genes are themselves repetitive elements, and a pair of such repeats on either side of the native *ori* would form an inter-replichore inverted repeat pair, which would promote inversions of the intervening DNA. If such inversions were to happen in natural isolates of *E. coli* replicating from the native *ori*, they would maintain strandedness, being catalysed by a pair of inter-replichore inverted repeats. However, this inversion in the context of replication from *oriX* caused several essential genes to switch from the leading to the lagging strand. This is probably an acceptable compromise. As emphasised earlier, rRNA synthesis accounts for the bulk of transcription. Ensuring that the highly transcribed rRNA genes stay on the leading strand even at the cost of other essential mRNA genes switching to the lagging strand is a fair bargain. Given enough time, *E. coli* would probably discover additional mutations that help to manage head-on DNAP-RNAP collisions even at essential genes.

Taken together, the way in which a bacterial chromosome replicates establishes a difference in gene dosage between *ori*- and *ter*-proximal regions of the chromosome in a growth rate-dependent manner. The selection arising from this contributes to the evolution of gene organisation, especially in fast-growing organisms. Conflicts between DNAP and RNAP add a further layer of constraint on the genome, determining how many and which types of genes are encoded on which strand of DNA, and how this can vary across clades of bacteria.

Thus, transcription is a crucial first step in the reading of the genome. The regulation of transcription is intricate, involving a vast network of regulators and the regulated. This helps bacteria to adapt to external environments as well as to changing genetic circumstances, such as the introduction and integration of a potentially expensive horizontally-acquired gene. Physiological adaptation through the regulation of gene expression meets genetic evolution when regulators evolve, and this appears to be a common phenomenon—especially early during adaptation to a new environment. Adaptation by mutations of regulators will have to balance the advantage such mutations provide against the collateral damage that they can cause by altering the expression of genes unrelated to the present adaptive challenge. Transcription not only provides fast physiological adaptations, but—along with chromosome replication—is also a factor determining the manner in which genes are organised around the bacterial chromosome.

153 R.T. Veetil, N. Malhotra, A. Dubey, and A.S.N. Seshasayee, 'Laboratory Evolution Experiments Help Identify a Predominant Region of Constitutive Stable DNA Replication Initiation', *mSphere* 5 (2020), e00939–19. https://doi.org/10.1128/msphere.00939-19