A Genome-Wide Evolutionary Simulation of the Transcription-Supercoiling Coupling

Théotime Grohens[1] Sam Meyer[2]
Guillaume Beslon[1]

[1]INRIA Beagle Team, Artificial Evolution and Computational Biology,
Université de Lyon, Inria, ECL, INSA Lyon, Université Claude Bernard Lyon 1,
Université Lumière Lyon 2, CNRS, LIRIS UMR 5205, F-69622, France
[2]Université de Lyon, INSA Lyon, Université Claude Bernard Lyon 1, CNRS,
UMR5240 MAP, F-69622, France
theotime.grohens@insa-lyon.fr

March 23, 2022

Abstract

DNA supercoiling, the level of under- or overwinding of the DNA polymer around itself, is widely recognized as an ancestral regulation mechanism of gene expression in bacteria. Higher levels of negative supercoiling facilitate the opening of the DNA double helix at gene promoters, and thereby increase gene transcription rates. Different levels of supercoiling have been measured in bacteria exposed to different environments, leading to the hypothesis that variations in supercoiling could be a response to changes in the environment. Moreover, DNA transcription has been shown to generate local variations in the supercoiling level, and therefore to impact the transcription rate of neighboring genes.

In this work, we study the coupled dynamics of DNA supercoiling and transcription at the genome scale. We implement a genome-wide model of gene expression based on the transcription-supercoiling coupling. We show that, in this model, a simple change in global DNA supercoiling is sufficient to trigger differentiated responses in gene expression levels via the transcription-supercoiling coupling. Then, studying our model in the light of evolution, we demonstrate that this non-linear response to different environments, mediated by the transcription-supercoiling coupling, can serve as the basis for the evolution of specialized phenotypes.

Keywords: DNA supercoiling, gene transcription, genome, evolution.

1 Introduction

The DNA molecule is a double-stranded polymer of nucleotides that plays a fundamental role in life. It is shaped as a double helix which rotates around itself at a rate of around one turn per 10.5 base pairs (Krogh et al., 2018). However, when subject to physical forces, it can become overwound or underwound, or writhe around itself, in a process known as DNA supercoiling; the supercoiling level σ is measured as the density of extra turns (or coils) per base pair (Duprey and Groisman, 2021). In bacterial cells, DNA is usually slightly underwound (Lal et al., 2016), with a negative value of σ. The basal supercoiling level depends on the species, but a typical value is $\sigma_{basal} \approx -0.066$ in *Escherichia coli* (Crozat et al., 2005). DNA supercoiling is tightly regulated by a class of enzymes called topoisomerases. The main topoisomerases are topoisomerase I and gyrase: gyrase uses ATP to maintain DNA in a negative supercoiling state by adding negative coils, while topoisomerase I relaxes supercoiling and does not need ATP (Martis B. et al., 2019).

DNA supercoiling furthermore plays an important role in bacterial cells as an ancestral regulator of gene activity (Dorman and Dorman, 2016). Indeed, as shown in figure 1A and 1B, high negative supercoiling levels, or hypercoiling ($\sigma < \sigma_{basal}$), favor higher expression of bacterial genes, as the thermodynamic reaction of promoter opening required to begin transcription is facilitated (El Houdaigui et al., 2019). Conversely, high supercoiling levels, or DNA relaxation ($\sigma > \sigma_{basal}$), reduce gene expression. DNA hypercoiling and relaxation have been shown to affect the expression of 10% and 13% respectively of the genome of *Streptococcus pneumoniae*, in which supercoiling-sensitive genes can see up to 5-fold increases or decreases in their expression level when subjected to DNA relaxation (de la Campa et al., 2017). In *Escherichia coli*, DNA relaxation has been shown to affect the expression of around 7% of the genes (Peter et al., 2004). Moreover, in some bacterial species such as *Buchnera aphidicola*, an obligate aphid endosymbiont with a highly reduced genome (Viñuelas et al., 2007), gene regulation still takes place even in the near-total absence of traditional transcription factors; in these species, DNA supercoiling is suspected to be the main, if not the sole, regulatory mechanism (Brinza et al., 2013).

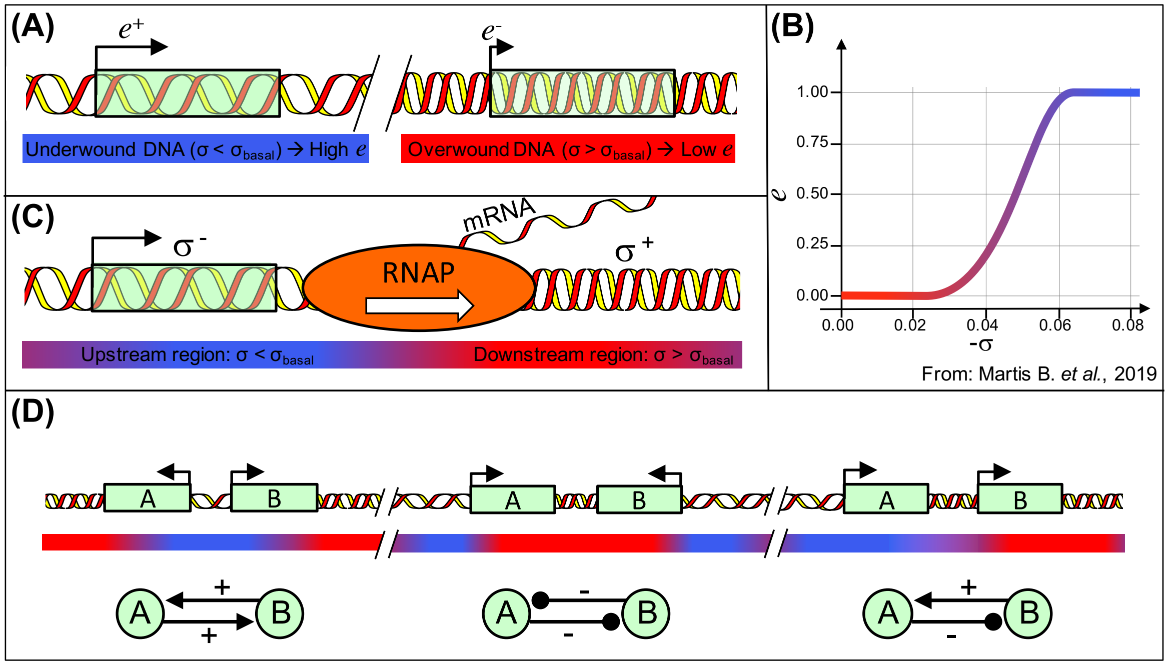
(A)
e+
e-
Underwound DNA (σ < σbasal) → High e
Overwound DNA (σ > σbasal) → Low e
(B)
1.00
0.75
e 0.50
0.25
0.00
0.00
0.02
0.04
0.06
0.08
-σ
From: Martis B. et al., 2019
(C)
σ -
mRNA
RNAP
σ +
Upstream region: σ < σbasal
Downstream region: σ > σbasal
(D)
A
B
+
-

Figure 1: **A**. When DNA is underwound ($\sigma < \sigma_{basal}$, left), gene transcription rates are higher than when DNA is overwound ($\sigma > \sigma_{basal}$, right). **B**. Promoter activity (equivalently, transcription level) e increases with the level of negative supercoiling $-\sigma$. **C**. The transcription of a gene by RNA polymerase (RNAP) generates a decrease in supercoiling upstream of the transcribed gene, and an increase downstream of the transcribed gene. **D**. Transcription-supercoiling coupling: the sign of the interaction between neighboring genes depends on their relative orientation.

1.1 Dynamic Properties of DNA Supercoiling

DNA supercoiling is under the influence of both internal and external constraints. It varies both in time, during the lifecycle of the bacterium, which alternates between growth and stationary phases (Krogh et al., 2018), and in space, as different regions of the chromosome experience different supercoiling levels (Lal et al., 2016; Junier and Rivoire, 2016). In bacteria, the supercoiling homeostasis is mainly regulated by the topoisomerases, but nucleoid-associated proteins (NAPs) such as *H-NS* also play a topological role by preventing the relaxation of superhelical stress at their fixation points in the genome, resulting in topological domains that have different supercoiling levels (Krogh et al., 2018). Moreover, the global supercoiling level can change as a genome-wide response to extracellular stresses, such as changes in pH, osmolarity, temperature, or oxidative stress (Duprey and Groisman, 2021).

Crucially, the transcription process itself plays a role in local supercoiling level variations. Indeed, when an RNA polymerase (RNAP in figure 1C) transcribes a gene, it follows the helical twist of the DNA template, but its rotation is hampered by frictional drag (Ma and Wang, 2016). The RNA polymerase therefore acts as a topological barrier and generates an accumulation of positive supercoiling through overwinding of the DNA molecule downstream of the transcribed region, and of negative supercoiling through underwinding upstream of the transcribed region, in what has been called the twin-domain model of DNA supercoiling (Liu and Wang, 1987), as shown in figure 1C.

1.2 The Transcription-Supercoiling Coupling

As transcription is itself regulated by supercoiling (figures 1A and B), these interactions result in a possible dynamic coupling between the transcription levels of neighboring genes, which has been named the transcription-supercoiling coupling (Martis B. et al., 2019). As previously described, when a gene is transcribed (figure 1C), DNA upstream of the transcribed gene is underwound and the level of negative supercoiling increases (σ decreases), while DNA downstream of the transcribed gene is overwound and the level of negative supercoiling decreases (σ increases). This, in turn, affects the transcription of neighboring genes, as a higher level of negative supercoiling at the promoter of a gene increases its transcriptional activity. Therefore, a highly transcribed gene can increase the transcription level of the genes that are upstream of itself, and decrease the transcription level of the genes that are downstream of itself. The influence of this coupling on gene transcription is represented in figure 1D: when two neighboring genes lie in diverging orientations (figure 1D, left), the transcription of each gene generates a local increase in negative supercoiling around the other gene, thereby increasing the transcription level of that gene; each gene therefore reinforces the transcription of the other gene. Conversely, when two neighboring genes face each other in converging orientations (figure 1D, center), each gene is located downstream of the RNA polymerase during the transcription of the other gene, leading to a decrease in negative supercoiling and therefore a lower transcription level. In that case, each gene inhibits the other gene. Finally, if two genes are in a colinear orientation (figure 1D, right), the downstream gene up-regulates the upstream gene, and the upstream gene down-regulates its downstream neighbor.

Several mathematical and computational models have been proposed to describe the effect of the transcription-supercoiling coupling on the expression level of neighboring genes. In Meyer and Beslon (2014), a quantitative model of the supercoiling level at a locus of interest is proposed. DNA transcription is regulated by the opening energy of DNA around gene promoters, which directly depends on the supercoiling level. In this model, the reciprocal influence of neighboring genes can be obtained by computing the difference in transcription levels due to supercoiling and the subsequent variation in supercoiling, and iterating this system until a fixed point is reached. El Houdaigui et al. (2019)

describe a more detailed stochastic model of DNA transcription involving explicit RNA polymerases and topoisomerases. The transcription level of a genomic region of interest is simulated using discrete time steps, during which RNA polymerases attach to the DNA template, progress along the transcribed region while generating positive supercoiling downstream and negative supercoiling upstream, and detach from the DNA, relaxing supercoiling constraints.

These models however limit themselves to mechanistic descriptions of the local interaction between genes, but do not try to generalize to the whole-genome scale nor to an evolutionary time frame. Yet, the dense gene content of bacteria suggests that the transcription-supercoiling coupling can generate a global transcriptional interaction network through the propagation of local supercoiling variations. Indeed, in bacteria, the distances between the beginning of two consecutive genes average around 1,000 bp (Blattner, 1997). This is low enough to connect multiple genes through the transcription-supercoiling coupling, as the typical distance at which this interaction operates is around a few thousand base pairs on each side of the transcribed gene (El Hanafi and Bossi, 2000). Measurements of the expression level of neighboring genes in bacteria have moreover experimentally demonstrated the existence of a coupling between transcription and supercoiling in specific gene systems, such as the ilvIH-leuO-leuABCD region in *Escherichia coli* or *Salmonella typhimurium* (El Hanafi and Bossi, 2000; Sobetzko, 2016; Dorman and Dorman, 2016). Finally, experimental evidence of this coupling has also been obtained in a synthetic system using the *ilvY* and *ilvC* promoters of *E. coli* placed in diverging orientations on a plasmid, in which a decrease (resp. increase) in *ilvY* expression was correlated with a decrease (resp. increase) in *ilvC* expression (Rhee et al., 1999).

In addition to a localized influence, global gene regulation through changes in the DNA supercoiling level has been shown to exist in nature. An example of this is *Buchnera aphidicola*, an endosymbiotic bacteria with a streamlined genome, in which control of the supercoiling level has evolved to be one of the main regulatory mechanisms (Brinza et al., 2013). Moreover, in *Dickeya dadantii*, a plant pathogenic bacteria, different genomic regions exhibit markedly different responses to changes in supercoiling (Muskhelishvili et al., 2019), allowing the expression of pathogenic genes only in stressful environments. Finally, in the setting of experimental evolution, mutations in the regulation of supercoiling have been shown to drive the evolutionary response of *Escherichia coli* strains. In the Long-Term Evolution Experiment (LTEE) (Lenski et al., 1991), 12 populations of *Escherichia coli* have been maintained for over 80,000 generations, evolving and adapting to a glucose-limited environment. In this experiment, parallel increases in the level of negative supercoiling been measured in 10 of the populations (Crozat et al., 2010), suggesting that a more negatively supercoiled genome confers an evolutionary advantage for that environment. Furthermore, mutations in two genes regulating DNA supercoiling, *topA* and *fis*, have been identified as the genetic basis for this phenotypic change, and have been verified to confer a fitness advantage (Crozat et al., 2005) when inserted into the genetic background of the ancestral strain. These results suggest a strong selection pressure to tune the level of DNA supercoiling to the new environment of the LTEE.

In conclusion, DNA supercoiling plays an important evolutionary role. It can serve as the support for the evolution of particular chromosomal organizations as a way to trigger certain sets of genes based on the change in supercoiling caused by specific environments, and its regulation is itself subject to selective pressure when adapting to new environments.

Both the importance of supercoiling regulation and the detailed mechanisms of the transcription-supercoiling coupling at the local scale have well been studied, but a thorough analysis of the genome-wide effect of the transcription-supercoiling coupling on gene expression, and of its possible evolutionary use by natural selection, remains missing. In this work, we describe a new model which incorporates a high-level model of global supercoiling regulation and of the transcription-supercoiling coupling within an *in silico* experimental evolution setting. Using this model, we first investigate the non-linear variation in gene transcription levels at the whole-genome scale in response to variations in the global supercoiling level. Then, we study the evolutionary trajectory of gene activation patterns in individuals subjected to different environments.

We show that in our model, a genome-scale gene interaction network emerges from local interactions, and creates a reaction norm in response to the change of a single parameter, the global supercoiling level, caused by different environments. Moreover, we demonstrate that, using genomic inversions as the only mutation operator, and therefore only changing the relative positions and orientations of genes on the genome, evolution can select genomes displaying qualitatively different phenotypes in different environments, characterized by different global supercoiling conditions.

This article is an extended version of Grohens et al. (2021).

2 A Genome-Wide Model of the Transcription-Supercoiling Coupling

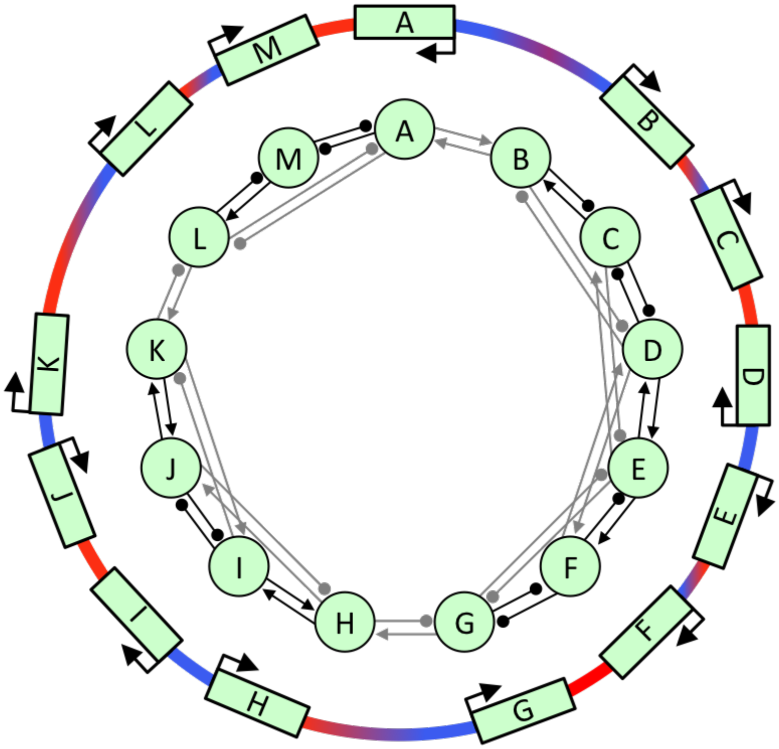

Figure 2: Genes along an example genome and local variations in supercoiling (outer ring), and the associated gene interaction network (inner ring). The outer ring color shows locally high ($\sigma > \sigma_0$, red) or low ($\sigma < \sigma_0$, blue) supercoiling levels due to gene transcription. In the inner ring, closer genes interact more strongly (black arrows) than genes that are farther apart (gray arrows), either positively or negatively depending on their relative orientations.

Our model is written in Python and consists in an individual-based simulation, the source code of which is available at `https://gitlab.inria.fr/tgrohens/evotsc/-/tree/alife-journal`. It is also preserved for long-term archival (Grohens, 2021) using the Software Heritage archive (Di Cosmo, 2020). An individual in the model is represented by a circular genome (representative of most bacterial genomes), comprising a fixed number of genes, separated by non-coding intergenic regions. Each gene is described by the following characteristics: its locus on the genome, its orientation, and its basal transcription (or expression) level. As we are mainly interested in the interplay between supercoiling and transcription, we voluntarily do not make the difference between gene expression levels, understood as mRNA or protein concentrations, and transcription levels, the immediate rate of mRNA production. Indeed, assuming a separation of timescales between the fast equilibrium of the transcription-supercoiling coupling, and the slow degradation of mRNAs, the concentration of a given mRNA is directly proportional to the transcription rate of its source gene.

Figure 2 shows the role played by the transcription-supercoiling coupling in an example genome. It includes the local supercoiling variations due to gene transcription, and the resulting gene interaction network, with each gene possibly activating or inhibiting its neighbors, depending on their relative orientations. Importantly to our approach, here genes do not interact only with their closest neighbors, but also with more distant genes, as is assumed to be the case in the gene-rich bacterial genomes (remember that *E. coli* genes are around one thousand base pairs apart, and that transcription-generated supercoiling propagates around a few thousand base pairs on each side of the transcription site).

2.1 Mathematical Description of the Model

We model the transcription-supercoiling coupling between an individual's genes as a system of equations, which relate the supercoiling level at the locus of each gene σ_i (for i ranging from 1 to n, the number of genes of the individual), and the expression level of every gene e_i. The parameters of the system are described by the genome of the individual, as will be detailed below.

In our model, the supercoiling at a given locus on the genome depends on three factors: the individual's basal supercoiling level σ_{basal}, the variation in supercoiling due to environmental conditions σ_{env}, and the variation in supercoiling due to the transcription of the neighboring genes. We compute this local variation in supercoiling at the locus of each gene with the help of a gene interaction matrix, whose coefficient at position (i, j) describes the influence of gene j on gene i. The coefficients are given by the following equation:

$$\frac{\partial \sigma_i}{\partial e_j} = \eta \cdot c \cdot \max(1 - \frac{d(i,j)}{d_{max}}, 0) \tag{1}$$

More precisely, the interaction level between two genes depends on the relative orientation of the genes, as the transcription of a gene increases supercoiling at the locus of downstream genes and decreases supercoiling at the locus of upstream genes (remember that an increase in supercoiling means a decrease in transcription). Therefore, we choose $\eta = 1$ if gene i is downstream of gene j and $\eta = -1$ otherwise (if $i = j$, $\eta = 0$ as a gene does not interact with itself). The interaction level also depends on gene distance, as genes that are further apart on the genome interact less strongly, so the strength of the interaction linearly decreases with the intergenic distance $d(i, j)$, and reaches 0 when $d(i, j) = d_{max}$, the maximum distance above which the interaction vanishes. Finally, an interaction coefficient c is applied to adjust the strength of the coupling.

Using this interaction matrix, we compute the level of supercoiling σ_i at the locus of every gene, which depends on the transcription level of all the other genes, on the basal supercoiling and on the environmental supercoiling:

$$\sigma_i = \sigma_{basal} + \sigma_{env} + \sum_{j=1}^{n} \frac{\partial \sigma_i}{\partial e_j} e_j \tag{2}$$

The transcription level e_i of every gene as a function of total supercoiling is then modeled with a sigmoidal activation curve, following El Houdaigui et al. (2019). The equation is given by:

$$e_i = \frac{1}{1 + e^{(\sigma_i - \sigma_0)/\varepsilon}} \tag{3}$$

In this equation, σ_0 is a parameter that represents the inflexion point of the sigmoid, that is the supercoiling level at which the gene is at half its maximum transcription rate, and ε a scaling factor that represents the strength of the dependence of the transcription level on the supercoiling level.

Finally, in order to obtain the phenotype of an individual, we numerically compute a solution to the system of equations 2 and 3, using a fixed point algorithm. This solution represents the state (of gene expression and supercoiling at every locus) towards which the individual would converge over time. Let $f(e_i)$ be the function that computes new supercoiling levels σ_i' from e_i using equation 2, then computes new expression levels e_i' from the new σ_i' using equation 3, and finally returns e_i'. In order to compute a fixed point of f, that is a set of transcription levels e_i^* such that $f(e_i^*) = e_i^*$, we start with the basal transcription levels e_i^0 that are a property of each gene, and iterate the sequence $e_i^{t+1} = \frac{1}{2}(e_i^t + f(e_i^t))$, until the difference between two successive iterations is below a given threshold. In our setting, this algorithm has empirically always converged to a solution that is a stable fixed point of the function, and that is therefore interpretable from a biological perspective.

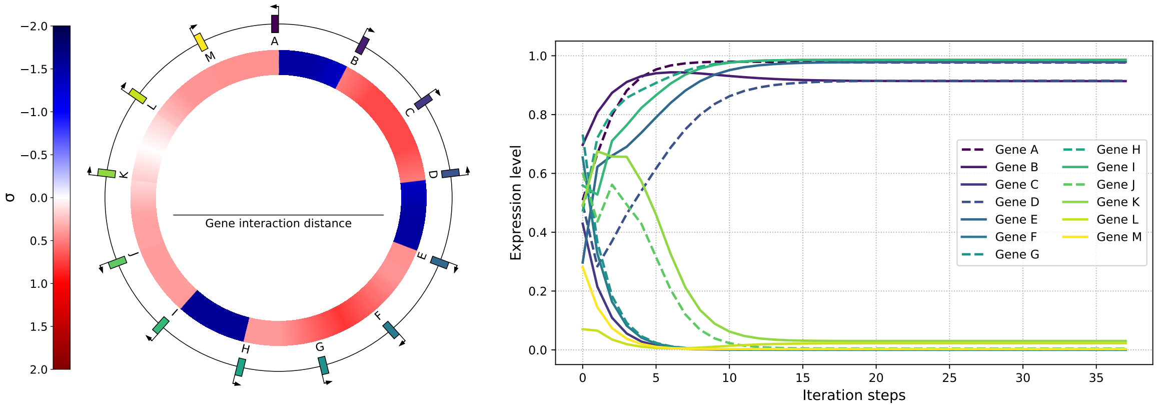

Figure 3: Left: genome (outer ring) and stable state level of supercoiling σ (inner ring) of an example individual with 13 genes in the model. Right: transcription levels of the individual's genes during the iterations of the fixed point computation, in an environment given by $\sigma_{env} = 0.05$. Solid lines represent genes in forward orientation, and dashed lines genes in reverse orientation.

Figure 3 shows the genome (left, outer ring) of an example individual with a genome of 13,000 bp and $n = 13$ genes evenly spaced along the genome, and with a basal supercoiling of $\sigma_{basal} = -0.06$. The basal transcription level of each gene is randomly chosen between 0 and 1, and the iterations of the fixed point algorithm giving the final gene transcription levels are shown on the right. In this individual, the non-linear effect of the interaction between neighboring genes is clearly visible. Indeed, six genes (A, B, D, E, H, and I) end up at a high transcription rate at the fixed point (or solution) of the system, while the others end up at low transcription rates. These activated genes can be grouped into 3 pairs (A and B, D and E, H and I), all of which are pairs of adjacent genes in divergent orientations. Even though gene D has a low (around 0.3) basal transcription rate, it eventually reaches a high transcription state because of its positive interaction with gene E. Conversely, genes F and G start with a high transcription rate, but are repressed by their neighbors H and E, and are therefore silenced when the system converges. We can also observe complex behaviors in the model, as the gene expression levels pass through very different states during convergence to the solution. Indeed, the transcription level of gene K initially increases due to its interaction with gene J, but both genes end up in a low transcription state, as they are inhibited by the very active gene I. The final supercoiling level along the genome (left, inner ring) moreover demonstrates the effect of the transcription-supercoiling coupling on local supercoiling. Highly transcribed genes, such as A and B, generate a large variation in the supercoiling level on their upstream and downstream sides, and the positive feedback loop between genes in divergent pairs is made clear by the very high negative value of the supercoiling level between each gene in these two pairs.

2.2 Effect of the Environmental Supercoiling on Gene Activation Levels

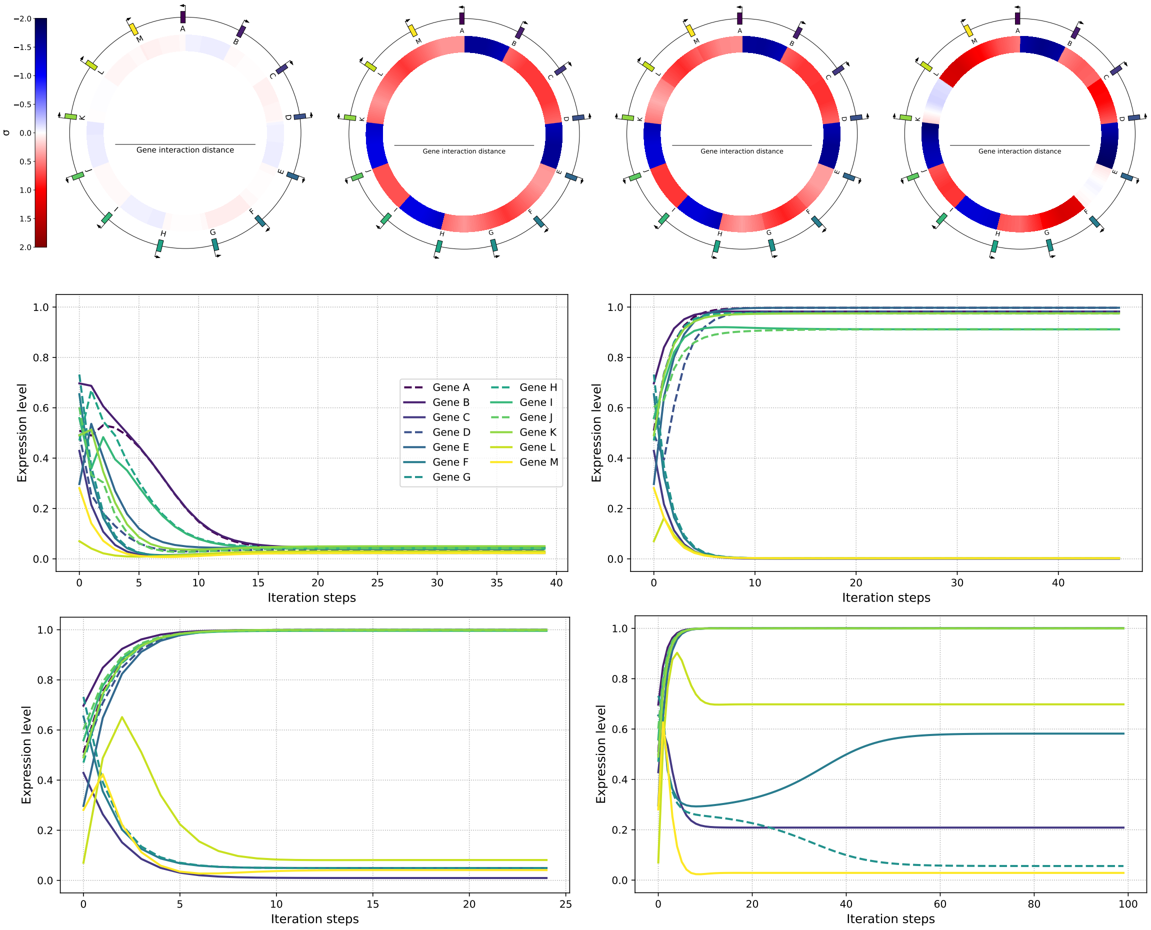

Figure 4: Influence of the environment supercoiling σ_{env} on the stable state local supercoiling level (top row) and gene transcription levels (bottom rows) of the example individual. From left to right and top to bottom: at $\sigma_{env} = 0.1$, no genes are activated ($e > 0.5$); at $\sigma_{env} = 0.0$ and at $\sigma_{env} = -0.1$, 8 genes are activated; at $\sigma_{env} = -0.2$, 10 genes are activated. Lower values of σ_{env} result in the activation of more genes, reflecting the *in vivo* effect of higher negative supercoiling.

Figure 4 captures the influence of the environmental change in supercoiling σ_{env} on the local supercoiling level due to the transcription-supercoiling coupling (top row) and on the repartition of genes between the activated and inhibited states (bottom rows), again using the example individual already shown in figure 3. From left to right and top to bottom: at a high value of $\sigma_{env} = 0.1$, meaning that DNA is severely overwound compared to normal, no gene at all is activated (meaning that $e > 0.5$).

As the external influence of the environment on supercoiling decreases to $\sigma_{env} = 0$, corresponding to normal relaxation of DNA, and then to $\sigma_{env} = -0.1$, 8 of the 13 genes of the individual reach an activated state. Finally, for $\sigma_{env} = -0.2$, there is a strong environmental pressure towards high gene transcription levels, and most genes are indeed activated; however, even at this level of σ_{env}, some genes remain shut down, because of the high amount of positive supercoiling (in red) generated by the transcription of their neighbors.

2.3 Influence of Relative Gene Positions on Gene Activation Levels

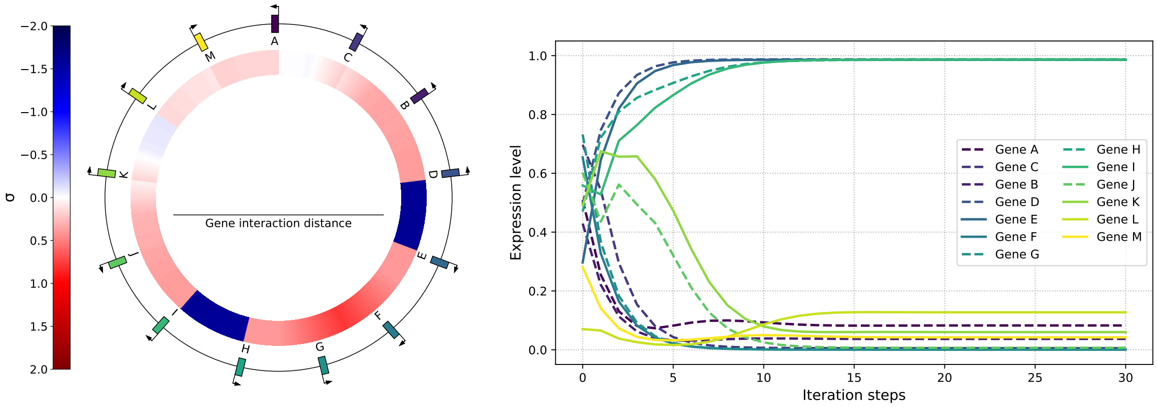

Figure 5: Genome, local supercoiling and gene expression levels of a new individual obtained from the individual in figure 3 by switching the positions and orientations of genes B and C.

Figure 5 again shows the local supercoiling and gene expression levels of the individual in figure 3, after reversing the positions and orientations of genes B and C. This is an example of a genomic inversion, which will be presented in further detail in section 3.3. The start point of this inversion falls between genes A and B and its end point between genes C and D; this results in the reversal of segment [BC] relative to the rest of the genome. Here, we can see that the diverging orientation that was present between genes A and B has vanished, replaced by a set of genes in colinear orientation, from A to D. This genomic reorganization results in the loss of the activation of genes A and B, as gene B is now more strongly inhibited by gene D due to its closer genomic location, and as genes A and B are not in a positive feedback loop due to diverging orientations any longer; only the pairs of genes D and E, and H and I, remain activated.

Based on these observations, we can confirm that in our model, the transcription-supercoiling coupling generates complex networks of genome-wide interactions between genes, and that these networks are directly dependent on the architecture of the genome.

3 An Evolutionary Genome-Wide Model of the Transcription-Supercoiling Coupling

Having shown that transcriptional activity depends on the organization of the genome, we now question to which extent evolution can simultaneously leverage the organization of the genome and the transcription-supercoiling coupling in order to adapt gene regulatory activity to different environments. Indeed, as has been observed in *Dickeya dadantii* (Muskhelishvili et al., 2019), different phenotypes can evolve as a response to different supercoiling levels induced by the environment, and the transcription-supercoiling coupling could play a role in enabling the existence of this reaction norm.

In this section, we expand our model into an evolutionary simulation. At each generation of the simulation, all individuals are evaluated and their fitness values are computed, based on their gene transcription levels. Then, the individuals of the new generation are chosen by picking their ancestor from the current generation, with a probability proportional to the ancestor's fitness. The model is panmictic, meaning that any individual in the population can be chosen as the ancestor of any new individual. Finally, during replication, the genome of each new individual stochastically undergoes a number of mutations, before the new individual is evaluated again; importantly, these mutations do not impact genes themselves, but only the spatial organization of the genome: gene orientations, syntenies, and intergenic distances.

3.1 Evolutionary Model: Evolution in Two Separate Environments

We model the evolution of populations of individuals that experience two different environments, named A and B. Each environment is defined by its value of σ_{env}, respectively σ_A and σ_B, which represent the change in the supercoiling level due to the environment (Dorman and Dorman, 2016). In order to have environments with distinct effects, we choose a value of $\sigma_A = 0.1$, for which isolated genes are effectively inhibited (as in the top-left panel of figure 4), and a value of $\sigma_B = -0.1$, for which some but not all genes are activated (bottom-left panel).

We separate genes into three classes, based on the environments in which they must be activated: either in both environment A and environment B (AB genes), only in environment A (A genes), or only in environment B (B genes). These classes allow us to define optimal phenotypes for both environments: in environment A, both A and AB genes should be activated, whereas B genes should be inhibited. Conversely, in environment B, only B and AB genes should be activated, but not A genes.

3.2 Fitness

In order to compute the fitness of an individual, we define an optimal phenotype $\tilde{e}^A$ (resp. $\tilde{e}^B$), corresponding to the vector of the expected expression level $\tilde{e}_i^A$ for each gene i in environment A (resp. environment B). We choose an expected expression level of $\tilde{e} = 1$ for genes that should be activated, which corresponds to the maximum possible expression level of a gene in our model. Similarly, we choose $\tilde{e} = 0$ for genes that should be inhibited, which is the minimum expression level that is attainable. Then, in each environment, we compute the gap g_A (resp. g_B), or average square distance of the individual's gene transcription levels e^A (the vector constituted by the transcription level e_i^A of each gene i) to the optimal levels $\tilde{e}^A$ (resp. e^B and $\tilde{e}^B$). The gap g_A is computed as follows:

$$g_A(e^A) = \frac{1}{n} \sum_{i=1}^{n} (e_i^A - \tilde{e}_i^A)^2 \tag{4}$$

The gap g_B is computed in the same way. Finally, we compute the fitness of the individual by summing the gap in each environment, and applying an exponential scaling: $f = e^{-k(g_A+g_B)}$, where k is a scaling factor representing the selection pressure. A higher value of k means that well-adapted individuals, those which have a smaller gap, will have an even higher fitness value compared to other individuals; we typically use $k = 50$, meaning that a small decrease in the gap compared to other individuals yields a large reproductive advantage.

3.3 Mutational Operator: Genomic Inversions

We introduce only one kind of mutation in our model, which is genomic inversions: we choose two breakpoints randomly on the genome, and reverse the genomic content between these points. Genes are then reinserted in the genome in the opposite orientation and order, taking care to update all intergenic distances appropriately. Note that in our model, genes have a length of zero and the break points can therefore not fall inside a gene. Moreover, an inversion has no effect if both breakpoints fall between two neighboring genes (as only an intergenic region would be affected), but can impact any number of genes otherwise. Genomic inversions hence affect gene syntenies and orientations, and therefore affect gene expression levels as presented in subsection 2.3. When mutating a genome during reproduction, we draw the number of inversions k to perform from a Poisson law with parameter $\lambda = 2$, giving an average of 2 inversions between an individual and its ancestor; the probability of not undergoing any mutations is $P(k = 0) = e^{-\lambda} \approx 0.136$.

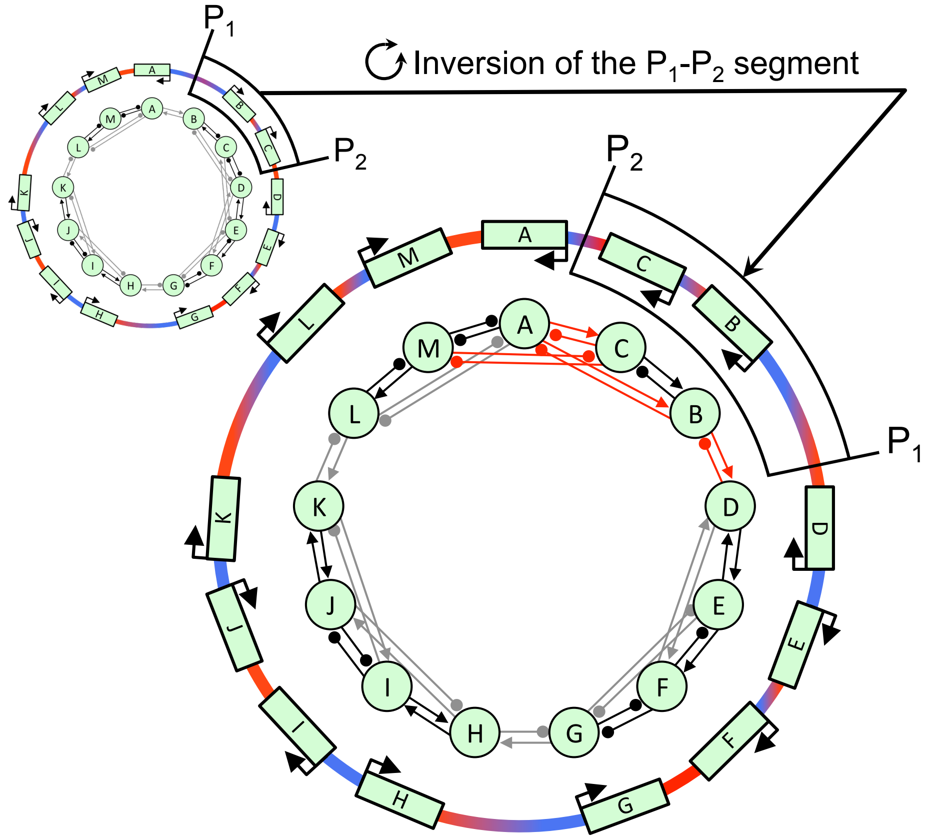

Figure 6: Result of the inversion of a genomic segment containing genes B and C from the individual presented in figure 2. The gene interactions which have changed due to the inversion are drawn in red. This illustration genome corresponds to the actual individual in our model presented in figure 5.

Figure 6 presents a genome obtained by performing an inversion on the genome shown in figure 2. As a result of this inversion, genes B and C have been switched from the forward to the backward orientation, and the intergenic distances between A and C on the one hand, and B and D on the other hand, have been modified; however, the relative orientation of B and C, and hence their interaction subnetwork, remain unchanged. This results in changes to the gene interaction network: instead of mutual activation between genes A and B and mutual inhibition between genes C and D, all four of those genes now lie in colinear orientations, in which each of these genes activates its upstream neighbor but represses its downstream neighbor.

3.4 Experimental Setup and Parameter Values

We initialized the simulation with a clonal population of $N = 100$ copies of an initial individual with the following genome: 60 genes in random orientations, uniformly distributed along a 60,000 bp genome, and equally divided between the AB, A and B classes. We chose a maximum interaction distance of $d_{max} = 2500$, meaning that each gene initially interacts with its 2 closest neighbors in each direction through the transcription-supercoiling coupling. Note that as inversions may change intergenic distances, genes can become closer or further apart during evolution. We set the basal supercoiling level σ_{basal} to the average supercoiling level in *E. coli* of -0.06 (Crozat et al., 2005), and σ_0 to -0.06 as well, so that in the absence of other sources of supercoiling (either environmental or through the coupling), the default activity level of a gene is 0.5. Finally, we set $c = 0.3$, in order to have comparable values for the variations in supercoiling due to the environment and due to the transcription-supercoiling coupling, and $\varepsilon = 0.03$, so that the variations in supercoiling have a qualitatively mild effect on gene expression.

In order to run the simulations, we evolved 15 different populations for 250,000 generations; the simulation lasted for approximately 48h on a computer with Intel Xeon E5-2640 v3 @ 2.60GHz CPUs, using around 100 MB of RAM per replicate.

3.5 Adaptation of Gene Expression Levels to Different Environments

Figure 7 summarizes the differences in the proportion of activated genes for each of the three sets of genes, between environments A and B, averaged over the 15 repetitions. In the figure, we consider a gene to be activated if its activity at the end of the lifecycle is over 0.5, and look at the average proportion of activated genes in the best individual of every replica. Recall that the evolutionary target for AB genes is an expression level of 1 in both environments, for A genes an expression level of 1 in environment A and 0 in B, and vice-versa for B genes. After 250,000 generations of evolution, individuals have acquired genomes that allow all AB genes to be activated in both environments, and that allow all B genes to be activated in environment B and inhibited in environment A. On average, over 60% of A genes are activated in environment A, which imposes a positive change in supercoiling ($\sigma_A = 0.1$) and makes gene activation harder. Conversely, less than 5% of A genes are activated in environment B, in which it is easier for genes to be activated ($\sigma_B = -0.1$). The final expression levels of A genes therefore show that specific sets of genes can be activated by the transcription-supercoiling coupling despite environmental hurdles.

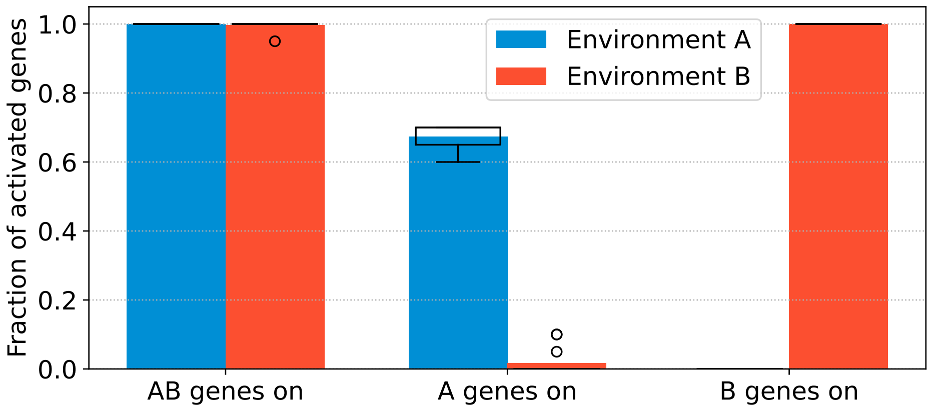

Figure 7: Fraction of activated genes of each type in each environment at the end of the lifecycle, averaged over the best individual in the last generation of each replica. Boxplots represent the median and quartiles, and dots flier data points. For A genes and B genes, activation levels differ depending on the environment: p-value 2.40×10^{-17} for A genes, and p-value $< 1 \times 10^{-25}$ for B genes (Student's t-test for dependent samples).

Furthermore, in each of the 15 replicates, the fitness of the best individual in the population increases continuously over the course of evolution, as shown in figure 8. As the fitnesses keep increasing until the end of the simulation, this suggests that fitter phenotypes remain reachable through further evolution by genomic rearrangements. The rhythm of evolution is however progressively slower and slower (note the logarithmic time scale in the figure), as the pool of available favorable mutations decreases.

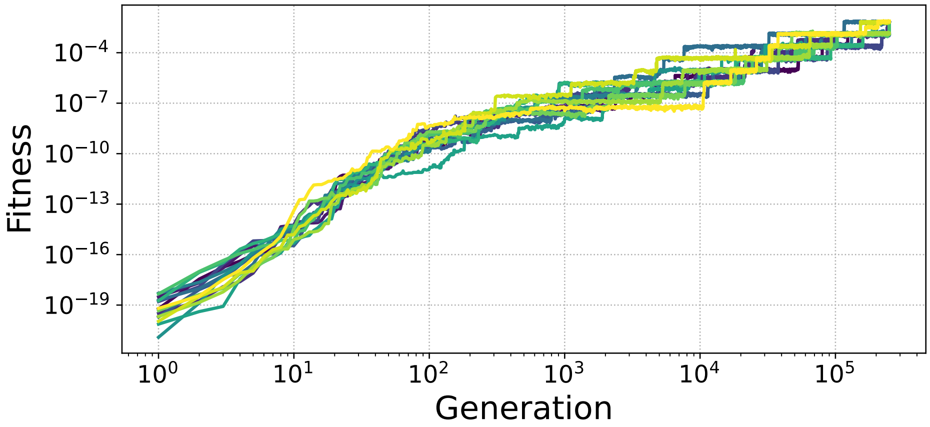

Figure 8: Evolution of the fitness of the best individual of each replicate at every generation.

Finally, details of the evolution of one of the 15 replicate populations are shown in figure 9. We can first see that the number of activated AB genes of the best individual at each generation quickly rises to 20 (out of 20 genes of that type) in both environment A and environment B; this shows that evolving a phenotype that is resistant to environmental perturbations, having genes that are always activated, is easy in the model. For A genes and B genes, we observe an asymmetric tendency during the course of evolution towards activation in the target environment, and inhibition in the opposite environment. However, the difference in the number of activated B genes between environment A and environment B is much higher than for A genes. As already mentioned above, this asymmetry comes from the different requirements expected of A genes and B genes: gene activation is easier in environment B than in environment A, as it is easier for a gene to become activated in an environment with a lower overall supercoiling level. A genes therefore have to be activated in a harder environment, and inhibited in a simpler environment, whereas B genes have to do the opposite.

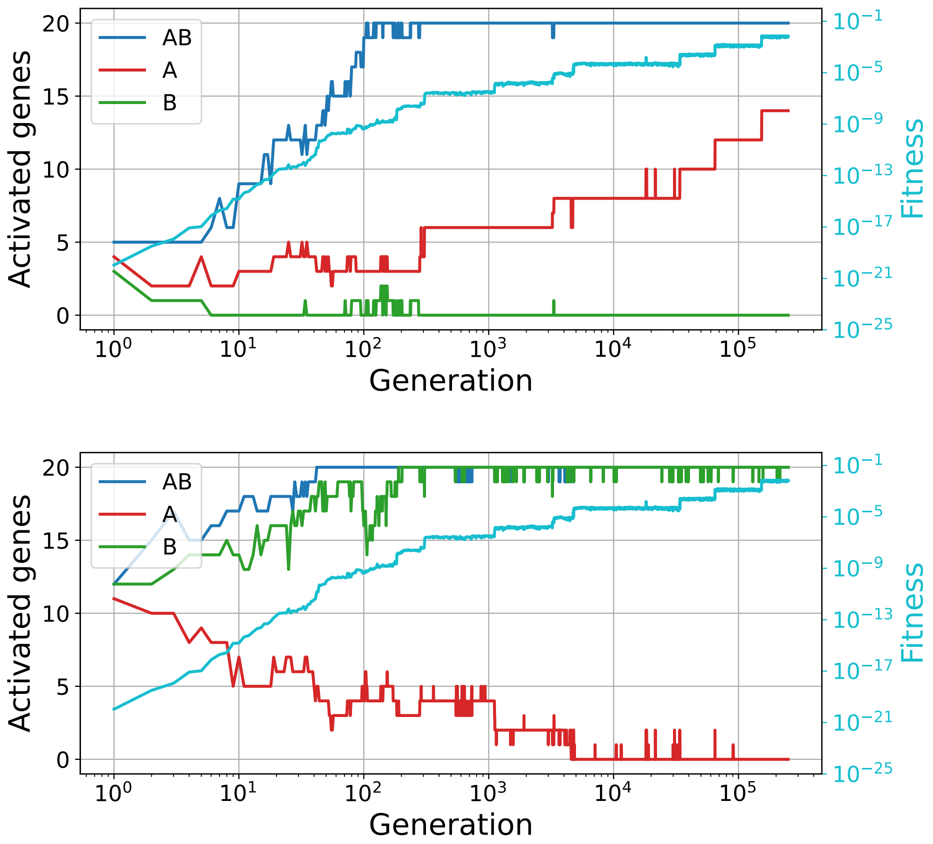

Figure 9: Number of activated genes of each type and fitness of the best individual at every generation of replicate 13, with a population size of $N = 100$, for 250,000 generations. The number of active AB genes increases until it reaches 20, in both environment A (top) and environment B (bottom). The number of active A (resp. B) genes increases in environment A (resp. B) and decreases in environment B (resp. A) over time, thus converging towards their evolutionary target.

This is shown in more detail in figure 10, which shows the supercoiling level and gene activation levels of the best individual of the last generation of replicate 13, in both environments. The phenotypes displayed in each environment present clearly distinct gene expression patterns. In environment A (top), nearly all genes converge directly towards their final state, whereas in environment B (bottom), most *A* genes (in red) and some *B* genes (in green) show a complex trajectory of activation levels before reaching the stable state. Moreover, genomic domains with markedly different supercoiling levels emerge through the transcription-supercoiling coupling, with both very overwound and very underwound zones. These domains also show qualitatively different responses to the different environments: in some domains, the supercoiling level is very similar (around gene 0, gene 15 or gene 55 for example), while in others supercoiling is completely different in each environment (between genes 20 and 35). This shows the plasticity of the response to environmental change at the local supercoiling level.

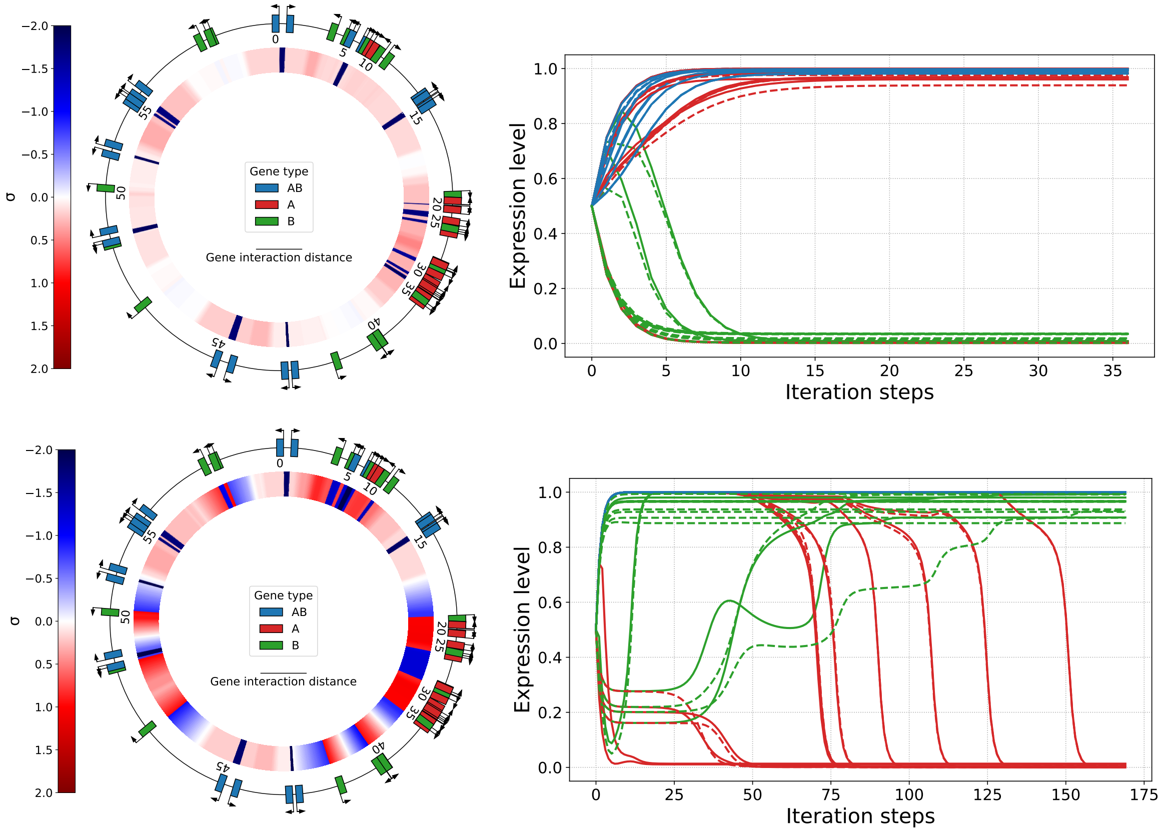

Figure 10: Local supercoiling along the genome and gene transcription levels of the best individual in replicate 13 after 250,000 generations. Environment A is on top and environment B at the bottom. *AB* genes are colored blue, *A* genes colored red, and *B* genes colored green.

Our experimental results show that, in a model of gene transcription that is structured around the transcription-supercoiling coupling, complex gene interaction networks can evolve. These gene interaction networks are sensitive to environmental variations, which are mediated in our model by a single parameter: σ_{env}, the amount of global supercoiling that is due to the environment.

3.6 Robustness of Gene Network Evolution

In order to ensure that our results remain experimentally valid over a broad range of parameter values, we ran additional sets of simulations, in which we changed respectively the sensitivity of gene promoters to supercoiling changes (ε in equation 3), the interaction coefficient used in computing the local supercoiling due to the transcription-supercoiling coupling (c in equation 1), and the strength of the change in supercoiling imposed by the environment (σ_A and σ_B). We chose sets of logarithmically-spaced values for each parameter, and ran 5 replicates of the evolution experiment for 250,000 generations for each parameter value. Note that, for extreme parameter values, gene expression levels did in some cases not converge to stable states by the maximum number of computation steps. In this situation, we chose to retain the gene expression levels at the last step as the phenotype of the affected individuals.

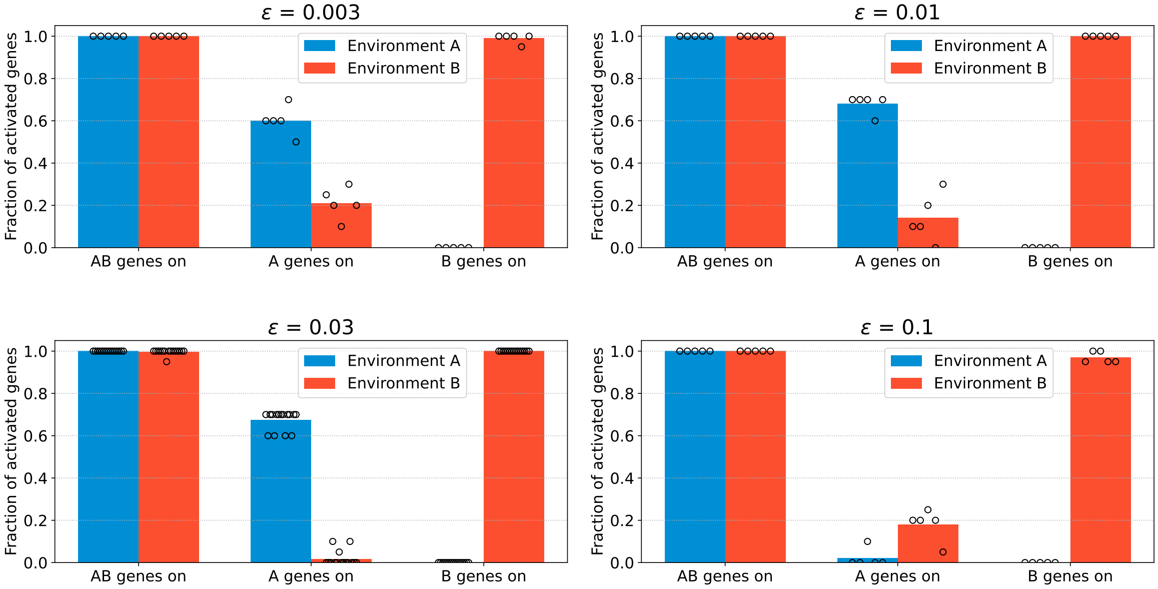

Figure 11: Average fraction of activated genes in each environment at the end of evolution, for increasing values of ε, from top to bottom and left to right. Every replicate is shown as a dot, and the bottom-left panel ($\varepsilon = 0.03$) recalls data from the main run (which has 15 replicates) for comparison. For all values of ε except 0.1, the behavior from the main run is qualitatively replicated.

The results of the additional simulations are presented in figures 11, 12 and 13. For ε, we chose values of $\varepsilon = 0.003$, $\varepsilon = 0.01$, and $\varepsilon = 0.1$, compared to an initial value of $\varepsilon = 0.03$, and the results are shown in figure 11. For the values of ε lower than the default (top row), representing a higher sensitivity of promoters to supercoiling, we observe the evolution of differentiated gene expression levels as in the main run (bottom-left panel), whereas for the higher value of ε (bottom-right panel), A genes are not expressed in environment A by the end of evolution. In this case, promoters are not sensitive enough to the supercoiling variations caused by the transcription-supercoiling coupling, and genes are unable to overcome the highly positive supercoiling of environment A.

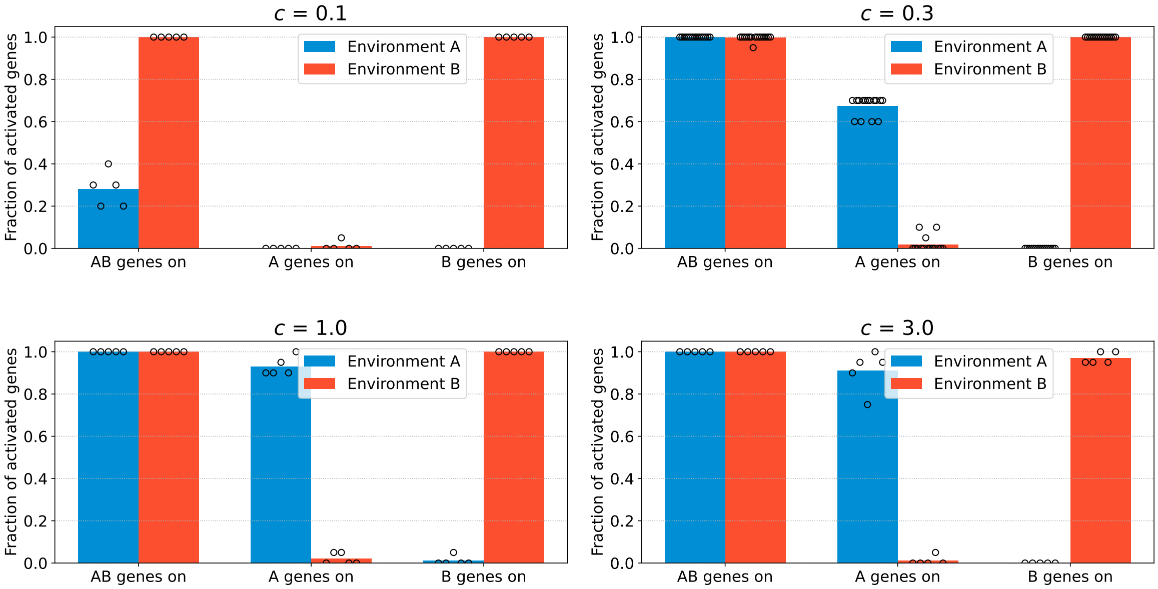

Figure 12: Average fraction of activated genes in each environment at the end of evolution, for increasing values of c, from top to bottom and left to right. Every replicate is shown as a dot, and the top-right panel ($c = 0.3$) recalls data from the main run for comparison. For all values of c except 0.1, the behavior from the main run is qualitatively replicated.

For c, we chose values of $c = 0.1$, $c = 1.0$, and $c = 3.0$, for an initial value of $c = 0.3$, and the results are shown in figure 12. Similarly to ε, when c is too low (top-left panel), genes do not interact strongly enough and a differentiated phenotype does not evolve depending on the environment, whereas higher values of c (bottom row) show the same evolutionary behavior as the main run (top-right panel).

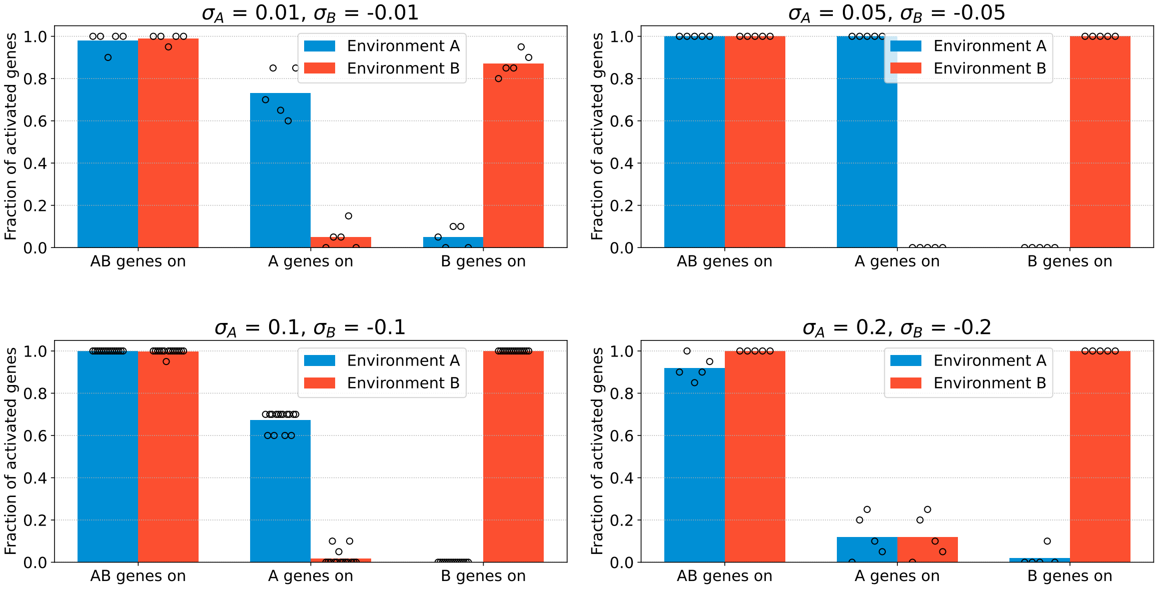

Figure 13: Average fraction of activated genes in each environment at the end of evolution, for more and more distinct environments σ_A and σ_B, from top to bottom and left to right. Every replicate is shown as a dot, and the bottom-left panel ($\sigma_A = 0.1$, $\sigma_B = -0.1$) recalls data from the main run for comparison. For all values except $\sigma_A = 0.2$ and $\sigma_B = -0.2$, the behavior from the main run is qualitatively replicated.

Finally, we also investigated different amplitudes in the difference in supercoiling level between the two environments, by choosing values of $\sigma_A = 0.01$, $\sigma_A = 0.05$ and $\sigma_A = 0.2$, and $\sigma_B = -\sigma_A$ respectively in each case (for an initial value of $\sigma_A = 0.1$ and $\sigma_B = -0.1$). We observe that, when $\sigma_A = 0.2$ (bottom-right panel), the environmental supercoiling constraint is too high and A genes are not activated in environment A by the end of the runs. However, for environments closer to each other than the default (top row), evolution is able to leverage the differences in supercoiling between these environments to evolve differentiated phenotypes, as in the main run (bottom-left panel), showing that our model remains sensitive to small changes in environmental supercoiling.

To conclude, in our model, the gene interaction network is therefore able to respond to different environments and can evolve an efficient regulation of gene expression under a broad range of parameter values, reinforcing the hypothesis that a supercoiling-mediated coupling between gene expression levels could indeed play a functional role in biological organisms.

4 Discussion and Perspectives

DNA supercoiling plays a fundamental role in the regulation of gene transcription in bacteria, and an important part of this role could be mediated by the local variations in supercoiling caused by the transcription-supercoiling coupling. While the influence of the global supercoiling level on gene transcription (Lal et al., 2016; Ma and Wang, 2016; Dorman and Dorman, 2016; Martis B. et al., 2019), the evolutionary importance of supercoiling regulation (Crozat et al., 2005, 2010; Duprey and Groisman, 2021) and the mechanistic details of the transcription-supercoiling coupling (Meyer and Beslon, 2014; El Houdaigui et al., 2019) have all already been studied, no existing work did to our knowledge tackle the question of the possible role of the transcription-supercoiling coupling at both the whole-genome scale and on an evolutionary time scale.

In this work, we have developed a genome-wide model of the influence of DNA supercoiling on gene transcription, incorporating both the global influence of the environment and the local variations due to the transcription-supercoiling coupling on the supercoiling level. We have shown that, in our model, complex interactions implicating several genes emerge from the coupling between supercoiling and transcription. Indeed, *A* genes display an activation pattern that would not be obtainable without the network of interactions that results from the coupling. Thanks to this network, *A* genes are activated in an environment where isolated genes would be inhibited, and inhibited in an environment where isolated genes would be activated. The transcription-supercoiling coupling therefore enables the selective activation or inhibition of specific sets of genes, providing a non-monotonic response to environmental variations through changes in the level of DNA supercoiling. Furthermore, we have shown, using an *in silico* experimental evolution approach, that natural selection can leverage this biophysical mechanism to selectively turn on or off several pools of genes, using only the very simple mutation operator of genomic inversions, that affect the relative positions and orientations of genes on the genome but do not change genome length or basal gene transcription rates, and that this behavior is able to evolve under a wide range of parameter values. This response of gene transcription levels to DNA supercoiling reflects a phenomenon which has been observed *in vivo* in the expression of pathogenicity-related genes in specific environments, such as the normally lethal inside of the macrophage for the mammalian pathogen *S. enterica* (Cameron et al., 2013), or plant tissue for *D. dadantii* (Hérault et al., 2014).

Our model voluntarily stays very simple, only incorporating the most important feature of the transcription-supercoiling coupling, the non-linear interaction between the expression levels of neighboring genes. This simplicity therefore hints at the possible pervasiveness of this regulation mechanism throughout the prokaryotic realm. Nonetheless, in order to go further and represent more accurately the diversity of gene behaviors found in real life, several more dimensions could be integrated to the model. At present, the target for genes in our model is bistability, meaning that genes should end up fully activated or fully inhibited. A biologically more plausible approach would be to relax this restriction and give genes arbitrary expression targets, in order to determine to which extent the transcription-supercoiling coupling is able to finely regulate gene expression. Furthermore, unlike in our model (in which all genes have the same response curve to DNA supercoiling), the genes of biological organisms can show different responses to the supercoiling level. These differences are in part caused by the GC content at the gene promoter (Forquet et al., 2021), and some genes can even respond in the opposite direction to changes in the supercoiling level, that is to say be activated rather than inhibited by DNA relaxation. This behavior is for instance present in the *gyrA* and *gyrB* genes that encode the gyrase subunits in *E. coli* (Peter et al., 2004). Moreover, while our model studies its role in an abstract transcription model, supercoiling intervenes during different parts of transcription initiation, and in transcript elongation and transcription termination as well (Martis B. et al., 2019). Incorporating such precise mechanistic processes into our model could give more accurate information on the link between the position of genes on the genome and their transcription rate. Similarly, increasing the number of genes of individuals in our model to match bacterial gene numbers might provide more fine-grained results, but is computationally intractable in the current implementation of the model. Furthermore, investigating the behaviors of individuals when they are placed successively in different environments, rather than evaluated separately in each environment, would also bring more information on the plasticity of the network of gene interaction levels that emerges from the transcription-supercoiling coupling. Finally, another valuable approach in order to bring this model closer to biology would be to incorporate it into a larger existing framework, such as the Aevol *in silico* experimental evolution platform (Rutten et al., 2019), which models the bacterial genome in much more detail, in order to leverage the power of a well-understood digital organism model.

References

Blattner, F. R. (1997). The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462.

Brinza, L., Calevro, F., and Charles, H. (2013). Genomic analysis of the regulatory elements and links with intrinsic DNA structural properties in the shrunken genome of Buchnera. *BMC Genomics*, 14(1):73.

Cameron, A. D. S., Kröger, C., Quinn, H. J., Scally, I. K., Daly, A. J., Kary, S. C., and Dorman, C. J. (2013). Transmission of an Oxygen Availability Signal at the Salmonella enterica Serovar Typhimurium fis Promoter. *PLoS ONE*, 8(12):e84382.

Crozat, E., Philippe, N., Lenski, R. E., Geiselmann, J., and Schneider, D. (2005). Long-Term Experimental Evolution in Escherichia coli . XII. DNA Topology as a Key Target of Selection. *Genetics*, 169(2):523–532.

Crozat, E., Winkworth, C., Gaffe, J., Hallin, P. F., Riley, M. A., Lenski, R. E., and Schneider, D. (2010). Parallel Genetic and Phenotypic Evolution of DNA Superhelicity in Experimental Populations of Escherichia coli. *Molecular Biology and Evolution*, 27(9):2113–2128.

de la Campa, A. G., Ferrándiz, M. J., Martín-Galiano, A. J., García, M. T., and Tirado-Vélez, J. M. (2017). The Transcriptome of Streptococcus pneumoniae Induced by Local and Global Changes in Supercoiling. *Frontiers in Microbiology*, 8:1447.

Di Cosmo, R. (2020). Archiving and Referencing Source Code with Software Heritage. In Bigatti, A. M., Carette, J., Davenport, J. H., Joswig, M., and de Wolff, T., editors, *Mathematical Software – ICMS 2020*, volume 12097, pages 362–373. Springer International Publishing, Cham.

Dorman, C. J. and Dorman, M. J. (2016). DNA supercoiling is a fundamental regulatory principle in the control of bacterial gene expression. *Biophysical Reviews*, 8(3):209–220.

Duprey, A. and Groisman, E. A. (2021). The regulation of DNA supercoiling across evolution. *Protein Science*, page pro.4171.

El Hanafi, D. and Bossi, L. (2000). Activation and silencing of leu-500 promoter by transcription-induced DNA supercoiling in the Salmonella chromosome: Transcription-dependent modulation of leu-500 promoter in topA mutants. *Molecular Microbiology*, 37(3):583–594.

El Houdaigui, B., Forquet, R., Hindré, T., Schneider, D., Nasser, W., Reverchon, S., and Meyer, S. (2019). Bacterial genome architecture shapes global transcriptional regulation by DNA supercoiling. *Nucleic Acids Research*, 47(11):5648–5657.

Forquet, R., Pineau, M., Nasser, W., Reverchon, S., and Meyer, S. (2021). Role of the Discriminator Sequence in the Supercoiling Sensitivity of Bacterial Promoters. *mSystems*, 6(4).

Grohens, T. (2021). Archive of the EvoTSC repository on the Software Heritage archive. `https://archive.softwareheritage.org/swh:1:rev:6fc36abc1661c295782886647c37ef05ffb9d357`.

Grohens, T., Meyer, S., and Beslon, G. (2021). A Genome-Wide Evolutionary Simulation of the Transcription-Supercoiling Coupling. *Artificial Life*, page 8.

Hérault, E., Reverchon, S., and Nasser, W. (2014). Role of the LysR-type transcriptional regulator PecT and DNA supercoiling in the thermoregulation of *pel* genes, the major virulence factors in *Dickeya dadantii*: *Dickeya dadantii* PecT protein and virulence thermoregulation. *Environmental Microbiology*, 16(3):734–745.

Junier, I. and Rivoire, O. (2016). Conserved Units of Co-Expression in Bacterial Genomes: An Evolutionary Insight into Transcriptional Regulation. *PLOS ONE*, 11(5):e0155740.

Krogh, T. J., Møller-Jensen, J., and Kaleta, C. (2018). Impact of Chromosomal Architecture on the Function and Evolution of Bacterial Genomes. *Frontiers in Microbiology*, 9:2019.

Lal, A., Dhar, A., Trostel, A., Kouzine, F., Seshasayee, A. S. N., and Adhya, S. (2016). Genome scale patterns of supercoiling in a bacterial chromosome. *Nature Communications*, 7(1):11055.

Lenski, R. E., Rose, M. R., Simpson, S. C., and Tadler, S. C. (1991). Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. *The American Naturalist*, 138(6):1315–1341.

Liu, L. F. and Wang, J. C. (1987). Supercoiling of the DNA template during transcription. *Proceedings of the National Academy of Sciences*, 84(20):7024–7027.

Ma, J. and Wang, M. D. (2016). DNA supercoiling during transcription. *Biophysical Reviews*, 8(S1):75–87.

Martis B., S., Forquet, R., Reverchon, S., Nasser, W., and Meyer, S. (2019). DNA Supercoiling: An Ancestral Regulator of Gene Expression in Pathogenic Bacteria? *Computational and Structural Biotechnology Journal*, 17:1047–1055.

Meyer, S. and Beslon, G. (2014). Torsion-Mediated Interaction between Adjacent Genes. *PLoS Computational Biology*, 10(9):e1003785.

Muskhelishvili, G., Forquet, R., Reverchon, S., Meyer, S., and Nasser, W. (2019). Coherent Domains of Transcription Coordinate Gene Expression During Bacterial Growth and Adaptation. *Microorganisms*, 7(12):694.

Peter, B. J., Arsuaga, J., Breier, A. M., Khodursky, A. B., Brown, P. O., and Cozzarelli, N. R. (2004). Genomic transcriptional response to loss of chromosomal supercoiling in Escherichia coli. *Genome Biology*, page 16.

Rhee, K. Y., Opel, M., Ito, E., Hung, S.-p., Arfin, S. M., and Hatfield, G. W. (1999). Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the ilvYC operon of Escherichia coli. *Proceedings of the National Academy of Sciences*, 96(25):14294–14299.

Rutten, J. P., Hogeweg, P., and Beslon, G. (2019). Adapting the engine to the fuel: Mutator populations can reduce the mutational load by reorganizing their genome structure. *BMC Evolutionary Biology*, 19(1):191.

Sobetzko, P. (2016). Transcription-coupled DNA supercoiling dictates the chromosomal arrangement of bacterial genes. *Nucleic Acids Research*, 44(4):1514–1524.

Viñuelas, J., Calevro, F., Remond, D., Bernillon, J., Rahbé, Y., Febvay, G., Fayard, J.-M., and Charles, H. (2007). Conservation of the links between gene transcription and chromosomal organization in the highly reduced genome of Buchnera aphidicola. *BMC Genomics*, 8(1):143.