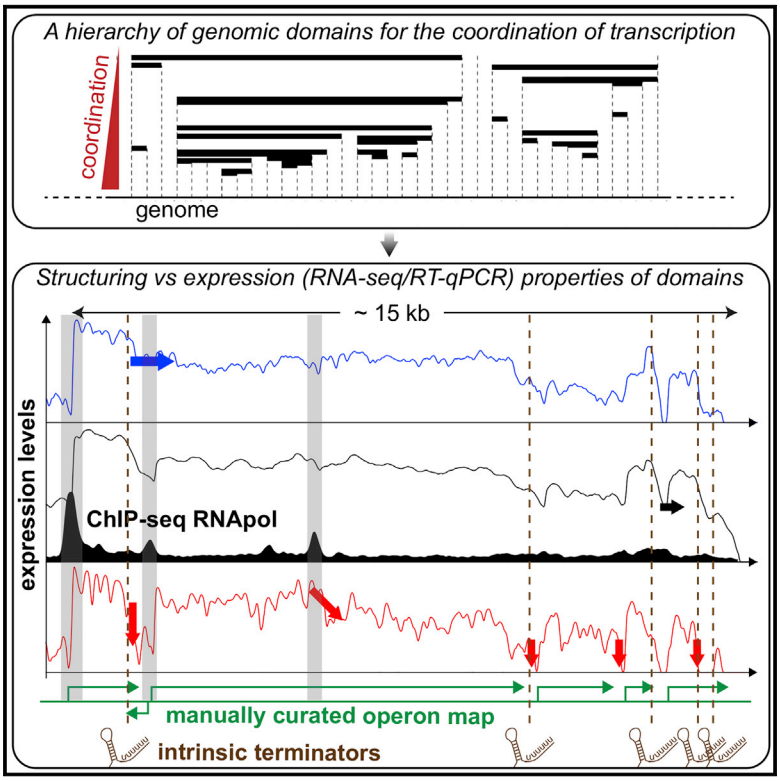
Article

Cell Systems

Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium

Graphical Abstract



Highlights

- Basal coordination of transcription is driven by pervasive transcription
- It is repressed by terminators, long intergenic regions, and stalled RNA polymerases
- Operon-like behaviors may strongly vary from condition to condition

Authors

Ivan Junier, E. Besray Unal, Eva Yus, Verónica Lloréns-Rico, Luis Serrano

Correspondence

ivan.junier@univ-grenoble-alpes.fr (I.J.), luis.serrano@crg.eu (L.S.)

In Brief

By analyzing multiple transcriptomes and structural features of a minimal bacterial genome, Junier et al. show that the basal coordination of transcription in bacteria relies on the capacity of RNA polymerases to transcribe consecutive genes and operons in one go. Stalled RNA polymerases, large intergenic regions, and intrinsic terminators delineate genomic domains of this coordination, for which operon-like behaviors may strongly vary from condition to condition.

Accession Numbers

E-MTAB-3771, E-MTAB-3772, E-MTAB-3773, E-MTAB-3783

Junier et al., 2016, Cell Systems 2, 391–401
June 22, 2016
http://dx.doi.org/10.1016/j.cels.2016.04.015

Cell Systems
Article

Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium

Ivan Junier,[1,6,*] E. Besray Unal,[2,6] Eva Yus,[3,4,6] Verónica Lloréns-Rico,[3,4] and Luis Serrano[3,4,5,*]
[1]CNRS & Université Grenoble Alpes TIMC-IMAG, 38000 Grenoble, France
[2]Institut für Pathologie, Charité - Universitätsmedizin Berlin, 10117 Berlin, Germany
[3]EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Doctor Aiguader 88, Barcelona 08003, Spain
[4]Universitat Pompeu Fabra, 08002 Barcelona, Spain
[5]Institució Catalana de Recerca i Estudis Avançats, Passeig Lluis Companys 23, 08010 Barcelona, Spain
[6]Co-first author
*Correspondence: ivan.junier@univ-grenoble-alpes.fr (I.J.), luis.serrano@crg.eu (L.S.)
http://dx.doi.org/10.1016/j.cels.2016.04.015

SUMMARY

Coordination of transcription in bacteria occurs at supra-operonic scales, but the extent, specificity, and mechanisms of such regulation are poorly understood. Here, we tackle this problem by profiling the transcriptome of the model organism *Mycoplasma pneumoniae* across 115 growth conditions. We identify three qualitatively different levels of co-expression corresponding to distinct relative orientations and intergenic properties of adjacent genes. We reveal that the degree of co-expression between co-directional adjacent operons, and more generally between genes, is tightly related to their capacity to be transcribed en bloc into the same mRNA. We further show that this genome-wide pervasive transcription of adjacent genes and operons is specifically repressed by DNA regions preferentially bound by RNA polymerases, by intrinsic terminators, and by large intergenic distances. Taken together, our findings suggest that the basal coordination of transcription is mediated by the physical entities and mechanical properties of the transcription process itself, and that operon-like behaviors may strongly vary from condition to condition.

INTRODUCTION

Transcriptional regulatory mechanisms can be broadly categorized into two classes. On one hand, response mechanisms can convert environmental cues into specific transcriptional responses. This occurs mostly through the action of dedicated transcription factors (TFs), as in the well-known case of the *lac* operon (Jacob et al., 1960). On the other hand, gene expression is continuously adjusted to adapt to varying environmental conditions, leading to quantitative relationships between the molecular content of cells and their growth rates (Scott and Hwa, 2011). "Physiological factors" (Berthoumieux et al., 2013), such as the concentration of RNA polymerases (RNAPs) (Klumpp and Hwa, 2008), the regulation of RNA degradation (Chen et al., 2015), and topological properties of DNA (Dorman, 1995; Hatfield and Benham, 2002; Travers and Muskhelishvili, 2005), are thus known to continuously affect, at a system level, the transcriptional activity of genes. The specificity, if any, of each of these mechanisms with respect to the set of co-regulated genes, nevertheless remains to be understood.

In bacteria, the coordination of transcription is strongly related to the linear organization of genomes (Képès et al., 2012). At the smallest scale, many co-regulated genes are thus found within operons so that they can be co-transcribed into the same mRNAs. Despite their apparent simplicity, the operons have nevertheless raised important questions, not only about their determination but also about their definition (Okuda et al., 2007; Güell et al., 2011; Mazin et al., 2014) and utility (de Lorenzo and Danchin, 2008; Junier, 2014). For instance, although certain operons are easily recognizable by their functional homogeneity (as, e.g., for the *lac* operon), many of them are composed of genes whose function appears unrelated (de Lorenzo and Danchin, 2008). Soon after the seminal work of Jacob and Monod, studies on operons such as *trp* also revealed the possibility to have specific internal regulation of termination (Yanofsky, 2000): mRNA-based intrinsic terminators may abort transcription midway, whereas competing mRNA secondary structures (anti-terminators) may attenuate this effect (Merino and Yanofsky, 2005; Santangelo and Artsimovitch, 2011). Together with the observation that the majority of operons actually contain alternative transcription start sites (TSSs) (Sharma et al., 2010; Cho et al., 2014), high-throughput data have thus revealed the presence, inside operons, of differential initiation and termination points (Okuda et al., 2007; Güell et al., 2011; Nicolas et al., 2012; Mazin et al., 2014). Yet the impact of these internal elements on the genome-wide coordination of transcription has remained unexplored.

Model systems such as the bacteriophage λ further revealed that operons may be part of larger functional genomic units with, in particular, the possibility of having subsequent operons transcribed in one go (Gottesman et al., 1980). RNAP can indeed override termination, which is called "transcriptional read-through" (TRT). TRT has actually been shown to be frequent and regulated by dedicated proteins (Stülke, 2002; Nudler and Gottesman, 2002), with the so-called ρ factor playing a major

role in many bacteria (Richardson, 2002). Transcriptional co-expression has thus been shown to extend beyond operons (Jeong et al., 2004; Carpentier et al., 2005; Nicolas et al., 2012). Yet the systematic identification of supra-operonic units and of their regulatory mechanisms remains an open problem.

A system-level understanding of transcriptional coordination thus requires abandoning, at least in the first stage, our preconception of the potential units that may come at play. A promising avenue along this line consists in analyzing in detail the genomic properties of proximal genes as a function of their degree of co-expression and to question the structural and regulatory properties that might be associated to the observed patterns (Ma et al., 2013). Here, we perform such analysis in *M. pneumoniae*, a model organism with a reduced genome (~820 kb) that offers ideal properties to address questions about the fundamental mechanisms that govern bacterial cell physiology (Güell et al., 2009; Kühner et al., 2009; Yus et al., 2009). In particular, although *M. pneumoniae* has two sigma factors (Torres-Puig et al., 2015), a tiny TF repertoire (Table S1) and no ρ factor (Himmelreich et al., 1996), it shows genome-wide complex specific regulatory patterns in response to different external perturbations (Güell et al., 2009). This suggests the existence of fundamental mechanisms ensuring coordination of transcription, different from TFs and from the ρ factor.

To test this hypothesis and to identify associated regulatory mechanisms, we analyzed RNA sequencing (RNA-seq) data obtained from *M. pneumoniae* under 115 different conditions (Figure 1A). To this end, we built a co-expression measure particularly well poised to highlight basal co-expression. Using a hierarchical clustering framework that is constrained to respect the 1D organization of the genome, we then reveal the existence of three qualitatively distinct levels of co-expression associated to different organizations of adjacent genes and to different properties of intergenic regions. We next show that the degree of co-expression between co-directional genes and operons is tightly related to the capacity of the RNAP to transcribe them as if they belonged to the same operon. We then reveal that such TRT is both ubiquitous and condition dependent and that it is repressed by DNA-bound RNAPs, strong intrinsic terminators, and large intergenic distances.

RESULTS

Profiling Basal Gene Expression across Conditions

We measured the transcriptional activity of 869 *M. pneumoniae* genes, of which 701 encode proteins (Lluch-Senar et al., 2015), across 141 conditions (282 samples). To this end, *M. pneumoniae* M129 (passage 34, NC_000912 reference genome in the National Center for Biotechnology Information [NCBI]) was grown in modified Hayflick medium and transformed by electroporation (Yus et al., 2009). RNA-seq data were then collected at various stages of the cell growth, after various perturbations and overexpression of different regulators (Table S2). The resulting transcription profiles were generally highly similar, even after shuffling the expression values between the conditions for each gene separately (light gray distribution in Figure 1A), showing that genes have mostly stable expression. To focus specifically on basal co-expression, we discarded "aberrant" transcription profiles using a network approach (Figure 1A). We identified a large set of 227 highly similar samples corresponding to 115 conditions (112 for which two technical replicates are present), the 29 remaining conditions being characterized by a particularly low level of transcription (Table S2).

A Hierarchical Genomic Analysis of Basal Co-expression

To analyze the basal coordination of transcription between all pairs (i,j) of genes, we built a specific measure of transcriptional co-expression, C_{ij}, hereafter referred to as "basal correlation." C_{ij} quantifies the tendency of the expression of the two genes to systematically vary in parallel (Figure S1), here among the 227 working samples. Specifically, it is equal to the difference between the number of pairs of samples for which the two genes co-vary and the number for which they vary in the opposite direction, normalized by the total number of possible pairs of samples. Compared with other correlation measures to which it is related, such as the Pearson correlation, C_{ij} is well suited to highlight the basal coordination of genes (Figure S1). In particular, $C_{ij} \approx 0$ indicates that the expression of the two genes varies as many times in the same direction as in the opposite direction. In contrast, $C_{ij} \approx 1$ indicates that they always vary in the same direction, whereas $C_{ij} \approx -1$ indicates a systematic tendency to vary in the opposite direction.

Figure 1B shows the resulting heatmap of C_{ij}, for which the genes are sorted according to their genomic position, and each pixel indicates the co-expression level between two genes. One can distinguish the presence of specific contiguous clusters of highly co-expressed genes with, on average, a genomic extension that typically extends up to 10 kbp (Figure 1B, bottom). This 10 kbp length scale is larger than the typical length scale of operons, corroborating that the co-expression of proximal genes extend beyond operons. It is actually similar to that found in *E. coli* and *B. subtilis* when performing a similar analysis of co-expression (Jeong et al., 2004; Carpentier et al., 2005; Junier and Rivoire, 2014).

To delineate in a more precise way the relationship between basal coordination of transcription and the established organization of *M. pneumoniae* into operons (see Experimental Procedures for their definition), we analyzed in detail the genomic organization of co-expression. To this end, we developed a hierarchical description of co-expression constrained to respect the linear organization of the genome. Briefly, we built a dendrogram in which pairs of adjacent genes were hierarchically fused on the basis of their co-expression (Figure 1C). Using this dendrogram, we defined domains of the genome, which we call Γ -domains, as the contiguous domains of genes for which all the adjacent genes have a co-expression larger than Γ (Figure 1C).

An analysis of Γ -domains for all possible values of Γ reveals that although these domains may coincide with operons at certain values of Γ, they are generally different. This can be qualitatively appreciated for specific clusters of genes, such as that of the F-ATPase machinery (Figure 1C). More quantitatively, we evaluated the capacity of Γ -domains to predict operons using our most recent manual annotation of operons as the ground truth (Table S1). To this end, we made Γ vary from 1 to -1 and we assessed both the specificity and the sensitivity of predictions. The resulting area under the receiver operating curve (AUC), which summarizes the balance between specificity and sensitivity by a single value, was equal to 0.76 (Figure 1D). Γ -domains

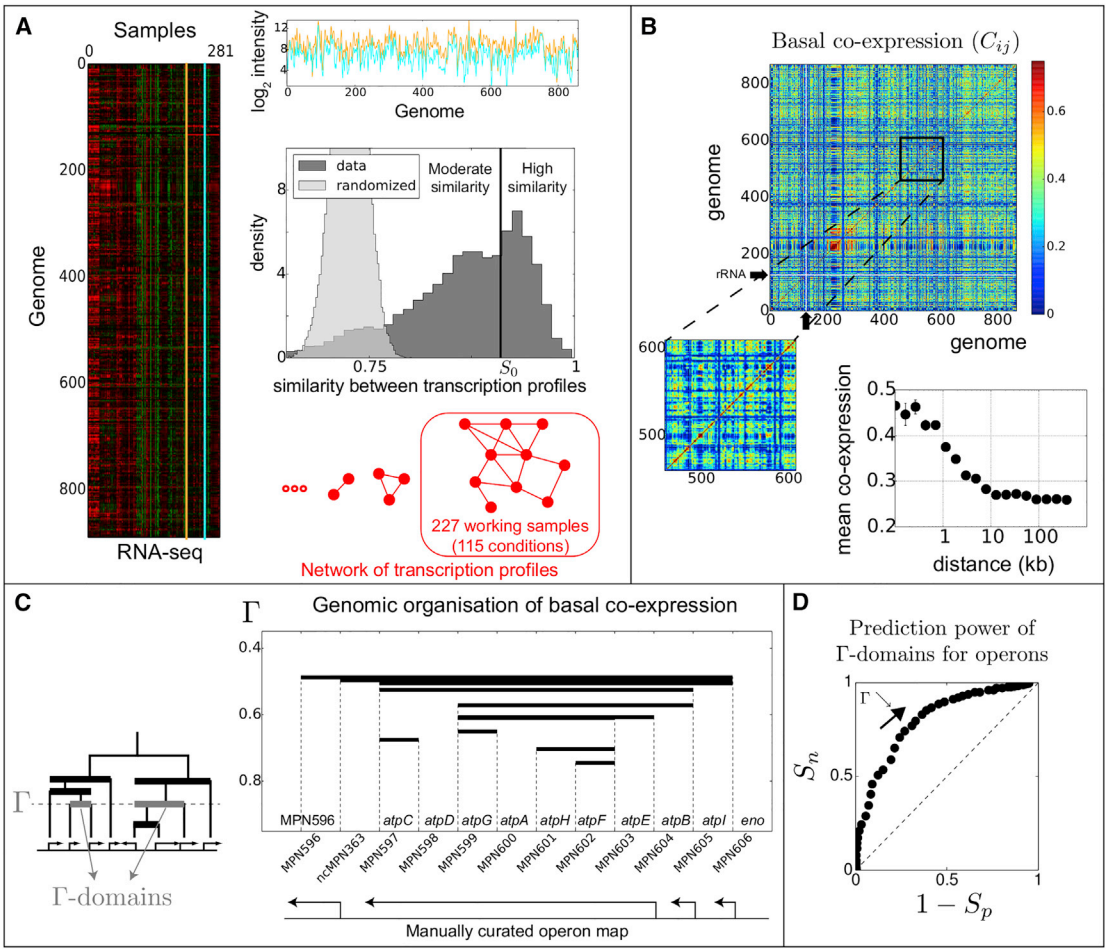


Figure 1. A Hierarchical Genomic Analysis of RNA-Seq Data across More Than 100 Conditions

(A) Given the initial set of RNA-seq samples (2 of which are shown at the top in cyan and orange), we computed all possible pairwise similarities (Pearson coefficient). These were in general high (dark gray distribution), even after shuffling, for each gene separately, the expression values between the conditions (light gray distribution). Given the bimodal shape of the resulting distribution, we defined a threshold (=0.91, vertical black line) above which profiles with larger similarities were connected to form a network, as schematically represented in red. The largest component of the network contained 227 samples (115 conditions), which we used to compute the basal co-expression.

(B) Top: heatmap of the basal co-expression for which the genes are sorted according to their genomic position. The black arrows indicate the position of the rRNAs, which were used to normalize the data and, hence, whose co-expression values were discarded (thin white lines). Bottom left: zoom in. Bottom right: average co-expression level between pairs of genes as a function of their genomic distance.

(C) Using a hierarchical clustering constrained to respect the linear organization of the genome, we built a dendrogram (bottom left) by fusing genes on the basis of their co-expression level. Γ -domains are maximal segments of the genome inside which all pairs of adjacent genes have a co-expression larger than Γ (gray thick lines; all thick lines correspond to a specific Γ -domain but for various values of Γ). They thus correspond to the clades of the dendrogram at the level Γ. As shown on the right panel for the F-ATPase genes, although different, Γ -domains share similarities with operons.

(D) Receiver-operating characteristic analysis to evaluate the predictive power of Γ -domains for operons (AUC = 0.76). S_n and S_p respectively indicate the sensitivity and specificity of the resulting domains.

are thus not perfect predictors (in which case the AUC would have been equal to 1), corroborating the necessity to analyze co-expression properties independently of our knowledge of operons, at least in the first stage.

Three Qualitatively Different Levels of Basal Co-expression

To better understand the hierarchical properties of co-expression, we analyzed the relative orientations of adjacent genes as a function of their co-expression (846 pairs analyzed for which expression values were available for the two genes) (Figure 2A). We also analyzed the tendency of co-directional genes to overlap (Figure 2B), as well as the intergenic distances between the non-overlapping ones (Figure 2C).

Notably, the three properties (relative orientation, overlapping, and distance) suggest a similar three-level organization of co-expression. Specifically, for co-expression > 0.6 (strong co-expression), 227 of 234 pairs of genes are co-directional, with a high proportion (~52%) of overlapping cases ($P \approx 4.10^{-11}$, hypergeometric test). For co-expression between 0.3 and 0.6 (moderate co-expression, 375 pairs), genes significantly tend to be co-directional ($P \approx 7.10^{-7}$) and to overlap ($P \approx 7.10^{-5}$), all

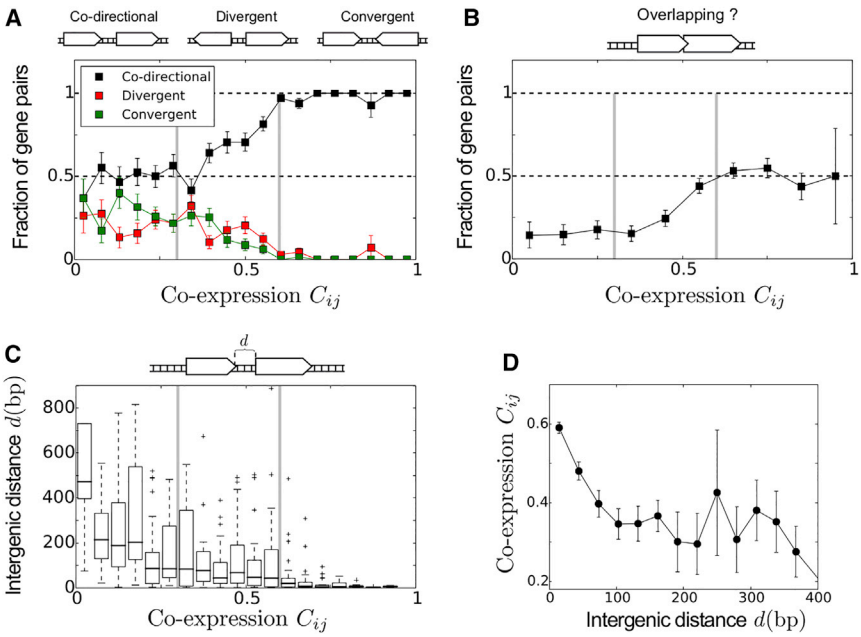


Figure 2. Evidence for the Existence of a Three-Level Organization in the Basal Coordination of Transcription

(A) The distribution of relative orientations of adjacent genes as a function of their co-expression reveals the existence of three qualitatively different levels of co-expression, with threshold occurring at ~0.3 and ~0.6 (vertical gray lines).

(B and C) A similar three-level organization can be distinguished both from the fraction of overlapping pairs (B) and from the distribution of the intergenic distances (d) that separate co-directional genes (C).

(D) Mean co-expression as a function of the distance separating co-directional adjacent genes, revealing a characteristic length scale of 100 bp below which co-expression is all the higher that the distance is small. Error bars correspond to SEM.

the more that co-expression is large. At this level, although intergenic distances between non-overlapping genes are larger than those with strong co-expression, they remain relatively small. Plotting the co-expression level of all pairs of non-overlapping co-directional genes as a function of their intergenic distance actually reveals the existence of a 100 bp length scale below which typical co-expression is larger than 0.3 and above which co-expression is low and statistically insensitive to distances (Figure 2D). Finally, below 0.3 (low co-expression, 237 pairs), there is no enrichment for a specific relative orientation of genes ($P \approx 1$, binomial test of the hypothesis that the probability of co-directionality is equal to 0.5). Moreover, intergenic distances between co-directional genes are large, exceeding 100 bp and reaching typically 400 bp at very low co-expression (Figure 2C).

Performing the same analyses using Pearson correlation led qualitatively to the same findings (Figure S2). The sharp delineation of the three different regimes as well as their correspondence between the different properties (relative orientation, overlapping, and distance) is less clear, though, than those obtained using the basal correlation.

Transcription En Bloc Coordinates Co-directional Genes and Operons

Because a significant number of operons have moderate co-expression levels and because TRT pervades the transcriptomes of bacteria (Wade and Grainger, 2014), we wondered whether TRT could explain the co-expression levels of co-directional genes belonging to distinct operons. We thus investigated the tendency of all adjacent co-directional genes belonging to two different operons to be transcribed as a single transcript (268 pairs analyzed). We examined the variation of expression in their intergenic region as a function of the variation of expression of the downstream gene (Figure 3A). Our rationale was that TRT, if present, should leave a trace on the expression of the intergenic region that precedes the downstream gene. We thus compared the co-expression between the downstream gene and the sense ($5' \rightarrow 3'$) intergenic region (co-expression C_S in Figure 3A) with that between the two genes (co-expression C); as a control, we considered the anti-sense ($3' \rightarrow 5'$) region (Figure 3A, right) (co-expression C_A). To prevent any bias arising from the transcription of the UTRs inside operons, for this analysis, we defined intergenic regions as the sequences that separate the transcription termination site (TTS) of the upstream gene and the TSS of the downstream gene.

For co-expression levels larger than ~0.3, we observed that the degree of co-expression between co-directional adjacent operons was higher when there was co-expression with the sense intergenic region (Figure 3A); in contrast, it did not show any dependency with the anti-sense expression (Figure 3A, right). The same analyses, but considering genes that are separated by more than 100 bp (Figure 3B), or using the Pearson correlation (Figure S2E), led to the same conclusions.

Next, to explicitly demonstrate the role of TRT in basal coordination of transcription, we first studied the efficiency, η, of TRT extending between two co-directional adjacent operons, say, X and Y with X preceding Y (Figure S3A). η was defined as the ratio between the RNA-seq expression levels measured at the TSS of Y and at the stop codon of the last cistron of X (independently whether this was associated to a well-defined terminator). We thus assumed that the RNA-seq level just preceding Y would result from the TRT of X and would be representative of the basal level of Y. According to this model, for which we provide below an experimental validation, the overall expression of Y is thus equal to the sum of its basal level coming from TRT, plus some contribution from its own TSS (Figure S3A).

We analyzed seven pairs of genes with various degrees of correlations and distances: two pairs with strong correlation, including one overlapping case (MPN155a-MPN155); four pairs with moderate correlation and distances larger than 100 bp, including two pairs with an intermediate gene located on the opposite strand of their intergenic region; and one pair with low co-expression. As shown in Figure S3B, η was close to 1 and varied little for pairs with strong correlation, but also

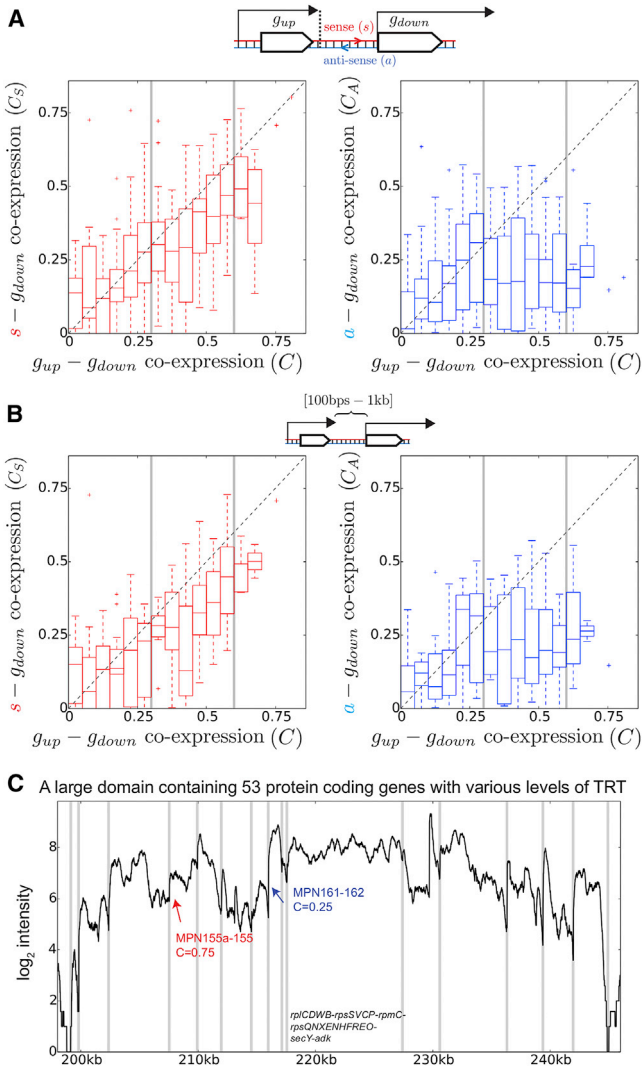


Figure 3. TRT at the Core of the Basal Coordination of Transcription

(A) Left: for pairs of co-directional adjacent genes belonging to different operons, we compare the co-expression, C_S, between the downstream gene and the sense (5′ → 3′) intergenic region with the co-expression, C, between the two genes. Right: as a control, we consider the anti-sense (3′ → 5′) region (co-expression C_A) instead of the sense region. Results show that for $C > 0.3$, C_S and C are strongly correlated, while C_A and C are not. Correlations for $C < 0.3$ might be explained by local concentration effects and the presence of pervasive transcription (Wade and Grainger, 2014).

(B) Same as in (A) but keeping only pairs of operons that are separated by more than 100 bp; distances are measured from the TTS of the upstream operon to the TSS of the downstream operon.

(C) Example of a large domain with a high-level background expression surrounding the ribosomal protein genes and containing 53 genes (15 operons) and for which 46 of the 52 pairs show a significant basal co-expression (> 0.3); for clarity, we indicate the composition of only the largest operon. Although the TSSs of most operons (vertical gray lines) can be distinguished by a steep fold change of the expression, real-time qPCR analysis confirms that TRT occurs between strongly co-expressed operons, as indicated in red for the pair MPN155a-MPN155. In contrast, TRT does not seem to occur at a significant level for low co-expression as in the case indicated in blue (see Figure S3 for details). The RNA-seq profile was obtained at 24 hr (late exponential) of the growth curve.

for the pair MPN160-MPN161 (highest moderate correlation, $C = 0.5$) except during heat shock. For the other pairs, η was both smaller and more variable. In particular, for the pair MPN161-MPN162 (low correlation), we observed a 2-fold variation during cold shock that poorly correlated to the expression variation of the downstream gene.

We then tested the validity of our TRT-based model of transcriptional coordination by confronting predictions of the model (using RNA-seq data) to direct measurements of transcripts using real-time quantitative PCR (qPCR). Specifically, for the aforementioned seven pairs, we measured the level of transcripts extending between the genes during cold shock, heat shock, and exponential growth (control). As shown in Figure S3C, the variations of extended TRT measured by real-time qPCR were qualitatively (quantitatively in most cases) similar to those predicted by the model. Notably, this was true for the cases with an intermediate gene on the opposite strand.

We thus conclude that TRT is ubiquitous and can explain, in principle, many of the significant co-expression levels of co-directional adjacent operons. These results also suggest that large pieces of genomes that extend beyond operons may be transcribed en bloc. An instructive example concerns the ribosome-encoding genes: these genes are surrounded by transcription-related genes and other biological pathways, apparently forming altogether a large domain containing more than 50 genes (corresponding to 15 manually annotated operons) with a high level of background expression (Figure 3C). Although some of this background is expected to result from the strong tendency of these promoters to initiate transcription, as demonstrated by our real-time qPCR analysis (Figure S3) it also results from TRT extending between operons. In this context, it is important to recognize that large domains of coordinated expression, in which several genes and operons may be transcribed in a row, remain compatible with the very presence of operons. This can be seen by the decrease of expression at the end of certain operons or by the presence of steep fold changes at their promoter (vertical gray lines in Figure 3C). The pair MPN155a-MPN155 (strong basal co-expression) provides a good example of this effect as it shows extended TRT between the two corresponding operons (Figures S3B and S3C) but also a sharp TSS at the downstream gene at late exponential phase (in red in Figure 3C) and at stationary phase (Figure S3B).

TRT Variations and Its Regulation

RNA-seq data and real-time qPCR show that TRT may vary not only along the genome but also among conditions (Figure S3). The ten-gene (four-operon) domain containing the heat shock gene (*grpE*) provides an insightful example of such variations (Figure 5C), with the operon containing *grpE* differentiating into two sub-operons during heat shock and distinct operons becoming transcribed as a single operon during cold shock (Figures S3B and S3C). Notably, both our RNA-seq analysis and our real-time qPCR measurements further suggest that TRT is globally enhanced during cold shock (Table S3; Figure S3), in accord with reports in *E. coli* and *B. subtilis* of the anti-terminator role of CspA cold-shock proteins (Bae et al., 2000; Stülke, 2002).

To systematically quantify TRT variations among conditions, we analyzed the behavior of the TTSs internal to pairs of co-directional genes belonging to different operons (233 TTSs

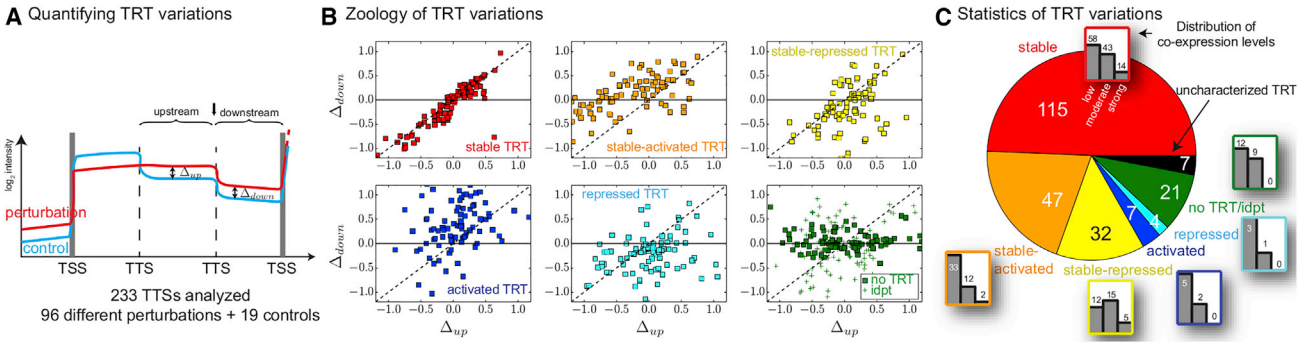


Figure 4. Quantification of TRT Variations

(A) For each pair of adjacent operons, we analyzed at the TTS of the last gene of the upstream operon (black arrow) the behavior of the downstream variation of expression (Δ_{down}) as a function of the upstream variation of expression (Δ_{up}); the corresponding regions were defined by the closest TTS or TSS on each side of the TTS of interest.

(B) We identified six types of TTS, for which an example of each type is shown in every panel; the 96 color points inside every panel correspond to the resulting behavior of the corresponding TTS for the 96 perturbations. To this end, we used two p values, P_1 and P_2, respectively associated to the null hypotheses that Δ_{down} and Δ_{up} are not linearly correlated and that on average, Δ_{down} is equal to Δ_{up}, and considered for significance thresholds a multiple hypothesis correction procedure (Supplemental Experimental Procedures). Stable TRT was then defined by a significant P_1 and a non-significant P_2, stable-activated (repressed) TRT by significant values of both P_1 and P_2 with $\Delta_{down} \geq \Delta_{up}$ ($\Delta_{down} \leq \Delta_{up}$), activated (repressed) TRT by a non-significant P_1 and a significant P_2 with $\Delta_{down} \geq \Delta_{up}$ ($\Delta_{down} \leq \Delta_{up}$), and the set "no TRT or independent TRT" by non-significant values of both P_1 and P_2.

(C) Distribution of the TTS types as identified in (B). Uncharacterized types (in black) correspond to those that did not fit the criteria of the p values. For each type, we show in addition the distribution of basal co-expression (low, moderate, or strong, indicated by the gray bars), revealing that only stable TRTs contribute to strong basal co-expression.

analyzed), independently of whether a well-defined terminator was associated to the TTS. For each TTS, we computed the variation of its downstream expression (Δ_{down}) as a function of the variation of its upstream expression (Δ_{up}) in response to perturbations (96 perturbations tested with respect to 19 controls; Figure 4A). Using this approach, we could identify at least six types of TTSs (Figure 4B). The three first types concern TTSs for which a statistically significant positive correlation exists between the values of Δ_{down} and Δ_{up} that are computed over the different perturbations (see legend of Figure 4 for details of the statistical analyses). They are respectively defined by $\Delta_{down} \simeq \Delta_{up}$ (stable TRT, in red in Figures 4B and 4C), $\Delta_{down} \geq \Delta_{up}$ (stable TRT plus some activation, in orange), and $\Delta_{down} \leq \Delta_{up}$ (stable TRT plus some repression, in yellow). For these TSSs, TRT thus tends to be maintained at a similar level, irrespective of the conditions. Notably, these correspond to ~85% of the total amount of TTSs (~50% if only considering $\Delta_{down} \simeq \Delta_{up}$) and appear to account for all strong co-expression levels (Figure 4C). The two next types correspond to activation only, with a majority of $\Delta_{down} \geq 0$ (in blue), and to repression only, with a majority of $\Delta_{down} \leq 0$ (in cyan), irrespective of the value of Δ_{up}. Together with the TTSs having independent or no apparent statistically significant variations of Δ_{down} (in green), these three last types contribute mainly to low co-expression levels.

Altogether, these results thus corroborate both the ubiquity of TRT and its major role in basal co-expression. The observation of pairs having both low co-expression and stable TRTs also suggest that TRT does not systematically extend to the next operon.

Finally, to apprehend whether TRT is "stochastic" or specifically regulated, we analyzed the behavior of the TTSs upon each perturbation. We found three interesting results (see Table S3 for details). First, in accord with our real-time qPCR experiments (Figure S3), we observed that the variations of TRT (activation or repression) for a given perturbation tend to be the same for all TTSs, suggesting that TRT is specifically regulated at a genome-wide level. Second, we found a larger number of perturbations with TRT activation. These include cold shock, osmotic shock, and novobiocin treatments, whereas low pH and heat stresses tend to repress TRT. Finally, by identifying the TTSs and the corresponding perturbations for which the variation of TRT was extreme (Supplemental Experimental Procedures), we found that conditions for which a large number ($\geq$12) of TTSs had an extreme behavior were strongly enriched in novobiocin (gyrase inhibitor) perturbations ($P \simeq 8 \times 10^{-7}$, hypergeometric test) and strongly depleted in single-gene perturbations ($P \simeq 2 \times 10^{-4}$) (Table S3). Because novobiocin targets topoisomerases and, hence, modify DNA supercoiling, these results suggest that the mechanical properties of DNA and its interaction with RNAPs might play a crucial role in TRT variations (see the following discussion for further details).

The Role of Genome Compactness and Intrinsic Terminators

Our observations of a TRT that depends strongly on conditions, with operons that can be transcribed uniformly, en bloc (super-operons), or differentially (sub-operons), raise at least two fundamental questions: what mechanisms are responsible (1) for promoting an operon-like transcription of adjacent genes and (2) for preventing it?

In answer to the first issue, using co-expression levels as a proxy of transcription en bloc, the results in Figures 2C and 2D suggest that compactness, with a distance between open reading frames smaller than 100 bp, may be required for efficient operon-like co-expression. Notably this length scale corresponds to the typical distance that is usually considered for operon prediction (McClure et al., 2013). We note, nevertheless,

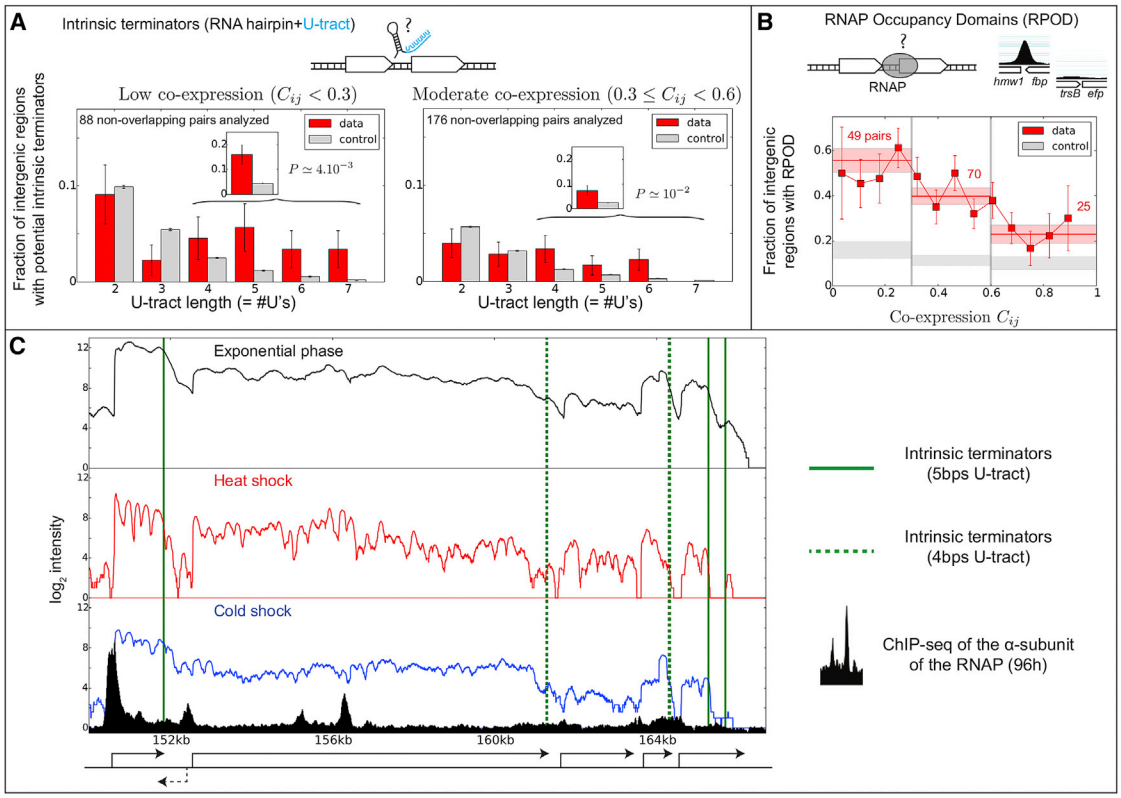


Figure 5. Intergenic Properties of Co-directional Genes Relevant to Delineate Domains of Transcription En Bloc

(A) Fraction of intergenic regions containing a potential intrinsic terminator for low co-expression levels (left) and for moderate co-expression levels (right). Potential terminators were defined as RNA hairpins immediately followed by a U tract. Several lengths (N_U) of the U tracts were analyzed (x axis of the bar plots). As a null model, we considered intergenic regions that were shifted by various amounts of base pairs (gray bars; Supplemental Experimental Procedures), allowing us to evaluate the statistical significance of the results (error bars indicate SEM). Insets show the results by cumulating the cases in which $N_U \geq 4$, revealing an enrichment that is absent with shorter U tracts ($N_U < 4$).

(B) Fraction of intergenic regions containing a RPOD as a function of the basal coordination of transcription. The red bands indicate the SEM computed over the whole region; the red numbers indicate the number of corresponding pairs among the 386 pairs of non-overlapping genes analyzed. The gray bands indicate the same values but for data for which the positions of the intergenic regions were globally shifted by an arbitrary amount of base pairs.

(C) RNA-seq profiles of a large ten-gene (four-operon) domain around the heat shock gene (*grpE*) showing condition-dependent TRT; one additional gene (dashed arrow) is present on the opposite strand. Bottom, in black: ChIP-seq profile of the α-subunit of the RNAP (data obtained at 96 hr), revealing in particular the presence of a large RPOD at the start of the domain. Vertical green lines, positions of strong intrinsic terminators as identified in (A).

that pairs of genes with intergenic regions larger than 100 bp can have a high level of TRT (Figures 3 and S3). Our analysis also shows that compactness alone is not sufficient, because a substantial number (52) of pairs of co-directional genes with low co-expression levels are separated by less than 100 bp (among which 17 pairs concern overlapping genes).

In answer to the second question, let us first mention that although compactness properties call for an important role of distances on the capacity of the RNAP to transcribe multiple genes in a row, co-expression does not depend primarily on distances when these exceed 100 bp (Figure 2D). To better understand the differences between intermediate co-expression levels and low co-expression levels, we thus investigated the possible impact of ρ -independent intrinsic terminators found within mRNA sequences (ρ -dependent termination is absent in *M. pneumonia*). Canonical intrinsic terminators consist of an RNA hairpin followed by a U tract, a combination that is believed to favor the disruption of the mRNA-DNA template hybridization necessary for the RNAP to process transcription (Peters et al., 2011). We thus evaluated the presence, in the intergenic regions of all pairs of co-directional genes, of RNA hairpins that were immediately followed by U tracts of various lengths ($N_U \geq 2$); as a control, we considered intergenic regions that were translated by an arbitrary amount of base pairs, which allowed us handling distance effects of intergenic regions, a longer sequence being more likely to contain an RNA hairpin (Figures S4A and S4B), independently whether the latter plays a functional role.

We found that more than 15% of the gene pairs with low co-expression contained an RNA hairpin with $N_U \geq 4$ (Figure 5A), a proportion that was highly significant with respect to the control ($P \approx 4.10^{-3}$, two-sided t test with unequal variances). Note here that the absence of an enrichment of terminators with shorter U tracts ($N_U < 4$) corroborates previous observations that long U tracts are needed to have efficient termination (Chen et al., 2013). Similar trends were observed for

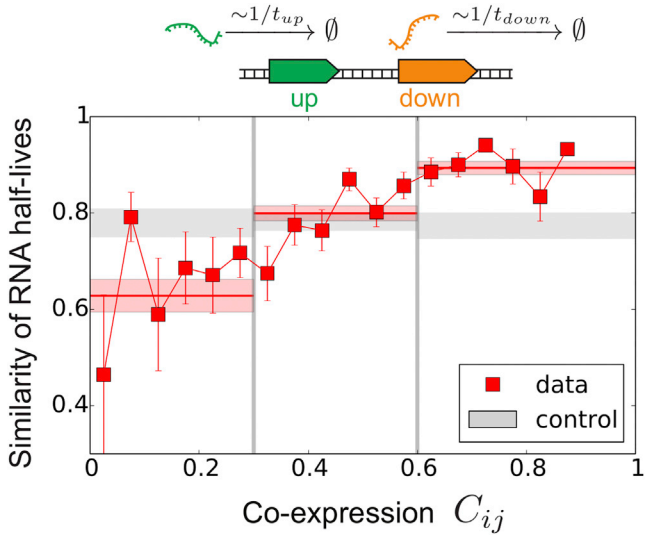


Figure 6. Relative Stability of Transcripts of Adjacent Co-directional Genes

The relative stability is defined as $1 - (|t_{up} - t_{down}| / |t_{up} + t_{down}|)$, with t_{up} and t_{down} the transcript half-lives of the upstream and downstream genes, respectively; this parameter is therefore close to 1 for similar half-lives and close to 0 for very different ones. The red bands indicate the SEM computed over the corresponding region of co-expression. The gray bands indicate the same values but for a random set of pairs of genes.

intermediate co-expression levels, although involving a lower fraction (typically half) of gene pairs, in accord with the fact that at this level, TRT is expected to occur in a larger subset of conditions.

Correlation with RNAP Occupancy Domains and Transcript Half-Lives

According to the above analysis, more than 80% of the co-directional gene pairs with low co-expression do not contain any strong intrinsic terminator in their intergenic region. Non-perfect U tracts or more complex termination signals that are yet to be identified might explain part of this low co-expression. The action of nucleoid-associated proteins (NAPs) such as H-NS in *E. coli* (Singh et al., 2014) could also be invoked. Data from our lab nevertheless show that *M. pneumoniae* contains only one NAP, IHF (gene MPN529), with a low copy number (<100).

To better apprehend the mechanisms related to the repression of TRT, we thus performed chromatin immunoprecipitation sequencing (ChIP-seq) analysis of the α-subunit of the RNAP. Consistent with results obtained in *E. coli* (Mooney et al., 2009), ChIP-seq profiles from cells in the stationary phase revealed the presence of well-defined peaks corresponding to preferentially RNAP occupancy domains (RPOD) (Figures 5B and 5C; Table S4). Notably, we found that the majority of gene pairs with low co-expression contain RPODs in their intergenic regions (Figure 5B). For intermediate co-expression, RPODs are less present but remain over-compared to strong co-expression. Note that in contrast to hairpins, the presence of RPODs does not depend on the intergenic distances (Figure S4C), meaning that larger intergenic distances cannot simply explain these results.

Because the average level of transcription strongly depends on RNA degradation, we eventually compared the half-lives of transcripts between adjacent genes. We found a remarkable correlation between the degree of transcriptional coordination and the similarity of half-lives (Figure 6).

DISCUSSION

From Operonic Transcription to Stochastic Condition-Dependent Transcription En Bloc

Using a correlation measure well poised to quantify basal co-expression and applying it to RNA-seq data obtained in more than 100 different conditions, we have revealed the existence in *M. pneumoniae* of three distinct levels of basal coordination of transcription (strong, moderate, and low), corresponding to three qualitatively different properties for the relative orientations and intergenic regions of adjacent genes. In accord with the major role of operons in the coordination of gene expression, we have found that strong basal co-expression requires adjacent genes to be co-directional. We have also found the existence of a 100 bp length scale, below which an operon-like behavior appears to be quasi-systematic and above which co-expression depends strongly on the sequence and structural properties of the intergenic region. In particular, although pairs of adjacent genes with low basal co-expression do not show any preferential relative orientations, ~70% of the intergenic region of the co-directional pairs either contain a domain preferentially occupied by RNAPs (RPODs) or strong terminators (~55% and ~15% of the cases, respectively).

By focusing specifically on co-directional adjacent genes, we have further revealed that the coordination of transcription is tightly related to the tendency of proximal genes to be transcribed en bloc, even though these genes may have not been categorized to belong to the same operon. Three extreme scenarios can then be considered (Figure S5A), which is in accord with our observation of the three qualitatively distinct co-expression levels. The transcription en bloc may be systematic (i.e., it occurs with probability close to 1 in all conditions), in which case the genes behave as canonical operons (green light in Figure S5A). Or it occurs from time to time, meaning that it can take place in specific conditions and be absent in others. Variations may also occur in a given condition, because transcription may terminate with a certain probability due, for example, to the presence of an intrinsic terminator. In this case, gene expression may present staircase-like patterns (Güell et al., 2009) (orange light). Finally, transcription en bloc can "never" occur, in which case genes must be considered to belong to different transcriptional units (red light).

In accord with the rich zoology of operon-related structures that have been described over the past decade (Okuda et al., 2007; Güell et al., 2009; Cho et al., 2009; Nicolas et al., 2012; Mazin et al., 2014) and with the ubiquitous presence of pervasive transcription (Wade and Grainger, 2014), our findings thus indicate that operon-like behaviors are often stochastic and condition dependent, with frequencies of occurrence that depend on intergenic sequences. In particular, transcriptional initiation may often occur on top of a background level of continuous expression. In this context, we surmise that one of the most fundamental mechanism for the coordination of transcription

relies on a high probability to have specific large domains of genes that are transcribed in a row, independently of the fact that these domains may contain several internal entry points and exit points for the RNAPs (see Figure S5B for a schematic representation of this model). These internal landmarks might then be used by the bacterium to adapt to a wide range of conditions (see, e.g., Figure 5C). They might also contribute to the activation of a given domain (see the following discussion).

Minimal Prescriptions for Generating Specific Domains of Transcriptional Coordination

Our scenario implies, on one hand, the existence of two mechanisms internal to the domains, which are a priori necessary to maintain a proper balance between transcripts. First, there should exist a mechanism that enhances the transcription of upstream genes whenever transcription is initiated within the domain, in order to avoid a gradient of transcripts along the domain (with downstream genes in larger quantity than upstream genes). Although at this stage we have no direct evidence of such phenomenon, this prediction suits the proposal, in bacteria, of a control of gene expression by DNA supercoiling (Dorman 1995; Hatfield and Benham, 2002; Travers and Muskhelishvili, 2005). It is also in accord with our observation of a strong impact of novobiocin (a gyrase inhibitor) treatment on TRT properties (Table S3). The negative supercoiling that is generated upstream of the transcribing RNAPs might indeed enhance the initiation of the upstream genes (Meyer and Beslon, 2014). Considering that these effects can propagate all the way up to the borders of the domain because of the long-range nature of the transmission of supercoiling constraints (Krasilnikov et al., 1999), an internal initiation event should in principle be able to activate the expression of the whole domain (Figure S5B), in particular without the additional action of TFs. Second, produced transcripts should have similar degradation rates, which we confirmed by analyzing RNA half-lives (Figure 6).

Well-defined domains of basal co-expression require, on the other hand, the ability, upstream, to prevent the activation of genes and, downstream, to terminate the transcription process. Supposing that the upstream activation is mainly the result of supercoiling transmission, stalled RNAPs, as suggested by the presence of RPODs (Reppas et al., 2006; Mooney et al., 2009), could act as topological barriers (Higgins, 2014) (see Figure 5C for a suggestive example). Downstream, in addition to the possibility of RPOD roadblocks, strong intrinsic terminators are expected to play an important role in terminating transcription (Figure 5). Other mechanisms can be contemplated, such as anti-sense transcription (Lybecker et al., 2014) or the action of small RNAs, although recent work from our lab shows that the latter have little impact on gene expression (Lloréns-Rico et al., 2016). Here, and more particularly in the absence of the ρ factor and of NAPs, which have been shown to prevent pervasive transcription (Singh et al., 2014), the mechanisms at the core of the basal coordination of transcription in *M. pneumoniae* thus appear to rely solely on the physical entities (RNAP and mRNA) and mechanical properties of the transcription process itself.

Local Concentration Effects

Although a strong terminator can efficiently prevent TRT, it may prevent co-expression only partially. This can be seen for instance by the adjacent genes 5S rRNA (Mpnr03) and MPN095, which is the unique pair showing both strong co-expression $(C = 0.68)$ and the presence of a strong intergenic terminator. Although some specific processing of rRNA might occur, overriding of the terminal signal, as suggested by the high level of co-expression with the sense intergenic region $(C_S = 0.55)$, might explain the strong co-expression. Local concentration effects of RNAPs might also contribute, more particularly because of the high expression level of the 5S rRNA. In such situations, intergenic distances might play a crucial role in the isolation of adjacent genes. Specifically, compared to the 20–30 nm size of the RNAP, the 130 bp that separate the TTS of the 5S rRNA from the TSS of MPN095 correspond to a maximal spatial distance of ~45 nm; 400 bp, the typical distance for pairs of genes with co-expression close to 0 (Figure 2C), correspond to ~135 nm.

Conclusions

Our scenario reckons with the intrinsic stochastic nature of transcriptional initiation, with the capacity of the RNAP to transcribe multiple operons in one go (Santangelo and Artsimovitch, 2011), and with the possible role of supercoiling to transmit regulatory properties, especially in a bacterium that is depleted in TFs (Zhang and Baseman, 2011; Dorman, 2011). It also opens new roads to understand the existence of preferential regions and promoters for the binding of RNAPs (Reppas et al., 2006; Mooney et al., 2009) and suggests that a large part of the specific basal coordination of transcription might rely exclusively on the interplay among RNAP, DNA, and mRNA.

Importantly, our findings appear to hold in a wide range of bacterial species. A similar three-level organization of co-expression, with the same properties of relative orientations and of intergenic distances (including the existence of a ~100 bp length scale), is indeed observed both in *E. coli* and in *B. subtilis*, (Figure S6). Domains of proximal genes that are conserved in phylogenetically distant bacteria have also been shown to correspond, both in *E. coli* and in *B. subtilis*, to domains of highly co-expressed genes and operons where TRT is particularly enhanced (Junier and Rivoire, 2016). Finally, we note that ρ -independent terminators, as well as attenuators of these terminators through, for example, the action of riboswitches, are often conserved among distant bacteria (Vitreschak et al., 2004; Merino and Yanofsky, 2005). Together with the dynamical interplay between DNA and RNAPs, they may thus correspond to ancestral mechanisms upon which the basal functioning of bacteria has been tinkered. In particular, TFs and other types of gene control such as the invertible DNA switches of Bacteroides (Kuwahara et al., 2004) may represent evolutionary solutions dedicated to specific needs related to the lifestyle of each bacterium.

EXPERIMENTAL PROCEDURES

RNA-seq and ChIP-seq Data

RNA isolation was performed using miRNeasy kits from Qiagen, and an in-column DNase treatment was included. RNA was measured using a Nanodrop (Thermo), and integrity was confirmed in a 6000 Nano chip Bioanalyzer (Agilent). We then used the TruSeq Stranded mRNA Sample Prep Kit v2 (Illumina) to obtain a paired-end strand-specific RNA-seq library. See Table S2 for further details of conditions.

ChIP-seq of RNAP (TAP-tagged; see Kühner et al., 2009) was performed as previously described (Yus et al., 2012).

TSSs and Manual Annotation of Operons and Sub-operons

We identified all mRNA TSSs from their associated tssRNAs (Yus et al., 2012). We distinguished productive promoters from short tssRNAs as explained previously (Lloréns-Rico et al., 2015). Regarding 3′ sites, we used strand-specific deep sequencing and tiling array data to define approximately their positions (Güell et al., 2009). We then used these data to refine our previously published operon map (Güell et al., 2009) (updated map in Table S1).

Real-Time qPCR of Regions Encompassing Distinct Operons

Cells were collected in the indicated conditions, and RNA was purified as described above. Retrotranscription and real-time qPCR of ~800 base long regions were done in one step with the GoTaq 1-Step RT-qPCR System (Promega). Oligos (Table S5) were used at 0.15 μM, and 25 ng total RNA was used as a template. mRNA of the stable gene MPN517 was used as control and reference.

Intrinsic Terminators and RPODs

Potential intrinsic terminators were defined as a RNA hairpin immediately followed by a U tract. RNA hairpins were identified as described previously (Mathews et al., 1999).

RPODs were identified by the presence of significant peaks (see Supplemental Experimental Procedures) in the ChIP-seq data of the RNAP α-subunit (gene MPN191) at 6 and 96 hr.

RNA Half-Lives

RNA half-lives were determined using a DNA gyrase inhibitor (novobiocin), which alters the chromosomal supercoiling releasing the RNAP, thus stopping transcription (Dorman, 2011). After novobiocin treatment, RNA was extracted at different time points, and RNA-seq was performed to determine transcript levels. Half-lives were estimated by fitting RNA decays using an exponential function.

ACCESSION NUMBERS

The accession numbers for sequencing data for RNA-seq and ChIP-seq have been deposited in the EMBL-EBI ArrayExpress Archive: E-MTAB-3771, E-MTAB-3772, E-MTAB-3773, and E-MTAB-3783.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.04.015.

AUTHOR CONTRIBUTIONS

I.J., E.Y., and L.S. conceived the analysis. E.Y. performed the experiments. I.J., E.B.U., E.Y., V.L.-R., and L.S. analyzed the data. All authors participated in writing the manuscript.

ACKNOWLEDGMENTS

I.J. is supported by an ATIP-Avenir grant (Centre National de la Recherche Scientifique). E.B.U. was co-funded by Marie Curie Actions. This work was supported by Fundación Marcelino Botin and the Spanish Ministerio de Economía y Competitividad (BIO2007-61762). This project was financed by Instituto de Salud Carlos III and co-financed by Federación Española de Enfermedades Raras under grant agreement PI10/01702 and the European Research Council and European Union's Horizon 2020 research and innovation program under grant agreements 634942 (MycoSynVac) and 670216 (MYCOCHASSIS). The Centre for Genomic Regulation acknowledges the support of the Spanish Ministry of Economy and Competitiveness, "Centro de Excelencia Severo Ochoa 2013-2017," SEV-2012-0208.

Received: July 22, 2015
Revised: January 18, 2016
Accepted: April 21, 2016
Published: May 26, 2016

REFERENCES

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

Berthoumieux, S., de Jong, H., Baptist, G., Pinel, C., Ranquet, C., Ropers, D., and Geiselmann, J. (2013). Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. Mol. Syst. Biol. *9*, 634.

Carpentier, A.-S., Torrésani, B., Grossmann, A., and Hénaut, A. (2005). Decoding the nucleoid organisation of Bacillus subtilis and Escherichia coli through gene expression data. BMC Genomics *6*, 84.

Chen, H., Shiroguchi, K., Ge, H., and Xie, X.S. (2015). Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. Mol. Syst. Biol. *11*, 781.

Chen, Y.-J., Liu, P., Nielsen, A.A., Brophy, J.A., Clancy, K., Peterson, T., and Voigt, C.A. (2013). Characterization of 582 natural and synthetic terminators and quantification of their design constraints. Nat. Methods *10*, 659–664.

Cho, B.-K., Zengler, K., Qiu, Y., Park, Y.S., Knight, E.M., Barrett, C.L., Gao, Y., and Palsson, B.Ø. (2009). The transcription unit architecture of the Escherichia coli genome. Nat. Biotechnol. *27*, 1043–1049.

Cho, B.-K., Kim, D., Knight, E.M., Zengler, K., and Palsson, B.Ø. (2014). Genome-scale reconstruction of the sigma factor network in Escherichia coli: topology and functional states. BMC Biol. *12*, 4.

de Lorenzo, V., and Danchin, A. (2008). Synthetic biology: discovering new worlds and new words. EMBO Rep. *9*, 822–827.

Dorman, C.J. (1995). 1995 Flemming Lecture. DNA topology and the global control of bacterial gene expression: implications for the regulation of virulence gene expression. Microbiology (Reading, England) *141*, 1271–1280.

Dorman, C.J. (2011). Regulation of transcription by DNA supercoiling in Mycoplasma genitalium: global control in the smallest known self-replicating genome. Mol. Microbiol. *81*, 302–304.

Gottesman, M.E., Adhya, S., and Das, A. (1980). Transcription antitermination by bacteriophage lambda N gene product. J. Mol. Biol. *140*, 57–75.

Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., et al. (2009). Transcriptome complexity in a genome-reduced bacterium. Science *326*, 1268–1271.

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

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

Higgins, N.P. (2014). RNA polymerase: chromosome domain boundary maker and regulator of supercoil density. Curr. Opin. Microbiol. *22*, 138–143.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C., and Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res. *24*, 4420–4449.

Jacob, F., Perrin, D., Sánchez, C., and Monod, J. (1960). L'opéron: groupe de gènes à expression coordonnée par un opérateur. CR Acad. Sci. Paris *250*, 1727–1729.

Jeong, K.S., Ahn, J., and Khodursky, A.B. (2004). Spatial patterns of transcriptional activity in the chromosome of Escherichia coli. Genome Biol. *5*, R86.

Junier, I. (2014). Conserved patterns in bacterial genomes: A conundrum physically tailored by evolutionary tinkering. Comput. Biol. Chem. *53*, 125–133.

Junier, I., and Rivoire, O. (2016). Conserved Units of Co-Expression in Bacterial Genomes: An Evolutionary Insight into Transcriptional Regulation. PLOS One. Published online May 19, 2016. http://dx.doi.org/10.1371/journal.pone.0155740.

Képès, F., Jester, B.C., Lepage, T., Rafiei, N., Rosu, B., and Junier, I. (2012). The layout of a bacterial genome. FEBS Lett. *586*, 2043–2048.

Klumpp, S., and Hwa, T. (2008). Growth-rate-dependent partitioning of RNA polymerases in bacteria. Proc. Natl. Acad. Sci. U S A *105*, 20245–20250.

Krasilnikov, A.S., Podtelezhnikov, A., Vologodskii, A., and Mirkin, S.M. (1999). Large-scale effects of transcriptional DNA supercoiling in vivo. J. Mol. Biol. *292*, 1149–1160.

Kühner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., et al. (2009). Proteome organization in a genome-reduced bacterium. Science *326*, 1235–1240.

Kuwahara, T., Yamashita, A., Hirakawa, H., Nakayama, H., Toh, H., Okada, N., Kuhara, S., Hattori, M., Hayashi, T., and Ohnishi, Y. (2004). Genomic analysis of Bacteroides fragilis reveals extensive DNA inversions regulating cell surface adaptation. Proc. Natl. Acad. Sci. U S A *101*, 14919–14924.

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

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., Bork, P., Glass, J.I., Serrano, L., and Lluch-Senar, M. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. Sci. Adv. *2*, e1501363.

Lluch-Senar, M., Delgado, J., Chen, W.-H., Lloréns-Rico, V., O'Reilly, F.J., Wodke, J.A., Unal, E.B., Yus, E., Martínez, S., Nichols, R.J., et al. (2015). Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. Mol. Syst. Biol. *11*, 780.

Lybecker, M., Bilusic, I., and Raghavan, R. (2014). Pervasive transcription: detecting functional RNAs in bacteria. Transcription *5*, e944039.

Ma, Q., Yin, Y., Schell, M.A., Zhang, H., Li, G., and Xu, Y. (2013). Computational analyses of transcriptomic data reveal the dynamic organization of the Escherichia coli chromosome under different conditions. Nucleic Acids Res. *41*, 5594–5603.

Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. *288*, 911–940.

Mazin, P.V., Fisunov, G.Y., Gorbachev, A.Y., Kapitskaya, K.Y., Altukhov, I.A., Semashko, T.A., Alexeev, D.G., and Govorun, V.M. (2014). Transcriptome analysis reveals novel regulatory mechanisms in a genome-reduced bacterium. Nucleic Acids Res. *42*, 13254–13268.

McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C.A., Vanderpool, C.K., and Tjaden, B. (2013). Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res. *41*, e140.

Merino, E., and Yanofsky, C. (2005). Transcription attenuation: a highly conserved regulatory strategy used by bacteria. Trends Genet. *21*, 260–264.

Meyer, S., and Beslon, G. (2014). Torsion-mediated interaction between adjacent genes. PLoS Comput. Biol. *10*, e1003785.

Mooney, R.A., Davis, S.E., Peters, J.M., Rowland, J.L., Ansari, A.Z., and Landick, R. (2009). Regulator trafficking on bacterial transcription units in vivo. Mol. Cell *33*, 97–108.

Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., et al. (2012). Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science *335*, 1103–1106.

Nudler, E., and Gottesman, M.E. (2002). Transcription termination and antitermination in E. coli. Genes Cells *7*, 755–768.

Okuda, S., Kawashima, S., Kobayashi, K., Ogasawara, N., Kanehisa, M., and Goto, S. (2007). Characterization of relationships between transcriptional units and operon structures in Bacillus subtilis and Escherichia coli. BMC Genomics *8*, 48.

Peters, J.M., Vangeloff, A.D., and Landick, R. (2011). Bacterial transcription terminators: the RNA 3′-end chronicles. J. Mol. Biol. *412*, 793–813.

Reppas, N.B., Wade, J.T., Church, G.M., and Struhl, K. (2006). The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting. Mol. Cell *24*, 747–757.

Richardson, J.P. (2002). Rho-dependent termination and ATPases in transcript termination. Biochim. Biophys. Acta *1577*, 251–260.

Santangelo, T.J., and Artsimovitch, I. (2011). Termination and antitermination: RNA polymerase runs a stop sign. Nat. Rev. Microbiol. *9*, 319–329.

Scott, M., and Hwa, T. (2011). Bacterial growth laws and their applications. Curr. Opin. Biotechnol. *22*, 559–565.

Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., et al. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. Nature *464*, 250–255.

Singh, S.S., Singh, N., Bonocora, R.P., Fitzgerald, D.M., Wade, J.T., and Grainger, D.C. (2014). Widespread suppression of intragenic transcription initiation by H-NS. Genes Dev. *28*, 214–219.

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

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

Travers, A., and Muskhelishvili, G. (2005). DNA supercoiling - a global transcriptional regulator for enterobacterial growth? Nat. Rev. Microbiol. *3*, 157–169.

Vitreschak, A.G., Rodionov, D.A., Mironov, A.A., and Gelfand, M.S. (2004). Riboswitches: the oldest mechanism for the regulation of gene expression? Trends Genet. *20*, 44–50.

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

Yanofsky, C. (2000). Transcription attenuation: once viewed as a novel regulatory strategy. J. Bacteriol. *182*, 1–8.

Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.-H., Wodke, J.A., Güell, M., Martínez, S., Bourgeois, R., et al. (2009). Impact of genome reduction on bacterial metabolism and its regulation. Science *326*, 1263–1268.

Yus, E., Güell, M., Vivancos, A.P., Chen, W.-H., Lluch-Senar, M., Delgado, J., Gavin, A.C., Bork, P., and Serrano, L. (2012). Transcription start site associated RNAs in bacteria. Mol. Syst. Biol. *8*, 585.

Zhang, W., and Baseman, J.B. (2011). Transcriptional response of Mycoplasma genitalium to osmotic stress. Microbiology *157*, 548–556.

Cell Systems, Volume *2*

Supplemental Information

Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium

Ivan Junier, E. Besray Unal, Eva Yus, Verónica Lloréns-Rico, and Luis Serrano

Table of contents

Supplementary information for the methods
- Bacterial strains, culture conditions, RNA-sequencing data and chromatin immunoprecipitation
- TSS sites and manual annotation of operons and sub-operons
- Identification of conditions with similar transcriptomes to analyze basal transcriptional co-expression
- Matrix of basal co-expression
- Genomic co-expression dendrograms and corresponding gene domains
- Properties of adjacent genes and of their intergenic regions
- Co-expression of adjacent genes: highlighting the role of transcriptional read-through
- Transcriptional read-through analysis at the TTSs
- Real time quantitative PCR
- Details on ChIP-seq analysis
- Details on RNA half-life determination

Legends of supplementary tables (5)
Legends of supplementary figures (6)
References

Supplementary information for the methods

Bacterial strains, culture conditions, RNA-sequencing data and chromatin immunoprecipitation

Mycoplasma pneumoniae M129 (passage 34) was grown in modified Hayflick medium and transformed by electroporation as previously described (Yus et al., 2009).

Cells in exponential (6 hours post-inoculation) or stationary phases (96 hours) were collected after various perturbations or by over-expressing different regulators in Qiazol (see Table S2). RNA isolation was performed following the manufacturers' instructions (miRNeasy kit from Qiagen), and an in-column DNase treatment was included. RNA was measured using a Nanodrop (Thermo) and integrity was confirmed in a 6000 Nano chip Bioanalyzer (Agilent). In order to obtain a paired-end strand–specific RNA-seq library, the TruSeq Stranded mRNA Sample Prep Kit v2 (Illumina) was employed according to the manufacturer's instructions. Briefly, 100 ng of total RNA was fragmented to approximately 300 bases. cDNA was synthesized using reverse transcriptase (SuperScript II, Invitrogen) and random primers. The second strand of the cDNA incorporated dUTP in place of dTTP. Double-stranded DNA was further used for library preparation. dsDNA was subjected to A-tailing and ligation of the barcoded Truseq adapters. All purification steps were performed using AMPure XP beads. Library amplification was performed by PCR using the primer cocktail supplied in the kit. Final libraries were analyzed using Agilent DNA 1000 chip to estimate the quantity and check size distribution, and were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems) prior to amplification with Illumina's cBot. Libraries were sequenced paired-end, 100 nts (2x50) on Illumina HiSeq 2500, in pools of 6 samples.

Chromatin immunoprecipiation (ChIP-seq) of RNAP (TAP-tagged, see (Kühner et al., 2009)) was performed as previously described (Yus et al., 2012). In this case the libraries were single-end and pooled in blocks of 12.

Resulting raw reads were mapped to the *M. pneumoniae* reference genome (NC_000912, NCBI) with MAQ software (default parameters, and one mismatch allowed) (Li et al., 2008). Counts per gene were extracted from the pileups using our genome annotation. In the case of RNA-seq, the expression per gene was extracted and then normalized, first by the length of the gene, second by the corresponding counts obtained for rRNAs (16S gene). Sequencing data have been deposited in the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/sra, accession numbers: E-MTAB-3771, E-MTAB-3772, E-MTAB-3773, E-MTAB-3783). Altogether, we generated 282 samples corresponding to 141 different conditions.

TSS sites and manual annotation of operons and sub-operons

Taking advantage that small non-coding RNAs (tssRNAs) are ubiquitously associated to transcription start sites (TSSs) we could identify all mRNA TSSs (Yus et al, 2012). Next, we used a previously described method (Lloréns-Rico et al., NAR 2015) to identify productive promoters at TSSs of mRNAs and non-coding RNAs and distinguish them from short tssRNAs (Yus et al, 2012). Regarding 3' sites, we used strand specific deep sequencing and tiling array data to define approximately their positions (Güell et al., 2009). The operon and sub-operon annotation that was published previously (Güell et al., 2009) has thus been refined with the last genome annotation published by our group (Wodke et al., 2015) (Table S1). Specifically, operons were defined manually by looking at microarrays, tiling arrays and deep sequencing of *M. pneumoniae* transcriptomes at 6 and 96h of the growth curve (Yus et al, 2012). They were defined as regions with a tssRNA, no internal tssRNA associated to a RNA level increase (<0.8 log2) and where RNA levels did not drop significantly between two consecutive genes (<0.8 log2). Sub-operons were defined as regions of an operon where different expression levels were found for consecutive genes, and/or having an internal tssRNA with a promoter associated to a gene. To this end, we used both deep sequencing and tiling array experiments (Güell et al., 2009; Yus et al, 2012).

Identification of conditions with similar transcriptomes to analyze basal transcriptional co-expression

In order to analyze basal transcriptional co-expression, we extracted a set of 227 similar samples out of the 282 initial ones, altogether corresponding to 115 different conditions. To this end, we first computed all pairwise similarities (Pearson coefficient) among the 282 initial transcriptomes. As a result, we obtained a bimodal distribution of similarities (Figure 1A), allowing us to define a threshold ($S_0 = 0.91$, vertical black line on Figure 1A) to separate pairs of conditions with high similarities ($S \geq S_{Sim}$) from pairs of conditions with moderate similarities ($S < S_0$). Using these similarities, we next built a network by connecting any two pairs of profiles with high similarity. The resulting network of profiles was composed of 24 disconnected components (schematically represented in Figure 1A), with the largest one containing 227 samples, which we used to analyze basal transcriptional co-expression.

Matrix of basal co-expression

To analyze basal co-expression between genes, we first defined the start and end of genes, which were given by the translation start codon and Stop codon for protein-coding genes, and by the TSS and transcription termination site (TTS) for small RNAs – for most analyses this definition prevented possible bias coming from a manual annotation of TSS; note also that in our analysis both protein-coding sequences and small RNAs are considered as genes. We next sorted these genes according to their middle position ((start+end)/2) and removed those that were included in other genes (like many small RNAs). From our initial set of 1083 genes, we eventually analyzed the expression of 869 genes, of which 701 encode proteins (Lluch Senar et al., 2015).

The 227 RNA-seq expression profiles were used to define a transcriptional activity a_{si} for every gene i in every sample s, by averaging the values associated with the corresponding RNA-seq reads. Using these activities, we defined for any pair (i,j) of genes a basal correlation $C_{ij} = \frac{\sum_{s,s'} sgn\left(a_{si}-a_{s'i}\right)\cdot sgn\left(a_{sj}-a_{s'j}\right)}{2M^2}$, where the sum runs over all pairs (s,s') of conditions ($M = 227$) and where the sign function, $sgn(x)$, is equal to 1 if $x \geq 0$ and -1 if $x < 0$. This basal correlation has a simple meaning: it corresponds exactly to the fraction of pairs of conditions for which the genes i and j vary in the same direction ($sgn(x_{si} - x_{s'i}) \cdot sgn\left(x_{sj} - x_{s'j}\right) = 1$) minus the fraction of condition pairs for which the genes vary in opposite directions ($sgn(x_{si} - x_{s'i}) \cdot sgn\left(x_{sj} - x_{s'j}\right) = -1$), independently of the amplitude of the variations (Figure S1). It is for instance close to 0 when genes are uncorrelated (same amount of pairs for the two sets). Notably, while C_{ij} might be regarded as a simplified form of a Pearson correlation to which it is tightly related, compared to the latter but also to other correlation measures that are more robust to outliers than Pearson correlation (Song et al., 2012), C_{ij} is more sensitive to basal co-expression, that is, to the systematic tendency for genes to have their expression vary in the same direction (Figure S1).

Genomic co-expression dendrograms and corresponding gene domains

From the basal co-expression matrix, we generated a dendrogram constrained to respect the 1D organization of the genome (Figure 1C). In this dendrogram, only pairs of genes that are adjacent along the chromosome can be connected, which was implemented by hierarchically fusing genes on the basis of their basal co-expression.

We defined the Γ-domains of the dendrogram as the resulting clades obtained by cutting the dendrogram at depth Γ, with Γ that can take all possible values in [-1,1]. Γ-domains thus correspond to the maximal segments of the genome inside which all pairs of adjacent genes have a basal co-expression larger than Γ (Figure 1C).

Properties of adjacent genes and of their intergenic regions

For every pair of adjacent genes along the DNA, we computed several properties as a function of their level of basal co-expression, including:

- *their relative orientation*, with co-directional genes aligned along the same strand, and genes belonging to opposite strands that can be either divergent or convergent, depending on whether their start-to-start distance is smaller or, respectively, larger than their end-to-end distance (Figure 2A).
- *whether genes overlap*, which occurs when genes share a common piece of DNA.
- *the distance between co-directional genes* (in base-pairs (bps)), which is given by the distance that separates the Stop (or TTS for the non-coding RNAs) of the upstream gene from the translation start codon (or TSS for the non-coding RNAs) of the downstream gene.
- *the presence of intrinsic terminators in the intergenic regions*. Potential intrinsic terminators were defined as a RNA hairpin immediately followed by a U-tract with at least 2 U's. RNA hairpins were identified as previously described (Mathews et al., 1999). To evaluate the statistical significance of the results, we considered a null model where the positions of the intergenic regions were translated by a certain amount of bps (t_{bp}). To provide statistical power, we performed this procedure 200 times, with t_{bp} taking equally separated values from 10 kbps to 400 kbps.
- *the presence of RNAP occupancy domains (RPOD) in the intergenic regions*. RPODs were identified by the presence of significant peaks in ChIP-seq data obtained for the α-subunit of the RNAP (gene MPN191) at 6 and 96h (Table S4) (see below for further details on ChIP-seq analyses performed in this work**)** Peaks were identified using a custom R implementation of the Matlab function "findpeaks". To evaluate the statistical significance of the results, we used the same procedure as that for the intrinsic terminators.
- *the relative stability of transcripts*, defined as $1 - \frac{|t_{up}-t_{down}|}{|t_{up}+t_{down}|}$, with t_{up} and t_{down} the transcript half-lives of the upstream and downstream genes, respectively; this parameter is close to 1 for similar half-lives and close to 0 for very different ones. RNA half-lives were determined experimentally using a DNA gyrase inhibitor

(Novobiocin), which alters the chromosomal supercoiling releasing the RNAP, thus stopping transcription (Yus et al., manuscript in preparation; see also Dorman, 2011). After treatment with Novobiocin, RNA was extracted at different time points and whole transcriptome sequencing by RNA-seq was performed to determine transcript levels. RNA decay was fitted to an exponential decay according to the following equation:$[RNA] = [RNA]_0 \cdot e^{-kt}$, from which the decay rate k was obtained. Half-lives were then calculated as $t_{1/2} = \log(2)/k$. See below for further details.

Co-expression of adjacent genes: highlighting the role of transcriptional read-through

To apprehend the mechanisms underlying the co-expression between adjacent co-directional genes, we compared the co-expression between these genes with that between the downstream gene and the sense intergenic region (Figure 3A). To this end, we analyzed the behavior of adjacent genes belonging to different operons and considered the intergenic region located between the TTS of the upstream operon and the TSS of the downstream operon. To further discard any effect coming from uncertainties in the identification of the TTS of the upstream gene, we considered only the second half of the intergenic region to measure the intergenic expression (similar results were obtained using the whole intergenic regions).

Transcriptional read-through analysis at the TTSs

To quantify the TRT occurring at the TTSs inside the pairs of adjacent co-directional genes (382 TTSs analyzed), first we defined the regions upstream and downstream each TTS. Regions upstream the TTS span from the TTS until the previous junction (either TSS or TTS) located in the same strand in the genome. Regions downstream the TTS span from the TTS until the next junction located in the same strand in the genome (Figure 4A). When these regions were longer than 1000 bases, they were trimmed to 1000 bases. Once the upstream and downstream regions were defined, expression was calculated for each of these regions at each of the 115 analyzed conditions. Expression was determined as the average number of read counts per base (in $\log_2$) across the entire region: $exp = \frac{1}{N}\sum_{n=1}^{N} log_2(RC_n)$, where $n = 1, 2, \dots, N$ bases in each region and RC_n represents the number of read counts at base n. After calculating all the expression values, we compared each condition with its corresponding control, to calculate Δ_{up} and Δ_{down} for each of the 382 TTS for 96 different perturbations. These represent the difference of expression between a given condition and its control – we thus used 19 (=115-96) conditions as a control. To assess the significance of the changes, we performed for each case a Student's t-test comparing the control and the perturbation. We considered as "extreme variations" those changes in which the t-test yielded a p-value smaller than 0.05, and the absolute difference of expression was larger than 2 standard deviations of the distribution of changes of the entire population.

Finally, to distinguish TTS types, for each and every TTS, given all its values of Δ_{down} and Δ_{up}, we computed two p-values, P_1 and P_2, respectively associated to the null hypotheses "Δ_{down} and Δ_{up} are not linearly (and positively) correlated" and "in average, Δ_{down} is equal to Δ_{up}". To this end, we used a Benjamini–Hochberg procedure to build two corresponding p-value thresholds, π_1^* and π_2^*, such that to work with a false discovery rate FDR=0.05. In this context, we considered P_1 and P_2 as cases showing statistical significance if $P_1 < \pi_1^*$ and $P_2 < \pi_2^*$, respectively.

Real time quantitative PCR and list of oligos

In order to demonstrate the presence of TRT between pairs of adjacent co-directional genes, real-time PCR of cDNA of ca. 800 bases regions encompassing the intergenic region and overlapping with the ORFs was performed. Briefly, cells were collected in the indicated conditions (exponential phase, heat shock at 43C for 30 min or cold shock at 15C for 15 min) and RNA was purified as described before. Retrotranscription and real-time quantitative PCR were done in one step (RT-qPCR) with the GoTaq® 1-Step RT-qPCR System (Promega) following the manufacturer's instructions. Two 10 µl reactions of two biological data were prepared. Oligos (Table S5) were used at 0.15 µM and 25 ng total RNA was used as template. An mRNA that usually doesn't show much variation (namely MPN517) was used as control and reference.

Details on ChIP-seq analysis

After the read mapping procedure, two curves were obtained corresponding to the plus and minus strand pileups of the *M. pneumoniae* chromosome.

For each of these curves, the signal was normalized with the signal of a control experiment (a ChIP-seq experiment in which the immunoprecipitation was performed only with the secondary antibody), so that the sample and the control experiments have equal baselines. Then, the signal from the control experiment was subtracted from the

RNAP signal. After the subtraction, noise was modeled as following a Gaussian distribution, and a threshold was set to reject all the values whose probability of being noise was greater than 1e-6. To check whether noise followed a Gaussian distribution, we performed ChIP-seq and control experiments of a wild-type strain of *M. pneumoniae*, without overexpression of any DNA-binding protein. We observed that our model held true and that after subtracting the control signal the values followed a Gaussian distribution. Then, a smoothing algorithm was applied to the processed data, and peaks were called separately in each of the strand curves. The peak calling was performed by using the "findpeaks" function with the following parameters, chosen to maximize the performance of the function in our datasets:

- Slope threshold = 0.0001 (minimum slope to consider in a peak)
- Amplitude threshold = 5 (minimum peak width)
- Smoothing width = 15 (number of points to consider to smooth the curve)
- Peak group =15 (number of data points to take to fit a peak)

After the peak calling in both strands, a further filtering step was applied. In ChIP-seq, it is expected to find the same peaks in both strands, but with the peak in the minus strand displaced to the right with respect to the peak in the minus strand. This is due to the fact that the read length in the sequencing procedure is usually smaller than the fragment length after sample sonication, and only the ends of the fragment are thus sequenced. Therefore, we associated each peak found in the plus strand to its corresponding peak in the minus strand, provided that the distance between the center of both peaks was smaller than 300bps. The peak position was then relocated to the mid-point between the associated partners. The mean inter-peak distance of all the matched peaks was calculated, as it is expected to be similar for all the peaks within the same experiment. For single peaks without associated partners in the opposite strand, the peak position was relocated according to this mean inter-peak distance. Finally, a score defining how well a pair of peaks matches this distance was given to each peak in the experiment. Single peaks were not assigned any score.

Details on RNA half-life determination

Transcription in bacterial cells can be modeled in a simple manner as the continuous balance between transcription production and degradation, according to the following equation: $\frac{d[RNA]}{dt} = \alpha - k[RNA]$, where α and k are the production and degradation rates, respectively. A straightforward manner of determining the degradation rate k is to make the production (α) equal to zero and then solve the differential equation to obtain that $[RNA] = [RNA]_0 \cdot e^{-kt}$. In order to experimentally make the transcription rate α equal to zero, we used a DNA gyrase inhibitor, Novobiocin. When applied to *M. pneumoniae* cells, it alters the chromosomal supercoiling, releasing the RNAP and thus stopping transcription. We confirmed that the RNAP was released off the chromosome by performing a ChIP-seq experiment of the RNAP after addition of the drug. Therefore, we treated *M. pneumoniae* cells in exponential growth phase with Novobiocin and extracted total RNA at different time points after the addition: 0 (as a control, without the drug), 2, 4, 6, 8, 10 and 15 minutes, with two biological replicates for each point. Whole transcriptome sequencing was performed and transcript levels were calculated for each of the samples. Transcript levels were transformed to copy numbers per cell using an experimentally determined adjust function (Maier et al., 2011, see below) and then to RNA concentrations, considering an approximate volume of 0.055μm^3 for *M. pneumoniae* (Hasselbring et al., 2006). After this transformation, the time-course values were adjusted to an exponential decay according to the formula $[RNA] = [RNA]_0 \cdot e^{-kt}$, and the degradation rates were determined for each gene. Given the degradation rates, we determined the half-live of all genes in *M. pneumoniae* as $t_{1/2}$=log(2)/k.

To compute copy numbers, short reads from each of the RNA-seq experiments were mapped to the reference genome of *M. pneumoniae* using MAQ (Li et al., 2008). Only one mismatch with the reference sequence was allowed. Reads mapping to more than one genomic position were discarded. After the mapping, a custom R script was used to calculate gene expression in CPKM (Counts Per Kilobase per Million reads mapped), a measure that is similar to RPKM. In this context, the experimental relationship between the copy number and the CPKM is the following: $CopyNumber = 2^{0.903 \cdot log_2(CPKM) - 7.9789}$. This equation was obtained after fitting RNA-seq data to the experimental values previously obtained for microarray data (Maier et al., 2011).

Legends of supplementary tables

Table S1. Related to the paragraph "TSS sites and manual annotation of operons and sub-operons" in Experimental procedures. Sheet 1: List of known or putative transcriptional regulators in *M. pneumoniae*. The last column indicates the name of the strains in which the TF is perturbed (see Table S2). Sheet 2: Manual operon and sub-operon annotation of the *M. pneumoniae* genome. The table indicates the following information for each of the manually annotated transcriptional units (operons and sub-operons): operon number, sub-operon ID, genes belonging to each sub-operon, TSS of the sub-operon, TTS of the sub-operon and strand.

Table S2. Related to the paragraph "RNA-seq and ChIP-seq data" in Experimental procedures. Sheet 1: list of RNA-seq experiments used in this work. For each sample, we indicate the strain (wt, M129 or mutant), transgene (indicates the gene that was overexpressed or mutated), timeOfGrowth_experimentPerformedAt in h (time of growth after inoculum), medium used, treatment (type of drug/perturbation), perturbant (drug, condition…), finalConcentration_perturbant (working dilution), durationOfPerturbation in min, Filtered? (in case it was left out of the analysis, see main Materials and Methods). Sheet 2: list of samples discarded for the co-expression analysis. Sheet 3: Corresponding list of conditions effectively used in the analysis of basal co-expression and of TRT variations. The last column indicates whether the condition was analyzed for TRT variation or if it corresponded to a control. The red names indicate that a single gene was perturbed, in contrast to more global perturbation (various stress shocks, Novobiocin treatments, etc…). The yellow boxes indicate that the perturbed gene is a putative TF (see Table S1).

Table S3. Related to Figure 4. Sheet 1: Leftmost list: conditions leading to an overall repression of TRT, that is, showing a tendency for having $\Delta_{down} \leq \Delta_{up}$. The average value of $\Delta_{down} - \Delta_{up}$ (third column) is computed over all the TTSs. The list is sorted according to the p-values of the bias of the distribution of $\Delta_{down} - \Delta_{up}$ (second column, one sample t-test value). The horizontal dashed and full lines respectively indicate the values where the false discovery rate (FDR) is equal to 0.005 and 0.05 (Benjamini–Hochberg procedure). Rightmost list: same thing but for conditions leading to an overall activation of TRT, that is, showing a tendency for having $\Delta_{down} \geq \Delta_{up}$. Sheet 2: Leftmost list: perturbations for which no pair of adjacent genes shows an extreme variation of TRT. Rightmost list: perturbations for which at least 12 pairs of adjacent genes show an extreme variation of TRT. The color codes are those of Table S2.

Table S4. Related to Figure 5. ChIP-seq peaks associated to RNAP (see Methods and Materials and Supp. Methods text for experimental procedures and for the identification of peaks). For each of the peaks, the following information is displayed: peak position (in bps); peak height (in arbitrary units); peak width (in bps covered); peak score, based on the confidence in the intra-peak distance (see Supplementary Methods); associated TSS(s), if any, otherwise is "NONE"; associated TSS strand(s), if any, otherwise is "NONE"; and time point of the corresponding experiment (6h or 96h).

Table S5. Related to Figures 4 and 5. Oligos used for the RT-qPCR.

Legends of supplementary figures

Figure S1. Related to Figure 1. Comparison of the Pearson correlation, the biweight midcorrelation (bicor) and our basal correlation. *Left column* – As indicated in the upper panel, the Pearson and bicor correlations (Song et al., 2012) are based on an analysis of the variations of the input signal with respect to a global property of the signal as schematically indicated by the arrows and the horizontal dashed value, the latter respectively representing the average (Pearson) and the median (bicor) values. Note that by construction bicor is more robust to the presence of outliers than Pearson as it provides an analysis of the variations with respect to the median value instead of the average value of the signal (Song et al., 2012). In contrast, our basal correlation is computed by considering *equally* the variations between all possible pairs of conditions. This is indicated by the +/- 1 value for the various variations obtained with different amplitudes. *Right column* – Four different stylized datasets showing the robustness of our correlation in the identification of basal co-expression. From top to bottom: a) for two signals that differ by a small random noise, the three correlations are close to 1; b) for uncorrelated signals, they are close to 0; c) in this dataset, the two signals are uncorrelated, except for conditions 50 to 59 where there is a global shift of the signal; this dataset thus corresponds to a globally low basal co-expression with, nevertheless, a similar shift for the conditions 50 to 59. Notably, the Pearson and bicor correlations indicate a significant co-expression, whereas the Basal co-expression does not; d) here the two signals are perfectly synchronized, except for conditions 50 to 59 where there is an overall opposite shift; this dataset thus corresponds to a globally strong basal co-expression with, nevertheless, an opposite shift for the conditions 50 to 59. Notably, the Pearson and bicor correlations indicate a negative co-expression value, whereas the Basal co-expression indicates a significant positive value.

Figure S2. Related to Figure 2. A,B,C,D: Same as Figure 2 but using Pearson correlation. E: same as Figure 3A but using Pearson correlation.

Figure S3. Related to Figure 3 and 4. A) Simple model of co-regulation of adjacent operons involving TRT with efficiency η. In this model, we suppose that for N_X transcripts of the upstream operon (X), the ηN_X transcripts obtained after TRT extend to the downstream operon (Y). As a consequence, the expression level $[Y]$ is equal to the sum of the expression level resulting from the TSS of Y (denoted by δ) plus the contribution of the read-through that can be measured just before the TSS of Y (denoted by $[R]$). B) Estimation of η for 7 different pairs of genes (the 3 pairs on the first line are in the vicinity of the ribosomal cluster) using RNA-seq data obtained in 3 different conditions (exponential phase (Expo), cold shock (CS), heat shock (HS)) – note that we also added the RNA-seq profiles of the stationary phase to show more clearly the TSSs of the downstream operons (for clarity, the profiles in the figure were translated such that the mean value of the exponential phase was equal to 7). The profiles were normalized with respect to the expression of the stable gene MPN517 (same normalization as in RT-qPCR) and two values of η corresponding to two replicates were reported in each case. Mean expressions were taken to be equal to $2^{\text{RNA-seq intensity}}$, $[X]$ was measured as the expression at the stop codon of the upstream gene (indicated by the vertical dashed gray line) and $[R]$ just before the TSS of the downstream gene (the TSS was specifically refined by hand in each case as indicated by the color arrows). For the overlapping case (MPN155a-MPN155; MPN155a is a new small protein described in Lluch Senar et al., 2015), η was set to 1; note that for the strongly correlated pair MPN227-MPN228, we observe high and stable values of η as well, although the genes do not overlap. In the case of MPN161-MPN162 (low correlation), one can observe a poor correlation between the changes in η (and $[R]$) and the changes in $[Y]$ for the cold shock experiment, likely indicating that only a small amount of TRT actually extend to the downstream gene. C) According to the model, the level of transcripts extending from the first operon to the second operon should correspond to $[R]$. By performing a RT-qPCR of extended transcripts, that is, of sequences that encompass the intergenic regions and that overlap with the ORF of the genes (small drawing at the bottom), we estimated quantitatively the variation of extended TRT in cold shock and heat shock with respect to the exponential phase (RT-qPCR data were normalized with respect to the stable gene MPN517). Remarkably, the two approaches (model and RT-qPCR) led to similar results, both qualitatively and quantitatively (error bars indicate 95% confidence intervals); note here that $[R]$ was estimated from the RNA-seq data by considering the minimum value of the RNA-seq profile in the region $[O_{up}, TSS]$, where O_{up} indicates the position of the RT-qPCR oligo in the upstream ORF (see small drawing). The order of panels correspond to the order of panels in (B). Overall, we can conclude that TRT is globally enhanced during cold shock, while it tends to be reduced during heat shock.

Figure S4. Related to Figure 5. A) Number of hairpins in the intergenic regions of co-directional genes as a function of their co-expression level. B) Number of hairpins in the intergenic regions of co-directional genes as a function of

the length of the region. The control corresponds to positions of the intergenic regions that were shifted by 10 kbps. These results show that without any additional constrains such as, e.g., the presence of U-tracts (see Figure 5A), the tendency observed on panel A is mainly an effect of the fact that the lower the co-expression, the larger the intergenic region. C) Fraction of intergenic regions containing an RPOD as a function of the length of the intergenic regions.

Figure S5. Related to Figure 5. – A) A simple model of condition-dependent transcription en bloc capturing the 3-level organization of co-expression, according which the RNAP has three possibilities after the transcription of a gene (or an operon):

1. It can systematically continue the transcription process (green light). In this case the system is reminiscent of an operon unit, although the downstream gene may contain a TSS as indicated by the small red arrow.
2. It can continue transcription only from time to time (orange light). Such stochastic transcription en bloc can occur within a given condition, given rise to a sub-operon pattern as schematically represented on the figure and as shown in Figure 5C. Variations of the capacity of transcribing en bloc can also occur between conditions as shown in Figure 5C, in which case a specific regulatory mechanism should be involved.
3. It never transcribes the two genes en bloc, in any condition. In this case, the genes might behave independently, provided that local concentration effects are not too strong.

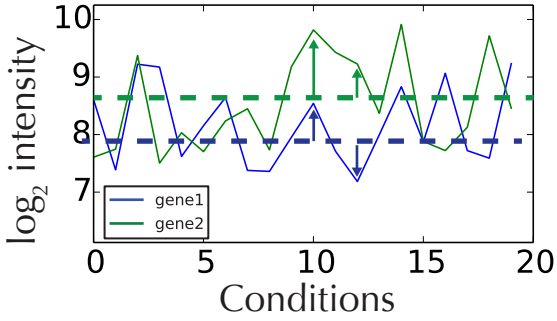
B) In a scenario of a transcription en bloc, the upstream operons should be more prone to transcription initiation, otherwise downstream operons would be more transcribed than upstream ones, leading to a gradient of gene expression within the domain. This phenomenon could be explained by the fact that the upstream negative supercoiling produced by a transcribing RNAP can enhance the activation of operons by favoring the melting of DNA promoters (Meyer & Beslon, 2014). In this context, RPODs can act as topological barriers upstream the domain, while both intrinsic terminators and RNAPs can prevent transcriptional read-through downstream the domain.

Figure S6. Related to Figure 2. We performed the same analysis as that reported in Figure 2 for *E. coli* and *B. subtilis*. For *E. coli*, we used micro-array data obtained across 466 conditions, for more than 4000 genes (McClure et al., 2013). For *B. subtilis*, we used RNA-seq data obtained across 269 conditions (Nicolas et al., 2012). Following our network approach to discard possible outliers (see main text), we identified thresholds (vertical black lines) around 0.7 in *E. coli* and around 0.9 in *B. subtilis* (leftmost panels). In *E. coli*, the resulting network was composed of a single connected component, meaning that we considered the whole set of conditions in this case. In *B. subtilis*, the largest component contained 120 conditions. Using these conditions to compute co-expression among genes, we obtain qualitatively the same results as in Figure 2, both in *E. coli* and in *B. subtilis*, although with different thresholds for the 3-level organization of the co-expression of adjacent genes – note here that only protein-encoding genes were considered in these studies.

References

1. Dorman, C. J. (2011). Regulation of transcription by DNA supercoiling in Mycoplasma genitalium: global control in the smallest known self-replicating genome. *Molecular Microbiology*, *81*(2), 302–304.
2. Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., et al. (2009). Transcriptome complexity in a genome-reduced bacterium. *Science*, *326*(5957), 1268–1271.
3. Hasselbring, B. M., Jordan, J. L., Krause, R. W., Krause, D. C. (2006), Terminal organelle development in the cell wall-less bacterium Mycoplasma pneumoniae. *PNAS*, 103(44):16478-83.
4. Kühner, S., van Noort, V., Betts, M. J., Leo-Macias, A., Batisse, C., Rode, M., et al. (2009). Proteome organization in a genome-reduced bacterium. *Science*, *326*(5957), 1235–1240.
5. Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858.
6. Lloréns-Rico V., Lluch-Senar M., & Serrano, L. (2015). Distinguishing between productive and abortive promoters using a random forest classifier in *Mycoplasma pneumoniae,* Nucleic Acids Res. 2015 : gkv170v1-gkv170.
7. Lluch Senar, M., Delgado, J., Chen, W.-H., Lloréns-Rico, V., O'Reilly, F. J., Wodke, J. A., et al. (2015). Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium.

Molecular Systems Biology, *11*(1), 780.
8. Maier, T., Schmidt, A., Güell, M., et al. (2011). Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Molecular Systems Biology*, 7:511.
9. Mathews, D. H., Sabina, J., Zuker, M. (1999), Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure, J. Mol. Biol. (1999) 288, 911–940.
10. McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C. A., et al. (2013). Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Research*, *41*(14), e140.
11. Meyer, S., & Beslon, G. (2014). Torsion-Mediated Interaction between Adjacent Genes. *PLoS Computational Biology*, *10*(9), e1003785.
12. Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., et al. (2012). Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus subtilis. *Science*, *335*(6072), 1103–1106.
13. Song, L., Langfelder, P., & Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, *13*(1), 328.
14. Wodke, J. A. H., Alibés, A., Cozzuto, L., Hermoso, A., Yus, E., Lluch Senar, M., et al. (2015). MyMpn: a database for the systems biology model organism Mycoplasma pneumoniae. *Nucleic Acids Research*, *43*(Database issue), D618–23.
15. Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.-H., et al. (2009). Impact of genome reduction on bacterial metabolism and its regulation. *Science*, *326*(5957), 1263–1268.
16. Yus, E., Güell, M., Vivancos, A. P., Chen, W.-H., Lluch Senar, M., Delgado, J., et al. (2012). Transcription start site associated RNAs in bacteria. *Molecular Systems Biology*, *8*, 585.

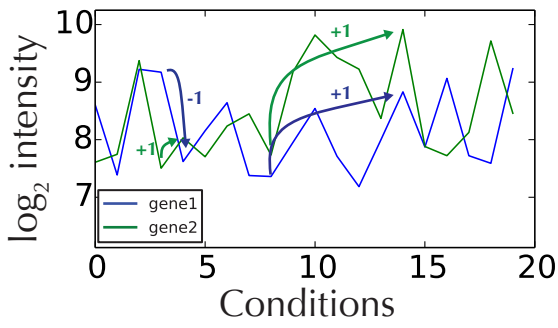


$$\text{Pearson} = \frac{\sum_s \bar{a}_{s1}\bar{a}_{s2}}{\sqrt{(\sum_s \bar{a}_{s1}^2)(\sum_s \bar{a}_{s2}^2)}} \qquad \bar{a}_{si} = a_{si} - \sum_s a_{si}/M$$

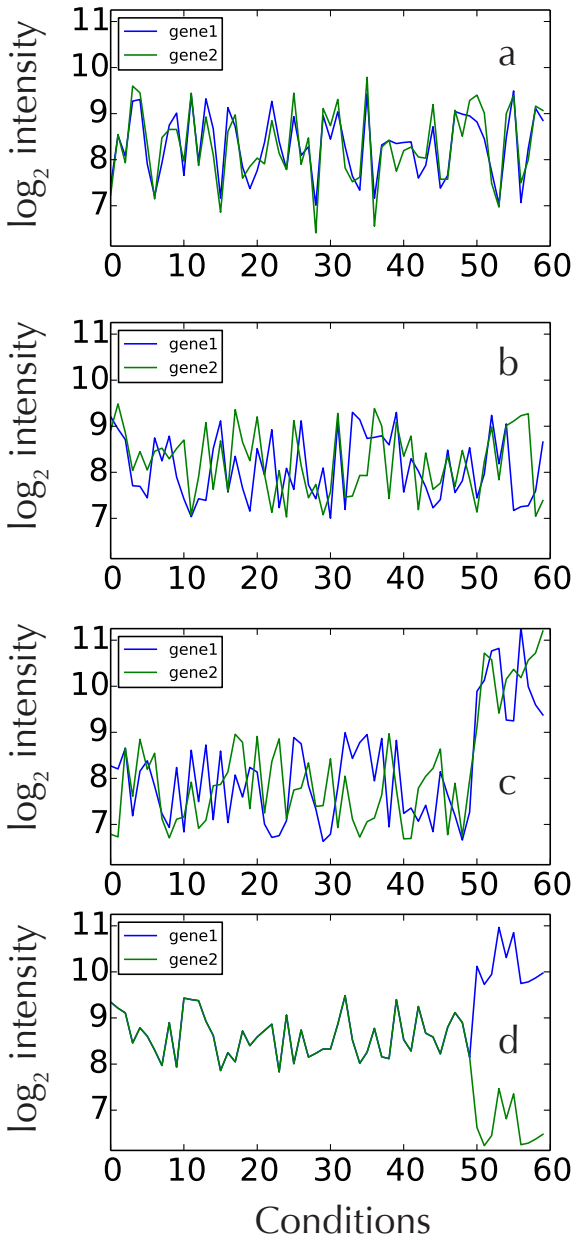
$$\text{bicor} = \frac{\sum_s \tilde{a}_{s1}\tilde{a}_{s2}}{\sqrt{(\sum_s \tilde{a}_{s1}^2)(\sum_s \tilde{a}_{s2}^2)}} \qquad \tilde{a}_{si} = w_{si}(a_{si} - \text{med}(a_s))$$

$\text{med}(a_s)$ = median of expression values

w_{si} = weigth involving the median and the median absolute deviation of gene expressions



$$\text{basal} = \frac{\sum_{s,s'} \text{sgn}(a_{s1} - a_{s'1}) \cdot \text{sgn}(a_{s2} - a_{s'2})}{2M^2}$$



Pearson = 0.89
bicor = 0.88
basal = 0.73

Pearson = 0.01
bicor = 0.02
basal = 0.01

Pearson = 0.61
bicor = 0.51
basal = 0.25

Pearson = −0.28
bicor = −0.1
basal = 0.44

Figure S1

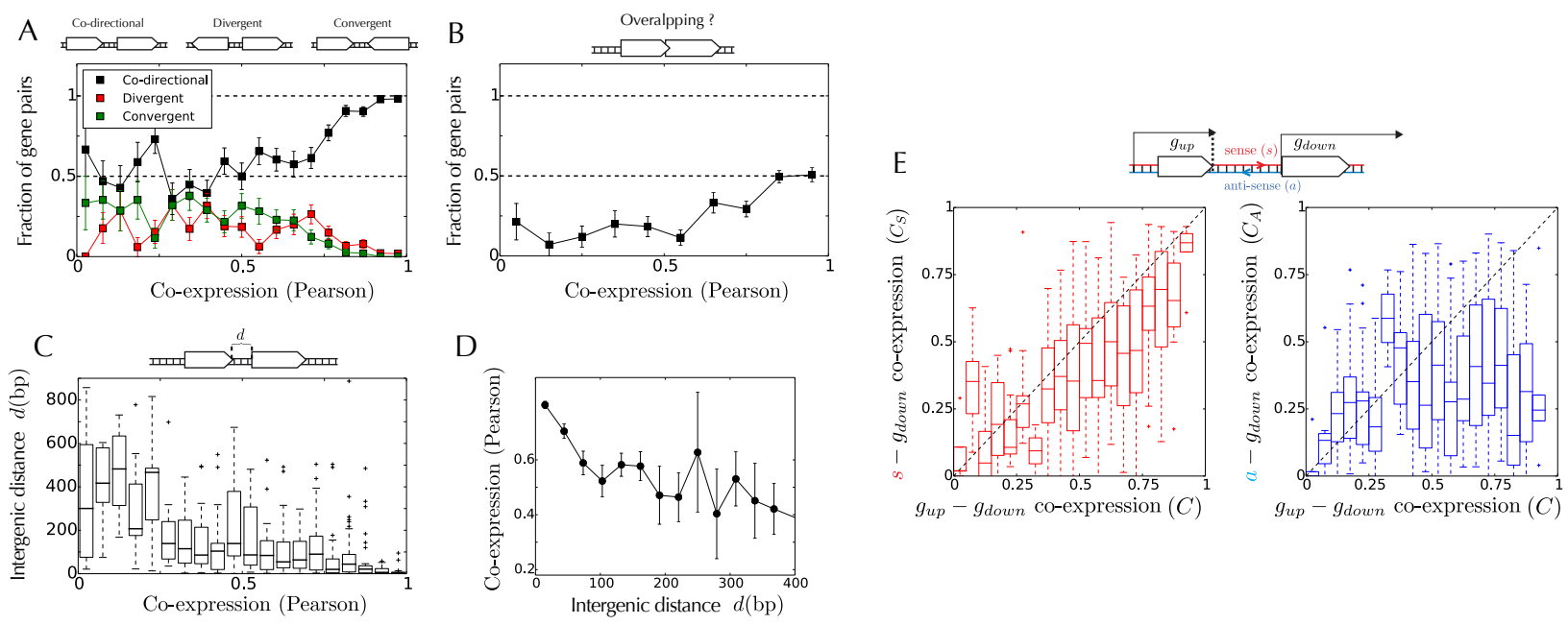

A
Co-directional
Divergent
Convergent
Co-directional
Divergent
Convergent
Fraction of gene pairs
Co-expression (Pearson)
B
Overalpping ?
Fraction of gene pairs
Co-expression (Pearson)
C
d
Intergenic distance d(bp)
Co-expression (Pearson)
D
Co-expression (Pearson)
Intergenic distance d(bp)
E
g_{up}
sense (s)
g_{down}
anti-sense (a)
$s - g_{down}$ co-expression (C_S)
$g_{up} - g_{down}$ co-expression (C)
$a - g_{down}$ co-expression (C_A)
$g_{up} - g_{down}$ co-expression (C)

Figure S2

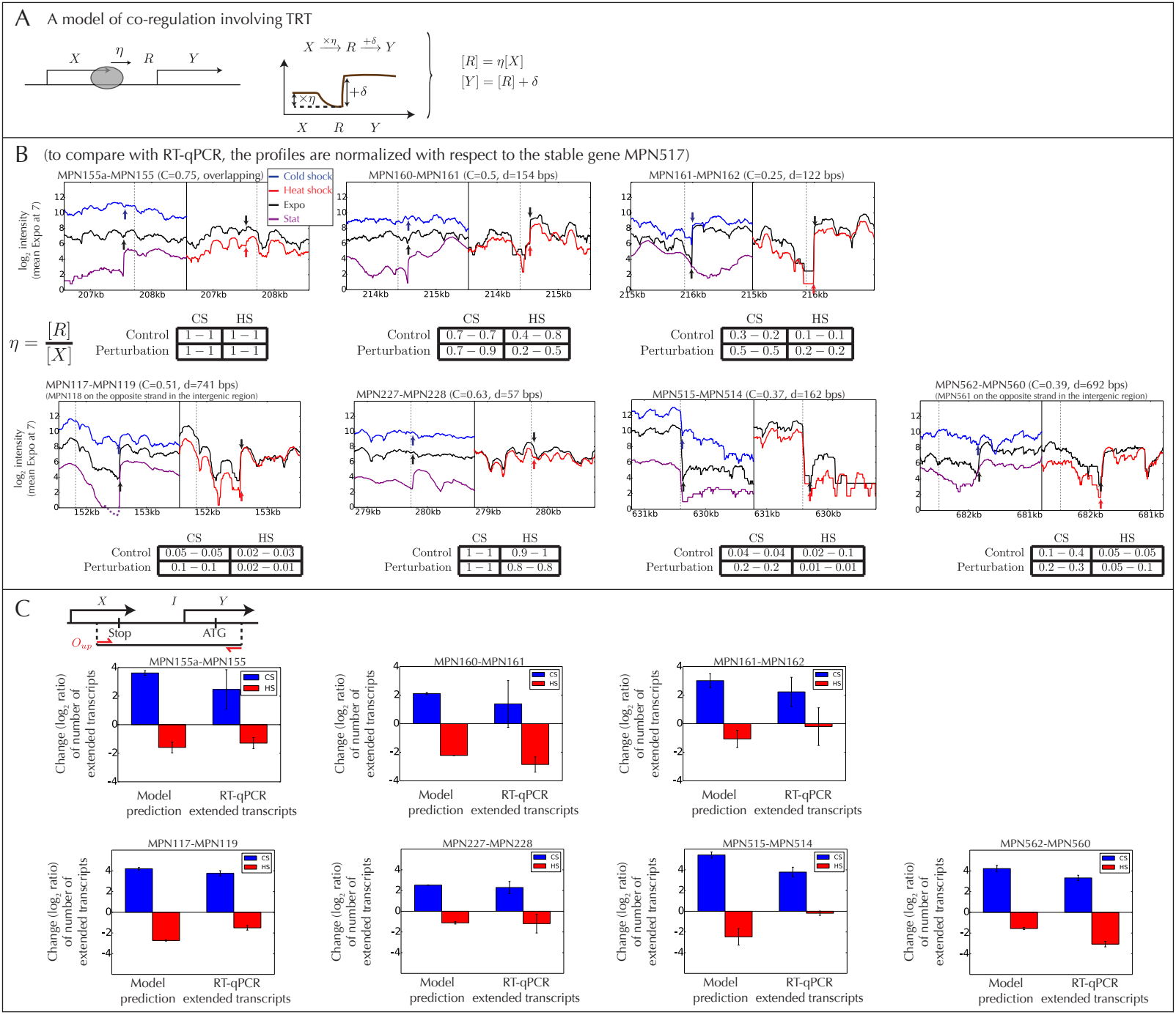

A
A model of co-regulation involving TRT
X
η
R
Y
$X \xrightarrow{\times\eta} R \xrightarrow{+\delta} Y$
$+\delta$
$\times\eta$
$[R] = \eta[X]$
$[Y] = [R] + \delta$
B
(to compare with RT-qPCR, the profiles are normalized with respect to the stable gene MPN517)
MPN155a-MPN155 (C=0.75, overlapping)
Cold shock
Heat shock
Expo
Stat
log2 intensity (mean Expo at 7)
MPN160-MPN161 (C=0.5, d=154 bps)
MPN161-MPN162 (C=0.25, d=122 bps)
$\eta = \frac{[R]}{[X]}$
CS HS
Control 1 − 1 1 − 1
Perturbation 1 − 1 1 − 1
CS HS
Control 0.7 − 0.7 0.4 − 0.8
Perturbation 0.7 − 0.9 0.2 − 0.5
CS HS
Control 0.3 − 0.2 0.1 − 0.1
Perturbation 0.5 − 0.5 0.2 − 0.2
MPN117-MPN119 (C=0.51, d=741 bps)
(MPN118 on the opposite strand in the intergenic region)
MPN227-MPN228 (C=0.63, d=57 bps)
MPN515-MPN514 (C=0.37, d=162 bps)
MPN562-MPN560 (C=0.39, d=692 bps)
(MPN561 on the opposite strand in the intergenic region)
CS HS
Control 0.05 − 0.05 0.02 − 0.03
Perturbation 0.1 − 0.1 0.02 − 0.01
CS HS
Control 1 − 1 0.9 − 1
Perturbation 1 − 1 0.8 − 0.8
CS HS
Control 0.04 − 0.04 0.02 − 0.1
Perturbation 0.2 − 0.2 0.01 − 0.01
CS HS
Control 0.1 − 0.4 0.05 − 0.05
Perturbation 0.2 − 0.3 0.05 − 0.1
C
X I Y
Stop ATG
O_{up}
MPN155a-MPN155
MPN160-MPN161
MPN161-MPN162
Change (log2 ratio) of number of extended transcripts
CS HS
Model prediction
RT-qPCR extended transcripts
MPN117-MPN119
MPN227-MPN228
MPN515-MPN514
MPN562-MPN560

Figure S3

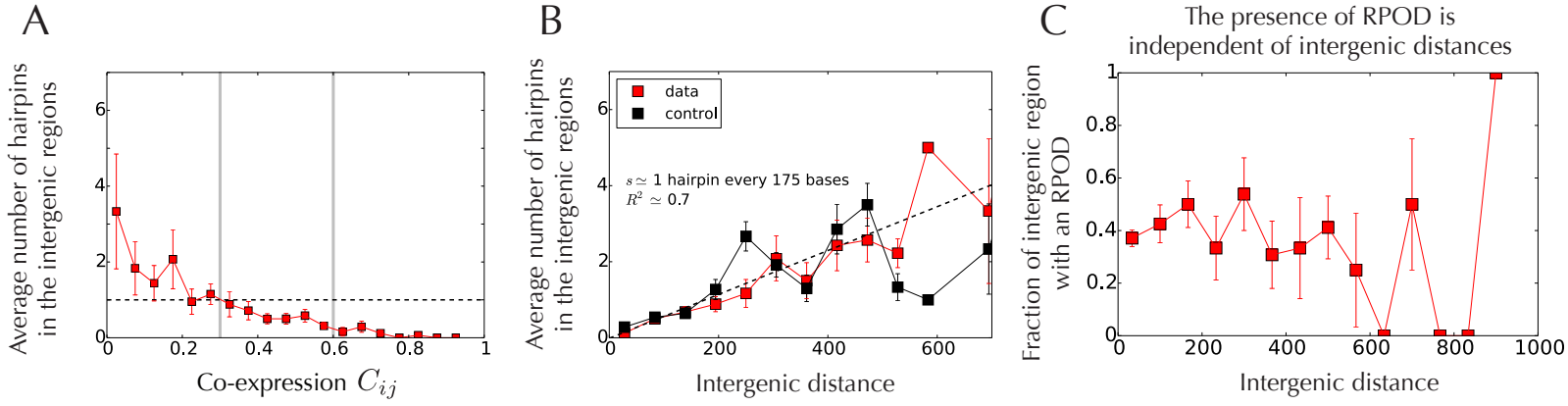
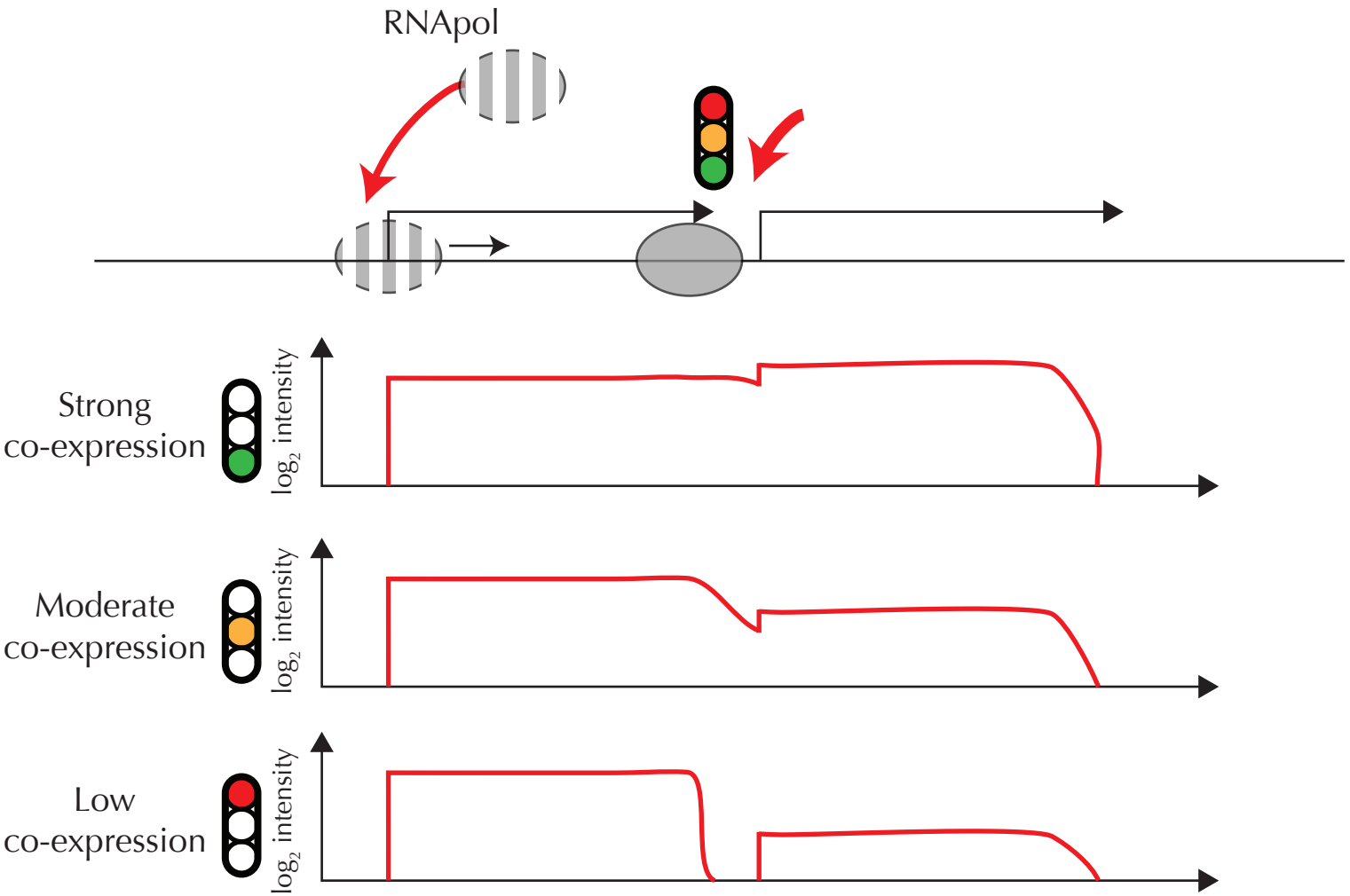

A
Average number of hairpins in the intergenic regions
Co-expression C_{ij}
B
Average number of hairpins in the intergenic regions
data
control
$s \simeq$ 1 hairpin every 175 bases
$R^2 \simeq 0.7$
Intergenic distance
C
The presence of RPOD is independent of intergenic distances
Fraction of intergenic region with an RPOD
Intergenic distance

Figure S4

A) Various degrees of co-expression, depending on the frequency of transcriptional read-through



B) Physical mechanims associated to the transcription en bloc of a specific large domain

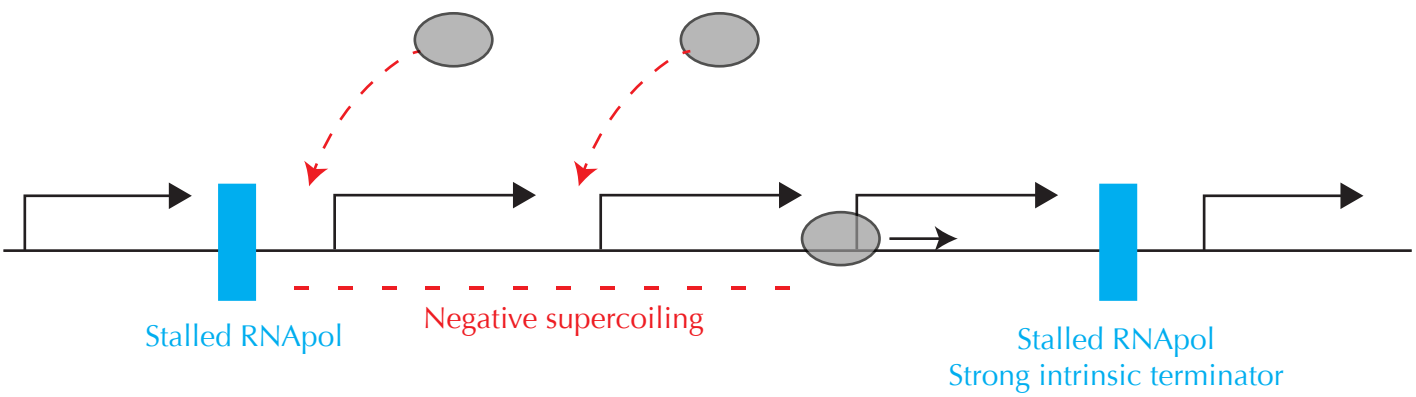
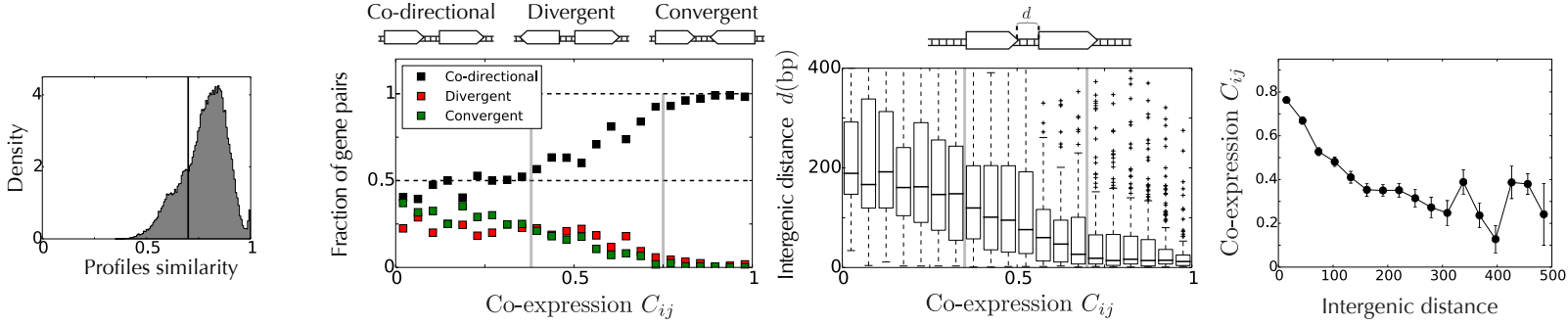


Figure S5

E. coli
(micro-array data)



B. subtilis
(RNA-seq data)

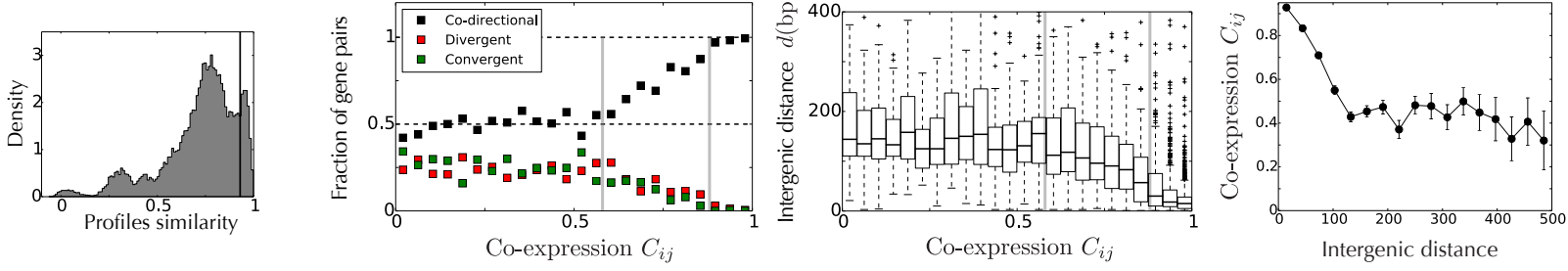


Figure S6