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

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.

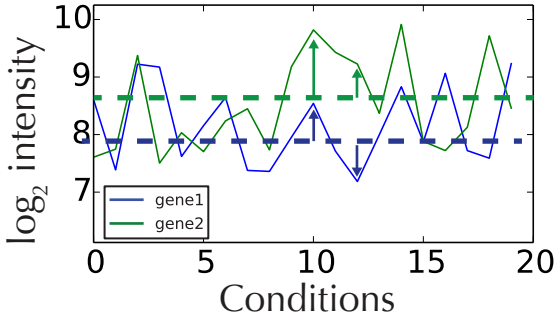
B) In a scenario of a transcription en bloc, the upstream operons should be more prone to transcription initiation, otherwise downstream operons would be more transcribed than upstream ones, leading to a gradient of gene expression within the domain. This phenomenon could be explained by the fact that 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.

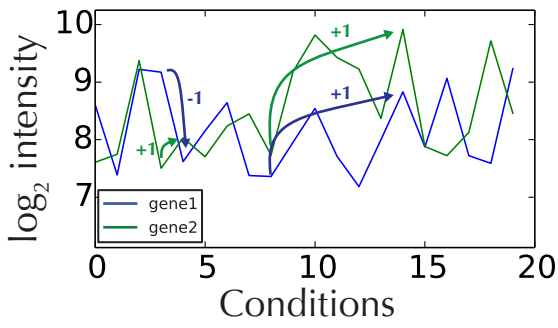


$$\text{Pearson} = \frac{\sum_s \bar{a}_{s1}\bar{a}_{s2}}{\sqrt{(\sum_s \bar{a}_{s1}^2)(\sum_s \bar{a}_{s2}^2)}} \qquad \bar{a}_{si} = a_{si} - \sum_s a_{si}/M$$

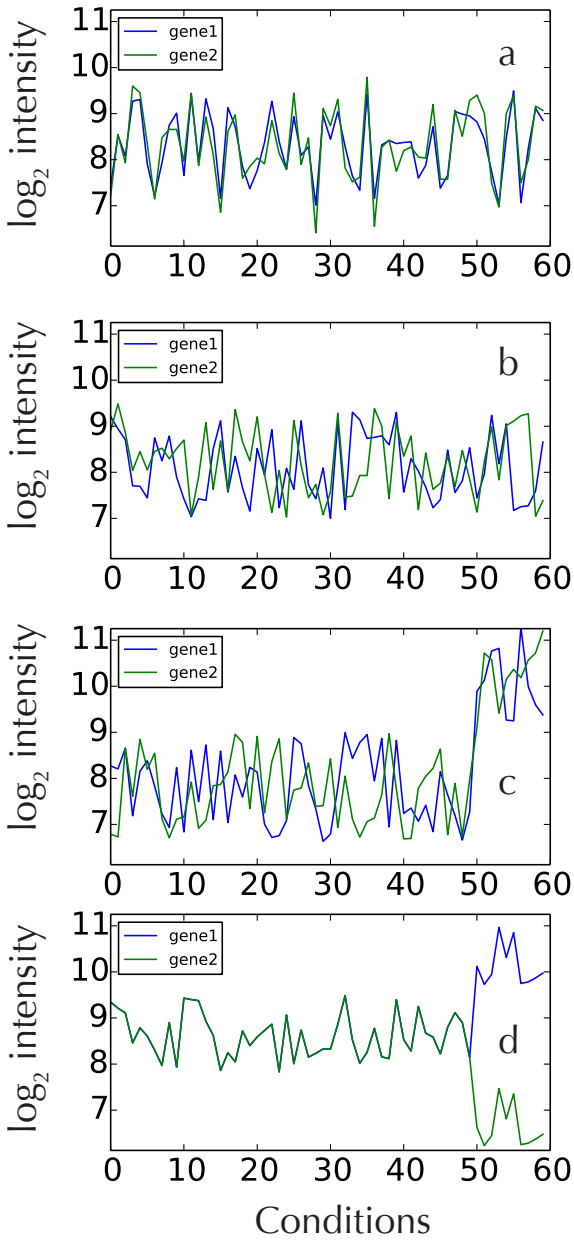
$$\text{bicor} = \frac{\sum_s \tilde{a}_{s1}\tilde{a}_{s2}}{\sqrt{(\sum_s \tilde{a}_{s1}^2)(\sum_s \tilde{a}_{s2}^2)}} \qquad \tilde{a}_{si} = w_{si}(a_{si} - \text{med}(a_s))$$

$\text{med}(a_s)$ = 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

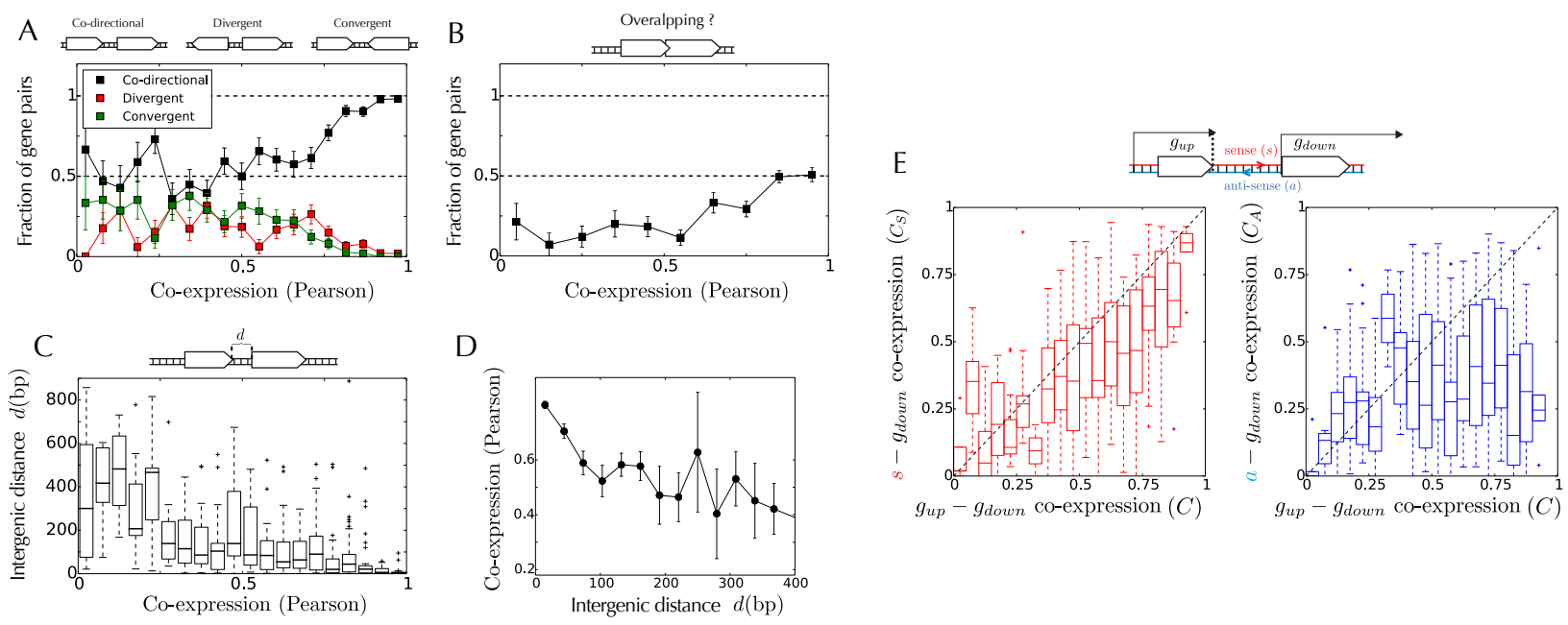

A
Co-directional
Divergent
Convergent
Co-directional
Divergent
Convergent
Fraction of gene pairs
Co-expression (Pearson)
B
Overalpping ?
Fraction of gene pairs
Co-expression (Pearson)
C
d
Intergenic distance d(bp)
Co-expression (Pearson)
D
Co-expression (Pearson)
Intergenic distance d(bp)
E
g_{up}
sense (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

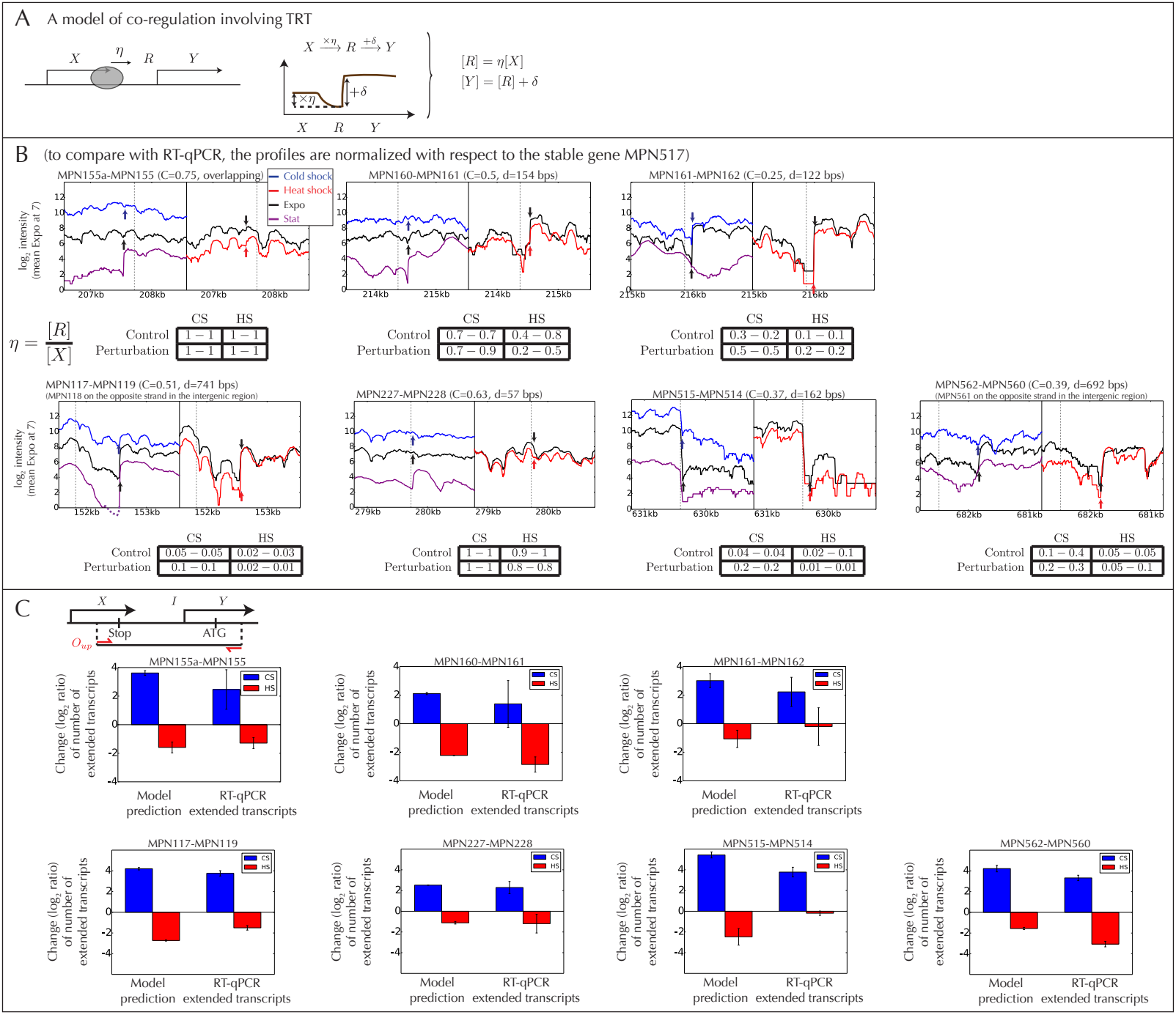

A
A model of co-regulation involving TRT
X
η
R
Y
$X \xrightarrow{\times\eta} R \xrightarrow{+\delta} Y$
$+\delta$
$\times\eta$
$[R] = \eta[X]$
$[Y] = [R] + \delta$
B
(to compare with RT-qPCR, the profiles are normalized with respect to the stable gene MPN517)
MPN155a-MPN155 (C=0.75, overlapping)
Cold 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

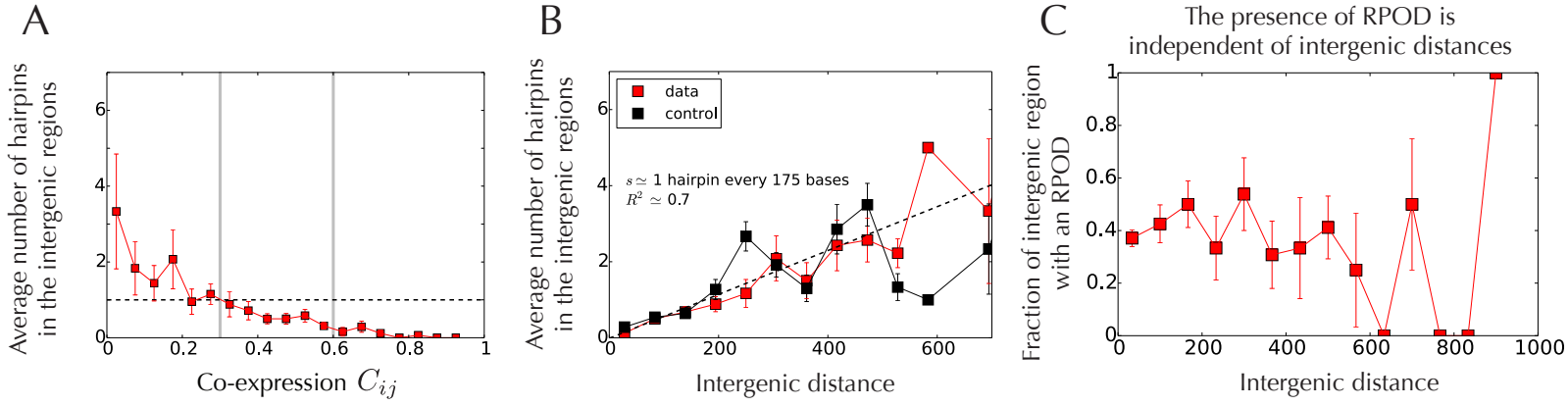
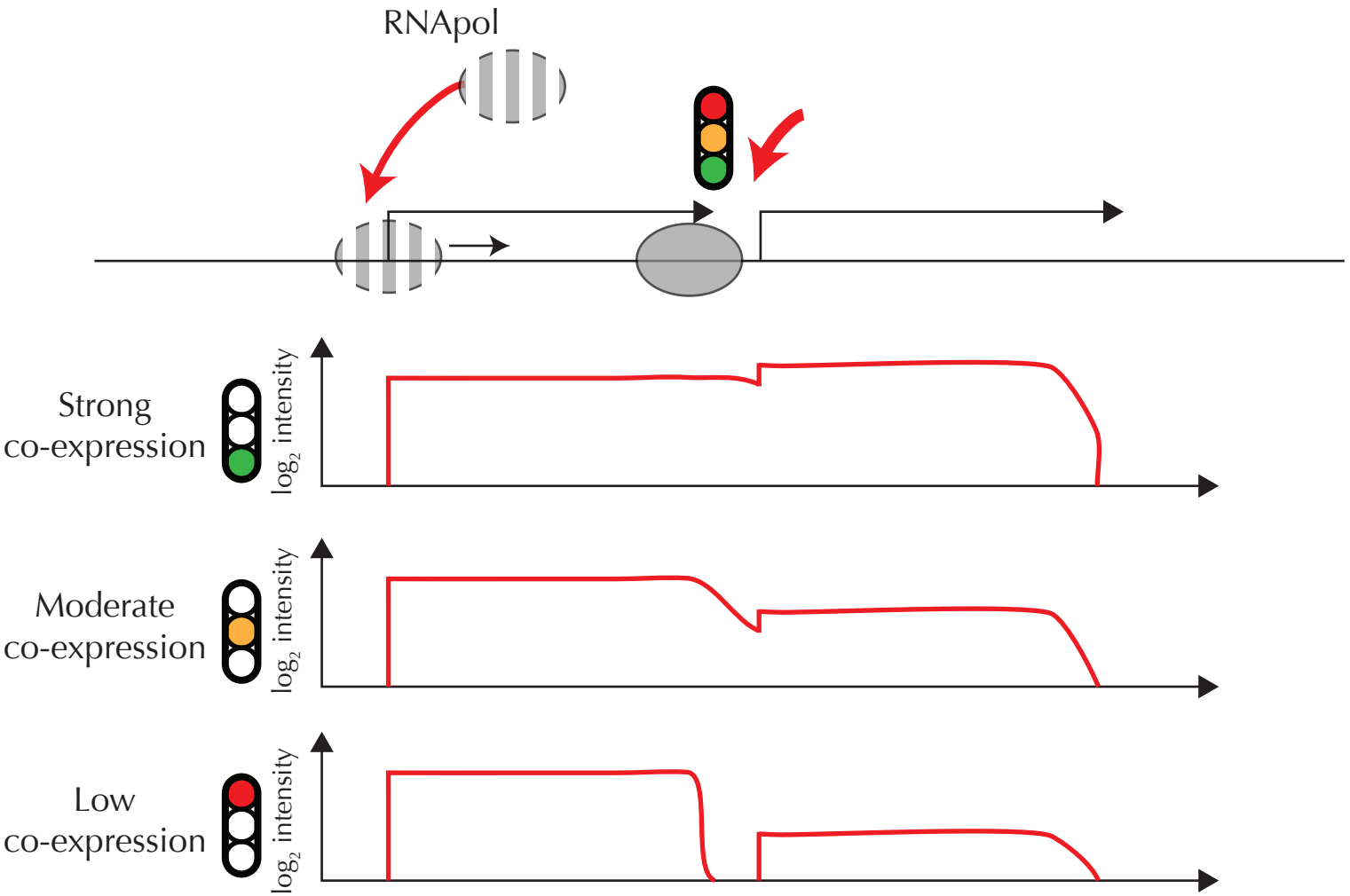

A
Average number of hairpins in the intergenic regions
Co-expression C_{ij}
B
Average number of hairpins in the intergenic regions
data
control
$s \simeq$ 1 hairpin every 175 bases
$R^2 \simeq 0.7$
Intergenic distance
C
The presence of RPOD is independent of intergenic distances
Fraction of intergenic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

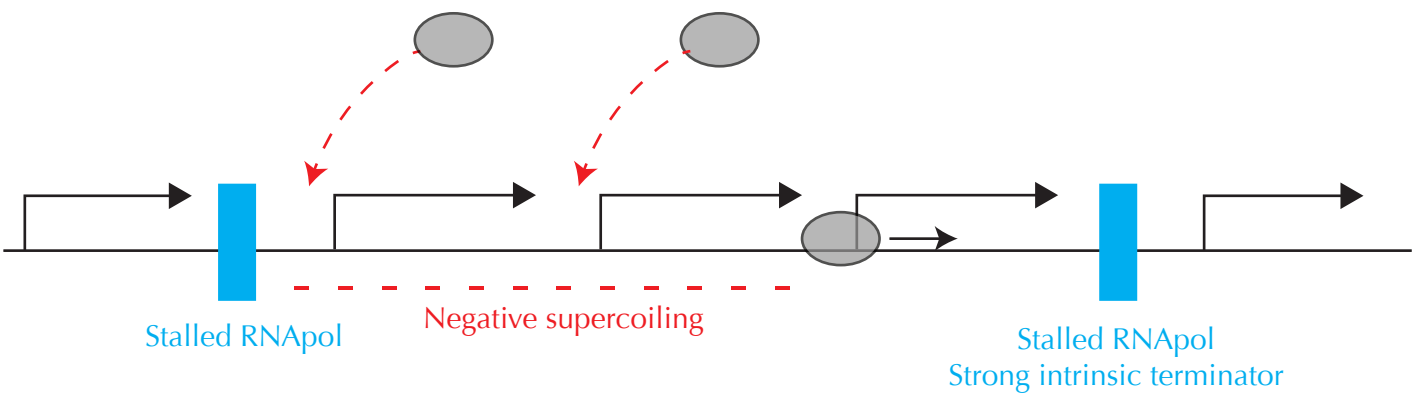
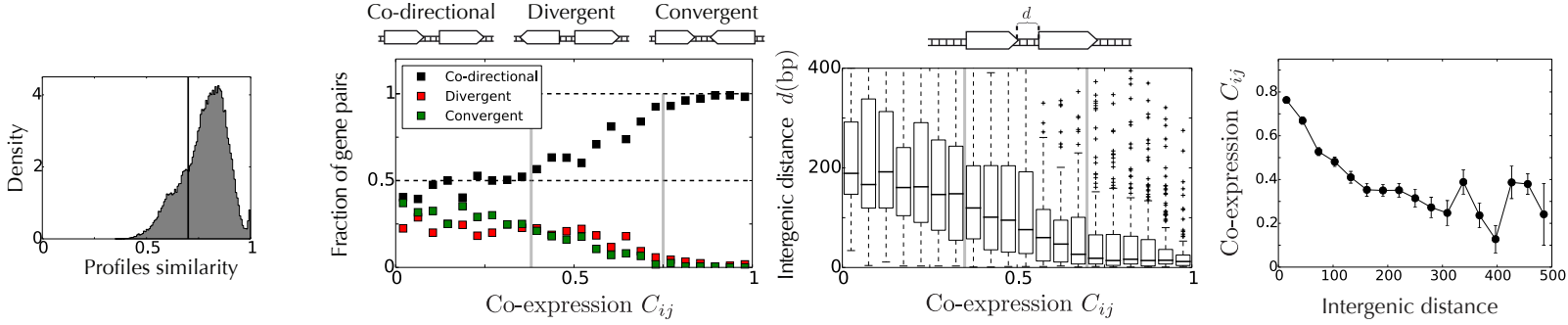


Figure S5

E. coli
(micro-array data)



B. subtilis
(RNA-seq data)

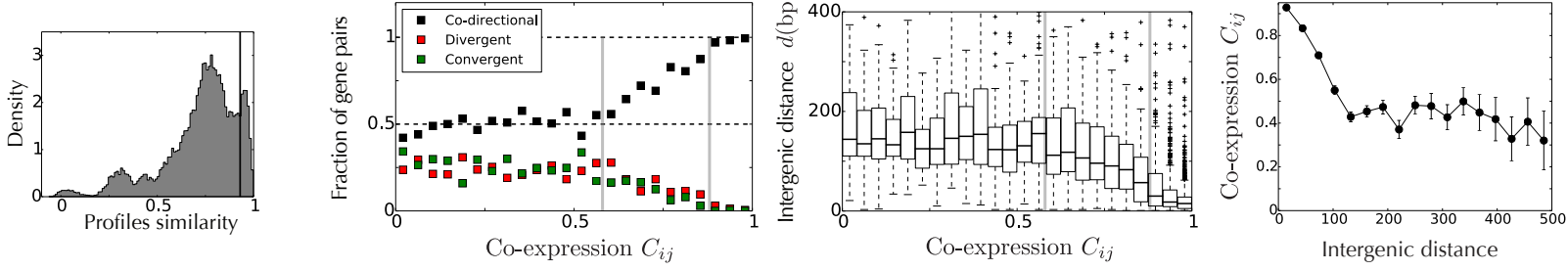


Figure S6