Large-scale analysis of the global transcriptional regulation of bacterial gene expression

Pablo Yubero and Juan F Poyatos

Logic of Genomic Systems Laboratory, CNB-CSIC, Madrid, Spain

ABSTRACT

Changes in bacterial gene expression depend not only on the operation of specific transcription factors but also on the regulation exerted by the global physiological condition of the cell. To characterize the influence of this global regulation constitutive genes appear as the most valid model, that can also help us distinguish the impact of physiology on genome organization. However, only a few genes have been investigated so far. Here, we present a large-scale approach to characterize the regulation of the global program by quantifying chromosomal promoter activities of over 700 constitutive genes in *Escherichia coli*. We identify four response classes (growth-rate dependencies) of which one follows the hyperbolic pattern previously seen. Within this class, we sort genes with respect to their responsiveness to growth rate and examine the association of responsiveness with gene order. Genes showing a particularly sensitive linear response to growth are located near the origin of replication, even when controlling for the increase in the number of replication forks. We then postulate that global regulation could act as a driving force to shape genomic architecture. Evidence is obtained by examining the relation between position conservation of *E. coli* genes in 100 bacterial species with the number of replication rounds, maximal growth rate and environmental variability of the species' habitat. The response to the physiological state of the cell results therefore in an additional feature contributing to bacterial genome organization.

INTRODUCTION

Transcription regulation is one of the fundamental mechanisms by which bacteria adapt gene expression to changing environmental conditions. Apart from the specific action mediated by transcription factors (TFs), expression is modulated by a global regulatory program determined by the physiological condition of the cell. Initial studies correlated this condition to the availability of core constituents of the expression machinery [free RNA polymerase, tRNAs, ribosomes, etc. (Kjeldgaard et al. 1958; Schaechter et al. 1958)], but many other interacting components can play a role (Kubitschek 1974; Liang et al. 1999). These works also provided an effective protocol to describe the influence of all these elements. The dependence on physiology was found to be linked to growth rate at balanced growth, independent of the particular media. Therefore, the examination of the global program reduced to the quantification of expression response to growth rate.

That global physiology complements specific regulation matters in many aspects. Indeed, growth rate dependencies can interfere directly with genetic circuits and change their operation, for example, by shifting the bistability regime of a switch or allowing for different antibiotic resistance strategies (Klumpp et al. 2009; Deris et al. 2013). Costs of synthetic genetic circuits on cell physiology and the impact of the latter on the function of the circuits made the subject also relevant in applied areas, e.g., Synthetic Biology (Scott et al. 2010). Beyond "simple" genetic circuits, the interplay of global regulation and cell resource allocation can modify many essential features at the system level (Klumpp and Hwa 2008; Peebo et al. 2015). In fact, mechanistic approaches revealed that the global regulatory program contributes to determining fundamental trade-offs involving the finiteness of the cellular size, energy, and ribosomal fraction (Weiße et al. 2015).

To examine the impact of this program the choice of constitutive genes as the primary model is clear: promoters of these genes lack any interaction with TFs, and thus, they are *a priori* constantly available for transcription initiation. Therefore, constitutive genes are subject only to physiological regulation. An alternative approach is to mutate the TF binding sites of non-constitutive genes to assess the separate (mutant) and combined (wild-type) effect of global and specific regulation. Studies applying these approximations included however only a few genes (Berthoumieux et al.

2013; Gerosa et al. 2013; Kochanowski et al. 2017), and thus we lack a large-scale evaluation of global regulation.

Beyond its evaluation, it is also intriguing to examine to what degree global regulation impacts bacterial genomic organization. One of the factors contributing to this regulation is copy number, which is relevant since gene dosage depends on the proximity to the origin of replication *oriC* and the growth rate. This is due to the overlap of many replication rounds at fast growth rates (multifork replication). Indeed, the position and copy number of ribosomal genes in *Escherichia coli* is tuned to maintain fast growth rates (Gyorfy et al. 2015). We could nevertheless ask if the global program affects genomic structure excluding the copy number. Two scenarios can be hypothesized. One in which promoters that are less sensitive to global regulation are located closer to *oriC* –to gain sensitivity associated to the copy number from the expression increase linked to multifork replication. And a second scenario in which promoters that are intrinsically sensitive, i.e., excluding the copy number contribution, are closer to the *oriC* to enhance their activity with growth rate. In the first situation, the influence of the global program is mainly linked to copy number whereas in the second copy number strengthens the dependence. Either solution would reveal design principles of genome architecture.

In this work, we analyze the global regulation program in *E. coli* at a large scale, by introducing a procedure that enables us to examine the response to this program of ~ 700 constitutive genes with the use of time series data. For a validated subset of these genes, we quantify the most sensitive to the global program and observe that they are significantly located near the origin of replication. This would then indicate that the position of these constitutive genes could be particularly conserved in species experiencing faster or more variable growth rates (both situations in which this basal mode of regulation could be essential). We examine evidences in this respect with the analysis of the correlation between position conservation of the corresponding *E. coli* genes in 100 bacterial species and the number of replication rounds, maximal growth rate, and environmental variability of the species' habitat.

RESULTS

Quantifying chromosomal promoter activity at a large scale.

To quantify the promoter activity of chromosomal genes (PA_{chr}) we followed an approach that makes use of promoter activity measurements obtained with low-copy plasmids (PA_{pl}). This is of interest as the availability of a fluorescent library in *E. coli* (Zaslaver et al. 2006) could then be used to determine PA_{chr} at a large scale. We first decoupled the copy number signal (g_{pl}) that contributes to PA_{pl} (Fig. 1A). Since the replication of these plasmids is synced to the cell cycle (del Solar et al. 1998; Morrison and Chattoraj 2004), the signal is proportional to the number of terminal regions in the chromosome (*ter*).

We can thus consider Cooper and Helmstetter's model (Cooper and Helmstetter 1968) describing the copy number of a chromosomal gene (g_{chr}) for a given growth rate (μ) to the case of a gene at *ter* ($g_{chr} = 2^{\mu[C(1-m)+D]}$, recall that m represents the normalized distance to the origin of replication of the gene) to obtain the plasmid copy number: $g_{pl} = k2^{\mu D} \propto g_{ter}$, with $m_{ter} = 1$ and k representing a proportionality constant [$k = 5$ from now on (Gerosa et al. 2013); the values of C and D were obtained by interpolation from experimental measurements (Bremer and Dennis 1996)]. This led us to the promoter activity per gene copy number, pa = PA_{pl}/g_{pl}, where the effect of copy number is excluded, and to the chromosomal promoter activity $PA_{chr} = \text{pa} \times g_{chr}$ (Fig. 1A.B).

Figures 1C-D show the resulting promoter activities of two example cases [using experimental data from (Zaslaver et al. 2009)]; genes *rph* and *hisL* located at distances of $m_{rph} = 0.04$ and $m_{hisL} = 0.80$, respectively. Differences in promoter activity become more relevant for *rph* because of its proximity to *oriC* compared to *hisL*. In this way, the distinction between the promoter activity per gene copy (pa, Fig. 1A) and chromosomal promoter activity (PA_{chr}, Fig. 1A) emphasizes the added effect of multifork replication depending on genomic context.

Constitutive promoters show a variety of responses to global regulation.

We applied the previous approach to characterize the global program at a large scale. Constitutive genes appear as the most suitable model, given the absence of any specific regulation acting on them. However, characterizing the response of constitutive promoters in a traditional manner, i.e.,

from balanced growth measurements in different carbon sources, limits in part the scalability of the approach. We used alternatively measurements of promoter activity during dynamic changes of growth rate (in a specific carbon source) since recent results showed that these measures correlate well with those observed with different carbon sources at balanced growth (Gerosa et al. 2013).

We thus processed the time series data of the set of 708 constitutive genes included in (Zaslaver et al. 2009)(Methods). Instead of measuring hundreds of genes in many distinct carbon sources, we considered data during exponential and late-exponential growth (within the first 5hrs) in the same growth medium (Zaslaver et al. 2009) to obtain profiles of promoter activity and growth rate; $\mathrm{PA}_{pl}(\mu)$ profiles (Supplementary Figure S1, Methods). Data derived in this way can be decoupled from their plasmid context in order to get chromosomal, $\mathrm{PA}_{chr}(\mu)$, and per gene, $\mathrm{pa}(\mu)$, profiles (previous section).

After computing $\mathrm{PA}_{chr}(\mu)$ we grouped all resulting profiles into four classes (applying a clustering algorithm, Methods, Fig. 2A). Classes 1 and 2 correspond to promoters whose activity increases or decreases respectively with growth rate, whereas class 3 corresponds to promoter activities that remain mostly constant across growth rates. Finally, class 4 includes promoters with a characteristic profile that has maximum promoter activity at intermediate growth rates. These classes are robust whether $\mathrm{PA}_{chr}(\mu)$ or $\mathrm{pa}(\mu)$ profiles are used for the classification (Supplementary Figure S2, and Table S1). Note that only genes from the first class (56%) behave as it is expected from foundational and recent works (Liang et al. 1999; Gerosa et al. 2013; Kochanowski et al. 2017).

To test this procedure, we experimentally validated the promoter activity profiles by measuring $\mathrm{PA}_{chr}(\mu)$ of 12 promoters –chosen among all four classes– from balanced growth data in 10 different growth media (Methods, Supplementary Figure S3). Figure 2B shows the experimental results of *maoP, rsd, rpsT, prmB, rssB, amyA, yaaA, ghrA, nudF, cpsG, argQ* and *mutT* promoters (a brief description of these genes is available in Supplementary Table S2). We observe that the approach of inferring PA_{chr} profiles from time series on a single growth medium is particularly robust for class 1 promoters (Supplementary Figure S4). Overall, this suggests that during the first hours of growth arrest after balanced growth, when entering saturation phase, the expression of constitutive

promoters is still mostly determined by the decrease availability of global resources. Promoters that deviate from this can be suspected of having unknown regulatory mechanisms.

Promoters sensitive to global regulation are located closer to the origin of replication. Beyond the previous classification, we noted different genes within class 1 promoters with distinct sensitivity to the global regulation. To quantify sensititivity, we fitted $\mathrm{PA}_{chr}(\mu)$ profiles to a Michaelis-Menten equation:

$$\mathrm{PA}_{chr}(\mu) = \frac{V_m \mu}{K_m + \mu}, \tag{1}$$

where V_m is the maximum promoter activity and K_m is the growth rate at which $\mathrm{PA}_{chr}(\mu)$ is half maximal; note that μ records the global program, and that the different responses emphasize a promoter-specific rather than an unspecific pattern (Liang et al. 1999; Klumpp and Hwa 2008; Gerosa et al. 2013). Next, we classified as linear (Fig. 3A) or saturable (Fig. 3B) profiles those with $K_m > 3$ dbl/h and $0.1 < K_m < 3$ dbl/h, respectively (Methods).

In the case of promoters with linear profiles, we defined the sensitivity to the global program as the slope of the $\mathrm{PA}_{chr}(\mu)$ profile, such that larger values stand for larger increases in promoter activity for fixed changes in growth rate. In the case of saturable promoters, we took K_m as a proxy of their sensitivity to the global program. Note that for smaller values of K_m the promoter activity becomes near saturation at smaller growth rates, thus becoming less sensitive to changes in growth rate [we also computed sensitivities of $\mathrm{pa}(\mu)$ profiles determined in an analogous manner; Methods, Supplementary Figures S5-S6 and Supplementary Table S1].

We then asked if there exists an association between sensitivity and chromosomal location, given that one of the factors that influence these responses is multifork replication (relevant near *oriC*). Figure 3C-D shows the running average of the sensitivities to the global program along the chromosome of constitutive promoters with linear and saturable profiles for both $\mathrm{PA}_{chr}(\mu)$ and $\mathrm{pa}(\mu)$. We observed that the sensitivity of linear profiles decreases linearly with the distance to *oriC* more abruptly and more significantly when considering PA_{chr} than pa. In the case of saturable

constitutive promoters, we notice that only when considering PA_{chr} there is a significant peak at $m < 0.20$ of the chromosome ($p < 0.05$).

Global regulation acts as a gene conservation force.

It seems then that in several situations those genes most sensitive to the global program tend to be closer to the origin of replication in *E. coli*. We examined next if these genes maintain their proximity (to *oriC*) in other species as a function of some characteristics of the species: their minimal doubling time, the variability of the environment where they live and, *a priori* anticipated, the capacity for multifork replication. All these features impact on physiology and, in this way, species presenting them could be utilizing the sensitivity to the global program as a supplementary mode of regulation.

To evaluate this assumption, we performed a homolog search across 100 species to compute the corresponding chromosomal displacement (Methods). Displacements of the half most growth-rate-dependent genes near *oriC* ($m < 0.2$) are compared against the null hypothesis, i.e., sensitivity does not affect displacement. This is scored by the probability of finding a larger mean displacement of gene groups of the same size chosen randomly among all constitutive promoters at $m < 0.2$ [for linear or saturated growth-rate dependences (Fig.4A,B)]. Smaller values of this score, termed the conservation measure, represent non-conserved locations of promoters.

We studied next the association between the conservation measure and three main species features: Environmental variability (env), relevance of multifork replication (R), and maximal growth rate (as the inverse of the minimal doubling time, 1/mindt). Environmental variability was based on an earlier environmental classification (Parter et al. 2007), while minimal doubling time with genome size was estimated to compute R; the ratio between the maximal chromosome's replication time and the minimal doubling time as a measure of the importance of multifork replication effects for an organism (Couturier and Rocha 2006). For each class of promoter dependence (linear and saturated, pa and PA_{chr}) we measured the correlation between the corresponding conservation measure and env, R, or 1/mindt. We obtained a significant correlation between conservation and R and 1/mindt, respectively, for lineal promoters (0.3607,0.3467,0.3765,0.4287, all Spearman ρ with

p<0.01; the first two values relate to PA_{chr} and the second pair to pa). Note also that we controlled for phylogenetic distance in all previous associations (Methods; the previous ρ denoted partial Spearman's rank correlation). Therefore, genes exhibiting a particularly sensitive linear response tend to be conserved in species with larger values of R and fast growth rates. Correlations with R are stronger when the global program includes the multifork effect (PA_{chr}; 0.3607 *vs.* 0.3467), as expected from the definition of R, while maximal growth rate (1/mindt) is more relevant when not including the multifork dosage effect (pa; 0.3765 *vs.* 0.4287). The conservation of saturable promoters was not significant.

DISCUSSION

We quantified growth-rate dependencies of over 700 constitutive genes in *E. coli*, arguably the best gene collection to explore the effects of the physiological state of the cell (the global program) on gene expression. This is based on an approach that obtains promoter activity as if reporters had been inserted in the chromosome and that characterizes growth-rate dependencies from dynamical data. Both features reduce the costs and difficulties of large-scale experiments.

Notably, half of the constitutive promoters that we examine present a Michaelis–Menten rate law confirming earlier reports (Liang et al. 1999; Gerosa et al. 2013). That we verify this class with experiments in which the dependency is obtained using conventional approaches (growth rate being modified with the utilization of different carbon sources; data obtained at steady state) supports our strategy. However, we also find three other patterns that differ. This does not seem associated with the approach itself, as our experimental characterization of these responses did not recover hyperbolic profiles. These genes could be somehow dependent on the growth nutrient or be perhaps subject to additional layers of regulation (e.g., differential expression under different sigma factors; Supplementary Figure S7). Moreover, the success in predicting the response of about 50% of promoters demonstrates the significance of the global program beyond balanced-growth (Berthoumieux et al. 2013).

Within those promoters exhibiting a Michaelis–Menten profile we distinguish subsets that reveal especially sensitive to growth rate. Could this strong sensitivity to physiology influence

gene location? We obtain that the most sensitive (constitutive) genes are selectively located in the chromosome. Indeed, genes with either linear or saturable profiles show larger sensitivities to growth rate within 20% of the replichore. This pattern is partially maintained when we control for the multiple replication fork effect, i.e., when we consider pa(μ) profiles instead of $\mathrm{PA}_{chr}(\mu)$. We thus propose a model in which multifork effects and the global program (excluding gene copy) work in combination; promoters that are most growth-rate dependent in *E. coli* benefit from a larger increase in gene expression at large growth rates (Supplementary Figure S8).

The fact that *E. coli* coordinates different mechanisms to obtain a multiplicative effect of enhanced expression of genes near *oriC* might not necessarily be a universal property of bacterial genomes. In fact, this precise coupling might have been selected in bacteria that are subject to variable growth rates, or bacteria that reach a large number of simultaneous replication rounds (for which multifork gene dosage fluctuations are strong). A new study found two fundamental bacterial reproduction strategies, the first relying on (metabolically) efficient but slow growth and a second that relies on inefficient but fast growth (Roller et al. 2016). Of the two strategies, the latter perhaps exploit the coordination of these mechanisms.

Recent studies show the importance between gene expression with genome architecture (Block et al. 2012; Bryant et al. 2014). In fact, the increase in gene dosage due to bacterial multifork replication appears as an additional control mechanism of natural genetic circuits (Slager and Veening 2016; Bar-Ziv et al. 2016). However, the relevance of genome organization goes beyond gene dosage fluctuations in fast growth (Sobetzko et al. 2012; Soler-Bistué et al. 2017), and it may be influenced by chromosomal structure (Sobetzko et al. 2012). Moreover, correlations between R and maximal growth rate with gene location conservation should be taken cautiously as known doubling times are biased by lab controlled environments (Gibson et al. 2018). Additional factors that affect gene distribution is gene essentiality (Rocha and Danchin 2003). In this context, our study presents physiology as an additional aspect to consider if we are to understand the organization of the bacterial genome.

METHODS

Promoter activity data and validating experiments. We obtained time series of optical density and promoter activity from public available data (Zaslaver et al. 2009) that used a library of *E. coli* promoters expressing a fast-folding fluorescent protein and cloned in a low-copy plasmid (Zaslaver et al. 2006). We considered only data of those experiments using minimal medium with 0.5%(w/v) glucose and supplemented amino acids, and did not include experiments in which the strains did not grow (within one standard deviation of the mean growth curve) or whose promoter activity was constant and equal to zero.

We used this very library (*E. coli* K-12 MG1655 strain) in our validation experiments. The reporter strains of genes *maoP, rsd, rpsT, prmB, rssB, amyA, yaaA, ghrA, nudF, cpsG, argQ* and *mutT* were retrieved from frozen stocks, plated, and grown overnight in LB agar. Isolated colonies were grown overnight in the specific medium, then diluted 1/20 and pre-cultured for ~5h. Then, 96-well flat transparent plates containing 190 µl of the specific medium were inoculated 1/20 with the pre-culture and added 50 µl of mineral oil to prevent evaporation. Optical density (600 nm) and fluorescence (535 nm) were assayed in a Victor x2 (Perkin Elmer) at 10min intervals for 8h (growth at 30ºC with shaking). Cultures were grown in M9 minimal medium with kanamycin (50 µg/ml) to which either glucose, arabinose, lactose, glycerol or maltose was added to a final concentration of 0.5%(w/v). The five carbon media were also supplemented in the second set of experiments with amino acids to a final concentration of 0.2%(w/v), thus making 10 different nutrient conditions in total.

Data processing and modeling. Growth rate time series were computed as the two-point finite differences of log(OD), $\mu(t) = \Delta \log(\mathrm{OD})/\Delta t$, and promoter activities were computed as the two-point finite difference in time of fluorescence per OD unit, $\mathrm{PA}_{pl}(t) = \Delta\mathrm{GFP}/\Delta t/\mathrm{OD}$. Balanced-growth data was computed from the mean time-series measurements of three technical replicates as the average value in a 2h time-window during observable exponential growth. Optical density and promoter activity are in arbitrary units that are proportional to cell number density and protein production per OD and time units, respectively.

Automatic clustering was performed with the normalized growth rates and promoter activities

(PA_{chr} and pa) with respect to their maximum. The Euclidean pairwise distance was used to compute the linkage matrix with the unweighted average distance, which was then automatically divided into 50 clusters of which were rejected those with less than 2% of the sampled promoters. Clusters were then grouped together by visual interpretation resulting in the four classes presented in the main text. Promoter activity profiles of cluster 1 (either from pa and PA_{chr}) were then fit to Eq.1 by means of non-linear least squares method. We fine-tuned the automatic classification and depending on the value estimated for K_m we distinguish between linear ($K_m > 3$ dbl/h), saturable ($0.1 < K_m < 3$ dbl/h) and constant ($K_m < 0.1$ dbl/h) profiles. The slope of linear profiles was obtained from a linear least squares fit. Figure 3C-D shows the running averages of the sensitivities in a window of 10 genes.

Regulatory network and gene annotation analysis. Information about genes, sigma factors, and the regulatory network was downloaded from RegulonDB (Gama-Castro et al. 2016).

Environmental variability, R and phylogenetic distances The classification of over 100 species depending on the variability of their environment used in this work was previously published (Parter et al. 2007). Classes of increasing environmental variability are obligate, specialized, aquatic, facultative, multiple and terrestrial. *E. coli* is found in the facultative class. From the original list, species with more than one chromosome, species that could not be found in the phylogenetic tree (see below), and species without a published unified genome in NCBI database were not considered in this study (Supplementary Table S3). The importance of multifork dosage increase due to multifork replication in a given species, termed R, is obtained as the ratio of chromosomal replication time by the minimal doubling time for each bacteria. In fact, it is proportional to the maximum number of overlapping replication rounds. Values of R and the minimal doubling time were retrieved for 60 of the 100 species (Couturier and Rocha 2006). Phylogenetic distances from *E. coli* were computed from the phylogenetic tree of Lang et al. (2013) (Supplementary Figure S9) and are measured in arbitrary units. These variables correlate with each other and specially with phylogenetic distance to *E. coli* (Supplementary Figure S10), which is why we corrected for phylogenetic inertia by using partial linear correlations.

Origins of replication and homology search. For the location of the origins of replication of most genomes, we used DoriC v7.0 (Gao et al. 2013) (a database of bacterial and archaeal genomes available at http://tubic.tju.edu.cn/doric/ – the update of September 15th, 2017). For genomes for which the origin of replication was not directly available, *Blochmannia Floridanus* and *Methanosarcina Acetivorans*, we used the web-tool that DoriC offers for its identification. The results had expected values $E = 0$ and $E = 3 \times 10^{-9}$, respectively (Supplementary Table S3). Moreover, the replication terminus *ter* was set half the genome length away from the origin of replication as is done in related works. We obtained the homolog sequences (and their location) of the 708 constitutive genes of *E. coli* from Blastp, results with expected values above $E = 10^{-3}$ were discarded (Pearson 2013). We quantified the conservation of the half most sensitive genes to growth rate, located at $m < 0.2$, for a given species as the probability of finding a smaller mean displacement in 10^4 random selections among all homologs found at $m < 0.2$ (independently of their sensitivity to growth rate). This protocol controls for a possible general conservation of genes near *oriC*, and for different numbers of homologs found in the set of species.

ACKNOWLEDGMENTS

This work was supported by PhD fellowship BES-2016-079127 (P.Y.) and grant FIS2016-78781-R (J.F.P.) from the Spanish Ministerio de Economía y Competitividad and the European Social Fund.

REFERENCES

Bar-Ziv, R., Voichek, Y., and Barkai, N. (2016). "Dealing with Gene-Dosage Imbalance during S Phase." *Trends in Genetics*, 32(11), 717–723.

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." *Molecular Systems Biology*, 9(1).

Block, D. H. S., Hussein, R., Liang, L. W., and Lim, H. N. (2012). "Regulatory consequences of gene translocation in bacteria." *Nucleic Acids Research*, 40(18), 8979–8992.

Bremer, H. and Dennis, P. (1996). “Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate.” *E Coli Salmonella Cell Mol Biol*, 2.

Bryant, J. A., Sellars, L. E., Busby, S. J. W., and Lee, D. J. (2014). “Chromosome position effects on gene expression in Escherichia coli K-12.” *Nucleic Acids Research*, 42(18), 11383–11392.

Cooper, S. and Helmstetter, C. E. (1968). “Chromosome replication and the division cycle of Escherichia coli Br.” *Journal of Molecular Biology*, 31(3), 519–540.

Couturier, E. and Rocha, E. P. C. (2006). “Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes.” *Molecular Microbiology*, 59(5), 1506–1518.

del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M., and Díaz-Orejas, R. (1998). “Replication and Control of Circular Bacterial Plasmids.” *Microbiology and Molecular Biology Reviews*, 62(2), 434–464.

Deris, J. B., Kim, M., Zhang, Z., Okano, H., Hermsen, R., Groisman, A., and Hwa, T. (2013). “The Innate Growth Bistability and Fitness Landscapes of Antibiotic-Resistant Bacteria.” *Science*, 342(6162), 1237435.

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., Castro-Mondragón, J. A., Medina-Rivera, A., Solano-Lira, H., Bonavides-Martínez, C., Pérez-Rueda, E., Alquicira-Hernández, S., Porrón-Sotelo, L., López-Fuentes, A., Hernández-Koutoucheva, A., Del Moral-Chávez, V., Rinaldi, F., and Collado-Vides, J. (2016). “RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond.” *Nucleic Acids Research*, 44(D1), D133–143.

Gao, F., Luo, H., and Zhang, C.-T. (2013). “DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes.” *Nucleic Acids Research*, 41(D1), D90–D93.

Gerosa, L., Kochanowski, K., Heinemann, M., and Sauer, U. (2013). “Dissecting specific and global transcriptional regulation of bacterial gene expression.” *Molecular Systems Biology*, 9, 658.

Gibson, B., Wilson, D. J., Feil, E., and Eyre-Walker, A. (2018). “The distribution of bacterial doubling times in the wild.” *Proc. R. Soc. B*, 285(1880), 20180789.

Gyorfy, Z., Draskovits, G., Vernyik, V., Blattner, F. F., Gaal, T., and Posfai, G. (2015). “Engineered ribosomal RNA operon copy-number variants of E. coli reveal the evolutionary trade-offs shaping rRNA operon number.” *Nucleic Acids Research*, 43(3), 1783–1794.

Kjeldgaard, N. O., Maaloe, O., and Schaechter, M. (1958). “The transition between different physiological states during balanced growth of Salmonella typhimurium.” *Journal of General Microbiology*, 19(3), 607–616.

Klumpp, S. and Hwa, T. (2008). “Growth-rate-dependent partitioning of RNA polymerases in bacteria.” *Proceedings of the National Academy of Sciences*, 105(51), 20245–20250.

Klumpp, S., Zhang, Z., and Hwa, T. (2009). “Growth Rate-Dependent Global Effects on Gene Expression in Bacteria.” *Cell*, 139(7), 1366–1375.

Kochanowski, K., Gerosa, L., Brunner, S. F., Christodoulou, D., Nikolaev, Y. V., and Sauer, U. (2017). “Few regulatory metabolites coordinate expression of central metabolic genes in Escherichia coli.” *Molecular Systems Biology*, 13(1), 903.

Kubitschek, H. E. (1974). “Constancy of the ratio of DNA to cell volume in steady-state cultures of Escherichia coli B-r.” *Biophysical Journal*, 14(2), 119–123.

Lang, J. M., Darling, A. E., and Eisen, J. A. (2013). “Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices.” *PLOS ONE*, 8(4), e62510.

Liang, S. T., Bipatnath, M., Xu, Y. C., Chen, S. L., Dennis, P., Ehrenberg, M., and Bremer, H. (1999). “Activities of constitutive promoters in Escherichia coli1.” *Journal of Molecular Biology*, 292(1), 19–37.

Morrison, P. F. and Chattoraj, D. K. (2004). “Replication of a unit-copy plasmid F in the bacterial cell cycle: a replication rate function analysis.” *Plasmid*, 52(1), 13–30.

Parter, M., Kashtan, N., and Alon, U. (2007). “Environmental variability and modularity of bacterial metabolic networks.” *BMC Evolutionary Biology*, 7, 169.

Pearson, W. R. (2013). “An Introduction to Sequence Similarity (“Homology”) Searching.” *Current*

protocols in bioinformatics, 03.

Peebo, K., Valgepea, K., Maser, A., Nahku, R., Adamberg, K., and Vilu, R. (2015). "Proteome reallocation in Escherichia coli with increasing specific growth rate." *Mol. BioSyst.*, 11(4), 1184–1193.

Rocha, E. P. C. and Danchin, A. (2003). "Essentiality, not expressiveness, drives gene-strand bias in bacteria." *Nature Genetics*, 34(4), 377–378.

Roller, B. R. K., Stoddard, S. F., and Schmidt, T. M. (2016). "Exploiting rRNA operon copy number to investigate bacterial reproductive strategies." *Nature Microbiology*, 1(11), 16160.

Schaechter, M., Maaloe, O., and Kjeldgaard, N. O. (1958). "Dependency on medium and temperature of cell size and chemical composition during balanced grown of Salmonella typhimurium." *Journal of General Microbiology*, 19(3), 592–606.

Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z., and Hwa, T. (2010). "Interdependence of Cell Growth and Gene Expression: Origins and Consequences." *Science*, 330(6007), 1099–1102.

Slager, J. and Veening, J.-W. (2016). "Hard-Wired Control of Bacterial Processes by Chromosomal Gene Location." *Trends in Microbiology*, 24(10), 788–800.

Sobetzko, P., Travers, A., and Muskhelishvili, G. (2012). "Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle." *Proceedings of the National Academy of Sciences*, 109(2), E42–E50.

Soler-Bistué, A., Timmermans, M., and Mazel, D. (2017). "The Proximity of Ribosomal Protein Genes to oriC Enhances Vibrio cholerae Fitness in the Absence of Multifork Replication." *mBio*, 8(1), e00097–17.

Weiße, A. Y., Oyarzún, D. A., Danos, V., and Swain, P. S. (2015). "Mechanistic links between cellular trade-offs, gene expression, and growth." *Proceedings of the National Academy of Sciences*, 112(9), E1038–E1047.

Zaslaver, A., Bren, A., Ronen, M., Itzkovitz, S., Kikoin, I., Shavit, S., Liebermeister, W., Surette, M. G., and Alon, U. (2006). "A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*." *Nature Methods*, 3(8), 623–628.

Zaslaver, A., Kaplan, S., Bren, A., Jinich, A., Mayo, A., Dekel, E., Alon, U., and Itzkovitz, S. (2009). “Invariant Distribution of Promoter Activities in Escherichia coli.” *PLOS Computational Biology*, 5(10), e1000545.

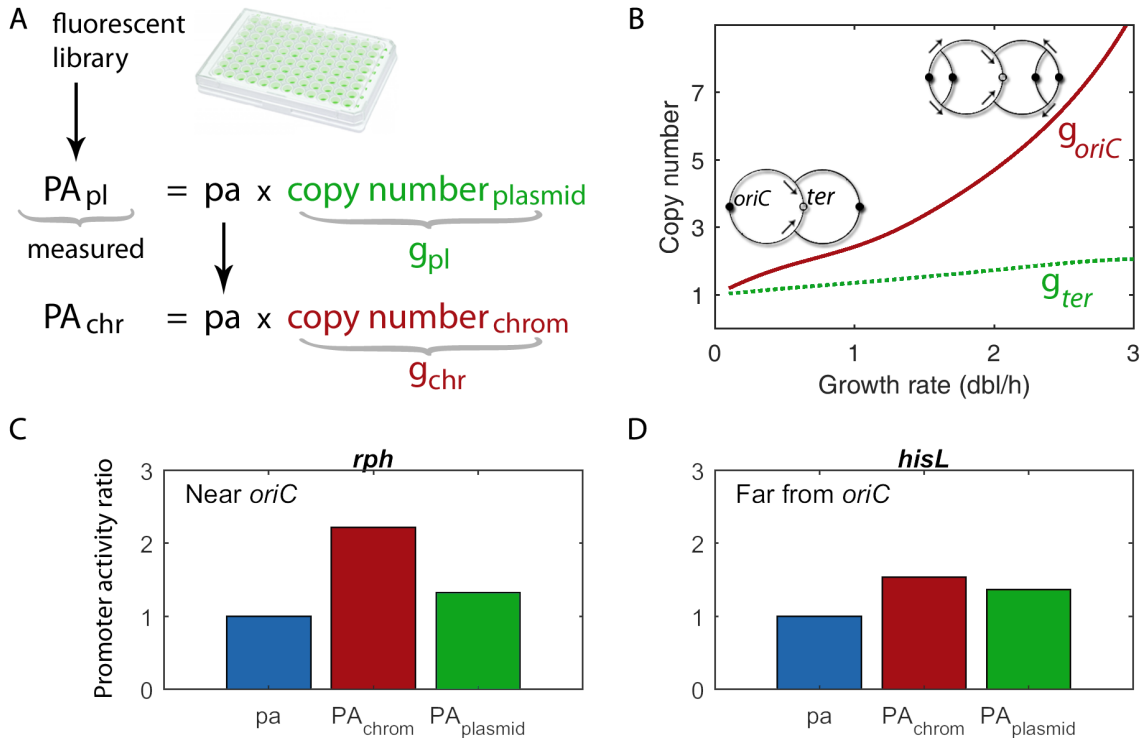

Fig. 1. Decoupling promoter activity from gene copy number. (A) Promoter activity per single gene copy, pa, can be obtained from experimental data of promoter activity quantified with a plasmid library, PA_{pl}, once the plasmid copy number g_{pl} is known. With this, one can calculate the promoter activity of a chromosomal gene, PA_{chr}, using Cooper and Helmstetter's model (Cooper and Helmstetter 1968). (B) Chromosomal multifork replication makes g_{chr} dependent on both growth rate and gene location in the chromosome. At a faster growth rate, the number of origins of replication *oriC*s (red solid line and black dots) increases due to the overlap in time of multiple replication rounds. Arrows show the direction of replication forks. In the case of plasmids with low copy number, as the one used in the plasmid library (pSC101), g_{pl} is proportional to the number of terminal regions (*ter*s) in the chromosome (green dotted line). (C,D) Relative differences in promoter activity (pa, PA_{chr}, PA_{pl}) for two genes at different chromosomal locations for a fixed growth rate (normalized to the corresponding pa). Genes (*rph* and *hisL*) are located at distances $m_{rph} = 0.04$ and $m_{hisL} = 0.80$ from *oriC*. Observe that the correction in promoter activity is larger for genes near *oriC*. PA_{pl} was obtained in balanced growth at a rate $\mu \simeq 0.9$ dbl/h (Zaslaver et al. 2009).

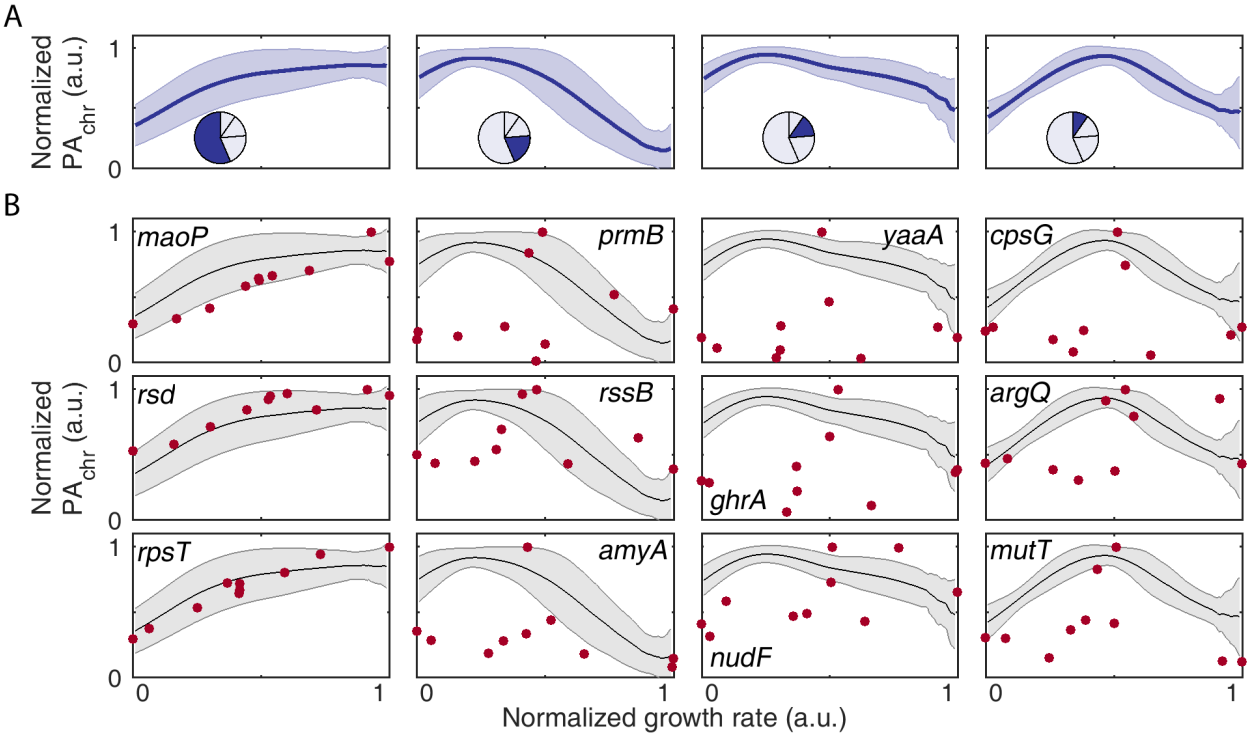

Fig. 2. Chromosomal promoter activities of constitutive genes follow different growth rate dependencies. (A) Typical normalized profile of each class (solid line) and standard deviation (shaded). Inset: pie chart with the fraction of promoters found in each class. Note that these classes were obtained from time series on a single growth medium (Zaslaver et al. 2009). (B) We validated previous classes with experimental measurements of 12 promoter activities at balanced growth in 10 different growth media. The typical profile of their corresponding classes is shown in black solid line and grey shaded area.

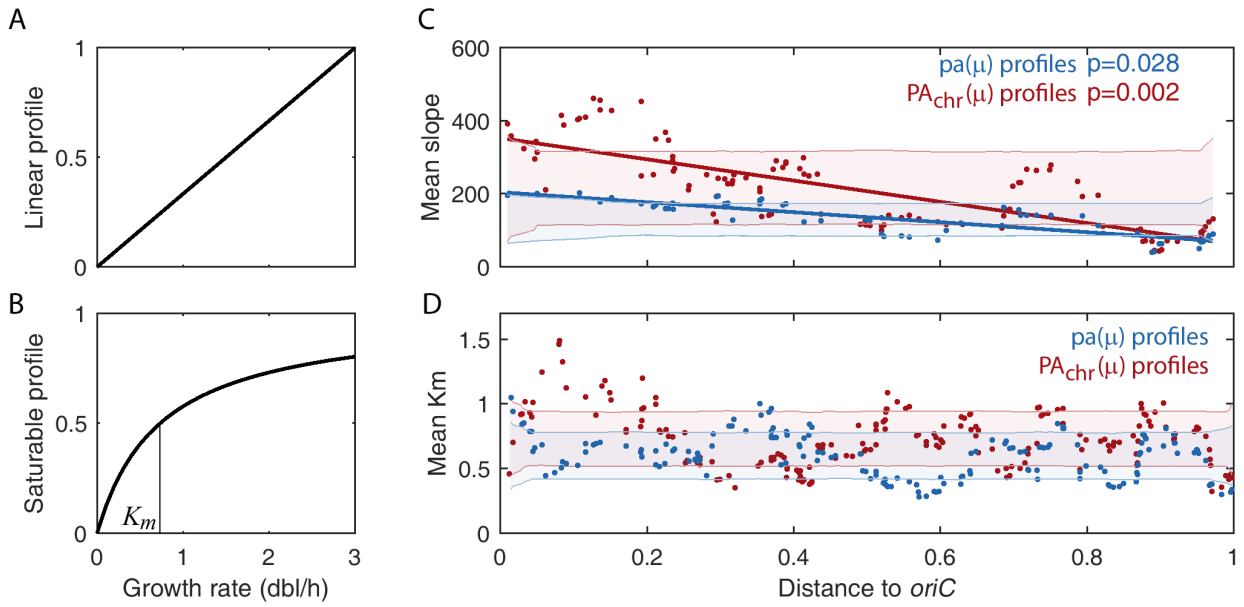

Fig. 3. Promoters that are most sensitive to growth rate are located closer to the origin of replication. Two profiles of promoter activity can be identified: linear (A) and saturable (B). Sensitivity to the global program is proxied by the slope in the case of linear profiles, and the growth rate at which activity is half maximal (K_m) in the case of saturable profiles. (C,D) Running averages of the sensitivity to the global program of promoters with linear and saturable profiles respectively of PA_{chr} (red dots) and pa (blue dots). The sensitivity of linear profiles decreases linearly with the distance to *oriC* (solid lines); this pattern is only significantly observed in saturable profiles when considering PA_{chr}. Shading denotes one standard deviation of sensitivities obtained from a permutation test with 10^4 randomizations.

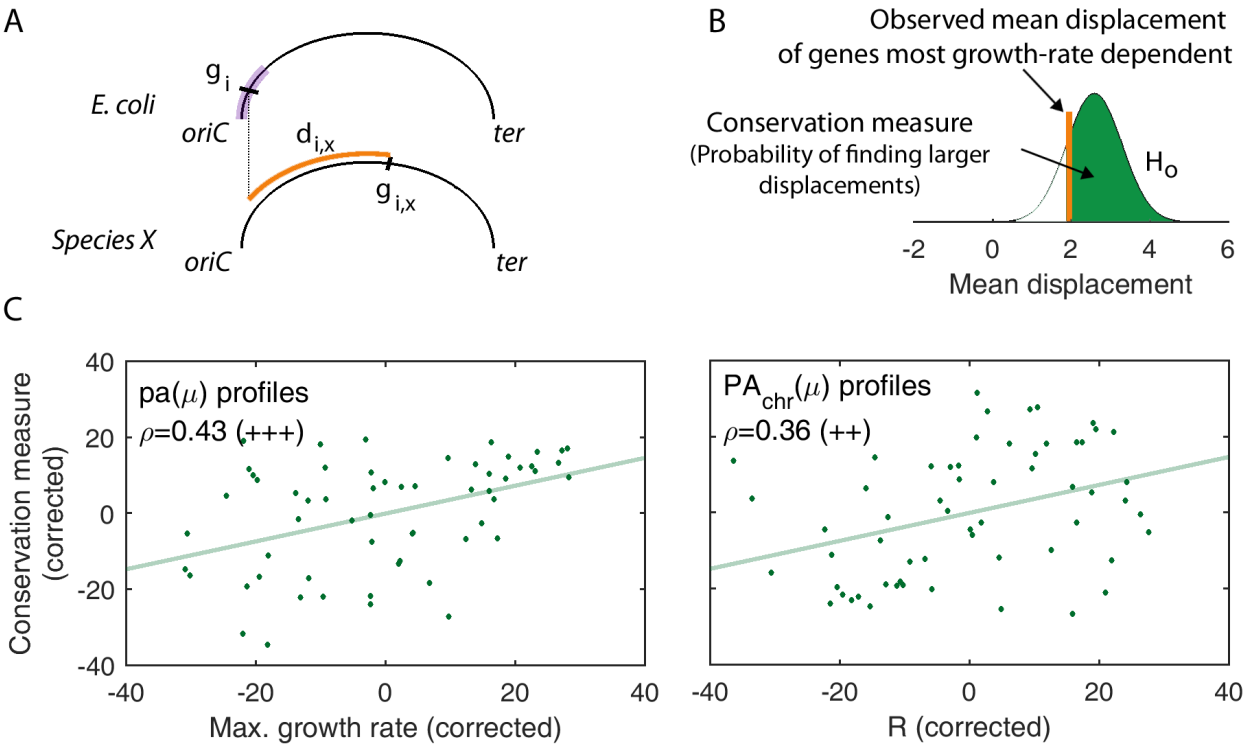

Fig. 4. Conservation of linear constitutive genes near oriC that are most dependent on the global program correlates with the maximum growth rate and R. (A) Gene's location conservation is computed from the displacement of a gene in *E. coli* (g_i) within the $m < 0.2$ region (purple), with respect to its homolog in other species ($g_{i,X}$). (B) In every species, the observed mean displacement of genes that are most dependent on the global program and are located at $m =< 0.20$ is tested against the displacement of the rest of constitutive genes at $m =< 0.20$ (H_o). (C) The most predictive partial correlations (Spearman ρ, and light green solid line) of the conservation measure of the half most growth-rate-dependent lineal profiles near *oriC* in *E. coli* were obtained with R, for $PA_{chr}(\mu)$ profiles, and the maximum growth rate, for pa(μ) profiles. Variables are corrected for phylogenetic inertia (Methods). Significance of correlations: (++) p<0.01 and (+++) p<0.001.

- Supplementary Information -
Large-scale analysis of the global transcriptional regulation of bacterial gene expression

Pablo Yubero and Juan F Poyatos

Logic of Genomic Systems Laboratory, CNB-CSIC, Madrid, Spain

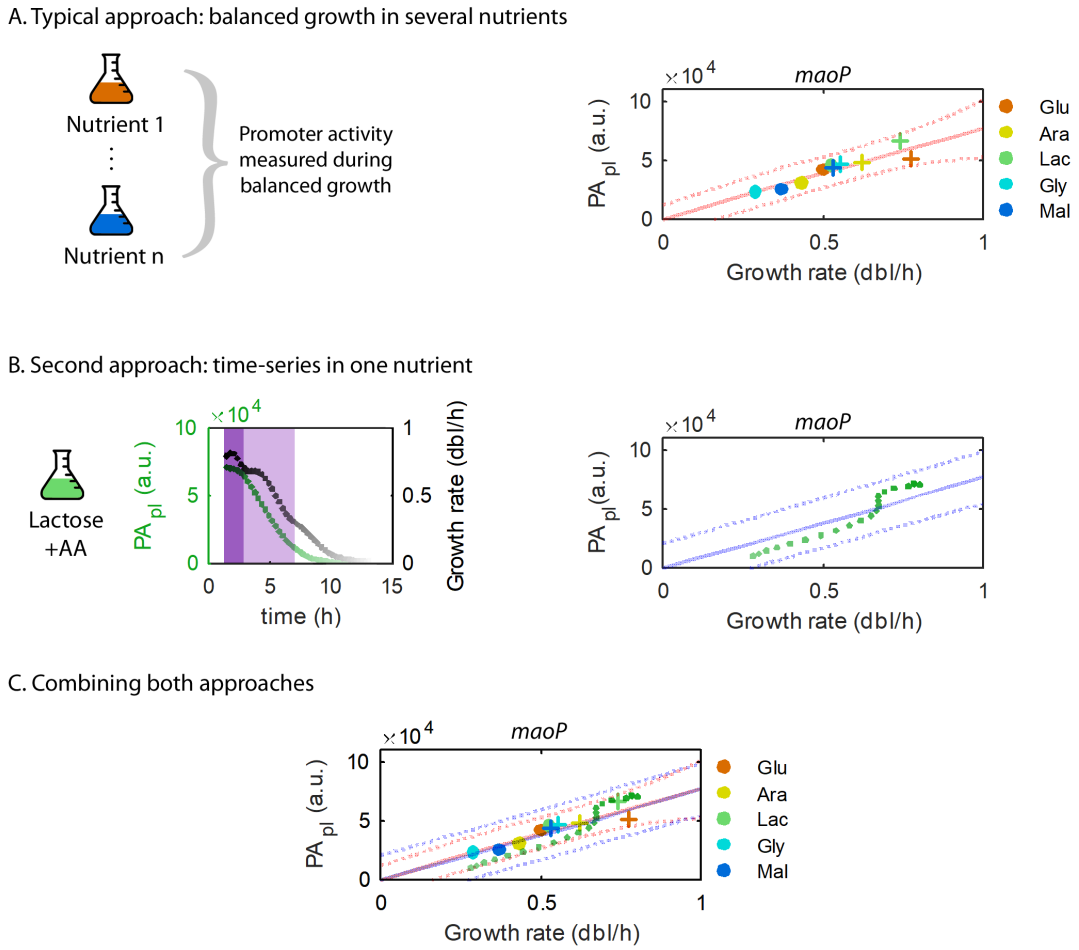

Figure S1. Comparison between two different approaches to compute promoter activity profiles. (A) The traditional approach consists in obtaining data pairs (PA_{pl}, μ) from cultures during balanced growth in a variety of growth media. Experimental profile obtained for the promoter of gene *maoP* in minimal medium with five different carbon sources: glucose (orange), arabinose (yellow), lactose (green), glycerol (light blue), maltose (dark blue) and supplemented, or not, with amino acids (AA; crosses and circles respectively; Methods). The best fit to Eq.(1) is also shown with the 95% confidence interval (red solid and dotted lines respectively). (B) We can also obtain promoter activity profiles from growth time-series in a single growth medium (e.g. lactose+AA). We consider promoter activity (in green) and growth rate (in black) points during early and late exponential phase (dark and light purple shade) to obtain a similar promoter activity profile (Methods). The best fit to Eq.(1) and the 95% confidence bounds are also shown (blue solid and dotted lines respectively). (C) Superposition of the two profiles obtained previously by different means. These two approaches yield similar qualitative and quantitative profiles in this and other cases, for promoters of genes in class 1 (Supplementary Figures S4).

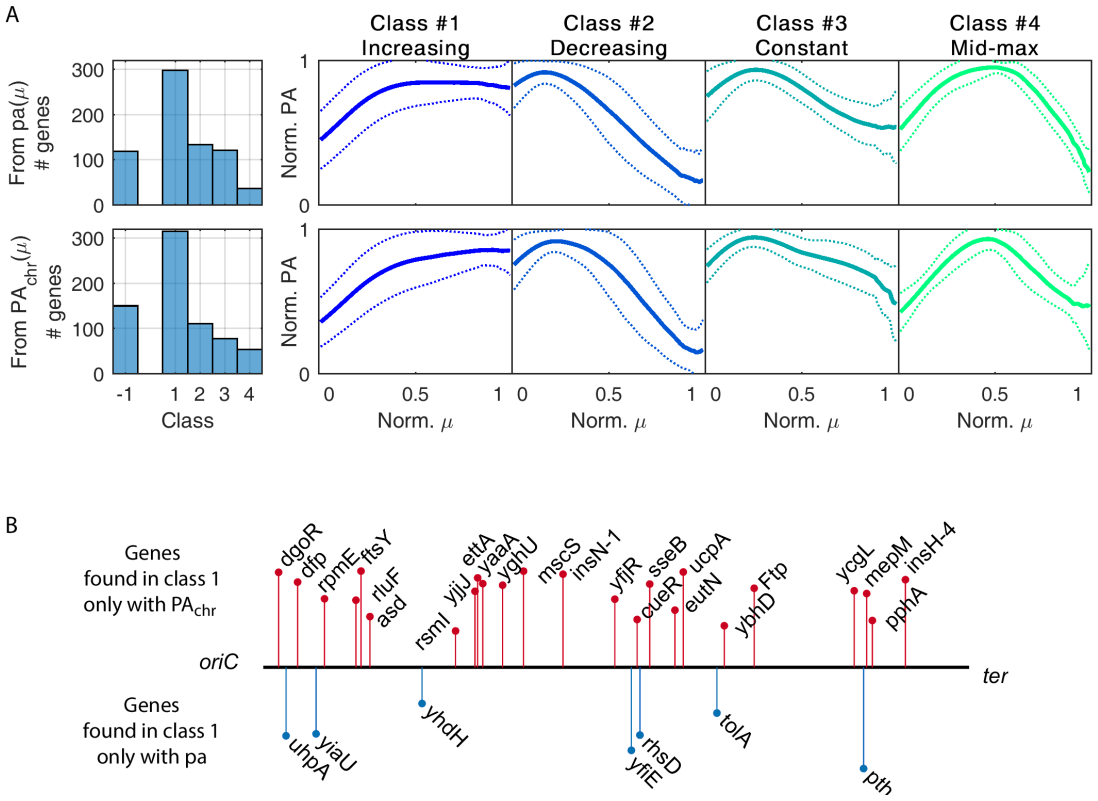

Figure S2. Comparison of the results of the automatic clustering algorithm when considering pa(μ) and PA$_{chr}$(μ) profiles. (A) Histograms with the number of promoters found in each class (where class -1 are unclassified promoters) and mean profiles (solid lines) with one standard deviation (dotted lines). (B) Differences in the composition of class 1 when considering pa(μ) and PA$_{chr}$(μ) profiles. Although 291 promoters are found in both cases, here we show the 24 (and 7) promoters, and their location, that are only found when considering PA$_{chr}$(μ) (and pa(μ)) profiles in red (and blue). Different heights are used for clarity.

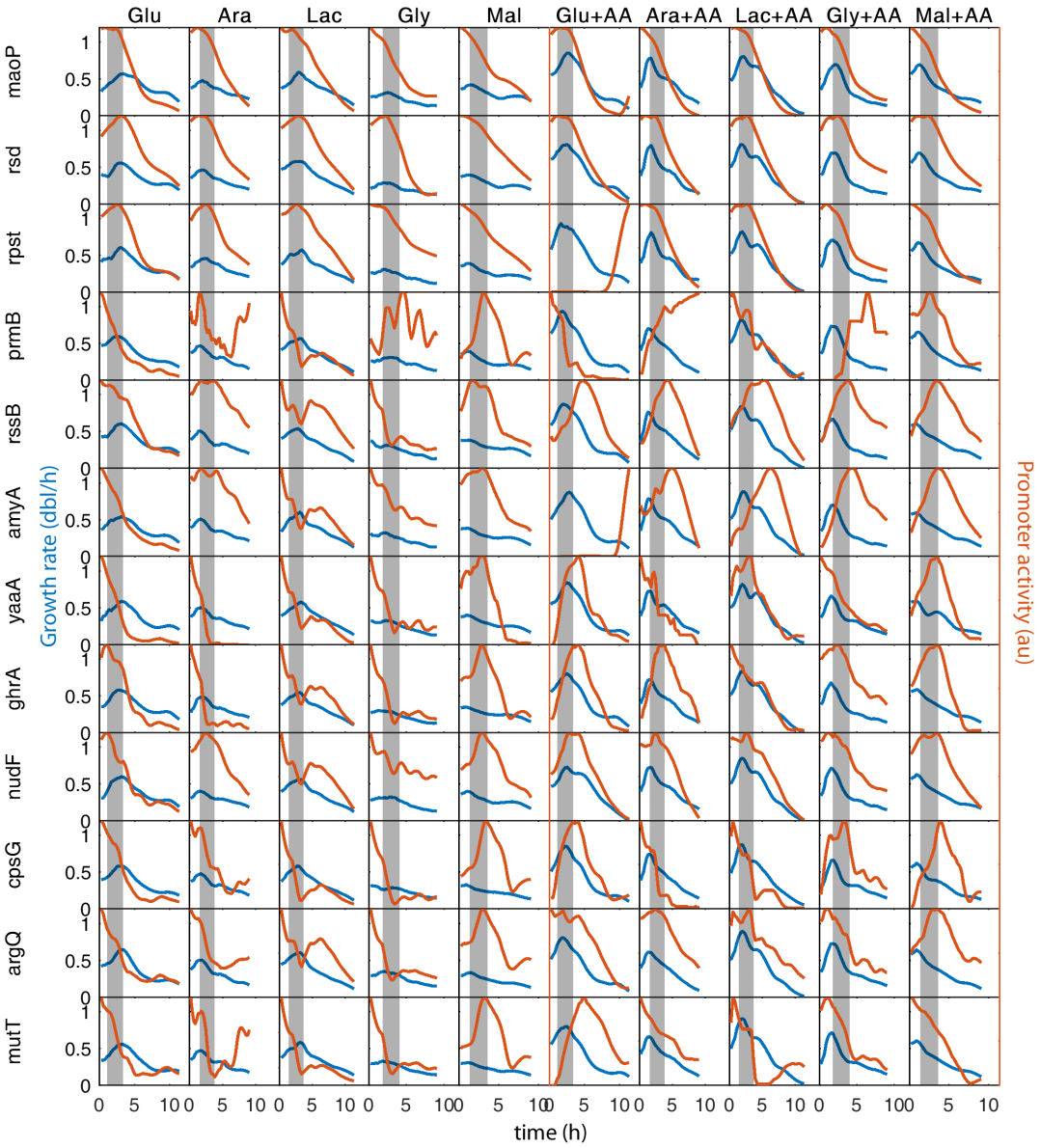

Figure S3. Time-series of the experimental measurements of growth rate (left y-axis, blue lines) and promoter activity from a plasmid reporter (right y-axis, orange lines). Three genes chosen from each class (rows), *maoP*, *rsd* and *rpsT* from class 1; *prmB*, *rssB* and *amyA* from class 2; *yaaA*, *ghrA* and *nudF* from class 3; *cpsG*, *argQ* and *mutT* from class 4), were assayed in ten different growth conditions (columns; Methods) for about 8h (x-axis). The grayed area highlights the 2h window of observed exponential phase.

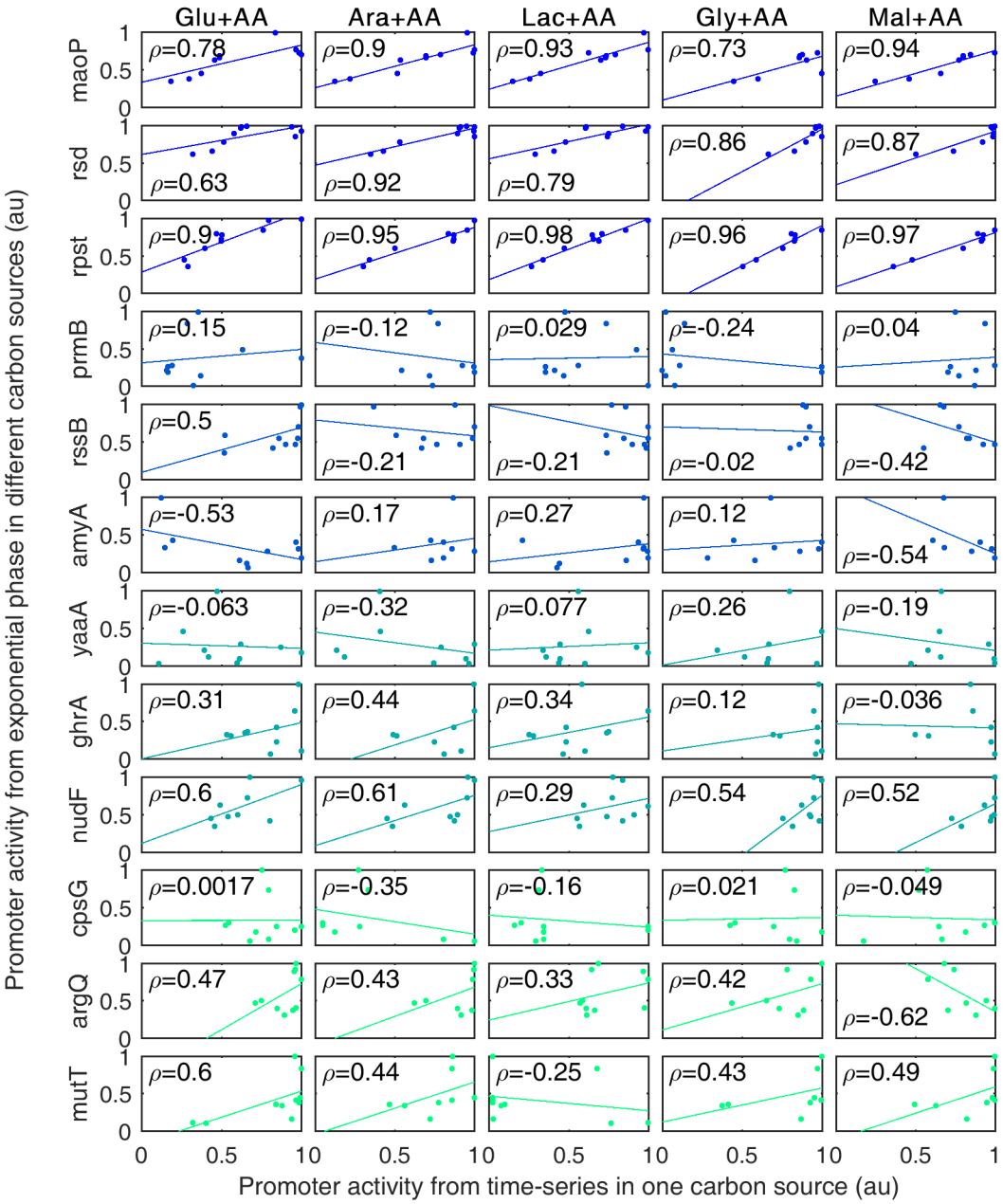

Figure S4. Correlation between $PA_{pl}(\mu)$ measured from exponential phase in 10 different growth media (y-axis) and from time-series growth in one carbon source (x-axis). We tested three genes of each class (rows; color indicates the class as in Supplementary Figure S1) in 5 different carbon sources (columns). Each point represents a growth rate. Promoter activities from time-series data are obtained by linear interpolation at the corresponding growth rates.

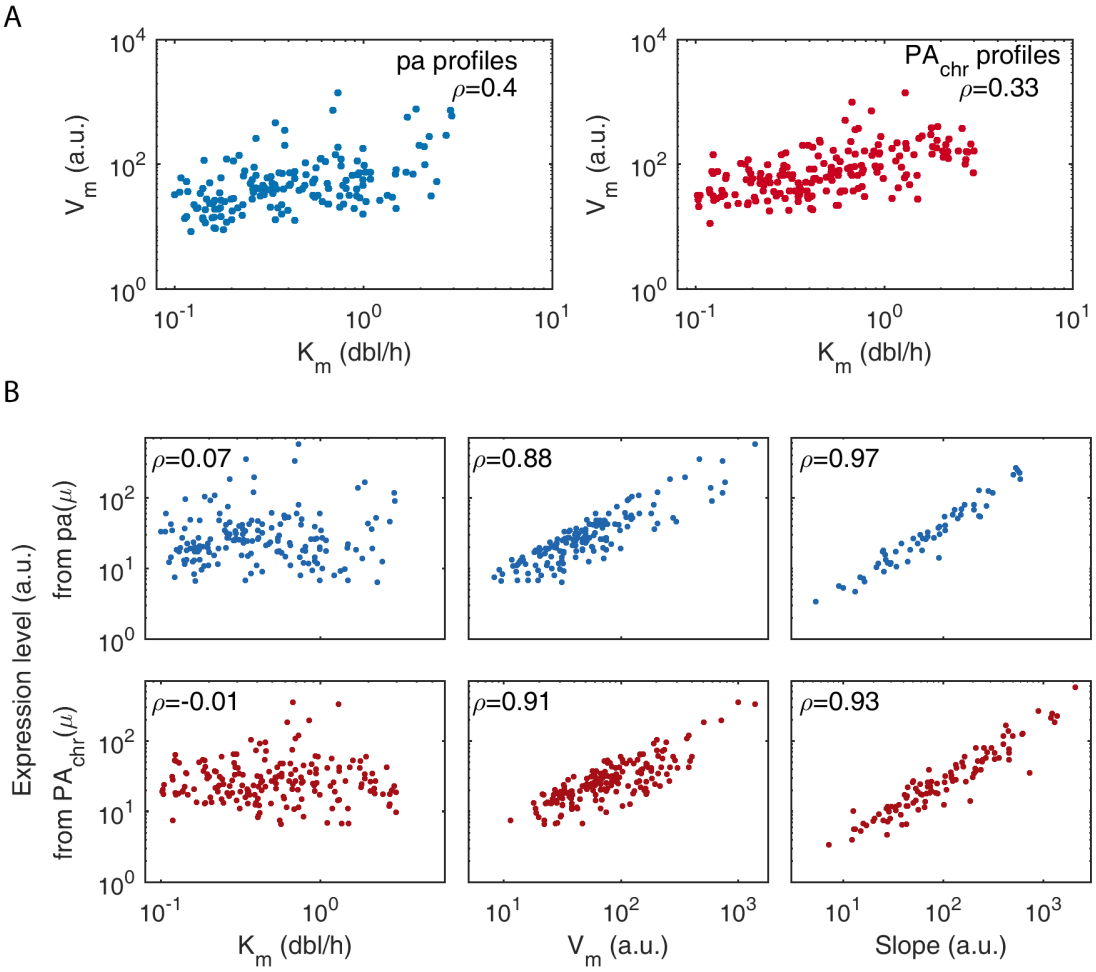

Figure S5. Correlation between the parameters obtained from fitting pa(μ) and PA$_{chr}$(μ) to Eq.(1). (A) There is a negligible correlation between the strength of a saturable promoter (V_m) and its sensitivity to growth rate (K_m). (B) The expression level of a promoter, measured as its activity at μ = 0.5 dbl/h, does not correlate with the sensitivity of saturable promoters K_m. However, expression level correlates well with the maximum activity of saturable promoters, and with the slope of lineal promoters. In all panels ρ is Pearson's linear correlation coefficient, and a log-log scale is used for clarity as the values shown span several orders of magnitude.

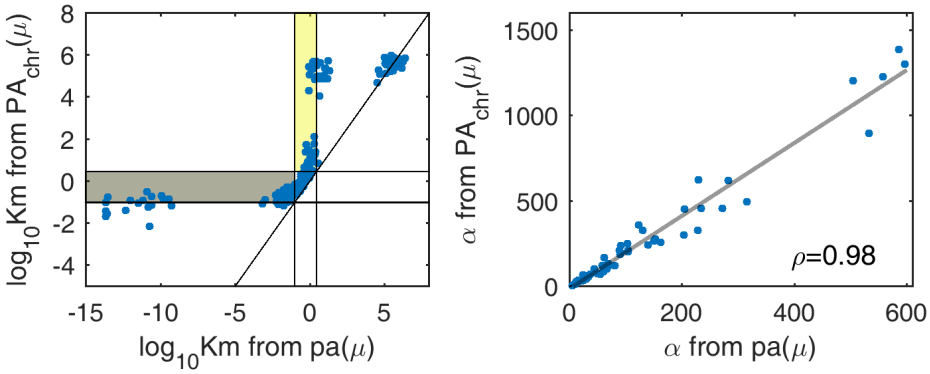

Figure S6. Different sensitivities are obtained from pa(μ) and PA$_{chr}$(μ) profiles. (A) Scatter of the parameter K_m of the 291 promoters belonging to class 1 when computed from pa(μ) and PA$_{chr}$(μ). The region where promoters whose pa is constant (saturable) but whose PA$_{chr}$ is saturable (linear) is marked in grey (yellow). A log-log scale is used for the sake of clarity as the parameters cover several orders of magnitude. (B) Slopes of linear profiles computed from pa and PA$_{chr}$ correlate linearly. Observe that in general, the addition of the chromosomal multifork replication effects increases the sensitivity to the global regulation program of constitutive promoters of class 1.

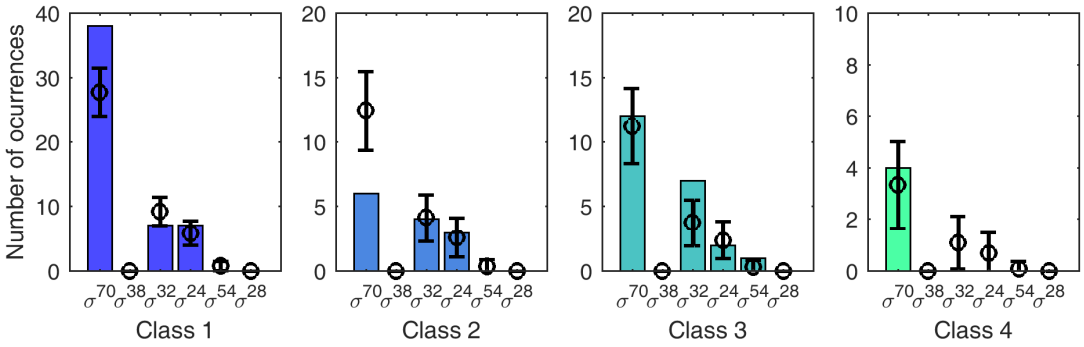

Figure S7. Histograms of the number of constitutive genes in each class whose expression is modulated by a sigma factor. Sigma factors modulate the selectivity of the RNAP for certain promoters under different situations: housekeeping (σ^{70}), general stress (σ^{38}), cytoplasmic stress (σ^{32}), extracytoplasmic stress (σ^{24}), nitrogen stress (σ^{54}) and flagellar genes (σ^{28}). Here we show the number of observed genes (vertical bars) within each sigmulon (x-axis) and the expected values under 1e4 randomizations (black circles, error bars correspond to one standard deviation).

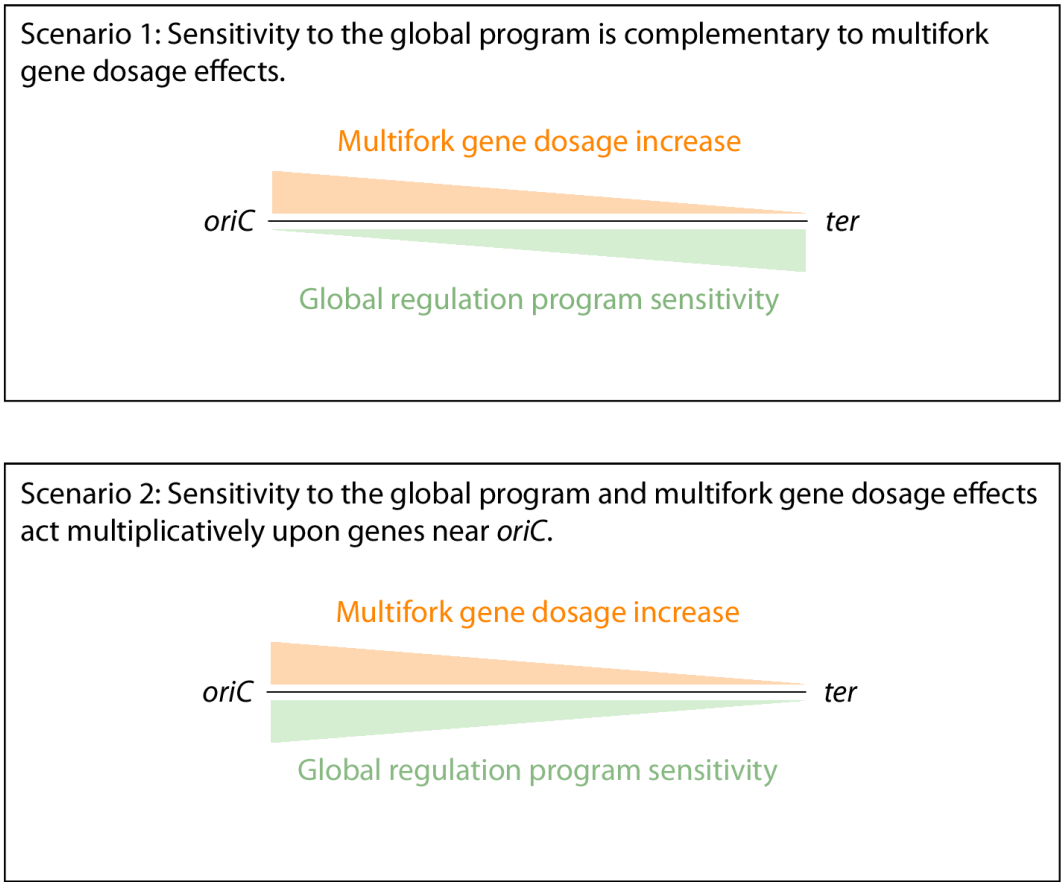

Figure S8. Two possible scenarios for the effect of multifork gene dosage as part of the global program. (Top) Scenario 1. The most sensitive genes to the global program locate near the *ter* region (symbolized by a green gradient; note that global regulation here excluded the impact of gene dosage). In contrast, the multifork effect is stronger near *oriC* (orange gradient). The gene dosage effect is consequently not coupled to a strong sensitivity. (Bottom) Scenario 2. The most sensitive genes to the global regulation (excluding gene dosage) locate in the *oriC* region. In this case, the effect of gene dosage is linked to the strong sensitivity.

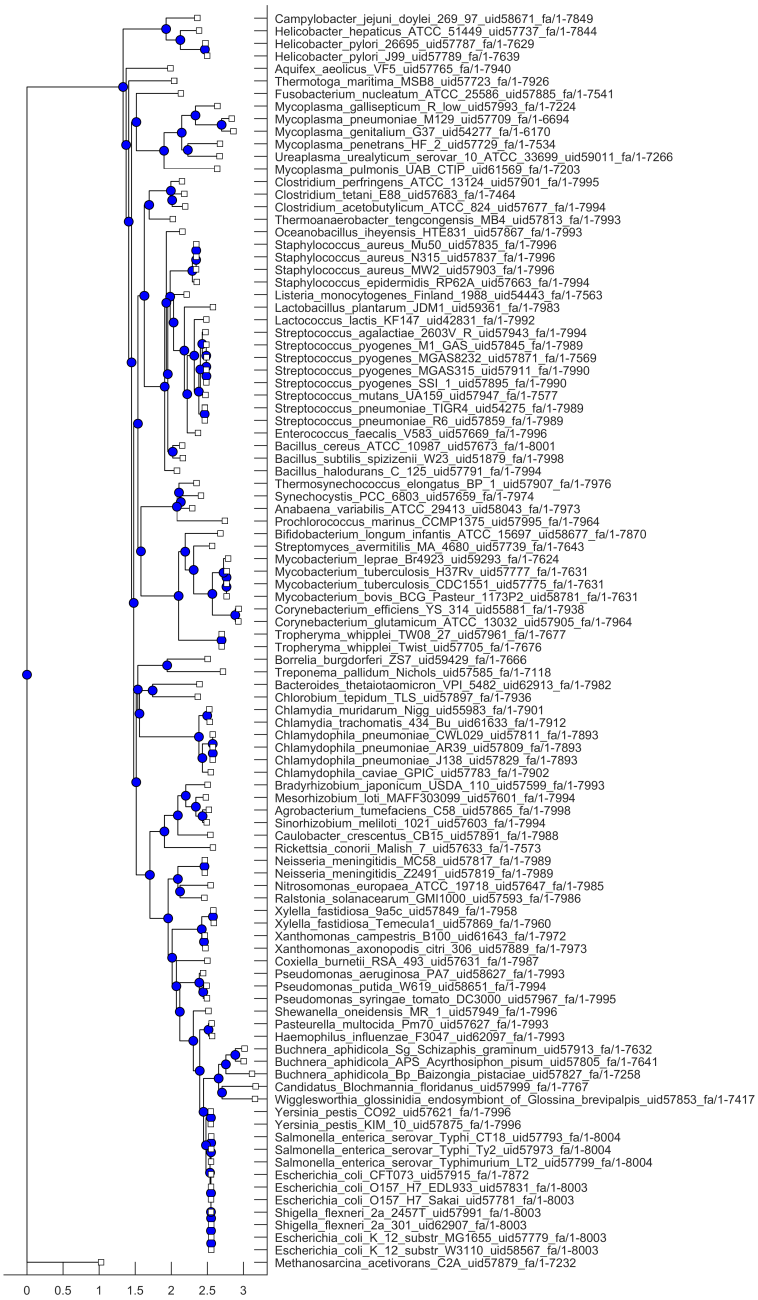

Figure S9. Phylogenetic tree of the 100 species used in this work.

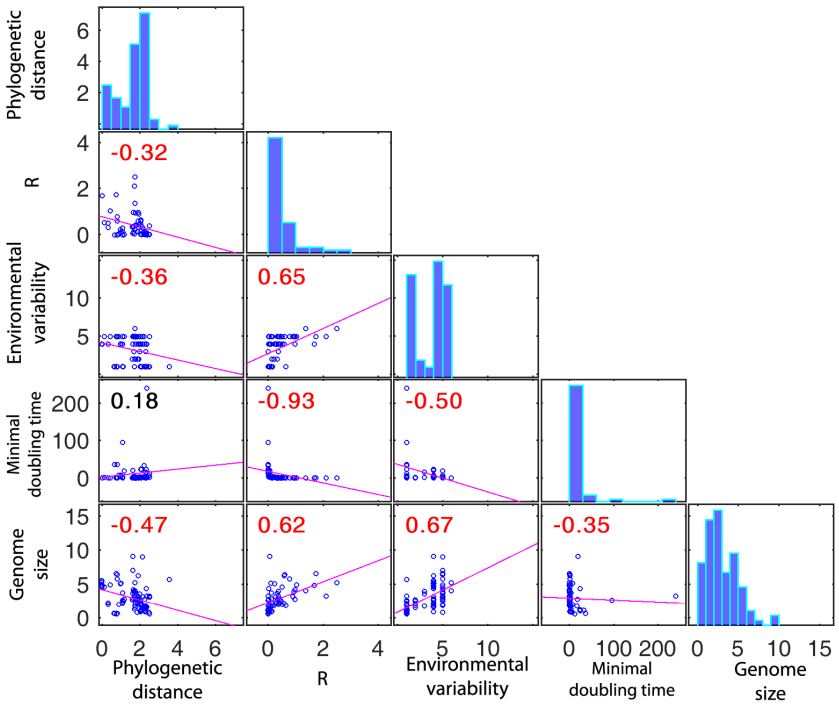

Figure S10. Spearman rank correlations between possible explanatory variables for conservation measures. All correlations between phylogenetic distance, R (i.e. the relevance of multifork effects), environmental variability, minimal doubling time and genome size in Mbp (Methods). Values in red denote significant correlations ($p<0.05$). Histograms of the values are shown in the diagonal.