Limits to a classic paradigm: most transcription factors in E. coli regulate genes involved in multiple biological processes

Journal Article

Author(s):
Ledezma Tejeida, Daniela Elizabeth; Altamirano-Pacheco, Luis; Fajardo, Vicente; Collado-Vides, Julio

Publication date:
2019-07-26

Permanent link:
https://doi.org/10.3929/ethz-b-000358034

Originally published in:
Nucleic Acids Research 47(13), https://doi.org/10.1093/nar/gkz525

6656–6667 *Nucleic Acids Research, 2019, Vol. 47, No. 13* *Published online 13 June 2019*
doi: 10.1093/nar/gkz525

Limits to a classic paradigm: most transcription factors in *E. coli* regulate genes involved in multiple biological processes

Daniela Ledezma-Tejeida[1,2,*], Luis Altamirano-Pacheco[1], Vicente Fajardo[1] and Julio Collado-Vides[1,3,*]

[1]Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico, [2]Department of Biology, Institute of Molecular Systems Biology, ETH Zürich, Zurich, Switzerland and [3]Department of Biomedical Engineering, Boston University, Boston, MA, USA

Received April 19, 2019; Revised May 29, 2019; Editorial Decision May 31, 2019; Accepted June 04, 2019

ABSTRACT

Transcription factors (TFs) are important drivers of cellular decision-making. When bacteria encounter a change in the environment, TFs alter the expression of a defined set of genes in order to adequately respond. It is commonly assumed that genes regulated by the same TF are involved in the same biological process. Examples of this are methods that rely on coregulation to infer function of not-yet-annotated genes. We have previously shown that only 21% of TFs involved in metabolism regulate functionally homogeneous genes, based on the proximity of the gene products' catalyzed reactions in the metabolic network. Here, we provide more evidence to support the claim that a 1-TF/1-process relationship is not a general property. We show that the observed functional heterogeneity of regulons is not a result of the quality of the annotation of regulatory interactions, nor the absence of protein–metabolite interactions, and that it is also present when function is defined by Gene Ontology terms. Furthermore, the observed functional heterogeneity is different from the one expected by chance, supporting the notion that it is a biological property. To further explore the relationship between transcriptional regulation and metabolism, we analyzed five other types of regulatory groups and identified complex regulons (i.e. genes regulated by the same combination of TFs) as the most functionally homogeneous, and this is supported by coexpression data. Whether higher levels of related functions exist beyond metabolism and current functional annotations remains an open question.

INTRODUCTION

Transcription factors (TFs) are important drivers of bacterial decision-making. They convey environmental cues into the gene expression machinery by binding to specific metabolites. In turn, they promote the recruitment, or dismissal, of the RNA polymerase in a defined set of promoters. Classical examples, such as *lacI* (1) or *trpR* (2), introduced the notion that individual TFs mediate defined responses, such as lactose utilization or tryptophan biosynthesis. The assumption followed that all the genes regulated by the same TF (termed regulons (3,4)) were involved in the same biological process. There are now various examples of widely studied TFs that are involved in more than one biological process, but they are considered special cases, termed 'global regulators' (5), and the wide variety of processes in which they take part is easily rationalized by the large sizes of these regulons. Moreover, it is common to refer to TFs by the biological process they are involved in, such as 'regulator of response to oxidative stress.'

The increasing availability of data on regulatory interactions and gene expression has allowed the development of several algorithms that predict new regulatory interactions (6–8), new gene functions (6,9,10), or describe modules of coregulated genes (11,12). Given the complexity of the data, the methods rely on several assumptions, including: (i) genes regulated by the same TF will be involved in the same biological process and (ii) genes regulated by the same TF are expected to be coexpressed. These suppositions are mostly used to either enrich the results with true positives or evaluate the efficiency of the method. However, there has been no systematic study to date that quantitatively analyzed whether this is a general property of TFs or an attribute of a few classical examples.

We have shown before that local TFs have a gradient of functional homogeneity and that less than one-fourth of TFs have a one-to-one correspondence with biological pro-

*To whom correspondence should be addressed. Tel: +41 44 633 7839; +52 777 313 9877; Fax: +52 777 313 9877; Email: dledezma@lcg.unam.mx
Correspondence may also be addressed to Julio Collado-Vides. Email: collado@ccg.unam.mx

cesses in terms of the connectivity of the metabolic subnetworks directly affected by the regulatory action of the TF (13). Here, we show that the observed gradient is not a result of missing metabolite-protein interactions, or low-confidence interactions, and is different from what would be expected by chance. We provide support that the gradient is a biological property by reassessing the functional homogeneity of each regulon using Biological Process Gene Ontology (GO) terms. In order to find an alternative to general regulons (i.e. genes regulated by the same TF irrespective of effect or additional regulation) for predictive algorithms, we explore four other types of regulons and quantify their functional complexity by two different methods. Our results indicate that complex regulons, defined as a group of genes regulated by the same combination of TFs, are the most functionally homogeneous type of regulons. Finally, we measure the coexpression of genes in each type of regulon and show that, consistent with our results, complex regulons are the most highly coexpressed.

MATERIALS AND METHODS

GENSOR unit assembly

GENSOR units were assembled using the semiautomatic pipeline reported in (13). Data from RegulonDB v10.0 (14) were obtained from two custom-made datasets available at GitHub and the TF-transcription unit (TF-TU) and TU-genes datasets available at the RegulonDB website. All data from EcoCyc (15) were obtained using the Pathway Tools software v22.0 (16), the Perlcyc API, and custom Perl scripts. Canonical metabolic pathways and enzymatic regulatory interactions were obtained from EcoCyc. Enzymatic regulatory interactions were only added to a GENSOR unit if the regulatory metabolite and the regulated enzyme were already present in the GENSOR unit. A modified version of the pipeline was used to assemble GENSOR units using as input a custom group of genes.

Connectivity

Connectivity was calculated in three steps: (i) all enzymes in the GENSOR unit were identified. (ii) All metabolic fluxes in the GENSOR unit were identified by connecting reactions that shared a metabolite, irrespective of it being a substrate or a product. (iii) Enzymes participating in metabolic fluxes of more than one reaction were identified and termed 'connected enzymes.' (iv) The following formula was applied:

$$C = \frac{E_c}{E_t + (M_f - 1)}$$

where E_c is the number of connected enzymes (calculated in iii), E_t is the total number of enzymes in the GENSOR unit (calculated in i) and M_f is the number of metabolic fluxes within the GENSOR (calculated in ii). Since the expectation is that all enzymes are involved in one metabolic flux, any extra metabolic fluxes are penalized in the denominator (see Appendix S1). Essentially, connectivity reflects the fraction of enzymes in the GENSOR unit that cooperate with others in a metabolic flux (i.e. a pathway), penalized by the number of unexpected fluxes. Enzymatic regulation is considered in the metric by additionally labeling as 'connected enzymes' any enzyme that is regulated by a metabolite also present in the GENSOR unit. Final calculations were performed using a custom Perl script available at GitHub.

Metabolic pathways, TUs and gene ontology terms

Genes belonging to metabolic pathways were automatically retrieved using Pathway Tools v22.0. TUs were obtained from RegulonDB downloadable datasets, and those with only 1 gene were eliminated. Gene Ontology (GO) terms (17,18) were obtained from the PortEco Filtered Annotation File version 24/05/2017, available at the AmiGO website. Members of a GO term were expanded to include the genes that directly belong to it, plus all the genes that belong to its children terms. All analyses were performed using only the Biological Process branch of the ontology.

Identification of dominant GO terms

We obtained the fraction of genes in a GENSOR unit present in each of the 2860 Biological Process (BP) GO terms. The GO term with the highest fraction of genes was selected as the dominant GO term.

In case of ties, the most specific term (farthest from the root) was selected. Only GENSOR units with more than one gene annotated in the BP branch of the ontology were considered in the analysis. Genes that were not annotated in at least one term of the BP branch were excluded from the analysis. Genes that belonged to multiple GO terms were considered in all their terms.

Regulatory groups and regulons

Regulatory groups were defined as depicted in Table 2. Regulons were identified through a custom Perl script using information from RegulonDB in the following manner (see also Supplementary Figure S9):

- *General regulons.* One regulon per TF; includes all the genes that have at least one annotated binding site for the TF in their promoter region.
- *Strict regulons.* One or two regulons per TF; includes the subset of genes in a general regulon that are regulated under the same effect (activation/repression).
- *Simple regulons.* Zero or one regulon per TF; includes all the genes with identified binding sites for only one TF in their promoter region.
- *Complex regulons.* Zero, one or more regulons per TF; includes all the genes that have annotated binding sites for the same combination of TFs in their promoter region.
- *Conformation regulons.* Zero, one or more regulons per TF; includes all the genes that have annotated binding sites in their promoter region for the same functional conformation of a TF, either in complex with a metabolite (*holo* conformation) or by itself (*apo* conformation).
- *Conformation + effect regulons.* Zero, one or more regulons per TF; includes the subset of genes in a conformation regulon that are regulated under the same effect (activation/repression).

Global TFs (ArcA, CRP, IHF, Fis, FNR, HNS, Lrp) were considered in all regulatory groups, except general regulons, since by definition (5) they are involved in more than one biological process and could introduce bias to our results. GENSOR units based on regulatory effect of the TF were assembled using as starting point all the genes in any regulon of the above-mentioned regulatory groups, except for complex regulons. Genes were separated into two groups: repressed and activated genes. Genes known to be dually regulated by the same TF were considered in both groups. GENSOR units were assembled for each subgroup. Complex regulons were classified in activated or repressed depending on the effect of all the TFs involved in the regulon. For example, the regulon CRP(–)/AraC(–)/XylR(–) would be classified as repressed. Complex regulons where TFs had different regulatory effects (e.g. CRP(-)/AraC(+)/XylR(–)) were omitted from the analysis. All regulons were calculated from RegulonDB datasets using a custom Perl script available at GitHub.

Randomization of regulons

Sets of random regulons were created using a custom Perl script. For each regulatory unit, the script created a list of the genes that belonged to all of its regulons and then recreated each regulon by assigning random genes from the list until the regulon reached its original number of genes. Assignment of random genes allowed repetition. The process was repeated 100 times for each regulatory unit, in order to obtain 100 sets of random regulons (Appendix S2). Random regulons for connectivity analysis were created with the same algorithm, with the exception that the number of enzymes was maintained, as opposed to number of genes. Enzymes were identified using Pathway Tools v22.0.

Coexpression of regulons

Coexpression analysis was performed using expression data from the COLOMBOS compendia (19) across 4077 microarray contrasts. The Spearman correlation of gene expression across all conditions was computed for all possible combinations of gene pairs in a regulon. The median of the obtained Spearman correlations was calculated and used as the representing coexpression value of the regulon. Regulons with less than two genes were excluded from the analysis.

RESULTS

Assembly of GENSOR units and calculation of connectivity

We used the GENSOR unit framework (13) to further analyze the relationship between regulons and the metabolic effects of their gene products. A general regulon was defined as the group of genes directly regulated by a TF, regardless of the effect (positive, negative, or dual) of the TF and regulation from other TFs. For each TF, we automatically retrieved from RegulonDB (14) its known effectors, active and inactive conformations, regulated genes, and the effects of the regulatory interactions. From EcoCyc (15), we retrieved the gene products and any protein complex that belonged to a GENSOR unit. If the gene products were enzymes, we extracted the catalyzed reactions and their substrates, products, and directionality. Finally, we included canonical metabolic pathways in a simplified way. In each GENSOR unit, we linked pairs of metabolites that were present in the same canonical pathway, taking into account the directionality of the pathway (i.e., that one metabolite can be transformed into the other). Only one meta-reaction (called 'complementary pathway reaction' in RegulonDB) was added to link the metabolites, regardless of the number of intermediate reactions between them in the pathway (Supplementary Figure S3). The end result was a multilevel network that included TUs, proteins, protein complexes, and metabolites; such a network was termed a Genetic Sensory Response unit, or GENSOR unit for short. GENSOR units integrate the transcriptional and metabolic level in a single network, providing a higher-level view of their interplay and their physiological relevance (Supplementary Figure S4).

We have previously reported (13) a connectivity metric that measures the functional homogeneity of a GENSOR unit in terms of the ability of its metabolic reactions to create a metabolic flux by sharing substrates or products, as in a pathway. In brief, connectivity takes into account the number of enzymes (E_c) whose catalyzed reactions create a metabolic flux, the total number of enzymes (E_t) and the total number of metabolic fluxes (M_f) present in the GENSOR unit (see Materials and Methods). The connectivity formula returns a value from 0 to 1. Zero indicates total functional heterogeneity, where none of the reactions in the GENSOR unit happen consecutively. A value of 1 indicates a paradigmatic GENSOR unit where all the reactions are involved in the same metabolic flux and therefore are functionally related (Appendix S1). We calculated the connectivity of 201 GENSOR units and eliminated those with less than two enzymatic reactions to avoid artificial values of 0. The resulting connectivity distribution (Figure 1A), based on a different version of RegulonDB, replicated previous results (13) where the largest proportion of GENSOR units had a connectivity value of 1, followed by those with connectivity of 0 and a continuum in between.

Functional heterogeneity in regulons is not explained by enzymatic regulation, low-confidence regulatory interactions, or chance

Once we reproduced our previous results, we explored artifacts that could be responsible for the observed connectivity gradient. As a first step, we examined the role of enzymatic regulation in the connectivity of GENSOR units. It is known that enzymatic regulation plays an important role in regulatory genetic programs (20). For example, anthranilate synthase is the enzyme that catalyzes the first step of tryptophan biosynthesis, and it can be allosterically inhibited by L-tryptophan, the end product of the pathway (21). This type of interactions can create functional links between enzymes and metabolites that are missed in the connectivity metric but are important because they increase the functional homogeneity of GENSOR units. We retrieved 422 enzyme-metabolite interactions annotated in EcoCyc and that were included in 87 GENSOR units. The connectiv-

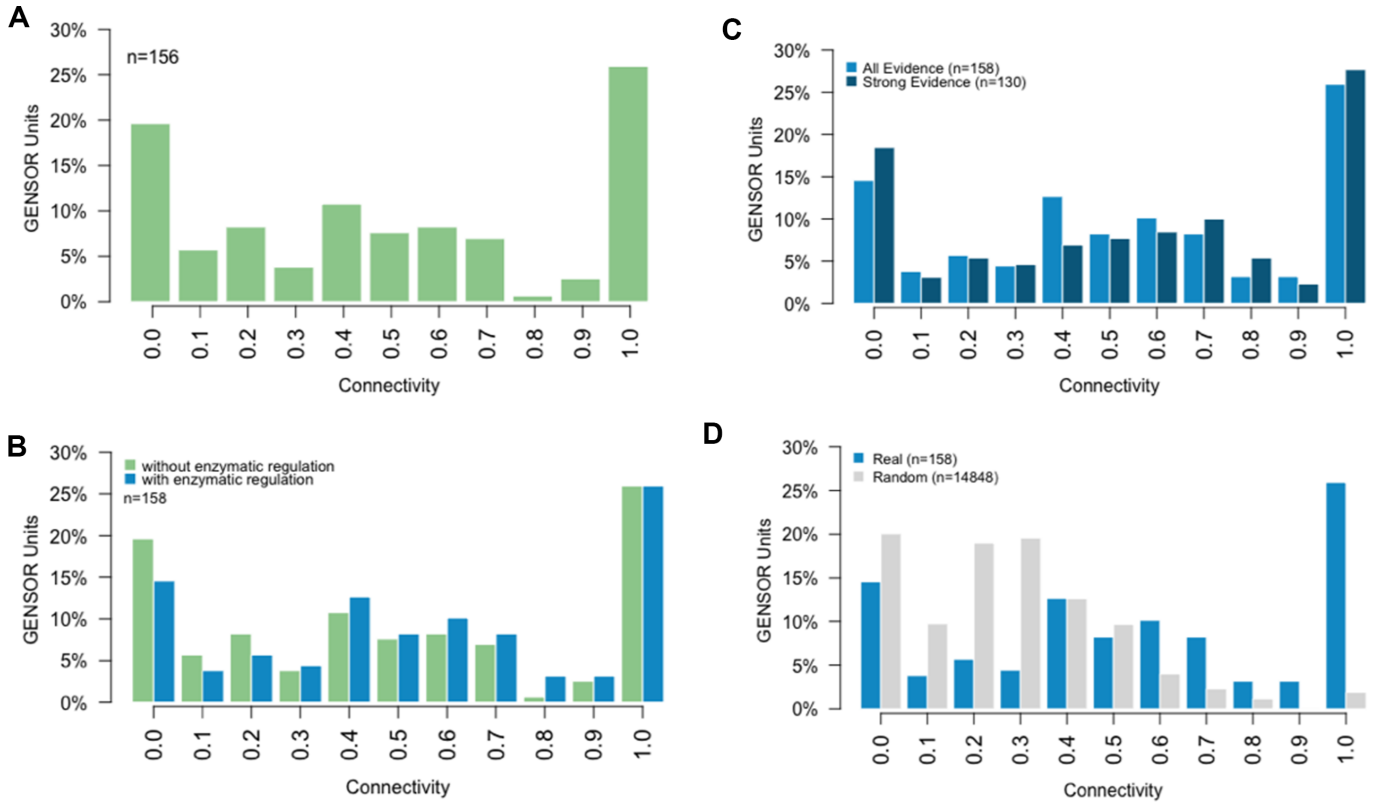

Figure 1. Connectivity analysis. (**A**) Connectivity distribution of GENSOR units calculated using the previously reported algorithm on later database versions. (**B**) Distribution of modified connectivity to consider enzymatic regulation in GENSOR units, compared to (A). (**C**) Connectivity distribution of GENSOR units containing only regulatory interactions with strong evidence, compared to GENSOR units including all reported regulatory interactions. (**D**) Mean connectivity distribution of GENSOR units assembled from 100 random regulons, compared to real GENSOR units. All distributions in (C) and (D) include enzymatic regulation.

ity calculation was slightly modified by expanding the definition of a *connected enzyme* (E_c in Equation 1), from 'an enzyme whose catalyzed reaction shares at least one substrate or product with another reaction in the GENSOR unit' to 'an enzyme whose catalyzed reaction shares at least one substrate or product with another reaction in the GENSOR unit or is allosterically regulated by the substrate or product of the reaction catalyzed by another enzyme in the GENSOR unit.' The resulting connectivity distribution using the modified connectivity (Figure 1B) did not differ significantly from the original distribution (Wilcoxon–Mann–Whitney test; P-value > 0.05). This result suggests that the gradient of functional complexity originally observed is not a consequence of missing metabolite-enzyme interactions. Nevertheless, metabolite-enzyme interactions add useful information to GENSOR units, so we included them and used the modified version of connectivity in all further analyses.

The gradient of connectivity could also be explained by spurious regulatory interactions. RegulonDB classifies each regulatory interaction as strong or weak according to the evidence provided by the methods used to identify it (22). For instance, binding experiments of the purified TF are considered strong evidence, while changes in gene expression in a TF mutant strain are considered weak evidence. Certain combinations of multiple independent weak evidence can also add up to a strong evidence. It is possible that interactions with low levels of confidence, such as those inferred from a mutant phenotype (i.e. the TF mutant strain shows a phenotype in which the regulated gene is involved) do not happen in reality and introduce noise in the connectivity of the GENSOR unit that includes them. We used RegulonDB evidence codes to eliminate all the regulatory interactions with weak evidence and then we assembled high-confidence GENSOR units. If the connectivity gradient observed is being biased by low-confidence interactions, the connectivity distribution of high-confidence GENSOR units should differ. Figure 1C shows that this is not the case, since the distribution is not significantly different (Wilcoxon–Mann–Whitney; P-value > 0.05), even considering that the total GENSOR units tested decreased by 18% and the average size of the assembled GENSOR units decreased from 7.5 enzymes to 5.4. As we have shown before (13), there was no correlation between the connectivity value of a GENSOR unit and its size (Supplementary Figure S5). Eighty-three percent of regulatory interactions with strong evidence included evidence of gene expression changes related to TF activity. The high percentage supports that our high-confidence GENSOR units not only rely on TF binding evidence, but also reflect functional interactions. These results show that low-confidence interactions are not introducing bias to our results and are not responsible for the connectivity gradient.

The prevalence of the connectivity gradient led us to explore whether it could be expected by chance, meaning that we are not measuring a biological property but the result of a metric's artifact. In order to do this, we took the complete set of genes in the Transcriptional Regulatory Network (TRN) and recreated the 201 regulons by randomly selecting genes from the set. The only property of the regulons that remained was the number of enzymes, because, as mentioned before, the connectivity value does not depend on the GENSOR unit size (Supplementary Figure S5). We assembled GENSOR units using the random regulons as starting points and obtained their connectivity distributions. For determination of statistical significance, we repeated this process 100 times and obtained the mean connectivity distribution (Figure 1D). The resulting distribution was significantly different from the original connectivity distribution (Wilcoxon–Mann–Whitney; P-value < 2.2e-16), suggesting that the connectivity gradient observed in regulons is not an artifact. The three previous results support that the gradient of connectivity reflects underlying biological principles, reinforcing the notion that functional homogeneity is not a general property of individual regulons.

The gene ontology also reveals functional heterogeneity in regulons

So far we have supported the functional heterogeneity of GENSOR units in terms of the connectivity metric. Next, we explored whether other methods of functional quantification yield the same results. Our connectivity metric reflects the functional homogeneity of a GENSOR unit under the assumption that proximity in the metabolic network implies similar function, as in a metabolic pathway. This assumption is commonly accepted, as enzymes and metabolites that are present in the same metabolic pathway (e.g. KEGG or EcoCyc pathways), and therefore are close in the larger metabolic network (23), work together to produce a final product. Nevertheless, there are other approaches to describe functional homogeneity, and of relevance are those based on the GO terms (17,18), a functional classification of genes mainly based on homology and phenotypic effects of genes. One of the most used approaches is GO Enrichment Analysis, which relies on statistical analysis to identify functional terms that are over- or underrepresented in a set of genes of interest, given the background of the functional annotations of each gene in the genome (24,25). Essentially, the GO Enrichment Analysis answers the question, 'Which biological processes are significantly enriched in a GENSOR unit?' Unfortunately, the results are not a true reflection of functional homogeneity, because genes tend to belong to more than one GO term and different subsets of genes in a GENSOR unit could account for different enriched processes, giving no direct information on whether there is one process where all the genes in the GENSOR unit are working together. For a better reflection of functional homogeneity, we focused on two different questions: 'How general is the biological process that can simultaneously describe all the genes in a GENSOR unit?' and 'What is the fraction of a GENSOR unit that can be explicitly explained by a biological process?'

To answer the first question, we identified the GO term that described all the genes in each GENSOR unit and focused on the tree structure of the ontology to quantify the specificity of the term. We obtained, for each GENSOR unit, the subset G of genes that are present in the Biological Process (BP) branch of GO. Eighty-two percent of the 2702 genes present in any GENSOR unit are present in the BP branch. To avoid uninformative results, 15 GENSOR units with less than two annotated genes were excluded from the analysis. For the remaining 186 GENSOR units, we identified the farthest downstream GO term from the root of the ontology that included all the genes in the GENSOR unit, in other words, the biological process that is most representative of, or dominant, in the GENSOR unit. Since terms closer to the root are more general in their definition, picking the term farther from it implies that it will also be the most informative. The root term, Biological Process, was labeled as level 1, its immediate children as level 2, and so on until level 11. The worst-case scenario would be a GENSOR unit whose most representative GO term is in level 1 of the ontology, indicating that no other term in the ontology can fully describe that GENSOR unit. This scenario would also imply functional heterogeneity, given that the genes have functions spread across the ontology.

To validate this interpretation, we obtained the dominant GO term of global regulators (ArcA, CRP, Fis, FNR, HNS, IHF, Lrp), which by definition are involved in several biological processes. The dominant GO term of all global regulators was indeed the level 1 term Biological Process, the most general term. Results for the complete set of GENSOR units showed that 69.1% also had the level 1 term Biological Process as the dominant GO term (Figure 2A). Additionally, 9.2% of GENSOR units had a level 2 GO term as its best descriptor. Taken together, these results agree with connectivity results in that around three-fourths of GENSOR units are functionally heterogeneous. The high percentage of functionally heterogeneous GENSOR units is not a consequence of the ontology having more terms in levels 1 and 2; in fact, these levels only include 20 terms, accounting for 0.005% of total BP GO terms (Supplementary Figure S6A). It is neither due to tested genes being present only in levels 1 and 2, since 99.8% of tested genes are present in terms from levels 3 and higher.

Another explanation could be that the more general terms are the only ones with enough genes annotated to describe all the genes in a GENSOR unit, but that is also not the case, since 124 GO terms of level 3 and higher contained more genes than the largest GENSOR unit (Supplementary Figure S6B-C). A possible functional bias is that TFs are annotated in the ontology with terms related to transcription, as opposed to the processes they regulate. To explore this possibility, we excluded autoregulated TFs from the analysis. Results were not significantly affected, since 74.4% of GENSOR units still obtained a dominant GO term in levels 1 or 2. It is also possible that expecting a complete regulon to be explained by a single GO term is extremely stringent, and perhaps two GO terms are enough to explain most regulons. However, that is not the case, since allowing two dominant GO terms only increased by 30.8% our interpretative power, leaving 47.5% of GENSOR units still dispersed in three or more biological processes. This anal-

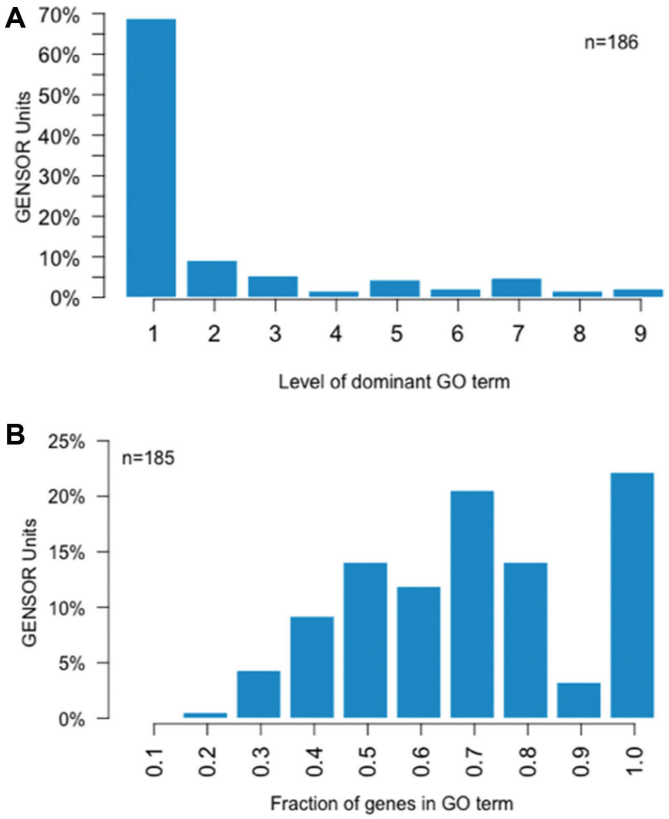

Figure 2. GO analysis. (**A**) Distribution of levels of the GO term that best describes each GENSOR unit. (**B**) Distribution of the highest fraction of genes of each GENSOR unit that are present in the same GO term. Only descriptive GO terms (level 3 and higher) were considered in this analysis. In both analyses, GENSOR units with less than two genes annotated in the ontology were excluded.

ysis supports the notion that functional heterogeneity is a general property of regulons.

The second question, 'What is the highest fraction of a GENSOR unit that can be explicitly explained by a biological process?' was answered by eliminating the two most general levels of the ontology and repeating the analysis. By eliminating the most general processes, we identified the most representative GO term of each GENSOR unit that was still informative. We obtained the fraction of genes in the GENSOR unit that were included in the best GO term; the closer the fraction is to 1, the more functionally homogeneous the GENSOR unit. The best scenario would be GENSOR units with a value of 1, meaning that all of their genes are involved in an informative Biological Process. Figure 2B shows that the gradient of complexity observed in the connectivity analysis was also present here: most GENSOR units cannot be entirely described through a single, informative Biological Process, and the fraction of genes that belong to the most representative process varies from 0.2 to 1. Notably, level 6 is the most populated in the ontology (Supplementary Figure S6A), with 25.5% of all terms belonging to it, but only 10.3% of GENSOR units had their most representative GO term at that level (Supplementary Figure S6D), suggesting that the analysis was not biased by properties of the ontology. Level 3, the most general in this analysis, was the most represented level (36.7%) in GENSOR units (Supplementary Figure S6D), further supporting their functional heterogeneity. The fraction of genes in the dominant GO term was not affected by the size of the GENSOR unit (Supplementary Figure S7).

Functionally heterogeneous GENSOR units are not the less studied TFs. For example, MetJ is a TF associated with biosynthesis and transport of methionine (hence its name). As expected, its regulated genes are included in methionine-related GO terms such as 'methionine biosynthetic process' or 'protein methylation,' but they only include a small fraction of the regulon. Interestingly, MetJ's dominant GO term is 'organonitrogen compound metabolic process,' which includes 58% of the regulated genes and is not directly related to methionine.

As mentioned before, connectivity and GO analyses rely on different properties of functional relationships, but both suggest that only around a quarter of TFs regulate functionally homogeneous genes. Scores do not show any clear correlation between the metrics (Supplementary Figure S8), but they overlapped in 15 GENSOR units that scored perfect functional homogeneity in both analyses: AllS, BetI, BirA, CynR, DhaR, FabR, FeaR, GcvA, HipAB, MazE, MazE-MazF, MhpR, TreR, XapR and YqjI. The dominant GO term of these regulons and a more detailed description of their GENSOR units can be found in Table 1. Effectors are known for nine of them. In most cases, the annotated biological process is directly related to the effector, for example, AllS is involved in the 'allantoin assimilation pathway' and it binds to allantoin (26). In other cases, the relationship is more indirect but still present, for instance, BetI's effector is choline (27) and is annotated as being involved in 'response to stress' because choline can be converted into glycine betaine, a thermo- and osmoprotectant (28) by genes directly regulated by BetI. These GENSOR units reflect the most functionally local TFs: they are involved in a single biological process and placed at the bottom of the TRN hierarchy since they do not regulate other TFs.

Complex regulons are the most functionally homogeneous type of regulons

Having further supported the notion that the genes directly regulated by a TF are not generally involved in the same biological process, we focused on exploring the functional homogeneity of other types of regulons, defined slightly differently. Our underlying assumption was that transcriptional regulation plays a central role in cellular decision-making: when a cell is faced with a change in the environment, a coherent response must be orchestrated mainly through the action of TFs. The resulting hypothesis is that there is a type of regulatory unit in the TRN that also acts as a functional unit. We selected three previously reported (29) types of regulatory groups: simple, complex, and strict regulons (Table 2). Their definitions (Table 2, Supplementary Table S1, Supplementary Figure S9) combine two properties of TFs: their effects on genes and their shared occupation of promoters with other TFs. Based on previous reports on the relevance of TF conformation information (29), we also considered the groups of genes directly regulated by a specific TF-

Table 1. Functionally homogenous GENSOR units. Regulated biological processes shown are obtained from the most representative GO term identified. Summaries describe active conformations and end-metabolites of the metabolic fluxes mediated by the TF of the GENSOR unit

GENSOR unit	Known Effectors	Dominant GO term (level of term)	Mechanistic summary of GENSOR unit	Genes in regulon
AllS	allantoin	allantoin assimilation pathway (7)	AllS, in the presence of allantoin, positively regulates the expression of genes required to produce ammonium and oxalurate.	3
BetI	choline	response to stress (3)	BetI, by itself, negatively regulates the expression of genes required to produce betaine and A(H2).	4
BirA	biotinyl-5′-AMP	monocarboxylic acid biosynthetic process (7)	BirA, in the presence of biotinyl-5′-AMP, negatively regulates the expression of genes required to produce coenzyme A, 5′-deoxyadenosine, AcpP, adenosylhomocysteine, S-adenosyl-4-methylthio-2-oxobutanoate, oxidized [2Fe-2S] ferredoxin, methionine, malonyl-[acp] methyl ester, unsulfurated [sulfur donor] and biotin.	5
CynR	cyanate	cyanate catabolic process (6)	CynR, in the presence of cyanate, positively regulates the expression of genes required to produce carbamate.	4
DhaR		organic substance metabolic process (3)	DhaR, by itself, positively regulates the expression of genes required to produce DHAP, PtsH and pyruvate	4
FabR		monocarboxylic acid metabolic process (7)	FabR, by itself, negatively regulates the expression of genes required to produce 3-oxo-decanoyl-[acp], 3-oxo-dodecanoyl-[acp], trans tetradec-2-enoyl-[acp], trans hexadecenoyl-[acp], (2E)-dodec-2-enoyl-[acp], 3-ketopimelyl-[acp] methyl ester, 3-oxo-cis-delta7-tetradecenoyl-[acp], 3-oxo-cis-delta9-hexadecenoyl-[acp], 3-oxo-octanoyl-[acp], 3-oxo-palmitoyl-[acp], 3-oxo-hexanoyl-[acp], 3-oxo-myristoyl-[acp], trans hex-2-enoyl-[acp], crotonyl-[acp], acetoacyl-ACP and acetoacetyl-[acp]	2
FeaR		cellular biogenic amine catabolic process (7)	FeaR, by itself, positively regulates the expression of genes required to produce perhydrol, oxopropanal, RCHO, ammonium and phenylacetate.	2
GcvA	purine, glycine	nitrogen compound metabolic process (3)	GcvA, by itself, dually regulates the expression of genes required to produce [glycine-cleavage complex H protein] N6-dihydrolipoyl-L-lysine, methylene-H4PteGlu(n) and ammonium.	5
HipAB		organic substance metabolic process (3)	HipAB, by itself, negatively regulates the expression of genes required to produce ppGpp and pppGpp.	5
MazE		nucleobase-containing compound catabolic process (5)	MazE, by itself, negatively regulates the expression of genes required to produce cytidylate, dAMP, 5′-IMP, 5′-UMP, dGMP, dCMP, dTMP, xanthosine-5-P, dIMP, mononucleotide and dUMP.	3
MazE-MazF		nucleobase-containing compound catabolic process (5)	MazE-MazF, by itself, negatively regulates the expression of genes required to produce cytidylate, 5′-IMP, 5′-UMP, dCMP, dAMP, dIMP, dGMP, xanthosine-5-P, mononucleotide, dTMP and dUMP.	3
MhpR	3-(2,3-dihydroxyphenyl) propanoate, 3-(3-hydroxyphe nyl)propanoate	aromatic compound catabolic process (5)	MhpR, in the presence of 2,3-DHP, positively regulates the expression of genes required to produce succinate, (2Z)-2-hydroxypenta-2,4-dienoate, acetyl-CoA, fumarate and pyruvate MhpR, in the presence of 3HPP, positively regulates the expression of genes required to produce succinate, (2Z)-2-hydroxypenta-2,4-dienoate, acetyl-CoA, fumarate and pyruvate	6
TreR	trehalose 6-phosphate, trehalose	cellular metabolic process (3)	TreR, by itself, negatively regulates the expression of genes required to produce glucose-6-P, PtsH and glucopyranose.	2

effector complex and those directly regulated by a specific TF-effector complex under the same effect. For instance, the genes activated by the complex TyrR-phenylalanine belong to a different regulon than those repressed by TyrR-tyrosine. In total, we analyzed six types of regulatory groups (Table 2, Supplementary Table S1, Supplementary Figure S9). We obtained the groups of genes in the TRN derived from each definition and used them as the starting points to assemble GENSOR units. As positive controls, we also used the GENSOR unit assembly pipeline on groups of genes defined by pathways and GO terms. As negative controls, we generated 100 sets of random gene groups for each type of regulatory grouping (see Materials and Methods). To compare the functional homogeneity of regulatory groups, we obtained their connectivity distribution and identified dominant GO terms as described in the previous sections.

Connectivity values (Figure 3A and Table 2) showed that the highest-scoring regulatory group, with a median of 0.8, contains complex regulons, that is, groups of genes regulated by the same combination of TFs. The rest of the reg-

Table 1. Continued

GENSOR unit	Known Effectors	Dominant GO term (level of term)	Mechanistic summary of GENSOR unit	Genes in regulon
XapR	xanthosine	nucleobase-containing small molecule metabolic process (4)	XapR, in the presence of xanthosine, positively regulates the expression of genes required to produce xanthine, hypoxanthine, guanine and nicotinamide ribose. XapR, by itself, positively regulates the expression of genes required to produce xanthine, hypoxanthine, guanine and nicotinamide ribose.	2
YqjI	nickel, Fe+2	cellular response to stimulus (3)	YqjI, by itself, negatively regulates the expression of genes required to produce (2,3-dihydroxybenzoylserine)3, Fe+2 and a siderophore.	2

Table 2. Definitions of regulatory units and controls used to assemble GENSOR units

Regulatory groups/controls	Definition	Total regulons
General Regulons	Genes directly regulated by a TF. Irrespective of effect, TF conformation, or coregulation with other TFs.	201
Strict Regulons	Genes directly regulated by a TF, under the same effect (+/−), irrespective of TF conformation, or coregulation with other TFs.	294
Simple Regulons	Genes directly regulated one and only one TF. Irrespective of effect or TF conformation.	107
Complex Regulons	Genes directly regulated by a combination of TFs, irrespective of effect or TF conformation.	398
Conformation Regulons	Genes directly regulated by a specific conformation (TF-effector complex) of a TF, irrespective of effect, or coregulation with other TFs.	221
Conformation + effect regulons	Genes directly regulated by a specific conformation (TF-effector complex) of a TF, under the same effect (+/−), or coregulation with other TFs.	304
GO term	Genes annotated in the same GO term, plus all the genes on its children terms	2860
Pathway	Genes that belong to the same metabolic pathway	420
Transcription unit	Genes transcribed on the same mRNA molecule	1037

ulatory groups had a median connectivity of 0.6, except for simple regulons, with a lower value of 0.5 (Supplementary Table S2). Pathways, the positive controls, had a median of 1, as expected. All regulatory groups had a distribution significantly different from random values (Supplementary Figure S10). Consistent with connectivity results, GO analysis also scored complex regulons as the most functionally homogeneous regulons, with a median fraction of genes of 1.0 (Figure 3B). The rest of the regulons showed a median of ∼0.7, with the lowest being general regulons, at 0.67 (Supplementary Table S2). The positive controls, groups of genes belonging to a GO term, had only 1.0 values, as expected. Again, all regulatory groups had a distribution significantly different from random values (Supplementary Figure S11). It is possible that the regulatory effect (activation/repression) of the TF over the regulated genes could be playing an important role in our results. To explore this possibility, we split all regulons into repressed or activated genes, assembled GENSOR units for each subgroup and obtained their connectivity and GO analysis values (Supplementary Figure S12). Interestingly, including the regulatory effect did not have a significant impact in the analysis, nor altered our previous conclusions. Complex regulons are the regulatory group with the highest proportion of small sized regulons (Supplementary Figure S13), but neither connectivity nor GO analysis show a correlation between size of the regulon and heterogeneity (Supplementary Figure S14). Through connectivity and GO analysis, two independent evaluations, we demonstrated that complex regulons are the most functionally homogeneous regulatory group of those tested, which fits with a model of TF coordination to orchestrate a cellular decision. Our results suggest that functionality in the TRN should be understood in terms of the cooperation between TFs and not at the level of individual TFs.

Gene expression data support functional homogeneity of complex regulons

To place our results within a more biologically relevant context, we quantified the coexpression of each regulon of each type of regulatory group. Our underlying assumption was that functionally related genes should be more frequently coexpressed than random genes, given that the execution of a biological process requires the presence of all the genes involved in it. Therefore, complex regulons, as the most functionally homogeneous type of regulatory unit, should also have the most coexpressed regulons. We measured coexpression by using the COLOMBOS database (19), a compendium of microarray experiments that includes expression data for 4321 genes across 4077 contrast conditions. For each regulon, we selected all the possible gene pairs and calculated the Spearman correlation of their expression values across all conditions. The median of the correlations of all gene pairs in a regulon was used as the coexpression score of the complete regulon. The Spearman correlation has been previously reported as the best statistic for coexpression analysis (30). As a control, we included TUs, genes that are transcribed on the same mRNA molecule and therefore coexpressed. Results showed that most regulatory groups have a median coexpression correlation co-

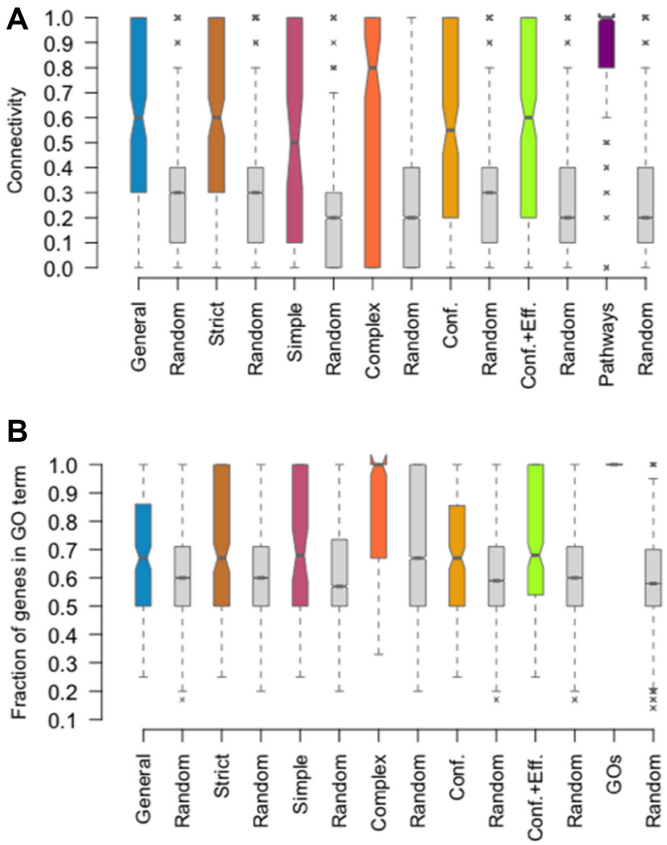

Figure 3. Functional homogeneity of different regulatory units. (**A**) Boxplots of connectivity distributions of regulatory groups and 100 randomizations of each set of regulons. Pathways are shown as control. (**B**) Boxplots of distributions of the highest fraction of genes of each GENSOR unit that are present in the same GO term, for each type of regulatory group and 100 randomizations of the genes that belong to each regulon. GO terms are shown as control. Conf stands for Conformation Regulons and Conf+Eff stands for Conformation + effect regulons. Regulon definitions can be found on Table 2.

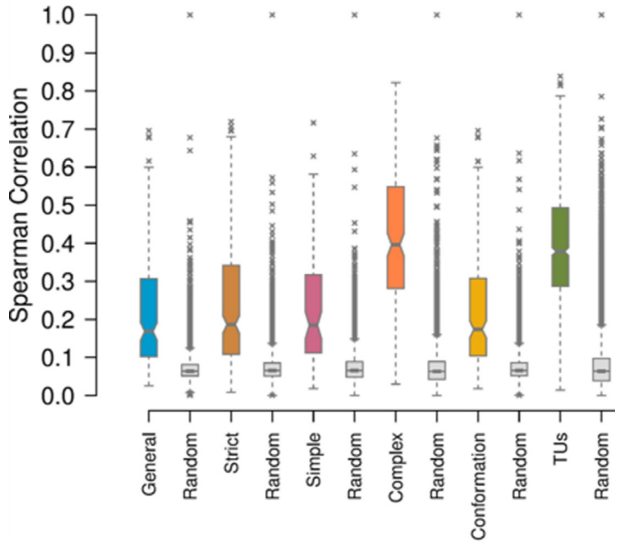

Figure 4. Boxplots of the distribution of coexpression values for each regulatory group, compared to their random sets. Transcription units (TUs) are shown as control.

efficient of around 0.2 (Figure 4A), except for complex regulons, with a median of 0.40, higher than the TU median of 0.38. Similar to connectivity and GO analysis results, considering the regulatory effect of the TF over the regulated genes did not appear to have a significant impact on the results (Supplementary Figure S15). Given that correlation coefficients are not very high, we were not able to make inferences about coexpression properties of different regulatory groups. However, we have obtained evidence that complex regulons are the most coexpressed regulatory unit, as frequently coexpressed as TUs, which agrees with their observed functional homogeneity.

DISCUSSION

The common model of transcriptional regulation mediated by TFs states that an individual TF detects a specific signal, changes conformation, and activates/represses a fixed set of genes. In turn, the regulated genes jointly orchestrate a response to the presence of the initial signal. Genomic studies allow the reevaluation of models, in order to identify the true general principles. We have shown before (13) that feedback between signals and the orchestrated response is a common occurrence. In this work, we expanded our analysis on the functional homogeneity of regulons by considering enzymatic regulation, the quality of data annotations, expectations by chance, and other definitions of functionality and regulons. All results confirmed that only one-fourth of known TFs regulate functionally homogeneous genes, demonstrating that genes directly regulated by a TF are not generally involved in the same biological metabolic process. Additionally, we have shown that complex regulons are the most homogeneous regulatory unit in the TRN, which is supported by their also being the most frequently coexpressed.

The functional heterogeneity measured here assumes that all the regulatory interactions in a regulon happen simultaneously, mainly due to the lack of information on growth conditions under which each regulatory interaction is active. It is possible that once this information becomes available, estimates of functional homogeneity of condition-dependent regulons will increase. However, there are reports of ChIP-seq (8), ChIP-exo (31,32), and microarray (7) experiments where regulated genes are involved in functions unrelated to the phenotype being studied, suggesting that even under very specific conditions our conclusions stand: direct targets of an individual TF are not necessarily involved in the same process. The most practical implication of our results is that coregulation is a dangerous assumption from which to propagate functional annotations for less-studied genes, given that this will be correct in only ~25% of the instances, as has been shown in this analysis of a comprehensive collection of regulons in *E. coli* (Figures 1B, 2B). Confidence that coregulation implies shared function can increase to around 50% if annotations are propagated to genes regulated by the same combination of TFs (Supplementary Figures S10C and S11C). TFs are commonly classified as local or global; one of the common features of global regulators is that the gene products they directly

regulate are involved in several functional classes. Results presented here show that involvement in many functional classes is also a property of local TFs, suggesting that this criterion should not be used alone for the classification of newly discovered TFs. It is certainly difficult to believe that our observations are unique to the biology of *E. coli*.

From a wider perspective, we consider three main possible explanations for the functional heterogeneity observed in general regulons: (i) we did not evaluate the regulatory groups that drive defined biological processes; (ii) there is not yet a functional framework that can describe the relationship between a regulon and its physiological effects; (iii) regulation of gene expression does not rely only on TFs. The first explanation relies on the fact that we only tested six regulon definitions. Although we considered several mechanistic properties of TFs in these definitions, it is possible that there is a definition of a regulon that has perfect correspondence with known biological processes, but we have yet to identify it.

The second explanation, that our current functional classifications of genes are not able to capture the functional regulatory logic, relies on the notion that regulation drives cellular decision-making. So far, function has been studied from the perspective of biochemical properties, homology, and phenotypic effects of mutants (17,33,34), not always taking into account regulation. Although in this work we did not comprehensively consider all available functional annotations, other widely used classifications, such as COG functional terms (34) and MultiFun terms (35), rely on similar evidences as those used by the GO Consortium, and we chose the latter given its higher level of curation and maintenance efforts. It is noteworthy that in the original MultiFun publication (35) it was mentioned that regulation was not considered in the gene classification because even operons are not always related in metabolic terms. In fact, connectivity and GO analysis of TUs also show some functional heterogeneity (Supplementary Figure S16). There could be an unexplored level of functional complexity where TFs are the main drivers, and more work is needed to explain how apparently different processes are in fact part of a larger response to a specific signal. Isolated examples of this interpretation efforts exist (36,37), and they provide evidence of our lack of a standardized functional vocabulary to interpret regulation; for instance, each ChIP-seq experiment requires a new effort by an expert curator. In the Jacob/Monod paradigm, it is counterintuitive that two genes regulated by the same TF are not involved in a common function. It may well be that the results shown here are more due to the current limitations in gene annotations and that the heterogeneity observed shows our ignorance of a regulatory logic waiting to be discovered.

The third explanation, that TFs are not the main drivers of cellular decision-making, is more feasible when one places TFs in the wider context of the cell. Bacteria do not rely solely on TFs to regulate their physiology; they also depend on ribosome abundance (38), RNA polymerase availability (39) and intrinsic stochasticity (40), not to mention posttranslational and posttranscriptional modifications. There are estimates for the TF Cra, suggesting that it only accounts for 32% of changes in gene expression of target genes in central metabolism (41). In this bigger picture, individual TFs are just one of the players in cellular decision-making, a resource to fine-tune the expression of genes that are not necessarily involved in the same process.

The observation that complex regulons are the most homogeneous regulatory unit agrees with the third explanation, since it supports that genetic programs are encoded beyond individual TFs. In this scenario, general regulons show the regulatory potential of TFs, but the specific subset of genes in the regulon that is expressed at a certain time is defined by the combinatory logic of the TFs bound to each gene's promoter. The possible combinations of TFs, promoters, number of sites, TF effects and order of TF binding far exceed the possible biological processes regulated by a simple model of one signal — one TF — one response (42–46). It has been shown that the large number of possible combinations allows for faster evolution of regulatory networks (43), which is known to happen (47,48). The high functional homogeneity of complex regulons highlights the importance of further exploring the combinatory logic of TFs on promoters. Although the TRN of *E. coli* has been widely studied, we are just beginning to explore the complexity of its relationships to metabolism and physiology. Dissection of the molecular decision-making processes associated with changes of growth conditions at a genomic level is now possible with current technologies and will no doubt bring an invaluable resource to further expand our understanding of microbial cell biology.

DATA AVAILABILITY

All regulon datasets, random regulons, custom scripts, and the raw data used in this study are available at GitHub [https://github.com/dledezma/functional_homogeneity]. GENSOR units of general regulons are available at RegulonDB [http://regulondb.ccg.unam.mx/central_panel_menu/integrated_views_and_tools/gensor_unit_groups].

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Luis José Muñiz-Rascado, Heladia Salgado and César Bonavides-Martínez for their skillful technical support and Mishael Sánchez-Pérez, Laura Gómez-Romero, Socorro Gama-Castro and Elad Noor for insightful discussions. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

FUNDING

Universidad Nacional Autónoma de México (UNAM); National Institutes of Health (NIH) [R01GM110597]; Consejo Nacional de Ciencia y Tecnología (CONACYT) Fronteras de la Ciencia [Fronteras 15]. Funding for open access charge: NIH [R01GM110597].

Conflict of interest statement. None declared.

REFERENCES

1. Pardee,A.B., Jacob,F. and Monod,J. (1959) The genetic control and cytoplasmic expression of "Inducibility" in the synthesis of β-galactosidase by E. coli. *J. Mol. Biol.*, **1**, 165–178.
2. Gunsalus,R.P. and Yanofsky,C. (1980) Nucleotide sequence and expression of Escherichia coli trpR, the structural gene for the trp aporepressor. *Proc. Natl. Acad. Sci. U.S.A.*, **77**, 7117–7121.
3. Maas,W.K. and Clark,A.J. (1964) Studies on the mechanism of repression of arginine biosynthesis in Escherichia coli: II. Dominance of repressibility in diploids. *J. Mol. Biol.*, **8**, 365–370.
4. Neidhardt,F.C. and Frederick,C. (1987) Escherichia coli and Salmonella typhimurium: cellular and molecular biology American Society for Microbiology. 2nd edition. **1**, 408.
5. Martínez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.*, **6**, 482–489.
6. Wu,W.S. and Li,W.H. (2008) Identifying gene regulatory modules of heat shock response in yeast. *BMC Genomics*, **9**, 1–15.
7. Göhler,A.K., Kökpinar,Ö., Schmidt-Heck,W., Geffers,R., Guthke,R., Rinas,U., Schuster,S., Jahreis,K. and Kaleta,C. (2011) More than just a metabolic regulator - elucidation and validation of new targets of PdhR in Escherichia coli. *BMC Syst. Biol.*, **5**, 197.
8. Fitzgerald,D.M., Bonocora,R.P. and Wade,J.T. (2014) Comprehensive mapping of the escherichia coli flagellar regulatory network. *PLoS Genet.*, **10**, e1004649.
9. Brohée,S., Janky,R., Abdel-Sater,F., Vanderstocken,G., André,B. and Van Helden,J. (2011) Unraveling networks of co-regulated genes on the sole basis of genome sequences. *Nucleic Acids Res.*, **39**, 6340–6358.
10. Liu,B., Zhou,C., Li,G., Zhang,H., Zeng,E., Liu,Q. and Ma,Q. (2016) Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses. *Sci. Rep.*, **6**, 1–11.
11. Lemmens,K., Dhollander,T., De Bie,T., Monsieurs,P., Engelen,K., Smets,B., Winderickx,J., De Moor,B. and Marchal,K. (2006) Inferring transcriptional modules form ChIP-chip, motif and microarray data. *Genome Biol.*, **7**, R37.
12. Pérez-Rueda,E., Tenorio-Salgado,S., Huerta-Saquero,A., Balderas-Martínez,Y.I. and Moreno-Hagelsieb,G. (2015) The functional landscape bound to the transcription factors of Escherichia coli K-12. *Comput. Biol. Chem.*, **58**, 93–103.
13. Ledezma-Tejeida,D., Ishida,C. and Collado-Vides,J. (2017) Genome-Wide mapping of transcriptional regulation and metabolism describes information-processing units in escherichia coli. *Front. Microbiol.*, **8**, 1466.
14. Santos-Zavaleta,A., Salgado,H., Gama-Castro,S., Sánchez-Pérez,M.S., Gómez-Romero,L.G., Ledezma-Tejeida,D., Santiago García-Sotelo,J., Alquicira-Hernández,K., Muñiz-Rascado,L.J., Peña-Loredo,P. *et al.* (2019) RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res.*, **47**, D212–D220.
15. Keseler,I.M., Mackie,A., Santos-Zavaleta,A., Billington,R., Bonavides-Martínez,C., Caspi,R., Fulcher,C., Gama-Castro,S., Kothari,A., Krummenacker,M. *et al.* (2017) The EcoCyc database: Reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res.*, **45**, D543–D550.
16. Karp,P.D., Latendresse,M., Paley,S.M., Krummenacker,M., Ong,Q.D., Billington,R., Kothari,A., Weaver,D., Lee,T., Subhraveti,P. *et al.* (2015) Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, **17**, 877–890.
17. Consortium, T.G.O. (2000) Gene ontology: Tool for the identification of biology. *Nat. Genet.*, **25**, 25–29.
18. Gene,T. and Consortium,O. (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, 331–338.
19. Moretto,M., Sonego,P., Dierckxsens,N., Brilli,M., Bianco,L., Ledezma-Tejeida,D., Gama-Castro,S., Galardini,M., Romualdi,C., Laukens,K. *et al.* (2016) COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. *Nucleic Acids Res.*, **44**, D620–D623.
20. Chubukov,V., Gerosa,L., Kochanowski,K. and Sauer,U. (2014) Coordination of microbial metabolism. *Nat. Rev. Microbiol.*, **12**, 327–340.
21. Pabst,M., Kuhn,J. and Somerville,R. (1973) Feedback regulation in the anthranilate aggregate from wild type and mutant strains of Escherichia coli. *J. Biol. Chem.*, **248**, 901–914.
22. Weiss,V., Medina-Rivera,A., Huerta,A.M., Santos-Zavaleta,A., Salgado,H., Morett,E. and Collado-Vides,J. (2013) Evidence classification of high-throughput protocols and confidence integration in RegulonDB. *Database*, **2013**, 1–15.
23. Feist,A.M., Henry,C.S., Reed,J.L., Krummenacker,M., Joyce,A.R., Karp,P.D., Broadbelt,L.J., Hatzimanikatis,V. and Palsson,B. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 1–18.
24. Yon Rhee,S., Wood,V., Dolinski,K. and Draghici,S. (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.*, **9**, 509–515.
25. Mi,H., Huang,X., Muruganujan,A., Tang,H., Mills,C., Kang,D. and Thomas,P.D. (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, **45**, D183–D189.
26. Rintoul,M.R., Cusa,E., Baldomà,L., Badia,J., Reitzer,L. and Aguilar,J. (2002) Regulation of the Escherichia coli allantoin regulon: coordinated function of the repressor AllR and the activator AllS. *J. Mol. Biol.*, **324**, 599–610.
27. Rkenes,T.P., Lamark,T. and Strøm,A.R. (1996) DNA-binding properties of the BetI repressor protein of Escherichia coli: the inducer choline stimulates BetI-DNA complex formation. *J. Bacteriol.*, **178**, 1663–1670.
28. Caldas,T., Demont-Caulet,N., Ghazi,A. and Richarme,G. (1999) Thermoprotection by glycine betaine and choline. *Microbiology*, **145**, 2543–2548.
29. Gutierrez-Rios,R.M., Rosenblueth,D.A., Loza,J.A., Huerta,A.M., Glasner,J.D., Blattner,F.R. and Collado-Vides,J. (2003) Regulatory network of escherichia coli: Consistency between literature knowledge and microarray profiles. *Genome Res.*, **13**, 2435–2443.
30. Pannier,L., Merino,E., Marchal,K. and Collado-Vides,J. (2017) Effect of genomic distance on coexpression of coregulated genes in E. coli. *PLoS One*, **12**, 1–20.
31. Seo,S.W., Kim,D., Szubin,R. and Palsson,B.O. (2015) Genome-wide reconstruction of OxyR and SoxRS transcriptional regulatory networks under oxidative Stress in Escherichia coli K-12 MG1655. *Cell Rep.*, **12**, 1289–1299.
32. Seo,S.W., Kim,D., O'Brien,E.J., Szubin,R. and Palsson,B.O. (2015) Decoding genome-wide GadEWX-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in Escherichia coli. *Nat. Commun.*, **6**, 7970.
33. Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K.B., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T. *et al.* (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. *Nucleic Acids Res.*, **34**, 1–9.
34. Galperin,M.Y., Makarova,K.S., Wolf,Y.I. and Koonin,E. V. (2015) Expanded Microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
35. Riley,M. (1993) Functions of the gene products of Escherichia coli. *Microbiol. Rev.*, **57**, 862–952.
36. Aquino,P., Honda,B., Jaini,S., Lyubetskaya,A., Hosur,K., Chiu,J.G., Ekladious,I., Hu,D., Jin,L., Sayeg,M.K. *et al.* (2017) Coordinated regulation of acid resistance in Escherichia coli. *BMC Syst. Biol.*, **11**, 1.
37. Gao,Y., Yurkovich,J.T., Seo,S.W., Kabimoldayev,I., Dräger,A., Chen,K., Sastry,A. V, Fang,X., Mih,N., Yang,L. *et al.* (2018) Systematic discovery of uncharacterized transcription factors in Escherichia coli K-12 MG1655. *Nucleic Acids Res.*, **46**, 10682–10696.
38. Borkowski,O., Goelzer,A., Schaffer,M., Calabre,M., Mäder,U., Aymerich,S., Jules,M. and Fromion,V. (2016) Translation elicits a growth rate-dependent, genome-wide, differential protein production in Bacillus subtilis. *Mol. Syst. Biol.*, **12**, 870.
39. 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.
40. Munsky,B., Neuert,G. and van Oudenaarden,A. (2012) Using gene expression noise to understand gene regulation. *Science*, **336**, 183–187.
41. 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*. *Mol. Syst. Biol.*, **13**, 903.
42. Buchler,N.E., Gerland,U. and Hwa,T. (2003) On schemes of combinatorial transcription logic. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 5136–5141.
43. Mayo,A.E., Setty,Y., Shavit,S., Zaslaver,A. and Alon,U. (2006) Plasticity of the cis-regulatory input function of a gene. *PLoS Biol.*, **4**, 555–561.
44. Hunziker,A., Tuboly,C., Horváth,P., Krishna,S. and Semsey,S. (2010) Genetic fl exibility of regulatory networks. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 12998–13003.
45. Ezer,D., Zabet,N.R. and Adryan,B. (2014) Physical constraints determine the logic of bacterial promoter architectures. *Nucleic Acids Res.*, **42**, 4196–4207.
46. Semsey,S. (2014) Mutations in transcriptional regulators allow selective engineering of signal integration logic. *MBio.*, **5**, 1–7.
47. Borneman,A.R., Gianoulis,T.A., Zhang,Z.D., Yu,H., Rozowsky,J., Seringhaus,M.R., Wang,L.Y., Gerstein,M. and Snyder,M. (2007) Divergence of transcription factor binding sites across related yeast species. *Science*, **317**, 815–819.
48. Shou,C., Bhardwaj,N., Lam,H.Y.K., Yan,K.K., Kim,P.M., Snyder,M. and Gerstein,M.B. (2011) Measuring the evolutionary rewiring of biological networks. *PLoS Comput. Biol.*, **7**, e1001050.