Corynebacterium glutamicum regulation beyond transcription: Organizing principles and reconstruction of an extended regulatory network incorporating regulations mediated by small RNA and protein-protein interactions

Juan M. Escorcia-Rodríguez[1], Andreas Tauch[2], and Julio A. Freyre-González[1,*]

[1] Regulatory Systems Biology Research Group, Laboratory of Systems and Synthetic Biology, Center for Genomic Sciences, Universidad Nacional Autónoma de México. Av. Universidad s/n, Col. Chamilpa, 62210. Cuernavaca, Morelos, México

[2] Centrum für Biotechnologie (CeBiTec). Universität Bielefeld, Universitätsstraße 27, 33615. Bielefeld, Germany

* Correspondence: jfreyre@ccg.unam.mx (J.A.F.G.)

Abstract: *Corynebacterium glutamicum* is a Gram-positive bacterium found in soil where the condition changes demand plasticity of the regulatory machinery. The study of such machinery at the global scale has been challenged by the lack of data integration. Here, we report three regulatory network models for *C. glutamicum*: *strong* (3040 interactions) constructed solely with regulations previously supported by directed experiments; *all evidence* (4665 interactions) containing the *strong* network, regulations previously supported by non-directed experiments, and protein-protein interactions with a direct effect on gene transcription; and *sRNA* (5222 interactions) containing the *all evidence* network and sRNA-mediated regulations. Compared to the previous version (2018), the *strong* and *all evidence* networks increased by 75 and 1225 interactions, respectively. We analyzed the system-level components of the three networks to identify how they differ and compared their structures against those for the networks of more than 40 species. The inclusion of the sRNAs regulations changed the proportions of the system-level components and increased the number of modules but decreased their size. The *C. glutamicum* regulatory structure contrasted with other bacterial regulatory networks. Finally, we used the *strong* networks of three model organisms to provide insights and future directions of the *C. glutamicum* regulatory network characterization.

Keywords: *Corynebacterium glutamicum*; regulatory interactions; regulatory network; curation; network inference; systems; modules; NDA; regulogs

---

1. Introduction

Corynebacterium glutamicum is a Gram-positive soil bacterium, industrially relevant due to its amino acid production proficiency. It is also a model organism for the study of regulatory networks [1], along with other bacteria such as *Escherichia coli*, *Bacillus subtilis*, and *Streptomyces coelicolor*. These model organisms are usually compared and diverse differences have been found (e.g., while *C. glutamicum* grows by apical elongation, *B. subtilis* and *E. coli* grow by lateral elongation [2]). Some aspects of the transcriptional regulatory mechanism of *C. glutamicum* have also been found to be different from those in other model organisms [3]. In contrast to *E. coli*, repression is the most common regulatory mechanism in C. glutamicum [4], and unlike *B. subtilis* and *E. coli* that have diauxic growth due to the preferential consumption of one carbon source over the others, *C. glutamicum* co-metabolizes glucose with several other carbon sources [3]. In terms of σ factors, *E. coli* and *C. glutamicum* have seven, while *B. subtilis* has 17, and over 60 σ factors have been found in *Streptomyces* species [5].

One of the challenges for the study of their transcriptional machinery at the global scale is the lack of data integration and the incompleteness of their global regulatory networks., despite being model

organisms [1]. The network incompleteness situation is worst for non-model organisms for which little or none is known about their transcriptional machinery. Even though high-throughput technologies speed up the reconstruction of regulatory networks, network models reconstructed solely with high throughput experiments present unusual structural properties when compared with other reconstructions performed mainly by conventional experiments (e.g., lower clustering coefficient [6]). Besides, the number of sequenced genomes scales rapidly, especially for bacteria, that even with high throughput experiments we cannot cope with all of them. Computational approaches for the inference of regulatory networks based on gene expression data are still emerging. Proof of that is their modest performance for model organisms in the DREAM5 challenge [7] and the inconsistency between gene expression data and the model used for regulatory networks [8]. Although a re-assessment with more complete networks and a larger number of model organisms is required [1]. An integrative approach of expression data and regulatory binding sites have shown to improve the prediction, but most of that improvement is by the binding site's approach, which provides more biological information (e.g., [8])

When inferring regulatory interactions with transcription factor (TF) binding sites data, the approaches can be classified into three major groups: phylogenetic footprinting, regulon expansion, and regulatory interactions transfer. The latter two approaches require previous regulatory information to increase the target genes (TGs) for the TFs in a network or transfer the regulatory information between organisms, respectively. On the other hand, phylogenetic footprinting does not require previous regulatory information but it is limited to the identification of co-regulated genes by a common TF. However, when the cognate regulator is unknown its identification is not trivial [9,10] due to the small size of the regulatory sequences and their overlap for some close homologous proteins. The transfer of regulatory interactions can be directly through orthology of both TF and TG conservation or by filtering for TF binding sites in the promoter region of the TG (also known as regulogs analysis [11]). The latter provides the best results helping to reduce spurious interactions that are no conserved in the organism of interest [12].

Previously, we studied the functional architecture of the *C. glutamicum* regulatory network with regulations by TFs binding to DNA acting at the level of transcription initiation (transcriptional regulatory network) and compared its connectivity distribution to those in *E. coli* and *B. subtilis* regulatory networks [13]. Since then, a plethora of works has continued unveiling novel transcriptional regulatory mechanisms in *C. glutamicum*. However, the study of the regulatory mechanisms has not been restricted to TF–DNA interactions. Some protein-protein interactions (PPi) are directly involved in transcription regulation (e.g., adenylated GlnK binding to AmtR (repressor) to release it from the DNA). Also, the inclusion of post-transcriptional regulations mediated by sRNAs into the global regulatory networks has been performed in other organisms (e.g., [14] in *E. coli* as a undirected network). Previous versions of the *C. glutamicum* transcriptional regulatory network have been used for the transfer of regulatory interactions to other corynebacterial strains hosted in the CoryneRegNet database [15], the construction of a model for the inference of the number of interactions once the regulatory networks are complete [6], an assessment of the NDA robustness to random remotion of nodes and interactions [13], as gold standard for the benchmarking of a network inference approach based on sequence data (unpublished results), and as reference for the identification of global regulators [16].

Here, we update the two previous transcriptional regulatory network models for *C. glutamicum* (Abasy IDs: 196627_v2018_s17 and 196627_v2018_s17_eStrong, hereinafter referred to as *all evidence* and *strong*, respectively) with hundreds of curated TF-DNA interactions, their effect, and their corresponding confidence level. In the *all evidence* network, we also included curated PPi having a direct effect on gene transcription, such as anti-σ–σ factor interactions and the formation of heteromeric regulatory complexes. We incorporated interactions mediated by regulatory small RNAs acting at the post-transcriptional level in a third network model (hereinafter referred to as *sRNA)*. We deposited all the three network models in the new v2.4 of Abasy Atlas. Our continuous curation of the *C. glutamicum* regulatory network has produced a set of five historical snapshots that together

recount the curation process spanning 11 years. These historical snapshots are also available in Abasy Atlas.

After this update, *C. glutamicum* moved from the fourth to the second position as the organism with the most complete regulatory network in Abasy Atlas, according to our recently published model of the total number of interactions a complete regulatory network has [6]. We discuss the global structural properties of the three network models in the context of the previous versions for the transcriptional regulatory models and more than 40 other bacterial networks from Abasy Atlas, the most complete collection of experimentally-validated regulatory networks [1]. We analyzed the organizing principles and the system-level components of the three networks to identify the effects of the inclusion of interactions supported by non-strong experiments, protein-protein interactions, and the post-transcriptional layer regulation by sRNAs. Finally, we use strongly supported regulatory networks from *S. coelicolor*, *B. subtilis*, and *E. coli* to gain knowledge of the DNA-binding TFs for which no TGs have been characterized in *C. glutamicum*, and we provide a list of potential interactions retrieved through a strict and conservative computational pipeline using the most precise tools to identify regulations.

1.1. A primer on analyzing regulatory networks

The concepts and procedures used in the field of network biology have been summarized and explained in-depth and with great clarity in previous works [17-21]. Nevertheless, in this section we summarized the state of the knowledge and main concepts required to analyze the relationship between structure and function of regulatory networks.

The abstraction of a regulatory network can be represented as a group of nodes and directed arcs. The nodes represent the entities of the network (commonly genes or sRNAs), and the arcs represent the direction of the interaction between two nodes. For example, the requirement of GlxR for the transcription of *ramA* can be represented as *glxR* → *ramA*, while a negative effect (such as GlxR on *acnR* transcription) is usually represented as *glxR* ⊣ *acnR*. We use the gene symbol (or locus tag in case no name has been assigned yet) to consistently represent the sequence of the interactions, for example, *sigA* → *glxR* ⊣ *acnR*, and so on (the housekeeping σ factor is required for the transcription of *glxR*, and GlxR hinders the transcription of *acnR*). Nodes representing other biological entities can also be included in the network. The *C. glutamicum* networks herein reported contain three types of nodes: genes, heteromeric protein complexes, and sRNAs. Heteromeric protein complexes are conformed by two or more regulatory proteins transcribed by different genes and are included in the network to reduce redundancy and improve the representation accuracy [1]. The effect of the sRNAs regulatory interactions is carried out at the post-transcriptional level. These interactions are included in the networks with an sRNA label in their corresponding Abasy ID [1]. The importance of the inclusion of sRNAs in bacterial regulatory networks is relatively recent [14] and there is little information regarding these types of interactions in bacterial regulatory networks.

Once the interactions are merged to form a global regulatory network, we can compute the **connectivity degree** of the nodes (k), which represents the number of interactions of a node with the rest of the network, regardless of the direction. In some scenarios, the connectivity degree could be more informative if the direction of the interactions is considered. The out-degree ($\mathbf{k_{out}}$) of a node is the number of nodes it regulates. The nodes with a k_{out} greater than zero are defined as regulators. The k_{out} is the most applied connectivity in regulatory networks (e.g., for the identification of proteins required for the transcription of a large fraction of the network: global regulators). The in-degree ($\mathbf{k_{in}}$) is the number of regulators involved in the transcription of a given gene/sRNA. An exception is the incoming interactions in heteromeric complexes that represent the formation of the complex instead of their regulation, despite that the relationship is causal as the presence of the subunits is required to produce the heteromeric complex. Hence, the heteromeric complexes have incoming interactions only from the subunits required for its conformation, while the subunits have outgoing interactions only to the heteromeric complexes they are part of [1]. These types of interactions are underrepresented in the network and therefore not specified in most cases. The $\mathbf{k_{max}}$ is defined as the largest connectivity value of the network and equals the k_{out} of the global regulator with the largest

set of TGs. The **auto-regulations** represent a direct transcriptional effect of the regulator onto its own coding sequence.

The average clustering coefficient quantifies the modularity of a network. This structural property is an example where the direction of the interactions is disregarded as modularity is defined as the degree to which the components of a system are separated or combined. The **clustering coefficient** of a node is defined as the fraction of its neighboring nodes that are connected to each other relative to the potential interactions that could exist among them. For example, a node A having as neighbors only the nodes B and C will have a clustering coefficient of one if an interaction exists between B and C (regardless of the direction of the interaction) because the potential number of interactions between the neighbors of A is only one. The clustering coefficient of A is zero if there is no interaction between B and C. Once the clustering coefficient is calculated for every node having at least two neighbors in the network, the values are averaged. For an illustrated example please see Box 1 in [19]. The **C(k)** shows a distribution of the average clustering coefficient for the nodes with connectivity k. Similarly, the distribution of the connectivity of the nodes is denoted as **P(k)**, provided by the probability of a node having k interactions. It has been previously debated whether the P(k) of real networks is truly governed by a power-law distribution where a few nodes have most of the interactions [22]. Recently, using several statistic methods we demonstrated that the regulatory networks truly follow a power-law distribution, they would fit other power law-like distributions better than a Poisson distribution regardless of the completeness of the network, and that the sole coefficient of determination (R^2) is a good proxy to assess the goodness-of-fit of the model [6].

A network component is a group of nodes in which every pair is connected by at least one path. Regulatory networks do not always comprise a single component. Commonly, small groups of nodes can be isolated from the rest of the network. This is frequently observed in non-model organisms for which only some groups of nodes have been studied. Whether regulatory networks are truly multicomponent, or this is only a consequence of network incompleteness, is still an open question. The **giant component** is the largest component of the network, and its size is determined by the number of nodes it covers. In regulatory networks the global TFs, such *sigA*, increase the fraction of nodes in the giant component. The higher the fraction of nodes in the giant component, the more cohesive the network is. The giant component of a network is the representative part of the network for most structural properties such as density. The **network density** is the fraction of interactions from the fully connected network (where every node would have a directed interaction to itself and every other node in the network) that exist in the actual network. The detection of a constrained space for the density values in bacterial regulatory networks [6] allowed us to infer the number of interactions expected once the curation of the network is completed [1], so we can identify some differences in the curation state of the regulatory networks.

Most of the definitions mentioned before are applied in the **κ-value** (Kappa value), which is defined as the point of the $C(k_{out})$ distribution where the change in the normalized k_{out} connectivity equals the change in the clustering but with the opposite sign. The κ-value is used as a threshold for the identification of global regulators and has shown high precision and sensitivity on different bacterial regulatory networks such as *E. coli* [23], *B. subtilis* [24], and *S. coelicolor* (unpublished results). While being conservative (high precision, low sensitivity) on an earlier version of the *C. glutamicum* regulatory network [13].

The global regulators shape the highest hierarchy in the diamond-shaped structure unveiled by the natural decomposition approach (NDA). The **NDA** is an *in silico* technique that deconstructs a regulatory network to naturally identify its structure and reconstructs it with the nodes classified into one of four classes: global regulators (**GR**), modular nodes (**Md**), intermodular nodes (**IM**), and basal machinery (**BM**). Global regulators are the TFs with a low clustering coefficient and a k_{out} greater than the κ-value. Once the GRs have been identified, the BM is also unveiled as the TGs that are regulated only by the GRs. The direct GR-BM regulation is required for fast responses without previous modulation of intermediates. GR and BM nodes and their interactions are removed from the network, as well as the rest of the nodes with $k_{out} = 0$ (putative structural genes). The remotion of these structural genes will lead to isolated groups of Md (modules) that work together for a common

purpose. Finally, the structural genes are re-inserted into the network preserving their original interactions, and they are included into the module of their regulators only if all their regulators are from the same module, otherwise as IM, integrating the signals from different modules. For further details about the NDA methodology please see figures 1-2 in reference [13] where the NDA is described and applied to an earlier version of the *C. glutamicum* transcriptional regulatory network. Noteworthy, this diamond-shaped hierarchy has been found structurally conserved even between phylogenetically distant organisms [24]. The NDA classification is robust to random remotion of interactions and nodes [13], but the curation state of the network can alter the class of some nodes. This applies mainly to the IM and the BM nodes that can be included in the Md class in a later (more complete) version of the network.

2. Materials and Methods

2.1. Curation and networks definition

Four types of interactions were defined to be considered in this new version of the C. glutamicum networks: 1) Homomeric-TF–DNA comprehending interactions between DNA-binding TFs (including σ factors) and DNA that alters the gene expression; 2) sRNA–RNA interactions, occurring at the post-transcriptional level modulating the concentration of the proteins; 3) Protein-protein interactions (PPi) class 1 (PPi-cI) defined as PPi with a causal regulatory effect such as anti-σ–σ interactions; and 4) PPi class 2 (PPi-cII), a form of TF–DNA interaction where the TF is a heteromeric protein complex with their cognate subunits-complex interactions. Two levels of confidence are defined for the interactions: strong if the interaction is supported by a TF-DNA direct binding experiment (e.g., footprinting with purified protein), and weak otherwise. Even though other types of interactions considered for this version might be supported by a direct experiment (e.g., yeast two-hybrid assay for PPi-cI), we only included homomeric-TF–DNA and heteromeric-TF–DNA interactions (PPi-cII) in the *strong* network. The *all evidence* network includes interactions supported by any experimental evidence, keeping the label "strong" only for those interactions taken from the *strong* network. For the *all evidence* network, all but the sRNA-mediated regulations are considered, while the sRNA network includes every type of interaction regardless of the experiment supporting them. The three networks reconstructed in this work were deposited in the new v2.4 of Abasy Atlas.

The curation of strong interactions was carried out manually by screening the PubMed library for publications describing regulatory interactions of *C. glutamicum*. Interactions were classified as strong when the respective paper contains experimental evidence of a TF-DNA interaction. In most cases, the TF of interest was purified and its direct interaction with DNA was demonstrated *in vitro*. Approaches like this also led to the experimental identification of the DNA binding site sequence. For the recovery of the weakly supported interactions, we reviewed the literature to identify TGs for the TFs already present in the *all evidence* network. We used as keywords "glutamicum", "regulon", "target genes", and the name symbol of the gene or its locus tag. Then, we followed a set of rules to include the interactions for every TF-TG pair of nodes: 1. an interaction does not exist in the network unless it is already in the previous version; 2. an interaction that is not part of the previous version does not exist unless there is experimental evidence to support the interaction; 3. an interaction supported solely by computational predictions is not included in any of the networks; 4. an interaction weakly supported by an experiment is part of the network until contradictory evidence is found (e.g., gene overexpression supported by microarrays data but invalidated by RT-PCR).

We included in the *sRNA* network the regulatory interactions by anti-sense sRNAs from the reference [25]. The authors include as anti-sense sRNA every sRNA that is transcribed in the opposite strand of a gene or starting within 100 nt of the 5′-end of an opposite CDS or within 60 nt from the 3′-end of an opposite CDS [25]. The authors identified two other types of sRNAs but regulatory interactions were only assigned to anti-sense sRNAs. For the name of the sRNAs, we used the nomenclature suggested by the authors: *cgb_xxxxx* to ease the identification of the nodes representing sRNAs in the *sRNA* network. The effect of the interactions was set to unknown: '?' and most of the sRNAs regulate the gene transcribed in the opposite DNA strand. We included the sRNAs as

independent nodes. We acknowledge that this artificially increases the genomic coverage for the *sRNA* network (counting twice the genes with an asRNA). However, assigning the interaction to the coding gene would be misleading and inflate the number of self-loops in the network even when the sRNAs might be transcribed through its own promoter. As previously discussed, the interaction coverage is a better proxy for network completeness than the genomic coverage [6]. Although the authors provide the σ factors required for the transcription of the sRNAs, we did not include these σ-DNA interactions as they were solely supported by DNA-binding motif computational predictions and we have identified a high number of false positives in the search of binding sites for the σ factors. Besides, interactions supported solely by computational predictions are not considered for the Abasy Atlas networks [1]. Interactions involving a protein-coding gene not mapping to a cgl-number or from another strain were not included in the networks but collected in a separated file (supplementary table 1).

2.2. Genome annotation and upstream sequences

Genome annotations used in this work were retrieved from NCBI [26] for the following organisms (accession code and version): *Corynebacterium glutamicum* ATCC 13032 (NC_006958.1), *Streptomyces coelicolor* A3(2) (NC_003888.3), *Bacillus subtilis* subsp. subtilis str. 168 (NC_000964.3), and *Escherichia coli* str. K-12 substr. MG1655 (NC_000913.3). Upstream (up to -300 to +50) sequences with reference to the start codon for the four genomes were retrieved from the RSAT suite [27] with the retrieve-seq tool preventing overlap with neighboring genes.

2.3. Regulatory networks for other organisms

All the regulatory networks used in this work were downloaded from Abasy Atlas, a large collection of manually curated transcriptional regulatory networks [1]. The set of non-redundant networks is defined as the most recent regulatory networks for each organism available in Abasy Atlas, resulting in a dataset of 42 regulatory networks for 42 bacterial strains. When using the non-redundant set for background for the here reported regulatory networks of *C. glutamicum*, the set includes the regulatory networks of all other organisms (41) plus the three herein reported networks.

2.4. System-level components

Nodes were classified into one of the four system-level component classes: GR, BM, Md, and IM were retrieved from Abasy Atlas. The classification of the nodes has been previously described [13]. Following, we briefly describe the NDA, the approach used for the classification of the nodes and modules identification: The κ-value is computed for the identification of GRs. Every node with a number of directly regulated TGs greater than the κ-value is classified as a GR and removed from the network as well as their interactions. The remotion of the global regulator nodes leads some nodes isolated. Those isolated nodes that are solely regulated by global regulators are classified as BM, representing structural components required for elemental functions such as the subunits for the RNA core polymerase. The nodes with no regulated genes in the remaining network are labeled as structural nodes and removed to identify an isolated group of nodes (modules) to be classified as Md. The nodes labeled as structural are reintegrated to the network as part of a module if all of their regulators belong to the same module, otherwise, as IM components, which integrate the signals from two or more modules dedicated to responding to different conditions.

2.5. Comparison of nodes and interactions of C. glutamicum with other bacterial regulatory networks

To quantify the fraction of *strong* interactions in each network, we computed the ratio of regulatory interactions classified as *strong* in each of the *all evidence* regulatory networks deposited in Abasy Atlas, including the *all evidence C. glutamicum* network herein reported and plot the distribution. We reconstructed the previously reported model developed to predict the size of regulatory networks [1] by using an expanded dataset including the herein reported *C. glutamicum* regulatory networks and robust linear regression. We then reassessed the goodness-of-fit of the

model by recomputing the adjusted coefficient of determination. Regulatory networks of *C. glutamicum* were highlighted in the distributions to ease identification and comparison with previous versions.

2.6. Global structural properties

All the structural properties reported in this work were retrieved from Abasy Atlas [1] version 2.4. For comparison with other bacteria, the values reported were normalized as follows: The number of auto-regulations was normalized by the number of regulatory nodes (those with the potential to have an auto-regulation). To ease the comparison of the density values in a plot, each of them was multiplied by 10. Please note that this modification is used only to compare the properties. The k_{max} was normalized by the number of nodes in the network (potential targets). The κ-value was normalized by the k_{max}. The size of the giant component was normalized by the number of nodes in the network. No normalization was applied to compare *C. glutamicum* network across versions and evidence levels. Instead, we used a log2-fold change ratio of the properties value relative to the corresponding value for the earliest network in case of different versions and the smallest network in case of comparing different evidence levels.

2.7. System-level components

Nodes classification, modules identification, and their annotation were retrieved from Abasy Atlas [1] version 2.4. For the graphic representation of the nodes classification, the values were computed using a log10 scale. For the representation of the module size, the actual values were used for the treemapping plot. For the distribution of the number of modules, the non-redundant set of regulatory networks from Abasy Atlas version 2.4 were used and the herein reported networks were highlighted and labeled to ease identification. For the comparison of the nodes in each NDA class for the three networks reported here, we used the Simpson similarity index defined as the number of common elements between two sets divided by the minimum of the two numbers. Hence, the similarity index can take values from zero (no overlap at all between the two sets) to one (one set is a subset of the other). For the interactions from GR and Md to the four classes, we computed the fraction of interactions between each class ignoring interactions from BM and IM classes (less than 1% of the network), which are attributed to missing interactions that will be included in the future curation of the network (e.g., *cgb_20925* regulating *sigA*). Matplotlib, Seaborn, Numpy, and Squarify libraries from Python were used to compute and plot the results.

2.8. Regulogs analysis

For the selection of source organisms, we used the last *strong* version of those organisms having strong regulatory networks. Namely *Escherichia coli* K-12 MG1655 (Abasy ID: 511145_v2020_sRDB18-13_eStrong), *Bacillus subtilis* strain 168 (Abasy ID: 224308_v2008_sDBTBS08_eStrong), and a curated *Streptomyces coelicolor* network with curated strong interaction until 2019 (unreported network). The regulogs analysis is based on the premise that regulatory sites are more conserved than the rest of non-coding sequences because are required for the cell to survive. Given the basis of the approach, the best strategy is to use phylogenetically closely-related organisms [11,28]. Unfortunately, model organisms for which a *strong* regulatory network is available are phylogenetically far from each other but we still can use them to study essential, conserved interactions [24]. The closest model organism with a highly-complete regulatory network is *Mycobacterium tuberculosis* (Abasy ID: 83332_v2018_s11-12-15-16), but the regulogs analysis has been previously used to transfer interactions in the opposite direction (from *C. glutamicum* to *M. tuberculosis*) [29], and the remaining interactions are mostly supported by weak evidence.

For the identification of orthologous genes, we used the OMA standalone software [30] with the genome sequences from NCBI (see above). We used the OMA classification of orthology relationship type and kept only the one-to-one orthology relationships. To construct the position weight matrices, we used MEME [31], Bioprospector [32], and MDscan [33] with the upstream sequence of TGs for

each TF with at least one *strong* evidence supporting the interaction. Upstream sequences were defined as up-to -300 to +50 bp relative to the start translation codon. Then, we used FIMO [34] to find individual matches of the matrices in the upstream sequences of the complete set of *C. glutamicum* one-to-one orthologous genes and considered only the motif using a p-value of $1x10^{-4}$ as a threshold to form TF-TG putative interactions. Gene identifiers for the TFs and TGs were mapped to the *C. glutamicum* genome annotation, and the interactions obtained with each of the three motif finding tools were integrated by a vote-counting approach, which has been found to improve the predictions [7], prioritizing the interactions considered as "more reliable" by the three motif finding tools.

3. Results and Discussion

3.1. The regulatory networks of C. glutamicum and potential applications

In this section, we report the new regulatory network models of *C. glutamicum*, their differences, and the statistics comparing them with the previous version, and discuss some potential applications of our network models. We reconstructed three regulatory network models: 1) The *strong* network (Abasy ID: 196627_v2020_s21_eStrong) conformed solely by DNA-binding TFs—mediated interactions that are supported by a direct experiment (e.g., footprinting with purified protein). 2) The *all evidence* network (Abasy ID: 196627_v2020_s21) conformed by every type of interaction at the transcriptional level that is supported by any experimental evidence and not discarded by any other. 3) The *sRNA* network (Abasy ID: 196627_v2020_s21_dsRNA) containing the *all evidence* network plus 545 post-transcriptional interactions mediated by regulatory sRNAs (**Figure 1**). The *strong* network is a subset of the *all evidence* network, while the *all evidence* network is a subset of the *sRNA* network (supplementary figure 1). We deposited the three reconstructed networks in the new v2.4 of Abasy Atlas, each of them providing a different level of completeness (supplementary figure 2), and useful in different scenarios. For example, even though the *strong* network is the smallest one, the confidence level of its interactions makes this network the best alternative to be used as a gold standard for benchmarking of approaches for the inference of directed regulatory networks (such as those based on regulatory binding sites). On the other hand, benchmarking of network inference tools based on transcriptomic data might tend to be penalized when using only the *strong* network, as it only contains direct TF-DNA interactions which cannot accurately be predicted based solely on transcriptomic data [8]. In that case, the *all evidence* network can be used as a gold standard, as it includes a broader scope of experimentally supported interactions that have not been reported as spurious. The *sRNA* network is the most comprehensive and therefore the best suited to study the biological regulatory mechanisms of *C. glutamicum*. Having reliable regulatory network models has proven being important even for synthetic biology, for example to engineer resource allocation by rationally modifying the transcriptional regulatory network [35].

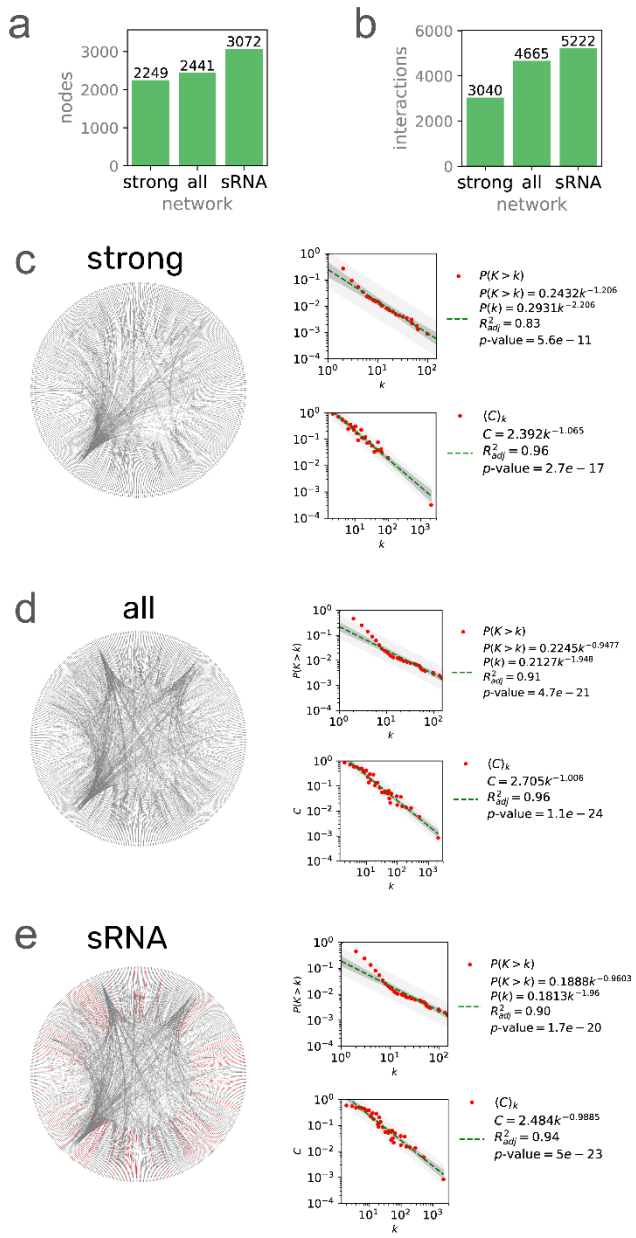


Figure 1. Three network models of the *C. glutamicum* regulatory network. The number of nodes (a) and interactions (b) for the three networks. Network, P(k) and C(k) distributions for the (c) 196627_v2020_s21_eStrong (*strong*), (d) 196627_v2020_s21 (*all evidence*), and (e) 196627_v2020_s21_dsRNA (*sRNA*) networks. Network plots were generated with Circos [36] using the leftmost gene/sRNA coordinates to sort the nodes clockwise. Nodes with no coordinates in the genome annotation were disregarded.

3.2. Global networks of C. glutamicum are quite different from other bacterial networks in terms of their structural properties

In this section, we analyze the global structural properties of the *C. glutamicum* regulatory networks in the context of the whole Abasy Atlas dataset. Previously, our group found a constrained complexity in the regulatory networks [6] and leveraged it to create a model for the inference of the size of regulatory interactions expected once the network curation is complete [1]. We identified a few networks falling outside of the prediction area (see figure 5 in [1]), being *C. glutamicum* one of those organisms, namely for the later versions containing the sigmulons of the housekeeping *sigA*. We found that this was a result of a low number of weakly supported interactions in contrast with other bacterial regulatory networks (**Figure 2**a). Mainly because the *C. glutamicum* regulatory

network has been highly curated in-house, giving preference to strongly supported interactions resulting in an overrepresentation of these interactions in contrast to other bacterial regulatory networks. The inclusion of weakly supported interactions fit better the *C. glutamicum* network into the model (**Figure 2**b). Note that the *strong* version of the network poorly follows the model as *sigA* directly regulates 85% of the network nodes. Besides, the fit of the *all evidence* network is affected by the inclusion of sRNA-mediated interactions (RNA in **Figure 2**b). This is a result of many sRNAs regulating only one gene in most cases.

Related to it, we expect the average clustering coefficient to decrease as the nodes/interactions ratio increases. The clustering coefficient of a node in the network is determined by the fraction of neighbors connected to each other. As expected, the average clustering coefficient of the *all evidence* network is higher than the other two types of networks for this year (**Figure 2**c) as it exhibits a better equilibrium (closest to 1) of the genomic/interaction coverage ratio (supplementary figure 2). Interestingly, despite the *C. glutamicum* networks exhibit a higher nodes/interactions ratio, they have a higher clustering coefficient than most of the bacterial regulatory networks (**Figure 2**c), perhaps because of a higher level of curation of the organism due to its biotechnological relevance. The density of the *C. glutamicum* networks is slightly lower than the rest of the bacterial regulatory networks. However, note that this difference is so small that even 10 times magnified the variance of the density values is very small (**Figure 2**b). This is expected due to the constraint governing the complexity of regulatory networks [6].

The fraction of nodes acting as transcriptional regulators is constrained in bacteria, beyond considering only the DNA-binding TFs (**Figure 2**c). The *C. glutamicum* regulatory network models show a different behavior, while the network including sRNA-mediated interactions falls on the upper boundary (~25%), the other two networks fall on the lower boundary of the distribution (5%), even when the latter includes most of the DNA-binding TFs of *C. glutamicum*. For most organisms, the k_{max} is below 50% of the nodes in the network. However, the regulatory networks for *C. glutamicum* are outliers in the distribution (**Figure 2**c) due to the *sigA* interactions. The size of the giant component can be represented by the fraction of the network it comprehends. For most regulatory networks, this fraction is close to one (**Figure 2**c). Especially in the case of *C. glutamicum*, whose networks with no sRNA regulation are practically a single component, showing the cohesiveness of these networks.

The κ-value is the threshold to identify global regulators. Every network has a different κ -value that relies on its hubness and modularity, but larger k_{max} values result in larger κ -values. To make the κ -values comparable, we normalized them by the k_{max} of the cognate network, allowing κ to take values between zero and one. Interestingly, the normalized κ-value seems to be also constrained to values lower than 0.25, and the values for the three networks of *C. glutamicum* overlap. This suggests that the κ -value is robust to the inclusion of weakly supported interactions and sRNAs. Besides, this agrees with previous analysis on the robustness of the inference of global regulators to random removal of nodes and interactions [13]. However, in-depth studies with other sampling approaches and other organisms are required. Auto-regulations in a regulatory network allow mechanisms to modulate themselves. A higher number of auto-regulations in the networks provide a faster response of the organism to the changing conditions [37]. *C. glutamicum* requires the adaptation to different media conditions in the soil, therefore a high number of auto-regulations is expected (**Figure 2**b, f) were the *strong* and the *all evidence* networks are above most regulatory networks. However, the fraction of auto-regulations in the network containing sRNA-mediated interactions is much lower because of the large number of regulatory sRNAs that bind to other RNA but not to themselves.

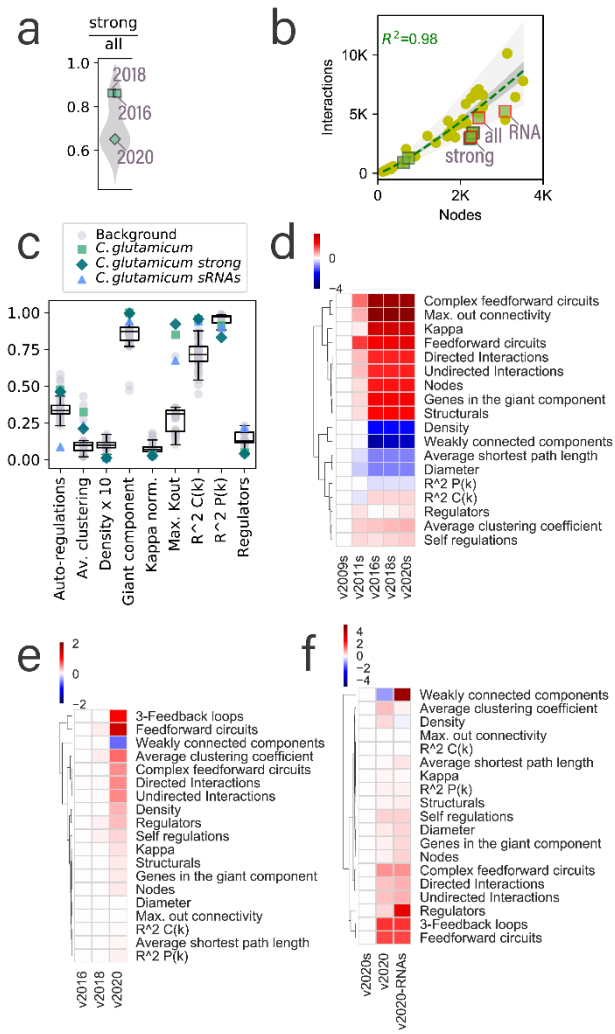


Figure 2. Structural properties of the *C. glutamicum* networks. (a) Distribution of the fraction of the strong interactions in the *all evidence* networks including at least one strong interaction. *C glutamicum* networks are highlighted and labeled. 2009 and 2011 versions of the *C. glutamicum* network are not included as they do not have a cognate *all evidence* network. (b) Inclusion of the three networks presented in this work in the previous model reported in [1] for the inference of the number of interactions for the regulatory networks. *C. glutamicum* networks are marked in green squares and the three networks reported in this work highlighted with a red outline and labeled. The rest of the data points (yellow dots) are the rest of the Abasy Atlas database used for reference. (c) Comparison of the *C. glutamicum* structural properties with the non-redundant set of bacterial networks. The boxplots were drawn including the non-redundant data set and the *C. glutamicum* networks reported in this work. (d) Heatmap values are the log2 fold change of the *C. glutamicum* regulatory networks for *strong* networks of the version 2011, 2016, 2018, and 2020 relative to the earliest strong version (2009). The v2009 column is included for clarity. Properties are clustered to ease the identification of those increasing, decreasing, or remaining virtually unchanged. Heatmaps (e) and (f) also represent the log2 fold change values relative to the leftmost column of (e) for versions of the *all evidence* network and (f) for the three different network models presented in this work to highlight the impact of the inclusion sRNAs-mediated interactions onto the structural properties of the network.

3.3. System-level components of the C. glutamicum regulatory networks

The regulation of gene transcription is organized in different hierarchical layers. Previously, we have described a large-scale modeling approach to characterize the nodes of a regulatory network, the NDA (natural decomposition approach). The NDA classifies each node of the network into one

of the four system-level components: GR, BM, Md, and IM. Regulatory networks having a diamond-shaped hierarchy have been found in different bacteria such as *E. coli* [23,24], *B. subtilis* [24], and a previous version of the *C. glutamicum* transcriptional regulatory network [13]. The hierarchy is divided into three layers (**Figure 3**a): the top layer composed solely of global regulators (coordination layer) is the smallest one and can directly regulate the four NDA classes. The middle layer (processing) is composed of Md and the BM, the two largest NDA components, both regulated by the coordination layer but only the Md class providing feedback to the top layer (i.e., some Md TFs regulates GRs). The last layer (integration) assimilates the combinatorial disparate signals provided by GR, and Md TFs belonging to different modules, into a single coordinated response essential to adapt to environmental changes.

Using the *all evidence* network as an example, the coordination layer is composed of nine GR (**Figure 3**a). As expected, the first GR when sorted by their K_{out} is the housekeeping σ factor (*sigA*), required for the transcription of 85% of the nodes in the network. Following, the dual regulator *hrrA* is involved in the transcription of 21% of the network. The rest of global regulators (and their corresponding rounded regulated network percentage) are *ramA* (11%), *glxR* (8%), *sigH* (6%), *ramB* (5%), *atlR* (4%), *mcbR* (4%), and *dtxR* (3%). The difference of regulated genes by the first and second global regulators is enormous, and this gap becomes smaller for the rest of the TFs. This is what provides the hierarchical structure to the network fitting a power-law distribution (a small fraction of nodes has most of the interactions). More than 66% of the *all evidence* networks are classified as BM. Examples of the BM are the *rpoA*, *rpoB*, *rpoC*, *rpoZ* genes coding for RNA polymerase subunits.

Please note that the BM class is constructed by non-regulators and is inferred based on their regulation solely by GR. Therefore, some of its members can be transferred to the Md or IM class if they are found to be regulated by a TF from the Md class. However, it is very unlikely for a structural gene belonging to the Md class to become part of the BM (because it requires to lose regulations mediated by an Md TF), and even less likely for IM because it would require the loss of at least two Md-mediated interactions. For these reasons, a regulatory network with high genomic coverage tends only to reduce the BM as more interactions are included. On the other hand, regulatory networks with low genomic coverage are highly likely to be lacking interactions by GR and their BM will increase with the genomic coverage. It was the case for the large increase in genomic coverage in a previous update of the *C. glutamicum* transcriptional regulatory network from 2011 (genomic coverage: ~24%) to 2016 (genomic coverage: ~71%), which was mainly due to the inclusion of the *sigA* sigmulon causing an increment of the BM from 60% to 77% of the network. The Md class is composed of ~28% (691/2441) of the network, divided into locally independent modules (see below). Finally, the IM class is composed of ~5% (117/2441) of the genes in the network, being all structural genes (non-regulators having $k_{out} = 0$).

The Md class is further divided into locally independent modules, groups of genes that are combinatorially expressed in response to specific media conditions. In the case of the *all evidence* network, the Md class is divided into 64 modules, 18 of them (28%) enriched with one or more biological functions (**Figure 3**b). We use a "guild-by-association" approach to assign a biological function to nodes that have no previous annotation due to poorly annotated orthologs but belong to enriched functionally modules (e.g., a module where all but one node has a GO annotation for DNA repair) [38]. The proportions for each NDA class are conserved in the network containing only strongly supported interactions, being BM the largest class followed by Md, IM, and lastly GR. On the other hand, when regulations mediated by sRNAs are integrated (*sRNA* network) to the *all evidence* network the proportions change for the BM and the Md classes, being the Md class the largest one (**Figure 3**c). The number of modules is largely increased with the inclusion of the sRNA regulations (**Figure 3**e), being an outlier in the distribution of the number of modules of bacterial regulatory networks, while the *strong* and *all evidence* networks have similar values. Even though the *sRNA* regulatory network is larger (figure 1) and every sRNA but *cgb_20925* is included in the Md class, this does not compensate for the number of modules in the network. This is observed when we compare the distribution of the size of the modules in the networks (**Figure 3**e). This is also a result of the sRNAs regulating many of the nodes that are solely regulated by *sigA* in the *all evidence*

network, transferring them from BM into the Md class and decreasing the BM class from 66.5% to 44.2% of the network.

Comparison of the size of the classes provides insights into their differences and similarities, contrasting the elements of each class contributes more to the comparative purpose. We used the Simpson similarity index to identify the overlap of two classes taking as reference the smallest one in each comparison. Thus, the Simpson similarity index for two sets, one being a subset of the other, is one. On the other hand, two sets having no overlap at all have an index of zero, and two sets where half of the smallest one is a subset of the largest one will have 0.5 as an index. For each NDA class, we computed the Simpson similarity index for every pair of networks and found that the *all evidence* and *sRNA* network are more similar to each other than to the *strong* network (**Figure 3**d). This is expected since the *all evidence* network is a subset of the *sRNA* network (supplementary figure 1). Please note that even though one network is a subset of the other, the NDA classification is performed independently for each network, therefore the class of a node can change from one network to another. Previous analysis of the robustness of the NDA classifications to random remotion of nodes and interactions showed the IM is the least conserved class [13]. Surprisingly, this was not the case in the class conservation across network models where the Md class was the least conserved (**Figure 3**d). This is caused by the inclusion of the sRNAs in the Md class. On the other hand, the similarity index of the IM class between the *all evidence* and the *sRNA* network is not affected because even though the number of intermodular nodes increased (from 117 to 194), one is a subset of the other. Consistent with the previous robustness analysis of the *C. glutamicum* network to random interactions remotion [13], the basal machinery is well conserved, while the GR is the most conserved class with a similarity index of one for the three comparisons between the networks. This is because the *all evidence* and the *sRNA* network have the same global regulators (listed in **Figure 3**a), and the *strong* network has four of these nine global regulators (*sigA*, *sigH*, *dtxR*, and *glxR*).

When analyzing the communication between classes (**Figure 3**f), most of the interactions in the network occur from GR → BM, followed by GR → Md, and Md → Md (regulations between modular TFs). For the *sRNA* network, the GR → BM is decreased, while the Md → Md interactions are increased due to the inclusion of the sRNAs in the Md class regulating nodes that used to be part of the BM but now are included in the Md class. The GR and IM classes have virtually the same fraction of regulations coming from GR and Md TFs in the *C. glutamicum* network, but further investigation in other organisms is required to assess the conservation of the proportions.

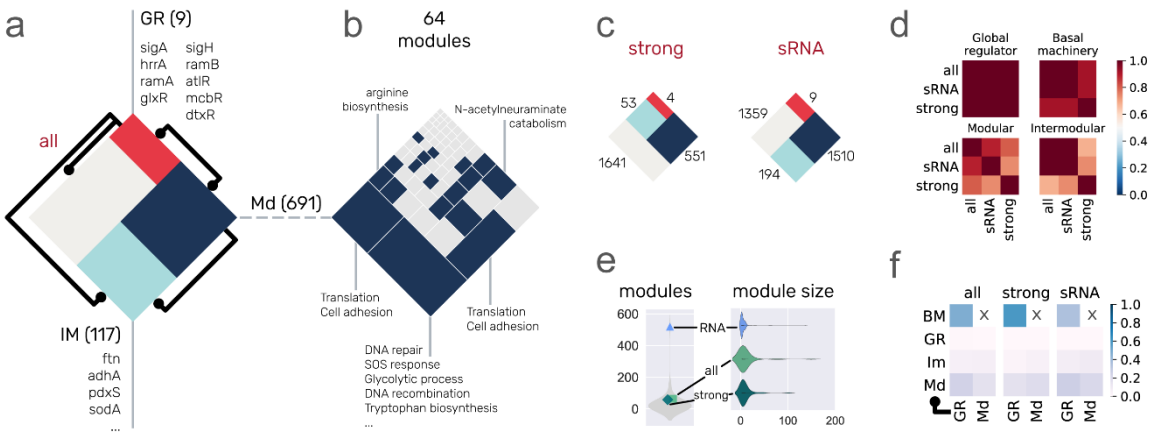


Figure 3. System-level classification of the networks. (a) The diamond represents the complete set of nodes in the network that are classified in one of the four classes: global regulators (red), modular (dark blue), intermodular (light blue), and basal machinery (gray, 1624 nodes). The size of the classes is proportional to the size of the *all evidence* network on a logarithmic scale. Black lines represent the interactions between the two classes. We listed the global regulators and some examples of intermodular. The modular class is further divided into 64 locally independent modules in the *all evidence* network (b). Modules enriched with a biological function are colored in blue. The size of the sections is proportional to the size of the modules. Similar to the *all evidence* network in panel (a),

panel (c) shows the proportion of the NDA classes for the *strong* and *sRNA* networks. (d) Heatmaps of similarity index between the three *C. glutamicum* networks for each one of the four NDA classes. The color bar shows that more than half of the nodes in the class are conserved for each class among the three networks, showing the precision of the network nodes classification. (e) Distribution of the number of modules and their size. Light gray distribution of the number of modules was drawn using the non-redundant set of networks including the three *C. glutamicum* networks. (f) The fraction of the network interactions between the four classes for each one of the *C. glutamicum* networks.

3.4. Recovering conserved interactions from other model organisms

The regulogs analysis is based on the premise that an interaction TF–TG from organism A is conserved in organism B iff B has an ortholog of the TF, an ortholog of the TG, and a binding site for TF in the promoter region of TG [11]. Because regulatory networks are highly plastic, a caveat of the regulogs analysis is the functional divergence of one of the components involved in the interaction, especially for the TF [39]. Therefore, this analysis is usually applied between phylogenetically closely related organisms and is useful to transfer interactions from one model organism to others, for example from *C. glutamicum* to other Corynebacteriales [15]. However, model organisms for the study of regulatory networks are phylogenetically far from each other, which allows the transfer of interactions from model organisms across several bacterial genera [10,15,40]. We restricted our source organisms to purely *strong* networks as they only contain directed TF-DNA interactions supported by at least strong evidence. Namely, *E. coli*, *B. subtilis*, and *S. coelicolor*. Please note that despite the high completeness of the network for *M. tuberculosis* [1], and the closeness to *C. glutamicum* in the phylogeny, we did not use this network as a source since it was mainly constructed using only high-throughput technologies without further confirmation with directed experiments. This causes an unusually lower clustering coefficient for the network (see figure 5 in [6]). Besides, *C. glutamicum* has been used as a source organism for the inference of regulatory interactions in *M. tuberculosis* [41]. We acknowledge the caveats of using distant organisms for the regulogs analysis, for this reason, we applied strict conditions during the complete workflow prioritizing precision at expense of losing many potential interactions.

Using *S. coelicolor*, *B. subtilis*, and *E. coli* as source organisms (**Figure 4**a), we aimed to identify conserved interactions despite their phylogenetic distance (especially for *B. subtilis* and *E. coli*). To do so, first, we identified the pair-wise genome-wide orthologs between the source organisms and *E. coli* with the OMA standalone software [30], and we kept only the one-to-one orthology relationships as they have a higher probability to be *bona fide* orthologs, more likely to conserve their functions [42]. We kept 1,117 one-to-one orthology relationships for *S. coelicolor* out of the total 2,480 (45%), 661 out of 1,480 (45%) for *B. subtilis*, and 641 out of 1,488 (43%) for *E. coli* (**Figure 4**b). There is a greater number of one-to-one orthologous genes with *S. coelicolor*, as expected due to its phylogenetic closeness compared with the other two source organisms. Just filtering orthologs, we restraint more than 50% of nodes to be included in the transferred interactions. The next filter is due to the completeness in the source networks since we can only transfer interactions between nodes already present in the source networks (**Figure 4**c). From there, we were primarily interested in TFs (white inner circles in **Figure 4**c) but we only considered those with at least one TG with a one-to-one ortholog in *C. glutamicum*, resulting in a total of 8, 7, and 13 potential TFs/regulons to be transferred from *S. coelicolor*, *B. subtilis*, and *E. coli*, respectively (colored inner circles in **Figure 4**c). However, the number of potential interactions to be transferred was reduced when we search for a TF binding site in the promoter sequences of the orthologous TG in *C. glutamicum*. 24 out of the 479 interactions from the *S. coelicolor* network were conserved along with the TF binding site, 17 out of 2,576 from the *B. subtilis* network, and 70 out of 4,653 from the *E. coli* network. We recovered more regulogs from *E. coli*, due to the completeness of the source network. We lost many interactions through the stringent filters we applied, but we expect those conserved interactions to be true positives. As mentioned above, the main goal of this interactions transfer is to detect interactions for the *C. glutamicum* TFs that are still missing in the network (supplementary figure 3) despite the

exhaustive work of the community to model the network. We retrieve interactions for a total of five DNA-binding TF not considered in the current curation state of the network (**Figure 4**e). Given that *C. glutamicum* regulatory mechanism is already one of the most studied and curated (supplementary figure 2), most of the TFs that were retrieved from regulogs were already present in the *all evidence* network (**Figure 4**e). However, in terms of interactions, 82 out of the 111 interactions were not present in any of the *C. glutamicum* curated networks (**Figure 4**f). There was poor overlap between the regulogs obtained from each organism and only one common TF between *E. coli* and *S. coelicolor* (Zur), and one between *E. coli* and *B. subtilis* (LexA) (**Figure 4**g-h).

Following, we describe some of the conserved regulations in *C. glutamicum*. From *S. coelicolor*, two interactions were conserved. The interaction of Zur (*cg2502*) regulating *cg0042* is already part of the *strong* network (supplementary figure 4). The other interaction is by RegX3 (*cg0484*), an essential response regulator of the SenX3-RegX3 two-component system [43]. RegX3 has a one-to-one orthology relationship with PhoP (*SCO4230*) from *S. coelicolor*, as well as the gene *amtB* (*cg2261*) with *SCO5583* for which a regulatory site for the PhoP ortholog in their upstream region is conserved. However, the interaction could not be transferred from *E. coli* or *B. subtilis* because a many-to-many orthology relationship was found for RegX3 in both organisms and therefore discarded. RegX3 has been characterized as a gene coding a regulator of phosphate-dependent gene expression in *Mycobacterium smegmatis* [44] and required for virulence in *M. tuberculosis* [45], but its regulon has not been characterized in *C. glutamicum*. PhoP represses *amtB* and other nitrogen genes in *S. coelicolor* [46]. Previous work showed that *amtB* is required for ammonium uptake in *C. glutamicum* [47]. A binding site for PhoP was found 87-69 bp upstream of the *cg2261* translation start codon. This agrees with the mechanism of *amtB* regulation in *S. coelicolor* binding upstream of the CDS and repressing its transcription by regulating a promoter in the upstream sequence from the binding site [46]. From *B. subtilis* an auto-regulation was fully conserved for LexA, which was already part of the *strong* network (supplementary figure 5). Cg1098 is an ortholog of SCO3129, a TetR family regulator involved in *S. coelicolor* osmotic stress [48]. In *S. coelicolor* it regulates the transcription of two (*SCO3128* and *SCO3130*) genes and its own. However, only the auto-regulation was fully conserved in *C. glutamicum*. Most of the characterized TetR family regulators regulate their own transcription [49].

From *E. coli* we recovered a total of 11 interactions, three of them already included in the *C. glutamicum* regulatory network. ArgR regulating *argC* (supplementary figure 6), LexA (cg2114) regulating *recA* (supplementary figure 7), and NrdR regulating *nrdI* (supplementary figure 8). While the first two interactions are already included in the *strong* network, the latter only is included in the *all evidence* network. The gene *cg1327* has *b1334* as an ortholog, coding for the FNR global regulator in *E. coli*. For this protein, the regulation of *hmp* (*cg3141*) and the auto-regulation were fully conserved. However, the *cg1327* gene is currently part of the basal machinery in the *C. glutamicum* network due to unreported characterization of its regulon. The gene *cg2899* codes for a regulator of the LysR family and is an ortholog of *b2537* (HcaR) in *E. coli*, regulating *hcaE* which is an ortholog of *cg2637* (*benA*) in *C. glutamicum*, only regulated by GlxR and BenR in the *all evidence* network. In contrast with *C. glutamicum*, in *E. coli hcaR* and *hcaE* are divergently transcribed sharing the same promoter recognized by HcaR. The gene *cg0350* encodes for GlxR ortholog to CRP in *E. coli*, both being global regulators in their corresponding networks. The regulation of CRP to *dadA* (*b1189*) is fully conserved in *C. glutamicum* for their orthologs GlxR regulating *cg3340*, which is currently regulated only by SigA. The other TG conserved is *cg2175* (with *b3167* as its ortholog in *E. coli*), which codes for a ribosome binding protein. However, none of the two targets were identified in a previous *in silico* analysis of the GlxR regulon in *C. glutamicum* [50]. The gene *cg1425* coding for LysG (ArgP encoded by *b2916* in *E. coli*) regulates *dnaA* that is not part of the current *C. glutamicum* network. However, none of the three interactions were conserved in *C. glutamicum*. DnaA, besides being the protein for the DNA replication initiation, is a transcriptional regulator that controls the transcription of its own coding gene and at least 10 others in *E. coli*. The auto-regulation and the regulation of the other four genes (*cg0004*, *cg0005*, *cg1525*, and *cg1550*) were fully conserved in *C. glutamicum* (supplementary table 2). Zur is encoded by *cg2502*, ortholog to *b0683* in *E. coli*. A regulation from Zur

to *cg2183* was recovered from the *oppC* gene in *E. coli*. The interactions are not part of the current networks for *C. glutamicum*. LldR is encoded by *cg3224*, ortholog to *b2980* (*glcC* in *E. coli*) that regulates *glcB*. The interaction was conserved in *C. glutamicum,* but not present in the current networks. Although LldR regulon has already 12 TGs.

These results show that even though some interactions already known in *C. glutamicum* are recovered, the rate of recovered interactions is low. Therefore, for long phylogenetic distances, it might be better to discriminate false positives after a mildly lax prediction. We noticed that most of the interactions are lost due to the conservative approach of using only one-to-one orthologs. A potential solution for this is the use of other orthology relationships with subsequent discrimination of false positives through the conservation of regulogs not only in *C. glutamicum* but also in other closely related organisms, conferring greater confidence values to those interactions conserved in more organisms.

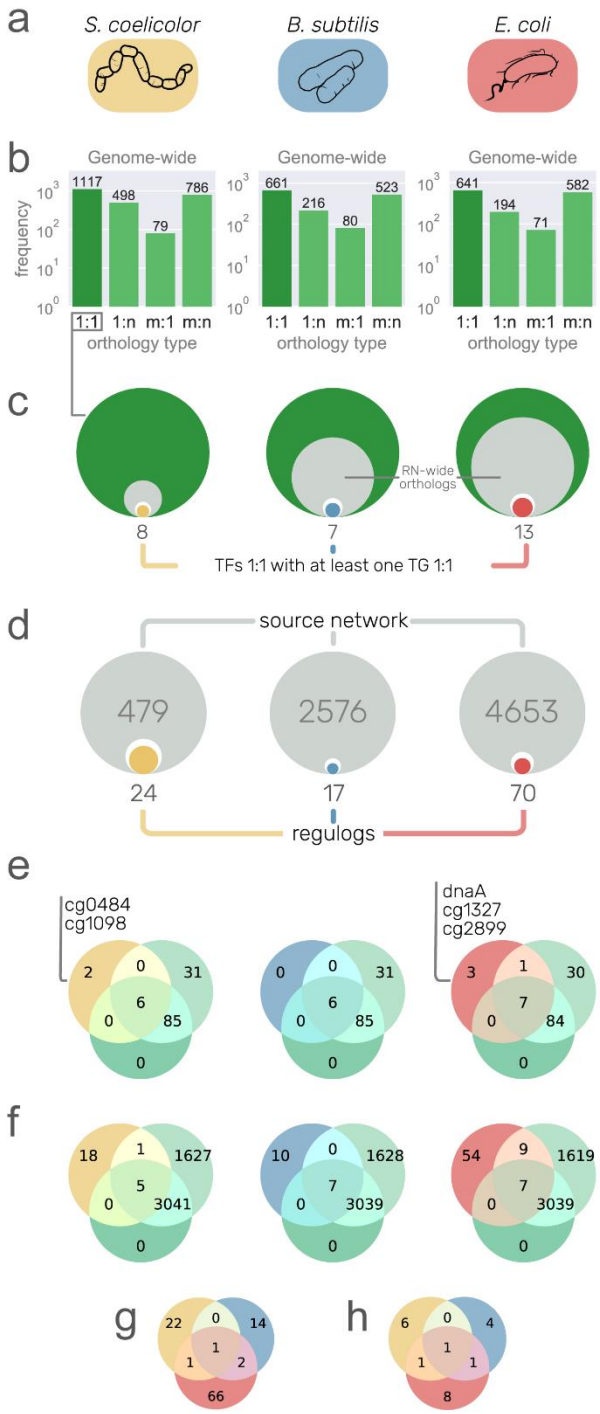


Figure 4. Putative regulons from other model organisms. (a) Networks with strong interactions of *S. coelicolor*, *B. subtilis*, and *E. coli*. used as a source of information. Rounded rectangles color is used to relate the organism in the rest of the figure. (b) Orthology relationship type between the source organisms and *C. glutamicum*. Only one-to-one relationships were used for downstream analysis. (c) Size comparison between the one-to-one orthology genes (green circles), the orthologs with at least one interaction in the source network (inner gray circle), transcription factor (TF) orthologs (inner white circle), and TF orthologs with at least one target gene (TG) with one-to-one orthology relationship (inner colored circles and numbers). (d) Size comparison between the source networks (gray circles with large gray numbers), TF–TG pairs conserved both as orthologs one-to-one in *C. glutamicum* (inner white circles), and the interactions conserved with a TF binding site in the promoter region of the TG (colored inner circles and numbers of regulogs). (e) Venn diagrams

showing the overlap of TFs between three sets: the *strong* network (green circle), the *all evidence* network (light green circle), and the interactions from the source organisms with the unique TFs listed. (f) Venn diagrams showing the overlap of interactions between the *strong* network, the *all evidence* network, and the regulogs network from the source organisms. (g) Euler diagrams showing no overlap between the regulogs and (h) poor overlap between their TFs.

4. Conclusions

In this work, we updated the *C. glutamicum* regulatory network by manual curation of literature. We also went beyond the regulation of transcription initiation to incorporate regulations mediated by protein-protein interactions and small RNAs. Three network models with different confidence levels were reconstructed and deposited in the new v2.4 of Abasy Atlas (https://abasy.ccg.unam.mx). Poor efforts have been carried out to provide consolidated, disambiguated, homogenized high-quality regulatory networks on a global scale, their structural properties, system-level components, and their historical snapshots to trace their curation process. We originally conceived Abasy Atlas to fill this gap by making a cartography of the functional architectures of the regulatory networks for a wide range of bacteria.

This work provides the most complete and reliable set of *C. glutamicum* regulatory networks, which can be used as the gold standard for benchmarking purposes and training data for modeling. The *C. glutamicum* regulatory networks have been meta-curated to avoid heterogeneity such as inconsistencies in gene symbols and heteromeric regulatory complexes representation. This enables large-scale comparative systems biology studies to understand the common principles and particular lifestyle adaptations of regulatory systems across bacteria and to implement those principles into future work such as the reverse engineering of regulatory networks. The historical snapshots deposited in Abasy Atlas allow to carry out network analyses at different incompleteness levels, making it possible to identify how a methodology is affected, to pinpoint potential bias and improvements, and to predict future results. Regulatory network models, gene information, and modules annotation can be downloaded from the "Downloads" section on Abasy Atlas (https://abasy.ccg.unam.mx/downloads). The same web page provides useful information about the downloadable files.

Supplementary Materials: Supplementary_File: contains images to support the discussion of the main text, and a table with several interactions in strains of *C. glutamicum* other than ATCC 13032.

Availability of supporting data: The data sets supporting the results of this article are available in the Abasy Atlas repository [https://abasy.ccg.unam.mx].

Author Contributions: Conceptualization, J.A.F.G.; Methodology, J.M.E.R., A.T. and J.A.F.G.; Software, J.M.E.R. and J.A.F.G; Validation, J.M.E.R., A.T. and J.A.F.G.; Formal Analysis, J.M.E.R. and J.A.F.G.; Investigation, J.M.E.R., A.T. and J.A.F.G.; Resources, J.A.F.G.; Data Curation, J.M.E.R., A.T. and J.A.F.G.; Writing – Original Draft Preparation, J.M.E.R.; Writing – Review & Editing, J.M.E.R., A.T. and J.A.F.G.; Visualization, J.M.E.R.; Supervision, J.A.F.G.; Project Administration, J.A.F.G.; Funding Acquisition, J.A.F.G.

Funding: This work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT-UNAM) [IN205918 to J.A.F.G.].

Acknowledgments: J.M.E.R. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 959406 from CONACYT.

Conflicts of Interest: The authors declare no conflict of interest.

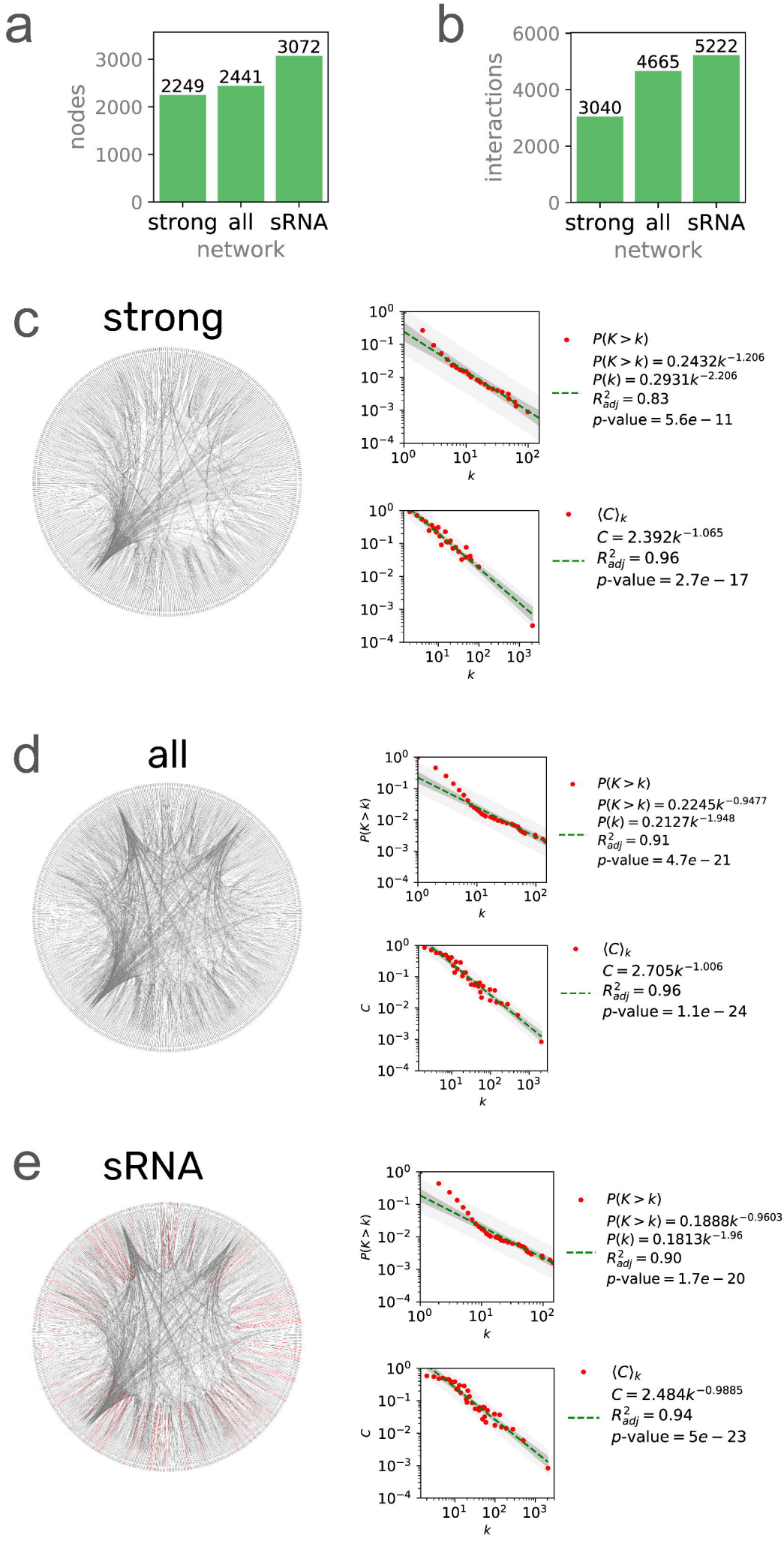
References

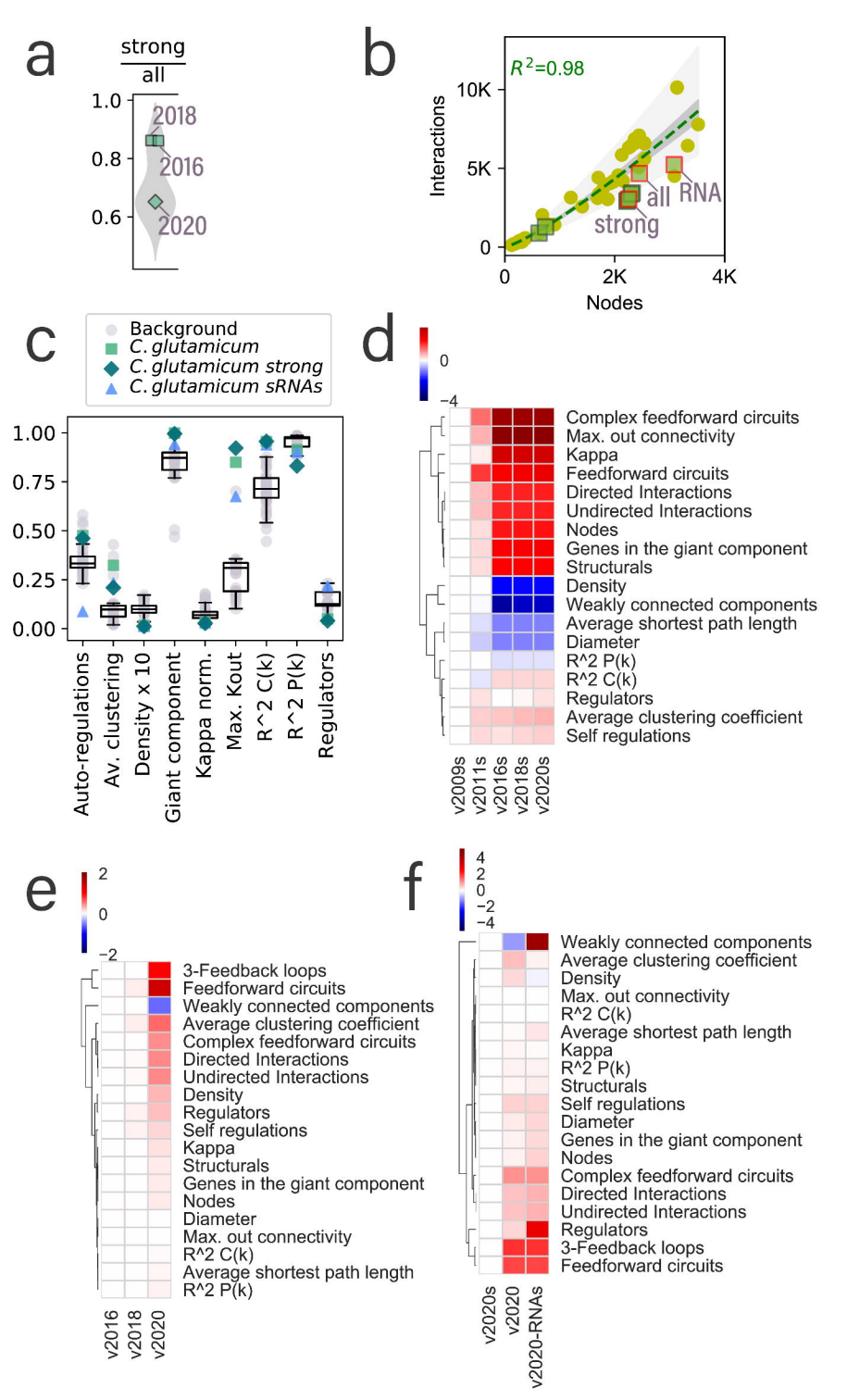
1. Escorcia-Rodriguez, J.M.; Tauch, A.; Freyre-Gonzalez, J.A. Abasy Atlas v2.2: The most comprehensive and up-to-date inventory of meta-curated, historical, bacterial regulatory networks, their completeness and system-level characterization. *Comput Struct Biotechnol J* **2020**, *18*, 1228-1237, doi:10.1016/j.csbj.2020.05.015.

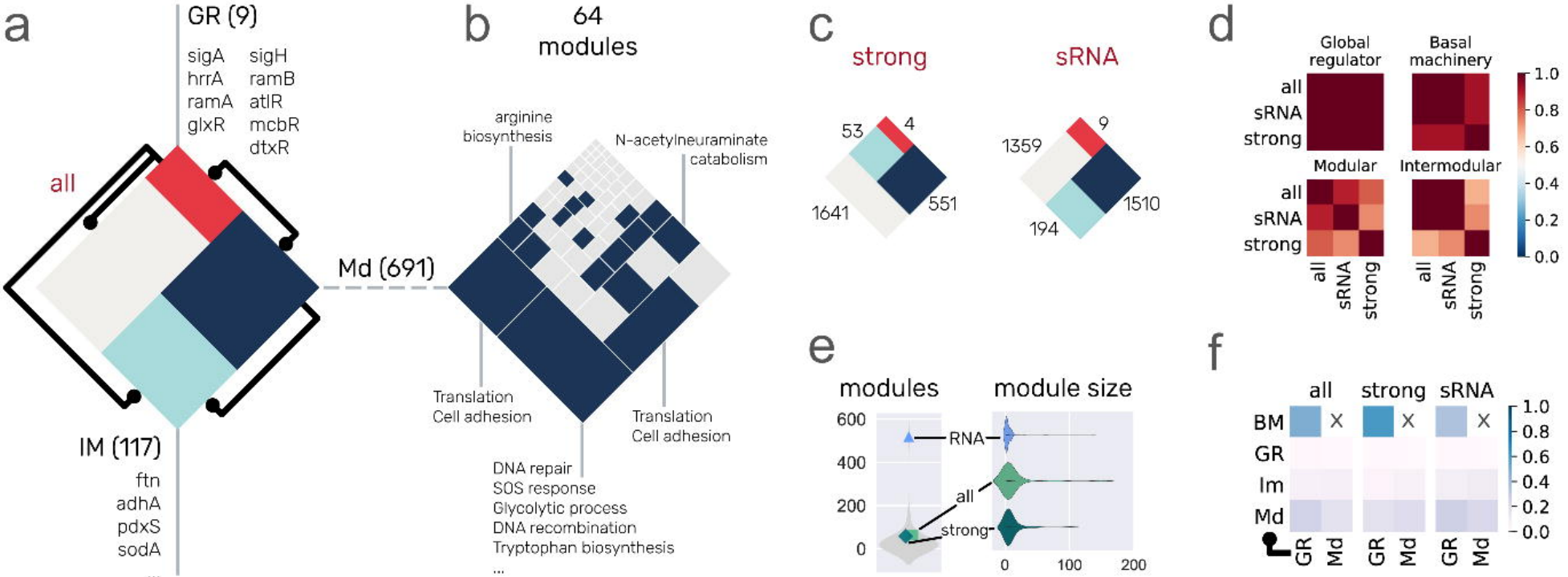
2. Donovan, C.; Schauss, A.; Kramer, R.; Bramkamp, M. Chromosome segregation impacts on cell growth and division site selection in Corynebacterium glutamicum. *PLoS One* **2013**, *8*, e55078, doi:10.1371/journal.pone.0055078.
3. Toyoda, K.; Inui, M. Global Transcriptional Regulators Involved in Carbon, Nitrogen, Phosphorus, and Sulfur Metabolisms in Corynebacterium glutamicum. In *Corynebacterium glutamicum: Biology and Biotechnology*, Inui, M., Toyoda, K., Eds. Springer International Publishing: Cham, 2020; 10.1007/978-3-030-39267-3_5pp. 113-147.
4. Brinkrolf, K.; Schroder, J.; Puhler, A.; Tauch, A. The transcriptional regulatory repertoire of Corynebacterium glutamicum: reconstruction of the network controlling pathways involved in lysine and glutamate production. *J Biotechnol* **2010**, *149*, 173-182, doi:10.1016/j.jbiotec.2009.12.004.
5. Pátek, M.; Dostálová, H.; Nešvera, J. Sigma Factors of RNA Polymerase in Corynebacterium glutamicum. In *Corynebacterium glutamicum: Biology and Biotechnology*, Inui, M., Toyoda, K., Eds. Springer International Publishing: Cham, 2020; 10.1007/978-3-030-39267-3_4pp. 89-111.
6. Campos, A.I.; Freyre-Gonzalez, J.A. Evolutionary constraints on the complexity of genetic regulatory networks allow predictions of the total number of genetic interactions. *Sci Rep* **2019**, *9*, 3618, doi:10.1038/s41598-019-39866-z.
7. Marbach, D.; Costello, J.C.; Kuffner, R.; Vega, N.M.; Prill, R.J.; Camacho, D.M.; Allison, K.R.; Consortium, D.; Kellis, M.; Collins, J.J., et al. Wisdom of crowds for robust gene network inference. *Nat Methods* **2012**, *9*, 796-804, doi:10.1038/nmeth.2016.
8. Larsen, S.J.; Rottger, R.; Schmidt, H.; Baumbach, J. E. coli gene regulatory networks are inconsistent with gene expression data. *Nucleic Acids Res* **2019**, *47*, 85-92, doi:10.1093/nar/gky1176.
9. Tan, K.; McCue, L.A.; Stormo, G.D. Making connections between novel transcription factors and their DNA motifs. *Genome Res* **2005**, *15*, 312-320, doi:10.1101/gr.3069205.
10. Rodionov, D.A. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem Rev* **2007**, *107*, 3467-3497, doi:10.1021/cr068309+.
11. Alkema, W.B.; Lenhard, B.; Wasserman, W.W. Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. *Genome Res* **2004**, *14*, 1362-1373, doi:10.1101/gr.2242604.
12. Kilic, S.; Erill, I. Assessment of transfer methods for comparative genomics of regulatory networks in bacteria. *BMC Bioinformatics* **2016**, *17 Suppl 8*, 277, doi:10.1186/s12859-016-1113-7.
13. Freyre-Gonzalez, J.A.; Tauch, A. Functional architecture and global properties of the Corynebacterium glutamicum regulatory network: Novel insights from a dataset with a high genomic coverage. *J Biotechnol* **2017**, *257*, 199-210, doi:10.1016/j.jbiotec.2016.10.025.
14. Nitzan, M.; Rehani, R.; Margalit, H. Integration of Bacterial Small RNAs in Regulatory Networks. *Annu Rev Biophys* **2017**, *46*, 131-148, doi:10.1146/annurev-biophys-070816-034058.
15. Parise, M.T.D.; Parise, D.; Kato, R.B.; Pauling, J.K.; Tauch, A.; Azevedo, V.A.C.; Baumbach, J. CoryneRegNet 7, the reference database and analysis platform for corynebacterial gene regulatory networks. *Sci Data* **2020**, *7*, 142, doi:10.1038/s41597-020-0484-9.
16. de Witt, J.; Oetermann, S.; Parise, M.; Parise, D.; Baumbach, J.; Steinbuchel, A. Global Regulator of Rubber Degradation in Gordonia polyisoprenivorans VH2: Identification and Involvement in the Regulation Network. *Appl Environ Microbiol* **2020**, *86*, doi:10.1128/AEM.00774-20.
17. Newman, M.E.J. The Structure and Function of Complex Networks. *SIAM Review* **2003**, *45*, 167-256, doi:10.1137/S003614450342480.

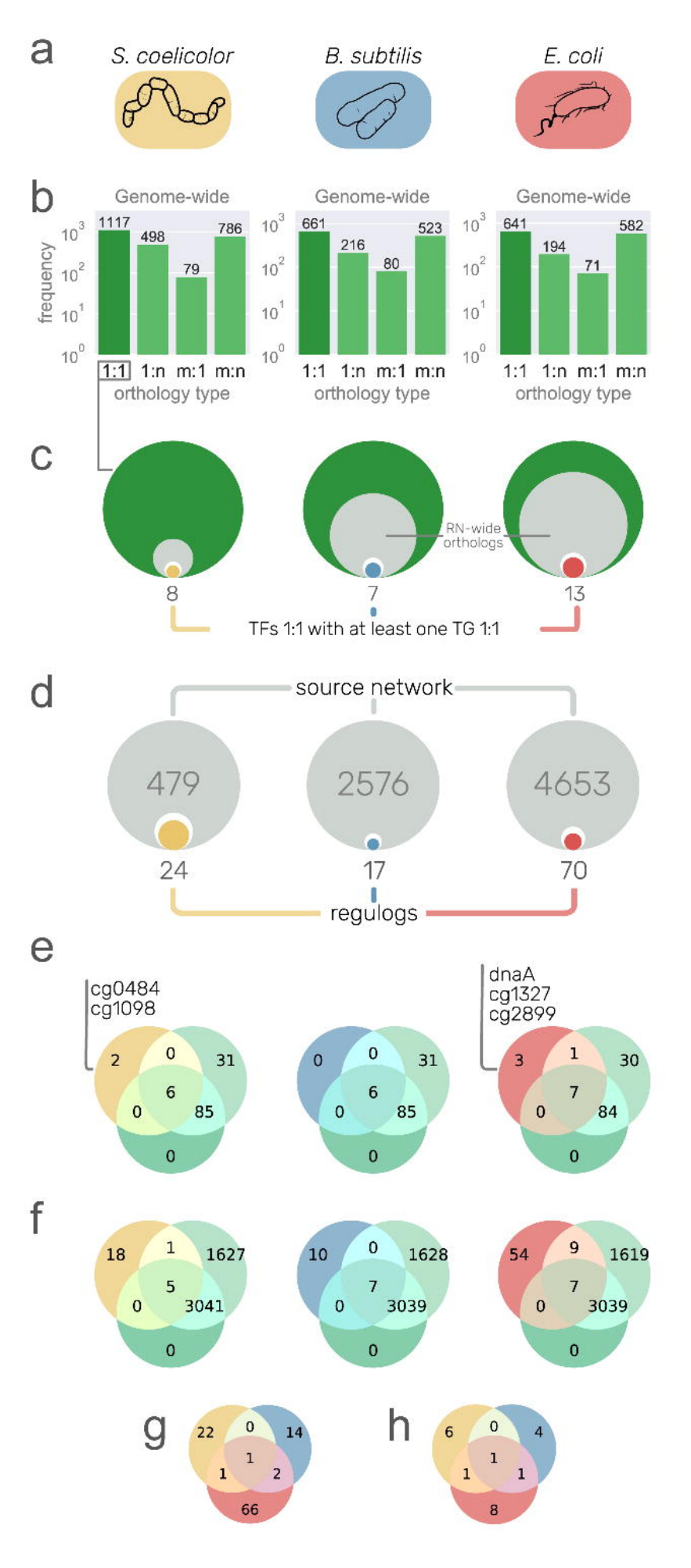
18. Dorogovtsev, S.N.; Mendes, J.F.F. The shortest path to complex networks. *arXiv:cond-mat/0404593* **2004**.
19. Barabasi, A.L.; Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **2004**, *5*, 101-113, doi:10.1038/nrg1272.
20. Mason, O.; Verwoerd, M. Graph Theory and Networks in Biology. *arXiv:q-bio/0604006* **2006**.
21. Costa, L.d.F.; Rodrigues, F.A.; Travieso, G.; Boas, P.R.V. Characterization of complex networks: A survey of measurements. *Advances in Physics* **2007**, *56*, 167-242, doi:10.1080/00018730601170527.
22. Lima-Mendez, G.; van Helden, J. The powerful law of the power law and other myths in network biology. *Mol Biosyst* **2009**, *5*, 1482-1493, doi:10.1039/b908681a.
23. Freyre-Gonzalez, J.A.; Alonso-Pavon, J.A.; Trevino-Quintanilla, L.G.; Collado-Vides, J. Functional architecture of Escherichia coli: new insights provided by a natural decomposition approach. *Genome Biol* **2008**, *9*, R154, doi:10.1186/gb-2008-9-10-r154.
24. Freyre-Gonzalez, J.A.; Trevino-Quintanilla, L.G.; Valtierra-Gutierrez, I.A.; Gutierrez-Rios, R.M.; Alonso-Pavon, J.A. Prokaryotic regulatory systems biology: Common principles governing the functional architectures of Bacillus subtilis and Escherichia coli unveiled by the natural decomposition approach. *J Biotechnol* **2012**, *161*, 278-286, doi:10.1016/j.jbiotec.2012.03.028.
25. Mentz, A.; Neshat, A.; Pfeifer-Sancar, K.; Puhler, A.; Ruckert, C.; Kalinowski, J. Comprehensive discovery and characterization of small RNAs in Corynebacterium glutamicum ATCC 13032. *BMC Genomics* **2013**, *14*, 714, doi:10.1186/1471-2164-14-714.
26. Coordinators, N.R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **2018**, *46*, D8-D13, doi:10.1093/nar/gkx1095.
27. Nguyen, N.T.T.; Contreras-Moreira, B.; Castro-Mondragon, J.A.; Santana-Garcia, W.; Ossio, R.; Robles-Espinoza, C.D.; Bahin, M.; Collombet, S.; Vincens, P.; Thieffry, D., et al. RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res* **2018**, *46*, W209-W214, doi:10.1093/nar/gky317.
28. Baumbach, J.; Rahmann, S.; Tauch, A. Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms. *BMC Syst Biol* **2009**, *3*, 8, doi:10.1186/1752-0509-3-8.
29. Krawczyk, J.; Kohl, T.A.; Goesmann, A.; Kalinowski, J.; Baumbach, J. From Corynebacterium glutamicum to Mycobacterium tuberculosis--towards transfers of gene regulatory networks and integrated data analyses with MycoRegNet. *Nucleic Acids Res* **2009**, *37*, e97, doi:10.1093/nar/gkp453.
30. Altenhoff, A.M.; Levy, J.; Zarowiecki, M.; Tomiczek, B.; Warwick Vesztrocy, A.; Dalquen, D.A.; Muller, S.; Telford, M.J.; Glover, N.M.; Dylus, D., et al. OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res* **2019**, *29*, 1152-1163, doi:10.1101/gr.243212.118.
31. Bailey, T.L.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **1994**, *2*, 28-36.
32. Liu, X.; Brutlag, D.L.; Liu, J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* **2001**, 127-138.
33. Liu, X.S.; Brutlag, D.L.; Liu, J.S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* **2002**, *20*, 835-839, doi:10.1038/nbt717.
34. Grant, C.E.; Bailey, T.L.; Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **2011**, *27*, 1017-1018, doi:10.1093/bioinformatics/btr064.
35. Lastiri-Pancardo, G.; Mercado-Hernandez, J.S.; Kim, J.; Jimenez, J.I.; Utrilla, J. A quantitative method for proteome reallocation using minimal regulatory interventions. *Nat Chem Biol* **2020**, *16*, 1026-1033, doi:10.1038/s41589-020-0593-y.

36. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: an information aesthetic for comparative genomics. *Genome Res* **2009**, *19*, 1639-1645, doi:10.1101/gr.092759.109.
37. Rosenfeld, N.; Elowitz, M.B.; Alon, U. Negative autoregulation speeds the response times of transcription networks. *J Mol Biol* **2002**, *323*, 785-793, doi:10.1016/s0022-2836(02)00994-4.
38. Ibarra-Arellano, M.A.; Campos-Gonzalez, A.I.; Trevino-Quintanilla, L.G.; Tauch, A.; Freyre-Gonzalez, J.A. Abasy Atlas: a comprehensive inventory of systems, global network properties and systems-level elements across bacteria. *Database (Oxford)* **2016**, *2016*, doi:10.1093/database/baw089.
39. Price, M.N.; Dehal, P.S.; Arkin, A.P. Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLoS Comput Biol* **2007**, *3*, 1739-1750, doi:10.1371/journal.pcbi.0030175.
40. Novichkov, P.S.; Kazakov, A.E.; Ravcheev, D.A.; Leyn, S.A.; Kovaleva, G.Y.; Sutormin, R.A.; Kazanov, M.D.; Riehl, W.; Arkin, A.P.; Dubchak, I., et al. RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics* **2013**, *14*, 745, doi:10.1186/1471-2164-14-745.
41. Baumbach, J. CoryneRegNet 4.0 - A reference database for corynebacterial gene regulatory networks. *BMC Bioinformatics* **2007**, *8*, 429, doi:10.1186/1471-2105-8-429.
42. Altenhoff, A.M.; Studer, R.A.; Robinson-Rechavi, M.; Dessimoz, C. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol* **2012**, *8*, e1002514, doi:10.1371/journal.pcbi.1002514.
43. Bott, M.; Brocker, M. Two-component signal transduction in Corynebacterium glutamicum and other corynebacteria: on the way towards stimuli and targets. *Appl Microbiol Biotechnol* **2012**, *94*, 1131-1150, doi:10.1007/s00253-012-4060-x.
44. Glover, R.T.; Kriakov, J.; Garforth, S.J.; Baughn, A.D.; Jacobs, W.R., Jr. The two-component regulatory system senX3-regX3 regulates phosphate-dependent gene expression in Mycobacterium smegmatis. *J Bacteriol* **2007**, *189*, 5495-5503, doi:10.1128/JB.00190-07.
45. Parish, T.; Smith, D.A.; Roberts, G.; Betts, J.; Stoker, N.G. The senX3-regX3 two-component regulatory system of Mycobacterium tuberculosis is required for virulence. *Microbiology (Reading)* **2003**, *149*, 1423-1435, doi:10.1099/mic.0.26245-0.
46. Rodriguez-Garcia, A.; Sola-Landa, A.; Apel, K.; Santos-Beneit, F.; Martin, J.F. Phosphate control over nitrogen metabolism in Streptomyces coelicolor: direct and indirect negative control of glnR, glnA, glnII and amtB expression by the response regulator PhoP. *Nucleic Acids Res* **2009**, *37*, 3230-3242, doi:10.1093/nar/gkp162.
47. Walter, B.; Kuspert, M.; Ansorge, D.; Kramer, R.; Burkovski, A. Dissection of ammonium uptake systems in Corynebacterium glutamicum: mechanism of action and energetics of AmtA and AmtB. *J Bacteriol* **2008**, *190*, 2611-2614, doi:10.1128/JB.01896-07.
48. He, X.; Li, H.; Pan, Y.; Wang, L.; Tan, H.; Liu, G. SCO3129, a TetR family regulator, is responsible for osmotic stress in Streptomyces coelicolor. *Synth Syst Biotechnol* **2018**, *3*, 261-267, doi:10.1016/j.synbio.2018.10.012.
49. Cuthbertson, L.; Nodwell, J.R. The TetR family of regulators. *Microbiol Mol Biol Rev* **2013**, *77*, 440-475, doi:10.1128/MMBR.00018-13.
50. Kohl, T.A.; Tauch, A. The GlxR regulon of the amino acid producer Corynebacterium glutamicum: Detection of the corynebacterial core regulon and integration into the transcriptional regulatory network model. *J Biotechnol* **2009**, *143*, 239-246, doi:10.1016/j.jbiotec.2009.08.005.


a
nodes
3000
2000
1000
0
2249
2441
3072
strong
all
sRNA
network
b
interactions
6000
4000
2000
0
3040
4665
5222
strong
all
sRNA
network
c
strong
$P(K>k)$
$P(K>k) = 0.2432k^{-1.206}$
$P(k) = 0.2931k^{-2.206}$
$R^2_{adj} = 0.83$
p-value $= 5.6e-11$
k
$\langle C \rangle_k$
$C = 2.392k^{-1.065}$
$R^2_{adj} = 0.96$
p-value $= 2.7e-17$
k
d
all
$P(K>k)$
$P(K>k) = 0.2245k^{-0.9477}$
$P(k) = 0.2127k^{-1.948}$
$R^2_{adj} = 0.91$
p-value $= 4.7e-21$
k
$\langle C \rangle_k$
$C = 2.705k^{-1.006}$
$R^2_{adj} = 0.96$
p-value $= 1.1e-24$
k
e
sRNA
$P(K>k)$
$P(K>k) = 0.1888k^{-0.9603}$
$P(k) = 0.1813k^{-1.96}$
$R^2_{adj} = 0.90$
p-value $= 1.7e-20$
k
$\langle C \rangle_k$
$C = 2.484k^{-0.9885}$
$R^2_{adj} = 0.94$
p-value $= 5e-23$
k


a
strong
all
1.0
0.8
0.6
2018
2016
2020
b
R^2=0.98
10K
5K
0
Interactions
0
2K
4K
Nodes
all
RNA
strong
c
Background
C. glutamicum
C. glutamicum strong
C. glutamicum sRNAs
1.00
0.75
0.50
0.25
0.00
Auto-regulations
Av. clustering
Density x 10
Giant component
Kappa norm.
Max. Kout
R^2 C(k)
R^2 P(k)
Regulators
d
0
−4
Complex feedforward circuits
Max. out connectivity
Kappa
Feedforward circuits
Directed Interactions
Undirected Interactions
Nodes
Genes in the giant component
Structurals
Density
Weakly connected components
Average shortest path length
Diameter
R^2 P(k)
R^2 C(k)
Regulators
Average clustering coefficient
Self regulations
v2009s
v2011s
v2016s
v2018s
v2020s
e
2
0
−2
3-Feedback loops
Feedforward circuits
Weakly connected components
Average clustering coefficient
Complex feedforward circuits
Directed Interactions
Undirected Interactions
Density
Regulators
Self regulations
Kappa
Structurals
Genes in the giant component
Nodes
Diameter
Max. out connectivity
R^2 C(k)
Average shortest path length
R^2 P(k)
v2016
v2018
v2020
f
4
2
0
−2
−4
Weakly connected components
Average clustering coefficient
Density
Max. out connectivity
R^2 C(k)
Average shortest path length
Kappa
R^2 P(k)
Structurals
Self regulations
Diameter
Genes in the giant component
Nodes
Complex feedforward circuits
Directed Interactions
Undirected Interactions
Regulators
3-Feedback loops
Feedforward circuits
v2020s
v2020
v2020-RNAs


a
GR (9)
sigA
hrrA
ramA
glxR
sigH
ramB
atlR
mcbR
dtxR
all
Md (691)
IM (117)
ftn
adhA
pdxS
sodA
...
b
64 modules
arginine biosynthesis
N-acetylneuraminate catabolism
Translation
Cell adhesion
Translation
Cell adhesion
DNA repair
SOS response
Glycolytic process
DNA recombination
Tryptophan biosynthesis
...
c
strong
sRNA
53
4
1641
551
1359
9
194
1510
d
Global regulator
Basal machinery
Modular
Intermodular
all
sRNA
strong
1.0
0.8
0.6
0.4
0.2
0.0
e
modules
module size
600
400
200
0
RNA
all
strong
0
100
200
f
all
strong
sRNA
BM
GR
Im
Md
GR
Md
1.0
0.8
0.6
0.4
0.2
0.0


a
S. coelicolor
B. subtilis
E. coli
b
Genome-wide
frequency
1117
498
79
786
661
216
80
523
641
194
71
582
1:1
1:n
m:1
m:n
orthology type
c
RN-wide orthologs
8
7
13
TFs 1:1 with at least one TG 1:1
d
source network
479
2576
4653
24
17
70
regulogs
e
cg0484
cg1098
dnaA
cg1327
cg2899
f
g
h