Reconstruction of a Global Gene Regulatory Network for Streptomyces Coelicolor: Curation, Inference, and Assessment

Andrea Zorro-Aranda
Department of Chemical Engineering, Universidad de Antioquia

Juan Miguel Escorcia-Rodríguez
Regulatory Systems Biology Research Group, Laboratory of Systems and Synthetic Biology, Center for Genomics Sciences, Universidad Nacional Autónoma de México

José Kenyi González-Kise
Regulatory Systems Biology Research Group, Laboratory of Systems and Synthetic Biology, Center for Genomics Sciences, Universidad Nacional Autónoma de México

Julio Augusto Freyre-González (jfreyre@ccg.unam.mx)
Regulatory Systems Biology Research Group, Laboratory of Systems and Synthetic Biology, Center for Genomics Sciences, Universidad Nacional Autónoma de México

Research Article

Keywords: Streptomyces coelicolor, Gene Regulatory Networks, Network Inference, Biological Networks, Natural Decomposition Approach, Abasy Atlas

Posted Date: September 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-869179/v1

Reconstruction of a global gene regulatory network for *Streptomyces coelicolor*: Curation, inference, and assessment

Andrea Zorro-Aranda[1,3], Juan Miguel Escorcia-Rodríguez[1], José Kenyi González-Kise[1,2], and Julio Augusto Freyre-González[1,*]

[1]Regulatory Systems Biology Research Group, Laboratory of Systems and Synthetic Biology, and [2]Undergraduate Program in Genomic Sciences, Center for Genomics Sciences, Universidad Nacional Autónoma de México. Av. Universidad s/n, Col. Chamilpa, 62210. Cuernavaca, Morelos, México. [3]Bioprocess Research Group, Department of Chemical Engineering, Universidad de Antioquia, Calle 70 No. 52-21, Medellín, Colombia.

***Corresponding author:** jfreyre@ccg.unam.mx (JAF-G)

[§]Present address:

[¶]These authors contributed equally to this work

Abstract

Streptomyces coelicolor A3(2) is a model microorganism for the study of Streptomycetes, antibiotic production, and secondary metabolism in general. However, little effort to globally study its transcription has been made even though *S. coelicolor* has an outstanding variety of regulators among bacteria. We manually curated 29 years of literature and databases to assemble a meta-curated experimentally-validated gene regulatory network (GRN) with 5386 genes and 9707 regulatory interactions (~41% of the total expected interactions). We performed network inference using several methods (including two proposed here) for motif detection in DNA sequences and GRN inference from transcriptomics to tackle network incompleteness, using a community integration approach to reduce false positives. Further, we contrasted the structural properties and functional architecture of the networks to assess the predictions' reliability, finding the inference from DNA sequence data to be the most trustworthy. The inferences allowed us to identify putative novel transcription factors for key *Streptomyces* antibiotic regulatory proteins (SARPs). Finally, we studied the conservation of the system-level components between *S. coelicolor* and *Corynebacterium glutamicum* and identified the basal machinery as the common signature between the two organisms. The curated networks were deposited in Abasy Atlas (https://abasy.ccg.unam.mx/) while the inferences are available as supplementary material.

Keywords: *Streptomyces coelicolor*, Gene Regulatory Networks, Network Inference, Biological Networks, Natural Decomposition Approach, Abasy Atlas

Background

Streptomycetes, the largest genus within the actinomycetes, are biotechnologically relevant organisms, which are producers of around half of natural antibiotics in current use[1]. However, according to the analysis of genome mining, less than 10% of antibiotics that could be produced by actinomycetes are currently used[2]. Their production could be enhanced not only by experimental technologies such as genetic manipulation but also by a deeper knowledge of their secondary metabolism and transcriptional regulation. *Streptomyces coelicolor* A3(2) has become the model microorganism for the study of antibiotic production and secondary metabolism in general[3]. Before its sequencing, it was already known that *S. coelicolor* produces the red-pigmented antibiotic Undecylprodigiosi (RED), the blue-pigmented actinorhodin (ACT), and the calcium-dependent antibiotic (CDA). However, sequencing revealed more than 20 biosynthetic gene clusters (BGCs). Most of the metabolites produced by these clusters and their regulation are still unknown[4].

Secondary metabolism is controlled by a network of regulators at many levels, from global to cluster situated regulators (CSRs). Most CSRs control their own BGC, however, some of them can bind to multiple BGCs causing a cross-cluster regulation[4]. Sequencing of *S. coelicolor* A3(2) revealed 7825 genes, 965 of them (~12.33%) code for proteins with a predicted regulatory function. From those, 65 genes coding for sigma factors, a remarkably high number among bacteria, of which ~70% (45/65) are ECF (extra-cytoplasmic function) sigma factors, suggesting independent regulation of diverse stress response regulons[5]. Besides, it counts with many two-component systems (TCSs), 85

sensor kinases, and 79 response regulators, also related to stress response[5]. The difference between sensor kinases and response regulators suggests a cross-talking among them. Noteworthy, *S. coelicolor* genome codes for several putative regulators that do not belong to families outside *S. coelicolor*[5].

An appropriate understanding of the *S. coelicolor* regulation requires it to be studied systematically at both local and global scale. At global scale, GRN can be represented as a directed graph where nodes represent genes, and edges represent the regulatory interactions among the transcription factors (TFs) and their target genes (TGs). Previous comprehensive reviews have been focused on specific morphological differentiation and metabolic processes[4,6-10]. Therefore, a curation of the GRN at a global scale still is missing. Even though network curation could be approached with text mining[11], it would still require manual intervention for those articles where interactions are not clearly defined. Alternatively, GRN inference has been applied in diverse bacteria to provide a deeper understanding of their regulatory mechanisms, propose selective experimental validation of putative interactions, analyze bacterial GRN evolution, and build biological models for biotechnological processes[12-16]. A GRN inference for *S. coelicolor* was performed by Castro-Melchor, et. al in 2010 using ARACNE and applying module validation through the identification of consensus DNA sequences[17]. However, the resulting network was not assessed with any gold standard (GS) available at the time, and no thorough study of its structural properties was performed. Moreover, benchmarking studies of network inference methods have shown the poor predictive power of using a single GRN inference tool[18].

Here, we performed a collection and curation of the experimentally-validated transcriptional regulatory interactions of *S. coelicolor* A3(2) and classified them based on the confidence level of their supporting evidence. Further, we integrated this curated GRN with previous GRNs from DBSCR (http://dbscr.hgc.jp/) and Abasy Atlas[19]. Then, we applied the natural decomposition approach (NDA) to identify their system-level components and unveiled different biology aspects of *S. coelicolor* regulation. Next, we applied several tools to infer novel interactions, three based on DNA binding sites for the TFs, and five based on gene expression along with two modifications proposed by the authors. We integrated the predictions using a community approach, which has been reported as the best strategy to reduce the number of false positives[18]. Then we used the most reliable curated network as a GS for the validation of the inferred GRNs. We further assessed the inferred networks through their structural properties and found that the NDA [20] is a valuable tool for GRNs dissection and comparison. From the best-rated inferred network, we proposed new TF candidates for the direct regulation of some of the key S*treptomyces* antibiotic regulatory proteins (SARPs) in *S. coelicolor.* Finally, we applied the meta-curated network of *S. coelicolor* to study the conservation of the system-level components with those of its phylogenetically related *C. glutamicum* as an application of the curated network.

Results and discussion

Reconstruction of the most complete experimentally-validated regulatory network for *S. coelicolor*

We curated a total of 124 papers retrieved from PubMed and Google Scholar queries covering a span of 29 years (from 1990 to July of 2019) (see Figure 1 and Supplementary Table 1). We collected a total of 9714 regulatory interactions (out of the 23908 expected interactions in the complete GRN as predicted by Abasy Atlas v2.4) among 5331 genes. We perceive a notable increment in the number of interactions and papers after the *S. coelicolor* genome was completely sequenced (2012)[5]. This facilitates the study of its genome and regulation, being 2012 the year with most publications (see Figure 1). We classified the interactions according to their experimental evidence, following the RegulonDB scheme[21,22], labeling the interactions as "strong" or as "weak" according to the methodology of the experiment performed. "Strong" evidence is assigned to experiments that prove a physical regulatory interaction among the TF and the TG, and "weak" when there is no evidence of direct interaction. For experiments that were not in the RegulonDB scheme, such as DNA-affinity capture assay (DACA)[23], we analyzed their methodology to classify them either as "strong" or "weak" evidence (Supplementary Figure 1 and Supplementary Table 2b).

Afterward, we gathered these interactions along others from the databases, RegTransBase[24] available at Abasy Atlas database (https://abasy.ccg.unam.mx)[19], and DBSCR (http://dbscr.hgc.jp/). We processed these curated interactions to construct the corresponding GRNs, removing redundancy by mapping the gene identifiers to locus tags and merging interactions while preserving the information about the effect and confidence level of the supporting experiments, as previously reported[19]. From our curation, we reconstructed a total of seven curated networks with different levels of confidence and completeness. 1) *Curated_FL* with a total of 9454 unique interactions, from which ~5% (493/9454) are "strong". 2) *Curated_FL(cS)* with the 493 "strong" interactions from *Curated_FL.* 3) *Curated_DBSCR* with the 341 interactions from DBSCR and used the ~34% (115/341) "strong" interaction to reconstruct 4) *Curated_DBSCR(S)*. 5) *Curated_RTB* is the network from the RegTransBase database with 330 interactions, all of them labeled as "weak" since their experimental evidence was not available. Later, we merged *Curated_FL, Curated_DBSCR,* and *Curated_RTB* into 6) *Curated_FL-DBSCR-RTB* a meta-curated network comprising a total of 5386 genes and 9707 non-redundant regulatory interactions, which is the most extensive experimental GRN of *S. coelicolor* up to date. From this meta-curation, we filtered the "strong" interactions to reconstruct 7) *Curated_FL(cS)- DBSCR(S)*. All curated networks are further described in Table 1.

The functional architecture of the *S. coelicolor* GRN

We applied the Natural Decomposition Approach (NDA) on all the curated networks to reveal the functional architecture and to elucidate the regulatory and biological function of some of the genes and interactions curated. The NDA is a biological-mathematical criterion to identify the structure of the GRN[20], classifying every node from the GRN into one of the four structural classes: i) global regulators (GR), coordinating genes from different metabolic pathways[25]; ii) modular genes, group of genes working together to carry out a biological function[20,26]; iii) intermodular genes, integrating at the promoter level the response from different modules[20,27]; and iv) genes constituting the basal

machinery of the cell. We decided to further study the NDA analysis *Curated_RTB-FL-DBSCR* since it is the most complete GRN.

The NDA analysis of the meta-curated network *Curated_RTB-FL-DBSCR* revealed 20 GRs (0.37% of the 5386 network genes), 502 modular genes (9.32%), 18 intermodular genes (0.33%), and 4846 basal machinery genes (89.97%). There is a large module (Module 16) that at the same time is divided into 12 submodules, where the largest submodule (Module 16.1) has 313 genes, conforming to a total of 46 modules and submodules (Supplementary Figure 2a). To analyze the GRs identified in the meta-curated network, we reviewed the literature to identify the TFs that have been previously reported as global or pleiotropic regulators in *S. coelicolor*. *Martín et al.*[28] reported a detailed description of the cross-talking between the global regulators in *S. coelicolor* and other *Streptomyces*. The review provides a list of genes considered as global and wide-domain regulators, due to the hundreds of genes they regulate and the multiple effects they produce[28]. Nine out of the 20 (45%) GRs identified by the NDA were reported as such in this review. We further screened the literature to identify GRs or pleiotropic regulators reported in individual papers (Supplementary Table 3). We found 20 pleiotropic TFs or GRs reported individually, from which 13 (65%) were categorized as GRs by the NDA. See the "Global regulators" section in Supplementary File 1 for further description of the GRs identified.

This analysis also revealed 18 intermodular genes. Some of their promoters integrate the signals of different GR related to carbon, nitrogen, and phosphate metabolism. This as in the case of *glnA* (SCO2198), *glnII* (SCO2210) and the *amtB-glnK-glnD* (SCO5583-85) operon, which are known to be mediators between the nitrogen and phosphate metabolism trough the binding of their GR to these intermodular genes promoters [29]. Others integrate signals from primary and secondary metabolism, or morphological differentiation and antibiotic production. A further description of these genes can be found in Supplementary File 1 in the section "Intermodular genes". Moreover, the functional annotation of the modules identified by the NDA also provides a new functional hypothesis for genes whose function is currently unknown using a guilt-by-association strategy (as previously described[30]). From the 46 modules and submodules in the GRN, 26% (12/46) are annotated (Supplementary Table 4a). Most of the modules annotated are related to cellular metabolism, organic substances metabolism, and biosynthetic processes, which are fundamental processes for every cell (Supplementary Figure 2b). From the annotation of these modules, we were able to suggest novel functional characterizations for 79 genes that were not previously annotated in GOA[31] (Supplementary Table 4). Gene classification provided by the NDA can be downloaded from the Abasy Atlas website (https://abasy.ccg.unam.mx/) and further described in the supplementary material (Supplementary File 1).

GRN inference based on binding sites identification performs better than that based on transcriptomics

Despite the exhaustive curation, the meta-curated network *Curated_FL-DBSCR-RTB* has regulatory information for only ~65% (5386/7825) of the genome *S. coelicolor* (network genomic coverage) and has ~41% (9707/23908) of its expected total interactions (network interaction coverage or completeness) according to the Abasy Atlas classification[19], considering both "Strong" and "Weak" interactions. We leveraged the large corpus of high-throughput data available to computationally

infer missing regulatory interactions to expand our GRN reconstruction. The inference was performed from two different approaches. For the first approach, we performed a regulon reconstruction, through the *de novo* prediction of TF binding sites and linked them to downstream genes. The regulon reconstruction was based on the network *Curated_FL(cS)-DBSCR(S)* using three methods for motif discovery: MEME, Bioprospector, and MDScan (see Methods). For the second approach, we performed a GRN inference from transcriptomic data, from different sources. First, we used the microarray data consisting of 371 samples available from the COLOMBOS Database[32]. Second, we obtain microarray and RNA-Seq data from NCBI GEO[33]. As there is not a consensus method for microarray data integration from different platforms, we decide to take the platform with the largest dataset (137 samples), which was an Affymetrix platform. This data was taken raw, and Robust Multi-chip Averaging (RMA) normalization was performed[34] and the data was batch-effect corrected[35]. For the case of RNA-Seq, the data is available with different types of normalization for each series. Therefore, we decided to take the largest dataset as well (54 samples). We used seven methods for GRN inference based on the gene expression data: CLR, Friedman, GENIE3, Inferelator, MRNET, Statmodel, and TIGRESS (see Methods and Supplementary Table 5a). These methods were selected based on their performance in a previous benchmarking of GRN inference methods[18], their availability, complete documentation, and maintenance. Since most of the experiments in the curation were performed on the *S. coelicolor* A3(2) strain M145, which is plasmid-free, we restricted the inference to interactions among genes of the chromosome.

The inference form expression data was performed over the 5 datasets available: i) COLOMBOS (colombos), ii) RNA-Seq (rnaseq), iii) Affymetrix raw (raw), iv) Affymetrix with RMA normalization (rma), and v) Affimetrix with batch-effect correction (rmabatch). To provide insights on the quality of the predictions, the inferred GRNs were assessed using the network *Curated_FL(cS)-DBSCR(S)* as GS, and the AUPR were computed for each one of the networks. We assessed the inferred GRN based mostly on the AUPR since it is more informative for imbalanced datasets[36] as it is the case of GRNs inference[37] (Supplementary Figure 3a,b). From the evaluation, despite a large amount of data, the prediction with the dataset from COLOMBOS performs poorly with respect to the other data sets. From the Affymetrix data, the rmabatch set performed the best with a very similar result to the rnaseq dataset. However, the latter comes from one unique study, while the Affymetrix data comes from different studies with more diverse data. This, since we believe a more diverse data will allow us to identify a higher quantity of regulatory interactions. Then, we selected the Affymetrix rmabatch as our dataset for the final inference from transcriptomic data.

Next, we performed the same evaluation for the inferred GRNs form binding sites along with the AUROC (see Figure 2a and Supplementary Figure 4). From the AUPR, it is evident that in general GRN inference from binding sites performed better than the inference from expression data. For the inference from binding sites, MEME performed better than the other methods. For the inference from gene expression data, TIGRESS performed better, followed closely by GENIE3 and Inferelator. For the assessment, all the inferred networks were pruned to the 23908 best scoring interactions among genes part of the GS, since it is the number of interactions expected in the final network for *S. coelicolor*[38]. Nevertheless, the GS has only 387 interactions, which is ~1.6% of the 23908 regulatory interactions expected in the complete regulatory network of *S. coelicolor*[19]. For this reason, the assessment only reflects the capacity of the methods to infer the interactions in the GS, while novel interactions (actual interactions not part of the incomplete GS) are labeled as false

positives. Moreover, as the GS was used as prior for the regulon extension, it might provide an advantage for the network predicted by motif discovery. Because of its approach, inference by motif discovery predicts direct regulatory interactions, while inference from transcriptomic data predicts both direct and indirect ones without distinction. Thus, as the GS is only built by direct interactions, it is expected that networks predicted by motif discovery predict more of this type of interactions than *the prediction from transcriptomic data*. However, using the "non-strong" GRN as GS could drive to a bigger problem because indirect regulatory interactions are not adequate to assess causal interactions.

Given that the current GS is still quite incomplete, and we cannot do proper discrimination among the different inferred networks, instead only using the single best method, we decided to build a community network for each one of the approaches. i) *Inferred_BS* for the prediction from binding sites; ii) *Inferred_Exp* for the prediction from expression data; iii) *Inferred_BS-Exp*, a community network from both previous community networks; and iv) *Inferred_All*, a community network built mixing both approaches (see Table 1). For the latter, we used the three methods for binding site inference and Statmodel, GENIE3, and TIGRESS from expression-based GRNs (due to their superior performance) to balance both approaches. *Inferred_BS* outperformed the rest of the community networks at both AUPR and AUROC (see Figure 2a). However, it was outperformed by MEME at both metrics. Given MEME's outstanding performance (see Figure 2a), we used it to perform a statistical validation of "weak" interactions supported by ChIP-data, similarly as proposed in[22] (see Methods). A total of 55 "weak" interactions were reclassified as "strong" (see Figure 2b and Supplementary Table 5b). We found one of these interactions (SCO4230-SCO4878) already reported as "strong" in the DBSCR database (*Curated_DBSCR(S)*). These statistically-validated interactions were merged with the "strong" interactions from *Curated_FL* and from the meta-curated network *Curated_FL-DBSCR-RTB* into two networks: *Curated_FL(S)* and *Curated_FL(S)-DBSCR(S).* We reassessed the network predictions with *Curated_FL(S)-DBSCR(S)* and the results remained virtually the same (Supplementary Figure 5-6).

Inferred networks have a similar structure to the largest curated networks

Even though the AUPR and AUROC metrics allow the assessment of the predictions, both metrics heavily rely on the ranking of the predicted interactions. Moreover, the GS is not complete and missing interactions would be still classified as false positives, decreasing the score more as higher its ranking is. We also assessed the inferences in terms of their structural properties and compared them against the curated networks to compensate for such drawbacks. Note that this approach has its own caveats. The global structural properties of the network might be different once the GS is complete, this can be approached by comparing the predictions to all the curated networks, each of them with different completeness. Also, two networks could have the same topology with different node entities. For this reason, we use the topological assessment in complement with the AUPR and AUROC metrics to identify the best prediction.

One of the main characteristics of biological networks is that they are scale-free and hierarchically modular. Same characteristics that our curated networks have been proved to possess (Supplementary File 1). Therefore, as an initial approach, we asked whether the inferred networks are scale-free too. Scale-free networks are networks whose degree (nodes' connectivity) distribution $\boldsymbol{P(k)}$ follows a power law, $\boldsymbol{P(k) \sim k^{-\alpha}}$, with $\boldsymbol{2 < \alpha < 3}$ [20,38,39]. This since with an $\boldsymbol{\alpha =}$

$\mathbf{2}$ appears a unique global regulator (hub-and-spoke network) and for $\boldsymbol{\alpha} > \mathbf{3}$ scale-free networks lose most of their characteristic properties[39]. First, to compute this α for the inferred networks, we performed a robust linear regression over a log-log plot of the complementary cumulative degree distribution and corrected the exponent accordingly (see Table 2 and Supplementary Figures 7-11). All inferred networks' degree distribution seems to follow a power law according to the adjusted coefficient of determination. Nevertheless, the data points in Inferred_All appear to be divided into three regions with different tendencies, instead of the two that are present in the other networks (Supplementary Figures 7-11). Usually, this type of network is divided into two regions, the region of the nodes (genes) with a low degree, and the one with nodes with a high degree[40]. The appearance of a third region might be a consequence of merging networks from methods with different approaches. This could affect the structural properties of the merged networks, while communities from the same approach appear to have more similar structural properties. In the case of Inferred_BS-Exp, the construction of communities ahead by each approach create more compatible networks in terms of structure that can be conveniently mixed.

Next, we wanted to confirm that their degree distribution followed a power law, so we contrasted each potential power law of the inferred networks the distributions against alternative fat-tailed probability distributions (Power-law, exponential, stretched exponential, lognormal, and truncated power-law) using Kolmogorov–Smirnov tests[38,41] to confirm that the degree distributions follow a power law (Supplementary Table 6a). We found the degree distribution of the inferred networks adjusted better to a power-law distribution than to an alternative distribution. Then, we computed a maximum likelihood estimation for the exponent (α) of the power law and found that most of them are between two and three, except for *Inferred_All* showing it as an anomalous scale-free network (Supplementary Table 6b). Perhaps caused by the mixing of networks with diverse structural properties. Nevertheless, we could consider all inferred networks to be scale-free.

However, we wanted to check the other properties of scale-free networks (see Table 2)[39,42]. The four community networks have small average shortest path lengths and a high clustering coefficient. *Inferred_BS* has the smallest average short path length, while *Inferred_All* has the highest average clustering coefficient (see Table 2). Scale-free networks also present an ultra-small world effect, which implies that the average path length is proportional to $\ln(\ln(N))$ (N is the number of nodes in the network). This is the case for all the inferred networks. Another characteristic of GRNs is their hierarchical modularity, which implies a diamond-shaped hierarchical organization as has been revealed by the NDA[20]. In a scale-free network, this implies that the clustering coefficient depending on the degree (C) follows a power law as $C(k) \sim k^{-1}$[39]. *Inferred_BS* has the exponent closest to -1 (0.92), with the best R^2. Even though *Inferred_BS* seems to be the network that behaves closest to a GRN, all networks have similar values, which makes it difficult to discern the most reliable inferred network.

To perform a more thorough comparison of their structural graph properties, we include several other (Supplementary Table 7). We clustered the vectors of structural properties for the curated and community-inferred networks (see Figure 3a). The clustering partitions the networks into two major groups. The first one containing the curated networks and *Inferred_BS*, while the second group contains the other inferred networks. The first group is in turn also divided into two groups: one with the two largest curated networks and *Inferred_BS,* and the other one with the remaining curated networks. The reason for this may be due to the size of the networks (see Table 1). When

standardizing the property values by max-min feature scaling, the two largest curated networks were clustered with the predictions (Supplementary Figure 12), which could be also due to the size of the network. To reduce the network size influence we used the network dissimilarity measure proposed by *Schieber et al*.[43]. We considered the third term which makes the distance measure robust to graph size in terms of the number of nodes (genes)[43] (see Figure 3b). Even with this metric, the two largest curated networks were clustered with the inferred networks. This might be a consequence of the high fraction of maximum out-connectivity and structural genes in the largest curated networks, like those found in the inferred networks. This shows that the inferred networks are similar that the most complete curated networks in terms of structure, suggesting their reliability.

The Natural Decomposition Approach identifies regulatory networks similarity despite their different completeness level

We compared all curated and community inferred networks based on the Simpson similarity of the four components proposed by the NDA: global regulators, modular genes, intermodular genes, and the basal machinery (see Figure 4). The Simpson similarity index measures the size of the core of two sets with reference to the smallest one[44]. Note that two identical sets, being one a subset of the other one will have a score of 1. On the other hand, two completely different sets will have a score of 0.

When comparing the global regulators (see Figure 4a), there is not a distinct division among the networks. *Infered_BS-Exp* and *Infered_Exp* have a similar correlation with all the curated networks, slightly higher with *Curated_DBSCR(S)*, *Curated_FL*, and *Curated_FL-DBSCR-RTB*. These two inferred networks have the highest amount of GR, 116, and 114 respectively; thus, the other GRs predicted could be easily a subset of them. For the case of *Infered_BS*, it has the highest correlation with the "strong" networks. This is expected since these networks have only direct regulatory interactions, as the interactions predicted in *Inferred_BS*, while there is no evidence of direct regulation for transcriptomic-based inferred interactions. This can affect the measurement of the effect of GRs over the rest of the genes since GRs regulating not so many targets from different processes might not be predicted as such in the "strong" networks. However, when their indirect influence is represented in the network, their ranking as GRs is noticeable.

When analyzing the modular genes, there are two major groups (see Figure 4b): the major group, with the curated networks, is divided into two subgroups, one with *Curated_RTB*, *Curated_DBSCR,* and its "strong" version *Curated_DBSCR(S),* and the second subgroup contains the integration proposed in this work. Interestingly, the meta-curated network *Curated_FL-DBSCR-RTB* correlates very well with all the smaller networks it contains, from which we could deduce that modular genes are conserved despite the addition of new regulatory interactions. In the second group, composed of the inferred networks, we can see there is not a high correlation among them. *Inferred_BS* is the closest to the curated networks, while Inferred_Exp and Inferred_BS-Exp have a high correlation. This tells us that the interactions in *Inferred_Exp* have a larger influence on the module configuration of *Inferred_BS-Exp* than *Inferred_BS*. The difference between the curated and inferred networks might come from the fact that inferred networks have a greater number of GRs and a much lower number of modular genes when compared with the curated networks (Supplementary Table 8).

Intermodular genes are the less conserved NDA class (see Figure 4c). There is overlap only among the smallest curated networks, all share the intermodular de gene SCO5877, which appears as a TF in the other curated networks. Moreover, an overlap among *Curated_FL* and *Curated_FL-DBSCR-RTB*, which share most of the interactions. Thus, is expected that they also share most of the intermodular genes. Note that the networks *Curated_DBSCR(S)* and *Inferred_BS-Exp* are not included in the clustering since they did not present any intermodular genes. Finally, when analyzing the basal machinery (see Figure 4d), the larger curated networks are grouped on one side, next to the inferred networks, and finally the smallest curated networks with *Curated_RTB* as an outgroup. Even though *Inferred_BS* is grouped with the other inferred networks, it has a higher correlation with *Curated_FL(S)* and *Curated_FL(S)-DBSCR(S)*, which again evidence the similarity among these three networks.

Assessment of the global regulators' inference

In the NDA, the identification of global regulators is a key step in the classification of every node in the GRN. Previously, it has been reported a high overlap between the predictions of global regulators by the NDA and those reported in the literature for *E. coli*[20], *B. subtilis* Because of the NDA algorithm, the identification of global regulators is a key step in the classification of every node in the GRN. Previously, it has been reported a high overlap between the predictions of global regulators by the NDA and those reported in the literature for *E. coli*[20], *B. subtilis*[45], and *C. glutamicum*[27]. We used the set of GRs reported by *Martín et. Al.*[28], besides those reported in independent articles, and the union of both sets (Supplementary Table 3). Then we assessed the predictions of the GR using the Matthews correlation coefficient (MCC) (see Figure 5) We used the MCC score as it is more informative and reliable than F1 for binary classification evaluation[46] but using the F1 score we obtained consistent results (Supplementary Figure 13).

Curated_FL and *Curated_RTB-FL-DBSCR* have the best performance in GR prediction. However, the "strong" networks have a slightly smaller score even having much less genomic coverage. This shows that the GRs are very robust to perturbations in the network as previously shown[20]. On the other hand, despite the high coverage of the inferred networks, the performance of the predictions with such networks was poor. This could be, as it was mentioned before, due to the great amount of GR predicted by these networks, which would cause a high proportion of false positives, affecting the score. *Inferred_BS* produced the most conservative prediction (lowest false-positives rate) among the inferred network (Supplementary Figure 14 and Supplementary Table 8).

GRN inference from transcription factor binding sites proves to be the most reliable approach and allows the prediction of new TFs for the most studied SARPs

Even though other similar studies suggest the integration of inference approaches as the most suitable methodology for GRN reconstruction[18,47,48], because of the analysis performed in this paper, we consider *Inferred_BS* as the most reliable inferred network. From the evaluation against the GS, where it presented the highest AUPR and AUROC among the community networks, through its structural properties along with the NDA analysis where it showed the most similar configuration to a GRN reconstruct from biological experiments. Moreover, it has the largest genomic coverage among all the networks, which would be advantageous for a deeper study of transcriptional regulation in *S. coelicolor*. Therefore, we decided to use *Inferred_BS* to further study the regulation of the SARPs of the most studied antibiotics in *S. coelicolor*: ActII–orf4, RedD/RedZ, CpkO (also

known as KasO), and CdaR, which regulate the production of ACT, RED, yCPK, and CAD, respectively[3] (Supplementary Table 9). A total of 13 new interactions for the SARPs were predicted, providing us a great opportunity to find new targets to manipulate the *S. coelicolor* antibiotic production.

Next, we describe some of the TFs predicted for the SARPs: For *actII-orf4* (SCO5085) only one regulator was inferred, MacR (SCO2120) which is the response regulator of the TCS MacRS. This TCS has been proved to activate ACT production. Nevertheless, a ChIP-qPCR analysis was not able to prove an *in vivo* interaction between MacR and *actII-orf4*, although a direct binding was not tested as *Inferred_BS* predicted[49]. For *redD* (SCO5877) two new regulators were predicted, LipR (SCO0712) and ActII-orf4 (SCO5085). LipR is related to AfsR (SCO4426)[50], homolog to the SARPs, and activator of the ACT and RED production[51]. Moreover, its mutant affects ACT production[50], which makes it plausible to affect RED production as well. It has been suggested that ActII-orf4 might regulate the production of other antibiotics[4], which could be by binding directly to their CSR. For *redZ* (SCO5881) five new regulators were predicted, among them is GluR (SCO5778) which has been shown to affect RED production. Nevertheless, it has been shown that GluR does not bind directly to *redZ*, thus it could be more an indirect regulation[52]. Another one, StgR (SCO2964) has been shown, by an RT-qPCR experiment, to be a repressor of redD[53], thus this repression could be through the direct binding to *redZ*. HpdA (SCO2928) and HpdR (SCO2935) are related to Tyrosine catabolism, which produces important precursors for antibiotic biosynthesis[54]. Moreover, HpdA has been shown to activate *actII-ORF4,* therefore might have a more direct role in RED production. In the case of *cdaR* (SCO3217), we have four predicted regulators, among them, OsdR (SCO0204) and RamR (SCO6685). Both are related to the response to stress and the development of *S. coelicolor*[55,56]. SsgR (SCO3925) regulates the sporulation and morphological differentiation[57]. These all processes are highly related to antibiotic production. Finally, for *cpkO/kasO* (SCO6280) six new regulators were inferred, among them OsdR (SCO0204), LipR (SCO0712), and StgR (SCO2964) were described before. Another one is NnaR (SCO2958), which regulates spore formation and antibiotic production[58]. We refer the reader to the supplementary material for the complete list of the predicted regulators in every predicted network (Supplementary Table 9).

Comparative analysis with *Corynebacterium glutamicum* shows coherent system-level components conservation

The diamond-shaped structure identified by the NDA is conserved between *E. coli* and *B. subtilis*[45]. As an application of the meta-curated network, we studied the conservation of its system-level components, comparing it against the *C. glutamicum* network, which is phylogenetically related to *S. coelicolor,* and a model organism for the study of GRNs[27]. We applied the regulogs analysis[59] with one-to-one orthology relationships to alleviate network incompleteness and make them comparable. As prior networks, we used *Curated_FL(S)-DBSCR(S)* (534 interactions) for *S. coelicolor* and 196627_v2020_s21_eStrong from Abasy Atlas[19] (2941 interactions) for *C. glutamicum*[16]. After the regulogs analysis, we ended up with 2966 interactions in *C. glutamicum* and 692 interactions in *S. coelicolor*.

We used the complemented networks to identify GRN-wide orthologous relationships defined as the orthologous present in the GRN of the respective organism. We obtained a total of 188 GRN-wide orthologous relationships from a total of 995 1:1 orthologs identified by OrthoFinder[60]. We applied the NDA analysis to both GRNs to identify the system-level components and computed the

fraction of the GRN-wide orthologous in each combinatory relationship between the components (see Figure 6a). We found that most of the GRN-wide orthologous (54%) are classified as basal machinery in both organisms. This is expected since 73% and 74% of the genes correspond to the basal machinery in the complemented networks of *C. glutamicum* and *S. coelicolor*, respectively. Besides, the distribution of the genes in the chromosome of *S. coelicolor* shows a central core, where are genes likely related to primary functions such as DNA replication, transcription, translation, and amino-acid biosynthesis; and likely non-essential genes such as secondary metabolism are in the chromosome arms[5]. More than 59% (111/188) of the GRN-wide orthologs conserved the same class in both organisms (Figure 6a) showing high conservation of the NDA classification.

We studied the pairwise Simpson similarity index between the four classes between the two organisms to remove the problem of the imbalanced classes in both organisms (see Figure 6b). GR is the class with the highest conservation rate, the orthologs of seven of the eight GRs in *C. glutamicum* are also GRs in *S. coelicolor* (see Figure 6b,c). The conservation between the same class in the two organisms is also high for the basal machinery, while poor for the modular genes. For the case of intermodular genes, even though the networks were complemented with information from the other network, they are not conserved at all (see Figure 6b). Previous work reported intermodular genes as the least conserved of the system-level components[27]. Intermodular genes are the most likely responsible for giving the TRN flexibility and increase evolvability by scouting different combinations of regulatory interactions between physiological functions so the organism could adapt better to environmental changes[45]. These results agree with a previous analysis of the robustness of the NDA to a random node and edge remotion showing GR and intermodular genes as the most and least conserved classes, respectively (see Figure 8 in[27]).

On the other hand, 24% of the GRN-wide orthologs that are modular genes in *C. glutamicum* were classified in *S. coelicolor* as basal machinery. This could be due to three possible reasons[45]: i) the basal machinery genes in *S. coelicolor* are misclassified and further research is needed to find the missing regulatory interactions (see Figure 6c) that will integrate some of the basal machinery genes into a module. ii) The GRs controlling *C. glutamicum* genes are not yet identified as GRs. iii) Genes in *S. coelicolor* need a more direct regulation because of their physiological function (high plasticity of transcriptional regulation). The previous comparison between *S. coelicolor, Mycobacterium tuberculosis,* and *Corynebacterium diphtheriae* showed a synteny among the whole chromosome of these last two microorganisms and the core of the one of *S. coelicolor*[5]. *C. glutamicum* is phylogenetically closely related to *M. tuberculosis* and *C. diphtheriae*, with roughly similar genome size. Therefore, a similar result would be expected. Furthermore, as more classical experiment data become available, new regulations for the currently basal machinery would turn those genes into the modular class. However, a deeper analysis of diverse factors such as genome size, the niche of the organisms, and a wider range of organisms are required to further study the robustness of the NDA analysis.

Conclusions

A meta-curated regulatory network for *S. coelicolor* (*Curated_RTB-FL-DBSCR)* was reconstructed from a collection and curation of regulatory interactions experiments in literature and databases. From the NDA analysis of the meta-curated network, we could identify 20 global regulators, of which 95% (19/20) have already been reported as global or pleiotropic regulators. 18 intermodular genes,

some of them found to be involved in more than one biological process. 46 modules and submodules were identified, allowing to propose the function for 79 genes without previous functional annotation. However, this network is ~40% of the complete regulatory network, which evidences a lack of information related to *S. coelicolor* transcriptional regulation. Especially for interaction experimentally supported by "strong" evidence, only ~2% (387/23908) of the complete network. However, the meta-curated network *Curated_RTB-FL-DBSCR* already portrays accurately the highest hierarchical level of *S. coelicolor* regulation. GRN inference from transcriptomic and DNA sequence data was performed and the inference from TF binding sites identification showed to be the best approach according to interactions inference assessment, topological assessment, and systems-level comparison. From this network 13 new TFs were predicted to bind directly to five of the principal SARPs, most of which previously proved to affect indirectly antibiotic production, but not proof of direct effect, or be related to stress response or morphological differentiation. We compared *S. coelicolor* network to *C. glutamicum* GRN and found high conservation only for the basal machinery, which might be a result of the high plasticity of the transcriptional regulation.

Material and methods

Transcriptional regulatory interactions curation

We performed a comprehensive review of the literature to identify experimentally-supported transcriptional regulatory interactions in *Streptomyces coelicolor* A3(2). We searched peer-reviewed articles in Google Scholar and PubMed using the keywords "*Streptomyces coelicolor*" AND "transcriptional" and its variations AND "regulation" and its variations. In the case where reviews were found, their references were followed to the original research papers. Then, we performed the curation and organized the interactions (Supplementary File 1). Experiments were classified according to their methodology and their names were standardized for the sake of clarity and easier evidence classification. We merged these interactions with two previously curated networks, one was reconstructed from an XML provided by the DBSCR team and the other one from RegTransBase [24] available at the Abasy Atlas website. These datasets are available from http://dbscr.hgc.jp/ and https://abasy.ccg.unam.mx. Abasy Atlas is a database of meta-curated bacterial GRNs for nine species including *S. coelicolor*[19]. It also provides historical snapshots for other model organisms such as *Escherichia coli, Bacillus subtilis, Corynebacterium glutamicum,* and *Mycobacterium tuberculosis*[19].

GRN inference from transcription factor binding sites

To extend the regulons for the TFs identified in the literature, we used the set of "strong" interactions as prior (*Curated_FL(cS)-DBSCR(S)*). We reconstructed a position weight matrix (PWM) for every TF in the "strong" network using the non-overlapping up to -300 to +50 bp (with reference to the translation start codon) upstream regions of their TGs as input for three motif discovery algorithms. Namely, i) MEME, an extension of the expectation-maximization algorithm for fitting finite mixture models[61]; ii) BioProspector, based on multiple Gibbs sampling[62]; and iii) MDscan, that employs a heuristic word-enumeration approach combined with statistical modeling[63]. Then, we used FIMO[64] (p-value threshold = $1x10^{-4}$) to identify TF-TG interactions. As most of the interactions curated are from *S. coelicolor* A3(2) strain M145 (plasmid-free), we excluded genes that are not part of the chromosome.

GRN inference from transcriptomic data

We downloaded first the transcriptomic dataset for *S. coelicolor* available at the COLOMBOS database [32]. Then, we also download data from the NCBI Gene Expression Omnibus (GEO)[33]. From there we download an Affymetrix dataset (Platform GPL9417) and an RNA-Seq dataset (GPL26763. Afterward, we normalize the Affymetrix data using Robust Multi-chip Averaging (RMA) with the affy package[65] and used the gPCA package[66] to identify a batch effect in the data, which was corrected with Combat from the sva package[67], all of them are packages for R. The data counts on 137 transcriptomes for 7738 genes. As in the case of GRN inference from transcription factor binding sites, we only considered genes from the chromosome. We selected the best inference methods according to their outstanding performance in the DREAM challenge[18]. Moreover, we selected methods that have an implementation in R or Matlab and were well documented. The inference methods selected were: i) CLR[68], a method that applies mutual information; ii) GENIE3[69], which applies tree-based regression and feature selection; iii) Inferelator[70], which applies regression and variable selection; iv) MRNET[71], which applies the maximum relevance/minimum redundancy algorithm; and v) TIGRESS[72] which applies LARS combined with stability selection. Along with these methods, we used two modifications we propose in this work, Friedman, and Statmodel (Supplementary File 1). We provided all methods with a list of 137 TFs from the meta-curated network *Curated_FL-DBSCR-RTB* to infer causality[18,72].

Integration of individual inferences into a community GRN

To increase the precision of the predictions we used a community approach[18] integrating individual predictions from different algorithms. First, the individual predictions are sorted by their confidence score, keeping the most reliable ones at the beginning of the prediction list. Then, the average of the rank positions in the predictions is given as the community score for each interaction. For missing interactions in a prediction list, the position is equal to the size of the prediction list + 1. All community networks were pruned to the 23908 first interactions with the highest score, which is the predicted size of the complete GRN of *S. coelicolor* reported by Abasy Atlas v2.4[19] according to the model developed in[38]. The model is constantly being updated on the Abasy Atlas website by the addition of new networks and interactions, which will cause a slight variation in the number of interactions[19].

Assessment of the inferred GRN

We computed the area under the precision-recall curve (AUPR) and the area under the receiver operating characteristics curve (AUROC) to assess our predictions using in-house scripts. The AUPR depicts the precision (1) as a function of the recall (2) obtained by the predictor. The AUROC depicts the relation between the recall, also called the true positive rate (TPR) (2), and the false positive rate (FPR) (3). Note that unknown actual interactions between genes in the GS will still be considered as FP[37]. For this reason, interactions involving genes that are not part of the GS were not considered.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = TPR = \frac{TP}{TP + FN} \quad (2)$$

$$FPR = \frac{FP}{FP + TN} \quad (3)$$

Statistical validation for interactions supported by ChIP experiments.

We performed a statistical validation approach for ChIP-data[22] for those interactions supported by ChIP experiments and no "strong" experiment. Because of the lack of data in some articles, instead of performed the statistical validation finding the motifs from the ChIP data directly as in[22], we used the "strong" interactions as seed to construct the matrix models for the TFs. We used MEME[61] to build a position weight matrix (PWM) for each TF with at least 3 "strong" interactions. Then, we used FIMO with the matrix for each TF to scan the non-overlapping up to -300 to +50 bp upstream regions with reference to the translation start sites of *S. coelicolor*. We kept those TF binding sites with p-value < $1x10^{-4}$ and used these interactions to compare them with those supported by ChIP technologies, keeping the intersection of both sets (interactions supported by both ChIP and motif finding approaches).

Network similarity

We computed the characteristic structural properties for GRNs reported as global properties on the Abasy Atlas database[19]. Namely, regulators ($k_{out} > 0$) (%), direct regulatory interactions, self-regulation (%), maximum out-connectivity (%), network density, weakly connected components, genes in the giant component (%), feedforward circuits, complex feedforward circuits, 3-Feedback loops, average shortest path length, network diameter, average clustering coefficient, adjusted coefficient of determination (R^2_{adj}) of $P(k)$, and R^2_{adj} of $C(k)$. Then we used pairwise Pearson correlation among the profiles of the structural properties of the networks and cluster them according to the Euclidean distance among the correlations using Ward's method. To compute the pairwise network dissimilarity we used a Python implementation adapted from that proposed by *Schieber et al.*[43] with the parameters proposed by the authors (0.45, 0.45, 0.1), being the last one required to discriminate network size in terms of nodes (genes)[43]. Then, we clustered the networks by using Euclidean distances among its dissimilarity values using Ward's variance minimization algorithm as the linkage method.

System-level components

We applied the Natural Decomposition Approach (NDA), a biological-mathematical criterion to identify the components of the diamond-shaped structure of the GRNs[20], on every curated and inferred network. See the supplementary methods (Supplementary File 1) for a brief description of the NDA, and[20] for further details. The assessment of the prediction of GRs was performed using in-house scripts for MCC, precision, and F1-score. Scores were computed for each GRs prediction obtained by the NDA with the different networks analyzed in this work. As GS we used the GRs previously reported in the literature from a review[28], from individual publications, and both (Supplementary Table 3).

Comparative analysis against *C. glutamicum*

We used as prior the strong regulatory networks Curated_FL(S)-DBSCR(S) for *S. coelicolor*, and 196627_v2020_s21_eStrong from the Abasy Atlas database for *C. glutamicum*[19], considering only the interactions between two genes both mapping to a locus tag. We used MEME to construct a PWM for every TF with at least three TGs using their upstream sequences. These sequences were defined as the non-overlapping regions of up to -300 to +50 bp with reference to the translation start codon and were obtained with retrieve-seq from RSAT[73]. Then, we used FIMO with the PWM of the TFs from *S. coelicolor* to find individual occurrences with a p-value < $1x10^{-4}$ in the upstream sequences of *C. glutamicum*. The same was done in the opposite direction. With this, we seek to alleviate network incompleteness by extrapolating known interactions from an organism to the other[59]. Predicted interactions were sorted by p-value and only the best scoring result was conserved for redundant interactions. Afterward, we used Orthofinder to find one-to-one ortholog relationships between both organisms. We used OrthoFinder due to its high accuracy[60]. The orthologs were used to further filter FIMO predictions to conserve interactions in which both TF and TG have a one-to-one orthologous relationship in the other organism. We considered the original "strong" network interactions at the beginning of the interactions list. The NDA was applied to both expanded GRNs to identify ortholog systems and only the genes with one-to-one orthologs in the other organism's network (GRN-wide orthologs) were considered in the analysis.

Availability of supporting data

The data set(s) supporting the results of this article are included within the article, its additional files, or in Abasy Atlas at https://abasy.ccg.unam.mx/.

References

1 Hoskisson, P. A. & van Wezel, G. P. Streptomyces coelicolor. *Trends Microbiol* **27**, 468-469, doi:10.1016/j.tim.2018.12.008 (2019).

2 Mast, Y. & Stegmann, E. Actinomycetes: The Antibiotics Producers. *Antibiotics (Basel)* **8**, doi:10.3390/antibiotics8030105 (2019).

3 Chen, S. *et al.* Roles of two-component system AfsQ1/Q2 in regulating biosynthesis of the yellow-pigmented coelimycin P2 in Streptomyces coelicolor. *FEMS Microbiol Lett* **363**, doi:10.1093/femsle/fnw160 (2016).

4 McLean, T. C., Wilkinson, B., Hutchings, M. I. & Devine, R. Dissolution of the Disparate: Co-ordinate Regulation in Antibiotic Biosynthesis. *Antibiotics (Basel)* **8**, doi:10.3390/antibiotics8020083 (2019).

5 Bentley, S. D. *et al.* Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141-147, doi:10.1038/417141a (2002).

6 Bednarz, B., Kotowska, M. & Pawlik, K. J. Multi-level regulation of coelimycin synthesis in Streptomyces coelicolor A3(2). *Appl Microbiol Biotechnol* **103**, 6423-6434, doi:10.1007/s00253-019-09975-w (2019).

7 Bibb, M. 1995 Colworth Prize Lecture. The regulation of antibiotic production in Streptomyces coelicolor A3(2). *Microbiology* **142 (Pt 6)**, 1335-1344, doi:10.1099/13500872-142-6-1335 (1996).

8 Chater, K. F. Regulation of sporulation in Streptomyces coelicolor A3(2): a checkpoint multiplex? *Curr Opin Microbiol* **4**, 667-673, doi:10.1016/s1369-5274(01)00267-3 (2001).

9 Bibb, M. J. Regulation of secondary metabolism in streptomycetes. *Curr Opin Microbiol* **8**, 208-215, doi:10.1016/j.mib.2005.02.016 (2005).
10 Martin, J. F. & Liras, P. Cascades and networks of regulatory genes that control antibiotic biosynthesis. *Subcell Biochem* **64**, 115-138, doi:10.1007/978-94-007-5055-5_6 (2012).
11 Zitnik, S., Zitnik, M., Zupan, B. & Bajec, M. Sieve-based relation extraction of gene regulatory networks from biological literature. *BMC Bioinformatics* **16 Suppl 16**, S1, doi:10.1186/1471-2105-16-S16-S1 (2015).
12 Novichkov, P. S. *et al.* RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics* **14**, 745, doi:10.1186/1471-2164-14-745 (2013).
13 Leyn, S. A. *et al.* Comparative genomics and evolution of transcriptional regulons in Proteobacteria. *Microb Genom* **2**, e000061, doi:10.1099/mgen.0.000061 (2016).
14 Metris, A. *et al.* SalmoNet, an integrated network of ten Salmonella enterica strains reveals common and distinct pathways to host adaptation. *NPJ Syst Biol Appl* **3**, 31, doi:10.1038/s41540-017-0034-z (2017).
15 Staunton, P. M., Miranda-CasoLuengo, A. A., Loftus, B. J. & Gormley, I. C. BINDER: computationally inferring a gene regulatory network for Mycobacterium abscessus. *BMC Bioinformatics* **20**, 466, doi:10.1186/s12859-019-3042-8 (2019).
16 Escorcia-Rodríguez, J. M., Tauch, A. & Freyre-González, J. A. Corynebacterium glutamicum regulation beyond transcription: Organizing principles and reconstruction of an extended regulatory network incorporating regulations mediated by small RNA and protein-protein interactions. *bioRxiv*, 2021.2001.2007.423633, doi:10.1101/2021.01.07.423633 (2021).
17 Castro-Melchor, M., Charaniya, S., Karypis, G., Takano, E. & Hu, W. S. Genome-wide inference of regulatory networks in Streptomyces coelicolor. *BMC Genomics* **11**, 578, doi:10.1186/1471-2164-11-578 (2010).
18 Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat Methods* **9**, 796-804, doi:10.1038/nmeth.2016 (2012).
19 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* **18**, 1228-1237, doi:10.1016/j.csbj.2020.05.015 (2020).
20 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* **9**, R154, doi:10.1186/gb-2008-9-10-r154 (2008).
21 Santos-Zavaleta, A. *et al.* 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, doi:10.1093/nar/gky1077 (2019).
22 Weiss, V. *et al.* Evidence classification of high-throughput protocols and confidence integration in RegulonDB. *Database (Oxford)* **2013**, bas059, doi:10.1093/database/bas059 (2013).
23 Park, S. S. *et al.* Mass spectrometric screening of transcriptional regulators involved in antibiotic biosynthesis in Streptomyces coelicolor A3(2). *J Ind Microbiol Biotechnol* **36**, 1073-1083, doi:10.1007/s10295-009-0591-2 (2009).
24 Cipriano, M. J. *et al.* RegTransBase--a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. *BMC Genomics* **14**, 213, doi:10.1186/1471-2164-14-213 (2013).

25 Gottesman, S. Bacterial regulation: global regulatory networks. *Annu Rev Genet* **18**, 415-441, doi:10.1146/annurev.ge.18.120184.002215 (1984).
26 Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47-52, doi:10.1038/35011540 (1999).
27 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* **257**, 199-210, doi:10.1016/j.jbiotec.2016.10.025 (2017).
28 Martín, J. F., Santos-Beneit, F., Sola-Landa, A. & Liras, P. in *Stress and Environmental Regulation of Gene Expression and Adaptation in Bacteria* (ed Frans J. de Bruijn) 257-267 (John Wiley & Sons, Inc., 2016).
29 Martin, J. F. *et al.* Cross-talk of global nutritional regulators in the control of primary and secondary metabolism in Streptomyces. *Microb Biotechnol* **4**, 165-174, doi:10.1111/j.1751-7915.2010.00235.x (2011).
30 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**, doi:10.1093/database/baw089 (2016).
31 Camon, E. *et al.* The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* **32**, D262-266, doi:10.1093/nar/gkh021 (2004).
32 Moretto, M. *et al.* COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. *Nucleic Acids Res* **44**, D620-623, doi:10.1093/nar/gkv1251 (2016).
33 Clough, E. & Barrett, T. The Gene Expression Omnibus Database. *Methods Mol Biol* **1418**, 93-110, doi:10.1007/978-1-4939-3578-9_5 (2016).
34 Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15, doi:10.1093/nar/gng015 (2003).
35 Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127, doi:10.1093/biostatistics/kxj037 (2007).
36 Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432, doi:10.1371/journal.pone.0118432 (2015).
37 Siegenthaler, C. & Gunawan, R. Assessment of network inference methods: how to cope with an underdetermined problem. *PLoS One* **9**, e90481, doi:10.1371/journal.pone.0090481 (2014).
38 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* **9**, 3618, doi:10.1038/s41598-019-39866-z (2019).
39 Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-113, doi:10.1038/nrg1272 (2004).
40 Barabási, A.-L. & Pósfai, M. *Network Science*. Edición: 1 edn, (Cambridge University Press, 2016).
41 Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-Law Distributions in Empirical Data. *SIAM Review* **51**, 661-703, doi:10.1137/070710111 (2009).
42 Khanin, R. & Wit, E. How scale-free are biological networks. *J Comput Biol* **13**, 810-818, doi:10.1089/cmb.2006.13.810 (2006).
43 Schieber, T. A. *et al.* Quantification of network structural dissimilarities. *Nat Commun* **8**, 13928, doi:10.1038/ncomms13928 (2017).

44 Simpson, G. G. Mammals and the nature of continents. *American Journal of Science* **241**, 1-31, doi:10.2475/ajs.241.1.1 (1943).
45 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* **161**, 278-286, doi:10.1016/j.jbiotec.2012.03.028 (2012).
46 Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6, doi:10.1186/s12864-019-6413-7 (2020).
47 Marbach, D. *et al.* Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. *Genome Res* **22**, 1334-1349, doi:10.1101/gr.127191.111 (2012).
48 Lihu, A. & Holban, S. A review of ensemble methods for de novo motif discovery in ChIP-Seq data. *Brief Bioinform* **16**, 964-973, doi:10.1093/bib/bbv022 (2015).
49 Liu, M. *et al.* Novel Two-Component System MacRS Is a Pleiotropic Regulator That Controls Multiple Morphogenic Membrane Protein Genes in Streptomyces coelicolor. *Appl Environ Microbiol* **85**, doi:10.1128/AEM.02178-18 (2019).
50 Valdez, F., Gonzalez-Ceron, G., Kieser, H. M. & Servi, N. G. L. The Streptomyces coelicolor A3(2) lipAR operon encodes an extracellular lipase and a new type of transcriptional regulator. *Microbiology (Reading)* **145 (Pt 9)**, 2365-2374, doi:10.1099/00221287-145-9-2365 (1999).
51 Aigle, B., Wietzorrek, A., Takano, E. & Bibb, M. J. A single amino acid substitution in region 1.2 of the principal sigma factor of Streptomyces coelicolor A3(2) results in pleiotropic loss of antibiotic production. *Mol Microbiol* **37**, 995-1004, doi:10.1046/j.1365-2958.2000.02022.x (2000).
52 Li, L., Jiang, W. & Lu, Y. A Novel Two-Component System, GluR-GluK, Involved in Glutamate Sensing and Uptake in Streptomyces coelicolor. *J Bacteriol* **199**, doi:10.1128/JB.00097-17 (2017).
53 Mao, X. M. *et al.* Positive feedback regulation of stgR expression for secondary metabolism in Streptomyces coelicolor. *J Bacteriol* **195**, 2072-2078, doi:10.1128/JB.00040-13 (2013).
54 Yang, H. *et al.* The tyrosine degradation gene hppD is transcriptionally activated by HpdA and repressed by HpdR in Streptomyces coelicolor, while hpdA is negatively autoregulated and repressed by HpdR. *Mol Microbiol* **65**, 1064-1077, doi:10.1111/j.1365-2958.2007.05848.x (2007).
55 Urem, M. *et al.* OsdR of Streptomyces coelicolor and the Dormancy Regulator DevR of Mycobacterium tuberculosis Control Overlapping Regulons. *mSystems* **1**, doi:10.1128/mSystems.00014-16 (2016).
56 Nguyen, K. T. *et al.* A central regulator of morphological differentiation in the multicellular bacterium Streptomyces coelicolor. *Mol Microbiol* **46**, 1223-1238, doi:10.1046/j.1365-2958.2002.03255.x (2002).
57 Traag, B. A., Kelemen, G. H. & Van Wezel, G. P. Transcription of the sporulation gene ssgA is activated by the IclR-type regulator SsgR in a whi-independent manner in Streptomyces coelicolor A3(2). *Mol Microbiol* **53**, 985-1000, doi:10.1111/j.1365-2958.2004.04186.x (2004).
58 Amin, R., Reuther, J., Bera, A., Wohlleben, W. & Mast, Y. A novel GlnR target gene, nnaR, is involved in nitrate/nitrite assimilation in Streptomyces coelicolor. *Microbiology* **158**, 1172-1182, doi:10.1099/mic.0.054817-0 (2012).

59 Alkema, W. B., Lenhard, B. & Wasserman, W. W. Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. *Genome Res* **14**, 1362-1373, doi:10.1101/gr.2242604 (2004).
60 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238, doi:10.1186/s13059-019-1832-y (2019).
61 Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).
62 Liu, X., Brutlag, D. L. & Liu, J. S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, 127-138 (2001).
63 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* **20**, 835-839, doi:10.1038/nbt717 (2002).
64 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).
65 Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-315, doi:10.1093/bioinformatics/btg405 (2004).
66 Reese, S. E. *et al.* A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics* **29**, 2877-2883, doi:10.1093/bioinformatics/btt480 (2013).
67 Leek JT, J. W., Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, Torres LC. (R package version 3.38.0, 2020).
68 Faith, J. J. *et al.* Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**, e8, doi:10.1371/journal.pbio.0050008 (2007).
69 Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5**, doi:10.1371/journal.pone.0012776 (2010).
70 Bonneau, R. *et al.* The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol* **7**, R36, doi:10.1186/gb-2006-7-5-r36 (2006).
71 Meyer, P. E., Kontos, K., Lafitte, F. & Bontempi, G. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinform Syst Biol*, 79879, doi:10.1155/2007/79879 (2007).
72 Haury, A. C., Mordelet, F., Vera-Licona, P. & Vert, J. P. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst Biol* **6**, 145, doi:10.1186/1752-0509-6-145 (2012).
73 Nguyen, N. T. T. *et al.* RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res* **46**, W209-W214, doi:10.1093/nar/gky317 (2018).

Acknowledgments

This work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT-UNAM) [IN205918 and IN202421 to JAF-G]. J.M.E.-R. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM); he received fellowship 959406 from CONACYT.

Authors' contributions

Conceptualization, A.Z.A., J.M.E.R., and J.A.F.G.; Methodology, A.Z.A., J.M.E.R., J.K.G.K., and J.A.F.G.; Software, A.Z.A., J.M.E.R., J.K.G.K. and J.A.F.G; Validation, A.Z.A., J.M.E.R., and J.A.F.G; Formal Analysis, A.Z.A., J.M.E.R., J.K.G.K., and J.A.F.G; Investigation, A.Z.A., J.M.E.R., J.K.G.K., and J.A.F.G; Resources, J.A.F.G.; Data Curation, A.Z.A.; Writing – Original Draft Preparation, A.Z.A., J.M.E.R, and J.K.G.K.; Writing – Review & Editing, A.Z.A., J.M.E.R, 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.

Additional Information

Competing interests

The authors declare that they have no competing interests.

Supplementary Information

- Supplementary File 1.docx: Word file with the supplementary figures 1 – 14. NDA and structural analysis of the *Curated_FL-DBSCR-RTB* network. Supplementary methods and references.
- Supplementary File 2.xlsx: Excel file with supplementary tables 1 – 9.
- Supplementary File 3.zip: Compressed folder with all the flat files of the inferred networks.

Figure legends

Figure 1. Interactions curated from literature for *Streptomyces coelicolor* A3(2). a) Number of publications per year and b) Number of interactions reported per year.

Figure 2. a) AUROC and AUPR for each of the methods and the community networks. b) Number of interactions statistically validated by TF

Figure 3. Network comparative by structural properties. a) Pearson correlation of the profile of structural properties listed in Supplementary Figure 12. b) D-value from [43] to identify network similarity.

Figure 4. Simpson index for NDA analysis for all curated and inferred networks. a) Global regulators. b) Modular genes. c) Intermodular genes. d) Basal machinery genes.

Figure 5. MCC for global regulators predicted by NDA for each of the curated and inferred networks. Scores ≥ 0.5 are represented in white numbers and meta-curations are marked with an asterisk. a) Gold standard curated from reference [28]. b) Gold standard curated from independent publications (Supplementary Table 3).

Figure 6. Conservation of the systems-level components between *S. coelicolor* and *C. glutamicum.*

Tables

Table 1. Description of networks used in this work.

Network	Abasy ID	Genes	Interactions	Description

Curated_RTB	100226_v2015_sRTB13	311	330	Network from RegTransBase database
Curated_DBSCR	100226_v2015_sDBSCR15	273	341	Network from Database of transcriptional regulation in *Streptomyces coelicolor* and its closest relatives.
Curated_DBSCR(S)	100226_v2015_sDBSCR15_eStrong	112	115	Filtration of interactions with strong evidence from the DBSCR network.
Curated_FL	100226_v2019_sFL	5331	9454	Network from the collection and curation performed for this work.
Curated_FL(cS)	Not Reported	347	438	Filtration of interactions with strong evidence from the FL network (cS=curated strong)
Curated_FL(S)	100226_v2019_sFL_eStrong	396	493	Filtration of interactions with strong evidence from the FL network along with statistically validated interactions.
Curated_FL-DBSCR-RTB	100226_v2019_sFL-DBSCR15-RTB13	5386	9707	Meta-curation of RTB, DBSCR and FL networks.
Curated_FL(cS)-DBSCR(S)	Not Reported	387	480	Filtration of interactions with strong evidence from the meta-curated network.
Curated_FL(S)-DBSCR(S)	100226_v2019_sFL-DBSCR15_eStrong	435	534	Filtration of interactions with strong evidence from meta-curated networks along with statistically validated interactions.
Inferred_BS	Available as a supplementary file	6263	23908	Inferred GRN from binding sites prediction.
Inferred_Exp	Available as a supplementary file	4739	23908	Inferred GRN from transcriptomic data.
Inferred_BS-Exp	Available as a supplementary file	4763	23908	Community network from Inferred_BS and Inferred_Exp.
Inferred_All	Available as a supplementary file	3804	23908	Community network from all the inference methods.

Table 2. Network properties for inferred networks

Property	**Inferred_BS**	**Inferred_Exp**	**Inferred_BS-Exp**	**Inferred_All**	**Curated_FL-DBSCR-RTB**
Number of nodes (N)	6263	4739	4763	3804	5386
$\ln(\ln(N))$	2.17	2.14	2.14	2.11	2.15
Average shortest path length	2.86	3.38	3.38	3.11	2.84
Average clustering coefficient	0.213	0.385	0.385	0.470	0.182
$\alpha\ (P(k))$	1.861	1.952	1.955	1.968	1.742
R^2_{adj}	0.87	0.92	0.92	0.91	0.84
$\alpha\ (C(k))$	0.924	0.767	0.742	0.729	1.142

R^2_{adj}	0.79	0.58	0.54	0.68	0.89

Figures

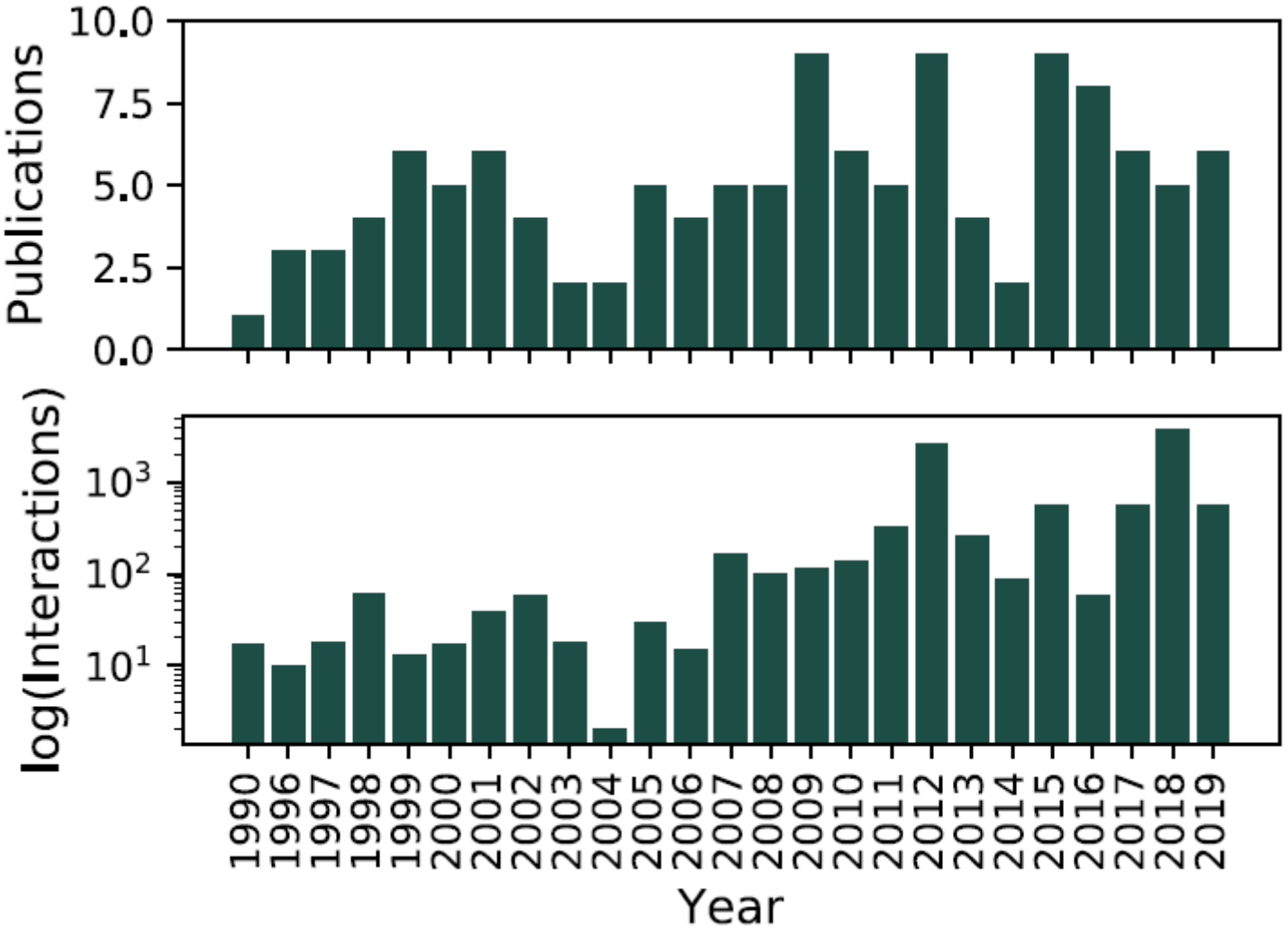

Figure 1

Interactions curated from literature for Streptomyces coelicolor A3(2). a) Number of publications per year and b) Number of interactions reported per year.

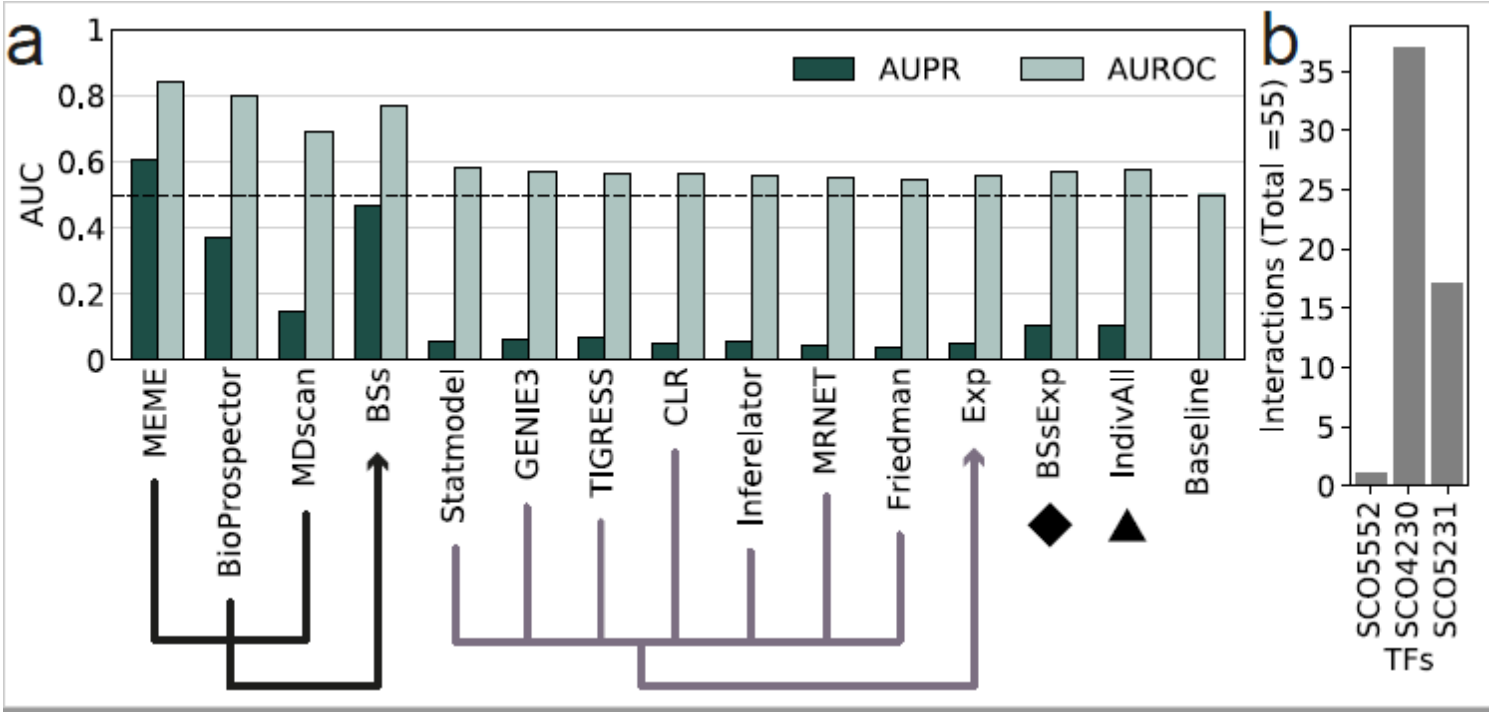

Figure 2

a) AUROC and AUPR for each of the methods and the community networks. b) Number of interactions statistically validated by TF

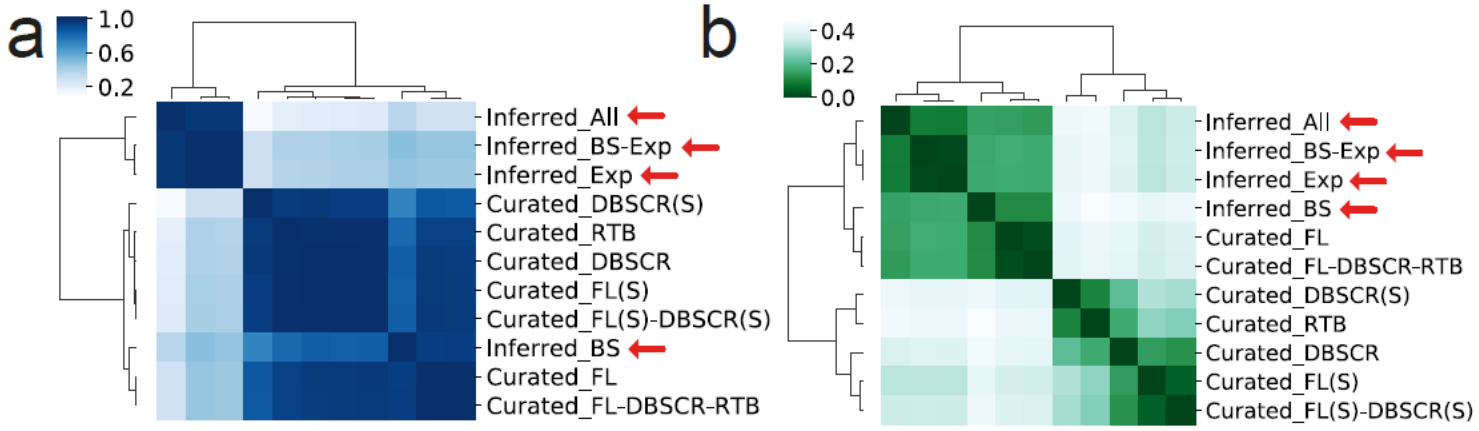

Figure 3

Network comparative by structural properties. a) Pearson correlation of the profile of structural properties listed in Supplementary Figure 12. b) D-value from 43 to identify network similarity.

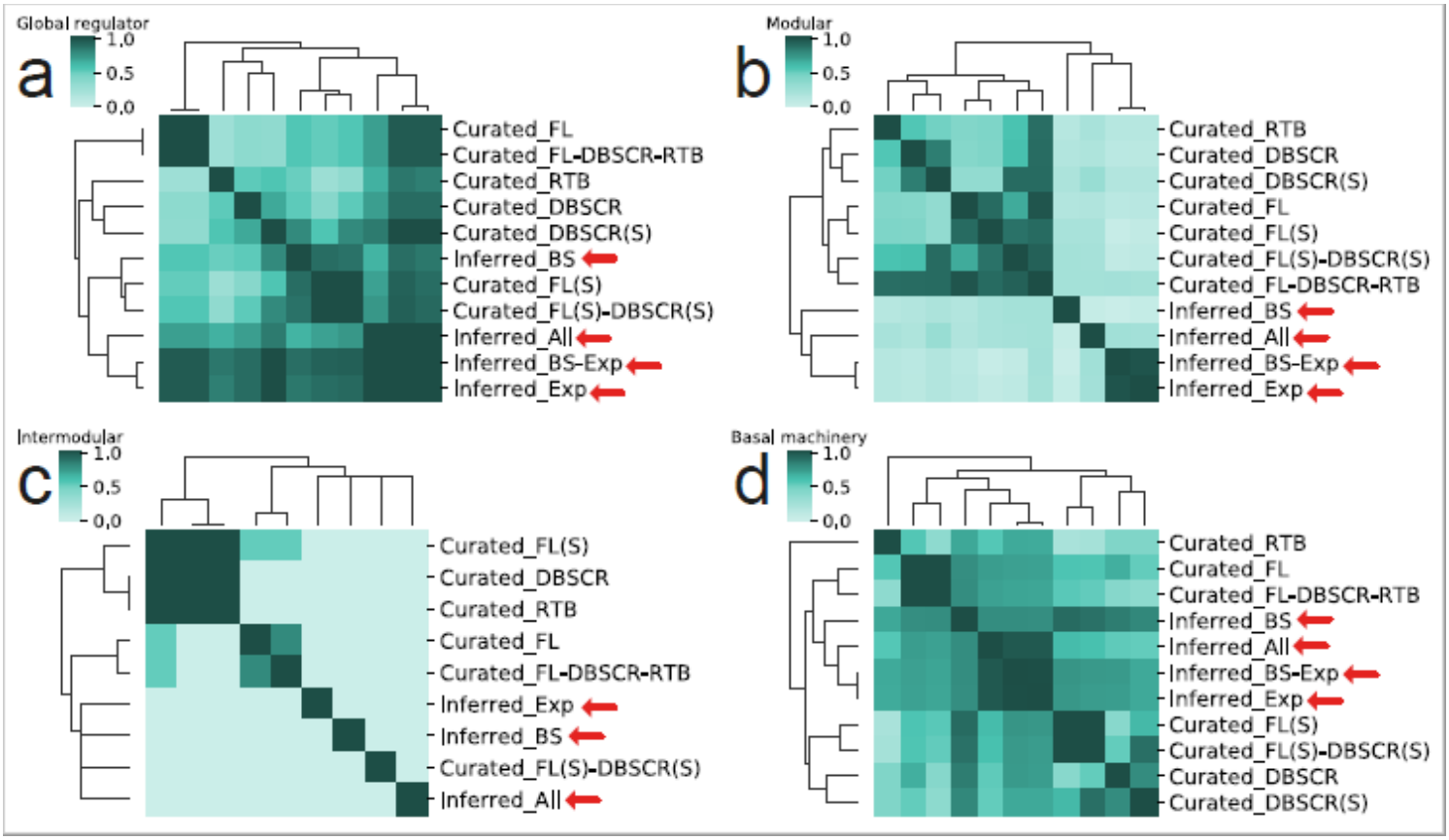

Figure 4

Simpson index for NDA analysis for all curated and inferred networks. a) Global regulators. b) Modular genes. c) Intermodular genes. d) Basal machinery genes.

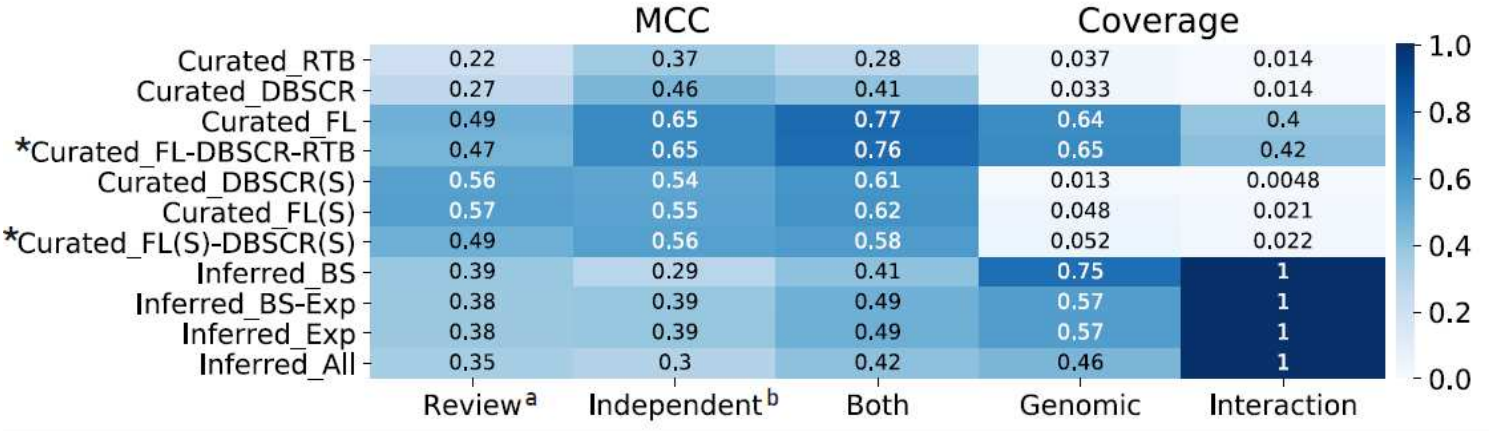

Figure 5

MCC for global regulators predicted by NDA for each of the curated and inferred networks. Scores ≥ 0.5 are represented in white numbers and meta-curations are marked with an asterisk. a) Gold standard curated from reference 28. b) Gold standard curated from independent publications (Supplementary Table 3).

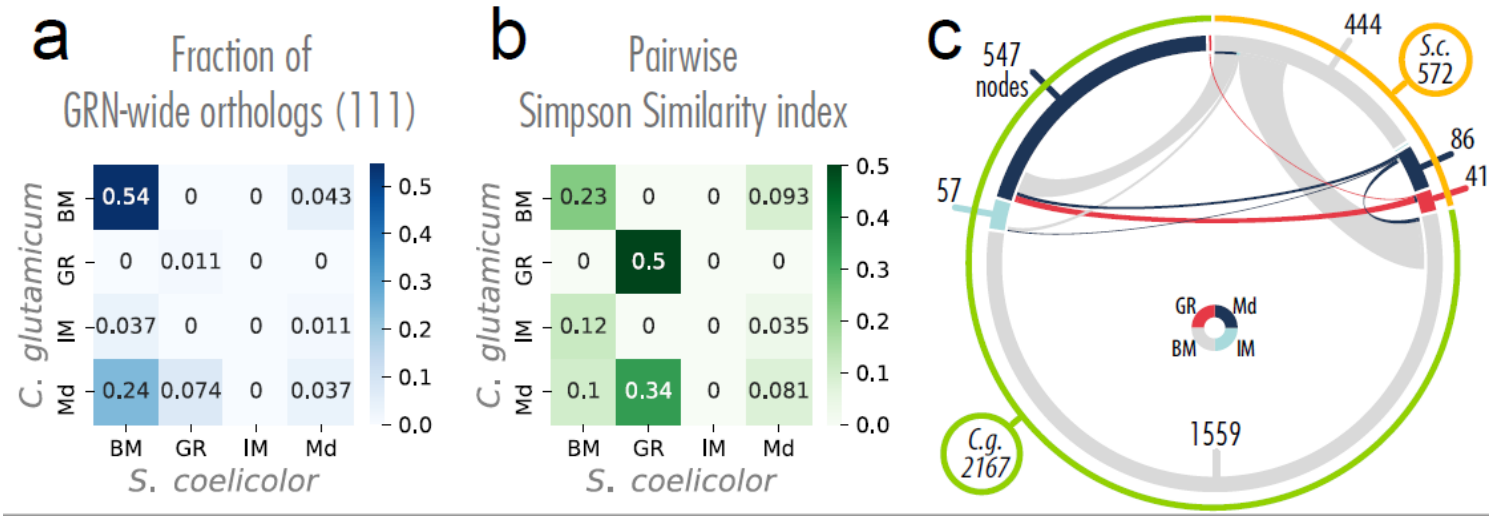

Figure 6

Conservation of the systems-level components between S. coelicolor and C. glutamicum.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1.docx
- SupplementaryFile2.xlsx
- SupplementaryFile3.xlsx