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

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

***Corresponding author:** jfreyre@ccg.unam.mx (JAFG)

Abstract

Some organism-specific databases about regulation in bacteria have become larger, accelerated by high-throughput methodologies, while others are no longer updated or accessible. Each database homogenize its datasets, giving rise to heterogeneity across databases. Such heterogeneity mainly encompasses different names for a gene and different network representations, generating duplicated interactions that could bias network analyses. Abasy (**A**cross-**ba**cteria **sy**stems) Atlas consolidates information from different sources into meta-curated regulatory networks in bacteria. The high-quality networks in Abasy Atlas enable cross-organisms analyses, such as benchmarking studies where gold standards are required. Nevertheless, network incompleteness still casts doubts on the conclusions of network analyses, and available sampling methods cannot reflect the curation process. To tackle this problem, the updated version of Abasy Atlas presented in this work provides historical snapshots of regulatory networks. Thus, network analyses can be performed at different completeness levels, making possible to identify potential bias and to predict future results. We leverage the recently found constraint in the complexity of regulatory networks to develop a novel model to quantify the total number of regulatory interactions as a function of the genome size. This completeness estimation is a valuable insight that may aid in the daunting task of network curation, prediction, and validation. The new version of Abasy Atlas provides 76 networks (204,282 regulatory interactions) covering 42 bacteria (64% Gram-positive and 36% Gram-negative) distributed in 9 species (*Mycobacterium tuberculosis, Bacillus subtilis, Escherichia coli, Corynebacterium glutamicum, Staphylococcus aureus, Pseudomonas aeruginosa, Streptococcus pyogenes, Streptococcus pneumoniae*, and *Streptomyces coelicolor*), containing 8,459 regulons and 4,335 modules.

Database URL: https://abasy.ccg.unam.mx/

Keywords: systems biology, regulatory networks, historical snapshots, completeness, modules, global regulators, intermodular genes, meta-curation

Background

Regulation at the gene transcription level is a fundamental process for bacteria to adapt to different media conditions and to cope with adverse environments. This process is mediated mainly by transcription factors (TFs), proteins capable to promote or hinder the transcription of their target genes (TGs). A TF-coding gene and its TGs conform a regulon, multiple regulons can be assembled to construct a gene regulatory network (GRN) where nodes and edges depict genes and interactions, respectively. Given the different specificity across TFs, they can contribute to organism adaptation in different levels which provides hierarchical and modular properties to GRNs in bacteria [1].

The increasing number of experimental strategies to study the transcriptional machinery [2] has allowed the community to unveil novel regulatory interactions. Despite curation efforts, many interactions remain buried in publications and not integrated into a GRN yet. Organism-specific databases offer expertise and often are the primary resource for further research on the organism of interest. Such databases include RegulonDB [3] for *Escherichia coli*, DBTBS [4] and SubtiWiki [5] for *Bacillus subtilis*, CoryRegNet [6] for *Corynebacterium glutamicum* and MtbRegList [7] for *Mycobacterium tuberculosis*. Nonetheless, many of those databases are no longer updated or accessible [8]. Besides, the availability of multiple organism-specific databases gives rise to heterogeneity, which could bias results when cross-organisms analyses are performed. Such heterogeneity encompasses different names for the same gene and different network representations. This is even a problem for a single organism when complementary databases are integrated.

The analysis of global properties through multiple bacteria have revealed similarities among them [9-14]. Nonetheless, those studies have been limited to only a few organisms and results need to be validated with the most complete GRNs [15]. Besides, the study of the effect of network incompleteness on network structural analyses has been hindered by the limitations in databases to identify when a set of novel interactions is reported, and the experimental evidence supporting those interactions. Since no GRN curation model has been developed, works to study this phenomenon have been limited to simulate the curation process by decomposition or reconstruction of the GRNs by different random models [16,17].

Diverse databases cope with information inconsistency, such as CollecTF [18] for experimentally-validated TF binding sites in bacteria, and GSDB [19] for 3D chromosome and genome topological structures. Other resources integrating and homogenizing experimentally-validated data with computational predictions include STRING [20] for protein-protein interaction networks, SwissRegulon [21] for regulatory sites in prokaryotes and eukaryotes organisms, PRODORIC [22] for DNA binding sites for prokaryotic TFs, RegNetwork [23] for transcriptional and posttranscriptional regulatory relationships for human and mouse, and Network Portal (http://networks.systemsbiology.net/) for coregulation networks. But poor efforts have been carried out to provide consolidated, disambiguated, homogenized high-quality GRNs on a global scale, their structural properties, system-level components, and their historical snapshots to trace their curation process.

Abasy Atlas v1.0 was originally conceived to fill this gap by making a cartography of the functional architectures of a wide range of bacteria [12]. Our database provides a comprehensive atlas of

annotated functional systems (hereinafter also referred to as modules), statistical and structural network properties, and system-level elements for reconstructed and meta-curated (homogeneous and disambiguated) GRNs across 42 bacteria, including pathogenically and biotechnologically relevant organisms. Abasy Atlas is the first database in providing predictions of global regulators, basal machinery genes, members of functional modules, and intermodular genes based on the system-level elements predicted for the natural decomposition approach (NDA) in several bacteria [9,11-13]. The NDA is a biologically-motivated mathematical approach leveraging the global structural properties of a GRN to derive its architecture and classify its genes into one of the four above mentioned categories of system-level elements. Abasy Atlas was also designed to provide statistical and structural properties characterizing the GRNs, such as their associated power laws, percentage of regulators, network density and giant component size, and the number of feedforward and feedback motifs among others.

In this work, we present the expanded version of Abasy (**A**cross-**ba**cteria **sy**stems) Atlas, which consolidates information from different sources into historical snapshots of meta-curated GRNs in bacteria. Each historical snapshot represents the integrated knowledge we had about a GRN at a given time point. The new Abasy Atlas v2.2 makes possible to study the effect of network incompleteness across bacteria on diverse GRNs analyses, to identify potential bias and improvements, and to predict future results with more complete GRNs. Besides, Abasy Atlas GRNs integrates regulation mediated by regulatory proteins, small RNAs, sigma factors and regulatory complexes to better understand the biological systems [24]. This global representation of the GRNs eases their use because the organism-specific databases usually represent each network in a different file and different format, which can convolute the parsing of the network flat files and the integration of information.

While most proteins regulate gene transcription as homodimeric complexes, the regulation of gene expression can also be achieved by heteromeric complexes, being their subunits encoded by different genes. Despite previous integrative approaches merging different level components [25-27], heterodimeric complexes have not been properly represented in most of them nor databases. One of the most common representations is to assign the regulations to each subunit, leading to a duplicated representation of the interaction in the GRNs. The new Abasy Atlas v2.2 provides a homogeneous representation for heteromeric complexes, when information is available, preserving the regulatory information and avoiding duplicated, misleading interactions.

In summary, Abasy Atlas v2.2 provides historical snapshots of reconstructed and meta-curated GRNs across bacteria, their completeness level, topological properties, and system-level components, enabling network completeness-dependent analyses for multiple organisms. Besides, the homogeneity of gene symbols, interactions confidence level, and network representation allow Abasy Atlas GRNs to be used as gold standards for benchmarking purposes, such as those to assess GRN predictions and theoretical models. In the section “Functionality”, we describe studies that would be benefited from the functionality of Abasy Atlas v2.2 [28-35].

Abasy Atlas does not intend to replace organism-specific databases containing regulatory interactions with biological information such as regulatory sites. Conversely, it fills a gap by providing a consolidated version of bacterial GRNs on a global scale, their structural properties, system-level components, and their historical snapshots to trace their curation process. Abasy

Atlas is cross-linked to diverse external databases providing biological, genomic, and molecular details. Cross-links to organism-specific databases included as a source for each GRN are also provided. From there, the user can further inquire about biological considerations such as binding sites annotation, TF conformation, genome annotation, and chromosomal conformation. All essential data when studying the molecular mechanisms and evolution of GRNs in bacteria. In this way, Abasy Atlas serves as an across-organisms database coping with information inconsistency and providing high-quality GRNs on a global scale.

Remarkable uses of previous versions of Abasy Atlas [12] comprise the characterization of *C. glutamicum* GRN [13], the integration of gene regulatory interactions to metabolism to identify the relevant TGs suitable for strain improvement [36], and comparative genomic analyses to characterize the transcriptome profile of *Corynebacterium pseudotuberculosis* in response to iron limitation [37]. Abasy Atlas v2.0 was used to identify evolutionary constraints on the complexity of GRNs enabling the study of three models to predict the total number of genetic interactions [14]. The latter allowed to compute an interaction coverage as a proxy of network completeness, which improves the biased network genomic coverage (fraction of the genome in the network). Abasy Atlas V2.2 could be useful to improve these works since more complete GRNs provide more information regarding transcriptional regulation in medically and biotechnologically relevant organisms such as *M. tuberculosis* and *C. glutamicum*. Also, to improve models developed with the previous version of Abasy, such as the novel network completeness model presented in the section “Estimating GRNs completeness by leveraging their constrained complexity”.

A primer on the natural decomposition approach: predicting global regulators, modular genes shaping functional systems, basal machinery genes, and intermodular genes

Abasy Atlas was designed to provide annotations of the modules and system-level elements integrating each GRN. These predictions are computed by using the NDA. The NDA is a large-scale modeling approach characterizing the circuit wiring and its global architecture. It defines a mathematical-biological framework providing criteria to identify the four classes of system-level elements shaping GRNs: global regulators, modular genes shaping functional systems, basal machinery genes, and intermodular genes. Studies have shown that regulatory networks are highly plastic [38]. Despite this plasticity, by applying the NDA our group has found that there are organizational principles conserved by convergent evolution in the GRNs of phylogenetically distant bacteria [11]. The high predictive power of the NDA has been proven in previous studies by applying it to the phylogenetically distant *E. coli* [9], *B. subtilis* [11], and *C. glutamicum* [13], and by comparing it with other methods to identify modules [39].

The NDA defines objective criteria (e.g., the κ-value to identify global regulators) to expose functional systems and system-level elements in a GRN, and rules to reveal its functional architecture by controlled decomposition (Supplementary Figure 1). It is based on two biological premises [10,11]: (1) a module is a set of genes cooperating to carry out a particular physiological function, thus conferring different phenotypic traits to the cell. (2) Given the pleiotropic effect of global regulators, they must not belong to modules but rather coordinate them in response to general-interest environmental cues.

According to the NDA, every gene in a GRN is predicted to belong to one out of four possible classes of system-level elements, which interrelate in a non-pyramidal, three-tier, hierarchy shaping the functional architecture [10-13] as follows (Supplementary Figure 2): (1) Global regulators are responsible for coordinating both the (2) basal cell machinery, composed of strictly globally regulated genes and (3) locally autonomous modules (shaped by modular genes), whereas (4) intermodular genes integrate, at the promoter level, physiologically disparate module responses eliciting combinatorial processing of environmental cues.

Construction and content

Abasy Atlas current content

Abasy Atlas v2.2 provides the most complete set of experimentally curated GRNs across bacteria. Abasy Atlas represents regulatory interactions by using network models where nodes represent genes or regulatory protein complexes, and directed links depict regulatory interactions. Since the release of Abasy Atlas v1.0 in 2016 [12], the number of GRNs has increased from 50 to 76 (+52%) covering 42 bacteria (64% Gram-positive and 36% Gram-negative) distributed in 9 species (*Mycobacterium tuberculosis*, *Bacillus subtilis*, *Escherichia coli*, *Corynebacterium glutamicum*, *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Streptococcus pyogenes*, *Streptococcus pneumoniae*, and *Streptomyces coelicolor*) and 41 strains (Figure 1A and Supplementary Figure 3).

These 76 GRNs comprise 204,282 regulatory interactions (+160%) organized into 8,459 (+128%) regulons and 4,335 modules (+144%). We homogenized the representation of heteromeric TFs and their subunits and obtained a total of 12 heteromeric TFs, all of them in the GRN of *E. coli* K-12. However, this paves the way for a homogeneous representation of GRNs that will be propagated to more organisms in a future version of Abasy Atlas, when information regarding heteromeric TFs for these organisms is available. A total of 20 historical snapshots for the model organisms *M. tuberculosis*, *B. subtilis*, *E. coli,* and *C. glutamicum* were also included in the Abasy Atlas v2.2.

Unique machine-readable, user-friendly identifiers for each GRN reconstruction

Studies using GRNs from organism-specific databases usually cite the source database. However, while some articles specify the GRNs used [28,39], others do not [9,40]. This drives to a reproducibility problem when the database updates the GRN and does not provide the historical snapshots. To cope with this problem, a machine-readable and user-friendly identifier was assigned to each network to ease reporting and identification when using the database.

Network identifiers are constructed as follows: Five fields are separated by an underscore, three are mandatory and two are optional. The first field represents the NCBI taxonomy ID of the organism (mandatory). The second field, preceded by a “v”, which stands for version, is the year when the network was reconstructed (mandatory). The field starting with an “s” provides information about the sources from which the network was reconstructed (mandatory). The confidence level of the evidence supporting the regulatory interactions is described by an optional

field starting with an "e". When this field is omitted means that the reconstruction contains all the available interactions disregarding the confidence level of evidence, whereas "strong" is used for those GRNs reconstructed only with interactions validated by direct experimental evidence. An optional description field, preceded by a "d", enables to include keywords such as "sRNA" for GRNs containing sRNAs-controlled regulons (Figure 1B).

The source field, that starting with an "s", is composed by a database name abbreviation and year when meta-curated from databases, and the last two digits of the publication year when curated from literature (see Supplementary Table 1 for a complete list of data sources abbreviations and references). On the "Browse" page of Abasy Atlas, the user can identify the source for each GRN, as well as for the subnetworks when the GRN is a meta-curation from different sources.

Historical snapshots of the GRNs

Network theory-based approaches to study the organizing principles governing GRNs have been pointed to be biased by the curation process and incompleteness [16,41]. Nevertheless, those studies have been mainly applied to subnetworks sampled by different random computational algorithms that cannot reproduce faithfully the curation process by the scientific community. To bring an alternative solution to this problem, we have been curating organism-specific databases and literature during the construction of Abasy Atlas in different time points for several organisms (hereinafter referred to as historical snapshots). Namely, nine historical snapshots for *E. coli,* four for *C. glutamicum,* four for *B. subtilis,* and three for *M. tuberculosis* (Figure 2).

Each historical snapshot represented in Figure 2 is the most complete version of the GRNs at that time point. However, individual GRNs are also available. For example, the historical snapshot of the GRN of *B. subtilis* in 2017 (224308_v2017_sDBTBS08-15-SW18, Figure 2) integrates regulatory interactions from two organism-specific databases (DBTBS [42] and SubtiWiki [5]) and one article [43] (Figure 3). The individual GRNs are available with their own network ID (224308_v2008_sDBTBS08_eStrong, 224308_v2017_sSW18, and 224308_v2015_s15, respectively). Note that the GRN from DBTBS is also the first historical snapshot for *B. subtilis* (Figure 2), and GRNs from different sources do not need to be from the same year since a new historical snapshot integrates every previous GRNs. The network integration and homogenization from different sources enables cross-bacteria analyses with the historical snapshots.

We will continue querying organism-specific databases and curating literature periodically to obtain more complete versions of the GRNs. Also, we will extend the historical snapshots to other organisms as information will be available.

Meta-curation of GRNs: Quality control coping with inconsistency and preserving information from the different sources

The heterogeneity in gene symbols and network representations often conduces to redundancy and loss of information. As a consequence, this heterogeneity can result in misleading network reconstructions. The meta-curation process mainly consists of homogenizing gene symbols and

network representation before merging interactions from different sources. To cope with gene symbols disagreement among regulatory datasets from different sources, we gathered gene name, locus tag, and synonyms for each gene in the GRNs. Then, we developed an algorithm to map gene symbols onto unambiguous canonical gene names and locus tags. This allowed us to remove a total of 223 redundant nodes and 412 redundant interactions from the current set of GRNs (Supplementary Figure 4). We refer the reader to version 1.0 of Abasy Atlas for further information about the gene symbols disambiguation algorithm [12]. For the graphical network representation, we use the unambiguous canonical gene name when available or locus tag. This eases to identify genes of interest. However, the mapping of gene identifiers allows the user to use the search box with different gene symbols and synonyms mapping to the same gene and navigate through the neighborhood of the gene of interest.

Abasy Atlas also provides the confidence level supporting each interaction since GRNs composed with different confidence-levels may bias their structural properties [14]. Therefore, a "strong" or "weak" confidence level is assigned to each interaction according to an expanded scheme based on the one proposed by RegulonDB [44,45]. The basic idea of the confidence level scheme is to label as "strong" only those interactions with direct, non-ambiguous experimental support such as DNA binding of purified TF [45]. Besides, the meta-curated networks that merge regulons from different sources also integrate the effect and the evidence level. This makes the GRNs from Abasy Atlas the most complete collection of homogenous versions in contrast to those individual GRNs available in organism-specific databases.

One of the main caveats of consolidating networks is the non-machine readable, heterogeneous way to represent the information about the way a TF regulates a specific TG and the evidence supporting such interaction, mainly for community-updated databases. To tackle this problem, we manually curate those attributes from different sources when available. Thus, Abasy Atlas makes possible to know in a homogenous fashion whether a TF promotes or hinders its TGs transcription even for interactions from a community-updated database such as SubtiWiki. Therefore, if the same interaction from a different source share effect but diverge on evidence, the interaction and the "strong" evidence is conserved since one directly experimentally validated interaction is enough to classify the edge as "strong" [45]. On the other hand, in case of different effects and the same evidence level, both effects are conserved in a single dual interaction to avoid redundancy. In the case that both attributes are different, only the "strong" interaction is conserved (Supplementary Figure 5). This meta-curation process allows us to reconstruct the most complete GRNs available preserving information from the different complementary sources (Figure 3).

Meta-curation of GRNs: Quality control filtering spurious interactions by reassessing the confidence level of each interaction

We perform a meta-curation process to reduce the number of spurious interactions, thereby reassessing the confidence level of the interactions. Although networks with "weak" evidence are a valuable resource to study the transcriptional regulation, only directly experimentally validated interactions offer the reliability needed to use GRNs as gold standards. Abasy Atlas eases the selection of gold standards for benchmarking purposes through ready-to-download filtered "strong" GRNs (Supplementary Figure 6).

Using the historical snapshots of the *E. coli* GRNs, we analyzed how often a regulatory interaction identified by a "weak" methodology was validated as "strong" evidence. We found that the number of interactions identified for each methodology varies in a wide range, as well as its fraction of predictions validated as "strong" (Figure 4A). Namely, "inferred computationally without human oversight" (ICWHO) is the evidence with the lowest fraction of validated interactions (Figure 4A and Supplementary Figure 7). On the other hand, "RNA-polymerase footprinting" (RPF) is the only methodology having a 100% of interaction validated as "strong" evidence, and > 50% of "gene expression analysis" (GEA) predictions have been validated despite being the "weak" evidence with the highest number of predictions.

We further analyzed the effect of the interactions with ICWHO as its unique evidence, and found that most of these interactions were present in the 2013 and 2014 time points but no longer in 2015 or later. Being this the reason for the outstanding completeness of these network reconstructions and its unusual system-level elements proportions (Figure 4B). For this reason, we decided to exclude predictions being supported only by the ICWHO evidence in Abasy Atlas. This analysis highlights the capability of the system-level properties to assess GRNs quality. It is important to note that despite the small fraction of validated interactions inferred by "non-traceable author statement" (NTAS) (Supplementary Figure 7), we did not remove interactions supported only by this evidence since the number of predicted interactions is very small (Figure 4A).

Estimating GRNs completeness by leveraging their constrained complexity

The ability to quantify the total number of interactions in the complete GRN of an organism is a valuable insight that will leverage the daunting task of curation, prediction, and validation by enabling the inclusion of prior information about the network structure. Besides, the ability to track the completeness, quantified as the fraction of the known interactions from the total number in the complete network (interaction coverage), through different historical snapshots could allow to develop models on how new regulatory interactions are discovered and to provide a framework to assess network analysis and network inference tools. But poor efforts have been directed towards the longstanding problem of how to assess the completeness of these networks. Traditionally, network genomic coverage has been used as a proxy of completeness. The genomic coverage of a regulatory network is the fraction of genes in the network relative to the genome size. Nevertheless, this measure poses potential biases as it neglects regulatory redundancy and the combinatorial nature of gene regulation, thus potentially overestimating network completeness.

For example, the addition of a global regulon or sigmulon (perhaps discovered by high-throughput methodologies) to a quite incomplete regulatory network could bias the genomic coverage. Assume you have a regulatory network with a genomic coverage of 15% (600 / 4000) and 700 interactions. You then found a paper reporting the promoter mapping for the corresponding housekeeping sigma factor, whose sigmulon has 3000 genes (400 of which were already in the original network). Next, you found that 100 out of the 3000 interactions in the global sigmulon already exist in our original network. You then integrate all the remaining 2900 new interactions to your original network to found that your resulting network has a new genomic coverage of 80%

(3200 / 4000) and 3600 interactions. This new high genomic coverage may suggest a highly complete network but it is indeed the same quite incomplete original network plus a single global sigmulon. To clarify this, assume that the total number of interactions in the complete network is 10000, then the completeness of this new network is 36% (3600 / 10000). Whereas the curation of a single housekeeping sigmulon increased the completeness ~30% (3600 / 10000 - 700 / 10000), the new completeness is still low, and the genomic coverage is highly overestimating when is used as a proxy for completeness. Therefore, to correctly state the completeness of a regulatory network, it is fundamental to estimate the total number of interactions. Two recent works have simultaneously provided estimations on the size of GRNs [14,46].

On one hand, the RegulonDB team carried out an exploratory analysis [46]. They used a single version of the *E. coli* regulatory network and high-throughput datasets of binding experiments for around 15 TFs. By assuming a linear model they found an upper-bound estimate of 45759 regulatory interactions. They claimed that only one-third of the ~46000 will affect gene expression, concluding that the complete network comprises only around 13000 interactions.

Alternatively, our group recently explored the constraints on several structural properties of the 71 regulatory networks deposited in Abasy Atlas v2.0 [14]. We found that the network density (d) as a function of the number of genes (n) follows a power law as $d \sim n^{-\gamma}$ with $\gamma \approx 1$. Since 1972, a seminal paper by Robert May showed that the frontier between dynamical stability and instability for a complex system follows a power law as $d \sim n^{-1}$, relating complexity quantified via the density of interactions and the number of variables (the size of the system) [47]. The density of interactions (network density) is the fraction of potential interactions that are real interactions, thus a constraint in network density implies a constraint in the total number of interactions in the complete network. As we found that density is constrained in GRNs, we explored three possible models to predict the total number of interactions as a function of the number of genes (see Figure 4 in [14]): edge regression (assuming linearity, $R^2 = 0.90$), density invariance (assuming an invariant density, $R^2 = 0.86$) and density proportionality (assuming an exponential decay, $R^2 = 0.91$). All the models had a good fit to the data ($0.86 \leq R^2 \leq 0.91$), with small differences between them. These models predicted that the total number of interactions in the complete *E. coli* regulatory network is ~10000, ~14000, and ~11000, respectively.

After publication, we reformulated the problem. As regulatory networks are directed and self-regulations are allowed, the maximum number of possible interactions (I_{max}) is n^2 as each of the n genes could regulate to other n genes including itself (self-regulation). The density of a regulatory network must be then computed as

$$d = \frac{I}{I_{\max}} = \frac{I}{n^2}$$

By introducing this equation into the power law found for the density of the Abasy Atlas networks ($d \sim n^{-\gamma}$), we derived another power law modeling the total number of interactions in the regulatory network as a function of the number of genes as

$$I = dn^2 \sim n^{-\gamma} n^2 \sim n^{2-\gamma}$$

This model has a better fit to data (Figure 5, $R^2 = 0.98$) than the previous three models, and allows us to compute the total number of interactions in the regulatory network of an organism as $I_{total} \sim$

(genome size)$^{2-\gamma}$. We implemented this model in Abasy Atlas v2.2 to provide estimations on the completeness of each regulatory network, including confidence intervals. The power-law model predicts that the complete *E. coli* regulatory network will have 11,656 total regulatory interactions. This model can learn the tendency in the number of interactions, and it improves as more regulatory networks are included in Abasy Atlas. That is one of the reasons motivating us to continue expanding Abasy Atlas by adding new organisms and historical snapshots.

Homogeneous representation for heteromeric transcription factor complexes

Even though heterodimeric regulatory complexes are not overrepresented in regulatory networks, some of them are global regulators and their interactions control up to ~10% of the genome and represent a valuable percent of the whole network (~6% in *E. coli* GRNs). IHF is a global regulator histone-like protein of *E. coli* that regulates transcription as a heterodimeric complex that is shaped by two different proteins: IhfA and IhfB. Although both subunits can form homodimeric complexes, the affinity for DNA is much lower [48], and no regulation in such fashion has been reported. For this reason, assigning the regulatory activity to each subunit (a gene-gene representation, Figure 6B) is a misleading representation. Additionally, the RpoS sigma factor allows the transcription of both subunits conforming IHF, which in turn also regulates its subunits (Figure 6A). Such interesting autoregulation cannot be properly represented in a gene-gene based representation (Figure 6B). Conversely, a representation of the IHF heteromeric complex regulating *ihfA* and *ihfB* is better as it depicts the IHF conformation and links them to the TFs regulating their transcription.

This representation is also useful for subunits of heteromeric regulatory complexes that can exhibit regulation in a homodimeric fashion, such as the *relB* product regulating *relE*, *hokD*, and its transcription both as a homodimer and as part of the RelBE complex with *relE* (Figure 6c). This RelE-RelB toxin-antitoxin system in *E. coli* [49] is not properly represented in a gene-gene network (Figure 6d) as it shows regulatory activity by the *relE* product on its own. This representation eases the application of the networks as gold standards for inference methods such as those based on the DNA sequence and TF binding sites prediction. For analysis requiring GRNs composed only by genes, Abasy Atlas provides the required information to identify the classification of each biological entity (Supplementary Figure 8). Currently, Abasy Atlas comprises a total of 12 heteromeric TFs, all of them in the meta-curated GRN of *E coli* K-12 obtained from RegulonDB [46]. Future development includes the addition of heteromeric TFs in those organisms where this information is available.

Updates for model organisms

Corynebacterium glutamicum ATCC 13032

The PubMed database was screened to find papers published between January 2017 and August 2018 and describing new transcriptional regulatory interactions of *C. glutamicum*, in addition to the comprehensive data set previously deposited in Abasy Atlas [13]. Four new regulators of different types have been examined in detail, exerting in total 63 new direct transcriptional

interactions. Moreover, the predicted regulatory role of the AraC/XylR-type protein Cg2965 (PheR) has been confirmed by experimental data [50,51]. PheR activates the expression of the *phe* gene (*cg2966*) encoding phenol hydroxylase, allowing *C. glutamicum* to degrade phenol by a meta-cleavage pathway. Electrophoretic mobility shift assays (EMSAs) demonstrated a direct interaction of the purified PheR protein with the *phe* promoter region [51]. The MarR-type regulator CrtR (Cg0725) is encoded upstream and in divergent orientation of the carotenoid biosynthesis operon *crtEcg0722crtBIYEb* in *C. glutamicum*. DNA microarray experiments revealed that CrtR acts as a repressor of the *crt* operon. Additional EMSAs with purified CrtR showed that CrtR binds to a region overlapping the –10 and –35 promoter sequences of the *crt* operon [52].

The two-component system EsrSR (Cg0707/Cg0709) controls a regulon involved in the cell envelope stress response of *C. glutamicum* [53]. Interestingly, the integral membrane protein EsrI (Cg0706) acts as an inhibitor of EsrSR under non-stress conditions. The resulting three-component system EsrISR directly regulates a broad set of genes, including the *esrI-esrSR* locus itself, and genes encoding heat shock proteins (*clpB*, *dnaK*, *grpE*, *dnaJ*), ABC transporters and putative membrane-associated or secreted proteins of unknown function. Among the target genes of EsrSR is moreover *rosR* (*cg1324*) encoding a hydrogen peroxide-sensitive transcriptional regulator of the MarR family and playing a role in the oxidative stress response of *C. glutamicum* [53,54].

The extracytoplasmic function sigma factor SigD (Cg0696) is a key regulator of mycolate biosynthesis genes in *C. glutamicum* [55]. Chromatin immunoprecipitation coupled with DNA microarray (ChIP-chip) analysis detected SigD-binding regions in the genome sequence, thus establishing a consensus promoter sequence for this sigma factor. The conserved DNA sequence motif 5'-GTAAC-$N_{17(16)}$-CGAT-3' was found in all ChIP-chip peak regions and presumably corresponds to the –35 and –10 promoter regions recognized by SigD. The *rsdA* (*cg0697*) gene, located immediately downstream of *sigD*, is under direct control of a SigD-dependent promoter and encodes the corresponding SigD anti-sigma factor [55].

The WhcD protein (Cg0850) interacts with WhiA (Cg1792) to exert jointly an important regulatory effect on cell division genes of *C. glutamicum* [56]. WhiA is an exceptional transcriptional regulator as it has been classified as a distant homolog of homing endonucleases that retained only DNA binding activity [57]. Binding of the WhcD-WhiA complex to the promoter region of the cell division gene *ftsZ* was observed by EMSAs using purified fusion proteins, although WhcD alone did not bind to the genomic DNA. The sequence motif 5'-GACAC-3' was found to be important for binding of the WhcD-WhiA complex to the DNA. Also, loss of the DNA-binding activity of WhiA in the presence of an oxidant indicated a regulatory role for this protein to control cell division of *C. glutamicum* under oxidative stress conditions [56].

These interactions were merged with the previous version of the GRN for *C. glutamicum* and included as a new historical snapshot (196627_v2018_s17) with 2317 genes (73.8% of genomic coverage) and 3444 interactions (45.8% of interaction coverage) (Figure 2). The "strong" version of the network was also included, containing a total of 2237 genes (71.3% of genomic coverage) and 2969 interactions (39.5% of interaction coverage).

Mycobacterium tuberculosis H37Rv

Chauhan et al. [58] reported 41 experimentally validated interactions among sigma factors and transcribed genes in the human pathogen *M. tuberculosis*. These interactions were added to the most recent *M. tuberculosis* GRNs and deposited in Abasy Atlas. The regulations among the sigma factors and TGs constitute a valuable contribution to the understanding of how *M. tuberculosis* sigma factors regulate their expression and therefore, their cellular concentrations to compete for the available RNA polymerases. Historical snapshots for the years 2015, 2016, and 2018 are available so far (Figure 2).

Bacillus subtilis subtilis 168

Interactions from the most recent big update of SubtiWiki [5] were merged with the last version of Abasy Atlas including interactions from DBTBS [4] and a non-database hosted publication [43]. The result represents a new time point in the *B. subtilis* GRN history. Until now, a total of four historical snapshots are available for this representative Gram-positive organism (Figure 2), being the last one the GRN with the highest genomic coverage in Abasy Atlas.

Escherichia coli K-12 MG1655

RegulonDB [46] is one of the first organism-specific databases for transcriptional regulation data and it continues being updated. This makes *E. coli* the organism with a higher number of historical snapshots. Meta-curated GRNs from 2003 to 2018 depict the effect of the curation process in this Gram-negative model organism (Figure 2). The meta-curation of the GRNs in Abasy Atlas reassesses the confidence level of the interactions (see "Construction and content"), and integrates the regulations by TFs, sRNAs, and sigma factors from RegulonDB into a global regulatory network.

Utility and discussion

User interface

From the "Home" page you can find the description and statistics of Abasy Atlas, as well as links of interest. In the "Browse" page you can find the species for which a global GRN is deposited in Abasy Atlas, along with the number of items (networks) for such species. Further, you can click on the species to identify the strains available and even the confidence level you need. After the selection of the strain and the confidence level, you will find the historical snapshots available for the GRN of interest, as well as additional information such as the genomic and interaction coverage, data sources, and fraction of the system-level components predicted by the NDA (Supplementary Figure 9). By clicking on "Global properties" you will find statistical and structural properties characterizing the GRN of interest. Such properties include the number of transcription factors, network density, size of the giant component, number of feedforward and feedback motifs, among others. On the same page, you can find the plots for degree, out-degree and

clustering coefficient distributions (Supplementary Figure 10). We fitted these distributions to a power-law using robust linear regression of log-log-transformed data with Huber's T for M-estimation. This overcomes the negative effect of outliers, in contrast to ordinary least squares, which is highly sensitive to outliers in data.

You can directly search for a specific gene in the upper-right box from any page. Once you are visualizing the subnetwork of interest, using the interactive panel (Supplementary Figure 11) you can customize the visualization with several buttons and download the subnetwork as a high-definition PNG image, as well as the JSON file. Every global network can be downloaded from the "Downloads" page (Supplementary Figure 6). Regulatory networks are provided in JSON data-interchange format, including NDA predictions and, when available, effect and evidence supporting regulatory interactions. JSON is an open standard file format, which is a lightweight, language-independent, widely used, data-interchange format supported by > 50 programming languages (e.g., Python, R, Matlab, Perl, Julia, JavaScript, PHP) through a variety of readily available libraries. JSON uses human-readable text to store and transmit data objects consisting of attribute-value pairs and array data types. The JSON data files downloadable from Abasy Atlas are readily importable into Cytoscape for further analyses. Gene information and module annotation flat files in tab-separated-value file formats are also available for download. Information on how to parse the JSON files is available in the "Downloads" page. The citation policy, and the methodology to identify the system-level elements and to predict the interaction coverage is available in the "About" page. You can find additional help on the "Help" page, and contact us on the "Contact" page for any subject, we will appreciate your feedback.

Functionality

Following, we describe some remarkable cases where this new version of Abasy Atlas could have been applied to improve the studies:

The DREAM5 consortium assessed to identify the best methodology to predict GRNs from gene expression data [28] using *E. coli* and *Staphylococcus aureus* as prokaryotic models. However, they did not study how its assessment was affected by network incompleteness. This analysis can be carried out by using the set of the historical snapshots for model organisms as gold standards. The same could be applied for other assessments such as identifying the best tools to predict TF binding sites [29], DNA motifs [29,30,59], and functional modules [31].

Further, Abasy Atlas could be used to extend those benchmarking studies to include more organisms. For example, DREAM5 considered only *E. coli* as a prokaryotic model to compute the overall score because a sufficiently large set of experimentally validated interactions for *S. aureus* did not exist at that time [28]. Currently, Abasy Atlas provides GRNs for 13 *S. aureus* strains, being USA300/TCH1516 the most complete one with 25 and 30.6% of genomic and interaction coverage, respectively.

In addition to benchmarking improvements, the comprehensive atlas of GRNs that Abasy Atlas provides could be applied to study the communication that exists between the regulation of gene transcription with other mechanisms such as protein-protein interactions and metabolism [32-34]. Even when only the regulation of gene transcription is studied, across-organisms information

provided by Abasy Atlas can be used to trace the evolution of the GRN in bacteria, and compare them using gene orthology and network alignment [35]. Future development of Abasy Atlas includes GRNs comparative analyses based on their structural properties.

Future development

Despite high-throughput strategies to study transcriptional regulation, there is a lack of novel interactions reported in contrast with earlier years (Figure 2). Besides, only a handful of organisms have been experimentally studied. Computational approaches have been a hopeful option for non-model organisms and a plethora of algorithms to infer GRNs have emerged. Nonetheless, many of them are based solely on statistical approaches lacking biological constraints to filter spurious interactions. Previous assessments of tools to infer GRNs have unveiled their poor performance but also have shed light on the possibility to increase precision by consensus approaches and biological constraints [28].

Future development of Abasy Atlas aims to include inferred non-model organisms GRNs in a conservative fashion by different consensus-based approaches and the application of currently available data to validate predicted networks by using GRN organizing constraints, such as the composition of system-level elements (Figure 4B) and network structural properties. The addition of heteromeric TFs for more organisms is also considered in the short-term future development. Mainly for the model organisms *C. glutamicum* and *B. subtilis* for which more information regarding regulation by heteromeric TFs is available. Also, historical snapshots for non-model organisms already available in Abasy Atlas, such *Streptomyces coelicolor* will be included, while continuing including additional historical snapshots for model organisms curated from the literature and organism-specific databases. Finally, a python library providing an API to allow programmatic access to Abasy Atlas, and a REST API are under development.

Conclusions

Beyond the regulon level, Abasy Atlas provides the most complete and reliable set of GRNs for many bacterial organisms, which can be used as the gold standard for benchmarking purposes and training data for modeling and network prediction. Besides, Abasy Atlas provides historical snapshots of regulatory networks. Therefore, network analyses can be performed with GRNs having different completeness levels, making it possible to identify how a methodology is affected by the incompleteness, to pinpoint potential bias and improvements, and to predict future results. Additionally, Abasy Atlas is the first database providing estimations on the completeness of GRNs, their global regulators, modules, and other system-level components. The estimation of the total number of regulatory interactions a GRN could have is a valuable insight that may aid in the daunting task of network curation, prediction, and validation. Furthermore, the prediction of the system-level elements in GRNs has allowed unraveling the complexity of these networks and provides new insights into the organizing principles governing them, such as the diamond-shaped, three-tier, hierarchy unveiled by the NDA. The GRNs in Abasy Atlas 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 aimed to understand the common organizing principles and particular lifestyle adaptations of regulatory systems across bacteria and to implement those principles into future work such as the reverse engineering of GRNs.

Availability and requirements

Abasy Atlas is available for web access at https://abasy.ccg.unam.mx. If you use any material from Abasy Atlas please cite properly. Use of Abasy Atlas and each downloaded material is licensed under a Creative Commons Attribution 4.0 International License. Permissions beyond the scope of this license may be available at jfreyre@ccg.unam.mx. **Disclaimer:** Please note that original data contained in Abasy Atlas may be subject to rights claimed by third parties. It is the responsibility of users of Abasy Atlas to ensure that their exploitation of the data does not infringe any of the rights of such third parties.

List of abbreviations

GRN – Gene regulatory network

TFs – Transcription factors

TGs – Target genes

NDA – Natural decomposition approach

Conflict of interest

The authors declare no conflicts of interest.

Funding

This work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT-UNAM) [IN205918 to JAFG].

Acknowledgments

We thank all anonymous reviewers for their positive and encouraging suggestions that helped to improve the manuscript. JMER was supported by an undergraduate fellowship from DGAPA-UNAM. He is also a PhD student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 959406 from CONACYT.

References

1. Barabasi, AL, Oltvai, ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101-113.
2. Geertz, M, Maerkl, SJ (2010) Experimental strategies for studying transcription factor-DNA binding specificities. Brief Funct Genomics 9: 362-373.
3. Gama-Castro, S, Salgado, H, Santos-Zavaleta, A, Ledezma-Tejeida, D, Muniz-Rascado, L, et al. (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res 44: D133-143.
4. Makita, Y, Nakao, M, Ogasawara, N, Nakai, K (2004) DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. Nucleic Acids Res 32: D75-77.
5. Zhu, B, Stulke, J (2018) SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism Bacillus subtilis. Nucleic Acids Res 46: D743-D748.
6. Pauling, J, Rottger, R, Tauch, A, Azevedo, V, Baumbach, J (2012) CoryneRegNet 6.0--Updated database content, new analysis methods and novel features focusing on community demands. Nucleic Acids Res 40: D610-614.
7. Jacques, PE, Gervais, AL, Cantin, M, Lucier, JF, Dallaire, G, et al. (2005) MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis. Bioinformatics 21: 2563-2565.
8. Wren, JD, Bateman, A (2008) Databases, data tombs and dust in the wind. Bioinformatics 24: 2127-2128.
9. Freyre-Gonzalez, JA, Alonso-Pavon, JA, Trevino-Quintanilla, LG, Collado-Vides, J (2008) Functional architecture of Escherichia coli: new insights provided by a natural decomposition approach. Genome Biol 9: R154.
10. Freyre-Gonzalez, JA, Trevino-Quintanilla, LG (2010) Analyzing Regulatory Networks in Bacteria. Nature Education 3: 24.
11. Freyre-Gonzalez, JA, Trevino-Quintanilla, LG, Valtierra-Gutierrez, IA, Gutierrez-Rios, RM, Alonso-Pavon, JA (2012) 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.
12. Ibarra-Arellano, MA, Campos-Gonzalez, AI, Trevino-Quintanilla, LG, Tauch, A, Freyre-Gonzalez, JA (2016) Abasy Atlas: a comprehensive inventory of systems, global network properties and systems-level elements across bacteria. Database (Oxford) 2016.
13. Freyre-Gonzalez, JA, Tauch, A (2017) 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.
14. Campos, AI, Freyre-Gonzalez, JA (2019) Evolutionary constraints on the complexity of genetic regulatory networks allow predictions of the total number of genetic interactions. Sci Rep 9: 3618.
15. Beber, ME, Muskhelishvili, G, Hutt, MT (2016) Effect of database drift on network topology and enrichment analyses: a case study for RegulonDB. Database (Oxford) 2016.
16. Lima-Mendez, G, van Helden, J (2009) The powerful law of the power law and other myths in network biology. Mol Biosyst 5: 1482-1493.
17. Sanz, J, Cozzo, E, Borge-Holthoefer, J, Moreno, Y (2012) Topological effects of data incompleteness of gene regulatory networks. BMC Syst Biol 6: 110.
18. Kilic, S, White, ER, Sagitova, DM, Cornish, JP, Erill, I (2014) CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. Nucleic Acids Res 42: D156-160.

19. Oluwadare, O, Highsmith, M, Cheng, J (2019) GSDB: a database of 3D chromosome and genome structures reconstructed from Hi-C data. bioRxiv: 692731.
20. Szklarczyk, D, Gable, AL, Lyon, D, Junge, A, Wyder, S, et al. (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47: D607-D613.
21. Pachkov, M, Balwierz, PJ, Arnold, P, Ozonov, E, van Nimwegen, E (2013) SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Res 41: D214-220.
22. Eckweiler, D, Dudek, CA, Hartlich, J, Brotje, D, Jahn, D (2018) PRODORIC2: the bacterial gene regulation database in 2018. Nucleic Acids Res 46: D320-D326.
23. Liu, ZP, Wu, C, Miao, H, Wu, H (2015) RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database (Oxford) 2015.
24. Greene, CS, Troyanskaya, OG (2010) Integrative systems biology for data-driven knowledge discovery. Semin Nephrol 30: 443-454.
25. Antiqueira, L, Janga, SC, Costa Lda, F (2012) Extensive cross-talk and global regulators identified from an analysis of the integrated transcriptional and signaling network in Escherichia coli. Mol Biosyst 8: 3028-3035.
26. Covert, MW, Xiao, N, Chen, TJ, Karr, JR (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. Bioinformatics 24: 2044-2050.
27. Wang, YC, Chen, BS (2010) Integrated cellular network of transcription regulations and protein-protein interactions. BMC Syst Biol 4: 20.
28. Marbach, D, Costello, JC, Kuffner, R, Vega, NM, Prill, RJ, et al. (2012) Wisdom of crowds for robust gene network inference. Nat Methods 9: 796-804.
29. Tompa, M, Li, N, Bailey, TL, Church, GM, De Moor, B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137-144.
30. Jayaram, N, Usvyat, D, AC, RM (2016) Evaluating tools for transcription factor binding site prediction. BMC Bioinformatics.
31. Saelens, W, Cannoodt, R, Saeys, Y (2018) A comprehensive evaluation of module detection methods for gene expression data. Nat Commun 9: 1090.
32. Simeonidis, E, Chandrasekaran, S, Price, ND (2013) A guide to integrating transcriptional regulatory and metabolic networks using PROM (probabilistic regulation of metabolism). Methods Mol Biol 985: 103-112.
33. Chandrasekaran, S, Price, ND (2010) Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 107: 17845-17850.
34. Banos, DT, Trebulle, P, Elati, M (2017) Integrating transcriptional activity in genome-scale models of metabolism. BMC Syst Biol 11: 134.
35. Zepeda, H, Considine, RV, Smith, HL, Sherwin, JR, Ohishi, I, et al. (1988) Actions of the Clostridium botulinum binary toxin on the structure and function of Y-1 adrenal cells. J Pharmacol Exp Ther 246: 1183-1189.
36. Koduru, L, Lakshmanan, M, Lee, DY (2018) In silico model-guided identification of transcriptional regulator targets for efficient strain design. Microb Cell Fact 17: 167.
37. Ibraim, IC, Parise, MTD, Parise, D, Sfeir, MZT, de Paula Castro, TL, et al. (2019) Transcriptome profile of Corynebacterium pseudotuberculosis in response to iron limitation. BMC Genomics 20: 663.
38. Price, MN, Dehal, PS, Arkin, AP (2007) Orthologous transcription factors in bacteria have different functions and regulate different genes. PLoS Comput Biol 3: 1739-1750.

39. Freyre-Gonzalez, JA, Manjarrez-Casas, AM, Merino, E, Martinez-Nunez, M, Perez-Rueda, E, et al. (2013) Lessons from the modular organization of the transcriptional regulatory network of Bacillus subtilis. BMC Syst Biol 7: 127.
40. Morrison, MD, Fajardo-Cavazos, P, Nicholson, WL (2019) Comparison of Bacillus subtilis transcriptome profiles from two separate missions to the International Space Station. NPJ Microgravity 5: 1.
41. Han, JD, Dupuy, D, Bertin, N, Cusick, ME, Vidal, M (2005) Effect of sampling on topology predictions of protein-protein interaction networks. Nat Biotechnol 23: 839-844.
42. Sierro, N, Makita, Y, de Hoon, M, Nakai, K (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Res 36: D93-96.
43. Arrieta-Ortiz, ML, Hafemeister, C, Bate, AR, Chu, T, Greenfield, A, et al. (2015) An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. Mol Syst Biol 11: 839.
44. Gama-Castro, S, Jimenez-Jacinto, V, Peralta-Gil, M, Santos-Zavaleta, A, Penaloza-Spinola, MI, et al. (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res 36: D120-124.
45. Weiss, V, Medina-Rivera, A, Huerta, AM, Santos-Zavaleta, A, Salgado, H, et al. (2013) Evidence classification of high-throughput protocols and confidence integration in RegulonDB. Database (Oxford) 2013: bas059.
46. Santos-Zavaleta, A, Salgado, H, Gama-Castro, S, Sanchez-Perez, M, Gomez-Romero, L, et al. (2019) RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. Nucleic Acids Res 47: D212-D220.
47. May, RM (1972) Will a large complex system be stable? Nature 238: 413-414.
48. Zulianello, L, de la Gorgue de Rosny, E, van Ulsen, P, van de Putte, P, Goosen, N (1994) The HimA and HimD subunits of integration host factor can specifically bind to DNA as homodimers. EMBO J 13: 1534-1540.
49. Gotfredsen, M, Gerdes, K (1998) The Escherichia coli relBE genes belong to a new toxin-antitoxin gene family. Mol Microbiol 29: 1065-1076.
50. Brinkrolf, K, Brune, I, Tauch, A (2006) Transcriptional regulation of catabolic pathways for aromatic compounds in Corynebacterium glutamicum. Genet Mol Res 5: 773-789.
51. Chen, C, Zhang, Y, Xu, L, Zhu, K, Feng, Y, et al. (2018) Transcriptional control of the phenol hydroxylase gene phe of Corynebacterium glutamicum by the AraC-type regulator PheR. Microbiol Res 209: 14-20.
52. Henke, NA, Heider, SAE, Hannibal, S, Wendisch, VF, Peters-Wendisch, P (2017) Isoprenoid Pyrophosphate-Dependent Transcriptional Regulation of Carotenogenesis in Corynebacterium glutamicum. Front Microbiol 8: 633.
53. Kleine, B, Chattopadhyay, A, Polen, T, Pinto, D, Mascher, T, et al. (2017) The three-component system EsrISR regulates a cell envelope stress response in Corynebacterium glutamicum. Mol Microbiol 106: 719-741.
54. Bussmann, M, Baumgart, M, Bott, M (2010) RosR (Cg1324), a hydrogen peroxide-sensitive MarR-type transcriptional regulator of Corynebacterium glutamicum. J Biol Chem 285: 29305-29318.
55. Toyoda, K, Inui, M (2018) Extracytoplasmic function sigma factor sigma(D) confers resistance to environmental stress by enhancing mycolate synthesis and modifying peptidoglycan structures in Corynebacterium glutamicum. Mol Microbiol 107: 312-329.

56. Lee, DS, Kim, P, Kim, ES, Kim, Y, Lee, HS (2018) Corynebacterium glutamicum WhcD interacts with WhiA to exert a regulatory effect on cell division genes. Antonie Van Leeuwenhoek 111: 641-648.
57. Knizewski, L, Ginalski, K (2007) Bacterial DUF199/COG1481 proteins including sporulation regulator WhiA are distant homologs of LAGLIDADG homing endonucleases that retained only DNA binding. Cell Cycle 6: 1666-1670.
58. Chauhan, R, Ravi, J, Datta, P, Chen, T, Schnappinger, D, et al. (2016) Reconstruction and topological characterization of the sigma factor regulatory network of Mycobacterium tuberculosis. Nat Commun 7: 11062.
59. Salgado, H, Peralta-Gil, M, Gama-Castro, S, Santos-Zavaleta, A, Muniz-Rascado, L, et al. (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res 41: D203-213.

Figures

Figure 1. Abasy Atlas content. (**A**) Completeness measured as genomic and interaction coverage for the GRNs in Abasy, 76 networks covering 42 bacteria distributed in 9 species. (**B**) Examples describing the format of the Abasy identifiers. The most complete *C. glutamicum* GRN (upper) filtered to contain only "strong" interactions, and the most recent, meta-curated *E. coli* GRN (lower).

Figure 2. Historical snapshots for GRNs of model organisms. The completeness of the network can be measured as genomic coverage (fraction of the genome included in the GRN, black triangles) and interaction coverage (fraction of the known interactions relative to the complete network, red circles). It is evident that for some networks genomic coverage overestimates completeness as some networks may be classified as almost completed in terms of genomic coverage whereas many interactions are still missing. For instance, the GRN for *C. glutamicum* in 2016 is a meta-curation of the network from 2011 and a set of interactions curated in [13] including the *sigA* housekeeping sigmulon. On the other hand, the GRN for *M. tuberculosis* in 2016 is the most complete in terms of interaction coverage (97.7%) since it integrates the network from 2015 with novel interactions curated from the literature.

Figure 3. Complementary sources to reconstruct the meta-curated GRN for *B. subtilis*. A poor overlap is observed between the different sources used to reconstruct the meta-curated GRN for *B. subtilis*, mainly for interactions. This highlights the need for the meta-curation since the organism-specific databases do not fully cover each other nor the dataset not previously hosted in any database. Abasy provides homogeneous meta-curations integrating all the available information.

Figure 4. (**A**) Number of interactions identified by methods described as "weak" in [3] and how many of these interactions have been validated by "strong" evidence. IGI (inferred from genetic interaction), TAS (traceable author statement), TASES (traceable author statement to experimental support), NTAS (non-traceable author statement), IC (inferred by curator), IHBCE (inferred by a human based on computational evidence), RFP (RNA-polymerase footprinting), ICA (inferred by computational analysis), IEP (inferred from expression pattern), IMP (inferred from mutant phenotype), BCE (binding of cellular extracts), AIPP (automated inference of promoter

position), HIPP (human inference of promoter position), AIBSCS (automated inference based on similarity to consensus sequences), ICWHO (inferred computationally without human oversight), HIBSCS (human inference based on similarity to consensus sequences), GEA (gene expression analysis) [59]. (**B**) Effect of removing spurious interactions through the meta-curation process. System-level elements (global regulators, modular, intermodular, and basal-machinery genes) values represent its fraction from the total genes in the *E. coli* GRN historical snapshots before and after removal of interactions supported only by the ICWHO evidence.

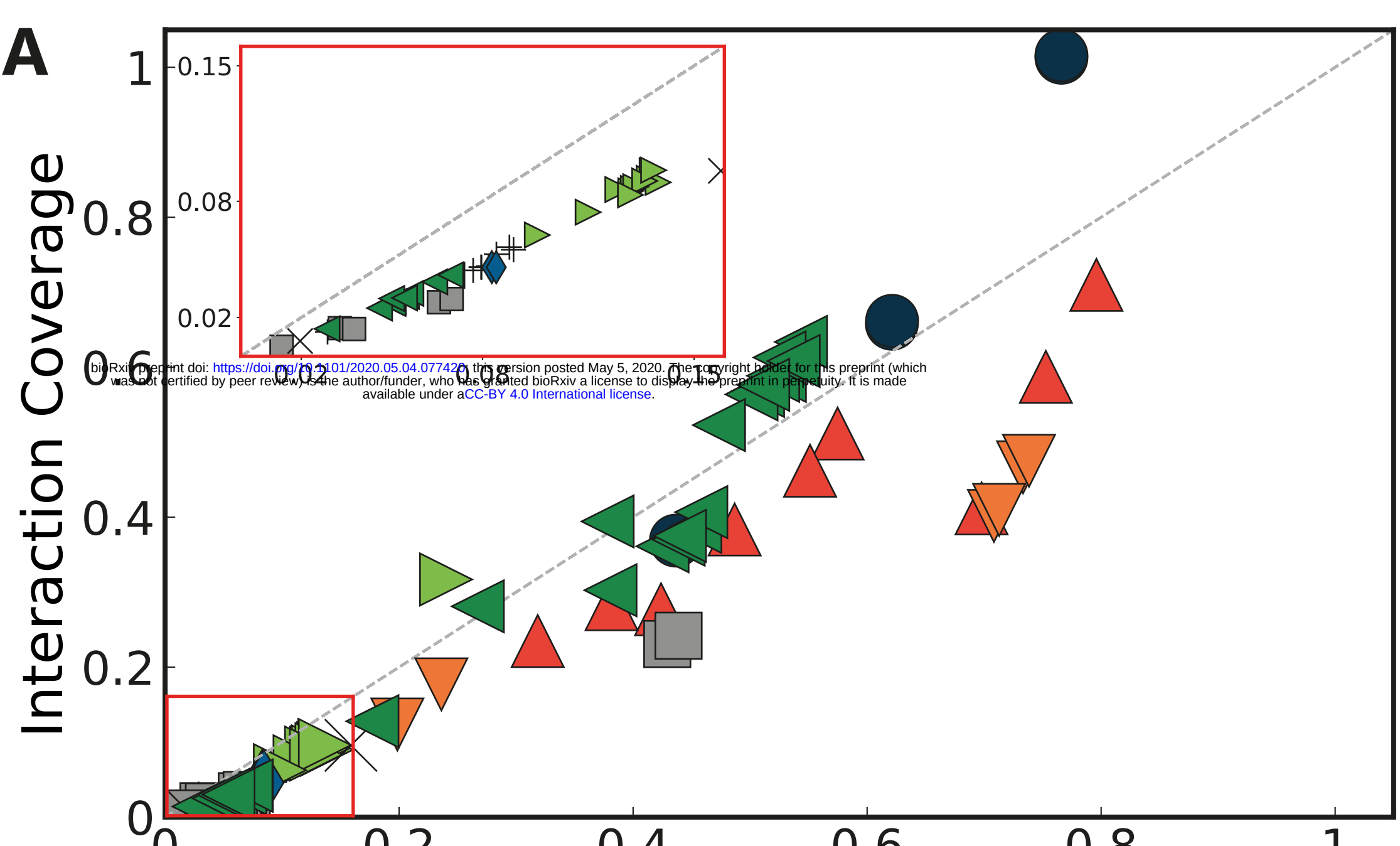
Figure 5**.** The constrained complexity of regulatory networks allows computing the total number of interactions. The number of interactions in the Abasy GRNs follows a power law with the number of genes as $I \sim n^{\gamma}$ with $\gamma = 2 - \alpha$ ($R^2 = 0.98$). This power law may be used to compute the total number of interactions (I_{total}) in the regulatory network of an organism as $I_{total} \sim (\text{genome size})^{2-\alpha}$.

Figure 6**.** Homogeneous network representation. Heteromeric-complex-base gene representation for IHF (**A**) and RelBE (**C**). Misrepresentation of gene expression regulation where heteromeric protein complexes are involved for IHF (**A**) and RelBE (**D**) systems. RelB can regulate itself as a homomeric-complex, and as a heteromeric-complex with *relE* (**C**). Besides, *relE* can regulate neither its transcription nor RelB transcription on its own, as could be misinterpreted from (**E**). This same misrepresentation is observed for the IHF complex where neither of the subunits has regulatory activity as a homomeric complex.

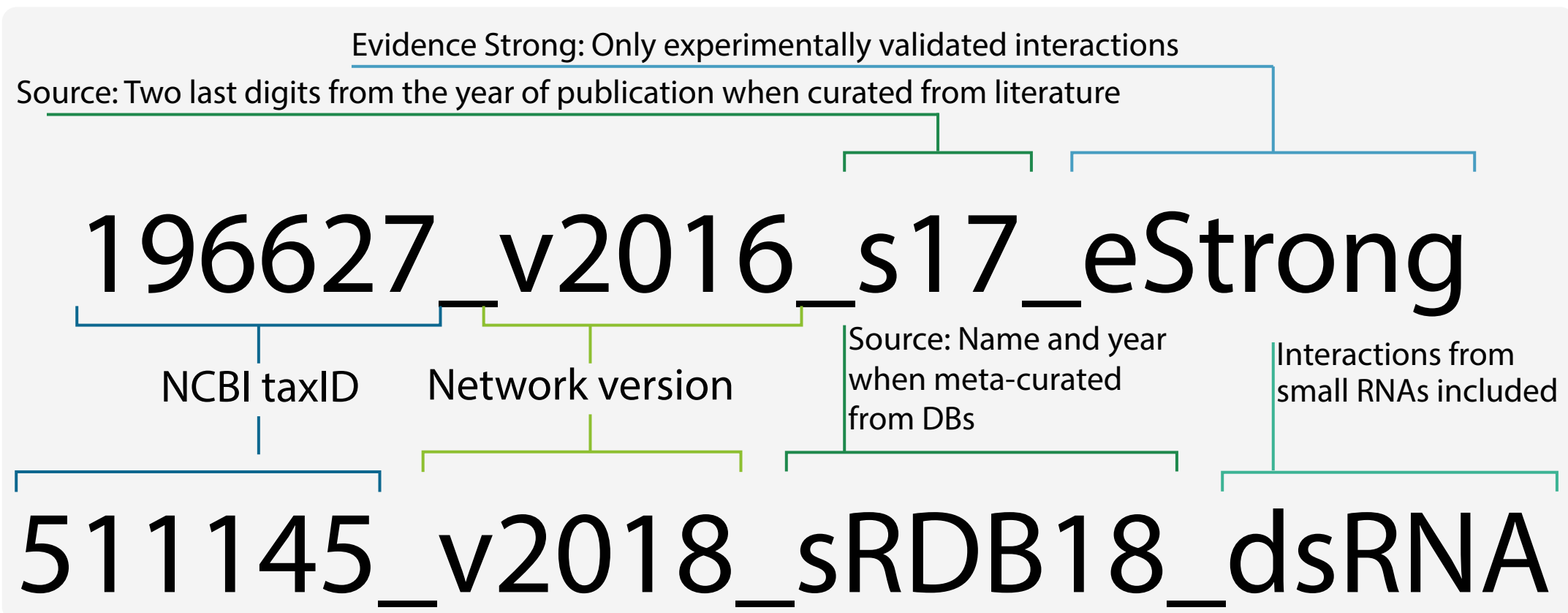
Supplementary data

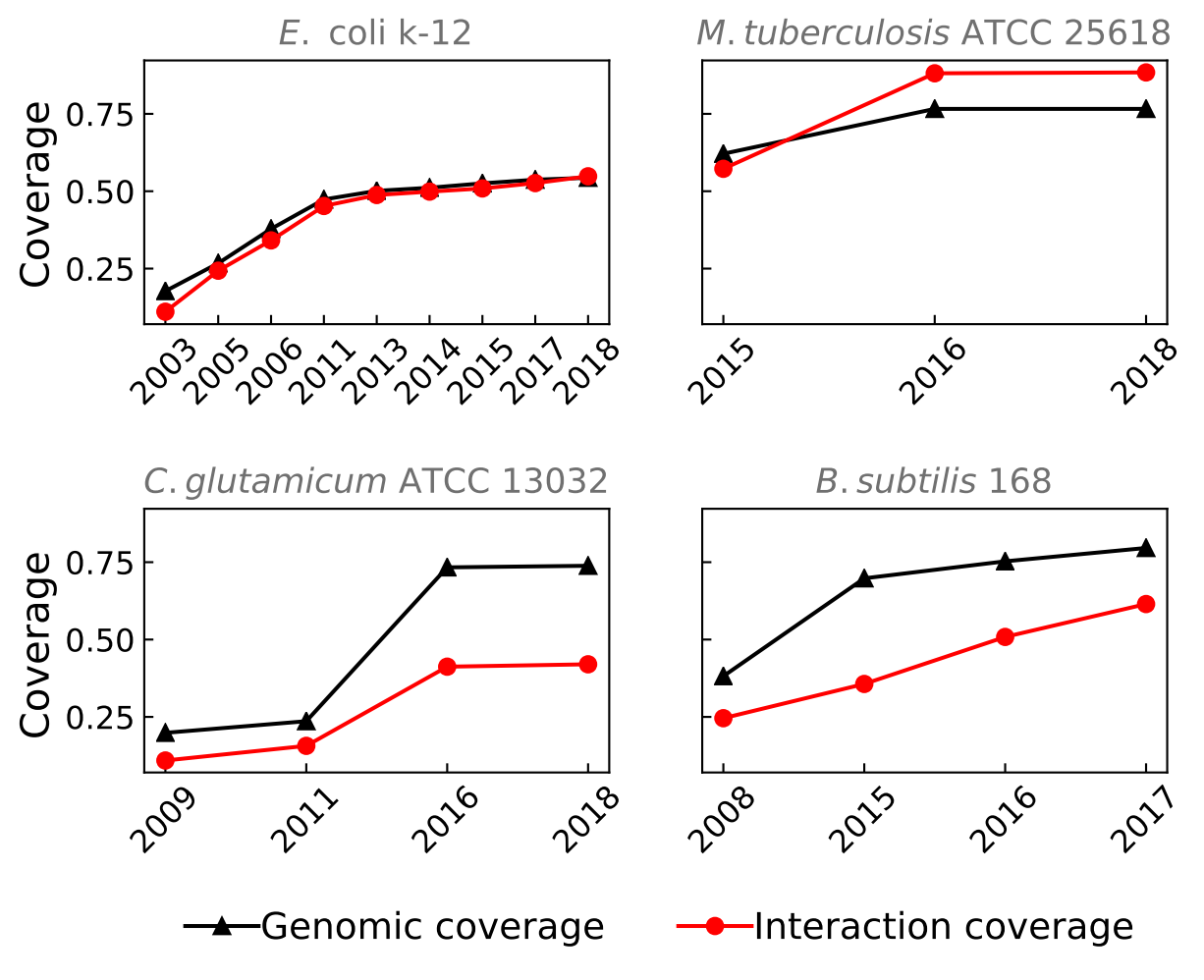
Supplementary information is available online.

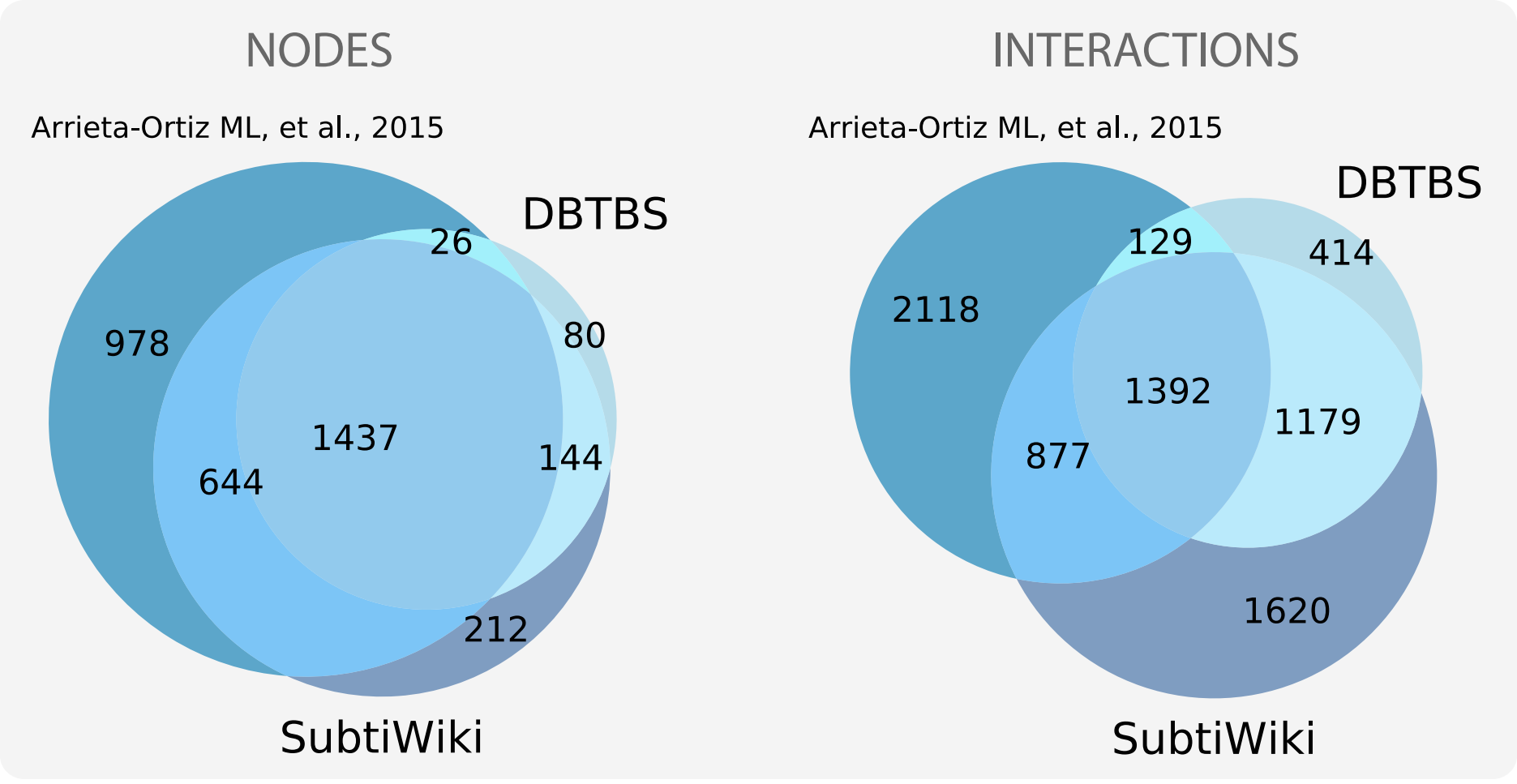
A

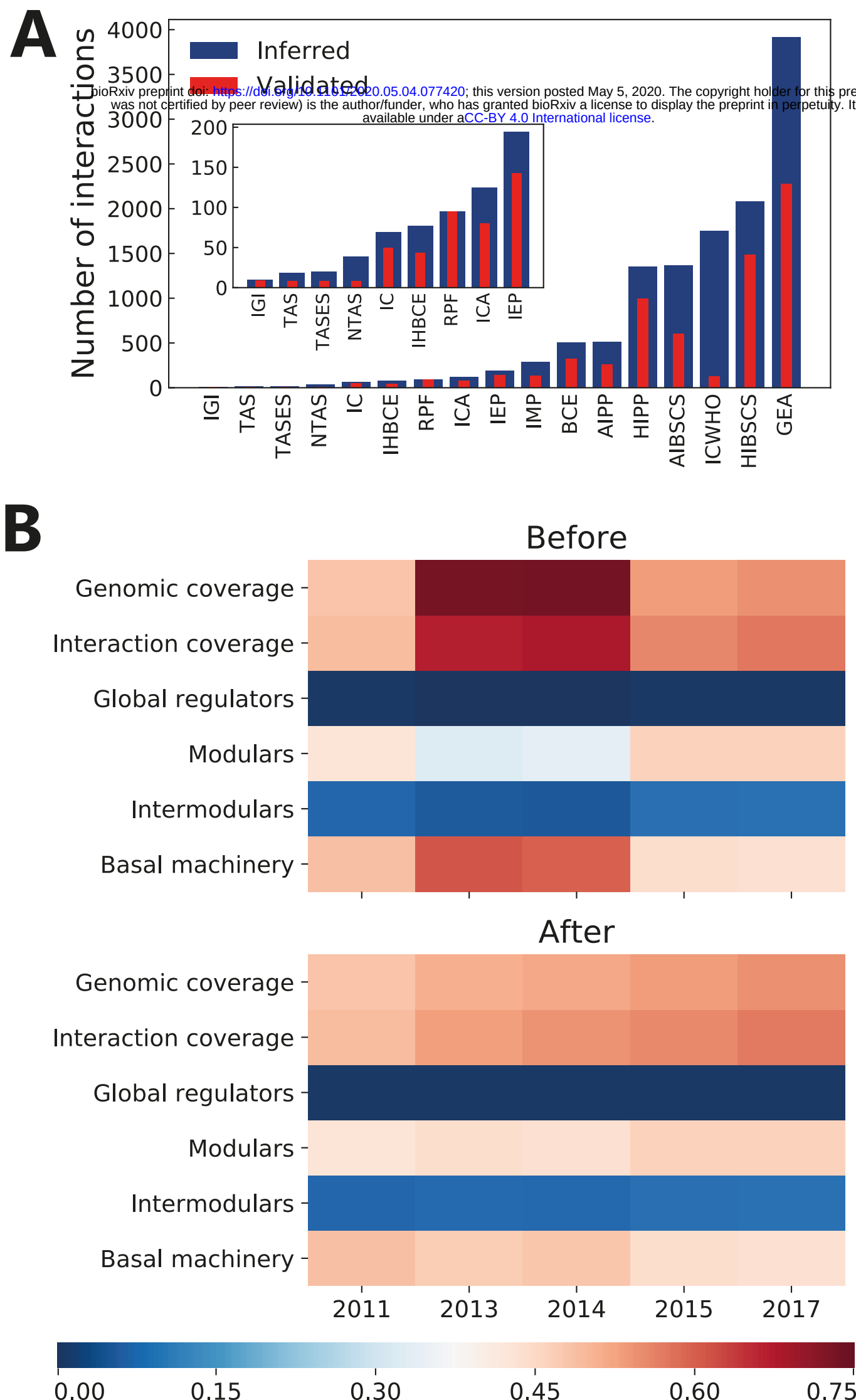


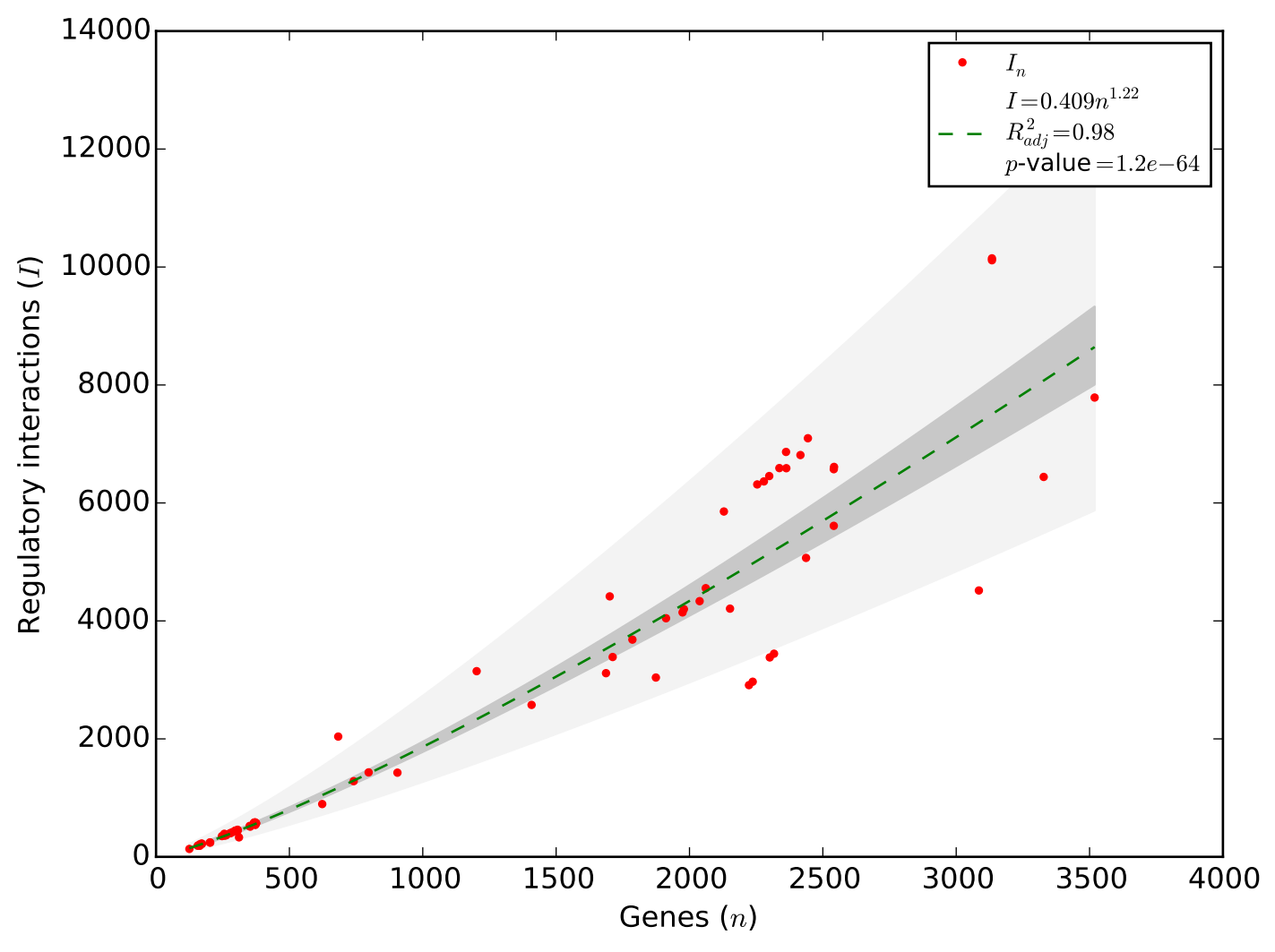
B

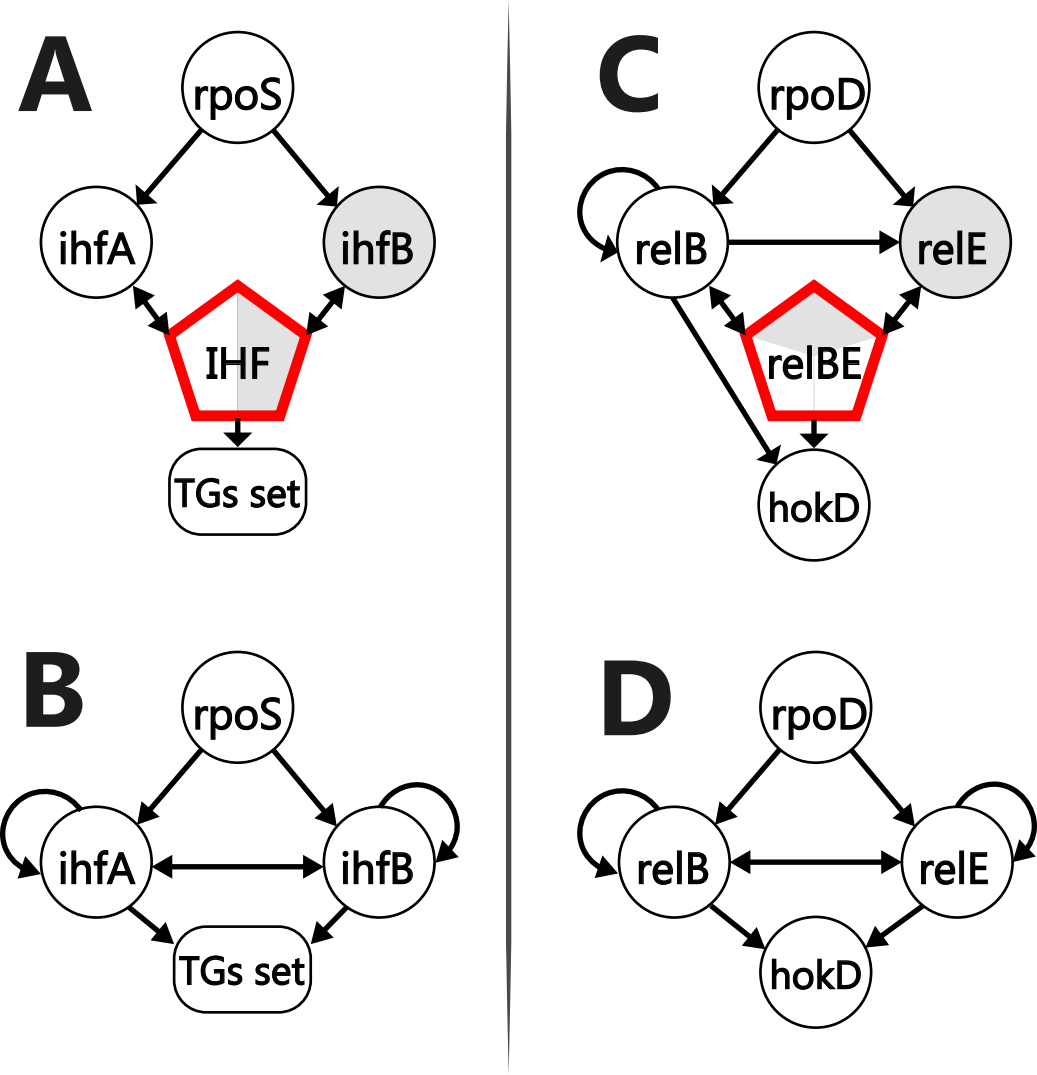



E. coli k-12
M. tuberculosis ATCC 25618
C. glutamicum ATCC 13032
B. subtilis 168
Coverage
0.75
0.50
0.25
2003
2005
2006
2011
2013
2014
2015
2017
2018
2015
2016
2018
2009
2011
2016
2018
2008
2015
2016
2017
Genomic coverage
Interaction coverage


NODES
Arrieta-Ortiz ML, et al., 2015
DBTBS
26
978
80
1437
144
644
212
SubtiWiki
INTERACTIONS
Arrieta-Ortiz ML, et al., 2015
DBTBS
129
414
2118
1392
1179
877
1620
SubtiWiki




I_n
$I = 0.409n^{1.22}$
$R^2_{adj} = 0.98$
p-value $= 1.2e{-}64$
Regulatory interactions (I)
Genes (n)
0
2000
4000
6000
8000
10000
12000
14000
0
500
1000
1500
2000
2500
3000
3500
4000


A
rpoS
ihfA
ihfB
IHF
TGs set
C
rpoD
relB
relE
relBE
hokD
B
rpoS
ihfA
ihfB
TGs set
D
rpoD
relB
relE
hokD