A Golden Standard for Transcription Factor-Gene Regulatory Interactions in *Escherichia coli* K-12: How reliable are they based on the methods supporting them?

Paloma Lara[1]*, Socorro Gama-Castro[1]*, Heladia Salgado[1]*, Claire Rioualen[1], Luis J. Muñiz-Rascado[1], Jair S. García-Sotelo[4], Víctor H. Tierrafría[1,2] and Julio Collado-Vides[1,2,3]**

*co-first authors
** corresponding: colladojulio@gmail.com

[1]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Avenida Universidad s/n, Cuernavaca 62210, Morelos, Mexico
[2]Department of Biomedical Engineering, Boston University, 44 Cummington Mall, Boston, MA 02215, USA
[3]Center for Genomic Regulation. The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Universitat Pompeu Fabra(UPF), Barcelona, Spain.
[4]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Querétaro 76230, Querétaro, Mexico.

keywords: regulatory interactions; *E.coli*; evidence codes; high throughput genomic methodologies

ABSTRACT

Genomics has strongly increased the available experimental strategies to identify the elements involved in regulatory networks, particularly those of the regulation of transcription initiation. As new methodologies emerge, a natural step is to compare their results with available knowledge obtained by previously established methodologies, such as the classic methods of molecular biology used to characterize transcription factor binding and regulatory sites, promoters, and transcription units.
Such previous knowledge is dispersed in the scientific literature, limiting their accessibility. Fortunately, in the case of *Escherichia coli* K-12, the best studied microorganism, we have been continuously gathering this knowledge from original scientific publications for the last 30 years , and have made it available to the public in two databases, RegulonDB and EcoCyc. More recently, we have also gathered knowledge generated by genomic high throughput (HT) methodologies as can be appreciated in the latest RegulonDB version 11.0, where, in addition to the objects and interactions identified by methods of molecular biology, users can find datasets of binding sites from ChIP-seq, ChIP-exo, gSELEX and DAP-seq HT technologies. This has motivated us to work on improving our evidence codes to facilitate their use as golden standards. In order to enhance the precision of knowledge representation, we implemented three alternatives to curate regulatory interactions, based on the level of detail provided by the experiments. We present them in a way that users can select different versions of objects based on properties such as the methods used to identify them (classic, computational predictions, HT methods); *in vivo* vs *in vitro* approaches; and degree of confidence (weak, strong, confirmed) of the evidence supporting them. The collection of regulatory interactions is analyzed based on their supporting evidence,

showing how classic methods still provide for the largest fraction of strong and confirmed sites, a distribution likely to change in the years to come as HT methods may become the dominant strategies to identify regulatory genomic components. We plan to keep updating and expanding these golden standard datasets as part of future RegulonDB releases.

I.- INTRODUCTION

Genomic sciences have strongly affected the landscape of available experimental strategies to identify a variety of collections such as transcription factor binding sites (TFBSs) and their subset of transcription factor regulatory sites (TFRSs) -those TFBSs with regulatory evidence- for a given transcription factor (TF); transcription start sites (TSSs), transcription termination sites (TTSs) as well as transcription units (TUs), all of these in principle under defined growth conditions. A major concern in our curation planning was how to deal with what we saw as a *tsunami* of data coming from post-genomic emerging high throughput (HT) methodologies, and how not to inundate and dilute the decades of previous work reflected in the corpus of knowledge supported by what we call classic molecular biology methods. These methods might not be scalable but are well appreciated since, as it is well known, they identify individual elements directly.

As mentioned before, we have been for the last 30 years, continuously extracting and gathering from original scientific publications in *Escherichia coli* K-12, knowledge about regulation of transcription initiation and operon organization in RegulonDB and EcoCyc. We have more recently also gathered knowledge generated by genomic HT methodologies as can be appreciated in the latest RegulonDB version 11.0 that, in addition to the collection from classic methods of molecular biology, also includes datasets of binding sites from ChIP-seq, ChIP-exo, gSELEX and DAP-seq technologies (1). In our curation work we have seen that this type of projects frequently compare their results with what is known in RegulonDB (2-17)

RegulonDB and EcoCyc accelerate access to knowledge. An example is their use in order to find the original publications supporting a specific object (for instance, a promoter, a regulatory site, a transcription factor). However, some objects have different properties that require different methods to be identified. For instance, locating the transcription start site involves methods different from those identifying the type of sigma factor of the promoter. Similarly, well characterized regulatory interactions (RIs) require support of the binding of the TF to a specific site in the genome on the one hand, as well as identifying the function of such TF-site in the activation or repression of the regulated promoter. All references, historically, were grouped for each object. We thought it would be more useful to separate, when possible, the evidence types and the corresponding references in complex objects. Thus, we have relatively recently separated the evidence and references supporting the binding of a TF to a site, from the evidence supporting the function of the regulatory interaction.

Furthermore, evidence codes used for years both in RegulonDB and EcoCyc, were not detailed enough to distinguish different methods. For instance: “Binding of purified proteins” or “gene expression analysis” do not provide any further detail.

Briefly, the need to easily distinguish objects based on the approach used (i.e. classic vs HT methods), the fact that RegulonDB sites are used as an index to evaluate the performance of novel methods, and the desire to improve the precision in literature access to specific properties of complex objects such as regulatory interactions (RIs) or promoters, motivated us to update the evidence codes behind the knowledge we gather on the regulation of transcription initiation. The new codes distinguish not only the class of methods, but also the specific methodology, distinguishing for instance ChIP-seq from ChIp-exo or gSELEX. We have been moving in this direction since a few years ago, but it is only in this paper that we report the rationale and new datasets available.
Once the new evidence types were defined, we reassessed the way they combine to determine what we call "confidence level", which, based on the whole collection of evidence behind an object, assigns it as either weak, strong or confirmed. Here we analyze the contribution of different sources (i.e. classic, HT and/or computational methods) in the confidence level of the currently available collection of RIs.

II.- RESULTS

A regulatory interaction is one of the major concepts together with transcription units and operons, that describes the knowledge of regulation of transcription initiation. As any piece of knowledge, it can be described at different levels of detail; at a low level we can say it is the triplet formed by the TF, the target gene and the positive or negative effect. The whole collection of currently known regulatory interactions at this high level is available as a downloadable file in http://regulondb.ccg.unam.mx/menu/download/datasets/index.jsp in the network interactions file "TF-gene interactions". At a high level of detail knowledge of regulatory interactions involves the TF, the effector that affects its binding and unbinding conformation, the precise TF regulatory binding site, the regulated promoter, the effect of the TF bound, either activating or repressing transcription initiation. It may also be expanded and include knowledge of the transcription unit(s) that the promoter transcribes and therefore the set of regulated genes, and finally the growing conditions (experimental and control) where such regulation has been identified. These fundamental concepts started to see their origin in the ´60s of last century in the work of Jacob and Monod and several other investigators that participated in the emergence of molecular biology in microbial organisms (18-20). The concepts used in our modeling of such knowledge both in RegulonDB and EcoCyc follow the updated definitions of these terms, given the expanded current knowledge of regulation of transcription initiation and operon organization as proposed in (21).

II.1- New evidence codes and their confidence level.

As mentioned before, we want to explicitly distinguish classic vs HT methods, and increase their precision to match with specific methods. This contrasts with previous evidence codes that we used in the early days of our curation work, that were not specific enough.

We updated our table of evidence types, and have in fact modified their description to explicitly include whether they are experimental methods, either HT or classic methods, or non-experimental such as computational predictions or author statements. We modified

the names of evidence codes to make them more informative. Since some objects are rather complex, particularly the regulatory interactions, we have separated the evidence for binding within sites, and the evidence for function within the RI itself. This also makes it easier for users to fish for specific references whenever they come from different publications. Each evidence type is associated with a specific code, that we built keeping it as short as possible but informative, and with prefixes indicating if it is an HT method. We added a link in RegulonDB that offers the name, description, evidence code, and confidence level (see below) of all evidence types, as well as whether they correspond to *in vitro* or *in vivo* binding experiments. **See:** **https://regulondb.ccg.unam.mx/evidenceclassification.** Although this table shows updated codes for regulatory interactions promoters and TUs, we focus in this paper only on the work around RIs. As can be seen, the new evidence types added are essentially those that support HT methods. In addition to the classic, HT and non-experimental classification, we also classified evidence types into *in vivo* vs *in vitro* binding.

Years ago we classified in RegulonDB the different evidence types into either weak or strong, depending on the confidence that the methods provide to support the existence of a piece of knowledge. The general principle is that strong confidence comes from experiments that provide clear physical evidence of the existence of the object. For instance, binding of purified proteins in the case of a given TF binding to its binding site is considered strong evidence; whereas binding of cellular extracts is considered weak evidence. A limitation to this initial approach was that even if some objects are identified either in the same paper, or through the years in more than one publication, by different methods, we did not have a process to add multiple weak evidence types and consider it a strongly supported object. This is contrary to a fundamental strategy in natural science, that precisely gives further support to knowledge that is reached from different, and ideally, independent, strategies or methods. Thus we embarked on a detailed analysis of which of the different methods behind the knowledge gathered in RegulonDB can be considered independent, because they use different assumptions, and/or different methodological strategies such that their potential sources of error are different (22). It is on the basis of these theoretical considerations that we built our algebra to combine multiple weak independent sources of methods into a strong confidence level. We also proposed the combination of independent strong evidence types to give birth to the new "confirmed" level of confidence.

In the current update we kept the same principles and criteria as defined in the 2013 paper and updated the three levels of confidence to all existing current evidence types. In our recent addition of HT-evidence codes, after discussion by the team in RegulonDB we decided to consider all HT-based methods for TF-binding identification as weak experimental evidence. This is due to the fact that these methods involve multiple steps of computational processing of the raw data, and these multiple steps can be performed by a variety of computational methods, each involving setting specific parameters and thresholds. Thus, processing the same raw data could potentially result in different final collections of objects, in addition, there is no consensus yet on a uniform processing pipeline used by the community. As a result, our current types of evidence, their classification in groups, and their level of confidence are summarized in Figure 1.

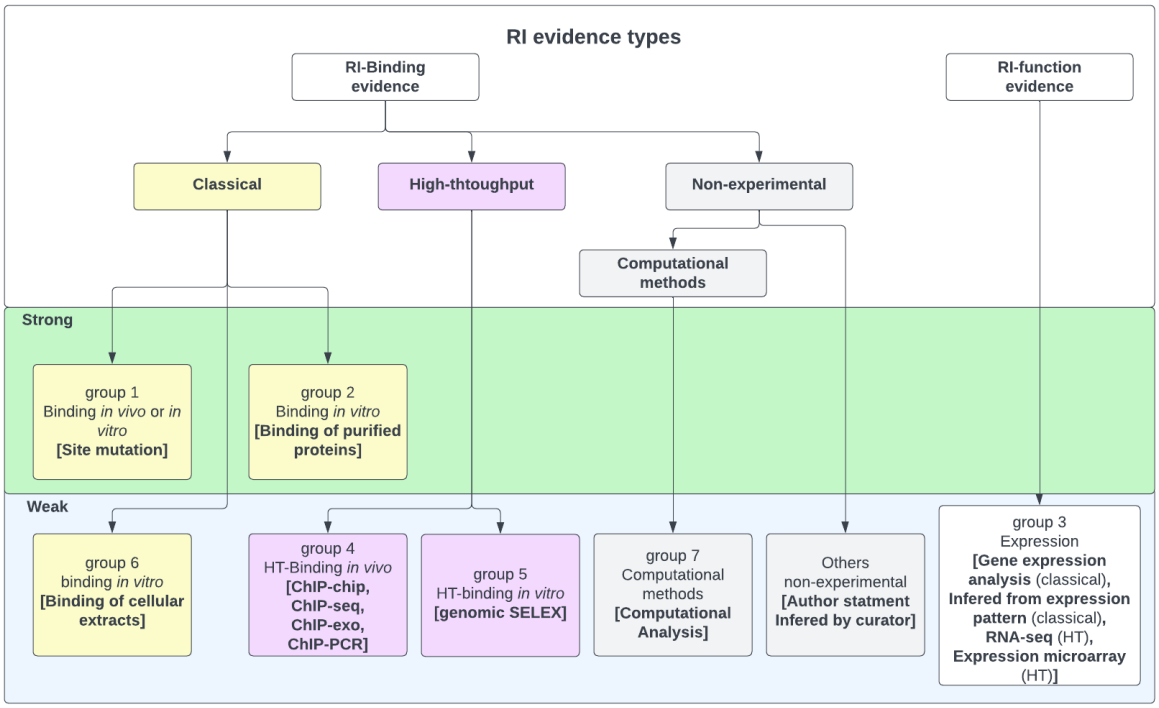

Figure 1. Evidence types, categories and methodological approaches.

In practice, we assigned each evidence type to one of 7 possible groups, and defined the combinations that upgrade the object confidence level using those group numbers (Figure 1). Evidence types in the same group are considered to share methodological bias and can not be combined to upgrade the confidence level of the associated object, while evidence types between different groups are considered independent, and their combination can upgrade the object confidence level. Currently, only the evidence types of the groups 1 and 2 are classified as strong, and the groups 4, 5, 6 and 7 contain weak evidence types, and the group of "others" do not contribute to confidence (Figure 1). The four general rules are:

1) Two independent binding-evidence types with confidence level weak (groups 4 to 7), and with a function-evidence type (group 3) upgrade the object confidence level to strong.
2) Two independent binding-evidence types with confidence level strong (groups 1 or 2) upgrade the object confidence level to confirmed.
3) Two independent binding-evidence types with confidence level weak (groups 4 to 7), and with a function-evidence type (group 3) in addition to a strong evidence type (group 1 or 2) upgrades the object confidence level to confirmed.
4) Four independent binding evidence types with a weak confidence level (groups 4 to 7), and with a function-evidence type (group 3) upgrades the object confidence level to confirmed.

Gene expression evidence is itself a prerequisite for defining a RI, however, several years ago, expression evidence types were not annotated. As a consequence, two independent evidence weak types can upgrade the object evidence to strong only when evidence of expression is not missing in RegulonDB.

Note in fact, that binding of purified protein and site mutation are currently the only evidence types with confidence level strong. There is no single evidence that supports the level of "confirmed". Site mutation is classified as a strong evidence type since it involves the identification of the sequence regulatory site and the ascertainment of that their alteration have an effect in the transcription, either *in vivo* (23,24), through a reporter gene or by *in vitro* transcription in the presence or absence of a determined TF (25). Binding of purified protein includes two similar methodologies: EMSA and footprinting, in which the TF binding to a specific sequence target is probed *in vitro*. Note that currently only HT methods are not sufficient to provide a confirmed confidence level, since there are only two independent HT groups.

The complete set of evidence types combinations that upgrades RIs confidence level can be found under the "regulatory interactions" of the "Stage II" section of the webpage : https://regulondb.ccg.unam.mx/evidenceclassification. For instance, ChIP-chip, ChIP-seq and ChIP-exo belong to group 4, whereas gSELEX belongs to group 5. The rule (4 / 5 / 3)-S means that any group 4 evidence plus any group 5 evidence, and any group 3 together upgrade two weak binding evidence types into a Strong confidence level.

II.3. Three representations of regulatory interactions: TF-promoter, TF-TU and TF-gene.

In addition to using updated concepts of gene regulation, a challenge we have been addressing when searching for the best possible way to represent knowledge is the need for intelligent ways to deal with partial knowledge. It is not uncommon for a curator to have to choose the least costly assumption when knowledge is lacking. For instance, years ago, since by definition a TU has a promoter, we added the so-called "phantom promoters" to those TUs that had no promoter. This was later eliminated to avoid confusion by users as suggested by Rick Gourse in an EcoCyc meeting. Another example is how to deal with the curation of RIs. Historically, we curated RIs affecting a given promoter, even when there was no such specific evidence. The curator uploaded the RI when the target gene had only one promoter, and if the target gene had two or more promoters, the new RI was mentioned in notes of the TU. It is important to note that our curation work has been evolving for more than 30 years now. In this long period we have added new objects, new features, and have improved our evidence codes, essentially improving the quality of knowledge representation as identified in different strategies and methods.

In order to ideally eliminate, or to minimize the need of additional assumptions made in our curation process, and thus improve knowledge representation, we defined three levels of description of regulatory interactions, which we use depending on the level of detail of knowledge available. We call these "RI types":

(1) The most precise knowledge is when there is evidence that identifies the regulated promoter affected by an RI. Most of these come from classic experiments. In these cases it is reasonable to deduce that all TUs associated with the regulated promoter are regulated by the new RI. We call these TF-promoter RIs.

(2) A less detailed description is when the regulated promoter is not known and there is evidence of a change in expression of a group of adjacent genes on the same strand, that matches with an existing TU with or without promoter. In such cases, we associate the new

RI to the existing TU. We call these TF-TU RIs. If it does not exist, we create a new TU without a promoter and with evidence of co-expression, and add to it the new RI.
(3) Finally, when the regulated promoter has not been identified, and there is evidence of differentially-regulated transcription of the downstream gene(s) from a TF binding site, we create a new RI for which the target is the gene. We call these TF-gene RIs.
As a result, we currently have three means of adding RIs depending on available knowledge in RegulonDB and EcoCyc: TF-promoter, TF-TU and TF-gene.

The curation of knowledge related to regulatory interactions exerted by a TF depends on several rules. The easy case is when there is not a previously annotated RI with the same TF and gene; in this case, a new RI at the adequate level is uploaded, according to the available knowledge. However, if there is a previous RI, of any of the three types, and the new and previous knowledge match, the new evidence is added to the existing RI. In order to avoid redundancy in the representation of these interactions among the three types, we follow the rules as explained to avoid creating duplicate objects. This is easy to say, but in practice, whether there are no redundant RIs will only be known when the missing pieces of information are identified. Furthermore, this also eliminates, or at least minimizes additional assumptions implied in our curation.

II. 4.- Anatomy of knowledge supporting the current collection of RIs.

Given the previous concepts we can now analyze in more detail the source of knowledge of the whole collection of RIs currently available in RegulonDB, what we call the anatomy of knowledge of regulatory interactions, which are the building blocks to construct the best known transcription regulatory network.

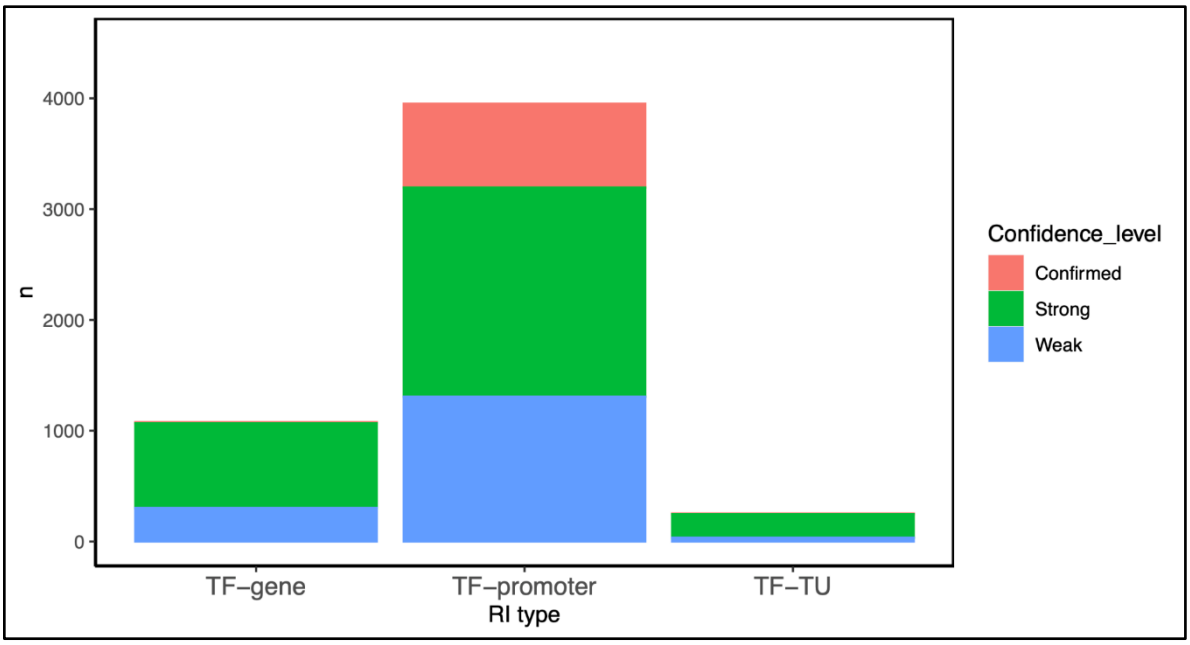

Figure 2. Number of RIs of each level of detail (also called “type” in the main text) and their confidence level highlighted in different colors.

The current version (11.1) of RegulonDB contains a total of 5,281 RIs, of which 769 (15%) have confirmed evidence, 2,865 (54%) strong and 1,646 (31%) have weak support. Combining the confirmed and strong levels includes almost 70% of all current RIs, a total of 3,634 RIs, which is the best available basis to compare with HT-based results.

Furthermore, currently, 75% of RIs are TF-promoter, 20% are TF-gene and only 5% are TF-TU, see Figure 2 and Table 1.

It is no surprise that the TF-promoter type is the one with the highest number of strong and confirmed levels of confidence, since, as mentioned before, these RIs come mostly from classic methodologies. The TF-gene type is mostly used to curate HT-supported RIs, as can be seen in Figure 3.

Table 1. Number of RIs grouped by type and by confidence level

RI type	Confidence level	# RIs	Total
TF-promoter	Confirmed	756 (19%)	3951 (100%)
	Strong	1888 (48%)	
	Weak	1307 (33%)	
TF-TU	Confirmed	5 (2%)	252 (100%)
	Strong	212 (84%)	
	Weak	35 (14%)	
TF-gene	Confirmed	8 (1%)	1077 (100%)
	Strong	765 (71%)	
	Weak	304 (28%)	

The third column shows the number of RIs of each subgroup, i.e. the first row corresponds to the number of TF-promoter RIs with confirmed confidence level. The percent value is relative to the total number of RIs of the corresponding type, indicated in the fourth column.

We analyzed the number of RIs with the different evidence categories: classical, HT or non-experimental evidence, among the three subsets of RIs grouped by type. As expected, the RIs of TF-gene and TF-TU types have HT evidence mainly, while the RIs with classical evidence are mainly of TF-promoter type (Figure 3, Table 2).
Furthermore, of 2,745 RIs with classical evidence, 242 (9%) have also HT evidence, that is, currently a small fraction of the classical RIs have been also determined by HT methodologies. Our analysis shows that HT methodologies have allowed the identification of new RIs, however since the promoter involved is not determined, most have been curated as TF-gene RIs (Figure 3, Table 2). However, it is relevant to say that to date, we

only have curated HT-RIs for which evidence for function and evidence of binding were reported in the same paper. An interesting alternative would be to extract from those HT-binding datasets, those that match with existing RIs in RegulonDB, since they already contain functional evidence. This will enrich the whole landscape of evidence for the existing RIs, but will not yet increase the number of RIs which is certainly a major goal of genomic research.

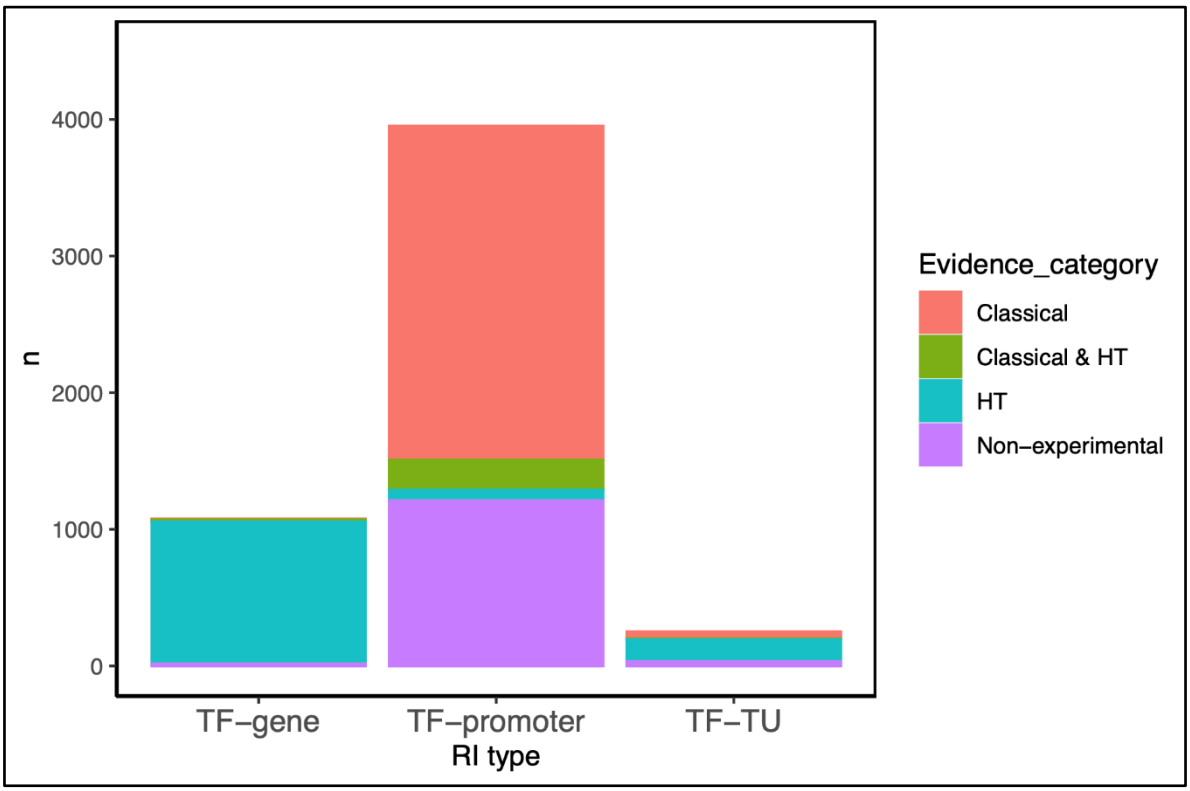

Figure 3. Number of RIs of type and their evidence category highlighted in different colors.

With the aim of better assessing the contribution of classical and HT evidence types to the confidence level of the current set of RIs, we plotted the number of RIs with the different confidence levels grouped by classical, HT, classical+HT (these three groups can or can not have computational evidence) and the RIs with only non-experimental evidence of binding (i. e. computational predictions) (Figure 4). We observe that all 769 RIs with a confirmed confidence level have classical evidence, and among those, 203 also have HT evidence. Of 2,865 RIs with a strong confidence level, 1,845 have classical evidence of binding, 982 have HT evidence and 38 have both (classical and HT). It is remarkable that among RIs with only HT experimental evidence, 77% (982) have a strong confidence level. As logic dictates, all RIs only with non-experimental evidence have a weak confidence level (Figure 4).

Table 2. Counts of RIs with classical, HT or non-experimental evidence and grouped by type.			
RI_type	**Evidence category**	**# Ris**	**Total**
TF-gene	Classical	6 (0.5%)	1077 (100%)
	HT	1040 (96.5%)	
	Classical & HT	16 (1.5%)	
	Non-experimental	15 (1.5%)	
TF-promoter	Classical	2444 (62%)	3952 (100%)
	HT	79 (2%)	
	Classical & HT	220 (5%)	
	Non-experimental	1209 (31%)	
TF-TU	Classical	53 (21%)	252 (100%)
	HT	161 (64%)	
	Classical & HT	6 (2%)	
	Non-experimental	32 (13%)	

The third column shows the number of RIs of each subgroup, i.e. the first row corresponds to the number of TF-gene RIs with classical evidence. The percent value is relative to the total number of RIs of the corresponding type, indicated in the fourth column.

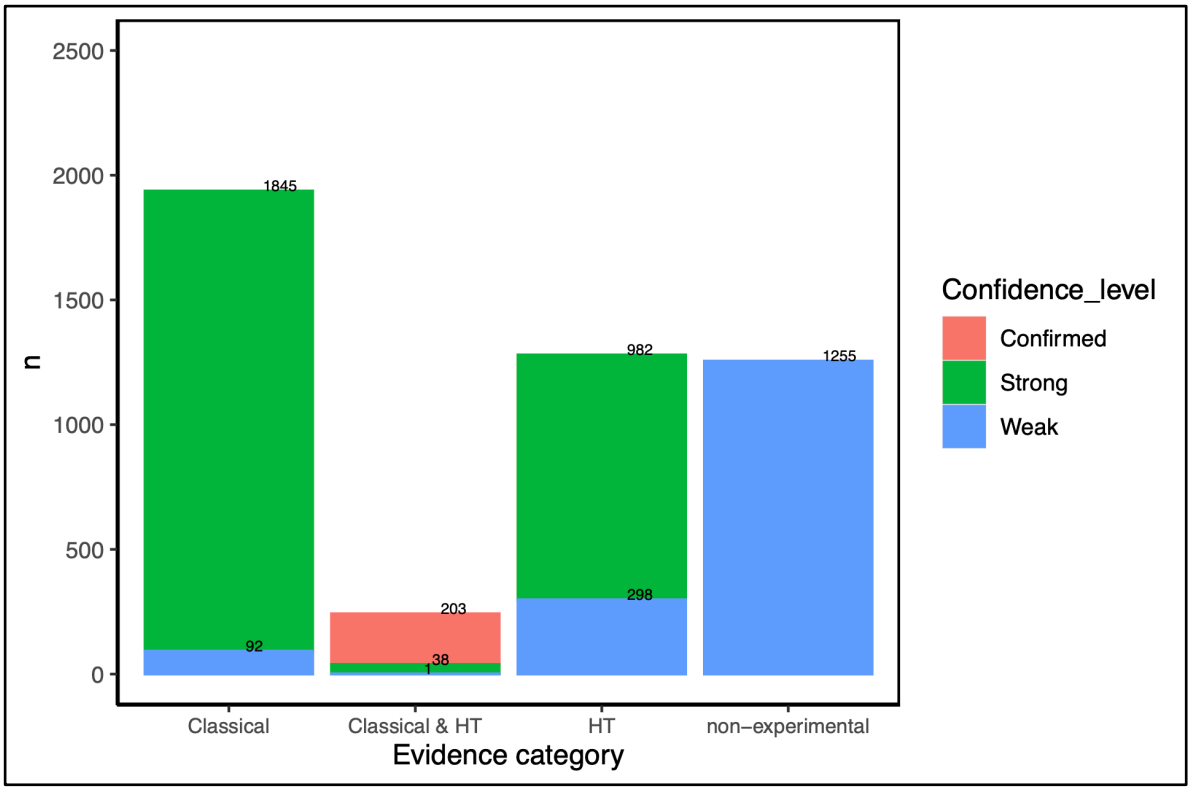

Figure 4. Number of RIs with classic, HT or non-experimental evidence and their confidence level highlighted in different colors.

Evidence types supporting current Ris

We graphed the number of RIs with each evidence type. Computational prediction, binding of purified proteins, ChIP-seq and site mutations are the binding evidence types with more RIs associated. On the other hand, gene expression analysis is the function evidence type with more RIs associated, followed by RNA-seq and expression microarrays (Figure 5). In Figure 5 we can see that binding of purified proteins with computational prediction is the combination with most RIs. However, when we analyze the evidence of RIs grouped by confidence levels different proportions are found (Figures 6-8).

Out of 769 RIs with a confirmed confidence level, 741 are supported by binding of the purified TF ("binding of purified proteins"), 653 have computational analysis and 617 have site mutations. And precisely the combination of these three binding-evidence types is the most frequent among the RIs with a confirmed level of confidence (Figure 6). Of 2,865 RIs with a strong confidence level, computational analysis is the most frequent evidence of binding with 2,151 RIs, followed by binding of purified proteins and ChIP-seq. The most frequent combination of binding-evidence types supporting RIs with a strong confidence level is the binding of purified proteins with computational analysis supporting 1,031 RIs, followed by ChIP-seq with computational analysis; and binding of purified protein only with 663 and 576 associated RIs, respectively (Figure 7).

We observed that most of RIs with a weak confidence level are supported by only one evidence (Figure 8), where the most frequent one is computational analysis, followed by ChIP-seq and binding of cellular extracts (Figure 8).

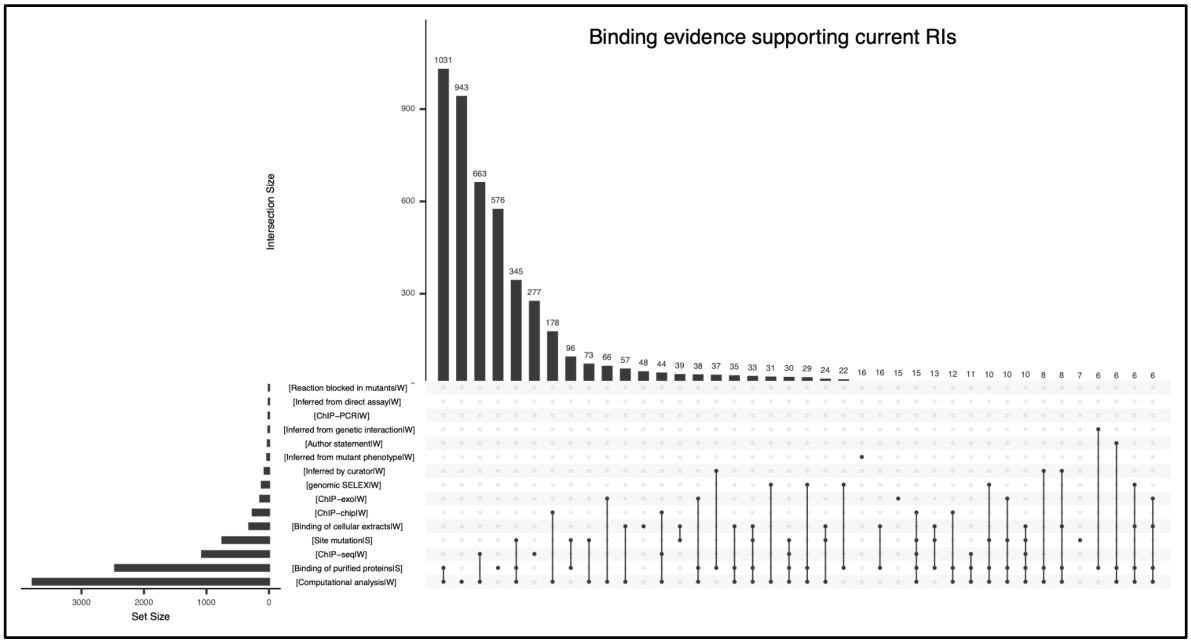

Figure 5. Combination of binding-evidence types supporting RIs. The number of RIs with each combination is indicated over the bar.

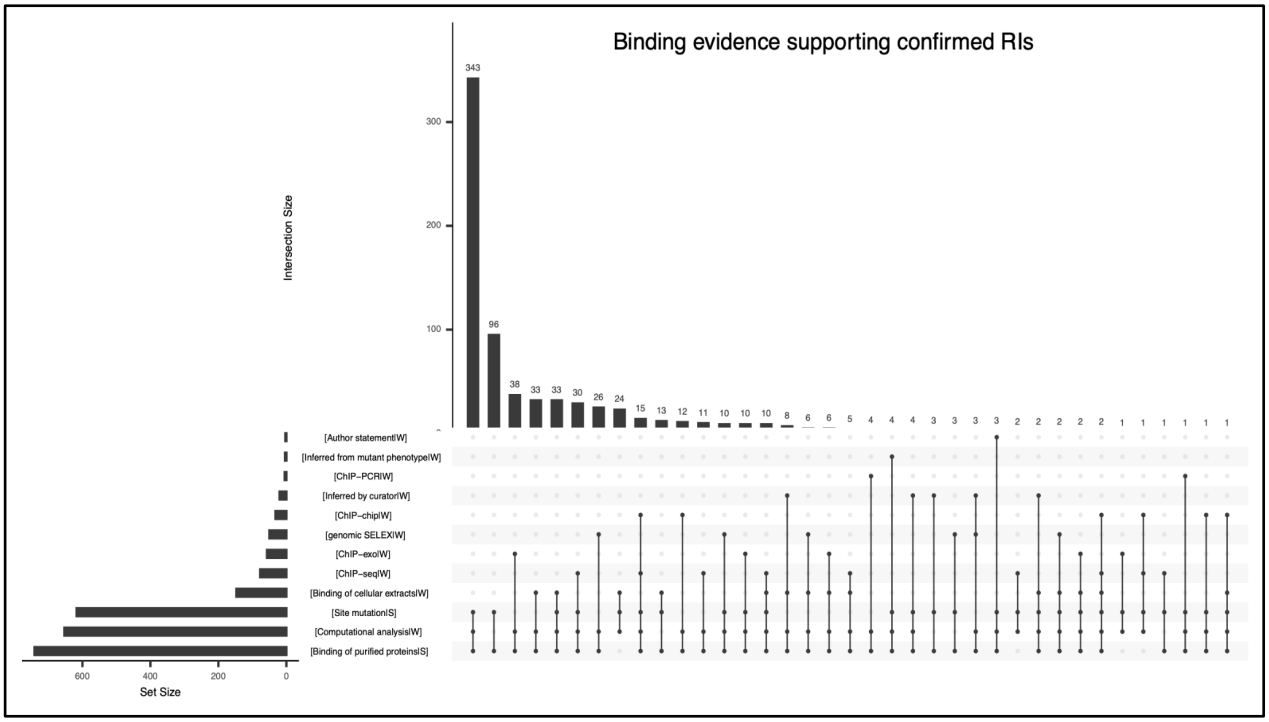

Figure 6. Combination of binding-evidence types supporting confirmed RIs. The number of RIs with each combination is indicated over the bar.

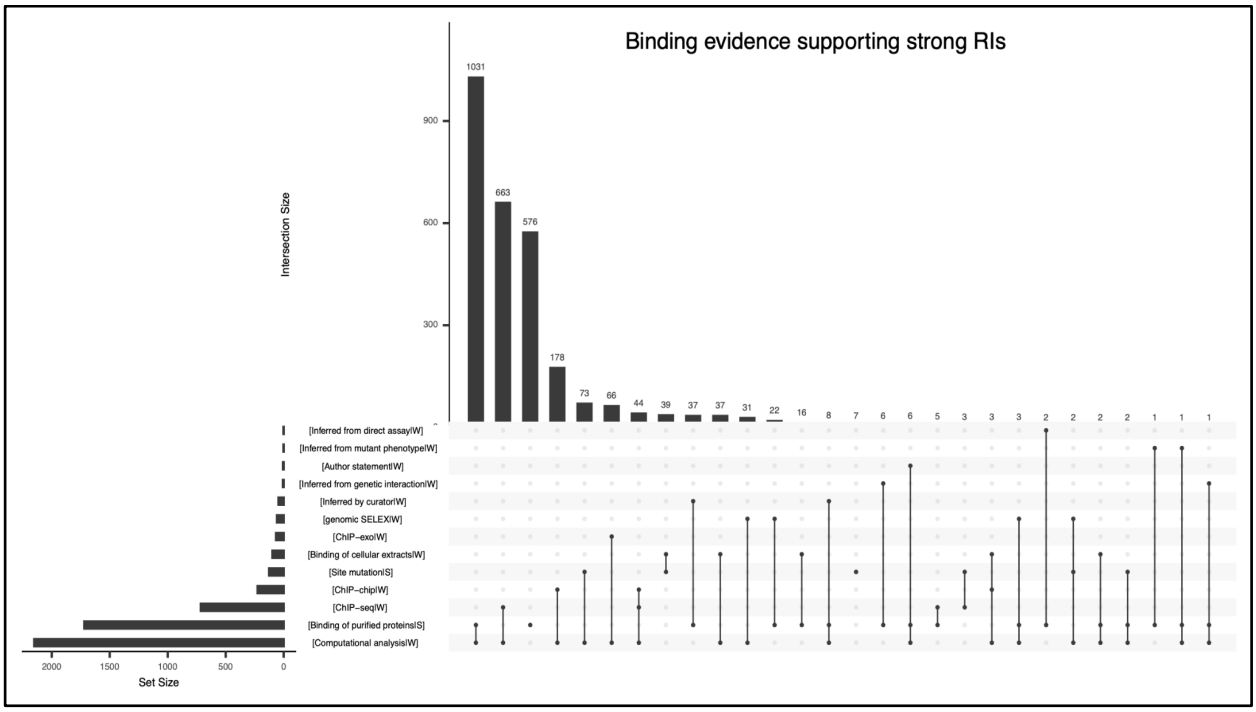

Figure 7. Combination of binding-evidence types supporting strong RIs. The number of RIs with each combination is indicated over the bar.

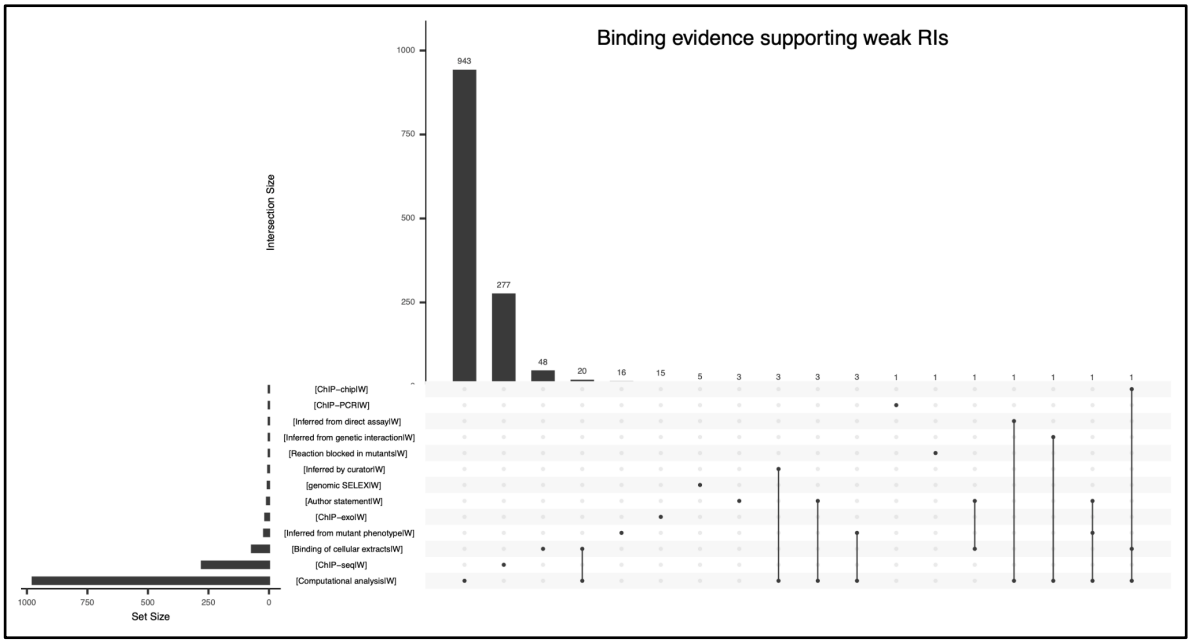

Figure 8. Combination of binding-evidence types supporting weak RIs. The number of RIs with each combination is indicated over the bar.

Dataset TF-gene interactions

As mentioned before, the dataset called network TF-gene interactions contains all the interactions described at this higher level of abstraction, that is to say, irrespective of the precise number of binding sites or RIs. All TF-gene-function interactions are derived from the current set of RIs, aggregating the three types of RIs (TF-promoter, TF-TU and TF-gene), and should not be confused with the specific RIs of TF-gene type implemented as mentioned, for purposes of a more precise curation. Those RIs curated as TF-TU, generate one TF-gene interaction for each gene in the TU; in the case of RIs curated at the TF-promoter level, one interaction is created for each gene in the TUs transcribed by the promoter; and finally, in the case of TF-gene RIs, it generates only one interaction.

As we have discussed in detail before, the level of confidence was assigned at the level of RIs, so the question arises of how to assign a confidence level when a TF-gene interaction may involve several RIs potentially with different confidence levels. We decided to assign the confidence level based on the RI with the highest confidence level among those supporting each interaction TF-gene-function. Figure 9 shows the number of TF-gene-function interactions with each confidence level. Currently, 17% of the interactions have confidence level confirmed, 48% strong and 35% weak.

Each interaction in the network dataset at the level of TF-gene is supported by one or several RIs.

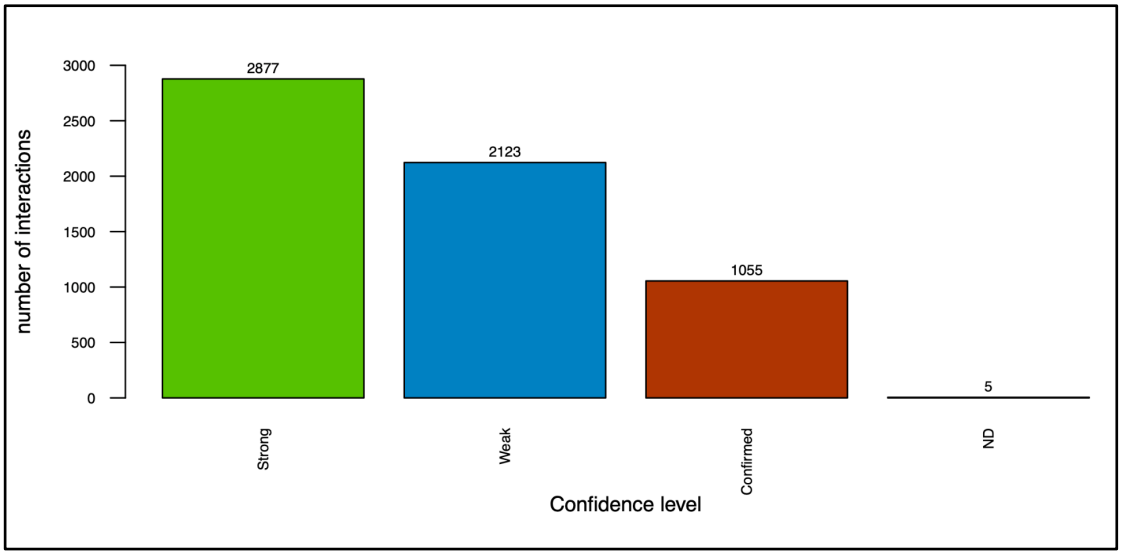

Figure 9. Confidence levels of interactions TF-gene-function in the Network TF-gene.

II.5. Golden standard options for users.

The value of comparing the results of new HT methods with the data in RegulonDB lies in the fact that RegulonDB contains the comprehensive collection of experiments performed during decades of classic methodologies. However, users may not be aware that we already have incorporated evidence from HT methods.

The current publicly available RegulonDB offers downloadable datasets grouping all RIs in https://regulondb.ccg.unam.mx/menu/download/datasets/index.jsp.
Columns 20 and 21 contain all the evidence types for binding and function of RIs, respectively. These can be used to filter and extract, for instance, the sub-collection supported only by classic methods. The same strategy could be used to select RIs supported by a specific evidence type. Furthermore, users can also sub–select RIs based on the confidence level, or on the different groups of methods as described in Figure 1. Furthermore, users may define their own rules and categories of different levels of confidence and use the whole collection of evidence types to classify each individual RI in a new classification of confidence levels.

IV. DISCUSSION.

Some decades ago, when knowledge about *E. coli* regulons was scarce, every piece of knowledge was relevant, even computational predictions or indirect inferences, were useful to glimpse the transcriptional regulation network governing gene regulation. Now, as a result of the constant accumulation of new knowledge, both classical knowledge and the explosion of HT data, it is both necessary and feasible to distinguish between the confidence levels and the evidence types behind each object, in this case of RIs. Similarly, we can distinguish RIs based on whether they were determined *in vivo* or *in vitro.*

Here we report the update of evidence codes types, their classification as classical, HT or non-experimental, and as *in vivo* or *in vitro*. Additionally, we update the confidence level to

each evidence type associated with RIs, and define rules for calculating the confidence level of the RIs according to the combinations of evidence types supporting them. Finally, we analyzed the distribution of RI types, confidence levels and evidence types in the current collection of RIs.

The evidence types codes were improved to facilitate their understanding and also to facilitate the identification of three major sources: classic experiments, HT methods, and non-experimental, such as computational predictions or author statements. This facilitates our current processing of all knowledge and allows filtering different collections of objects to cater to distinct user requirements. As mentioned before, for the sake of benchmarking HT methods, the knowledge obtained by classic methods of molecular biology available in RegulonDB is of paramount relevance to compare with.

The confidence levels weak or strong were assigned to each evidence type using specific criteria which may be subject to discussion. In the case of classical evidence, data comes from individual experiments focused on individual objects, so they were classified as strong, except the binding of cellular extracts which can be considered less specific, since such experiments do not eliminate the possibility of indirect effects. As already mentioned, HT binding-evidence types were classified as weak, since they involve several processing steps, including different bioinformatics options of methods and thresholds, making the final results more variable/dependent on the specific set of programs and variables used in their identification.
Non-experimental evidence types were also classified as weak, however, among them only computational analysis can be used in combination with other evidence types to upgrade the confidence level of the RIs associated, while author statements or inferences by curator do not. We know our classification can be improved and modified. For instance we do not assign a different confidence level to the *in vitro* vs the *in vivo* HT methodologies (i.e. DAP and gSELEX vs ChIP-seq and ChIP-exo). However, the data is there for such types of combinations.

As mentioned, the table at the level of TF-gene regulatory interactions available in the download files in RegulonDB merged all three RI types: TF-gene, TF-TU and TF-promoter which have different degrees of knowledge as described before. This is a complex table with many RIs with some blank columns as they contain or lack different components. However, all of them contain the minimum information to define an RI: the TF, the function or effect and the target (gene, TU or promoter). This file includes all evidence types supporting each RI (columns 20 and 21), as well as the RI confidence level. Depending on the method, users can include/exclude specific sub-collections based on the method and/or evidence types, and can also select subsets of RIs filtering by confidence levels. We plan in future versions to offer in RegulonDB mechanisms that facilitate the selection of different golden standard datasets based on either their confidence level, or on the classic vs HT experiment type.

We analyzed the composition of the current set of RIs, including the RI types, classical/HT and confidence levels proportions. We found that currently 75% of RIs are of the TF-promoter type, which was expected given that among these, 77% have experimental classical evidence, but 33% are not supported by experimental evidence. We consider that

this information should not be discarded, but depending on the purposes, users may delete those. For instance, we know that the classical evidence continues to dominate in the case of RIs. The HT methodologies have contributed mainly in the identification of new RIs, which we have curated as RIs TF-gene, because the precise regulated promoter has not been identified.

The update in confidence level assignment to RIs resulted in 769 RIs confirmed, 2,865 strong and 1,646 weak. The most RIs confirmed are of TF-promoter type. We found that all RIs with confidence level confirmed have at least one classical evidence.

One can imagine that the current pattern or anatomy of knowledge source will change in the near future as HT-experiments are likely to become the dominant source of knowledge of transcriptional regulation. It is important to know that until now HT evidence has only been assigned to those RIs whose evidence of binding and function was reported in the same publication, and therefore most likely in the same laboratory setting. As mentioned, an interesting possibility will be to add to existing RIs in RegulonDB and EcoCyc the binding HT evidence extracted from the comprehensive collections from different HT methods already gathered in RegulonDB v.11.0. This will, most likely, impact the overall confidence levels of the RI current set. A different issue will be to increase the number of RIs given such large collection of HT-supported binding sites that currently lack functional evidence.

We believe our combined work of redefining evidence codes and their combinations to define the confidence level, as well as the mechanism currently implemented to deal with partial knowledge of what constitutes an RI (the TF-promoter, TF-TU and TF-gene types), all together contribute to an increased in the quality of knowledge representation of what constitutes one of the best, if not the best, characterized transcriptional regulatory network of any microbial organism. In practice these improvements enable both RegulonDB and EcoCyc to incorporate HT-generated knowledge without "diluting" the valuable fraction of knowledge supported by classical molecular biology methods, since it is easy to dissect subsets based on their supporting methods. Furthermore, users can benefit from these improvements, by having a deeper understanding of the origin of the knowledge about transcriptional regulation encoded in RegulonDB and EcoCyc, and therefore should be able to make a better and more flexible use of it. Specifically, the results here presented facilitate a better use of this knowledge to benchmark results generated by existing or novel HT methods.
It is well known that the *E. coli* transcriptional regulatory network is one of the best characterized of any organism. It may well be also the best computationally represented corpus of knowledge of gene regulation.

METHODS

Updates in RegulonDB

Evidence update: In RegulonDB version 11.1, we made important evidence-related changes, including: 1) Evidence code. The evidence codes were made more informative, i.e. BPP was changed to EXP-IDA-BINDING-OF-PURIFIED-PROTEINS. 2) Evidence confidence level.

Evidence types were classified as “weak” or “strong” depending on whether they provide physical and direct proof of the existence of the object or interaction

Objects confidence level update: The confidence level for each regulatory interaction, promoter and transcription units was calculated and updated using their linked evidence and their additive evidence. The confidence level assignment to RIs is described below, we followed the same principles for the other objects. These changes are reflected in the RegulonDB interface as well as in the downloadable datasets.
Note that the availability of evidence types in the download datasets, as well as the more informative codes allows users to filter the evidence types as they wish, using or not using our flags of confidence level (weak, strong, confirmed).

Input for regulatory interactions.
The analysis was done using RegulonDB version 11.1 synchronized with Ecocyc version 26.0. In this version, the downloadable text file for Regulatory Interactions was made available and also the evidence catalog file. The formats and descriptions of these files are available at https://github.com/regulondbunam/download-data-files

Confidence level assignment to evidence types and to RIs
The confidence level was assigned to RIs by a process involving two stages:
Stage I: Each single evidence type was classified into weak or strong, based on the following criteria: The evidence types using classical methods were classified as strong since they provide physical direct proof of the existence of the interaction, except for “binding of cellular extracts”, classified as weak since it could involve indirect interactions. Evidence types from binding-HT category were classified as weak since they are not direct assays; on the contrary, HT- methods involve multiple steps of computational processing of the raw data. Computational evidence was considered weak. And “author statements” and “inferred by curator” are weak but do not contribute to confidence. Finally, evidence types involving gene expression provide evidence for the function (activation/repression) of TFs, a property different from binding. Thus, one such evidence is required for an RI to exist, but they do not contribute to the confidence levels which are focused on support for binding.
Stage II: Assignment of confidence level to RIs based on the set of their evidence types.
In this stage we use the concept “additive evidence” that in previous versions was called “cross-validation”. As was described by Weiss et al. (2013) (22), the confidence level of a biological entity depends on the combined evidence derived from mutually independent methods.
We grouped methods that could have similar sources of false positives. This resulted in seven independent evidence groups (Figure 1). The combinations of evidence groups that upgrade the RIs confidence levels were defined based on the four rules mentioned in section II.2. We call these combinations additive evidence, which define the final level of confidence assigned to each RI. The complete set of group-combinations that upgrades RIs confidence level can be found under the “regulatory interactions” of the “Stage II” section of the webpage: https://regulondb.ccg.unam.mx/evidenceclassification.

Analysis of current set of RIs

The analysis of RI types and associated evidence types, categories or approaches were performed using R (2022.06.23, version 4.2.1), Rstudio (2022.07.1, Build 554) and the ggplot2 (version 3.4.0) library. Scripts used can be found in: https://github.com/PGC-CCG/supplementary-material/tree/master/golden-standard

Acknowledgements.

We acknowledge suggestions by Suzanne Paley and Peter Karp on some evidence code descriptions; Pablo Peña Loredo for his help in software development; Shirley Alquicira Hernández, César Bonavides and Víctor Del Moral for IT support.

Funding

We acknowledge funding from Universidad Nacional Autónoma de México (UNAM), as well as funding by NIGMS-NIH grant number 5RO1GM131643. PL acknowledges a postdoctoral fellowship from DGAPA-UNAM.

Authors contributions:

P.L. Curation, redefinition of evidence codes, and their additive groups; analysis, writing, editing. S.G.C.: Curation, redefinition of evidence codes, and their additive groups, analysis, writing. H.S.: Software, validation, visualization, review and analysis. C.R. Validation, analysis, discussion. L.M.R.: Software for additive evidence, release process and downloading datasets. J.G.S.: Software for downloading datasets and evidence web pages. V.H.T. redefinition of evidence codes, and their additive groups; analysis. J.C.V.: Conceptualization, supervision, funding acquisition, review and writing.

REFERENCES

1. Tierrafría, V.H., Rioualen, C., Salgado, H., Lara, P., Gama-Castro, S., Lally, P., Gómez-Romero, L., Peña-Loredo, P., López-Almazo, A.G., Alarcón-Carranza, G. *et al.* (2022) RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in Escherichia coli K-12. *Microb Genom*, **8**, **PMID:** 35584008
2. Kahramanoglou, C., Seshasayee, A.S., Prieto, A.I., Ibberson, D., Schmidt, S., Zimmermann, J., Benes, V., Fraser, G.M. and Luscombe, N.M. (2011) Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic Acids Res*, **39**, 2073-2091, **PMID:** 21097887
3. Shimada, T., Yamamoto, K. and Ishihama, A. (2011) Novel members of the Cra regulon involved in carbon metabolism in Escherichia coli. *J Bacteriol*, **193**, 649-659, **PMID:** 21115656
4. Shimada, T., Fujita, N., Yamamoto, K. and Ishihama, A. (2011) Novel roles of cAMP receptor protein (CRP) in regulation of transport and metabolism of carbon sources. *PLoS One*, **6**, e20081, **PMID:** 21673794
5. Shimada, T., Bridier, A., Briandet, R. and Ishihama, A. (2011) Novel roles of LeuO in transcription regulation of E. coli genome: antagonistic interplay with the universal silencer H-NS. *Mol Microbiol*, **82**, 378-397, **PMID:** 21883529
6. Seo, S.W., Kim, D., Latif, H., O'Brien, E.J., Szubin, R. and Palsson, B.O. (2014) Deciphering Fur transcriptional regulatory network highlights its complex role

beyond iron metabolism in Escherichia coli. *Nat Commun*, **5**, 4910, **PMID:** 25222563
7. Shimada, T., Takada, H., Yamamoto, K. and Ishihama, A. (2015) Expanded roles of two-component response regulator OmpR in Escherichia coli: genomic SELEX search for novel regulation targets. *Genes Cells*, **20**, 915-931, **PMID:** 26332955
8. Ishihama, A., Shimada, T. and Yamazaki, Y. (2016) Transcription profile of Escherichia coli: genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Res*, **44**, 2058-2074, **PMID:** 26843427
9. Shimada, T., Saito, N., Maeda, M., Tanaka, K. and Ishihama, A. (2015) Expanded roles of leucine-responsive regulatory protein in transcription regulation of the Escherichia coli genome: Genomic SELEX screening of the regulation targets. *Microb Genom*, **1**, e000001, **PMID:** 28348809
10. Kim, D., Seo, S.W., Gao, Y., Nam, H., Guzman, G.I., Cho, B.K. and Palsson, B.O. (2018) Systems assessment of transcriptional regulation on central carbon metabolism by Cra and CRP. *Nucleic Acids Res*, **46**, 2901-2917, **PMID:** 29394395
11. Shimada, T., Ogasawara, H. and Ishihama, A. (2018) Single-target regulators form a minor group of transcription factors in Escherichia coli K-12. *Nucleic Acids Res*, **46**, 3921-3936, **PMID:** 29529243
12. Kroner, G.M., Wolfe, M.B. and Freddolino, P.L. (2019) Escherichia coli Lrp Regulates One-Third of the Genome via Direct, Cooperative, and Indirect Routes. *J Bacteriol*, **201**, **PMID:** 30420454
13. Anzai, T., Imamura, S., Ishihama, A. and Shimada, T. (2020) Expanded roles of pyruvate-sensing PdhR in transcription regulation of the Escherichia coli K-12 genome: fatty acid catabolism and cell motility. *Microb Genom*, **6**, **PMID:** 32975502
14. Choudhary, K.S., Kleinmanns, J.A., Decker, K., Sastry, A.V., Gao, Y., Szubin, R., Seif, Y. and Palsson, B.O. (2020) Elucidation of Regulatory Modes for Five Two-Component Systems in Escherichia coli Reveals Novel Relationships. *mSystems*, **5**, **PMID:** 33172971
15. Ishihama, A. and Shimada, T. (2021) Hierarchy of transcription factor network in Escherichia coli K-12: H-NS-mediated silencing and Anti-silencing by global regulators. *FEMS Microbiol Rev*, **45**, **PMID:** 34196371
16. Shimada, T., Furuhata, S. and Ishihama, A. (2021) Whole set of constitutive promoters for RpoN sigma factor and the regulatory role of its enhancer protein NtrC in Escherichia coli K-12. *Microb Genom*, **7**, **PMID:** 34787538
17. Baumgart, L.A., Lee, J.E., Salamov, A., Dilworth, D.J., Na, H., Mingay, M., Blow, M.J., Zhang, Y., Yoshinaga, Y., Daum, C.G. *et al.* (2021) Persistence and plasticity in bacterial gene regulation. *Nat Methods*, **18**, 1499-1505, **PMID:** 34824476
18. Colvill, A.J., Kanner, L.C., Tocchini-Valentini, G.P., Sarnat, M.T. and Geiduschek, E.P. (1965) Asymmetric RNA synthesis in vitro: heterologous DNA-enzyme systems; E. coli RNA polymerase. *Proc Natl Acad Sci U S A*, **53**, 1140-1147, **PMID:** 4958034
19. Jacob, F. and Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, **3**, 318-356, **PMID:** 13718526
20. Chamberlin, M. and Berg, P. (1962) Deoxyribo ucleic acid-directed synthesis of ribonucleic acid by an enzyme from Escherichia coli. *Proc Natl Acad Sci U S A*, **48**, 81-94, **PMID:** 13877961

21. Mejía-Almonte, C., Busby, S.J.W., Wade, J.T., van Helden, J., Arkin, A.P., Stormo, G.D., Eilbeck, K., Palsson, B.O., Galagan, J.E. and Collado-Vides, J. (2020) Redefining fundamental concepts of transcription initiation in bacteria. *Nat Rev Genet*, **21**, 699-714, **PMID:** 32665585
22. Weiss, V., Medina-Rivera, A., Huerta, A.M., Santos-Zavaleta, A., Salgado, H., Morett, E. and Collado-Vides, J. (2013) Evidence classification of high-throughput protocols and confidence integration in RegulonDB. *Database (Oxford)*, **2013**, bas059, **PMID:** 23327937
23. Baseggio, N., Davies, W.D. and Davidson, B.E. (1990) Identification of the promoter, operator, and 5' and 3' ends of the mRNA of the Escherichia coli K-12 gene aroG. *J Bacteriol*, **172**, 2547-2557, **PMID:** 1970563
24. Venkatesh, G.R., Kembou Koungni, F.C., Paukner, A., Stratmann, T., Blissenbach, B. and Schnetz, K. (2010) BglJ-RcsB heterodimers relieve repression of the Escherichia coli bgl operon by H-NS. *J Bacteriol*, **192**, 6456-6464, **PMID:** 20952573
25. Schneiders, T., Barbosa, T.M., McMurry, L.M. and Levy, S.B. (2004) The Escherichia coli transcriptional regulator MarA directly represses transcription of purA and hdeA. *J Biol Chem*, **279**, 9037-9042, **PMID:** 14701822